Translation Validation for Synchronous Data-flow Specification in the SIGNAL Compiler

Van Chan Ngo, Jean-Pierre Talpin, and Thierry Gautier

INRIA, 35042 Rennes, France
{firstname.lastname}@inria.fr

Abstract. We present a method to construct a validator based on translation validation approach to prove the value-equivalence of variables in the compilation of the SIGNAL compiler. The computation of output *signals* in a SIGNAL program and their counterparts in the generated C code is represented by a *Synchronous Data-flow Value-Graph* (SDVG). Our validator proves that every output signal and its counterpart variable have the same values by transforming the SDVG graph.

Keywords: Value-Graph, Graph Transformation, Formal Verification, Translation Validation, Certified Compiler, Synchronous Programs.

1 Introduction

Motivation A compiler is a large and very complex program which often consists of hundreds of thousands, if not millions, lines of code, and is divided into multiple sub-systems and modules. In addition, each compiler implements a particular algorithm in its own way. That results in two main drawbacks regarding the formal verification of the compiler itself. First, constructing the specifications of the actual compiler implementation is a long and tedious task. Second, the correctness proof of a compiler implementation, in general, cannot be reused for another compiler.

To deal with these drawbacks of formally verifying the compiler itself, one can prove that the source program and the compiled program are semantically equivalent, which is the approach of *translation validation* [13,12,5]. The principle of translation validation is as follows: the source and the compiled programs are represented in a common semantics. Based on the representations of the input and compiled programs, the notion of *"correct transformation"* is formalized. An automated *proof method* is provided to generate the *proof scripts* in case the compiled program implements correctly the input program. Otherwise, it produces a counter-example.

In this work, to adopt the translation validation approach, we use a value-graph as a common semantics to represent the computation of variables in the source and compiled programs. The "correct transformation" is defined by the assertion that every output variable in the source program and the corresponding variable in the compiled program have the same values.

The Language SIGNAL [3,7] is a synchronous data-flow language that allows the specification of multi-clocked systems. SIGNAL handles unbounded sequences of *typed* values $(x(t))_{t\in\mathbb{N}}$, called *signals*, denoted by x. Each signal is implicitly indexed by a logical *clock* indicating the set of instants at which the signal is present, noted C_x. At a given instant, a signal may be present where it holds a value, or absent where it holds no value (denoted by $\perp$). Given two signals, they are *synchronous* iff they have the same clock. In SIGNAL, a process (written P or Q) consists of the synchronous composition, noted $|$, of equations over signals x, y, z, written $x := y\, op\, z$ or $x := op(y, z)$, where op is an operator. Naturally, equations and processes are concurrent.

Contribution A SDVG symbolically represents the computation of the output signals in a SIGNAL program and their counterparts in its generated C code. The same structures are shared in the graph, meaning that they are represented by the same subgraphs. Suppose that we want to show that an output signal and its counterpart have the same values. In order to do that we simply check that they are represented by the same subgraphs, meaning they label the same node. We manage to realize this check by transforming the graph using some rewrite rules, which is called *normalizing* process.

Let A and C be the source program and its generated C code. Cp denotes the unverified SIGNAL compiler which compiles A into $C = Cp(A)$ or a compilation error. We now associate Cp with a validator checking that for any output signal x in A and the corresponding variable x^c in C, they have the same values (denoted by $\widetilde{x} = \widetilde{x^c}$). We denote this fact by $C \sqsubseteq_{val} A$.

```
1 if (Cp(A) is Error) return Error;
2 else {
3   if (C ⊑val A) return C;
4   else return Error;
5 }
```

The main components of the validator are depicted in Fig. 1. It works as follows. First, a shared value-graph that represents the computation of all signals and variables in both programs is constructed. The value-graph can be considered as a generalization of symbolic evaluation. Then, the shared value-graph is transformed by applying graph rewrite rules (the normalization). The set of rewrite rules reflects the general rules of inference of operators, or the optimizations of the compiler. For instance, consider the 3-node subgraph representing the expression $(1 > 0)$, the normalization will transform that graph into a single node subgraph representing the value `true`, as it reflects the constant folding. Finally, the validator compares the values of the output signals and the corresponding variables in the C code. For every output signal and its corresponding variable, the validator checks whether they point to the same node in the graph, meaning that their computation is represented by the same subgraph. Therefore, in the best case, when semantics has been preserved, this check has constant time complexity $\mathcal{O}(1)$. In fact, it is always expected that most transformations and

optimizations are semantics-preserving, thus the best-case complexity is important.

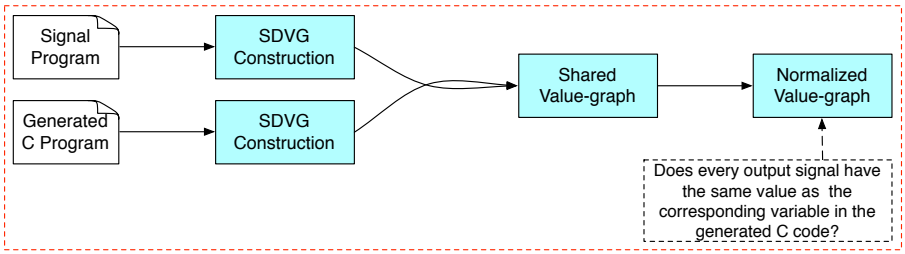

Fig. 1: SDVG Translation Validation Architecture

This work is a part of the whole work of the SIGNAL compiler formal verification. Our approach is that we separate the concerns and prove each analysis and transformation stage of the compiler separately with respect to ad-hoc data-structures to carry the semantic information relevant to that phase. The preservation of the semantics can be decomposed into the preservation of clock semantics at the *clock calculation and Boolean abstraction* phase, the preservation of data dependencies at the *static scheduling* phase, and value-equivalence of variables at the *code generation* phase. Fig. 2 shows the integration of this verification framework into the compilation process of the SIGNAL compiler. For each phase, the validator takes the source program and its compiled counterpart, then constructs the corresponding formal models of both programs. Finally, it checks the existence of the *refinement* relation to prove the preservation of the considered semantics. If the result is that the relation does not exist then a "compiler bug" message is emitted. Otherwise, the compiler continues its work.

Outline The remainder of this paper is organized as follows. In Section 2, we consider the formal definition of SDVG and the representation of a SIGNAL program and its generated C code as a shared SDVG. Section 3 addresses the mechanism of the verification process based on the normalization of a SDVG. Section 4 illustrates the concept of SDVG and the verification procedure. Section 5 terminates this paper with some related work, a conclusion and an outlook to future work.

2 Synchronous Data-flow Value-Graph

Let X be the set of variables which are used to denote the signals, clocks and variables in a SIGNAL program and its generated C code, and F be the set of function symbols. In our consideration, F contains usual logic operators (`not`, `and`, `or`), numerical comparison functions (`<`, `>`, `=`, `<=`, `>=`, `/=`), numerical operators (`+`, `-`, `*`, `/`), and gated ϕ-function [2]. A gated ϕ-function such as $x = \phi(c, x_1, x_2)$ represents a branching in a program, which means x takes the value of x_1 if the condition c is satisfied, and the value of x_2 otherwise. A constant is defined as a function symbol of arity 0.

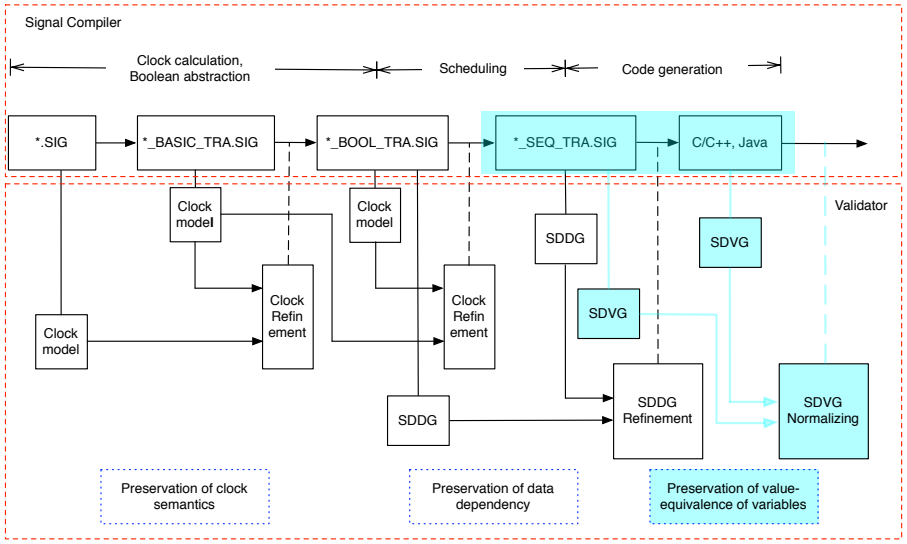

Fig. 2: The Translation Validation for the SIGNAL Compiler

Definition 1. *A* SDVG *associated with a* SIGNAL *program and its generated C code is a directed graph* $G = \langle N, E, l_N, m_N \rangle$ *where* N *is a finite set of nodes that represent clocks, signals, variables, or functions.* $E \subseteq N \times N$ *is the set of edges that describe the computation relations between nodes.* $l_N : N \longrightarrow X \cup F$ *is a mapping labeling each node with an element in* $X \cup F$. $m_N : N \longrightarrow \mathcal{P}(N)$ *is a mapping labeling each node with a finite set of clocks, signals, and variables. It defines the set of equivalent clocks, signals and variables.*

A subgraph rooted at a node is used to describe the computation of the corresponding element labelled at this node. In a graph, for a node labelled by y, the set of clocks, signals or variables $m_N(y) = \{x_0, ..., x_n\}$ is written as a node with label $\{x_0, ..., x_n\}$ y.

2.1 SDVG of SIGNAL Program

Let P be a SIGNAL program, we write $X = \{x_1, ..., x_n\}$ to denote the set of all signals in P which consists of input, output, state (corresponding to delay operator) and local signals, denoted by I, O, S and L, respectively. For each $x_i \in X$, $\mathbb{D}_{x_i}$ denotes its domain of values, and $\mathbb{D}_{x_i}^{\perp} = \mathbb{D}_{x_i} \cup \{\perp\}$ is the domain of values with the absent value. Then, the domain of values of X with absent value is defined as follows: $\mathbb{D}_X^{\perp} = \bigcup_{i=1}^{n} \mathbb{D}_{x_i} \cup \{\perp\}$. For each signal x_i, it is associated with a Boolean variable $\widehat{x_i}$ to encode its clock at a given instant t (`true`: x_i is present at t, `false`: x_i is absent at t), and $\widetilde{x_i}$ with the same type as x_i to encode its value. Formally, the abstract values to represent the clock and value of a signal can be represented by a gated ϕ-function, $x_i = \phi(\widehat{x_i}, \widetilde{x_i}, \perp)$.

Assume that the computation of signals in processes P_1 and P_2 is represented as shared value-graphs G_1 and G_2, respectively. Then the value-graph G of the

synchronous combination process $P_1|P_2$ can be defined as $G = \langle N, E, l_N, m_N \rangle$ in which for any node labelled by x, we replace it by the subgraph that is rooted by the node labelled by x in G_1 and G_2. Every identical subgraph is reused, in other words, we maximize sharing among graph nodes in G_1 and G_2. Thus, the shared value-graph of P can be constructed as a combination of the sub-value-graphs of its equations.

A SIGNAL program is built through a set of primitive operators. Therefore, to construct the SDVG of a SIGNAL program, we construct a subgraph for each primitive operator. In the following, we present the value-graph corresponding to each SIGNAL primitive operator.

Stepwise Function Consider the equation using the stepwise function $y := f(x_1, ..., x_n)$, it indicates that if all signals from x_1 to x_n are defined, then the output signal y is defined by applying f on the values of $x_1, ..., x_n$. Otherwise, it is assigned no value. Thus, the computation of y can be represented by the following gated ϕ-function: $y = \phi(\hat{y}, f(\widetilde{x_1}, \widetilde{x_2}, ..., \widetilde{x_n}), \perp)$, where $\hat{y} \Leftrightarrow \widehat{x_1} \Leftrightarrow \widehat{x_2} \Leftrightarrow ... \Leftrightarrow \widehat{x_n}$ (since they are *synchronous*). The graph representation of the stepwise function is depicted in Fig. 3. Note that in the graph, the node labelled by $\{\widehat{x_1}, ..., \widehat{x_n}\}\ \hat{y}$ means that $m_N(\hat{y}) = \{\widehat{x_1}, ..., \widehat{x_n}\}$. In other words, the subgraph representing the computation of $\hat{y}$ is also the computation of $\widehat{x_1}$, ..., and $\widehat{x_n}$.

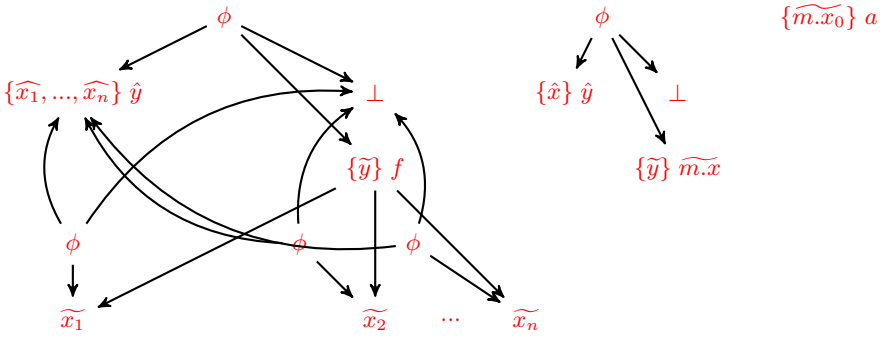

Fig. 3: The graphs of $y := f(x_1, ..., x_n)$ and $y := x\$1$ init a

Delay Consider the equation using the delay operator $y := x\$1$ init a. The output signal y is defined by the last value of the signal x when the signal x is present. Otherwise, it is assigned no value. The computation of y can be represented by the following nodes: $y = \phi(\hat{y}, \widetilde{m.x}, \perp)$ and $\widetilde{m.x_0} = a$, where $\hat{y} \Leftrightarrow \hat{x}$. $\widetilde{m.x}$ and $\widetilde{m.x_0}$ are the last value of x and the initialized value of y. The graph representation is depicted in Fig. 3.

Merge Consider the equation which corresponds to the merge operator $y := x$ default z. If the signal x is defined then the signal y is defined and holds the value of x. The signal y is assigned the value of z when the signal x is

not defined and the signal z is defined. When both x and z are not defined, y holds no value. The computation of y can be represented by the following node: $y = \phi(\hat{y}, \phi(\hat{x}, \widetilde{x}, \widetilde{z}), \bot)$, where $\hat{y} \Leftrightarrow (\hat{x} \vee \hat{z})$. The graph representation is depicted in Fig. 4. Note that in the graph, the clock $\hat{y}$ is represented by the subgraph of $\hat{x} \vee \hat{z}$.

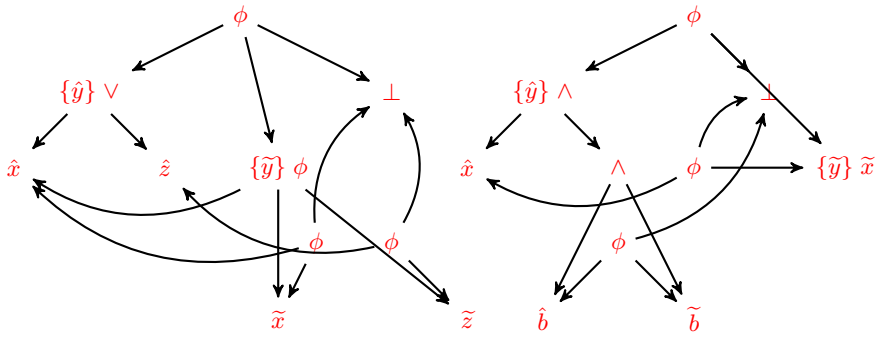

Fig. 4: The graphs of $y := x$ default z and $y := x$ when b

Sampling Consider the equation which corresponds to the sampling operator $y := x$ when b. If the signal x, b are defined and b holds the value `true`, then the signal y is defined and holds the value of x. Otherwise, y holds no value. The computation of y can be represented by the following node: $y = \phi(\hat{y}, \widetilde{x}, \bot)$, where $\hat{y} \Leftrightarrow (\hat{x} \wedge \hat{b} \wedge \widetilde{b})$. Fig. 4 shows its graph representation.

Restriction The graph representation of restriction process $P_1 \backslash x$ is the same as the graph of P_1.

Clock Relations Given the above graph representations of the primitive operators, we can obtain the graph representations for the derived operators on clocks as the following gated ϕ-function $z = \phi(\hat{z}, true, \bot)$, where $\hat{z}$ is computed as $\hat{z} \Leftrightarrow \hat{x}$ for $z := \hat{} x$, $\hat{z} \Leftrightarrow (\hat{x} \vee \hat{y})$ for $z := x \hat{}+\ y$, $\hat{z} \Leftrightarrow (\hat{x} \wedge \hat{y})$ for $z := x \hat{}*\ y$, $\hat{z} \Leftrightarrow (\hat{x} \wedge \neg\hat{y})$ for $z := x \hat{}-\ y$, and $\hat{z} \Leftrightarrow (\hat{b} \wedge \widetilde{b})$ for $z :=$ when b. For the clock relation $x \hat{}=\ y$, it is represented by a single node graph labelled by $\{\hat{x}\}\ \hat{y}$.

2.2 SDVG of Generated C Code

For constructing the shared value-graph, the generated C code is translated into a subgraph along with the subgraph of the SIGNAL program. Let A be a SIGNAL program and C its generated C code, we write $X_A = \{x_1, ..., x_n\}$ to denote the set of all signals in A, and $X_C = \{x_1^c, ..., x_m^c\}$ to denote the set of all variables in C. We added "c" as superscript for the variables, to distinguish them from the signals in A.

As described in [4,8,6,1], the generated C code of A consists of the following files:

- `A_main.c` is the implementation of the *main function*. It opens the IO communication channels by calling functions provided in `A_io.c`, and calls the *initialization function*. Then it calls the *step function* repeatedly in an infinite loop to interact with the environment.
- `A_body.c` is the implementation of the initialization function and the step function. The initialization function is called once to provide initial values to the program variables. The step function, which contains also the step initialization and finalization functions, is responsible for the calculation of the outputs to interact with the environment. This function, which is called repeatedly in an infinite loop, is the essential part of the concrete code.
- `A_io.c` is the implementation of the *IO communication functions*. The IO functions are called to setup communication channels with the environment.

The scheduling and the computations are done inside the step function. Therefore, it is natural to construct a graph of this function in order to prove that its variables and the corresponding signals have the same values. To construct the graph of the step function, the following considerations need to be studied. The generated C code in the step function consists of only the assignment and `if-then` statements. For each signal named x in A, it has a corresponding Boolean variable named C_x in the step function. Then the computation of x is implemented by a conditional `if-then` statement as follows:

```
1 if (C_x) {
2   computation(x);
3 }
```

If x is an input signal then its computation is the reading operation which gets the value of x from the environment. In case x is an output signal, after computing its value, it will be written to the IO communication channel with the environment. Note that the C programs use persistent variables (e.g., variables which always have some value) to implement the SIGNAL program A which uses volatile variables. As a result, there is a difference in the types of a signal in the SIGNAL program and of the corresponding variable in the C code. When a signal has the absent value, $\perp$, at a given instant, the corresponding C variable always has a value. This implies that we have to detect when a variable in the C code such that whose value is not updated. In this case, it will be assigned the absent value, $\perp$. Thus, the computation of a variable, called x^c, can fully be represented by a gated ϕ-function $x^c = \phi(C\_x^c, \widetilde{x^c}, \perp)$, where $\widetilde{x^c}$ denotes the newly updated value of the variable.

In the generated C code, the computation of the variable whose clock is the *master clock*, which ticks every time the step function is called, and the computation of some local variables (introduced by the SIGNAL compiler) are implemented using the forms below.

It is obvious that x is always updated when the step function is invoked. The computation of such variables can be represented by a single node graph labelled by $\{\widetilde{x^c}\}\ x^c$. That means the variable x^c is always updated and holds the value $\widetilde{x^c}$.

```
1 if (C_x) {
2   computation(x);
3 } else computation(x);
4 // or without if-then
5 computation(x)
```

Considering the following code segment, we observe that the variable x is involved in the computation of the variable y before the updating of x.

```
1 if (C_y) {
2   y = x + 1;
3 }
4 // code segment
5 if (C_x) {
6   x = ...
7 }
```

In this situation, we refer to the value of x as the previous value, denoted by $m.x^c$. It happens when a *delay* operator is applied on the signal x in the SIGNAL program. The computation of y is represented by the following gated ϕ-function: $y^c = \phi(C\_y^c, m.x^c + 1, \perp)$.

3 Translation Validation of SDVG

In this section, we introduce the set of rewrite rules to transform the shared value-graph resulting from the previous step. This procedure is called *normalizing*. At the end of the normalization, for any output signal x and its corresponding variable x^c in the generated C code, we check whether x and x^c label the same node in the resulting graph. The normalizing procedure can be adapted with any future optimization of the compiler by updating the set of rewrite rules.

3.1 Normalizing

Once a shared value-graph is constructed for the SIGNAL program and its generated C code, if the values of an output signal and its corresponding variable in the C code are not already equivalent (they do not point the same node in the shared value-graph), we start to normalize the graph. Given a set of term rewrite rules, the normalizing process works as described below. The normalizing algorithm indicates that we apply the rewrite rules to each graph node individually. When there are no more rules that can be applied to the resulting graph, we maximize the shared nodes, reusing the identical subgraphs. The process terminates when there exists no more sharing or rules that can be applied.

We classify our set of rewrite rules into three basic types: *general simplification rules*, *optimization-specific rules* and *synchronous rules*. In the following, we shall present the rewrite rules of these types, and we assume that all nodes in our shared value-graph are typed. We write a rewrite rule in form of term rewrite rules, $t_l \rightarrow t_r$, meaning that the subgraph represented by t_l is replaced

by the subgraph represented by t_r when the rule is applied. Due to the lack of space, we only present a part of these rules, the full set of rules is shown in the appendix.

```
1  // Input: G: A shared value-graph. R: The set of
2  // rewrite rules. S: The sharing among graph nodes.
3  // Output: The normalized graph
4  while (∃s ∈ S or ∃r ∈ R that can be applied on G) {
5    while (∃r ∈ R that can be applied on G) {
6      for (n ∈ G)
7        if (r can be applied on n)
8          apply the rewrite rule to n
9    }
10   maximize sharing
11 }
12 return G
```

General Simplification Rules The general simplification rules contain the rules which are related to the general rules of inference of operators, denoted by the corresponding function symbols in F. In our consideration, the operators used in the primitive stepwise functions and in the generated C code are usual logic operators, numerical comparison functions, and numerical operators. When applying these rules, we will replace a subgraph rooted at a node by a smaller subgraph. In consequence of this replacement, we will reduce the number of nodes by eliminating some unnecessary structures. The first set of rules simplifies numerical and Boolean comparison expressions. In these rules, the subgraph t represents a structure of value computing (e.g., the computation of expression $b = x \neq \texttt{true}$). These rules are self explanatory, for instance, with any structure represented by a subgraph t, the expression $t = t$ can always be replaced with a single node subgraph labelled by the value `true`.

$$
\begin{aligned}
= (t, t) &\rightarrow \texttt{true} \\
\neq (t, t) &\rightarrow \texttt{false}
\end{aligned}
$$

The second set of general simplification rules eliminates unnecessary nodes in the graph that represent the ϕ-functions, where c is a Boolean expression. For instance, we consider the following rules.

$$
\begin{aligned}
\phi(\texttt{true}, x_1, x_2) &\rightarrow x_1 \\
\phi(c, \texttt{true}, \texttt{false}) &\rightarrow c \\
\phi(c, \phi(c, x_1, x_2), x_3) &\rightarrow \phi(c, x_1, x_3)
\end{aligned}
$$

The first rule replaces a ϕ-function with its left branch if the condition always holds the value `true`. The second rule operates on Boolean expressions represented by the branches. When the branches are Boolean constants and hold different values, the ϕ-function can be replaced with the value of the condition c. Consider a ϕ-function such that one of its branches is another ϕ-function. The third rule removes the ϕ-function in the branches if the conditions of the ϕ-functions are the same.

Optimization-specific Rules Based on the optimizations of the SIGNAL compiler, we have a number of optimization-specific rules in a way that reflects the effects of specific optimizations of the compiler. These rules do not always reduce the graph or make it simpler. One has to know specific optimizations of the compiler when she wants to add them to the set of rewrite rules. In our case, the set of rules for simplifying constant expressions of the SIGNAL compiler such as:

$$+(cst_1, cst_2) \rightarrow cst, \text{ where } cst = cst_1 + cst_2$$
$$\wedge(cst_1, cst_2) \rightarrow cst, \text{ where } cst = cst_1 \wedge cst_2$$
$$\square(cst_1, cst_2) \rightarrow cst$$

where $\square$ denotes a numerical comparison function, and the Boolean value cst is the evaluation of the constant expression $\square(cst_1, cst_2)$ which can hold either the value `false` or `true`.

We also may add a number of rewrite rules that are derived from the list of *rules of inference* for propositional logic. For example, we have a group of laws for rewriting formulas with `and` operator, such as:

$$\wedge(x, \texttt{true}) \quad \rightarrow x$$
$$\wedge(x, \Rightarrow (x, y)) \rightarrow x \wedge y$$

Synchronous Rules In addition to the general and optimization-specific rules, we also have a number of rewrite rules that are derived from the semantics of the code generation mechanism of the SIGNAL compiler.

The first rule is that if a variable in the generated C code is always updated, then we require that the corresponding signal in the source program is present at every instant, meaning that the signal never holds the absent value. In consequence of this rewrite rule, the signal x and its value when it is present $\tilde{x}$ (resp. the variable x^c and its updated value $\widetilde{x^c}$ in the generated C code) point to the same node in the shared value-graph. Every reference to x and $\tilde{x}$ (resp. x^c and $\widetilde{x^c}$) point to the same node.

We consider the equation $pz := z\$1$ `init` 0. We use the variable $\widetilde{m.z}$ to capture the last value of the signal z. In the generated C program, the last value of the variable z^c is denoted by $m.z^c$. The second rule is that it is required that the last values of a signal and the corresponding variable in the generated C code are the same. That means $\widetilde{m.z} = m.z^c$.

Finally, we add rules that mirror the relation between input signals and their corresponding variables in the generated C code. First, for any input signal x and the corresponding variable x^c in the generated C code, if x is present, then the value of x which is read from the environment and the value of the variable x^c after the reading statement must be equivalent. That means $\widetilde{x^c}$ and $\widetilde{x}$ are represented by the same subgraph in the graph. Second, if the clock of x is also read from the environment as a parameter, then the clock of the input signal x is equivalent to the condition in which the variable x^c is updated. It means that we represent $\hat{x}$ and C_x^c by the same subgraph. Consequently, every reference to $\hat{x}$ and C_x^c (resp. $\widetilde{x}$ and $\widetilde{x^c}$) points to the same node.

4 Illustrative Example

Let us illustrate the verification process in Fig. 1 on the program DEC in Listing 1.1 and its generated C code DEC_step() in Listing 1.2.

In the first step, we shall compute the shared value-graph for both programs to represent the computation of all signals and their corresponding variables. This graph is depicted in Fig. 5.

```
1 process DEC=
2 (? integer FB;
3  ! integer N)
4  (| FB ^= when (ZN<=1)
5   | N := FB default (ZN-1)
6   | ZN := N$1 init 1
7   |)
8 where integer ZN init 1
9 end;
```

Listing 1.1: DEC in Signal

```
1  EXTERN logical DEC_step() {
2    C_FB = N <= 1;
3    if (C_FB) {
4      if (!r_DEC_FB(&FB)) return FALSE; // read input FB
5    }
6    if (C_FB) N = FB; else N = N - 1;
7    w_DEC_N(N); // write output N
8    DEC_step_finalize();
9    return TRUE;
10 }
```

Listing 1.2: Generated C code of DEC

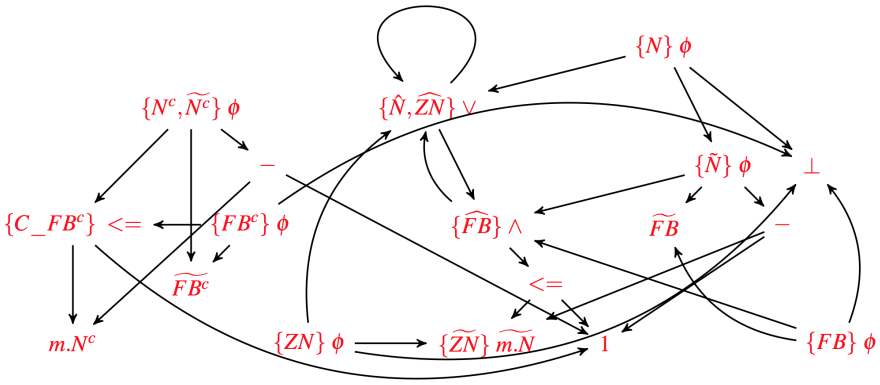

Fig. 5: The shared value-graph of DEC and DEC_step

Note that in the C program, the variable N^c ("c" is added as superscript for the C program variables, to distinguish them from the signals in the SIGNAL

program) is always updated (line (6)). In lines (2) and (6), the references to the variable N^c are the references to the last value of N^c denoted by $m.N^c$. The variable FB^c which corresponds to the input signal FB is updated only when the variable C_FB^c is `true`.

In the second step, we shall normalize the above initial graph. Below is a potential normalization scenario, meaning that it might have more than one normalization scenario, and the validator can choose one of them. For example, given a set of rules that can be applied, the validator can apply these rules with different order. Fig. 6 depicts the intermediate resulting graph of this normalization scenario, and Fig. 7 is the final normalized graph from the initial graph when we cannot perform any more normalization.

1. The clock of the output signal N is a master clock which is indicated in the generated C by the variable N^c being always updated. The node $\{\hat{N}, \widehat{ZN}\}\ \vee$ is rewritten into `true`.
2. By rule $\wedge(\texttt{true}, x) \rightarrow x$, the node $\{\widehat{FB}\}\ \wedge$ is rewritten into $\{\widehat{FB}\}\ <=$.
3. The ϕ-function node representing the computation of N is removed and N points to the node $\{\widetilde{N}\}\ \phi$.
4. The ϕ-function node representing the computation of ZN is removed and ZN points to the node $\{\widetilde{ZN}\}\ \widetilde{m.N}$.
5. The nodes $\widetilde{FB^c}$ and $\widetilde{FB}$ are rewritten into a single node $\{\widetilde{FB}\}\ \widetilde{FB^c}$. All references to them are replaced by references to $\{\widetilde{FB}\}\ \widetilde{FB^c}$.
6. The nodes $m.N^c$ and $\widetilde{m.N}$ are rewritten into a single node $\{\widetilde{m.N}\}\ m.N^c$. All references to them are replaced by references to $\{\widetilde{m.N}\}\ m.N^c$.

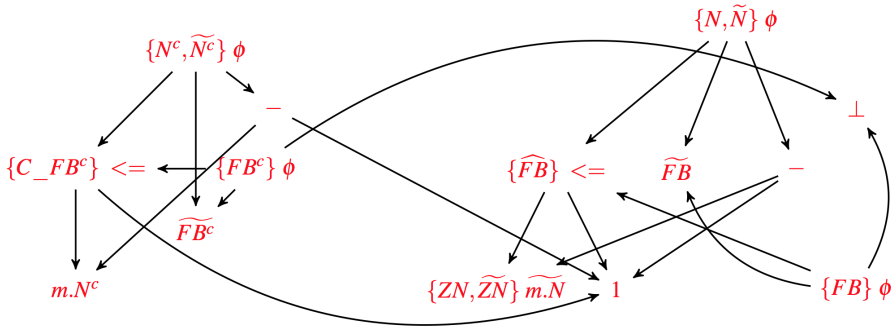

Fig. 6: The resulting value-graph of DEC and DEC_step

In the final step, we check that the value of the output signal and its corresponding variable in the generated code merge into a single node. In this example, we can safely conclude that the output signal N and its corresponding variable N^c are equivalent since they point to the same node in the final normalized graph.

5 Related Work and Conclusion

There is a wide range of works for value-graph representations of expression evaluations in a program. For example, in [16], Weise et al. present a nice summary

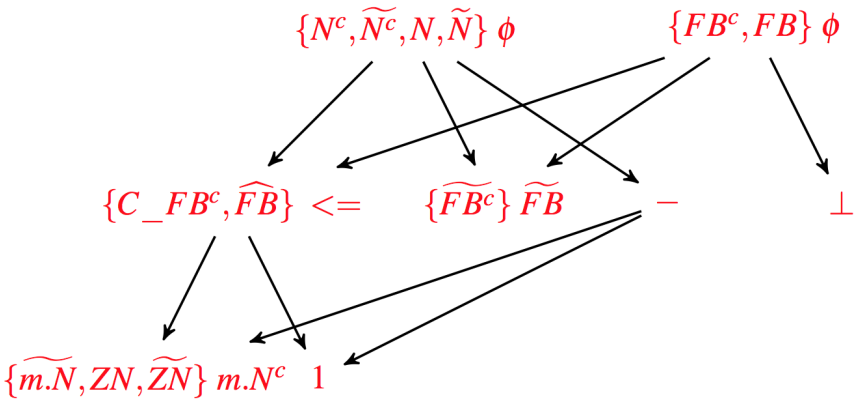

Fig. 7: The final normalized graph of DEC and DEC_step

of the various types of value-graph. In our context, the value-graph is used to represent the computation of variables in both source program and its generated C code in which the identical structures are shared. We believe that this representation will reduce the required storage and make the normalizing process more efficient than two separated graphs. Another remark is that the calculation of clocks as well as the special value, the absent value, are also represented in the shared graph.

Another related work which adopts the translation validation approach in verification of optimizations, Tristan et al. [15], recently proposed a framework for translation validation of LLVM optimizer. For a function and its optimized counterpart, they construct a shared value-graph. The graph is *normalized* (the graph is reduced). After the normalization, if the outputs of two functions are represented by the same sub-graph, they can safely conclude that both functions are equivalent.

On the other hand, Tate et al. [14] proposed a framework for translation validation. Given a function in the input program and the corresponding optimized version of the function in the output program, they compute two value-graphs to represent the computations of the variables. Then they transform the graph by adding equivalent terms through a process called *equality saturation*. After the saturation, if both value-graphs are the same, they can conclude that the return value of two given functions are the same. However, for translation validation purposes, our normalization process is more efficient and scalable since we can add rewrite rules into the validator that reflect what a typical compiler intends to do (e.g., a compiler will do the constant folding optimization, then we can add the rewrite rule for constant expressions such as three nodes subgraph $(1+2)$ is replaced by a single node 3).

The present paper provides a verification framework to prove the value-equivalence of variables and applies this approach to the synchronous data-flow compiler SIGNAL. With the simplicity of the graph normalization, we believe that translation validation of synchronous data-flow value-graph for the industrial compiler SIGNAL is feasible and efficient. Moreover, the normalization process

can always be extended by adding new rewrite rules. That makes the translation validation of SDVG scalable and flexible.

We have considered sequential code generation. A possibility is to extend this framework to use with other code generation schemes including cluster code with static and dynamic scheduling, modular code, and distributed code. One path forward is the combination of this work and the work on data dependency graph in [10,11,9]. That means that we use synchronous data-flow dependency graphs and synchronous data-flow value-graphs as a common semantic framework to represent the semantics of the generated code. The formalization of the notion of "correct transformation" is defined as the refinements between two synchronous data-flow dependency graphs and in a shared value-graph as described above.

Another possibility is that we use an SMT solver to reason on the rewriting rules. For example, we recall the following rules:

$$\phi(c_1, \phi(c_2, x_1, x_2), x_3) \rightarrow \phi(c_1, x_1, x_3) \text{ if } c_1 \Rightarrow c_2$$
$$\phi(c_1, \phi(c_2, x_1, x_2), x_3) \rightarrow \phi(c_1, x_2, x_3) \text{ if } c_1 \Rightarrow \neg c_2$$

To apply these rules on a shared value-graph to reduce the nested ϕ-functions (e.g., from $\phi(c_1, \phi(c_2, x_1, x_2), x_3)$ to $\phi(c_1, x_1, x_3)$), we have to check the validity of first-order logic formulas, for instance, we check that $\models (c_1 \Rightarrow c_2)$ and $\models c_1 \Rightarrow \neg c_2$. We consider the use of SMT to solve the validity of the conditions as in the above rewrite rules to normalize value-graphs.

References

1. P. Aubry, P. L. Guernic, and S. Machard. Synchronous distribution of signal programs. In *In Proceedings of the 29th Hawaii International Conference on System Sciences, IEEE Computer Society Press*, volume 1, pages 656–665, 1996.
2. R. Ballance, A. Maccabe, and K. Ottenstei. The program dependence web: A representation supporting control, data, and demand driven interpretation of imperative languages. In *In Proc. of the SIGPLAN'90 Conference on Programming Language Design and Implementation*, pages 257–271, 1990.
3. A. Benveniste and P. L. Guernic. Hybrid dynamical systems theory and the signal language. In *IEEE Transactions on Automatic Control*, volume 35(5), pages 535–546, 1990.
4. L. Besnard, T. Gautier, P. L. Guernic, and J.-P. Talpin. Compilation of polychronous data-flow equations. In *In Synthesis of Embedded Software Springer*, pages 01–40, 2010.
5. S. Blazy. Which c semantics to embed in the front-end of a formally verified compiler? In *Tools and Techniques for Verification of System Infrastructure, TTVSI*, 2008.
6. T. Gautier and P. L. Guernic. Code generation in the sacres project. In *In Towards System Safety, Proceedings of the Safety-critical Systems Symposium*, pages 127–149, 1999.
7. T. Gautier, P. L. Guernic, and L. Besnard. Signal, a declarative language for synchronous programming of real-time systems. In *Proc. 3rd. Conf. on Functional Programming Languages and Computer Architecture*, volume LNCS 274, 1990.

8. O. Maffeis and P. L. Guernic. Distributed implementation of signal: Scheduling and graph clustering. In *In 3rd International School and Symposium on Formal Techniques in Real-time and Fault-tolerant Systems*, volume LNCS 863, pages 547–566, 1994.
9. V. Ngo. Formal verification of a synchronous data-flow compiler: from signal to c. In *Ph.D Thesis* - http://tel.archives-ouvertes.fr/tel-01058041, 2014.
10. V. Ngo, J.-P. Talpin, T. Gautier, P. L. Guernic, and L. Besnard. Formal verification of compiler transformations on polychronous equations. In *In Proceedings of 9th International Conference on Integrated Formal Methods IFM 2012*, volume LNCS 7321, pages 113–127, 2012.
11. V. Ngo, J.-P. Talpin, T. Gautier, P. L. Guernic, and L. Besnard. Formal verification of synchronous data-flow program transformations toward certified compilers. In *In Frontiers of Computer Science, Special Issue on Synchronous Programming*, volume 7(5), pages 598–616, 2013.
12. A. Pnueli, O. Shtrichman, and M. Siegel. Translation validation: From signal to c. In *In Correct Sytem Design Recent Insights and Advances*, volume LNCS 1710, pages 231–255, 2000.
13. A. Pnueli, M. Siegel, and E. Singerman. Translation validation. In *In B. Steffen, editor, 4th Intl. Conf. TACAS'98*, volume LNCS 1384, pages 151–166, 1998.
14. R. Tate, M. Stepp, Z. Tatlock, and S. Lerner. Equility saturation: A new approach to optimization. In *In 36th Principles of Programming Languages*, pages 264–276, 2009.
15. J.-B. Tristan, P. Govereau, and G. Morrisett. Evaluating value-graph translation validation for llvm. In *In ACM SIGPLAN Conference on Programming and Language Design Implementation*, 2011.
16. D. Weise, R. Crew, M. Ernst, and B. Steensgaard. Value dependence graphs: Representation without taxation. In *In 21th Principles of Programming Languages*, pages 297–310, 1994.