Towards native higher-order remote procedure calls
Fredriksson, Olle; Ghica, Dan; Wheen, Bertram

DOI:
10.1145/2746325.2746332

Document Version
Peer reviewed version

Citation for published version (Harvard):

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

• Users may freely distribute the URL that is used to identify this publication.
• Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
• Users may use extracts from the document in line with the concept of ‘fair dealing’ under the Copyright, Designs and Patents Act 1988 (?)
• Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 05. Jun. 2020
Towards native higher-order remote procedure calls

Olle Fredriksson  Dan R. Ghica  Bertram Wheen
University of Birmingham, UK

Abstract

We present a new abstract machine, called DCESH, which models the execution of higher-order programs running in distributed architectures. DCESH implements a native general remote higher-order function call across node boundaries. It is a modernised version of SECD enriched with specialised communication features required for implementing the remote procedure call mechanism. The key correctness result is that the termination behaviour of the remote procedure call is indistinguishable (bisimilar) to that of a local call. The correctness proofs and the requisite definitions for DCESH and other related abstract machines are formalised using Agda. We also formalise a generic transactional mechanism for transparently handling failure in DCESHs.

We use the DCESH as a target architecture for compiling a conventional call-by-value functional language ("Floskel") which can be annotated with node information. Conventional benchmarks show that the single-node performance of Floskel is comparable to that of OCaml, a semantically similar language, and that distribution overheads are not excessive.

1. Native RPC and transparent distribution

The remote procedure call (RPC) [7] is a widely used mechanism for higher-level inter-process communication. An RPC is what its name suggests: a call to a procedure located on a remote node. The idea is that an RPC looks and acts like an ordinary procedure call, and performs the necessary message passing under the hood. However, the RPC mechanism tends to be bolted on top of a pre-existing language, as a library for example, rather than be seamlessly integrated into it. This leads to significant syntactic differences between calling a local library function and a remote function. Even if these syntactic differences can be smoothed using stubs that wrap remote calls into local calls [6] important differences still persist, of which the most important is that arguments must be of ground types.

Generalising RPC to all types and incorporating it seamlessly in the language has been considered but dismissed on grounds of potential inefficiencies [40]. However, considering that a number of technologies that trade efficiency for convenience have been rejected on similar grounds (from machine independent languages to functional programming, garbage collection or automated program verification – to name only a few), we decided it is time to revisit this issue. Because the execution of applications is increasingly and rapidly moving from single devices to networks of (often heterogeneous) devices a case can be made that such revisiting is timely.

A native RPC system, offering as an immediate benefit transparent and automated distribution, can be useful in domains where programmer effectiveness is more important than machine efficiency.

We attack this problem in a principled manner. We want the seamless integration of the RPC in the language to be not merely syntactic but semantic as well: the RPC is a native call of our language, on the same level as a conventional local call. This is realised by introducing a new kind of abstract machine which extends the conventional SECD machine [26], or rather a modernisation version of it, with communication primitives. These are not general-purpose low-level communication primitives but are especially designed to support the implementation of RPCs. Even though we aim for simplicity first, some of the technicalities, especially the proofs of correctness, are quite intricate. For this reason the abstract machine network framework and its correctness properties are fully formalised using the Agda language [34]. The technical challenge is not just one of handling complicated formalisms but also mathematical, the key correctness proof requiring the adaptation of the step-indexed relations technique [2] to bisimulation.

Note that we are language rather than system oriented. We assume the existence of a run-time infrastructure to handle system-level aspects associated with distribution such as failure detection, load balancing, global reset and initialisation, and so on. From a systems perspective we only deal with failure handling, which is particularly important in a distributed setting, using a general transactional approach.

The networks of abstract machines can serve as a target for the compilation of a conventional call-by-value language which we call Floskel. The interesting new thing about it is that there is almost nothing new about it. It uses a Haskell-like syntax but it has the semantics of the pure fragment of OCaml. Remote procedure calls are indistinguishable from native calls except for an annotation, which can be applied to any sub-term, indicating that it is to be executed on a different node. From the point of view of the language, and of the programmer, this annotation has no syntactic, semantic or typing implication. It is a mere pragma-like directive.

Prior work on native RPCs and seamless distribution is specialised to web programming [11], thus domain-specific. We aim to be as generic as possible in this context. Some existing work uses as a starting point interaction-based semantic paradigms which lend themselves naturally to a communication-centric implementation: Geometry of Interaction [17] and Game Semantics [18]. Such approaches have two significant disadvantages. The exotic operational behaviour makes it impossible to apply known optimisation techniques, and to interact with code compiled conventionally. This can be seen in the low performance of single-node execution of programs compiled using such techniques. On the other hand, the single-node compilation of Floskel is very similar to the conventional compilation of a language such as OCaml, and the bench-
marks indicate that the overhead required by the RPC run-time is not excessive.

In a nutshell, we believe that this paper can help make the case that functional languages with native RPCs could be a lot more useful than presumed.

1.1 Technical outline and contributions
Our main contributions are in the following areas:

Compiler and run-time We describe the syntax (Sec. 2.1) and implementation (Sec. 2.2) of Floskel, a general-purpose functional language with native RPCs. Our basis is a conventional compiler for such a language, and we show how it is modified to support RPCs and, additionally, ubiquitous functions, i.e. functions available on all nodes. Our benchmarks suggest that Floskel’s performance is comparable to the state of the art OCaml compiler (Sec. 2.4) for single-node execution.

Abstract machines The semantics of a core of Floskel has been formalised in Agda (Sec. 3) in the form of an abstract machine that can be used to guide an implementation. To achieve this we make gradual refinements to a machine, based on Landin’s SECD machine [26], that we call the CES machine (Sec. 3.1). First we add heaps for dynamically allocating closures, forming the CESH machine (Sec. 3.2); we show that the CES and CESH execution is bisimilar. We then add communication primitives (synchronous and asynchronous) by defining a general form of networks of nodes that run an instance of an underlying abstract machine (Sec. 3.3). Using these networks, we illustrate the idea of subsuming function calls by communication protocols by constructing a degenerate distributed machine, DCESH₁ (Sec. 3.4), that decomposes some machine instructions into message passing, but only runs on one node. Execution on the fully distributed CESH machine called DCESH (Sec. 3.5), is shown to be bisimilar to the CESH machine — our main theoretical result. Finally, we model a general-purpose fault-tolerant environment for DCESH-like machines (Sec. 4) by layering a transactional abstract machine that provides a simple commit-and-rollback mechanism for an abstract machine that may unexpectedly fail.

Formalisation in Agda The theorems that we present in this paper have been proved correct in Agda [34], an interactive proof assistant and programming language based on intuitionistic type theory. The definitions and proofs in this paper are intricate, so carrying them out manually would be error-prone, arduous and perhaps unconvincing. Agda has been a helpful tool in producing these proofs, providing a pleasant interactive environment in which to play with alternative definitions. To eliminate another source of error, we do not present our results in informal mathematics; the code blocks in Sec. 3 come directly from the formalisation.

The formalisation is organised as follows, where the arrows denote dependence, the lines with ~ symbols bisimulations, and the parenthesised numerals section numbers:

2. Floskel: a location-aware language

2.1 Syntax
At the core of the Floskel language is a call-by-value functional language with user-definable algebraic data types and pattern matching. Floskel is semantically similar to languages in the ML family, and syntactically similar to Haskell. The main syntactic difference between Floskel and Haskell is that pattern matching clauses are given without the leading function name, to avoid repetition, and that type annotations are given after a single colon, as in the following example:

\[
\begin{align*}
\text{map} &: (a \to b) \to [a] \to [b] \\
\text{f} &[] = [] \\
\text{f} (x::xs) = \text{f} x :: \text{map} \text{f} xs 
\end{align*}
\]

Node annotations An ordinary function definition, like map, is a ubiquitous function by default. This means that it is made available on all nodes in the system, and a call to such a function is always done locally – a plain old function call.

On the other hand, a function or sub-term defined with a node annotation, such as

\[
\begin{align*}
\text{query} @ \text{Database} &\ : \text{Query} \to \text{Response} \\
\text{f} @ A &\ : \ (\text{Query} \to \text{Response}) \to X \\
q &\ = \ \ldots \ \text{use} q \ldots 
\end{align*}
\]

can be applied to query yielding \( \text{f} \text{query} \) : \( X \).

Node annotations can also be applied to sub-expressions, as in the following example:

\[
\begin{align*}
\text{sum} [] &\ = \ 0 \\
\text{sum (x::xs)} &\ = \ x + \text{sum} \text{xs} \\
\text{xs} @ A &\ = \ \ldots \\
\text{ys} @ B &\ = \ \ldots \\
\text{result} @ C &\ = \ \langle \text{sum} \text{xs} \rangle @ A + \langle \text{sum} \text{ys} \rangle @ B 
\end{align*}
\]

Here we want to calculate the sum, on node \( C \), of the elements of two lists located on nodes \( A \) and \( B \). If the lists are lengthy, it is better to calculate the sums on \( A \) and \( B \), and to then send the final sum to \( C \), since this saves us having to send the full lists over the network.

2.2 Compilation
The Floskel compiler [1] currently targets C using the MPI library for communication, though other targets are possible since we do not make use of any exotic features of our target. Any compatible combination of low-level language and message passing library would work. Most of the compiler’s pipeline is standard for a functional language implementation. It works by applying a series of transformations to the source program until reaching a level low enough to be straightforwardly translated to C. Since the source language has pattern matching, it first compiles the pattern matching to simple case trees [4]. Local definitions are then lifted to the top-level using lambda lifting [25], and lastly the program is closure converted [31] to support partially applied functions.

Up until the lambda lifting, a node annotation is a constructor in the abstract syntax tree of the language’s expressions. The lambda lifter lifts such sub-expressions to the top-level such that
annotations are afterwards associated with definitions (and not expressions).

The main work specific to Floskel is done in the closure conversion and the run-time system that the compiled programs make use of.

**Closures** For function applications, the closure converter distinguishes between known functions – those that are on the top-level and have a known arity, and unknown functions – those that are provided e.g. as function arguments.

A known function \( f \) that is either ubiquitous or available on the same node as the definition that is being compiled is compiled to an ordinary function call if there are enough arguments. If there are not, and the function is ubiquitous we have to construct a *partial application* closure, which contains a pointer to the function and the arguments so far. The compiler maintains the invariant that unknown functions are always in the form of a closure, whose general layout is:

\[
\text{generic layout is:} \quad \text{payload} \quad \text{arity} \quad \text{gptr} \quad \text{fptr} \quad \ldots
\]

Since the function may require access to the payload of the closure, \( \text{gptr} \) is a function of arity \( \text{arity} + 1 \): when applying a closure \( c \) as above to arguments \( x_1, \ldots, x_\text{arity} \), the call becomes \( \text{gptr} (c, x_1, \ldots, x_\text{arity}) \) meaning that the function has access to the payload through \( c \). To construct the initial closure for a partial application of a function \( f \) of arity \( \text{arity} \) with \( \text{nargs} \) arguments, we have to conform to this rule, so we construct the closure \( \text{cl}_n (f, x_1, \ldots, y, x_{\text{nargs}}) \) where \( n = \text{arity} - \text{nargs} \) and \( f' \) is a new ubiquitous top-level function defined as follows:

\[
f' c \ x_1 \ldots \ x_n = \text{case } c \text{ of } (\ldots \ y \ldots ) \rightarrow f (y, \ldots, y_{\text{nargs}}, x_1, \ldots, x_n)
\]

A family of *apply* functions handle, in a standard way, applications of \( f \) arguments of unknown functions by inspecting the arity stored in the closure to decide whether to construct a new partial application closure with the additional arguments or to apply the function.

The field \( \text{fptr} \) is an integer identifier assigned to every function at compile-time and used as a system-wide identifier if the function is ubiquitous, or a node-specific identifier if not. If there are \( k \) ubiquitous functions they are assigned the first \( k \) identifiers, and the nodes of the system may use identifiers greater than \( k \) for their respective located functions. Determining if a function is ubiquitous is thus a simple comparison: \( \text{fptr} < k \). Additionally, every node has a table of functions that maps ubiquitous or local located function identifiers to local function pointers, which is used by the deserialiser.

If we have a saturated call to a known remote function, we make a call to the function \( \text{rApply}_\text{arity} \), defined in the run-time system (to be described). If we have a non-saturated call to a known remote or located function, we construct the closure \( \text{cl}_n (f, x_1, \ldots, y, x_{\text{nargs}}) \) where \( f' \) is a new ubiquitous top-level function defined as follows:

\[
f' c \ x_1 \ldots \ x_n = \text{case } c \text{ of } (\ldots \ y \ldots ) \rightarrow f (y, \ldots, y_{\text{nargs}}, x_1, \ldots, x_n)
\]

A family of *apply* functions handle, in a standard way, applications of \( f \) arguments of unknown functions by inspecting the arity stored in the closure to decide whether to construct a new partial application closure with the additional arguments or to apply the function.

The field \( \text{fptr} \) is an integer identifier assigned to every function at compile-time and used as a system-wide identifier if the function is ubiquitous, or a node-specific identifier if not. If there are \( k \) ubiquitous functions they are assigned the first \( k \) identifiers, and the nodes of the system may use identifiers greater than \( k \) for their respective located functions. Determining if a function is ubiquitous is thus a simple comparison: \( \text{fptr} < k \). Additionally, every node has a table of functions that maps ubiquitous or local located function identifiers to local function pointers, which is used by the deserialiser.

If we have a saturated call to a known remote function, we make a call to the function \( \text{rApply}_\text{arity} \), defined in the run-time system (to be described). If we have a non-saturated call to a known remote or located function, we construct the closure \( \text{cl}_n (f, x_1, \ldots, y, x_{\text{nargs}}) \) where \( f' \) is a new ubiquitous top-level function defined as follows:

\[
f' c \ x_1 \ldots \ x_n = \text{case } c \text{ of } (\ldots \ y \ldots ) \rightarrow f (y, \ldots, y_{\text{nargs}}, x_1, \ldots, x_n)
\]

A family of *apply* functions handle, in a standard way, applications of \( f \) arguments of unknown functions by inspecting the arity stored in the closure to decide whether to construct a new partial application closure with the additional arguments or to apply the function.

The field \( \text{fptr} \) is an integer identifier assigned to every function at compile-time and used as a system-wide identifier if the function is ubiquitous, or a node-specific identifier if not. If there are \( k \) ubiquitous functions they are assigned the first \( k \) identifiers, and the nodes of the system may use identifiers greater than \( k \) for their respective located functions. Determining if a function is ubiquitous is thus a simple comparison: \( \text{fptr} < k \). Additionally, every node has a table of functions that maps ubiquitous or local located function identifiers to local function pointers, which is used by the deserialiser.

If we have a saturated call to a known remote function, we make a call to the function \( \text{rApply}_\text{arity} \), defined in the run-time system (to be described). If we have a non-saturated call to a known remote or located function, we construct the closure \( \text{cl}_n (f, x_1, \ldots, y, x_{\text{nargs}}) \) where \( f' \) is a new ubiquitous top-level function defined as follows:

\[
f' c \ x_1 \ldots \ x_n = \text{case } c \text{ of } (\ldots \ y \ldots ) \rightarrow f (y, \ldots, y_{\text{nargs}}, x_1, \ldots, x_n)
\]

A family of *apply* functions handle, in a standard way, applications of \( f \) arguments of unknown functions by inspecting the arity stored in the closure to decide whether to construct a new partial application closure with the additional arguments or to apply the function.

The field \( \text{fptr} \) is an integer identifier assigned to every function at compile-time and used as a system-wide identifier if the function is ubiquitous, or a node-specific identifier if not. If there are \( k \) ubiquitous functions they are assigned the first \( k \) identifiers, and the nodes of the system may use identifiers greater than \( k \) for their respective located functions. Determining if a function is ubiquitous is thus a simple comparison: \( \text{fptr} < k \). Additionally, every node has a table of functions that maps ubiquitous or local located function identifiers to local function pointers, which is used by the deserialiser.

If we have a saturated call to a known remote function, we make a call to the function \( \text{rApply}_\text{arity} \), defined in the run-time system (to be described). If we have a non-saturated call to a known remote or located function, we construct the closure \( \text{cl}_n (f, x_1, \ldots, y, x_{\text{nargs}}) \) where \( f' \) is a new ubiquitous top-level function defined as follows:

\[
f' c \ x_1 \ldots \ x_n = \text{case } c \text{ of } (\ldots \ y \ldots ) \rightarrow f (y, \ldots, y_{\text{nargs}}, x_1, \ldots, x_n)
\]

A family of *apply* functions handle, in a standard way, applications of \( f \) arguments of unknown functions by inspecting the arity stored in the closure to decide whether to construct a new partial application closure with the additional arguments or to apply the function.

The field \( \text{fptr} \) is an integer identifier assigned to every function at compile-time and used as a system-wide identifier if the function is ubiquitous, or a node-specific identifier if not. If there are \( k \) ubiquitous functions they are assigned the first \( k \) identifiers, and the nodes of the system may use identifiers greater than \( k \) for their respective located functions. Determining if a function is ubiquitous is thus a simple comparison: \( \text{fptr} < k \). Additionally, every node has a table of functions that maps ubiquitous or local located function identifiers to local function pointers, which is used by the deserialiser.

If we have a saturated call to a known remote function, we make a call to the function \( \text{rApply}_\text{arity} \), defined in the run-time system (to be described). If we have a non-saturated call to a known remote or located function, we construct the closure \( \text{cl}_n (f, x_1, \ldots, y, x_{\text{nargs}}) \) where \( f' \) is a new ubiquitous top-level function defined as follows:

\[
f' c \ x_1 \ldots \ x_n = \text{case } c \text{ of } (\ldots \ y \ldots ) \rightarrow f (y, \ldots, y_{\text{nargs}}, x_1, \ldots, x_n)
\]

A family of *apply* functions handle, in a standard way, applications of \( f \) arguments of unknown functions by inspecting the arity stored in the closure to decide whether to construct a new partial application closure with the additional arguments or to apply the function.

The field \( \text{fptr} \) is an integer identifier assigned to every function at compile-time and used as a system-wide identifier if the function is ubiquitous, or a node-specific identifier if not. If there are \( k \) ubiquitous functions they are assigned the first \( k \) identifiers, and the nodes of the system may use identifiers greater than \( k \) for their respective located functions. Determining if a function is ubiquitous is thus a simple comparison: \( \text{fptr} < k \). Additionally, every node has a table of functions that maps ubiquitous or local located function identifiers to local function pointers, which is used by the deserialiser.
### 3. Abstract machines and networks

Having introduced the programming language, its compiler and its run-time system we now present the theoretical foundation for the correctness of the compiler. We start with the standard abstract machine model of CBV computation, which we refine, in several steps, into increasingly expressive abstract machines with heap and networking capabilities, while showing along the way that correctness is preserved, via bisimulation results. All definitions and theorems are formalised using the proof assistant Agda, the syntax of which we will follow. We give some of the key definitions and examples in Agda syntax, but most of the description of the formalisation is intended as a high-level guide to the Agda code, which is available online [1]. In order to help the reader navigate the code when a significant theorem or lemma is mentioned the fully qualified Agda name is given in a footnote. Note that we shall not formalise the whole of Floskel but only a core language which coincides with Plotkin’s (untyped) call-by-value PCF [37].

#### 3.1 The CES machine

The starting point is a variation of Landin’s SECD machine [26] called Modern SECD [27], which can be traced to the SECD machine variation of Henderson [22] and to the CEK machine of Felleisen [15], which we call CES (Agda module CES). Just like the machine of Henderson, it uses bytecode for the control component of the machine, and just like the CEK it places the continuations that originally resided in the dump directly on the stack, simplifying the machine configurations.

A CES configuration (Config) is a tuple consisting of a fragment of code (Code), an environment (Env), and a stack (Stack). Evaluation begins with an empty stack and environment, and then follows a stack discipline. Sub-terms push their result on the stack so that their super-terms can consume them. When (and if) the evaluation terminates, the program’s result is the sole stack element.

**Source language** The source language has constructors for lambda abstractions ($\lambda$, applications ($\langle f \ t \rangle$), and variables (var $n$), represented using De Bruijn indices [13], so a variable is a natural number. Additionally, we have natural number literals (lit $n$), binary operations (op $f\ t\ t'$), and a conditional (if0 $\mathit{then}\ t\ \mathit{else}\ t'$). Because the language is untyped, we can express a fixed-point combinator without adding additional constructors.

The machine operates on bytecode and does not directly interpret the source terms, so the terms need to be compiled before they can be executed. The main work of compilation is done by the function compile’, which takes a term $t$ and a fragment of code $c$ used as a postlude. The bold upper-case names (CLOS, VAR, and so on) are the bytecode instructions, which are sequenced using _;_. Instructions can be seen to correspond to the constructs of the source language, sequenced.

**Example 3.1.** To compile a term $t$ we supply END as a postlude: compile $t$ = compile’ $t$ END. The term $t = \langle x \ s\ \langle \lambda\ y\ s\ \lambda\ y\ r\ t\ \rangle \rangle$ is compiled as follows:

```
compile (\lambda\ y\ r\ t\ \rangle \rangle) = CLOS (VAR \ r\ \lambda\ \lambda\ y) ; CLOS (CLOS (VAR S\ \lambda\ \lambda\ y)) ; CLOS (V AR \ r\ \lambda\ \lambda\ y) ; CLOS (APPL \ r\ \lambda\ \lambda\ y) ; CLOS (END)
```

Environments (Env) are lists of values (List Value), which are either natural numbers (nat $n$) or closures (clos $\mathbf{e}$). A closure (Close) is a fragment of code paired with an environment (Code $\times$ Env). Stacks (Stack) are lists of stack elements (List StackElem), which are either values (val $v$) or continuations (cont $\mathbf{c}$), represented by closures.

Fig. 3 shows the definition of the transition relation for configurations of the CES machine. A note on Agda syntax: the instruction constructor names are overloaded as constructors for the relation; their usage is disambiguated by context.
curly braces are *implicit* and can be automatically inferred. Equality (propositional) is written \( \equiv \).

Stack discipline is clear in the definition of the transition relation. When e.g. `VAR` is executed, the CES machine looks up the value of the variable in the environment and pushes it on the stack. A somewhat subtle part of the relation is the interplay between the `APPL` instruction and the `RET` instruction. When performing an application, two values are required on the stack, one of which has to be a closure. The machine enters the closure, adding the value to the environment, and pushes a return continuation on the stack. The code inside a closure will be terminated by a `RET` instruction, so once the machine has finished executing the closure (and thus produced a value on the stack), that value is returned to the continuation. Note that it will be useful to establish that the CES machine is deterministic.

**Example 3.2.** We trace the execution of Ex. 3.1 defined above, which exemplifies how returning from an application works. Here we write \( a \rightarrow_{CES} \{ x \} b \) meaning that the machine uses rule \( x \) to transition from \( a \) to \( b \).

\[
\begin{align*}
\text{let } c_1 &= \text{VAR } 0 \text{ ; RET } \\
&\rightarrow_{CES} \{ \text{VAR } 0 \} \text{ ; RET } \\
\text{let } c_2 &= \text{CLOS } \{ \text{VAR } 0 \} \text{ ; RET } \\
&\rightarrow_{CES} \{ \text{VAR } 0 \} \text{ ; RET } \\
\text{let } c_1 &= \text{val } \{ \text{CLOS } (c_1, []) \} ; c_2 &= \text{val } \{ \text{CLOS } (c_2, []) \} \\
&\rightarrow_{CES} \{ \text{CLOS } (c_1, []) \} \text{ ; APPL } \{ \text{END } [1] \} , \{ c_1 \} \\
&\rightarrow_{CES} \{ \text{CLOS } (c_2, []) \} \text{ ; APPL } \{ \text{END } [1] \} , \{ c_2 \} \\
&\rightarrow_{CES} \{ \text{VAR } 0 \} \text{ ; RET } , \{ c_2 \} , \{ \text{cont } \{ \text{END } [1] \} \} \\
&\rightarrow_{CES} \{ \text{VAR } 0 \} \text{ ; RET } , \{ c_2 \} , \{ \text{cont } \{ \text{END } [1] \} \} \\
\text{The final result is therefore the second closure, } c_2. \\
\end{align*}
\]

The CES machine terminates with a value \( v \), written \( cfig \downarrow_{CES} v \) if it, through the reflexive transitive closure of \( \rightarrow_{CES} \), reaches the end of its code fragment with an empty environment, and \( v \) as its sole stack element. It terminates, written \( cfig \downarrow_{CES} \) if there exists a value \( v \) such that it terminates with the value \( v \). It diverges, written \( cfig \uparrow_{CES} \) if it is possible to take another step from any configuration reachable from the reflexive transitive closure of \( \rightarrow_{CES} \).

We do not prove formally that the compilation of CBV-PCF to the CES machine is correct, as it is a standard result [12].

### 3.2 CESH: A heap machine

In a compiler implementation of the CES machine targeting a low-level language, closures have to be dynamically allocated in a heap. However, the CES machine does not make this dynamic allocation explicit. Now we make it explicit in a new machine, called the CESH, which is a CES machine with an extra heap component in its configuration.

While heaps are not strictly necessary for a *presentation* of the CES machine, they are of great importance to us. The distributed machine that we will later define needs heaps for persistent storage of data, and the CESH machine forms an intermediate step between that and the CES machine. A CESH configuration is defined as

\[
\text{Config} = \text{Code } \times� \text{Env } \times� \text{Stack } \times� \text{Heap}
\]

where *Heap* is a type constructor for heaps parameterised by the type of its content. The only difference in the definition of the configuration constituents, compared to the CES machine, is that a closure value (the `clos` constructor of the *Value* type) does not contain an actual closure, but just a pointer (`Ptr`). The stack is as in the CES machine.

**Figure 3.** The definition of the transition relation of the CES machine.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>VAR</code></td>
<td>( \forall { n \ c \ e \ s } \rightarrow \text{lookup } n \ e \equiv \text{just } v \rightarrow (\text{VAR } n ; c, e, s) \rightarrow_{CES} (c, e, \text{val } v :: s) )</td>
</tr>
<tr>
<td><code>CLOS</code></td>
<td>( \forall { c' \ c \ e \ s } \rightarrow (\text{CLOS } c' ; c, e, s) \rightarrow_{CES} (c, e, \text{val } (\text{clos } (c', e)) :: s) )</td>
</tr>
<tr>
<td><code>APPL</code></td>
<td>( \forall { c \ e v c' e' s } \rightarrow (\text{APPL } c ; c, e, \text{val } v :: \text{val } (\text{clos } (c', e')) :: s) \rightarrow_{CES} (c', e', \text{val } v :: s) )</td>
</tr>
<tr>
<td><code>RET</code></td>
<td>( \forall { e v c' e' s } \rightarrow (\text{RET } e, \text{val } v :: \text{cont } (c, e') :: s) \rightarrow_{CES} (c, e', \text{val } v :: s) )</td>
</tr>
<tr>
<td><code>LIT</code></td>
<td>( \forall { n c e s } \rightarrow (\text{LIT } n ; c, e, \text{val } (\text{nat } n) :: s) \rightarrow_{CES} (c, e, \text{val } (\text{nat }n) :: s) )</td>
</tr>
<tr>
<td><code>OP</code></td>
<td>( \forall { f c e n_1 n_2 s } \rightarrow (\text{OP } f ; c, e, \text{val } (\text{nat } n_1) :: \text{val } (\text{nat } n_2) :: s) \rightarrow_{CES} (c, e, \text{val } (\text{nat } (f n_1 n_2)) :: s) )</td>
</tr>
<tr>
<td><code>COND-0</code></td>
<td>( \forall { c' c e s } \rightarrow (\text{COND } c' c ; e, \text{val } (\text{nat } 0) :: s) \rightarrow_{CES} (c, e, s) )</td>
</tr>
<tr>
<td><code>COND-1+n</code></td>
<td>( \forall { c' c e n s } \rightarrow (\text{COND } c' c ; e, \text{val } (\text{nat } (1 + n)) :: s) \rightarrow_{CES} (c', e, s) )</td>
</tr>
</tbody>
</table>

3.2.1 Correctness

To show that our definition of the machine is correct, we construct a bisimulation between the CES and CESH, which gives the similarity between the two machines, is almost equality. The difference is dealing with closure values, since the CESH stores pointers rather than closures. The relation for closure values must be parameterised by the heap of the CESH configuration, where the (dereferenced) value of the closure pointer is related to the CES closure.

Formally, the relation is constructed separately for the different components of the machine configurations. For bytecode it is equality, and for closures it is defined component-wise. Values are related only if they have the same heap constructor and related components: if the two values are number literals, they are related if they are equal; a CES closure and a pointer are related only if the pointer leads to a CESH closure that is in turn related to the CES closure. Environments are related if they have the same list spine and their values are pointwise related. The relation on stacks is defined similarly, using the relation on values and continuations. Finally, two configurations are related if their components are related.

In the formalisation we define heaps and their properties *abstractly*, rather than using a specific heap implementation. The first key property we require is that dereferencing a pointer in a heap...
where that pointer was just allocated with a value gives back the same value:

\[ \forall h \ x \rightarrow \text{let} (h', \text{ptr}) = h \triangleright x \text{ in } h' ! \text{ ptr} \triangleright \text{ just } x \]

We will require a preorder \( \subseteq \) for sub-heaps. The intuitive reading for \( h \subseteq h' \) is that \( h' \) can be used where \( h \) can, i.e. that \( h' \) contains at least the allocations of \( h \). The second key property that we require of a heap implementation is that allocation does not overwrite any previously allocated memory cells (\( \text{proj} \) means first projection):

\[ \forall h \ x \rightarrow h \subseteq \text{proj}_h (h \triangleright x) \]

For any two heaps \( h \) and \( h' \) such that \( h \subseteq h' \), if \( R_{\text{CES}} \text{ cf}_{\text{g}} (c, e, s, h) \), then \( R_{\text{CES}} \text{ cf}_{\text{g}} (c, e, s, h') \).

Our first correctness result is that \( R_{\text{CES}} \) is a simulation relation.\(^3\)

The proof is by cases on the CES transition, and, in each case, the CESH machine can make analogous transitions. The property mentioned above is then used to show that \( R_{\text{CES}} \) is preserved.

It is helpful to introduce the notion of a presimulation relation – a generalisation of a simulation relation that does not require the target states of the transitions to be related – formally defined as:

**Presimulation** \( \xrightarrow{\text{sim}} \xrightarrow{\text{sim}}' \xrightarrow{\text{sim}}' \text{R} = \)

\[ \forall a \ a' b \rightarrow (a \xrightarrow{\text{sim}} a') \rightarrow a \ R b \rightarrow \exists \lambda b' \rightarrow (b \xrightarrow{\text{sim}}' b') \]

Then, the inverse of \( R_{\text{CES}} \) is a presimulation.\(^4\) In general, if \( R \) is a simulation relation between states \( \rightarrow \) and \( \rightarrow' \), \( R^{-1} \) is a presimulation, and \( \rightarrow' \rightarrow' \) is deterministic at states \( b \) related to some \( a \), then \( R^{-1} \) is a simulation,\(^5\) from which it follows that \( R_{\text{CES}} \) is a bisimulation, because we have already established that the CESH machine is deterministic.

In particular, if \( R_{\text{CES}} \text{ cf}_{\text{g1}} \text{ cf}_{\text{g2}} \) then \( \text{cf}_{\text{g1}} \triangleleft_{\text{CES}} \text{nat} n \leftrightarrow \text{cf}_{\text{g2}} \triangleleft_{\text{CES}} \text{nat} n \text{ and cf}_{\text{g1}} \triangleleft_{\text{CES}} \leftrightarrow \text{cf}_{\text{g2}} \triangleleft_{\text{CES}} \).

To finalise the proof we note that there are configurations in \( R_{\text{CES}} \). One such example is the initial configuration for a fragment of code: For any \( c \), we have \( R_{\text{CES}} (c, [], [], [], s, \emptyset) \) (where \( \emptyset \) is the empty heap).

### 3.3 Network models

In this section we define models for networks with synchronous and asynchronous communication, that are parameterised by an underlying labelled transition system. Both kinds of networks are modelled by two-level transition systems, which is common in operational semantics for concurrent and parallel languages. A **global level** describes the transitions of the system as a whole, and a **local level** the local transitions of the nodes in the system.

Synchronous communication is modelled by **rendezvous**, i.e. two nodes have to be ready to send and receive a message at a single point in time. Asynchronous communication is modelled using a "message soup", representing messages currently in transit, that nodes can add and remove messages from, reminiscent of the Chemical Abstract Machine [5].

The model (Agda module Network) is parameterised by the underlying transition relation of the machines \( \rightarrow_{\text{Mach}} \rightarrow \). The sets \( \text{Node} \), \( \text{Machine} \), and \( \text{Msg} \) are additional parameters. Elements of \( \text{Node} \) will act as node identifiers, and we assume that these enjoy decidable equality – in MPI, they would correspond to the so called integer "node ranks". The **Type Machine** is the type of the nodes' configurations, and \( \text{Msg} \) the type of messages that the machines can send. The presence of the \( \text{Node} \) argument means that the configuration of a node may know about and can depend on its own identifier.

A synchronous network (SyncNetwork) is an indexed family of machines, \( \text{Node} \rightarrow \text{Machine} \), representing the nodes of the system. An asynchronous network (AsyncNetwork) is an indexed family of machines together with a list of pending messages \( (\text{Node} \rightarrow \text{Machine}) \times \text{List} \text{Msg} \).

### 3.4 DCESH\(_1\): A trivially distributed machine

In higher-order distributed programs containing location specifiers, we will sometimes encounter situations where a function is not available locally. For example, when evaluating the function \( f \) in the term \( f \, x \, A \) (\( g \, x \, B \)) we may need to apply the remotely available function \( g \). Our general idea is to do this by decomposing some instructions into communication. In the example, the function \( f \) may send a message requesting the evaluation of \( g \), meaning that the APPL instruction is split into a pair of instructions: APPL-send and APPL-receive.

This section outlines an abstract machine, called DCESH\(_1\), which decomposes all application and return instructions into communication. The machine is trivially distributed, because it runs as the sole node in a network, sending messages only to itself. Although it is not used as an intermediate step for the proofs, it is included because it illustrates this decomposition.

A configuration of the DCESH\(_1\) machine (\( \text{Machine} \)) is a tuple consisting of a possibly running thread (\( \text{Maybe Thread} \)), a closure heap (\( \text{Heap Closure} \)), and a “continuation heap” (\( \text{Heap Closure} \times \text{Stack} \)). Since the language is sequential we have at most one thread

---

\(^3\) CESH.Simulation.HeapUpdate.config
\(^4\) CESH.Simulation.simulation
\(^5\) CESH.Bisimulation.pretension-to-simulation
\(^6\) CESH.Bisimulation.termination-agrees, CESH.Bisimulation.divergence-agrees
\(^7\) Network.\(\rightarrow\text{Sync-to} \rightarrow \rightarrow\text{Async} \).
Figure 5. The definition of the transition relations for synchronous and asynchronous networks.

running at once. The thread resembles a CES configuration, Thread = Code × Env × Stack, but stacks are defined differently. A stack is now a list of values paired with an optional pointer (into the continuation heap), Stack = List Val × Maybe ContPtr (ContPtr is a synonym for Ptr). When performing an application, when CES would push a continuation on the stack, the DCESH machine is going to stop the current thread and send a message, which means that it has to save the continuation and the remainder of the stack in the heap for them to persist the thread’s lifetime.

The optional pointer in Stack is an element at the bottom of the list of values. Comparing it to the definition of the CES machine, where stacks are lists of either values or continuations (just closures), we can picture their relation: Whereas the CES machine stores the values and continuations in a single, contiguous stack, the DCESH machine stores first a contiguous block of values until reaching a continuation, at which point it stores (just) a pointer to the continuation closure and the rest of the stack.

The definition of closures, values, and environments are otherwise just like in the CESH machine. The machine communicates with itself using two kinds of messages, APPL and RET, corresponding to the instructions that we are replacing with communication.

Fig. 6 defines the transition relation for the DCESH machine, written \( m \xrightarrow{msg} m' \) for a tagged message msg and machine configurations m and m'. Most transitions are the same as in the CESH machine, just framed with the additional heaps and the just meaning that the thread is running. We elide them for brevity.

The interesting rules are the decomposed application and return rules. When an application is performed, an APPL message containing a pointer to the closure to apply, the argument value and a pointer to a return continuation (which is first allocated) is sent, and the thread is stopped (nothing). We call such a machine inactive. The machine can receive an application message if the thread is not running. When that happens, the closure pointer is dereferenced and entered, adding the received argument to the environment. The stack is left empty apart from the continuation pointer of the received message. When returning from a function application, the machine sends a return message containing the continuation pointer and the value to return.

On the receiving end of that communication, it dereferences the continuation pointer and enters it, putting the result value on top of the stack.

Example 3.3. We trace the execution of Ex. 3.1 in a synchronous network of nodes indexed by the unit type. Heaps with pointer mappings are written \( \{ptr \rightarrow element\} \). The last list shown in each step is the message list of the asynchronous network.

\[
\text{silent-step} : \forall \{i m'\} \rightarrow i \rightarrow \text{nodes} i \rightarrow m' \rightarrow \text{nodes} \xrightarrow{\text{Update nodes } i \text{ } m'}
\]

\[
\text{comm-step} : \forall \{s \rightarrow \text{msg } \text{sender' receiver'}\} \rightarrow \text{let nodes' } = \text{update nodes } s \text{ } \text{sender' } \text{in}
\]

\[
s \rightarrow \text{nodes } s \xrightarrow{\text{msg }} \text{sender' } \rightarrow r \rightarrow \text{nodes' } r \xrightarrow{\text{msg }} \text{receiver' } \rightarrow \text{nodes} \xrightarrow{\text{Sync}} \text{update nodes' } r \text{receiver' }.
\]

\[
\text{step} : \forall \{\text{nodes } m \text{msg}_{s} \text{msg}_{r} \{\text{msg}_{s} m' i\} \rightarrow \text{let (msg}_{s_{i}} \text{msg}_{r_{i}}\} = \text{detag msg}_{i} \in i \rightarrow \text{nodes } i \xrightarrow{\text{msg }} m' \rightarrow \text{nodes} \xrightarrow{\text{Async}} \text{update nodes } i \text{ } m', \text{msg}_{s} + \text{msg}_{r} + \text{msg}_{s_{i}} \}
\]

\[
\text{just (APPL ; END ; [ ]}, \text{closer ptr1 closer ptr1 }, \text{nothing} , h'_{ci}, \emptyset, [ ] \longrightarrow \text{(step APPL-send)}
\]

\[
\text{nothing, h'_{ci}, h_{iota}} , \{APPL , \text{ptr1 (closer ptr2) ptr_{iota}} \}
\]

\[
\longrightarrow \text{(step APPL-receive)}
\]

\[
\text{just (VAR } h_{\emptyset} \text{ }; \text{RET } \{\text{closer ptr2} , [ ]}, \text{just ptr_{iota}} , h'_{ci}, h_{iota} , [ ] \}
\]

\[
\longrightarrow \text{(step VAR refl)}
\]

\[
\text{just (RET , \text{closer ptr2} , [ ]}, \text{just ptr_{iota}} , h'_{ci}, h_{iota} , [ ] \}
\]

\[
\longrightarrow \text{(step RET-send)}
\]

\[
\text{nothing, h'_{ci}, h_{iota}} , \{RET , \text{ptr_{iota}} (closer ptr2) \}
\]

\[
\longrightarrow \text{(step RET-receive)}
\]

\[
\text{just (END ; [ ]}, \{\text{closer ptr2} , [ ]}, \text{nothing} , h'_{ci}, h_{iota} , [ ] \}
\]

Comparing this to Example 3.2 we can see that an APPL-send followed by an APPL-receive amounts to the same thing as the APPL rule in the CES machine, and similarly for the RET instruction.

3.5 DCESH: The distributed CESH machine

We have so far seen two refinements of the CES machine. We have seen CESH, that adds heaps, and DCESH1, that decomposes instructions into communication in a degenerate network of only one node. Our final refinement is a distributed machine, DCESH, that supports multiple nodes. The main problem that we now face is that there is no centralised heap, but each node has its own local heap. This means that, for supporting higher-order functions across node boundaries, we have to somehow keep references to closures in the heaps of other nodes. Another problem is efficiency; we would like a system where we do not pay the higher price of communication for locally running code. The main idea for solving these two problems is to use remote pointers, \( Rptr = \text{Ptr} \times \text{Node} \). Pointers paired with node identifiers signifying on what node’s heap the pointer is located. This solves the heap problem because we always know where a pointer comes from. It can also be used to solve the efficiency problem since we can choose what instructions to run based on whether a pointer is local or remote. If it is local, we run the rules of the CESH machine. If it is remote, we run the decomposed rules of the DCESH1 machine.

The final extension to the term language and bytecode will add support for location specifiers. We add a term construct \( i \ast i \), and an instruction \text{REMOTE c i} for its compilation. The location specifiers, \( r \ast i \), are taken to mean that the term \( r \) should be evaluated on node \( i \). For compilation, we require that the terms \( r \) in all location specification sub-terms \( r \ast i \) are closed. Terms where this does not hold are transformed automatically using lambda lifting [25] (transform every sub-term \( r \ast i \) to \( r = \langle \lambda x, f, t, i \rangle (f \circ t) \). The \text{REMOTE c i} instruction will be used to start running a code fragment \( c \) on node \( i \) in the network. We also extend the \text{compile'} function to handle the new term construct:

\[
\text{compile'} (\ast i) c = \text{REMOTE (compile' RET i) i ; c}
\]

Note that we reuse the RET instruction to return from a remote computation.

The definition of closures, values, environments and closure heaps are the same as in the CESH machine, but using \( Rptr \) instead of \( Ptr \) for closure pointers.
The stack combines the functionality of the CES(H) machine, permitting local continuations, with that of the DCESH\_1 machine, making it possible for a stack to end with a continuation on another node. A stack element is a value or a (local) continuation signified by the \texttt{val} and \texttt{cont} constructors. A stack (Stack) is a list of stack elements, possibly ending with a (remote) pointer to a continuation, \texttt{List StackElem \times Maybe ContPtr (where ContPtr = RPtr)}. Threads and machines are defined like in the DCESH\_1 machine. The messages that DCESH can send are those of the DCESH\_1 machine but using remote pointers instead of plain pointers, plus a message for starting a remote computation, \texttt{REMOTE c i rptr}. Note that sending a \texttt{REMOTE} message amounts to sending code in our formalisation, which is something that we would not like to do. However, because no code is generated at run-time, every machine can be “pre-loaded” with all the bytecode it needs, and the message only needs to contain a \texttt{reference} to a fragment of code.

Fig. 7 defines the transition relation of the DCESH machine, written \( i \vdash m \Rightarrow \text{msg}_m m' \) for a node identifier \( i \), a tagged message \( \text{msg}_m \) and machine configurations \( m \) and \( m' \). The parameter \( i \) is taken to be the identifier of the node on which the transition is taking place. For local computations, we have rules analogous to those of the CESH\_1 machine, so we omit them and show only those for remote computations. The rules use the function \( i \vdash h \Rightarrow x \) for allocating a pointer to \( x \) in a heap \( h \) and then constructing a remote pointer tagged with node identifier \( i \) from it. When starting a remote computation, the machine allocates a continuation in the heap and sends a message containing the code and continuation pointer to the remote node in question. Afterwards the current thread is stopped.

On the receiving end of such a communication, a new thread is started, placing the continuation pointer at the bottom of the stack for the later return to the caller node. To run the apply instruction when the function closure is remote, i.e. its location is not equal to the current node, the machine sends a message containing the closure pointer, argument value, and continuation, like in the DCESH\_1 machine. On the other end of such a communication, the machine dereferences the pointer and enters the closure with the received value. The bottom remote continuation pointer is set to the received continuation pointer. After either a remote invocation or a remote application, the machine can return if it has produced a value on the stack and has a remote continuation at the bottom of the stack. To do this, a message containing the continuation pointer and the return value is sent to the location of the continuation pointer. When receiving a return message, the continuation pointer is dereferenced and entered with the received value.

A network of abstract machines is obtained by instantiating the \texttt{Network} module with the \texttt{---Machine} relation. From here on \texttt{SyncNetwork} and \texttt{AsyncNetwork} and their transition relations refer to the instantiated versions.

Unsurprisingly, if all nodes in a synchronous network except one are inactive, then the next step is deterministic.\(^8\) Another key ancillary property of DCESH networks is that synchronous or asynchronous networks for single threaded computations behave essentially the same,\(^9\) which means it is enough to deal with the simpler synchronous networks.

DCESH network nodes can terminate with a value \( v \) (nodes \( \downarrow^\text{Sync} \) \( v \)), terminate (nodes \( \downarrow^\text{Sync} \)), or diverge (nodes \( \dagger^\text{Sync} \)). A network terminates with a value \( v \) if it can step to a network where only one node is active, and that node has reached the END instruction with the value \( v \) on top of its stack. The other definitions are analogous to those of the CES(H) machine.

3.5.1 Correctness

To prove the correctness of the machine, we will now establish a bisimulation between the CESH and the DCESH machines. To simplify this development, we extend the CESH machine with a dummy rule for the \texttt{REMOTE} \( c i \) instruction so that both machines run the same bytecode. This rule is almost a no-op, but since we are assuming that the code we run remotely is closed, the environment is emptied, and since the compiled code \( c \) will end in a \texttt{RET} instruction a return continuation is pushed on the stack.

\[
\begin{array}{c}
\text{REMOTE } c' i ; c, e, s, h \rightarrow (c', [], \text{cont } (c, e) :: s, h)
\end{array}
\]

The relation that we are about to define is, as before, \textit{almost} equality. But since values may be pointers to closures, it must be parameterised by heaps. A technical problem is that both machines use pointers, and the DCESH machine also uses remote pointers and has two heaps for each node: so the relation must be parameterised by all the heaps in the system. The extra parameter is a synonym for an indexed family of the closure and continuation heaps, \texttt{Heaps} \( = \text{Node} \rightarrow \text{DCESH.ClosHeap} \times \text{DCESH.ContHeap} \). The complexity of this relation justifies our use of mechanised reasoning.

The correctness proof itself is not routine. Simply following the recipe that we used before does not work. In the old proof, there can be no circularity, since that bisimulation was constructed inductively on the structure of the CES configuration. But now both systems, CESH and DCESH, have heaps where there is a potential

\(^8\) DCESH.Properties.determinism-Sync

\(^9\) DCESH.Properties.---Async---\rightarrow Sync
for circular references (e.g. a closure, residing in a heap, whose environment contains a pointer to itself), preventing a direct proof via structural induction. This is perhaps the most mathematically (and formally) challenging point of the paper. The solution lies in using the technique of step-indexed relations, adapted to the context of bisimulation relations [2]. The additional rank parameter records how many times pointers are allowed to be dereferenced.

The rank is used in defining the relation for closure pointers \( R_{ptr} \). If the rank is zero, the relation is trivially fulfilled. If the rank is non-zero then three conditions must hold. First, the CESH continuation heap of the location of the pointer, which is related to the CESH stack element. Finally, a CESH configuration and a DCESH stack may end in a pointer representing a remote machine that is related to the configuration.

In particular, if \( R_{sync} \) is a bisimulation relation \(^{11} \), then the inverse \( R_{sync}^{-1} \) of \( R_{sync} \) preserves the composition of \( R_{sync} \) with itself. This property is used to show that \( R_{sync} \) is preserved. It is quite immediate that the inverse of \( R_{sync} \) is a presimulation \(^{15} \) which leads to the main result that \( R_{sync} \) is a bisimulation.

In particular, if \( R_{sync} \) is a bisimulation relation, then \( R_{sync} \) holds for configurations in \( R_{sync} \). These final results complete the picture for the DCESH machine. We have established that we get the same fundamental result regardless of whether we choose to run a fragment of code using the CES, the CESH, or the DCESH machine.

4. Fault-tolerance via transactions

In this section we present a generic transaction-based method for handling failure which is suitable for the DCESH. Node state is “backed up” (commit) at certain points in the execution, and if an exceptional condition arises, the backup is restored (rollback).

This development is independent of the underlying transition relation, but the proofs rely on sequentiality. We assume that we have two arbitrary types \( Machine \) and \( Msg \), as well as a transition relation over them:

\[
\xrightarrow{\text{normal-step}} : Machine \rightarrow Tagged \text{Msg} \rightarrow Machine \rightarrow *
\]

Since we have no knowledge of exceptional states in \( Machine \), since it is a parameter, we define another relation,

\[
\xrightarrow{\text{crash}} \omega \rightarrow \omega \text{ as a thin layer on top of } \xrightarrow{\text{normal-step}} \omega \text{. The new definition is shown in Fig. 8 and adds the exceptional state nothing by extending the set of states of the relation to Maybe Machine. The fallible machine can make a normal-step transition from and to just ordinary Machine states, or it can crash which leaves it in the exceptional state. This means that we tolerate fail-stop faults as opposed to e.g. the more general Byzantine failures.}

The additional assumptions for sequentiality are that we have a decidable predicate, \( active : Machine \rightarrow * \) on machines, and the following functions:

\[
\begin{align*}
\text{inactivate-active} & : \forall \{ m, m’ \text{msg} \} \rightarrow \\
\text{inactive} & : Machine \rightarrow m’ \rightarrow \neg (active m) \times active m’ \\
\text{active-silent-active} & : \forall \{ m \rightarrow m’ \} \rightarrow
\end{align*}
\]

\(^{13} \) DCESH.Simulation-CESH.bisimulation-sync

\(^{14} \) DCESH.Simulation-CESH.termination-agrees-sync, DCESH.Simulation-CESH.divergence-agrees-sync
Figure 9. The definition of the transition relation for a crashing machine with backup.

\[
\begin{align*}
\text{silent-step} & : \forall \{m, m'\} \rightarrow \\
(just \ m \rightarrow \text{just} \ m') & \rightarrow ((m, n) \rightarrow \text{Backup} \rightarrow (m', n)) \\
\text{receive-step} & : \forall \{m, m' \text{ msg}\} \rightarrow \\
(just \ m \rightarrow \text{receive} \text{ msg} \rightarrow \text{just} \ m') & \rightarrow ((m, n) \rightarrow \text{receive msg} \rightarrow \text{Backup} \rightarrow (m', m')) \\
\text{send-step} & : \forall \{m, m' \text{ msg}\} \rightarrow \\
(just \ m \rightarrow \text{send} \text{ msg} \rightarrow \text{just} \ m') & \rightarrow ((m, n) \rightarrow \text{send msg} \rightarrow \text{Backup} \rightarrow (m', m')) \\
\text{recover} & : \forall \{\text{ m} \rightarrow \nothing\} \rightarrow \\
(just \ m \rightarrow \text{Crash} \rightarrow \nothing) & \rightarrow ((m, n) \rightarrow \text{Backup} \rightarrow (n, n))
\end{align*}
\]

These functions express the property that if a machine is invoked, i.e. it receives a message, then it must go from an active to an inactive state. If the machine then takes a silent step, it must remain active, and when it sends a message it must go back to being inactive. This gives us sequentiality; a machine cannot fork new threads, and cannot be invoked simultaneously in parallel.

As the focus here is on obvious correctness and simplicity, we abstract from the method of detecting faults in a separate node, and assume that it can be done (e.g. a heartbeat network). Similarly, we assume that we have a means of creating and restoring a backup of a node in the system; how this is done depends largely on the underlying system. We define a machine with a backup as Backup = \(\text{Machine} \times \text{Machine}\), where the second Machine denotes the backup. Using this definition, we define a backup strategy, given in Fig. 9. This strategy makes a backup just after sending and receiving messages. In the case of the underlying machine crashing, it restores the backup. Note that this is only one of many possible backup strategies. This one is particularly nice from a correctness point-of-view, because it makes a backup after every observable event, although it may not be the most performant.

We define binary relations for making transition with some tagged message, as follows:

- \(\text{m} \rightarrow \text{Machine} \rightarrow \text{active} \times \text{active} \text{m}'\)
- \(\text{active-send-inactive} : \forall \{\text{m} \text{ m' msg}\} \rightarrow \)
- \(\text{m} \rightarrow \text{send msg} \rightarrow \text{Machine} \rightarrow \text{active} \times \text{active} \text{m}'\)

Using these relations we can define the observable trace of run of a Machine (Backup), i.e. an element of the reflexive transitive closure of the above relations. First we define \(\text{IO}\), the subset of tagged messages that we can observe, namely \(\text{send}\) and \(\text{receive}\):

\[
\text{data} \ IO \ (A : \text{*}) \rightarrow \text{*} \text{ where} \\
\text{send} \ \text{receive} \ : \ A \rightarrow \ IO \text{ A}
\]

The following function now gives us the observable trace, given an element of \(\text{IO msg}^*\) (which is defined using list-like notation) by ignoring any silent steps.

\[
[()]_M = [] \\
([\text{\tau}, \_])_M = [\text{steps}]_M \\
([\text{send msg}, \_])_M = [\text{send msg纵 steps}]_M \\
([\text{receive msg}, \_])_M = [\text{receive msg纵 steps}]_M
\]

\([\cdot]_M\) is defined analogously. Given this definition, we can trivially prove that if we have a run \(\text{IO msg}_1 \rightarrow \text{IO msg}_2\) then there exists a run of the Backup machine that starts and ends in the same state and has the same observational behaviour.\(^{16}\) This is proved by constructing a crash-free Backup run given the Machine run. Obviously, the interesting question is whether we can take any crashing run and get a corresponding Machine run. The key to proving the result that we want, which more formally is that, given \(\forall \{\text{ bs : (b1,b1) Backup msg (m2,b2)}\}\), there is a run \([\text{ms : b1 Backup msg (m2,b2)}]\) with the same observational behaviour as \(\text{bs}\), is the following lemma\(^{17}\):

\[
\text{fast-forward-to-crash} : \forall \{\text{m1 m2 b1 b2 n}\} \rightarrow \\
(\text{s : (m1,b1) Backup msg (m2,b2)}) \rightarrow \\
\text{thread-crashes s} \rightarrow \text{length s} \leq n \rightarrow \\
\exists \lambda (s' : ((\text{b1,b1)) Backup \rightarrow (\text{m2,b2)})) \rightarrow \\
(\text{\neg thread-crashes s'}) \times ([s \rightarrow (s') \times (\text{length s'} \leq n))
\]

Here \(\text{thread-crashes}\) is a decidable property on backup runs, that ensures that, if \(\text{m1}\) is active, then it crashes and does a recovery step at some point before it performs an observable action. The proof of \(\text{fast-forward-to-crash}\) is done by induction on the natural number \(n\).

The above result can be enhanced further by observing that if the probability of a machine crash is not 1 then the probability of the machine eventually having a successful execution is 1. This means that the probability for the number \(n\) above to exist is also 1. This argument has not been formalised in Agda.

5. Related work

There is a multitude of programming languages and libraries for distributed or client-server computing. We focus mostly on those with a functional flavour. For surveys, see [28, 41]. Broadly speaking, we can divide them into those that use some form of explicit message passing, and those that have more implicit mechanisms for distribution and communication.

Explicit A prime example of a language for distributed computing that uses explicit message passing is Erlang [3]. Erlang is a very successful language used prominently in the telecommunications industry. Conceptually similar solutions include MPI [21] and Cloud Haskell [14]. The theoretically advanced projects Nomadic Pict [43] and the distributed join calculus [16] both support a notion of mobility for distributed agents, which enables more expressivity for the distribution of a program than the fairly static networks (with only ubiquitous functions being mobile) that our work uses. Session types have been used to extend a variety of languages, including functional languages, with better communication primitives [42]. Work on session types has been an inspiration also for us: the way that we compile a single program to multiple nodes can be likened to the projection operator in multiparty session types [24]. But in general, explicit languages are well-proven, but far away in the language design-space from the seamless distributed computing that we envision could be done using native RPC, because they place the significant burden of explicit communication on the programmer.

Implicit Our work generalises remote procedure call (RPC) [7] with full support for higher-order functions. In loc. cit. it is argued that emulating a shared address space is infeasible since it requires each pointer to also contain location information, and that it is
questionable whether acceptable efficiency can be achieved. These arguments certainly apply to our work, where we do just this. With the goal of expressivity in mind, however, we believe that we should enable the programmer to write the potentially inefficient programs that (internally) use remote pointers, because not all programs are performance critical. Furthermore, using a tagged pointer representation [30] for closure pointers means that we can tag pointers that are remote, and pay a very low, if any, performance penalty for local pointers.

Remote Evaluation (REV) [39] is another generalisation of RPC, siding with us on enabling the use of higher-order functions across node boundaries. The main differences between REV and our work is that REV relies on sending code, whereas we can do both, and that it has a more elaborate distribution mechanism.

The well-researched project Eden [29] and the associated abstract machine DREAM [8], which builds on Haskell, is a semi-implicit language. Eden allows expressing distributed algorithms at a high level of abstraction, and is mostly implicit about communication, but explicit about process creation. Eden is specified operationally using a two-level semantics similar to ours, but in the context of the call-by-need evaluation strategy. Kanor [23] is a project that similarly aims to simplify the development of distributed programs by providing a declarative language for specifying communication patterns inside an imperative host language.

Hop [38], Links [11], and ML5 [32] are examples of so called tierless languages that allow writing (for instance) the client and server code of web applications in unified languages with more or less seamless interoperability between them. The Links language shares our goal of unifying distributed programs into a single language with seamless interoperability between the nodes, but its focus is on web programming with client, server, and database. For that purpose it includes sub-languages for elegantly constructing and manipulating XML documents and for doing database queries. The behaviour of Links is specified using the client/server calculus [10], which is an operational semantics similar in purpose to our abstract machines, but, as its name suggests, limited to two nodes. The two nodes of the system are constructed with web-programming in mind and are not equal peers. The server is stateless to be able to handle a large number of clients and unexpected disconnections. The semantics operates on a first-order language and uses explicit substitutions. The client/server calculus also inspired work on the LSAM [33], that lifts the two-node limitation. A similar, but earlier, machine is that of dML [35]. The dML and LSAM machines are conceptually close to our machine, but described on a level that is not as readily implementable as our work, using explicit substitutions and synchronous message passing.

On the language side, our work draws inspiration from abstract machines for game semantics [17, 18] where an exotic compilation technique based on game semantics is used to implement a language like ours, but for call-by-name and without algebraic data types. Recent work shows a formalisation similar to ours, but based on the Krivine machine and the significantly less popular call-by-name evaluation strategy [19].

6. Conclusion

The main conceptual contribution of this work is a new abstract machine, or abstract machine network rather, called the DCESH. Its main feature is that function calls behave, from the point of view of the programmer, in the same way whether they are local or remote. Moreover, on a single node the behaviour of DCESH is that of a conventional SECD-like abstract machine. All correctness proofs have been formalised using Agda.

On this theoretical foundation it is natural to build a compiler for a (very conventional) CBV language where terms are location-aware. Compared to most of the literature on the topic, the thrust of this work is not a general-purpose functional language for location-aware computing but rather a location-aware compilation technique for general-purpose functional languages. The performance of the compiler, as suggested by the benchmarks, improves dramatically on previous work in terms of single-node behaviour, and the distribution overheads are not onerous.

RPC as a paradigm has been criticised for several reasons: its execution model does not match the native call model, there is no good way of dealing with failure, and it is inherently inefficient [40]. By taking an abstract machine model in which RPCs behave exactly the same as local calls, by showing how a generic transaction mechanism can handle failure, and by implementing a reasonably performant compiler we address all these problems head-on. We believe that we provide enough evidence for general native RPCs to be reconsidered in a more positive light.

6.1 Further work

The DCESH has the internal machinery required for parallel execution, but we restrict ourselves to sequential execution. In moving towards parallelism there are several language design (how to add parallelism?) and theoretical (is compilation still correct?) challenges. The language design aspects are too broad to discuss here, beyond emphasising that the thread-based mechanism of DCESH is indeed quite flexible. The only ingredient lacking is a synchronisation primitive, but that is not a serious difficulty. The theoretical challenges are mainly stemming out of the failure of the equivalence of synchronous and asynchronous networks in the presence of multiple pending messages. Whether we choose to stick to a synchronous network model, which can however be rather unrealistic, or we try to work directly in the more challenging environment of asynchronous networks, remains to be seen.

The current implementation does not need distributed garbage collection. The values which are the result of CBV evaluation are always sent, along with any required closures, to the node where the function using them as arguments is executed. With this approach local garbage collection suffices. Note that this is similar to the approach that Links takes. If a large data structure needs to be held on a particular node the programmer needs to be aware of this requirement and indirect the access to it using functions. However, if we wanted to automate this process as well, and prevent some data from migrating when it’s too large, the current approach to garbage collection could not cope, and distributed garbage collection would be required. Mutable references or lazy evaluation would also require it. Whether this can be done efficiently is a separate topic of research [36].

Whether code or data can or should be migrated to different nodes is a question that can be answered from a safety or from an efficiency point of view. The safety angle is very well covered by type systems such as ML5’s, which prevent the unwanted exportation of local resources. The efficiency point of view can also be dealt with in a type-theoretic way, as witnessed by recent work in resource-sensitive type systems [9, 20]. The flexibility of the DCESH in terms of localising or remoting the calls (statically or even dynamically) together with a resource oriented system can pave the way towards a highly-convenient automatic orchestration system in which a program is automatically distributed among several nodes to achieve certain performance objectives.

Finally, an Agda formalisation is given only for the abstract machine and its properties, which are the new theoretical contributions of the paper. However, a full formalisation of the compiler stack remains a long-term ambition.

[19] DCESH Properties: \(\xrightarrow{\text{Async}}\xrightarrow{\text{to}}\xrightarrow{\text{Sync}}\)
References


