Translating C# to Branching Symbolic Transducers

Olli Saarikivi\(^1\) and Margus Veanes\(^2\)

\(^1\) Helsinki Institute for Information Technology HIIT
Aalto University
olli.saarikivi@aalto.fi

\(^2\) Microsoft Research
margus@microsoft.com

Abstract

The paper looks at tooling aspects of transforming C# programs into symbolic transducers with branching rules (BSTs). The latter are used for describing list comprehensions that incorporate loop-carried state. One concrete application is log analysis where input streams of data are transformed into output streams of data via intermediate pipelines of transducers. The paper presents algorithms for translating C# to BSTs, and for exposing control state in BSTs.

1 Introduction

This paper discusses some of the algorithmic support underlying the tool introduced in \([14]\) that implements effectful comprehensions and introduces the notion of symbolic transducers with branching rules (BSTs). Effectful comprehensions provide an elegant way to describe list comprehensions that incorporate loop-carried state. As a motivation, consider the problem of analyzing logs. The log on the disk is compressed, and thus the user has to first decompress the input stream of bits into bytes. Then the bytes are decoded into characters, and finally sequences of characters are deserialized or parsed into objects in a higher-level language such as C#. Such processing from input stream of bits to output stream of bits with intermediate layers of objects is not uncommon today \([8, 1, 19, 5]\), and applying fusion to such pipelines can be beneficial \([13, 18]\). In order for such fusion techniques to be widely applicable to real world programs there must be an accessible way to specify effectful comprehensions.

While efficient fusion of transducers is important and improves efficiency akin to filter fusion \([13]\) and deforestation \([18]\), so is the aspect of transforming C# programs (that are used in the frontend) into transducers (that are used in the backend). This latter aspect and the underlying tool support and algorithms used for that is the primary focus of this paper. We present a C# interface for specifying effectful comprehensions that encapsulates state usage. The interface is similar to ones found in existing streaming libraries. We describe the algorithms that are used to translate programs that implement this interface into symbolic transducers. There are two levels of transformations. First we show how we translate C# programs into BSTs and then how we further transform the generated BSTs to expose control states, by eliminating register dependencies, where we study partial and full register exploration algorithms for BSTs.
2 Branching Symbolic Transducers

We here formally define branching symbolic transducers or BSTs and give examples of how BSTs capture behavior of programs. For the background logic of BSTs we assume a background structure that has an effectively enumerable background universe \( U \), and is equipped with a language of function and relation symbols with fixed interpretations.

We use \( \tau \), \( s \) and \( o \) to denote types, and we write \( U_\tau \) for the corresponding sub-universe of elements of type \( \tau \). The Boolean type is \( \text{bool} \), with \( U_{\text{bool}} = \{ \text{true}, \text{false} \} \), the integer type is \( \text{int} \), and the type of \( k \)-bit bit-vectors is \( \text{bv}_k \). The Cartesian product type of types \( \iota \) and \( o \) is \( \iota \times o \). We use \( \langle \ldots, \cdot, \ldots \rangle : \tau_1 \rightarrow \cdots \rightarrow \tau_n \rightarrow \tau_1 \times \cdots \times \tau_n \) as constructors for Cartesian product (i.e. tuple) types. For projecting the \( n \)-th element of a Cartesian product term \( x \) we use \( \pi_n(x) \).

The type \( \iota^* \) is the type for finite sequences of elements of type \( \iota \). The universe \( U_{(\iota^*)} \) is the Kleene closure \( (U_{\iota})^* \) of the universe \( U_{\iota} \). We also write type \( \iota^{\leq k} \) as a semantic subtype of \( \iota^* \) of sequences of elements of length at most \( k \geq 0 \).

Terms and formulas are defined by induction over the background language and are assumed to be well-typed. The type \( \tau \) of a term \( t \) is indicated by \( t : \tau \). Terms of type \( \text{bool} \), or Boolean terms, are treated as formulas, i.e., no distinction is made between formulas and Boolean terms. All elements in \( U \) are also assumed to have corresponding constants in the background language and we use elements in \( U \) also as constants. The set of free variables in a term \( t \) is denoted by \( FV(t) \). A term \( t \) is closed when \( FV(t) = \emptyset \), and closed terms \( t \) have Tarski semantics \([t] \) over the background structure. Substitution of a variable \( x : \tau \) in \( t \) by a term \( u : \tau \) is denoted by \( t[u/x] \).

A \( \lambda \)-term \( f \) is an expression of the form \( \lambda x.t \), where \( x : \iota \) is a variable, and \( t : o \) is a term such that \( FV(t) \subseteq \{ x \} \); the type of \( f \) is \( \iota \rightarrow o \); \([f] \) denotes the function that maps \( a \in \Sigma \) to \( [t[a/x]] \in \tau \). As a convention, \( f \) and \( g \) stand for \( \lambda \)-terms. A \( \lambda \)-term of type \( \iota \rightarrow \text{bool} \) is called a \( \iota \)-predicate. We write \( \varphi \) and \( \psi \) for \( \iota \)-predicates and, for \( a \in \Sigma \), we write \( a \in [\varphi] \) for \( [\varphi](a) = \text{true} \).

We often treat \([\varphi] \) as a subset of \( \Sigma \). Given a \( \lambda \)-term \( f = (\lambda x.t) : \iota \rightarrow o \) and a term \( u : \iota \), \( f(u) \) stands for \( (\lambda x.t)[u/x] \). A predicate \( \varphi \) is unsatisfiable when \([\varphi] = \emptyset \); satisfiable, otherwise.

The main building block of an BST is a rule. A rule is an expression that denotes a partial function corresponding to a straight-line conditional statement of a program that may yield outputs, produce updates and raise exceptions. We first provide an inductive definition of rules that omits type annotations. We then define additional well-formedness criteria and the semantics for rules.

- **Undef** is the exception rule.
- If \( f \) is a \( \lambda \)-term then **Base**(\( f \)) is a basic rule.
- If \( \varphi \) is a predicate and \( r_1 \), \( r_2 \) are rules then **Ite**(\( \varphi, r_1, r_2 \)) is an if-then-else (ite) rule.

We say that a rule \( r \) is well-formed with respect to the type \( \iota \rightarrow o \), denoted \( r : \iota \rightarrow o \), when one of the following conditions holds:

- \( r \) is the rule **Undef**.
- \( r \) is a rule **Base**(\( f : \iota \rightarrow o \)).
- \( r \) is a rule **Ite**(\( \varphi : \iota \rightarrow \text{bool}, r_1 : \iota \rightarrow o, r_2 : \iota \rightarrow o \)).

A rule \( r : \iota \rightarrow o \) represents a function \([r] \) from \( U_\iota \) to \( U_o \cup \{ \bot \} \). For all \( a \in U_\iota \):

\[
\begin{align*}
[r\text{Undef}]_{a} & \overset{\text{def}}{=} \bot \\
[r\text{Base}(f)]_{a} & \overset{\text{def}}{=} [f]_{[a]} \\
[r\text{Ite}(\varphi, r_1, r_2)]_{a} & \overset{\text{def}}{=} \left\{ \begin{array}{ll} 
[r_1]_{a}, & \text{if } a \in [\varphi]; \\
[r_2]_{a}, & \text{otherwise.}
\end{array} \right.
\end{align*}
\]
We now introduce the central definition of a symbolic branching transducer that uses the definition of rules.

**Definition 1.** A Symbolic Branching Transducer or BST $A$ with input type $\iota$, output type $o$ and state type $\tau$ is a tuple $(q^0, R, F)$, where

- $q^0 \in U_{\tau}$ is the initial state;
- $R$ is an input rule of type $(\iota \times \tau) \to (o^{\leq k} \times \tau)$, for some $k \geq 0$;
- $F$ is a final rule of type $\tau \to o^{\leq k}$, for some $k \geq 0$.

For a basic subrule $r = Base(\lambda(x,y).(f(x,y),g(x,y)))$ of the input rule, $f$ is called the yield and $g$ the update of $r$. A basic subrule of the final rule is called a final yield.

We write $p \xrightarrow{a/b\,\hat{\imath}\,\hat{o}_A} q$ for a concrete transition of $A$ such that $[R_A][a,p] = (b,q)$. Similarly, we write $q \xrightarrow{b\,\hat{\imath}\,\hat{o}_A}$ for a final output of $A$ such that $[F_A](q) = b$. Intuitively, a final output is a special case of an input-epsilon move of a classical finite state transducer into a final state, but it is algorithmically useful to keep final rules separate from general input-epsilon moves. Unlike input-epsilon moves in general, final rules do not affect the core algorithms, while providing a very convenient mechanism to yield additional outputs upon reaching the end of the input tape.

We write $A^{\sigma/\gamma/\tau}$ to indicate the input/output types $\sigma/\gamma$ and the state type $\tau$ of a BST $A$. In the following we use the abbreviations $\Sigma = U_\iota$, $\Gamma = U_o$ and $Q = U_{\tau}$.

The reachability relation $p \xrightarrow{a/b\,\hat{o}_A} q$ for $a \in \Sigma^*$, $b \in \Gamma^*$, and $p,q \in Q$ is defined through the closure under the following conditions, where ‘$\cdot$’ is concatenation of sequences:

- For all $q \in Q$, $q \xrightarrow{\cdot/\cdot\,\hat{o}_A} q$.
- If $p \xrightarrow{a/b\,\hat{o}_A} p' \xrightarrow{a/e\,\hat{o}_A} q$ then $p \xrightarrow{a-a/b-c\,\hat{o}_A} q$.

**Definition 2.** The transduction of a BST $A$, $\mathcal{J}_A$, is a function from $\Sigma^*$ to $\Gamma^* \cup \{\bot\}$:

$$\mathcal{J}_A(a) \overset{\text{def}}{=} \begin{cases} b \cdot c & \text{if } \exists q \in Q, b,c \in \Gamma^* \text{ such that } (q^0 \xrightarrow{a/b\,\hat{o}_A} q \xrightarrow{e\,\hat{o}_A}) \\ \bot & \text{otherwise} \end{cases}$$

BSTs are inherently deterministic and single-valued as rules are functions and according to Definition 2 all of the input is consumed before the final rule is applied.

The following example illustrates the use of BSTs on a typical string transformation scenario and illustrates the fragment of C# that we use for defining BSTs in this paper.

**Example 2.1.** The C# program in Figure 1 corresponds to a BST that decodes certain occurrences of pairs of digits between 5 and 9 by their corresponding ASCII letters. For example `DecodeDigitPairs("a77")` is "aM".

Let $f$ be the $\lambda$-term $\lambda(x,y).(((10\cdot(y-48))+(x-48)))$ and let $\varphi$ be the predicate $\lambda(x,y).('5' \leq x \leq '9')$. The tree of if else statements in the update method maps directly to the following input rule where we lift the $\lambda$-prefix to be in the front:

$$\lambda(x,y).\text{Ite}(y = 0, \text{Ite}(\varphi(x,y), \text{Base}([],x), \text{Base}([x],y)), \text{Ite}(\varphi(x,y), \text{Base}([f(x,y)],0), \text{Base}([x],y)))$$
partial class DecodeDigitPairs : Transducer<char, char> {
    char prev = 0;
    public override IEnumerable<char> Update(char x) {
        if (prev == 0) { // no previous digit was recorded
            if ('5' <= x && x <= '9') {
                prev = x; // store the digit
            } else {
                yield return x; // output directly
            }
        } else { // prev != 0 so prev is the previous digit
            if ('5' <= x && x <= '9') {
                yield return (char)((10*(prev-48))+(x-48));
                prev = 0;
            } else {
                yield return x; // output directly
            }
        }
    }
    public override IEnumerable<char> Finish() {
        if (prev != 0) yield return prev;
    }
}

Figure 1: Sample C# transducer.

Figure 2: Depiction of the BST in Figure 1. Dashed arrows correspond to final rules. Oval nodes correspond to branch conditions and rectangular nodes correspond to basic rules.

Similarly, the final rule corresponds to the Finish method:

$$\lambda y. Re(y \neq 0, [y], [])$$
The graphical illustration of the BST for DecodeDigitPairs is shown in Figure 2. All graphs in the paper are produced automatically from our analysis framework.

3 C# to BSTs

In this section we present a procedure for translating a transducer specified in C# (see Figure 1), into an equivalent BST (see Figure 2). The C# code is in the form of a class implementing the Transducer>I,O> interface in Figure 3. The code must:

• produce output via yield return statement,
• not reference variables apart from its parameters, local variables and non-static fields,
• not call functions outside the class or any non-pure functions (purity is checked).

To translate C# into BSTs the procedure has to be able to lift types and operations on them in C# into those in a background logic for a BST. If the background logic is defined by what is supported in Z3 the lifting could for example lift:

• int into 32-bit bitvectors,
• bool into the Boolean type,
• struct into tuples (or algebraic datatypes) of the component types,

The following explanation assumes that an appropriate lifting is available, but does not go into details.

We write a function that maps a₁,...,aₙ to b₁,...,bₙ as \{a₁ \mapsto b₁,...,aₙ \mapsto bₙ\}. Given a function \(f\) we write the modified function that maps \(a\) to \(b\) as \(f\{a \mapsto b\}\).

The entry point to the procedure is ToBST in Figure 4. Given a program \(P\) it constructs control flow graphs (CFGs) [6] for the Update and Finish methods (using GetCFG) and calls ToRULE_R (or ToRULE_F) to translate the C# code into rules for a BST. The state type \(\tau\) for the final BST is a Cartesian product type of the lifted field types. ToBST also maps (see line 4) the fields of \(P\) into an initial variable mapping, where each field is mapped to a term that projects the appropriate value out of the state. This initial mapping represents an identity transformation on the state. To construct the initial state \(q₀\), ToBST lifts the initial values of the fields of \(P\) into the background logic and constructs the appropriate product from them. Finally, ToBST returns a BST with \(q₀\), where the input and final rules implement the Update and Finish methods respectively.

The main procedure for translating C# into input rules is ToRULE_R in Figure 5. In addition to a basic block \(B\) from a CFG and the current variable assignment \(vars\), each call to ToRULE_R is passed a path constraint \(\varphi_{\text{path}}\). As ToRULE_R recursively calls itself to explore further basic blocks, the recursion structure will correspond to the tree of possible executions of the current
ToBST($P$)
1 \( (in, out) := (I, O) \) as defined by $P$’s base class Transducer $<I, O>$
2 \( B_{update} := \text{GetCFG}(\text{method } \text{IEnumerable } \text{<out}> \text{ Update}(in \text{ input}) \text{ from } P) \)
3 \( B_{finish} := \text{GetCFG}(\text{method } \text{IEnumerable } \text{<out}> \text{ Finish()} \text{ from } P) \)
4 \( \text{vars} := \{i\text{th field of } P \mapsto \lambda (x,y).\pi_i(y) \mid i = 1 \ldots m \}, \) where $m$ is the number of fields in $P$
5 \( R := \text{ToRule}_R(P, B_{update}, \text{vars} \{ \text{input} \mapsto \lambda (x,y).x \}, [], \text{true} \) \)
6 \( F := \text{ToRule}_F(P, B_{finish}, \text{vars}, [], \text{true} \) \)
7 \( q_0 := (z \text{ lifted into the background logic } | z \in \text{ the fields of } P) \)
8 \( \text{return } (q_0, R, F) \)

Figure 4: Translation of C# into a BST

ToRule$_R(P, B, \text{vars}, \bar{u}, \varphi_{path})$
1 \( \text{if } \neg \text{IsSat}(\varphi_{path}) \)
2 \( \text{return } \bot \)
3 \( \text{for each stmt in the statements of } B \)
4 \( (\text{vars}, \bar{w}) := \text{EvalStmt}(\text{stmt}, \text{vars}) \)
5 \( \bar{u} := \bar{u} + \bar{w} \)
6 \( \text{match the terminator of } B \)
7 \( \text{case if } (\text{cond}) \text{ Btrue else Bfalse:} \)
8 \( (\text{vars}, \bar{u}, \varphi_{cond}) := \text{Eval}(\text{cond}, \text{vars}) \)
9 \( R_{true} := \text{ToRule}_R(P, B_{true}, \text{vars}, \bar{u}, \varphi_{path} \land \varphi_{cond}) \)
10 \( R_{false} := \text{ToRule}_R(P, B_{false}, \text{vars}, \bar{u}, \varphi_{path} \land \neg \varphi_{cond}) \)
11 \( \text{if } R_{true} = \bot \text{ return } R_{false} \)
12 \( \text{elseif } R_{false} = \bot \text{ return } R_{true} \)
13 \( \text{else return } \text{Ite}(\varphi_{cond}, R_{true}, R_{false}) \) \)
14 \( \text{case goto } B_{target} : \)
15 \( \text{return } \text{ToRule}_R(P, B_{target}, \text{vars}, \bar{u}, \varphi_{path}) \)
16 \( \text{case yield break:} \)
17 \( \text{return } \text{Base}(\lambda (x,y).<\bar{u}, \{\text{vars}(f) \mid f \text{ of } P}>)) \)
18 \( \text{case throw:} \)
19 \( \text{return } \text{Undef} \)

Figure 5: Translation of a CFG into a rule

CFG. In each recursive call \( \varphi_{path} \) is the conjunction of branch constraints for the corresponding execution path. On line 1 satisfiability of \( \varphi_{path} \) is checked using an SMT solver to prune paths from the rule being constructed.

For obtaining final rules ToRule$_F$, which is not shown, is used. The only difference to ToRule$_R$ is that on line 17 the returned base rule does not specify a state update.

The basic block $B$ consists of a list of non-branching statements followed by a terminator. ToRule$_R$ executes the statements in the basic block by calling EvalStmt, which returns an updated variable assignment and a list of yields. The code on lines 6–19 that pattern matches on the terminator of $B$ handles the different types of control flow:
if else causes the exploration to branch into two recursive ToRule\(_R\) calls. The path constraint of the recursive calls to ToRule\(_R\) may end up being unsatisfiable, in which case the rule simplifies to the one from the other branch instead of an Ite-rule.

goto has one target and as such the recursive call is a tail call, i.e., for efficiency this call could just set the parameters in the current call and jump to the beginning of the procedure.

yield break terminates the current execution path. A Base-rule is constructed from the list of yields along the path and the state update as defined by the values in \(\text{vars}\) for the fields of \(P\).

throw results in an Undef-rule, indicating that the input was rejected.

This process of exploring an execution tree of the CFG while pruning unsatisfiable branches also supports looping constructs in C\# since these translate to a CFG with if else and goto terminators. As long as the loops terminate for all inputs and states (also unreachable ones), ToRule\(_R\) will also terminate. However, it is easy to use loops to specify very large rules, in which case ToRule\(_R\) may run out of memory or appear to hang.

Example 3.1. The following transducer formats unsigned 32-bit integers in decimal notation.

```csharp
partial class FormatInt : Transducer<uint, char> {
    public override IEnumerable<char> Update(uint x) {
        int digits = 10;
        int divisor = 1000000000;
        while (divisor > 1 && divisor > x) {
            divisor /= 10;
            --digits;
        }
        for (int i = 0; i < digits; ++i) {
            yield return (char)((x / divisor) % 10 + '0');
            divisor /= 10;
        }
        yield return '\n';
    }
}
```

The input rule for this transducer has ten different Base-rules (one for each number of digits). A transducer written in C\# without loops would be larger and the code would have more repetition.

The procedure for translating C\# statements is EvalStmt in Figure 6, which directly handles:

- yield return statements by returning the expression evaluated with Eval in the list of yields, and
- local variable definitions by returning an updated vars.

For statements which consist of just a C\# expression it calls Eval, which interprets the expression in the context of the current \(\text{vars}\) and returns an equivalent expression in the background logic. Since expressions may have side effects, Eval also returns an updated version of \(\text{vars}\).

The handling of function calls on lines 6–13 of Figure 6 calls for further explanation. To evaluate the function \(f\) its CFG is created and interpreted by a call to the procedure ToExpr
Translating C# to Branching Symbolic Transducers

**EVAL_STMT(stmt, vars)**

1. match stmt
   2. case yield return a:
      3. (vars, _, v_a) := EVAL(a, vars)
      4. return (vars, [v_a])
   5. case var a = b:
      6. (vars, result) := EVAL(b, vars)
      7. return (vars{a ↦ result}, ü)
   8. default:
      9. // The statement is an expression
      10. (vars, _) := EVAL(stmt, vars)
      11. return (vars, [])

**EVAL(expr, vars)**

1. result = ⊥
2. match expr
   3. case a = b:
      4. (vars, result) := EVAL(b, vars)
      5. vars := vars{a ↦ result}
   6. case f(a_1, ..., a_n):
      7. B_f := GETCFG(f)
      8. vars_f := {}
      9. for i = 1 ... n
         10. (vars, v) := EVAL(a_i, vars)
         11. vars_f := vars_f{ith parameter of f ↦ v}
      12. result := TOEXPR(B_f, vars_f, true)
      13. vars := vars_f
   14. case var if var is a variable:
      15. result := vars(var)
   16. case l if l is a literal:
      17. result := l lifted into the background logic
   18. case a + b:
      19. plus := + lifted into the background logic
      20. (vars, v_a) := EVAL(a, vars)
      21. (vars, v_b) := EVAL(b, vars)
      22. result := plus(v_a, v_b)
      23. return (vars, result)

Figure 6: Translation of C# expressions into a background logic

in Figure 7. The arguments are interpreted left-to-right (applying any side effects in vars) and an initial variable assignment vars_f mapping the parameters of f to the arguments is created.

TOEXPR is largely similar to TO_RULE_R in Figure 5, except that it constructs a formula in the background logic instead of rules. The main differences are in the terminators supported,
ToExpr\((B, \text{vars}, \varphi_{path})\)

1. if \(\neg\text{IsSAT}(\varphi_{path})\)
2. return \(\perp\)
3. for each stmt in the statements of B
4. \((\text{vars}, _) := \text{EVAL}(\text{stmt}, \text{vars})\)
5. match the terminator of B
6. case if (cond) Btrue else Bfalse:
7. \((\text{vars}, _, \varphi_{\text{cond}}) := \text{EVAL}(\text{cond}, \text{vars})\)
8. \(e_{\text{true}} := \text{ToExpr}(B, \text{true}, \text{vars}, \varphi_{path} \land \varphi_{\text{cond}})\)
9. \(e_{\text{false}} := \text{ToExpr}(B, \text{false}, \text{vars}, \varphi_{path} \land \neg \varphi_{\text{cond}})\)
10. if \(e_{\text{true}} = \perp\) return \(e_{\text{false}}\)
11. elseif \(e_{\text{false}} = \perp\) return \(e_{\text{true}}\)
12. else return ite(\(\varphi_{\text{cond}, e_{\text{true}}, e_{\text{false}}})\)
13. case goto Btarget:
14. return ToExpr(Btarget, \text{vars}, \varphi_{path})
15. case return a:
16. \((_, \text{result}) := \text{EVAL}(a, \text{vars})\)
17. return result

Figure 7: Translation of a CFG for a pure function into an expression in a background logic

with functions passed to ToExpr having to end in a return statement and yield statements
not being supported. Also note that on line 12 the ite being returned is a term in the
background logic instead of an Ite-rule. As the initial \(\text{vars}_f\) constructed by EVAL includes only
the parameters of \(f\), only pure functions are supported by ToExpr.

While the EVAL presented here uses ToExpr to inline function calls, ToExpr can also be
used to provide background definitions for functions. This would result in more compact terms
being created, but would prevent functions from being simplified to the context they are called from.

4 Register to Control State Exploration

In this section we develop an algorithm that allows us to eliminate either all or some of the
state registers used in a deterministic BST \(A\). In particular, we focus on two, most prominent
cases:

- \textit{full} exploration, and
- \textit{Boolean} exploration.

For the purpose of explaining the exploration algorithm, we extend \(A = (q^0, R, F)\) with a
component \(P\) that is a finite set of control states and an initial control state \(p^0 \in P\). The rules
\(R\) and \(F\) are extended to be maps from \(P\) to rules, and each basic subrule of the input rule
\(R\) has an additional control state component \(p \in P\). With this extension in mind, we write a
basic rule as \(\text{Base}(\text{yield}, \text{update}, p)\). We say that \(A\) is \textit{stateless} when the register type \(\tau\) is the
unit type \(T_0 (U_{T_0} = \{\})), i.e., registers are not used in a stateless BST, and thus \(R\) has the
Explore($A^{1/\omega \tau_1 \times \tau_2}$)

1. $p_0 := \pi_1(q_0^0)$
2. $q_0 := \pi_2(q_0^0)$
3. $S := \text{stack}(q_0)$
4. $P := \{p_0\}$
5. $\text{Add} := \lambda p. \text{if } p \notin P \text{ then } P := P \cup \{p\}; \text{Push}(S, p)$
6. $R := \{\to\}$
7. $F := \{\to\}$
8. while $S \neq \emptyset$
9. \quad $p := \text{Pop}(S)$
10. $R(p) := \text{Expl}(\lambda y.: \tau_2. \text{true}, \text{Inst}(\lambda y.: \tau_2. \text{true}, R_A, p), \text{Add})$
11. $F(p) := \text{Expl}(\lambda y.: \tau_2. \text{true}, \text{Inst}(\lambda y.: \tau_2. \text{true}, F_A, p), \text{Add})$
12. return $(P, p_0, q_0, R, F)$

Inst($\phi, R, p$)

1. match $R$
2. \quad case Undef: return Undef
3. \quad case Base($f, g$): return Base($\lambda y.f(p, y), \lambda y.g(p, y)$)
4. \quad case Ite($\psi, t, f$):
5. \quad \quad $\phi_t := \lambda y.\phi(y) \land \psi(p, y)$
6. \quad \quad $\phi_f := \lambda y.\phi(y) \land \neg\psi(p, y)$
7. \quad \quad if IsSat($\phi_t$) return Inst($\phi_f, f, p$)
8. \quad \quad elseif IsSat($\phi_f$) return Inst($\phi_t, t, p$)
9. \quad else return Ite($\lambda y.\psi(p, y), \text{Inst}(\phi_t, t, p), \text{Inst}(\phi_f, f, p)$)

Expl($\phi, R, \text{Add}$)

1. match $R$
2. \quad case Undef: return Undef
3. \quad case Ite($\psi, t, f$): return Ite($\phi, \text{Expl}(\phi \land \psi, t, \text{Add}), \text{Expl}(\phi \land \neg\psi, f, \text{Add})$
4. \quad case Base($f, g$):
5. \quad \quad $\psi := \lambda y z.\phi(y) \land (z = \pi_1(g(y)))$
6. \quad \quad $r := \text{Undef}$
7. \quad \quad while $\exists M \models \psi$
8. \quad \quad \quad $r := \text{Ite}(\lambda y.p = \pi_1(g(y)), \text{Base}(f, \lambda y.\pi_2(g(y)), p), r)$
9. \quad \quad \quad $\psi := \lambda y z.\psi(y, z) \land z \neq z^M$
10. \quad \quad \quad $\text{Add}(z^M)$
11. return $r$

Figure 8: Exploration algorithm of BSTs.

equivalent form

$$\{p_1 \mapsto r_1, p_2 \mapsto r_2, \ldots, p_{|R|} \mapsto r_{|R|}\}$$

where each rule $r_i$ corresponds to a conditional statement that may yield outputs and transition to new control states but does not make use of registers by storing intermediate results in registers. This extension is useful for separation of concerns, it helps to keep the control state
By full exploration of $A$, we mean a construction of a stateless BST $A^f$ such that $T_A = T_{A^f}$, i.e., $A$ and $A^f$ are equivalent. Full exploration is not always possible, because equivalence of stateless BSTs reduces to equivalence of symbolic finite transducers (SFTs), and equivalence of SFTs is decidable [16] modulo a decidable label theory, while equivalence of BSTs is undecidable already for very restricted decidable label theories. Even when full exploration is possible, $A^f$ may still be exponentially larger than $A$.

By Boolean exploration of $A$, we mean a construction of an BST $A^b$ such that $T_A = T_{A^b}$ where all Boolean registers of $A$ have been eliminated. For example, if the state type of $A$ is $(\text{bool} \times \text{bool}) \times \text{int}$ then the the state type of $A^b$ is int, i.e., the two Boolean registers have been eliminated by adding new control states.

Note that, in order to completely eliminate the symbolic update of a rule $Base([], \lambda x, y. \varphi(x))$, where $\varphi$ is a $\iota$-predicate, i.e., to replace $\varphi$ by $\lambda x. \text{true}$ (resp. $\lambda x. \text{false}$) we would need to decide if $\forall x \varphi(x)$ holds, i.e., $\neg \varphi$ is unsatisfiable, (resp. if $\forall x \neg \varphi(x)$ holds, i.e., $\varphi$ is unsatisfiable).

Algorithm. The generic exploration algorithm of BSTs is described in figure 8. The algorithm takes as its input a BST $A$, and assumes a projection of the state type $\tau$ of $A$ into two parts $\tau_1$ and $\tau_2$. We assume, without loss of generality, that $\tau = \tau_1 \times \tau_2$. The algorithm uses an SMT solver to solve satisfiability and to generate models for formulas.

The algorithm generates a new BST by exploring the rules with respect to $\tau_1$, effectively eliminating $\tau_1$, i.e. turning it into an explicit state. In order to avoid special cases, we may always assume that either $\tau_1$ or $\tau_2$ can be unit types $\text{T}_0$ ($\text{\{\}}$). Now, full exploration of $A$ corresponds to the case when $\tau_2$ is unit type, and Boolean exploration corresponds to the case when $\tau_1$ is a Cartesian combination of Boolean registers and $\tau_2$ is a Cartesian combination of all the non Boolean registers.

$\text{Inst}(\varphi, r, p)$ creates an instance of the rule $r$ with the path condition $\varphi$ with respect to the fixed register values given by $p$. For the exception rule this is a no-op. For a basic rule this is a partial instantiation of the yield and update with respect to $p$, where $\lambda y.f(p, y)$ instantiates the first projection of the state register with the value $p$. An important point for the rules is that unreachable rule instances are incrementally eliminated by deciding satisfiability of corresponding accumulated path conditions.

$\text{Expt}(\varphi, r, \text{add})$ is a form of partial exploration of $r$ the with respect to $\tau_1$ or the projection projection function. For the exception rule the operation is a no-op. For an if-then-else rule, the
step is a direct propagation of the concretizations of the branches. The core of the computation takes place during the concretization of basic rules.

**Theorem 4.1.** Let $A$ be a deterministic BST with state type $\tau_1 \times \tau_2$. If $\text{Explore}(A)$ terminates then the result is a BST that is equivalent to $A$ and whose state type is $\tau_2$.

We omit the formal proof of the theorem but note that termination of the algorithm depends on two factors: decidability of the background theory, and finiteness of the reachable subset of $U_{\tau_1}$. The first point is already needed in the \textsc{Inst} procedure that eliminates unsatisfiable branches. The second point is needed both, for termination of construction of $r$ in $\text{Expl}$, as well as for guaranteeing that the search stack is bounded in size. A sufficient condition for the second point is when the functions used for computing the first state projection have the \textit{finite-range} property, i.e., when $U_{\tau_1}$ can be assumed to be finite.

**Example 4.1.** The BST after full exploration of $\text{DecodeDigitPairs}$ from Figure 1, is illustrated in Figure 9. The unexplored BST (in Figure 2) has a single control state 0, while the fully explored BST has 6 control states.

## 5 Implementation

We have implemented the C# to BSTs and the register exploration algorithms in the Automata library, which is available under the MIT license. A version with our changes can be found at: https://github.com/OlliSaarikivi/Automata/

The C# frontend is implemented in the CSharpFrontend project. A string containing a class extending $\text{Transducer}<I, O>$ can be turned into an instance of $\text{STb}<\text{FuncDecl}, \text{Expr}, \text{Sort}>$ (a BST with Z3 formulas as its background logic) by calling:

```csharp
Microsoft.Automata.CSharpFrontend.CSharpParser.FromString(Z3Context, string)
```

For Boolean exploration $\text{STb}$ has an $\text{ExploreBools}()$ method.

## 6 Related Work

Symbolic transducers were introduced in flat form in [16] for analysis of string sanitizers with the main focus on \textit{symbolic finite transducers} or SFTs. The paper [14] develops composition algorithms for BSTs. Further work on symbolic transducers has focused on \textit{register exploration} and \textit{input grouping}. Input grouping tries to take advantage of grouping characters into larger tokens in order to avoid intermediate register usage, that has applications in decoder analysis [7] and parallelization [17].

Stream processing area has a large body of work [9, 10, 11, 12, 15]. Some libraries for streams provide APIs for expressing stateful operations. The Apache Flink [5] and Spark Streaming [4] distributed streaming engines both provide support for using state in stream operations and an associated framework for implementing fault tolerance in the presence of state. The Highland.js [3] and Conduit [2] are traditional stream libraries, which both provide a way to express stateful operations.
7 Conclusion

The translation of C# into BSTs in Section 3 allows a natural and compact way to specify effectful comprehensions as imperative code. Using a fragment of the host language for specification ensures a seamless integration by obviating impedance mismatches arising from differences in type systems.

The register exploration algorithm in Section 4 exposes control states in the BST, thus allowing the programmer to freely use C#'s native types for state while still permitting efficient application of BST algorithms that leverage control state, such as fusion [14].

References

