Proving Correctness of Compilers Using Structured Graphs

Patrick Bahr

Department of Computer Science, University of Copenhagen, Denmark
paba@di.ku.dk

Abstract. We present an approach to compiler implementation using Oliveira and Cook's structured graphs that avoids the use of explicit jumps in the generated code. The advantage of our method is that it takes the implementation of a compiler using a tree type along with its correctness proof and turns it into a compiler implementation using a graph type along with a correctness proof. The implementation and correctness proof of a compiler using a tree type without explicit jumps is simple, but yields code duplication. Our method provides a convenient way of improving such a compiler without giving up the benefits of simple reasoning.

1 Introduction

Verification of compilers – like other software – is difficult [13]. In such an endeavour one typically has to balance the "cleverness" of the implementation with the simplicity of reasoning about it. A concrete example of this fact is given by Hutton and Wright [10] who present correctness proofs of compilers for a simple language with exceptions. The authors first present a naïve compiler implementation that produces a tree representing the possible control flow of the input program. The code that it produces is essentially the right code, but the compiler *loses information* since it duplicates code instead of sharing it. However, the simplicity of the implementation is matched with a clean and simple proof by equational reasoning. Hutton and Wright also present a more realistic compiler, which uses labels and explicit jumps, resulting in a target code in linear form and without code duplication. However, the cleverer implementation also requires a more complicated proof, in which one has to reason about the freshness and scope of labels.

In this paper we present an intermediate approach, which is still simple, both in its implementation and in its correctness proof, but which avoids the loss of information of the simple approach described by Hutton and Wright [10]. The remedy for the information loss of the simple approach is obvious: we use a graph instead of a tree structure to represent the target code. The linear representation with labels and jumps is essentially a graph as well – it is just a very inconvenient one for reasoning. Instead of using unique names to represent sharing, we use the *structured graphs* representation of Oliveira and Cook [18]. This representation of graphs uses parametric higher-order abstract syntax [5] to represent binders, which in turn are used to represent sharing. This structure allows us to take the simple compiler implementation using trees, make a slight adjustment to it, and obtain a compiler implementation using graphs that preserves the sharing information.

In essence our approach teases apart two aspects that are typically combined in code generation: (1) the translation into the target language, and (2) generating fresh (label) names for representing jumps in the target language. By keeping the two aspects separate, we can implement further transformations, e.g. code optimisations, without having to deal with explicit jumps and names. Only in the final step, when the code is linearised, names have to be generated in order to produce explicit jump instructions. Consequently, the issues that ensue in this setting can be dealt with in isolation – separately from the actual translation and subsequent transformation steps.

Our main goal is to retain the simplicity of the correctness proof of the tree-based compiler. The key observation making this possible is that the semantics of the tree-based and the graph-based target language, i.e. their respective *virtual machines*, are equivalent after *unravelling* of the graph structure. More precisely, given the semantics of the tree-based and the graph-based target language as $exec_{\mathsf{T}}$ and $exec_{\mathsf{G}}$, respectively, we have the following equation:

$$exec_{\mathsf{G}} = exec_{\mathsf{T}} \circ unravel$$

We show that this correspondence is an inherent consequence of the recursion schemes that are used to define these semantics. In fact, this correspondence follows from the correctness of *short cut fusion* [8, 12]. That is, the above property is independent of the target language of the compiler. As a consequence, the correctness proof of the improved, graph-based compiler is reduced to a proof that its implementation is equivalent to the tree-based implementation modulo unravelling. More precisely, it then suffices to show that

$$comp_{\mathsf{T}} = unravel \circ comp_{\mathsf{G}}$$

which is achieved by a straightforward induction proof.

In sum, the technique that we propose here improves existing simple compiler implementations to more realistic ones using a graph representation for the target code. This improvement requires minimal effort – both in terms of the implementation and the correctness proof. The fact that we consider both the implementation and its correctness proof makes our technique the ideal companion to improve a compiler that has been obtained by calculation [16]. Such calculations derive a compiler from a specification, and produce not only an implementation of the compiler but also a proof of its correctness. The example compiler that we use in this paper has in fact been calculated in this way by Bahr and Hutton [3], and we have successfully applied our technique to other compilers derived by Bahr and Hutton [3], which includes compilers for languages with features such as (synchronous and asynchronous) exceptions, (global and local) state and non-determinism. Thus, despite its simplicity, our technique is quite powerful, especially when combined with other techniques such as the abovementioned calculation techniques.

In short, the contributions of this paper are the following:

- From a compiler with code duplication we derive a compiler that avoids duplication using a graph representation.
- Using short cut fusion, we prove that folds over graphs are equal to corresponding folds over the unravelling of the input graphs.
- Using the above result, we derive the correctness of the graph-based compiler implementation from the correctness of the tree-based compiler.
- We further simplify the proof by using free monads to represent tree types together with a corresponding monadic graph type.

Throughout this paper we use Haskell [14] as the implementation language.

2 A Simple Compiler

The example language that we use throughout the paper is a simple expression language with integers, addition and exceptions:

$$
\begin{array}{l}
\textbf{data}\ \mathit{Expr} = \mathit{Val}\ \mathit{Int} \mid \mathit{Add}\ \ \mathit{Expr}\ \mathit{Expr} \\
\qquad\qquad\quad \mid \mathit{Throw}\ \mid \mathit{Catch}\ \mathit{Expr}\ \mathit{Expr}
\end{array}
$$

The semantics of this language is defined using an evaluation function that evaluates a given expression to an integer value or returns *Nothing* in case of an uncaught exception:

$$
\begin{array}{lll}
\multicolumn{3}{l}{\mathit{eval} :: \mathit{Expr} \to \mathit{Maybe}\ \mathit{Int}} \\
\mathit{eval}\ (\mathit{Val}\ n) & = & \mathit{Just}\ n \\
\mathit{eval}\ (\mathit{Add}\ x\ y) & = & \textbf{case}\ \mathit{eval}\ x\ \textbf{of} \\
 & & \quad \mathit{Nothing} \to \mathit{Nothing} \\
 & & \quad \mathit{Just}\ n \ \ \to \textbf{case}\ \mathit{eval}\ y\ \textbf{of} \\
 & & \qquad\qquad\qquad \mathit{Nothing} \to \mathit{Nothing} \\
 & & \qquad\qquad\qquad \mathit{Just}\ m \ \ \to \mathit{Just}\ (n + m) \\
\mathit{eval}\ \ \mathit{Throw} & = & \mathit{Nothing} \\
\mathit{eval}\ (\mathit{Catch}\ x\ h) & = & \textbf{case}\ \mathit{eval}\ x\ \textbf{of} \\
 & & \quad \mathit{Nothing} \to \mathit{eval}\ h \\
 & & \quad \mathit{Just}\ n \ \ \to \mathit{Just}\ n
\end{array}
$$

This is the same language and semantics used by Hutton and Wright [10]. Like Hutton and Wright, we chose a simple language in order to focus on the essence of the problem, which in our case is control flow in the target language and the use of duplication or sharing to represent it. Moreover, this choice allows us to compare our method to the original work of Hutton and Wright whose focus was on the simplicity of reasoning.

The target for the compiler is a simple stack machine with the following instruction set:

$$
\begin{array}{llll}
\textbf{data}\ \mathit{Code} = & \mathit{PUSH}\ \mathit{Int}\ \mathit{Code} & \mid \mathit{ADD}\ \mathit{Code} & \mid \mathit{HALT} \\
 & \mid\ \mathit{UNMARK}\ \mathit{Code} & \mid \mathit{MARK}\ \mathit{Code}\ \mathit{Code} & \mid \mathit{THROW}
\end{array}
$$

The intended semantics (which is made precise later) for the instructions is:

– $\mathit{PUSH}\ n$ pushes the integer value n on the stack,
– ADD expects two integers on the stack and replaces them with their sum,
– $\mathit{MARK}\ c$ pushes the exception handler code c on the stack,
– UNMARK removes such a handler code from the stack,
– THROW unwinds the stack until an exception handler code is found, which is then executed, and
– HALT stops the execution.

For the implementation of the compiler we deviate slightly from the presentation of Hutton and Wright [10] and instead write the compiler in a style that uses an additional accumulation parameter c, which simplifies the proofs [9]:

$$
\begin{array}{ll}
\multicolumn{2}{l}{\mathit{comp}^{\mathsf{A}} :: \mathit{Expr} \to \mathit{Code} \to \mathit{Code}} \\
\mathit{comp}^{\mathsf{A}}\ (\mathit{Val}\ n) & c = \mathit{PUSH}\ n\ c \\
\mathit{comp}^{\mathsf{A}}\ (\mathit{Add}\ x\ y) & c = \mathit{comp}^{\mathsf{A}}\ x\ (\mathit{comp}^{\mathsf{A}}\ y\ (\mathit{ADD}\ c)) \\
\mathit{comp}^{\mathsf{A}}\ \mathit{Throw} & c = \mathit{THROW} \\
\mathit{comp}^{\mathsf{A}}\ (\mathit{Catch}\ x\ h) & c = \mathit{MARK}\ (\mathit{comp}^{\mathsf{A}}\ h\ c)\ (\mathit{comp}^{\mathsf{A}}\ x\ (\mathit{UNMARK}\ c))
\end{array}
$$

Since the code generator is implemented in this code continuation passing style, function application corresponds to concatenation of code fragments. To stress this reading, we shall use the operator $\triangleright$, which is simply defined as function application and is declared to associate to the right with minimal precedence:

$$
\begin{array}{l}
(\triangleright) :: (a \to b) \to a \to b \\
f \triangleright x = f\ x
\end{array}
$$

For instance, the equation for the Add case of the definition of $\mathit{comp}^{\mathsf{A}}$ then reads:

$$
\mathit{comp}^{\mathsf{A}}\ (\mathit{Add}\ x\ y)\ c = \mathit{comp}^{\mathsf{A}}\ x \triangleright \mathit{comp}^{\mathsf{A}}\ y \triangleright \mathit{ADD} \triangleright c
$$

To obtain the final code for an expression, we supply HALT as the initial value of the accumulator of $\mathit{comp}^{\mathsf{A}}$. The use of the $\triangleright$ operator to supply the argument indicates the intuition that HALT is placed at the end of the code produced by $\mathit{comp}^{\mathsf{A}}$:

$$
\begin{array}{l}
\mathit{comp} :: \mathit{Expr} \to \mathit{Code} \\
\mathit{comp}\ e = \mathit{comp}^{\mathsf{A}}\ e \triangleright \mathit{HALT}
\end{array}
$$

The following examples illustrate the workings of the compiler comp:

$$
\begin{array}{ll}
\mathit{comp}\ (\mathit{Add}\ (\mathit{Val}\ 2)\ (\mathit{Val}\ 3)) & \rightsquigarrow \mathit{PUSH}\ 2 \triangleright \mathit{PUSH}\ 3 \triangleright \mathit{ADD} \triangleright \mathit{HALT} \\
\mathit{comp}\ (\mathit{Catch}\ (\mathit{Val}\ 2)\ (\mathit{Val}\ 3)) & \rightsquigarrow \mathit{MARK}\ (\mathit{PUSH}\ 3 \triangleright \mathit{HALT}) \\
 & \qquad \triangleright \mathit{PUSH}\ 2 \triangleright \mathit{UNMARK} \triangleright \mathit{HALT} \\
\mathit{comp}\ (\mathit{Catch}\ \mathit{Throw}\ (\mathit{Val}\ 3)) & \rightsquigarrow \mathit{MARK}\ (\mathit{PUSH}\ 3 \triangleright \mathit{HALT}) \triangleright \mathit{THROW}
\end{array}
$$

For the virtual machine that executes the code produced by the above compiler, we use the following type for the stack:

$$
\begin{array}{l}
\textbf{type}\ \mathit{Stack} = [\mathit{Item}] \\
\textbf{data}\ \mathit{Item}\ \ = \mathit{VAL}\ \mathit{Int} \mid \mathit{HAN}\ (\mathit{Stack} \to \mathit{Stack})
\end{array}
$$

This type deviates slightly from the one for the virtual machine defined by Hutton and Wright [10]. Instead of having the code of an exception handler on the stack (constructor *HAN*), we have the continuation of the virtual machine on the stack. This will simplify the proof as we shall see later on. However, this type and the accompanying definition of the virtual machine that is given below is exactly the result of the calculation given by Bahr and Hutton [3] just before the last calculation step (which then yields the virtual machine of Hutton and Wright [10]). The virtual machine that works on this stack is defined as follows:

$$
\begin{array}{ll}
\mathit{exec} :: \mathit{Code} \to \mathit{Stack} \to \mathit{Stack} & \\
\mathit{exec}\ (\mathit{PUSH}\ n\ c) & s = \mathit{exec}\ c\ (\mathit{VAL}\ n : s) \\
\mathit{exec}\ (\mathit{ADD}\ c) & s = \textbf{case}\ s\ \textbf{of} \\
 & \qquad (\mathit{VAL}\ m : \mathit{VAL}\ n : t) \to \mathit{exec}\ c\ (\mathit{VAL}\ (n+m) : t) \\
\mathit{exec}\ \mathit{THROW} & s = \mathit{unwind}\ s \\
\mathit{exec}\ (\mathit{MARK}\ h\ c) & s = \mathit{exec}\ c\ (\mathit{HAN}\ (\mathit{exec}\ h) : s) \\
\mathit{exec}\ (\mathit{UNMARK}\ c) & s = \textbf{case}\ s\ \textbf{of}\ (x : \mathit{HAN}\ \_ : t) \to \mathit{exec}\ c\ (x : t) \\
\mathit{exec}\ \mathit{HALT} & s = s
\end{array}
$$

$$
\begin{array}{ll}
\mathit{unwind} :: \mathit{Stack} \to \mathit{Stack} & \\
\mathit{unwind}\ [\,] & = [\,] \\
\mathit{unwind}\ (\mathit{VAL}\ \_ : s) & = \mathit{unwind}\ s \\
\mathit{unwind}\ (\mathit{HAN}\ h : s) & = h\ s
\end{array}
$$

The virtual machine does what is expected from the informal semantics that we have given above. The semantics of *MARK*, however, may seem counterintuitive at first: as mentioned above, *MARK* does not put the handler code on the stack but rather the continuation that is obtained by executing it. Consequently, when the unwinding of the stack reaches a handler h on the stack, this handler h is directly applied to the remainder of the stack. This slight deviation from the semantics of Hutton and Wright [10] makes sure that *exec* is in fact a fold.

We will not go into the details of the correctness proof for the compiler *comp*. One can show that it satisfies the following correctness property [3]:

Theorem 1 (compiler correctness).

$$\mathit{exec}\ (\mathit{comp}\ e)\ [\,] = \mathit{conv}\ (\mathit{eval}\ e) \qquad \textit{for all } e :: \mathit{Expr}$$

$$
\textit{where} \quad
\begin{array}{l}
\mathit{conv}\ (\mathit{Just}\ n) = [\,\mathit{Val}\ n] \\
\mathit{conv}\ \mathit{Nothing} = [\,]
\end{array}
$$

That is, in particular, we have that

$$\mathit{exec}\ (\mathit{comp}\ e)\ [\,] = [\,\mathit{Val}\ n] \quad \iff \quad \mathit{eval}\ e = \mathit{Just}\ n$$

While the compiler has the nice property that it can be derived from the language semantics, the code that it produces is quite unrealistic. Note the duplication that occurs for generating the code for *Catch*: the continuation code c is inserted both after the handler code (in $comp^{\mathsf{A}}\ h\ c$) and after the *UNMARK* instruction. This is necessary since the code c may have to be executed regardless whether an exception is thrown in the scope x of the *Catch* or not.

This duplication can be avoided by using explicit jumps in the code. Instead of duplicating code, jumps to a single copy of the code are inserted. However, this complicates both the implementation of the compiler and its correctness proof [10]. Also the derivation of such a compiler by calculation is equally cumbersome.

The approach that we suggest in this paper takes the above compiler and derives a slightly different variant that instead of a tree structure produces a graph structure. Along with the compiler we derive a virtual machine that also works on the graph structure. The two variants of the compiler and its companion virtual machine only differ in the sharing that the graph variant provides. This fact allows us to derive the correctness of the graph-based compiler very easily from the correctness of the original tree-based compiler.

3 From Trees to Graphs

Before we derive the graph-based compiler and the corresponding virtual machine, we restructure the definition of the original compiler and the corresponding virtual machine. This will smoothen the process and simplify the presentation.

3.1 Preparations

Instead of defining the type *Code* directly, we represent it as the initial algebra of a functor. To distinguish this representation from the graph representation we introduce later, we use the name *Tree* for the initial algebra construction.

$$\mathbf{data}\ \mathit{Tree}\ f = \mathit{In}\ (f\ (\mathit{Tree}\ f))$$

The functor that induces the initial algebra that we shall use for representing the target language is easily obtained from the original *Code* data type:

$$\begin{aligned}\mathbf{data}\ \mathit{Code}_{\mathsf{F}}\ a = {} & \mathit{PUSH}_{\mathsf{F}}\ \mathit{Int}\ a \mid \mathit{ADD}_{\mathsf{F}}\ a \mid \mathit{HALT}_{\mathsf{F}} \\ \mid {} & \mathit{MARK}_{\mathsf{F}}\ a\ a \mid \mathit{UNMARK}_{\mathsf{F}}\ a \mid \mathit{THROW}_{\mathsf{F}}\end{aligned}$$

The type representing the target code is thus $\mathit{Tree}\ \mathit{Code}_{\mathsf{F}}$, which is isomorphic to *Code* modulo non-strictness. We proceed by reformulating the definition of *comp* to work on the type $\mathit{Tree}\ \mathit{Code}_{\mathsf{F}}$ instead of *Code*:

$$\begin{aligned}&\mathit{comp}^{\mathsf{A}}_{\mathsf{T}} :: \mathit{Expr} \to \mathit{Tree}\ \mathit{Code}_{\mathsf{F}} \to \mathit{Tree}\ \mathit{Code}_{\mathsf{F}} \\ &\mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ (\mathit{Val}\ n) \quad\ \ c = \mathit{PUSH}_{\mathsf{T}}\ n \triangleright c \\ &\mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ (\mathit{Add}\ x\ y) \quad c = \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ x \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ y \triangleright \mathit{ADD}_{\mathsf{T}} \triangleright c\end{aligned}$$

$$
\begin{array}{l}
comp^{\mathsf{A}}_{\mathsf{T}}\ \ Throw \qquad c = THROW_{\mathsf{T}} \\
comp^{\mathsf{A}}_{\mathsf{T}}\ (Catch\ x\ h)\ c = MARK_{\mathsf{T}}\ (comp^{\mathsf{A}}_{\mathsf{T}}\ h \triangleright c) \triangleright comp^{\mathsf{A}}_{\mathsf{T}}\ x \triangleright UNMARK_{\mathsf{T}} \triangleright c \\
comp_{\mathsf{T}} :: Expr \to Tree\ Code_{\mathsf{F}} \\
comp_{\mathsf{T}}\ e = comp^{\mathsf{A}}_{\mathsf{T}}\ e \triangleright HALT_{\mathsf{T}}
\end{array}
$$

Note that we do not use the constructors of $Code_{\mathsf{F}}$ directly, but instead we use *smart constructors* that also apply the constructor *In* of the type constructor *Tree*. These smart constructors serve as drop-in replacements for the constructors of the original *Code* data type. For example, $PUSH_{\mathsf{T}}$ is defined as follows:

$$
\begin{array}{l}
PUSH_{\mathsf{T}} :: Int \to Tree\ Code_{\mathsf{F}} \to Tree\ Code_{\mathsf{F}} \\
PUSH_{\mathsf{T}}\ i\ c = In\ (PUSH_{\mathsf{F}}\ i\ c)
\end{array}
$$

Lastly, we also reformulate the semantics of the target language, i.e. we define the function *exec* on the type $Tree\ Code_{\mathsf{F}}$. To do this, we use the following definition of a fold on an initial algebra:

$$
\begin{array}{l}
fold :: Functor\ f \Rightarrow (f\ r \to r) \to Tree\ f \to r \\
fold\ alg\ (In\ t) = alg\ (fmap\ (fold\ alg)\ t)
\end{array}
$$

The definition of the semantics is a straightforward transcription of the definition of *exec* into an algebra:

$$
\begin{array}{lll}
\multicolumn{3}{l}{execAlg :: Code_{\mathsf{F}}\ (Stack \to Stack) \to Stack \to Stack} \\
execAlg\ (PUSH_{\mathsf{F}}\ n\ c) & s = & c\ (VAL\ n : s) \\
execAlg\ (ADD_{\mathsf{F}}\ c) & s = & \mathbf{case}\ s\ \mathbf{of} \\
 & & \quad (VAL\ m : VAL\ n : t) \to c\ (VAL\ (n+m) : t) \\
execAlg\ THROW_{\mathsf{F}} & s = & unwind\ s \\
execAlg\ (MARK_{\mathsf{F}}\ h\ c) & s = & c\ (HAN\ h : s) \\
execAlg\ (UNMARK_{\mathsf{F}}\ c) & s = & \mathbf{case}\ s\ \mathbf{of}\ (x : HAN\ \_ : t) \to c\ (x : t) \\
execAlg\ HALT_{\mathsf{F}} & s = & s \\
\multicolumn{3}{l}{exec_{\mathsf{T}} :: Tree\ Code_{\mathsf{F}} \to Stack \to Stack} \\
\multicolumn{3}{l}{exec_{\mathsf{T}} = fold\ execAlg}
\end{array}
$$

From the correctness of the original compiler from Section 2, as expressed in Theorem 1, we obtain the correctness of our reformulation of the implementation:

Corollary 1 (correctness of $comp_{\mathsf{T}}$).

$$
exec_{\mathsf{T}}\ (comp_{\mathsf{T}}\ e)\ [\,] = conv\ (eval\ e) \qquad \textit{for all } e :: Expr
$$

Proof. Let $\phi :: Code \to Tree\ Code_{\mathsf{F}}$ be the function that recursively maps each constructor of *Code* to the corresponding smart constructor of $Tree\ Code_{\mathsf{F}}$. We can easily check that $comp_{\mathsf{T}}$ and $exec_{\mathsf{T}}$ are equivalent to the original functions *comp* respectively *exec* via ϕ, i.e.

$$
comp_{\mathsf{T}} = \phi \circ comp \quad \text{and} \quad exec_{\mathsf{T}} \circ \phi = exec
$$

Consequently, we have that $exec_{\mathsf{T}} \circ comp_{\mathsf{T}} = exec \circ comp$, and thus the corollary follows from Theorem 1. □

3.2 Deriving a Graph-Based Compiler

Finally, we turn to the graph-based implementation of the compiler. Essentially, this implementation is obtained from $comp_{\mathsf{T}}$ by replacing the type $Tree\ Code_{\mathsf{F}}$ with a type $Graph\ Code_{\mathsf{F}}$, which instead of a tree structure has a graph structure, and using explicit sharing instead of duplication.

In order to define graphs over a functor, we use the representation of Oliveira and Cook [18] called *structured graphs*. Put simply, a structured graph is a tree with added sharing facilitated by let bindings. In turn, let bindings are represented using parametric higher-order abstract syntax [5].

$$
\begin{aligned}
\mathbf{data}\ Graph'\ f\ v = {} & GIn\ (f\ (Graph'\ f\ v)) \\
\mid {} & Let\ (Graph'\ f\ v)\ (v \to Graph'\ f\ v) \\
\mid {} & Var\ v
\end{aligned}
$$

The first constructor has the same structure as the constructor of the $Tree$ type constructor. The other two constructors will allow us to express let bindings: $Let\ g\ (\lambda x \to h)$ binds g to the metavariable x in h. Metavariables bound in a let binding have type v; the only way to use them is with the constructor Var. To enforce this invariant, the type variable v is made polymorphic:

$$\mathbf{newtype}\ Graph\ f = MkGraph\ (\forall\ v\ .\ Graph'\ f\ v)$$

We shall use the type constructor $Graph$ (and $Graph'$) as a replacement for $Tree$. For the purposes of our compiler we only need acyclic graphs. That is why we only consider non-recursive let bindings as opposed to the more general structured graphs of Oliveira and Cook [18]. This restriction to non-recursive let bindings is crucial for the reasoning principle that we use to prove correctness.

We can use the graph type almost as a drop-in replacement for the tree type. The only thing that we need to do is to use smart constructors that use the constructor GIn instead of In, e.g.

$$
\begin{aligned}
& PUSH_{\mathsf{G}} :: Int \to Graph'\ Code_{\mathsf{F}}\ v \to Graph'\ Code_{\mathsf{F}}\ v \\
& PUSH_{\mathsf{G}}\ i\ c = GIn\ (PUSH_{\mathsf{F}}\ i\ c)
\end{aligned}
$$

From the type of the smart constructors we can observe that graphs are constructed using the type constructor $Graph'$, not $Graph$. Only after the construction of the graph is completed, the constructor $MkGraph$ is applied in order to obtain a graph of type $Graph\ Code_{\mathsf{F}}$.

The definition of $comp^{\mathsf{A}}_{\mathsf{T}}$ can be transcribed into graph style by simply using the abovementioned smart constructors instead:

$$
\begin{aligned}
& comp^{\mathsf{A}}_{\mathsf{G}} :: Expr \to Graph'\ Code_{\mathsf{F}}\ a \to Graph'\ Code_{\mathsf{F}}\ a \\
& comp^{\mathsf{A}}_{\mathsf{G}}\ (Val\ n) && c = PUSH_{\mathsf{G}}\ n \triangleright c \\
& comp^{\mathsf{A}}_{\mathsf{G}}\ (Add\ x\ y) && c = comp^{\mathsf{A}}_{\mathsf{G}}\ x \triangleright comp^{\mathsf{A}}_{\mathsf{G}}\ y \triangleright ADD_{\mathsf{G}} \triangleright c \\
& comp^{\mathsf{A}}_{\mathsf{G}}\ (Throw) && c = THROW_{\mathsf{G}} \\
& comp^{\mathsf{A}}_{\mathsf{G}}\ (Catch\ x\ h) && c = MARK_{\mathsf{G}}\ (comp^{\mathsf{A}}_{\mathsf{G}}\ h \triangleright c) \triangleright comp^{\mathsf{A}}_{\mathsf{G}}\ x \triangleright UNMARK_{\mathsf{G}} \triangleright c
\end{aligned}
$$

The above is a one-to-one transcription of $comp^{\mathsf{A}}_{\mathsf{T}}$. But this is not what we want. We want to make use of the fact that the target language allows sharing. In particular, we want to get rid of the duplication in the code generated for *Catch*.

We can avoid this duplication by simply using a let binding to replace the two occurrences of c with a metavariable c' that is then bound to c. The last equation for $comp^{\mathsf{A}}_{\mathsf{G}}$ is thus rewritten as follows:

$$\begin{aligned} comp^{\mathsf{A}}_{\mathsf{G}}\ (Catch\ x\ h)\ c = Let\ c\ (\lambda c' \to MARK_{\mathsf{G}}\ (comp^{\mathsf{A}}_{\mathsf{G}}\ h \triangleright Var\ c') \\ \triangleright comp^{\mathsf{A}}_{\mathsf{G}}\ x \triangleright UNMARK_{\mathsf{G}} \triangleright Var\ c') \end{aligned}$$

The right-hand side for the case *Catch x h* has now only one occurrence of c.

The final code generator function $comp^{\mathsf{A}}_{\mathsf{G}}$ is then obtained by supplying $HALT_{\mathsf{G}}$ as the initial value of the code continuation and wrapping the result with the *MkGraph* constructor so as to return a result of type $Graph\ Code_{\mathsf{F}}$:

$$\begin{aligned} &comp_{\mathsf{G}} :: Expr \to Graph\ Code_{\mathsf{F}} \\ &comp_{\mathsf{G}}\ e = MkGraph\ (comp^{\mathsf{A}}_{\mathsf{G}}\ e \triangleright HALT_{\mathsf{G}}) \end{aligned}$$

To illustrate the difference between $comp_{\mathsf{G}}$ and $comp_{\mathsf{T}}$, we apply both of them to an example expression $e = Add\ (Catch\ (Val\ 1)\ (Val\ 2))\ (Val\ 3)$:

$$\begin{aligned} comp_{\mathsf{T}}\ e \rightsquigarrow\ & MARK_{\mathsf{T}}\ (PUSH_{\mathsf{T}}\ 2 \triangleright PUSH_{\mathsf{T}}\ 3 \triangleright ADD_{\mathsf{T}} \triangleright HALT_{\mathsf{T}}) \\ & \triangleright PUSH_{\mathsf{T}}\ 1 \triangleright UNMARK_{\mathsf{T}} \triangleright PUSH_{\mathsf{T}}\ 3 \triangleright ADD_{\mathsf{T}} \triangleright HALT_{\mathsf{T}} \\ comp_{\mathsf{G}}\ e \rightsquigarrow\ & MkGraph\ (Let\ (PUSH_{\mathsf{G}}\ 3 \triangleright ADD_{\mathsf{G}} \triangleright HALT_{\mathsf{G}})\ (\lambda v \to \\ & MARK_{\mathsf{G}}\ (PUSH_{\mathsf{G}}\ 2 \triangleright Var\ v) \triangleright PUSH_{\mathsf{G}}\ 1 \triangleright UNMARK_{\mathsf{G}} \triangleright Var\ v)) \end{aligned}$$

Note that $comp_{\mathsf{T}}$ duplicates the code fragment $PUSH_{\mathsf{T}}\ 3 \triangleright ADD_{\mathsf{T}} \triangleright HALT_{\mathsf{T}}$, which is supposed to be executed after the catch expression, whereas $comp_{\mathsf{G}}$ binds this code fragment to a metavariable v, which is then used as a substitute.

The recursion schemes on structured graphs make use of the parametricity in the metavariable type as well. The general fold over graphs as given by Oliveira and Cook [18] is defined as follows:[1]

$$\begin{aligned} &gfold :: Functor\ f \Rightarrow (v \to r) \to (r \to (v \to r) \to r) \to (f\ r \to r) \to \\ &\qquad\qquad Graph\ f \to r \\ &gfold\ v\ l\ i\ (MkGraph\ g) = trans\ g \\ &\quad \textbf{where}\ trans\ (Var\ x) \quad = v\ x \\ &\qquad\quad\ \ trans\ (Let\ e\ f) = l\ (trans\ e)\ (trans \circ f) \\ &\qquad\quad\ \ trans\ (GIn\ t) \quad = i\ (fmap\ trans\ t) \end{aligned}$$

The combinator takes three functions, which are used to interpret the three constructors of *Graph'*. This general form is needed for example if we want to transform the graph representation into a linearised form [2], but for our purposes we only need a simple special case of it:

[1] Oliveira and Cook [18] considered the more general case of cyclic graphs, the definition of *gfold* given here is specialised to the case of acyclic graphs.

$$
\begin{aligned}
&\mathit{ufold} :: \mathit{Functor}\ f \Rightarrow (f\ r \to r) \to \mathit{Graph}\ f \to r \\
&\mathit{ufold} = \mathit{gfold}\ \mathit{id}\ (\lambda e\ f \to f\ e)
\end{aligned}
$$

Note that the type signature is identical to the one for *fold* except for the use of *Graph* instead of *Tree*. Thus, we can reuse the algebra *execAlg* from Section 3.1, which defines the semantics of $\mathit{Tree}\ \mathit{Code}_{\mathsf{F}}$, in order to define the semantics of $\mathit{Graph}\ \mathit{Code}_{\mathsf{F}}$:

$$
\begin{aligned}
&\mathit{exec}_{\mathsf{G}} :: \mathit{Graph}\ \mathit{Code}_{\mathsf{F}} \to \mathit{Stack} \to \mathit{Stack} \\
&\mathit{exec}_{\mathsf{G}} = \mathit{ufold}\ \mathit{execAlg}
\end{aligned}
$$

4 Correctness Proof

In this section we shall prove that the graph-based compiler that we defined in Section 3.2 is indeed correct. This turns out to be rather simple: we derive the correctness property for $\mathit{comp}_{\mathsf{G}}$ from the correctness property for $\mathit{comp}_{\mathsf{T}}$. The simplicity of the argument is rooted in the fact that $\mathit{comp}_{\mathsf{T}}$ is the same as $\mathit{comp}_{\mathsf{G}}$ followed by unravelling. In other words, $\mathit{comp}_{\mathsf{G}}$ only differs from $\mathit{comp}_{\mathsf{T}}$ in that it adds sharing – as expected.

4.1 Compiler Correctness by Unravelling

Before we prove this relation between $\mathit{comp}_{\mathsf{T}}$ and $\mathit{comp}_{\mathsf{G}}$, we need to specify what unravelling means:

$$
\begin{aligned}
&\mathit{unravel} :: \mathit{Functor}\ f \Rightarrow \mathit{Graph}\ f \to \mathit{Tree}\ f \\
&\mathit{unravel} = \mathit{ufold}\ \mathit{In}
\end{aligned}
$$

While this definition is nice and compact, we gain more insight into what it actually does by unfolding it:

$$
\begin{aligned}
&\mathit{unravel} :: \mathit{Functor}\ f \Rightarrow \mathit{Graph}\ f \to \mathit{Tree}\ f \\
&\mathit{unravel}\ (\mathit{MkGraph}\ g) = \mathit{unravel}'\ g \\
&\mathit{unravel}' :: \mathit{Functor}\ f \Rightarrow \mathit{Graph}'\ f\ (\mathit{Tree}\ f) \to \mathit{Tree}\ f \\
&\mathit{unravel}'\ (\mathit{Var}\ x) \quad = x \\
&\mathit{unravel}'\ (\mathit{Let}\ e\ f) = \mathit{unravel}'\ (f\ (\mathit{unravel}'\ e)) \\
&\mathit{unravel}'\ (\mathit{GIn}\ t) \quad = \mathit{In}\ (\mathit{fmap}\ \mathit{unravel}'\ t)
\end{aligned}
$$

We can see that *unravel* simply replaces *GIn* with *In*, and applies the function argument f of a let binding to the bound value e. For example, we have that

$$
\begin{aligned}
&\mathit{MkGraph}\ (\mathit{Let}\ (\mathit{PUSH}_{\mathsf{G}}\ 2 \triangleright \mathit{HALT}_{\mathsf{G}})\ (\lambda v \to \mathit{MARK}_{\mathsf{G}}\ (\mathit{Var}\ v) \triangleright \mathit{Var}\ v)) \\
&\overset{\mathit{unravel}}{\rightsquigarrow} \mathit{MARK}_{\mathsf{T}}\ (\mathit{PUSH}_{\mathsf{T}}\ 2 \triangleright \mathit{HALT}_{\mathsf{T}}) \triangleright \mathit{PUSH}_{\mathsf{T}}\ 2 \triangleright \mathit{HALT}_{\mathsf{T}}
\end{aligned}
$$

We can now formulate the relation between $\mathit{comp}_{\mathsf{T}}$ and $\mathit{comp}_{\mathsf{G}}$:

Lemma 1. $$comp_{\mathsf{T}} = unravel \circ comp_{\mathsf{G}}$$

This lemma, which we shall prove at the end of this section, is one half of the argument for deriving the correctness property for $comp_{\mathsf{G}}$. The other half is the property that $exec_{\mathsf{T}}$ and $exec_{\mathsf{G}}$ have the converse relationship, viz.

$$exec_{\mathsf{G}} = exec_{\mathsf{T}} \circ unravel$$

Proving this property is much simpler, though, because it follows from a more general property of *fold*.

Theorem 2. *Given a strictly positive functor f, a type c, and $alg :: f\ c \to c$, we have the following:*

$$ufold\ alg = fold\ alg \circ unravel$$

The equality $exec_{\mathsf{G}} = exec_{\mathsf{T}} \circ unravel$ is an instance of Theorem 2 where $alg = execAlg$. We defer discussion of the proof of this theorem until Section 4.2.

We derive the correctness of $comp_{\mathsf{G}}$ by combining Lemma 1 and Theorem 2:

Theorem 3 (correctness of $comp_{\mathsf{G}}$).

$$exec_{\mathsf{G}}\ (comp_{\mathsf{G}}\ e)\ [\,] = conv\ (eval\ e) \qquad \textit{for all } e :: Expr$$

Proof.
$$\begin{aligned} exec_{\mathsf{G}}\ (comp_{\mathsf{G}}\ e)\ [\,] &= exec_{\mathsf{T}}\ (unravel\ (comp_{\mathsf{G}}\ e)\ [\,] && \text{(Theorem 2)} \\ &= exec_{\mathsf{T}}\ (comp_{\mathsf{T}}\ e)\ [\,] && \text{(Lemma 1)} \\ &= conv\ (eval\ e) && \text{(Corollary 1)} \end{aligned}$$
□

We conclude this section by giving the proof of Lemma 1.

Proof (of Lemma 1). Instead of proving the equation directly, we prove the following equation for all $e :: Expr$ and $c :: Graph'\ Code_{\mathsf{F}}\ (Tree\ Code_{\mathsf{F}})$:

$$comp^{\mathsf{A}}_{\mathsf{T}}\ e \triangleright unravel'\ c = unravel'\ (comp^{\mathsf{A}}_{\mathsf{G}}\ e \triangleright c) \tag{1}$$

In particular, the above equation holds for all $c :: \forall\ v\ .\ Graph'\ Code_{\mathsf{F}}\ v$. Thus, the lemma follows from the above equation as follows:

$$\begin{aligned} & comp_{\mathsf{T}}\ e \\ = \quad & \{ \text{ definition of } comp_{\mathsf{T}} \} \\ & comp^{\mathsf{A}}_{\mathsf{T}}\ e \triangleright HALT_{\mathsf{T}} \\ = \quad & \{ \text{ definition of } unravel' \} \\ & comp^{\mathsf{A}}_{\mathsf{T}}\ e \triangleright unravel'\ HALT_{\mathsf{G}} \\ = \quad & \{ \text{ Equation (1) } \} \\ & unravel'\ (comp^{\mathsf{A}}_{\mathsf{G}}\ e \triangleright HALT_{\mathsf{G}}) \\ = \quad & \{ \text{ definition of } unravel \} \\ & unravel\ (MkGraph\ (comp^{\mathsf{A}}_{\mathsf{G}}\ e \triangleright HALT_{\mathsf{G}})) \\ = \quad & \{ \text{ definition of } comp_{\mathsf{G}} \} \\ & unravel\ (comp_{\mathsf{G}}\ e) \end{aligned}$$

We prove (1) by induction on e:

- Case $e = \mathit{Val}\ n$:

$$
\begin{aligned}
& \mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ (\mathit{Val}\ n) \triangleright c) \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{G}} \ \} \\
& \mathit{unravel}'\ (\mathit{PUSH}_{\mathsf{G}}\ n \triangleright c) \\
= & \quad \{ \text{ definition of } \mathit{unravel}' \ \} \\
& \mathit{PUSH}_{\mathsf{T}}\ n \triangleright \mathit{unravel}'\ c \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{T}} \ \} \\
& \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ (\mathit{Val}\ n) \triangleright \mathit{unravel}'\ c
\end{aligned}
$$

- Case $e = \mathit{Throw}$:

$$
\begin{aligned}
& \mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ \mathit{Throw} \triangleright c) \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{G}} \ \} \\
& \mathit{unravel}'\ \mathit{THROW}_{\mathsf{G}} \\
= & \quad \{ \text{ definition of } \mathit{unravel}' \ \} \\
& \mathit{THROW}_{\mathsf{T}} \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{T}} \ \} \\
& \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ \mathit{Throw} \triangleright \mathit{unravel}'\ c
\end{aligned}
$$

- Case $e = \mathit{Add}\ x\ y$:

$$
\begin{aligned}
& \mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ (\mathit{Add}\ x\ y) \triangleright c) \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{G}} \ \} \\
& \mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ x \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ y \triangleright \mathit{ADD}_{\mathsf{G}} \triangleright c) \\
= & \quad \{ \text{ induction hypothesis } \} \\
& \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ x \triangleright \mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ y \triangleright \mathit{ADD}_{\mathsf{G}} \triangleright c) \\
= & \quad \{ \text{ induction hypothesis } \} \\
& \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ x \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ y \triangleright \mathit{unravel}'\ (\mathit{ADD}_{\mathsf{G}} \triangleright c) \\
= & \quad \{ \text{ definition of } \mathit{unravel}' \ \} \\
& \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ x \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ y \triangleright \mathit{ADD}_{\mathsf{T}} \triangleright \mathit{unravel}'\ c \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{T}} \ \} \\
& \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ (\mathit{Add}\ x\ y) \triangleright \mathit{unravel}'\ c
\end{aligned}
$$

- Case $e = \mathit{Catch}\ x\ h$:

$$
\begin{aligned}
& \mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ (\mathit{Catch}\ x\ h) \triangleright c) \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{G}} \ \} \\
& \mathit{unravel}'\ (\mathit{Let}\ c\ (\lambda c' \to \mathit{MARK}_{\mathsf{G}}\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ h \triangleright \mathit{Var}\ c') \\
& \qquad\qquad\qquad\qquad \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ x \triangleright \mathit{UNMARK}_{\mathsf{G}} \triangleright \mathit{Var}\ c')) \\
= & \quad \{ \text{ definition of } \mathit{unravel}' \text{ and } \beta\text{-reduction } \} \\
& \mathit{unravel}'\ (\mathit{MARK}_{\mathsf{G}}\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ h \triangleright \mathit{Var}\ (\mathit{unravel}'\ c)) \\
& \qquad\qquad \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ x \triangleright \mathit{UNMARK}_{\mathsf{G}} \triangleright \mathit{Var}\ (\mathit{unravel}'\ c)) \\
= & \quad \{ \text{ definition of } \mathit{unravel}' \ \} \\
& \mathit{MARK}_{\mathsf{T}}\ (\mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ h \triangleright \mathit{Var}\ (\mathit{unravel}'\ c))) \\
& \quad \triangleright \mathit{unravel}'\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}\ x \triangleright \mathit{UNMARK}_{\mathsf{G}} \triangleright \mathit{Var}\ (\mathit{unravel}'\ c)) \\
= & \quad \{ \text{ induction hypothesis } \} \\
& \mathit{MARK}_{\mathsf{T}}\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ h \triangleright \mathit{unravel}'\ (\mathit{Var}\ (\mathit{unravel}'\ c))) \\
& \quad \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ x \triangleright \mathit{unravel}'\ (\mathit{UNMARK}_{\mathsf{G}} \triangleright \mathit{Var}\ (\mathit{unravel}'\ c)) \\
= & \quad \{ \text{ definition of } \mathit{unravel}' \ \} \\
& \mathit{MARK}_{\mathsf{T}}\ (\mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ h \triangleright \mathit{unravel}'\ c) \triangleright \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ x \triangleright \mathit{UNMARK}_{\mathsf{T}} \triangleright \mathit{unravel}'\ c \\
= & \quad \{ \text{ definition of } \mathit{comp}^{\mathsf{A}}_{\mathsf{T}} \ \} \\
& \mathit{comp}^{\mathsf{A}}_{\mathsf{T}}\ (\mathit{Catch}\ x\ h) \triangleright \mathit{unravel}'\ c \qquad \square
\end{aligned}
$$

4.2 Proof of Theorem 2

Theorem 2 states that folding a structured graph $g :: Graph\ f$ over a strictly positive functor f with an algebra *alg* yields the same result as first unravelling g and then folding the resulting tree with *alg*, i.e.

$$ufold\ alg = fold\ alg \circ unravel$$

Since *unravel* is defined as *ufold In*, the above equality follows from a more general law of folds over algebraic data types, known as *short cut fusion* [8]:

$$b\ alg = fold\ alg\ (b\ In) \quad \text{for all } b :: \forall\ c\ .\ (f\ \ c \to c) \to c$$

This law holds for all strictly positive functors f as proved by Johann [12]. Essential for its correctness is the polymorphic type of b.

For any given graph $g :: Graph\ f$, we can instantiate b with the function $\lambda a \to ufold\ \ a\ \ g$, which yields that

$$(\lambda a \to ufold\ \ a\ \ g)\ alg = fold\ alg\ ((\lambda a \to ufold\ \ a\ \ g)\ In)$$

Note that $\lambda a \to ufold\ \ a\ \ g$ has indeed the required polymorphic type. After applying beta-reduction, we obtain the equation

$$ufold\ alg\ g = fold\ alg\ (ufold\ In\ g)$$

Since g was chosen arbitrarily, and *unravel* is defined as *ufold In*, we thus obtain the equation as stated in Theorem 2:

$$ufold\ alg = fold\ alg \circ unravel$$

5 Other Approaches

5.1 Other Graph Representations

The technique presented here is not necessarily dependent on the particular representation of graphs that we chose. However, while other representations are conceivable, structured graphs have two properties that make them a suitable choice for this application: (1) they have a simple representation in Haskell and (2) they provide a convenient interface for introducing sharing, viz. variable binding in the host language.

Nevertheless, in other circumstances a different representation may be advantageous. For example the use of higher-order abstract syntax may have a negative impact on performance in practical applications. Moreover, the necessity of reasoning over parametricity may be inconvenient for a formalisation of the proofs in a proof assistant.

Therefore, we also studied an alternative representation of graphs that uses de Bruijn indices for encoding binders instead of parametric higher-order abstract syntax (PHOAS). To this end, we have used the technique proposed by

Bernardy and Pouillard [4] to provide a PHOAS interface to this graph representation. This allows us to use essentially the same simple definition of the graph-based compiler as presented in Section 3.2. Using this representation of graphs – PHOAS interface on the outside, de Bruijn indices under the hood – we formalised the proofs presented here in the Coq theorem prover[2].

5.2 A Monadic Approach

We briefly describe a variant of our technique that is based on free monads and a corresponding monadic graph structure. The general approach of this variant is similar to what we have seen thus far; however, the monadic structure simplifies some of the proofs. The details can be found in the companion report [2].

The underlying idea, originally developed by Matsuda et al. [15], is to replace a function f with accumulation parameters by a function f' that produces a *context* with the property that

$$f\ x\ a_1 \ldots a_n = (f'\ x)\langle a_1, \ldots, a_n \rangle$$

That is, we obtain the result of the original function f by plugging in the accumulation arguments $a_1, \ldots, a_n$ in to the context that f' produces.

In order to represent contexts, we use a free monad type $\mathit{Tree}_{\mathsf{M}}$ instead of a tree type *Tree*, where $\mathit{Tree}_{\mathsf{M}}$ is obtained from *Tree* by adding a constructor of type $a \to \mathit{Tree}_{\mathsf{M}}\ f\ a$. A context with n holes is represented by a type $\mathit{Tree}_{\mathsf{M}}\ f\ (\mathit{Fin}\ n)$ – where $\mathit{Fin}\ n$ is a type with exactly n distinct inhabitants – and context application is represented by the monadic bind operator $\ggg$=. The compiler is then reformulated as follows – using the shorthand $\mathit{hole} = \mathit{return}\ ()$:

$$
\begin{aligned}
&\mathit{comp}^{\mathsf{C}}_{\mathsf{M}} :: \mathit{Expr} \to \mathit{Tree}_{\mathsf{M}}\ \mathit{Code}_{\mathsf{F}}\ ()\\
&\mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ (\mathit{Val}\ n) &&= \mathit{PUSH}_{\mathsf{M}}\ n\ \mathit{hole}\\
&\mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ (\mathit{Add}\ x\ y) &&= \mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ x \gg \mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ y \gg \mathit{ADD}_{\mathsf{M}}\ \mathit{hole}\\
&\mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ (\mathit{Throw}) &&= \mathit{THROW}_{\mathsf{M}}\\
&\mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ (\mathit{Catch}\ x\ h) &&= \mathit{MARK}_{\mathsf{M}}\ (\mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ h)\ (\mathit{comp}^{\mathsf{C}}_{\mathsf{M}}\ x \gg \mathit{UNMARK}_{\mathsf{M}}\ \mathit{hole})
\end{aligned}
$$

As we only have a single accumulator for the compiler, we use the type $() \simeq \mathit{Fin}\ 1$ to express that there is exactly one type of hole.

Also graphs can be given monadic structure by adding a constructor of type $a \to \mathit{Graph}'_{\mathsf{M}}\ f\ v\ a$ to the data type Graph'. And the compiler $\mathit{comp}^{\mathsf{A}}_{\mathsf{G}}$ can be reformulated in terms of this type accordingly.

We can define fold combinators for the monadic structures as well. The virtual machines are thus easily adapted to this monadic style by simply reusing the same algebra *execAlg*. Again, one half of the correctness proof follows from a generic theorem about folds corresponding to Theorem 2. The other half of the proof can be simplified. In the corresponding proof of Lemma 1 it suffices to show the following simpler equation, in which $\mathit{unravel}'$ only appears once:

$$\mathit{comp}^{\mathsf{A}}_{\mathsf{T}} = \mathit{unravel}' \circ \mathit{comp}^{\mathsf{A}}_{\mathsf{G}}$$

[2] Available from the author's web site.

This simplifies the induction proof. While this proof requires an additional lemma, viz. that unravelling distributes over $\gg$, this lemma can be proved (once and for all) for any strictly positive functor f:

$$unravel'\ (g_1 \gg g_2) = unravel'\ g_1 \gg unravel'\ g_2$$

Unfortunately, we cannot exploit short cut fusion to prove this lemma because it involves a genuine graph transformation, viz. $\gg$ on graphs . However, with the representation mentioned in Section 5.1, we can prove it by induction.

Note that the full monadic structure of $Tree_{\mathsf{M}}$ and $Graph_{\mathsf{M}}$ is not needed for our example compiler since we only use the simple bind operator $\gg$, not $\gg\!=$. However, a different compiler implementation may use more than one accumulation parameter (for example an additional code continuation that contains the current exception handler), for which we need the more general bind operator.

6 Concluding Remarks

6.1 Related Work

Compiler verification is still a hard problem and in this paper we only cover one – but arguably the central – part of a compiler, viz. the translation of a high-level language to a low-level language. The literature on the topic of compiler verification is vast (e.g. see the survey of Dave [7]). More recent work has shown impressive results in verification of a realistic compiler for the C language [13]. But there are also efforts in verifying compilers for higher-level languages (e.g. by Chlipala [6]).

This paper, however, focuses on identifying simple but powerful techniques for reasoning about compilers rather than engineering massive proofs for full-scale compilers. Our contributions thus follow the work on calculating compilers [21, 16, 1] as well as Hutton and Wright's work on equational reasoning about compilers [10, 11].

Structured graphs have been used in the setting of programming language implementation before: Oliveira and Löh [17] used structured graphs to represent embedded domain-specific languages. That is, graphs are used for the representation of the source language. Graph structures used for representing intermediate languages in a compiler typically employ pointers (e.g. Ramsey and Dias [20]) or labels (e.g. Ramsey et al. [19]). We are not aware of any work that makes use of higher-order abstract syntax or de Bruijn indices in the representation of graph structures in this setting.

6.2 Discussion and Future Work

The underlying goal of our method is to separate the transformation to the target language from the need to generate fresh names for representing jumps. For a full compiler, we still have to deal with explicit jumps eventually, but we can do so in isolation. That is, (1) we have to define a function

$$linearise :: Graph\ Code_{\mathsf{F}} \to Code_{\mathsf{L}}$$

that transforms the graph-based representation into a linear representation of the target language, and (2) we have to prove that it preserves the semantics. The proof can focus solely on the aspect of fresh names and explicit jumps. Since *linearise* is trivial for all cases except for the let bindings of the graph representation, we expect that the proof can be made independently of the actual language under consideration.

While our method reduces the proof obligations for the graph-based compiler considerably, there is still room for improvement. Indeed, we only require a simple induction proof showing the equality $comp_{\mathsf{T}} = unravel \circ comp_{\mathsf{G}}$. But since the two compiler variants differ only in the sharing they produce, one would hope the proof obligation could be further reduced to the only interesting case, i.e. the case for *Catch* in our example. In a proof assistant such as Coq, we can indeed take care of all the other cases with a single tactic and focus on the interesting case. However, it would be desirable to have a more systematic approach that captures this intuitive understanding.

A shortcoming of our method is its limitation to acyclic graphs. Nevertheless, the implementation part of our method easily generalises to cyclic structures, which permits compilation of cyclic control structures like loops. Corresponding correctness proofs, however, need a different reasoning principle.

Acknowledgements

The author is indebted to Janis Voigtländer for his many helpful suggestions to improve both the substance and the presentation of this paper. In particular, the idea to use short cut fusion to prove Theorem 2 was his. The author would like to thank Nicolas Pouillard and Daniel Gustafsson for their assistance in the accompanying Coq development.

This work was supported by the Danish Council for Independent Research, Grant 12-132365, “Efficient Programming Language Development and Evolution through Modularity”.

References

[1] Ager, M.S., Biernacki, D., Danvy, O., Midtgaard., J.: From interpreter to compiler and virtual machine: A functional derivation. Tech. Rep. RS-03-14, BRICS, Department of Computer Science, University of Aarhus (2003)

[2] Bahr, P.: Proving correctness of compilers using structured graphs (extended version). Tech. rep., University of Copenhagen (2014)

[3] Bahr, P., Hutton, G.: Calculating correct compilers (2014), unpublished manuscript

[4] Bernardy, J.P., Pouillard, N.: Names for free: polymorphic views of names and binders. In: Proceedings of the 2013 ACM SIGPLAN symposium on Haskell. pp. 13–24. ACM (2013)

[5] Chlipala, A.: Parametric higher-order abstract syntax for mechanized semantics. In: Proceeding of the 13th ACM SIGPLAN International Conference on Functional Programming. pp. 143–156. ACM (2008)
[6] Chlipala, A.: A verified compiler for an impure functional language. In: Proceedings of the 37th annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages. pp. 93–106. ACM (2010)
[7] Dave, M.A.: Compiler verification: a bibliography. SIGSOFT Softw. Eng. Notes 28(6), 2–2 (Nov 2003)
[8] Gill, A., Launchbury, J., Peyton Jones, S.L.: A short cut to deforestation. In: Proceedings of the Conference on Functional Programming Languages and Computer Architecture. pp. 223–232. ACM (1993)
[9] Hutton, G.: Programming in Haskell, vol. 2. Cambridge University Press Cambridge (2007)
[10] Hutton, G., Wright, J.: Compiling exceptions correctly. In: Kozen, D. (ed.) Mathematics of Program Construction. Lecture Notes in Computer Science, vol. 3125, pp. 211–227. Springer Berlin / Heidelberg (2004)
[11] Hutton, G., Wright, J.: What is the meaning of these constant interruptions? J. Funct. Program. 17(06), 777–792 (2007)
[12] Johann, P.: A generalization of short-cut fusion and its correctness proof. Higher Order Symbol. Comput. 15(4), 273–300 (2002)
[13] Leroy, X.: Formal certification of a compiler back-end or: programming a compiler with a proof assistant. In: Conference record of the 33rd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages. pp. 42–54. ACM (2006)
[14] Marlow, S.: Haskell 2010 language report (2010)
[15] Matsuda, K., Inaba, K., Nakano, K.: Polynomial-time inverse computation for accumulative functions with multiple data traversals. In: Proceedings of the ACM SIGPLAN 2012 workshop on Partial evaluation and program manipulation. pp. 5–14. ACM (2012)
[16] Meijer, E.: Calculating Compilers. Ph.D. thesis, Katholieke Universiteit Nijmegen (1992)
[17] Oliveira, B.C.d.S., Löh, A.: Abstract syntax graphs for domain specific languages. In: Proceedings of the ACM SIGPLAN 2013 Workshop on Partial Evaluation and Program Manipulation. pp. 87–96. ACM (2013)
[18] Oliveira, B.C., Cook, W.R.: Functional programming with structured graphs. In: Proceedings of the 17th ACM SIGPLAN International Conference on Functional Programming. pp. 77–88. ACM (2012)
[19] Ramsey, N., Dias, J.a., Peyton Jones, S.: Hoopl: a modular, reusable library for dataflow analysis and transformation. In: Proceedings of the third ACM Haskell symposium on Haskell. pp. 121–134. ACM (2010)
[20] Ramsey, N., Dias, J.: An applicative control-flow graph based on huet's zipper. In: Proceedings of the ACM-SIGPLAN Workshop on ML. pp. 105 – 126 (2006)
[21] Wand, M.: Deriving target code as a representation of continuation semantics. ACM Trans. Program. Lang. Syst. 4(3), 496–517 (Jul 1982)