Programming and Proving with Distributed Protocols

ILYA SERGEY, University College London
JAMES R. WILCOX, University of Washington
ZACHARY TATLOCK, University of Washington

Distributed systems play a crucial role in modern infrastructure, but are notoriously difficult to implement correctly. This difficulty arises from two main challenges: (a) correctly implementing core system components (e.g., two-phase commit), so all their internal invariants hold, and (b) correctly composing standalone system components into functioning trustworthy applications (e.g., persistent storage built on top of a two-phase commit instance). Recent work has developed several approaches for addressing (a) by means of mechanically verifying implementations of core distributed components, but no methodology exists to address (b) by composing such verified components into larger verified applications. As a result, expensive verification efforts for key system components are not easily reusable, which hinders further verification efforts.

In this paper, we present Disel, the first framework for implementation and compositional verification of distributed systems and their clients, all within the mechanized, foundational context of the Coq proof assistant. In Disel, users implement distributed systems using a domain specific language shallowly embedded in Coq and providing both high-level programming constructs as well as low-level communication primitives. Components of composite systems are specified in Disel as protocols, which capture system-specific logic and disentangle system definitions from implementation details. By virtue of Disel’s dependent type system, well-typed implementations always satisfy their protocols’ invariants and never go wrong, allowing users to verify system implementations interactively using Disel’s Hoare-style program logic, which extends state-of-the-art techniques for concurrency verification to the distributed setting. By virtue of the substitution principle and frame rule provided by Disel’s logic, system components can be composed leading to modular, reusable verified distributed systems.

We describe Disel, illustrate its use with a series of examples, outline its logic and metatheory, and report on our experience using it as a framework for implementing, specifying, and verifying distributed systems.

1 INTRODUCTION

Real-world software systems, including distributed systems, are rarely built as standalone, monolithic pieces of code. Rather, they are composed of multiple independent modules, which are connected either by the linker or through communication channels. Such a compositional approach enables clean separation of concerns and a modular development process: in order to use one component within a larger system, one only needs to know what it does without requiring details on how it works. Unfortunately, the benefits of modular software development are not yet fully realized in the context of verified distributed systems.

Recent work has produced several impressive formal proofs of correctness for implementations of core distributed system components, ranging from consensus protocols to causally consistent key-value stores [16, 31, 41, 63]. These artifacts, while formally verified, are not immediately reusable in the context of larger verified applications. For example, to compose a linearizable database with a causally consistent cache [2], one would need a framework general enough to express both specifications and reason about their interaction, possibly...
in the presence of application-specific constraints. Furthermore, existing verified systems entangle implementation details with abstract protocol definitions, preventing independent evolution and requiring extensive refactoring when changes are made [63].

Finally, like all software, real-world distributed systems exist in an open world, and should be usable in multiple contexts by various clients, each of which may make different assumptions.

1.1 Towards Modular Distributed System Verification

Recent advances in the area of formal machine-assisted program verification demonstrated that composition, obtained by means of expressive specifications and rich semantics, is the key to producing scalable, robust and reusable software artifacts in correctness-critical domains, such as compilers [26, 50], operating systems [14, 24] and concurrent libraries [15, 49]. Following this trend, we identify the following challenges in designing a verification tool to support compositional proofs of distributed systems.

1) Protocol-program modularity. One should be able to define an abstract model of a distributed protocol (typically represented by a form of a state-transition system) without tying it to a specific implementation. Any purported implementation should then be proven to follow the protocol’s abstract model. This separation of concerns supports reuse of existing techniques for reasoning about the high-level behavior of a system, while allowing for optimized implementations, without redefining the high-level interaction protocol.

2) Modular program verification. Once proven to implement an abstract protocol, a program should be given a sufficiently expressive declarative specification, so that clients of the code never need to examine the implementation itself. Furthermore, it should be possible to specify and verify programs made up of parts belonging to different protocols (horizontal compositionality). This enables decomposing a distributed application into independently specified and proved parts, making verification scale to large codebases.

3) Modular proofs about distributed protocols. A single protocol may be useful to multiple different client applications, each of which may exercise the protocol in different ways. For instance, a “core” consensus protocol implementation can be employed both for leader election as well as for a replicated data storage. In this case, the invariants of the core protocol should be proved once and for all and then reused to establish properties of composite protocols. These composite protocols often require elaborating the core invariants with client-specific assumptions, but it would be unacceptable to re-verify all existing code under new assumptions for each different use of the core protocol. Instead, clients should be able to prove their elaborated invariants themselves by reasoning about the core protocol after the fact. This also ensures any existing program that follows the protocol is guaranteed to also satisfy the client’s new invariant. This decomposition between core protocols and elaborated client invariants reduces and parallelizes the proof engineering effort: the core system implementor verifies basic properties of the protocol and correctness of the implementation, while the system’s client proves the validity of their context-specific invariants.

This paper presents Disel, a mechanized framework for verification and implementation of distributed systems that aims to address these challenges.
1.2 What is Disel?

Disel is a verification framework incorporating ideas from dependent type theory, interactive theorem proving, separation-style program logics for concurrency, resource reasoning, and distributed protocol design.

From the perspective of a distributed protocol designer, Disel is a domain-specific language for defining a protocol $P$ in terms of its state-space invariants and atomic primitives (e.g., send and receive). These primitives implement specific transitions which synchronize message-passing with changes to the local state of a node. Described this way, the protocols are immediately amenable to machine-assisted verification of their safety and temporal properties [46, 62], and Disel facilitates these proofs by providing a number of higher-order lemmas and libraries of auxiliary facts.

From the point of view of a system implementor, Disel is a higher-order programming language, featuring a complete toolset of programming abstractions, such as first-class functions, algebraic datatypes, and pattern matching, as well as low-level primitives for message-passing distributed communication. Disel’s dependent type system makes programs protocol-aware and ensures that well-typed programs don’t go wrong; that is, if a program $c$ type-checks in the context of one or many protocols $P_1, \ldots, P_n$ (i.e., informally, $P_1, \ldots, P_n \vdash c$), then it correctly exercises and combines transitions of $P_1, \ldots, P_n$.

Finally, for a human verifier, Disel is an expressive higher-order separation-style program logic\(^1\) that allows programs to be assigned declarative Hoare-style specifications, which can be subsequently verified in an interactive proof mode. Specifically, one can check that, in the context of protocols $P_1, \ldots, P_n$, a program $c$ satisfies pre/postconditions $P$ and $Q$, where $P$ constrains the pre-state $s$ of the system, and $Q$ constrains the result $\text{res}$ and the post-state $s'$. The established pre-/postconditions can be then used for verifying larger client programs that use $c$ as a subroutine. Disel takes a partial correctness interpretation of Hoare-style specifications, thus focusing on verification of safety properties and leaving reasoning about liveness properties for future work.

We implemented Disel on top of the Coq proof assistant [6], making use of Coq’s dependent types and higher-order programming features. In the tradition of Hoare Type Theory (HTT) by Nanevski et al. [38–40] and its recent versions for concurrency [32, 37], we give the semantics to effectful primitives, such as send/receive, with respect to a specific abstract protocol (or protocols). Thus, we address challenge (1) by ensuring that any well-typed program is correct (i.e., respects its protocols) by construction, independently of which and how many of the imposed protocols’ transitions are taken and of any imperative state the program might use. This type-based verification method for distributed systems, which was motivated by a recent vision paper by Wilcox et al. [61], is different from more traditional techniques for establishing refinement [1, 16] between an actual implementation (the code) and a specification (an abstract protocol) via a simulation argument [35]. In comparison with the refinement-based techniques, the type-based verification method makes it easy to account for horizontal composition of protocols (necessary, e.g., for reasoning about remote procedure calls, as we will show in Section 2) and accommodate advanced programming features, such as higher-order functions.

As a program logic, Disel draws on ideas from separation-style logics for shared-memory concurrency [37, 55], allowing one to instrument programs with pre/postconditions and providing a form of the frame rule [47] with respect to protocols. For example, assuming that the state-spaces of $P_1$ and $P_2$ are disjoint, $P_1 \vdash c_1$ and $P_2 \vdash c_2$ together with the frame

\(^1\)The framework name stands for Distributed Separation Logic.
rule imply $P_1, P_2 \vdash C[c_1, c_2]$ for any well-formed program context $C$. This ensures that the composite program $C[c_1, c_2]$ can “span” multiple protocols, thus addressing challenge (2). The assumption of protocol state-spaces being disjoint might seem overly restrictive, but, in fact, it reflects the existing programming practices. For instance, the local state of a node responsible for tracking access permissions is typically different from the state used to store persistent data.

Disel further alleviates the issue of disjoint state and also addresses challenge (3) with two novel logical mechanisms, described in detail in Section 3. The first one supports the possibility of elaborating protocol invariants via an inference rule, WITHINV, allowing one to strengthen the assumptions about a system’s state, resulting in the strengthened guarantees, as long as these assumptions form an inductive invariant. Second, Disel supports “coupling” protocols via inter-protocol behavioral dependencies, which allow one protocol restricted logical access to state in another protocol, all while preserving the benefits of disjointness, including the frame rule. Dependencies are specified with the novel logical mechanism of inter-protocol send-hooks, allowing one to restrict interaction between a core protocol and its clients by placing additional preconditions on certain message sends. For example, a send-hook could disallow certain transitions of the client protocol unless a particular condition holds for the local state associated with the core protocol. These additional preconditions do not require re-verifying any core components.

To summarize, this paper makes the following contributions:

- Disel, a domain-specific language and the first separation-style program logic for the implementation and compositional verification of message-passing distributed applications for full functional correctness, supporting effectful higher-order functional programming style, as well as custom distributed protocols and their combinations;
- Two conceptually novel logical mechanisms allowing reuse of Hoare-style and inductive invariant proofs while reasoning about distributed protocols: (a) the WITHINV rule enabling elaboration of the protocol invariant in program specifications, and (b) send-hooks, providing a way to modularly verify programs operating in a restricted product of multiple protocols;
- A proof-of-concept implementation of Disel as a foundational (i.e., proven sound from first principles [3]) verification tool, built on top of Coq, as well as mechanized soundness proofs of Disel’s logical rules with respect to a denotational semantics of message-passing distributed programs;
- An extraction mechanism into OCaml and a trusted shim implementation, allowing one to run programs written in Disel on multiple physical nodes;
- A series of case studies implemented and verified in Disel (including the Two-Phase Commit protocol [60] and its client application), as well as a report on our experience of using Disel and a discussion on the executable code.

2 OVERVIEW

In this section we illustrate the Disel methodology for specifying, implementing, and verifying distributed systems by developing a simple distributed calculator. Disel systems are composed of concurrently running nodes communicating asynchronously by exchanging messages, which, as in real networks, can be reordered and dropped.

In the calculator system, each node $n$ is either a client (written $n \in \mathcal{C}$) or a server ($n \in \mathcal{S}$), and the system is parameterized over some expensive partial function $f$ with domain $\text{dom}(f)$.
Send-transitions

<table>
<thead>
<tr>
<th>Transition</th>
<th>Requires</th>
<th>Ensures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\tau_s)</td>
<td>((m, to))</td>
<td>((n \rightarrow (to, args)) \uplus rs)</td>
</tr>
<tr>
<td>sreq</td>
<td>(n \in C \land to \in S \land n \rightarrow rs \land m = (\text{Req}, args) \land args \in \text{dom}(f))</td>
<td>(n \rightarrow (to, args) \uplus rs)</td>
</tr>
<tr>
<td>aresp</td>
<td>(n \in S \land f(args) = v \land n \rightarrow (to, args) \uplus rs \land m = (\text{Resp}, v, args))</td>
<td>(n \rightarrow rs)</td>
</tr>
</tbody>
</table>

Receive-transitions

<table>
<thead>
<tr>
<th>Transition</th>
<th>Requires</th>
<th>Ensures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\tau_r)</td>
<td>((m, from))</td>
<td>((n \rightarrow (from, args)) \uplus rs)</td>
</tr>
<tr>
<td>treq</td>
<td>(n \in S \land n \rightarrow rs \land m = (\text{Req}, args))</td>
<td>(n \rightarrow (from, args) \uplus rs)</td>
</tr>
<tr>
<td>trresp</td>
<td>(n \in C \land n \rightarrow (from, args) \uplus rs \land m = (\text{Resp}, ans, args))</td>
<td>(n \rightarrow rs)</td>
</tr>
</tbody>
</table>

Fig. 2. Send- and receive-transitions of the distributed calculator protocol with respect to a node \(n\).

Given arguments \(args \in \text{dom}(f)\), a client can send a request containing \(args\) to a server, which will reply with \(f(args)\). Fig. 1 depicts an example execution for the calculator system with one server \(S\) and two clients, \(C_1\) and \(C_2\). Note that requests and responses may not be received in the order they are sent due to network reordering, and the server may service requests in any order (e.g., due to implementation details such as differing priorities among requests). However, the system should satisfy weak causality constraints, e.g., a client \(C\) should only receive a response \(f(args)\) if \(C\) had previously made a request for \(args\). In the remainder of this section we show how Disel enables developers to specify the abstract calculator protocol, implement several versions of server and client nodes that follow the protocol, and prove key invariants for reasoning about the system.

2.1 Defining a Calculator Protocol

A protocol in Disel provides a high-level specification of the interface between distributed system components. As with traditional program specifications, Disel protocols serve to separate concerns: implementations can refine details not specified by the protocol (e.g., the order in which to respond to client requests), invariants of the protocol can be proven separately (e.g., showing that calculator responses contain correct answers), and interactions between components within a larger system can be reasoned about in terms of their protocols rather than their implementations. Following the tradition established by Lamport [27], Disel protocols are defined as state-transition systems.

Fig. 2 depicts the state-transition system for the calculator example with two send-transitions and two receive-transition. Each transition is named in the first column: \(s\)-transitions are for sending and \(r\)-ones for receiving. Their pre- and postconditions (in the form of requires/ensures pairs) are given as assertions in the second and third columns respectively. These assertions are phrased in terms of the message being sent/received, recipient/sender (\(to/from\)), and the protocol-specific state of a node \(n\). For the calculator, the state for node \(n\) is a multiset of outstanding requests \(rs\), denoted as \(n \rightarrow rs\).
Protocol transitions synchronize the exchange of messages with changes in a node’s state. Preconditions in send-transitions specify requirements that must be satisfied by the local state of node $n$ for it to send message $m$ to recipient $to$ and postconditions specify how $n$’s state must be updated afterward. For example, the $sreq$ transition can be taken by a client node $n \in C$ to send a request message $(\text{Req}, \text{args})$ to server $to$ where $\text{args} \in \text{dom}(f)$ and, after sending, $n$ has added $(\text{to}, \text{args})$ to its state. Preconditions in receive-transitions specify requirements that must be satisfied by the local state of node $n$ for it to receive message $m$ from sender $from$ and postconditions specify how $n$’s state must be updated. For example, the $rreq$ transition can be taken by a server node $n$ to receive a request message $(\text{Req}, \text{args})$ from node $from$ where, after receiving, $n$ has added $(\text{from}, \text{args})$ to its state.

Notice that preconditions in send-transition can be arbitrary predicates, while the precondition of receive-transitions must be decidable (which we emphasize by using boolean conjunction $\&$ instead of propositional $\land$). This is because a program’s decision to send a message is active and corresponds to calling the low-level send primitive (described later in this section); the system implementer must prove such preconditions to use the transition. In contrast, receiving messages is passive and corresponds to using the low-level recv primitive (also described later in this section) that will react to any valid message. A message $m$ sent to node $n$ should trigger the corresponding receive transition only if $n$’s state along with the message satisfies the transition’s precondition. To choose such a transition unambiguously, we require that each message’s tag (e.g., Req and Resp) uniquely identifies a receive-transition that should be run. Combined with the decidability of receive-transition preconditions, this allows Disel systems to automatically decide whether a transition can be executed.

As defined, the calculator protocol prohibits several unwelcome behaviors. For instance, a server cannot send a response without a client first requesting it, since (a) servers only send messages via the $sresp$ transition, (b) $sresp$ requires $(\text{to}, \text{args})$ to be in the multiset of outstanding requests at the server, and (c) $(\text{to}, \text{args})$ can only be added to the set of outstanding requests once it has been received from a client. Also note that the precondition of $sreq$ requires that when a client sends a request to a server to compute $f(\text{args})$, $\text{args} \in \text{dom}(f)$. Similarly, the precondition of $sresp$ requires that when a server responds to a client request for $\text{args}$, it may only send the correct result $f(\text{args})$. In this case, the initial arguments $\text{args}$ are included into the response in order make it possible for the client to distinguish between responses to multiple outstanding requests.

The protocol also leaves several details up to the implementation. For example, the $sresp$ transition allows a server to respond to any outstanding request, not necessarily the least recently received. This flexibility allows for diverse implementation strategies and enables the implementation $I$ of a component to evolve without requiring updates to other components which only assume that $I$ satisfies its protocol.

This state-space and transitions define the calculator protocol $C$. Protocols are basic specification units in Disel, and, as we will soon see, a single program can “span” multiple protocols. Thus, we will annotate each protocol instance with a unique label $\ell_i$ (e.g., $C_{\ell_1}, C_{\ell_2}$).

### 2.2 From Protocols to Programs

The transitions in Fig. 2 define functions mapping a state, message, and node id to a new state. We can use these functions as basic elements in building implementations of distributed system components, but first we need to “tie” them to realistic low-level message sending/receiving primitives. We can then combine these basic elements, via high-level programming constructs, into executable programs.
In DISEL a programmer can define a new programming primitive based on a send- or receive transition using a library of transition wrappers, that decorate send/receive primitives with transitions of protocols at hand. The generic send(τs, ℓ) wrapper from this library takes a send-transition τs of a protocol identified by a label ℓ and yields a program that sends a message. For instance, from the description in Fig. 2 and DISEL's logic (discussed in Section 3), we can assign the following Hoare type (specification) to a “wrapped” transition sresp run by server n in the context of the protocol Cℓ:

\[ C_\ell \vdash^n \text{send}[sresp, ℓ](m, to) : \begin{cases} \{ n \in S \land n \rightarrow ((to, args) \cup rs) \land m = (\text{Resp}, f(args), args) \} & \{ n \rightarrow rs \land \text{res} = m \} \end{cases} \] (1)

The assertions in the pre/postconditions of the type (1) quantify implicitly over the entire global distributed state s (including previously sent messages), although the calculator protocol only constrains n’s local contents in s, which are referred using the “node n’s local state points-to” assertion of the form n \rightarrow \_\_. In particular, the specification ensures that the outstanding request (to, args) is removed from the local state of a node n upon sending a message. As customary in Hoare logic, all unbound variables (e.g., rs, args) are universally-quantified and their scope spans both the pre- and post-condition. The return value res, occurring freely in the postcondition of a wrapped send-transition, is the message sent. In most of the cases, we will omit the type of res for the sake of brevity.

DISEL’s type system ensures Hoare-style pre/postconditions in types are stable, i.e., invariant under possible concurrent transitions of nodes other than n. Stability often requires manual proving, but is indeed the case in the triple (1), as its pres/posts constrain only local state of the node n, which cannot be changed by other nodes. In general, Hoare triples in DISEL can refer to state of other nodes as well, as we will demonstrate in Section 4.

Using a wrapper recv for tying a receive-transition to a non-blocking receive command is slightly more subtle. In general, we cannot predict which messages from which protocols a node n may receive at any particular point during its execution. To address this, receive wrapper recv[T, L] specifies a set T of message tags and a set L of protocol labels; and only accept messages whose tag is in T for a protocol whose label is in L.\(^2\) The resulting primitive provides non-blocking receive: if there are no messages matching the criteria, it returns None and acts as an idle transition. Otherwise, it returns Some (from, m) for a matching incoming message m from sender from, chosen non-deterministically from those available. For example, we can assign the following Hoare type to a wrapper, associated with the tag Req of Cℓ:

\[ C_\ell \vdash^n \text{recv}[\{\text{Req}\}, \{ℓ\}] : \begin{cases} \{ n \in S \land n \rightarrow rs \} & \begin{cases} \text{if res = Some } (\text{from, (Req, args)}) \text{ then } n \rightarrow ((\text{from, args}) \cup rs) \land \langle\text{from, n, •, (Req, args)}\rangle \in MS_\ell \end{cases} \\
\text{else } n \rightarrow rs \end{cases} \] (2)

The postcondition of the type (2) demonstrates an important feature of DISEL’s Hoare-style specs: in the case of a received message, it existentially binds its components (i.e., from, args) in then-branch, and also identifies the message \langle from, n, •, (Req, args)\rangle in the message soup MS_\ell (which models both the current state and history of the network) of the post-state s’ wrt. the protocol C_\ell. Messages in DISEL’s model (described in detail in Section 3.1) are never “thrown away”; instead they are added to the soup, where they remain active (●) until received, at which points they become consumed (○).\(^3\)

---

\(^2\)Our implementation also allows “filtering” messages to be received with respect to their content.

\(^3\)This design choice with respect to message representation is common in state-of-the-art frameworks for distributed systems verification, e.g., IronFleet [16] and Ivy [44], as it simplifies reasoning about past events.
We can now employ the program (2) to write a blocking receive for request messages via
Disel’s built-in general recursion combinator \texttt{letrec} (explained in Section 3), assigning this
procedure the following specification:

\[
\begin{align*}
C_\ell \vdash^n \textbf{letrec} & \text{ receive\_req \ (\_ : unit) } \triangleq \\
& r \leftarrow \text{recv}\{\{\text{Req} \}, \{\ell\}\}; \\
& \text{if } r = \text{Some } (from, m) \\
& \text{then return } (from, m) \\
& \text{else receive\_req } () : \forall u : \text{unit. } \begin{cases} \\
& n \in S \land n \rightarrow rs \{ n \rightarrow (\langle \text{res.1, res.2} \rangle \uplus rs) \land \\
& \langle \text{res.1, n, \bullet, (Req, res.2) } \rangle \in MS_\ell \} \\
\end{cases}
\end{align*}
\]

The Hoare type of \texttt{receive\_req} describes it as a function, which takes an argument of type
unit and is safe to run in a state, satisfied by its precondition. The pre/postconditions of
\texttt{receive\_req} are derived from the type (2) by application of a typing (inference) rule for
fixpoint combinator, with an assistance of a human prover and according to the inference
rules of Disel, described in Section 3.2. Internally, \texttt{receive\_req} corresponds to an execution of
possibly several idle transitions, followed by one receive-transition. That is, when invoked, it
still follows \( C_\ell \)'s transitions: otherwise we simply could not have assigned a type to it at all!
In other words, a body of \texttt{receive\_req} is merely a combination of more primitive sub-programs
(namely, the “wrapped” non-blocking receive (2)) that are proven to be protocol-compliant.

2.3 Elaborating State-Space Invariants of a Protocol

Let us now use \texttt{receive\_req} to implement our first useful component of the system: a simple server,
which runs an infinite loop, responding to one
request message via a \texttt{receive\_req} body:

\[
\begin{align*}
1 & \textbf{letrec} \texttt{simple\_server} \ (\_ : \text{unit}) \triangleq \\
2 & (from, args) \leftarrow \texttt{receive\_req} () ; \\
3 & \textbf{let } v = f(args) \textbf{ in} \\
4 & \texttt{send}(sresp, \ell)(\langle \text{Resp, v, args, from}\rangle) ; \\
5 & \texttt{simple\_server} () \\
6 & \textbf{in} \texttt{simple\_server} ()
\end{align*}
\]

1:8 Ilya Sergey, James R. Wilcox, and Zachary Tatlock

letrec receive_batch \( (k : \text{nat}) \) :=
\[
\begin{align*}
\text{if } k &= k' + 1 \\
\text{then } \text{fargs} &\leftarrow \text{receive_req}(); \\
\text{rest} &\leftarrow \text{receive_batch} k'; \\
\text{return } \text{fargs} :: \text{rest} \\
\text{else return } []
\end{align*}
\]

letrec send_batch \( (rs : [((\text{Node, [\text{nat}]}))] \) :=
\[
\begin{align*}
\text{if } rs &= (\text{from}, \text{args}) :: rs' \\
\text{then let } v &= f(\text{args}) \text{ in} \\
\text{send}sresp, \ell((\text{Resp}, v, \text{args}), \text{from}); \\
\text{send}batch rs' \\
\text{else return } ()
\end{align*}
\]

letrec batch_server \( (bsize : \text{nat}) \) :=
\[
\begin{align*}
\text{reqs} &\leftarrow \text{receive_batch} bsize; \\
\text{send}batch \text{reqs}; \\
\text{batch_server} bsize
\end{align*}
\]

(a)

letrec memo_server \( (mmap : \text{map}) \) :=
\[
\begin{align*}
\text{let } ans &= \text{lookup mmap args in} \\
\text{if } ans \neq \bot \text{ then} \\
\text{send}\ [sresp, \ell](\text{Resp, ans, args}, \text{from}); \\
\text{memo_server mmap} \\
\text{else} \\
\text{let } ans &= f(\text{args}) \text{ in} \\
\text{send}sresp, \ell(m, (\text{Resp, ans, args})); \\
\text{let } mmap' &= \text{update mmap args ans in} \\
\text{memo_server mmap'}
\end{align*}
\]

(b)

Fig. 3. Batching (a) and memoizing (b) calculator servers defined on top of the protocol \( C'_\ell \).

allowing us to complete the verification of our looping server implementation, assigning it the following type (with the standard \text{False} postcondition due to non-termination):
\[
C'_\ell \vdash \text{simple_server} () : \{ n \in S \wedge n \Rightarrow rs \} \{ \text{False} \}
\]

Having a server loop assigned a specification (5) ensures that it faithfully follows the protocol’s transitions and does not terminate.

2.4 More Implementations for Cheap

With the elaborated protocol \( C'_\ell \), we can now develop and verify a variety of system components, reusing the previously developed libraries and enjoying the compositionality of specs, afforded by Hoare types quantifying over a distributed state and sent/received messages. It is still up to the programmer to verify those implementations in a Hoare style, but writing them does not require changing the protocol, only composing the verified subroutines.

Alternative servers. Fig. 3 presents two alternative looping server implementations. The first one processes requests in batches of a predefined size \( bsize \). This batching may cause \text{batch_server} to loop for an unbounded period, until \( bsize \) requests have been received, but this is perfectly safe. Once this is done, the batch is passed to the second subroutine, \text{send_batch}, which delivers the results. Finally, the server loop restarts. Another, more efficient server implementation \text{memo_server} uses memoization, implemented by means of store-passing style, in order to avoid repeating computations. It first checks whether the answer for a requested argument list is available in the memoization table \( mmap \), and, if so, sends it back to the client. Otherwise, it computes the answer and stores it in the local state, which is then passed to the next recursive call. Both implementations, when invoked with a suitable initial argument (batch size and an empty map, correspondingly), type-check against the same Hoare type as the simple server (5) and are verified directly from the specifications of their components in the context of \( C'_\ell \).

Implementing a calculator client. Let us now build and verify a simple client-side procedure that requests a computation and obtains the result.
It can be implemented as shown on the right. The program compute sends a request to a server serv and then runs a blocking procedure receive_res for a message with the Resp tag, implemented similarly to receive_req, and having, when invoked as a function, the following specification, stating that res is the received response:

\[
C_{\ell}^n \vdash \text{receive\_res}() : \{ n \in C \land n \rightarrow \{(\text{serv}, \text{args})\} \}
\]

Unfortunately, this type is not helpful to prove the desired spec of compute, stating that its result is equal to \(f\)(args); this dependency is not captured in (6)’s postcondition. In order to deliver a stronger postcondition of receive_res, we need to elaborate the protocol’s state-space assumption even further, proving the following invariant INV₂ inductive:

\[
\text{INV}_2(s) \equiv \forall m \in s \# MS_\ell, \ m = (n_1, n_2, -, (\text{Resp}, \text{ans}, \text{args})) \implies f(\text{args}) = \text{ans}
\]

What is left is to verify the implementation of receive_res in the context of \(C''_\ell\) \equiv \text{WithInv}(C_{\ell}, \text{INV}_2). The property INV₂ ensures that any answer carried by a Resp message is correct wrt. the corresponding arguments. Since the client has only one outstanding request at the moment it calls receive_res, it will only accept a message with an answer to that request. Thus, we can prove the following spec for the RPC compute:

\[
C_{\ell}^n \vdash \text{compute} \ (\text{args}, \text{serv}) : \{ n \in C \land n \rightarrow \emptyset \land \text{serv} \in S \land \text{args} \in \text{dom}(f) \}
\]

\[
\{ \text{res} = f(\text{args}) \land n \rightarrow \emptyset \}
\]

Server as a client. So far, we have only considered programs that operate in the context of a single protocol. However, it is common for realistic applications to participate in several systems. Disel accounts for such a possibility by providing an injection/protocol framing mechanism, inspired by the FCSL program logic [37], and allowing one to type-check a program in the context of several protocols with disjoint state-spaces. The disjointness of those does not mean the disjointness of the node sets: one node can be a part of several protocols, in which case its local state is divided among them.

As an example, let us implement yet another calculator server, this time using an \(\ell_1\)-labelled protocol run by a node \(n\), which, instead of calculating directly, delegates to a server \(n'\) in another protocol (labelled with \(\ell_2\), which we use to annotate the corresponding call to compute to emphasize the protocol it “belongs to”), in which \(n\) is a client. The code of deleg_server is almost identical to the code of simple_server and it has the following type in the context of two independent protocols, \(C''_{\ell_1}\) and \(C''_{\ell_2}\):

\[
C''_{\ell_1}, C''_{\ell_2} \vdash \text{deleg\_server} \ (n') : \{ (n \in S_{\ell_1} \land n \vdash rs) \ast (n \in S_{\ell_2} \land n' \in S_{\ell_2} \land n \vdash 0) \} \{ \text{False} \}
\]

In the pre-condition, the assertions about the nodes’ roles and local state are elaborated for specific constituent protocols, labeled with \(\ell_1\) and \(\ell_2\), correspondingly. Furthermore, we use the separating conjunction \(\ast\) in order to emphasize the disjointness of the protocol-specific local states, used to handle outstanding requests within two different protocols. As a server, \(n\) can have an arbitrary number of “outstanding responses” \(rs\) in its local state (hence \(n \vdash rs\)), but should start with an empty set of its own outstanding requests, thus \(n \vdash 0\).
Summary of the DISEL methodology. Our entire development of the calculator-aware applications (e.g., servers and clients) is outlined in Fig. 4. This is a general layout of structuring the development of applications in DISEL. In the figure, the top-down direction corresponds to elaborating the protocol invariants (so the specs of programs verified there can be directly reused further down), and the arrows denote dependencies between components.

2.5 Putting It All Together

DISEL programs can be extracted into OCaml code, linked with a trusted shim, and run. In order to do so, one needs to assign each participant node a program to run (some nodes might have no programs assigned) and provide an initial distributed configuration that instantiates the local state for each participant in each protocol and satisfies all imposed state-space invariants (e.g., (4) and (7)). The semantics of Hoare types in DISEL, defined in Section 3.3, specifies what does it mean for a program to be type-safe (i.e., correct) in a distributed setting: postconditions (even those constraining the global state) of well-typed programs are not affected by execution of programs running concurrently on other nodes, and such programs are always safe to run when their precondition is stable and satisfied.

As an illustration of one possible finalized protocol/program composition, Fig. 5 depicts the three calculator-based programs, described earlier, running concurrently by three different nodes, \(n_1\), \(n_2\), and \(n_3\), such that \(n_1\) and \(n_2\) communicate according to the protocol \(C''_{E_1}\), and \(n_2\) and \(n_3\) follow the protocol \(C''_{E_2}\). Solid arrows between nodes denote message exchange, with the time going from left to right. The initial local states for all the nodes/protocols are instantiated with empty lists of requests. Importantly, the code run by the nodes \(n_1\) and \(n_3\) has been verified separately, in simpler, smaller contexts, and only the implementation of \(n_2\)’s program \(deleg_server\) has been done in the composite context of two protocols. Our accompanying Coq development provides the complete implementation of the described programs in DISEL DSL, their extracted executable counterparts in OCaml, and mechanized proofs of all of the mentioned invariants and specifications.

3 DISTRIBUTED SEPARATION LOGIC

We next describe the formal model of the state and protocols, giving meaning to DISEL’s Hoare-style specifications in the context of multiple protocols with disjoint state-spaces and possible imposed inter-protocol dependencies.

3.1 State and Worlds

Distributed state and its components. The left part of Fig. 6 defines the components of the state, subject to manipulation by concurrently executing programs run by different nodes. Each global system state \(s\) is a finite partial mapping from protocol labels \(\ell \in \text{Lab}\) to
statelets. Each statelet represents a protocol-specific component, consisting of a “message soup” $MS$ and a per-node local state ($\text{DistLocState}$). The former represents a finite partial map from unique message identifiers to messages,$^4$ each of which carries its sender and recipient node ids, the payload $m$, which includes a tag, and a boolean indicating whether the message is already received ($\bullet$) or not yet ($\circ$). The per-node local state maps each node id into protocol-specific piece of local state, represented as a mapping from locations (isomorphic to natural numbers) to specific values. For instance, in the calculator system example from Section 2, all local states had the same type and each carried just one value, updated in the course of execution,—a multiset of outstanding requests—so we omitted the only location from assertions in the program specs.

Protocols, hooks and worlds. The right part of Fig. 6 shows the components of Disel protocols and worlds. A protocol $\mathcal{P}$ consists of a state-space coherence predicate $\text{coh}$, which defines the shape of the corresponding statelet (i.e., components of the per-node local state and message soup properties), and two finite sets of send- and receive transitions: $T_s$ and $T_r$, correspondingly. Each send-transition is defined by a tag of a message it can send, a precondition, and a step function. The precondition constrains the sender, the addressee, the message to be sent, and the local state of the sender. The step function, which is partially defined, describes the changes in the local state of the sender, assuming that the state satisfies the precondition. Each receive-transition comes with a tag, which uniquely identifies it in a specific protocol. Its precondition is decidable in order to allow the runtime to check it for applicability. Its step function is totally defined. We will use the notations $\tau.tag$, $\tau.pre$ and $\tau.step$ to refer correspondingly to the tag, precondition and step-components of a transition $\tau$, which might be either send- or receive-one.

A world $W$ is represented by a pair $(C,H)$, with its first component $C$ being a collection of protocols that are assigned unique labels. For instance, $\text{deleg server}$ from Section 2 was specified in the context of a world with two protocols with disjoint state-spaces, $C_{\ell_1}''$ and $C_{\ell_2}''$. The second component of a world $H$ contains client-imposed send-hooks. Each hook $h(l_s,l_c,m, to)$ is a predicate, relating a local state of a node $l_s$, which belongs to a core (or server) protocol, a local state $l_c$ of the same node from a client protocol, a content of a message $m$ to be sent and a potential recipient $to$. A hook-map $\text{Hooks}$ associates each hook

---

$^4$The uniqueness constraint is introduced to make the encoding easier in Coq, but our specs and proofs do not rely on it, and the implementation prevents using message ids as values in programs.

---

The programming language of Disel is component of the state constrained by pre- and postconditions of the corresponding Hoare distributed system state that we have used in the examples in Section 2, referring to particular e that returns the value of return node executing the command. The commands also include the standard monadic of a location v to identify the messages to react to. The L labels and transition tags. A decorated receive takes a set of tags T over by a meta-variable f effectful procedures are ranged over by a functional symbol F. The meta-variable F ranges over possibly recursive procedures. Non-interpreted effectful procedures are ranged over by a functional symbol f. Non-Hoare types are ranged over by a meta-variable T. The syntax of Disel commands is given below:

\[
\begin{align*}
  e & ::= \text{send}[\tau_s, \ell](e_m, e_{to}) \mid \text{recv}[T, L] \mid \text{read}_i(v) \mid x \leftarrow c_1; c_2 \mid \text{return } e \mid \text{if } e \text{ then } c_1 \text{ else } c_2 \mid F(e) \\
  F & ::= f \mid \text{letrec } f(x : T) \triangleq c
\end{align*}
\]

Commands include send, receive and read actions, decorated with the corresponding protocol labels and transition tags. A decorated receive takes a set of tags T and a set of protocol labels L to identify the messages to react to. The \text{read}_i(v) command is used to examine the contents of a location v of a local state with respect to the protocol labelled \ell, at the corresponding node executing the command. The commands also include the standard monadic return e that returns the value of e, a sequential composition \(x \leftarrow c_1; c_2\), implemented as a monadic bind (x may be omitted if not used in c2), a conditional statement, and an application \(F(e)\).

Program specifications. Fig. 7 provides the semantics of the assertions with respect to a distributed system state that we have used in the examples in Section 2, referring to particular component of the state constrained by pre- and postconditions of the corresponding Hoare specs. Specifically, a local state assertion \(n \overset{\ell}{\rightarrow} l\) allows one to refer to a specific component \ell of a local state of a node n (which might be different from the one running the code),

\[
\begin{align*}
  s \vdash n \overset{\ell}{\rightarrow} l & \iff \exists d, s(\ell) = (-, d) \land d(n) = l \\
  s \vdash P(MS_1) & \iff \exists MS, s(\ell) = (MS, -) \land P(MS) \\
  s \vdash P_1 \cdot P_2 & \iff \exists s_1, s_2, s = s_1 \uplus s_2 \land s_1 \vdash P_1 \land s_2 \vdash P_2 \\
  s \vdash \text{this } s' & \iff s = s'
\end{align*}
\]

Fig. 7. Semantics of Disel state assertions.
with respect to a protocol labelled $\ell$. The message soup selector $MS_2$ allows one to make statement about message soup of a specific protocol. Finally, the separating conjunction (*), allows one to decompose assertions in the presence of a composite state $s$, which can be represented as a disjoint union of sub-states $s_1 \uplus s_2$. The separating conjunction allows one to combine separately proved specifications wrt. multiple involved protocols, as we did when assigning the type (9) to deleg_server. As is customary in Separation Logic [47], the * operator distributes over plain conjunction for assertions that do not constrain state. this $s'$ allows one to assert that the immediate state is equal to a certain fixed state $s'$.

A command $c$ run by a node $n$ in a world $W$ satisfies a spec $W \vdash^n c : \{P\}\{Q\}$ if it is safe to execute $c$ from a global system state $s$ satisfying $P$, concurrently with programs on other nodes, $c$ respects the protocols and hooks from $W$, and returns a result value $res$, leaving the system in a state $s'$, such that $s' \models Q$ holds. Here and below, we assume that $res$ occurs freely in $Q$. All other unbound variables in $Q$ and $P$ are considered to be logical variables, whose scope spans both pre- and postcondition of the specification, with logical variables in $Q$ (except for $res$) being a subset of those in $P$. In order to describe an effect of an uninterpreted and potentially recursive procedure $f(x : T)$, we employ the following notation for parameterized Hoare specs: $W \vdash^n f(x) : \forall x : T.\{P\}\{Q\}$, where $x$ may occur freely in $P$ and $Q$. The Hoare-style logic of Disel will ensure that all intermediate program-level assertions, describing the global state from a perspective of a node $n$, which runs the code being verified, are stable [20, 57], i.e., closed under observable changes performed by all other nodes, involved into execution of the protocol, and, thus, captured by its definition.

Logic judgements and inference rules. The top part of Fig. 8 shows selected inference rules of Disel. In order to account for typed free program variables and functional symbols $f$, Disel’s judgements are stated in the presence of a typing context $\Gamma$, defined as follows:

$$\Gamma ::= \emptyset \mid \Gamma, x : T \mid \Gamma, f : \langle W, \forall x : T.\{P\}\{Q\}\rangle$$

Typing entries for procedures $f$ include the world $W$ in which their specification was derived. The top two rules, BIND and LETREC, demonstrate the use of typing contexts.

The next two rules, SEND_WRAP and RECEIVE_WRAP, are crucial for program verification in Disel, as they allow one to assign Hoare specifications to atomic decorated send- and receive-commands, instrumented with the suitable protocol annotations. Both rules require user-assigned pre/postconditions to be stable with respect to interference imposed by the protocols in the world $W$. The net effect of sending or receiving a message atomically is captured by the two auxiliary assertion tuples $Sent$ and $Received$, defined at the bottom of Fig. 8, which relate the states $s$ and $s'$ (captured via free logical variables) immediately before and after sending and receiving a message correspondingly.

Specifically, $Sent$ ensures that the precondition of the corresponding send-transition $\tau_s$, holds over the pre-state $s$, as well as all of the hook statements imposed by $H$, which is ensured by the auxiliary predicate $\text{HooksOk}$ defined below in the same figure. The immediate post-state $s'$ is the same as $s$, except for the local state of node $s(\ell)(n)$ of the node $n$ wrt. the protocol $\ell$, which is updated with the effect of the state transition $\tau_s$.step (we use the notation $s(\ell)(n)$ to refer directly to the local state of $n$ of in the second component of $s(\ell)$). Finally, the new message is added to the $\ell$-related message soup $MS_2$ of $s'$. In contrast with sending, receiving messages does not impose any non-trivial preconditions, but in case of a successfully received message (i.e., $res$ is not None), it allows one to learn a number of facts about the pre-state, as captured by the assertions of Received. For instance, the tag $t$ of a received message corresponds to the tag of the corresponding triggered receive-transition $\tau_t$.
of the $\ell$-labelled protocol, so the transition has changed the local state of $n$ accordingly, and also “consumed” the received message in the message soup $MS_\ell$. In conjunction with the protocol invariants, relating local state and message soup properties, this allows one to infer global assertions about the state of the network, as we have shown in Section 2.3.

The premises of these rules rely on the following definition of Hoare ordering $\sqsubseteq$, allowing one to strengthen the precondition $P_2 \Rightarrow P_1$ and weaken the postcondition $Q_1 \Rightarrow Q_2$, while accounting for the local scope of free logical variables in the assertions [25].

**Definition 3.1 (Hoare ordering).** For the given pairs preconditions $P_1, P_2$ and postconditions $Q_1, Q_2$, possibly containing free logical variables, we say $(P_1, Q_1) \sqsubseteq (P_2, Q_2)$ if $\forall s' \ (s \vdash \exists x_1. P_2 \Rightarrow s' \vdash \exists x_2. P_1) \land (\forall x_1. \text{res. } s \vdash P_1) \Rightarrow (\forall x_2. \text{res. } s \vdash P_2 \Rightarrow s' \vdash Q_2)$, where $x_i$ are the free logical variables of both $P_i$ and $Q_i$ correspondingly.

The rule READ is similar to the rules for sending and receiving messages, but it does not modify the local state in any way, observable by other nodes, which is what is ensured by the “atomic specification” in its premise.
The rule Frame is the key to horizontal compositionality with respect to involved protocols. It allows one to add a “framed in” world part \((C, H)\) (with the corresponding assertion \(R\), quantifying over components of \(C\)-relevant state) to a specification, assuming that all involved assertions are stable. This rule is inherently asymmetric due to the “hooking” component \(H\). Specifically, it allows any additions \((C, H)\) as long as hooks in \(H\) cannot invalidate preconditions of send-transitions of \(W\)’s protocols. This check, captured by the NotHooked auxiliary predicate defined at the bottom of Fig. 8, can be done syntactically on the domains of \(W\) and \(H\), just by checking the “intersection” of their “footprints”, very much in the spirit of ordinary Separation Logic. Furthermore, if \(H = \emptyset\), the rule Frame becomes symmetric and can be used to combine any two worlds that do not have mutual inter-protocol restrictions, which is what we did in Section 2.4 when implementing a delegating server. Typically, the world \(W\) contains a number of core protocols (e.g., for locking or replication), whereas the addition \((C, H)\) comes with client-specific protocols and restrictions imposed by the state \(wrt\. W\), so client applications have to be verified in a joint “large-footprint” world \(W \sqcup (C, H)\). In Section 4, we will demonstrate how to make such efforts reusable by exploiting Coq’s higher-order definitions and abstract predicates.

Finally, the rule WithInv allows one to elaborate the context assumptions \(wrt\.\) a specific protocol \(P_\ell\) and also the corresponding state assertions for any invariant \(I\), which is \(P_\ell\)-inductive, i.e., it, as an assertion, over the global network state, is preserved while any node invokes any allowed send- or receive-transitions of \(P_\ell\).\(^5\) Internally, the protocol combinator WithInv\((P_\ell)\) replaces the coherence predicate coh of the protocol \(P_\ell\) with a new one, elaborated with the inductive \(I\). The remaining rules, such as the rule of conjunction, function application, specification weakening etc, are standard and thus omitted.

### 3.3 Program Semantics and Logic Soundness

The semantics of programs and the soundness result in Disel are closely tied to the notion of protocol-aware network semantics. This is a non-deterministic small-step operational semantics, and its two transition rules are shown in Fig. 9 (ignore the gray boxes for now). All free variables in the rules other than \(s\), \(n\) and \(W\) are existentially quantified. That is, the SENDStep-rule will fire for a node \(n\) in a world \(W = (C, H)\) if there is a protocol \(P_\ell\) in \(C\) and there is a send-transition \(\tau_s\) in \(P_\ell\), such that the corresponding local state of the sender \(n\) and the message \(m\) satisfy its precondition and also all \(W\)’s hooks constraining \(\tau_s\) are satisfied. The resulting state will thus have its \(n\)-entry \(wrt\. P_\ell\) updated correspondingly, and a new message added to the soup \(MS\) with a fresh logical message id (omitted here for brevity). The rule RECEIVEStep is similar in that it looks for an active message \(m\) in the soup \(MS\) of an arbitrarily chosen protocol \(P_\ell\), such that \(n\) is its addressee, and its tag corresponds to a specific receive-transition \(\tau_r\) of \(P_\ell\). It then checks the precondition of \(\tau_r\) at \(n\)’s local state, and executes it, updating \(s\)’s local state and soup correspondingly.

One can notice the similarity between the network semantic rules SENDStep and RECEIVEStep and the inference rules SENDWrap and RECEIVEWrap from Fig. 8. This should not come as a surprise: indeed, the two mentioned inference rules provide a way to symbolically account for corresponding local executions of send- receive-transitions by a specific node, consistent with the network semantics.

We build the semantics of programs in Disel with respect to a specific node \(n\) and a world \(W\). To do so, we provide the semantics of wrappers for transitions via the following

---

\(^5\)The formal definition of inductive invariants is with respect to the protocol-aware network semantics, defined in Section 3.3, and is available in the accompanying Coq development.
semi-formal definitions (the formal ones are in our Coq code), accompanied by the natural adequacy result (Lemma 3.4).

Definition 3.3 (Send-wrapper). The operational semantics of a send-wrapper call \( w = \text{send}[\tau_s, \ell](m, to) \) is defined by fixing the grayed elements in the rule SendStep to be the wrapper’s arguments \( \tau_s, m, \ell, \) and \( to. \) The wrapper precondition \( w.pre \) is \( \tau_s.pre \) and its result is \( m. \)

Definition 3.3 (Receive-wrapper). The semantics of a receive-wrapper call \( \text{recv}[T, L] \) is defined by fixing the grayed elements in the rule RecvStep such that \( \ell \in L \) and \( \tau_r.tag \in T \) are chosen non-deterministically. The precondition \( w.pre \) is True and the result is the pair Some \((from, m)\) from \( m, \) if side conditions of \( \text{recv} \) are satisfied and there is a message in the soup matching some tag \( t \in T \) and a label \( \ell \in L, \) or None otherwise.

We use the notation \( s \sim^{\omega_n}_W s’ \) to indicate the effect of a wrapper \( w, \) executed by a node \( n, \) in a global system state \( s, \) such that \( s \vdash W, \) resulting in a new state \( s’. \)

Lemma 3.4 (Wrappers obey the network semantics). Let \( w \) be a send- or receive-wrapper call at a node \( n, \) in a world \( W, \) instantiated with valid arguments. Then for any global state \( s, \) such that \( W \vdash s, \) the resulting state \( s’ \) of a wrapper execution \( s \sim^{\omega_n}_W s’ \) is computable from \( s \) and \( w, \) and \( s \sim^{n}_W s’ \) holds.

A program execution in Disel can be thought of as a sequence of wrapper calls. Indeed, in a distributed system, every such execution at a specific node takes place concurrently with executions on other nodes, which will typically result in multiple possible outcomes for the global state \( s. \) To account for all such behaviors experienced by a program \( c \) running locally, we adopt the trace-based approach for semantics of sequentially-consistent concurrent programs [4]. We define a denotational semantics of a Disel command \( c \) as a (possibly infinite) set of finite partial execution traces \([c] = \{t_\kappa \mid t = [w_1, \ldots, w_n]\}, \) where each element \( w_i \) of a trace \( t \) is a transition wrapper call or an idle step (corresponding to reading local state) as it occurs during a single, potentially incomplete, sequential execution of \( c, \) and \( \kappa \in \{\perp, \text{done } v\}, \) where \( \perp \) indicates an incomplete execution of \( c, \) and \( \text{done } v \) stands for a complete execution returning a result value \( v. \) Thus, a trace \( t \) is generated by a program running at a node, so each of its element corresponds to a single, possible idle, transition, changing the global system state. Since all composite commands in Disel preserve monotonicity in the complete lattice of sets of traces, the semantics of a recursive procedure is defined as the least fixed point of the corresponding functional by the Knaster-Tarski.

Fig. 9. Transition rules of the network semantics.
theorem. That is, DISEL programs are not directly executable within Coq, but are rather extracted into the corresponding OCaml definitions, as we will outline in Section 5.

To give semantics for the Hoare types and formulate a type soundness result, we need several auxiliary definitions, relating program traces and system states. Those are directly inspired by modern concurrency logics [32, 37], and we refer the reader to our Coq code for fully formal definitions. We first define interference-reachable states from a system state \( s \) with respect to a node \( n \):

**Definition 3.5.** A state \( s' \) is interference-reachable from \( s \) wrt. a node \( n \) (denoted by \( s \xrightarrow{\mathcal{W}}_{W} s' \)) iff \( s = s' \) or there exist \( s'' \), \( n' \neq n \), such that \( s \xrightarrow{n'}_{W} s'' \) and \( s'' \xrightarrow{n} s' \).

We next define \( Q \)-satisfying safe traces wrt. a node \( n \), state \( s \), and an assertion \( Q \), as traces executing from \( s \) to the end under interference, so the final state and the result satisfy \( Q \):

**Definition 3.6.** A trace \( t_{\kappa} \) is post-safe for \( n, s \) and \( Q \) iff either

- \( t = [ ] \), \( \kappa = \text{done} \ v \) and \( \forall s', s \xrightarrow{\mathcal{W}}_{W} s' \implies s' \models [v/\text{res}]Q \), or
- \( t = w :: t' \), and for any \( s' \), such that \( s \xrightarrow{\mathcal{W}}_{W} s' \), the state \( s' \) satisfies \( w.\text{pre} \), and for any \( s'' \), such that \( s' \xrightarrow{n} s'' \), \( t'_{\kappa} \) is post-safe for \( n, s'' \) and \( Q \).

Finally, we define well-typed programs via our denotational semantics and post-safe traces.

**Definition 3.7 (Hoare Type Semantics).** \( W \vdash^{n} c : \{P\}\{Q\} \) iff for any \( s \), such that \( s \models P \), and for any trace \( t_{\kappa} \in [c] \), such that \( \kappa = \text{done} \ v \), the trace \( t_{\kappa} \) is post-safe for \( n, s \) and \( Q \).

Definition 3.7 implicitly incorporates fault-avoidance (safety) into the semantics of a type: if a program can be assigned a type, it will safely run from a state satisfying its precondition till the end or diverge, with each wrapper in its trace being able to execute, and the final state satisfying the postcondition. Our implementation comes with a number of lemmas, allowing one to reduce a derivation of a Hoare type for a composite program \( c \) to those of its components, corresponding precisely to inference rules (cf. Fig. 8) in program logics. The proofs of those lemmas with respect to the denotational semantics \([c]\) of specific programming constructs deliver the soundness result of DISEL as a logic:

**Theorem 3.8 (Soundness of DISEL logic).** If the type \( \emptyset ; W \vdash^{n} c : \{P\}\{Q\} \) can be derived in DISEL, the program \( c \) satisfies the spec \( W \vdash^{n} c : \{P\}\{Q\} \) according to Definition 3.7.
4 CASE STUDY: TWO-PHASE COMMIT AND ITS CLIENT APPLICATION

We now present a case study: an implementation and verification in Disel of the basic distributed Two-Phase Commit algorithm (TPC) [60, Chapter 19]. TPC is widely used in distributed systems to implement a centralized consensus protocol, whose goal is to achieve agreement among several nodes about whether a transaction should be committed or aborted (e.g., as part of a distributed database). Since the system may execute in an asynchronous environment where message delivery is unreliable and machines may experience transient crashes, achieving agreement requires care.

The goal of conducting this exercise for us was twofold: (a) to show that the protocol properties established for systems in the distributed systems community (e.g., consensus) are useful for Hoare-style reasoning about program composition and (b) to demonstrate that Disel’s protocols with disjoint state-space and hooks are sufficient for conducting modular proofs about core algorithms (e.g., TPC) and their client applications. To give a better taste of Disel-style programming and verification, in this section we abandon mathematical notation and show fragments of our development taken, with cosmetic adjustments, from our code.

4.1 The Protocol: Intuition and Formalization

The Two-Phase Commit protocol designates a single node as the coordinator, which is in charge of managing the commit process; other nodes participating in the protocol are participants. The protocol proceeds in a series of rounds, each of which makes a single decision. Each round consists of two phases; an example round execution is shown in Fig. 10. In phase one, the coordinator begins processing a new transaction by sending Prepare messages to all participants. Each participant responds with its local decision Yes or No. In the figure, both participants vote Yes, so the coordinator enters phase two by sending Commit messages to all participants, informing them of its decision to commit. If some participant had voted No, the coordinator would instead send Abort messages. In either case, participants acknowledge the decision by sending AckCommit or AckAbort to the coordinator. When the coordinator receives all acknowledgments, it knows that all nodes have completed the transaction.

The component of the coherence predicate constraining the local state \( l \) of each node \( n \) depending on its role, coordinator or a participant, is defined as follows:

**Definition** localCoh \((n: nid) := [Pred l |
if n == cn then \( \exists (r: round) (s: CState) (log: Log), l = st \leftrightarrow (r, s) \uplus lg \leftrightarrow log \)
else if n \( \in pts \)
then \( \exists (r: round) (s: PState) (log: Log), l = st \leftrightarrow (r, s) \uplus lg \leftrightarrow log \) else True].

According to the predicate localCoh, the local state of the coordinator (\( cn \) is a parameter bound at the level of the protocol description) consists of two globally defined locations, \( st \) and \( lg \), for storing the state of the coordinator and logs of messages corresponding to the local state, respectively. For participants, the local state is defined similarly, but with different state and log types specific to participants.

---

Definition c_send_step (r: round) (cs: CState) (log: Log) (to: node) := match cs with
(* Sending prepare-messages *)
| CSentPrep x tos ⇒
  (* switch for receiving responses *)
  if (* sent all messages *) then (r, CWaitPrepResp x [::], l)
  (* keep sending requests *)
  else (r, CSentPrep x (to :: tos), log)
(* ...more cases depending on cs and to... *)
end.

(* Waiting for prepare-responses *)
| CWaitPrepResp x ⇒
  if (* received all votes *) then (r, if (* all votes yes *) then CCommit x
  else CAbort x, log)
  else (r, CWaitPrepResp, log)
(* ...more cases depending on cs, tag, mbody... *)
end.

Fig. 12. Send and receive transitions of a coordinator in a Disel definition of the TPC protocol.

and log, which together store a round number r, a coordinator status s, and a log. The state of a participant (n ∈ pts) is similar, except that its status is a participant status. Finally, any node which is not the coordinator or a participant (e.g., a node participating only in other protocols) may have an arbitrary local state with respect to TPC.

The coordinator’s status can be in any of the seven states shown in shown in Fig. 11(a). Between rounds, the coordinator waits in the CInit status. From the initial state, the coordinator enters the CSentPrep phase and remains in it until all prepare-requests are sent, after which it switches into the receiving state CWaitPrepResp x for the data x. Upon receiving all response messages to the prepare-requests, the coordinator changes either to the commit-state or to the abort-state, notifying all of the participants about the decision and collecting the acknowledgements, eventually returning to the CInit state with an updated log. The participants follow a similar pattern to the coordinator’s, except that a participant sends messages only to or receives messages only from the coordinator.

Fig. 12 shows how to encode a few of the coordinator’s transitions. Recall that Disel transitions are computable functions that describe how to update the local state of the node when executing the transition. The figure shows the snippets of Disel code related to sending a prepare-request messages and receiving a corresponding response message from participants. In the latter case, depending on the responses, once all of them are collected, the coordinator switches to either CCommit or CAbort state.

4.2 Program Specification and Implementation

With the protocol in hand, we can now proceed to build programs that implement the coordinator and participant and assign them useful Hoare-style specifications. An implementation of a single round of the coordinator and its Hoare type are shown in Fig. 13. The function

```coq
coordinator_round (r log) DHT [cn, TPC] (...) represents a Hoare spec, whose logical variables are r and log. The spec is parametrized by the dedicated coordinator node id cn and a world with a single protocol instance TPC, with no hooks. The pre/postconditions (in parentheses) are encoded as Coq functions fun s ⇒... and fun res s' ⇒..., correspondingly, so the immediate pre/post-states s/s' are made explicit, similarly to using the connective this s.

The preconditions, which makes use of the local state getter loc cn s = ..., equivalent to the connective cn TPC ... from Fig. 7, requires that the coordinator is in the CInit state, with an arbitrary round number and log. The postcondition ensures that the local state has returned to CInit, the round number has been incremented, and the return value accurately reflects the decision made on the data, which is also reflected in the updated log. The code proceeds along the lines required by the protocol: it reads the round number from the
```
local state, sends requests, collects the responses and then, depending on the locally stored
result $b$, sends commit/abort messages, collecting the acknowledgements from participants.

4.3 Protocol Consistency and Inductive Invariant

The spec given to \texttt{coordinator\_round} in Fig. 13 only constrains the local state \texttt{loc} of the
coordinator, but in fact the protocol maintains stronger global invariants. For example, we
might like to conclude that between rounds, all logs are in agreement. This strong global
agreement property is not implied by the coherence predicate given above, so we must
prove an inductive invariant that implies it. Finding such inductive invariants is the art
of verification, and the process typically requires several iterations before converging on
a property that is inductive and implies the desired spec. Tools such as Ivy \cite{44} make
the process of finding an inductive invariant much more pleasant by providing automatic
assistance in debugging and correcting invariants, and it would be interesting to connect
\textsc{Disel} to Ivy, which we leave to the future work.

In this case, an invariant that closely follows the intuitive execution of the protocol (its
formulation can be found in our Coq files) suffices to prove the global log agreement property.
For example, when the coordinator is in the \texttt{CSendCommit} state, the invariant ensures that
all participants are either waiting to hear about the decision, have received the decision but
not acknowledged it, or have acknowledged the decision and returned to the initial state.
The invariant also implies a simple statement of \textit{global log agreement}, shown below:

\begin{verbatim}
Lemma cn_log_agreement d r log pt : coh d → TPCInv d →
  loc cn d = st → (r, CInit) ∪ lg → log →
  ∀ pt, pt ∈ pts → loc pt d = st → (r, PInit) ∪ lg → log.
\end{verbatim}

In other words, a coordinator $cn$ in the $CInit$ state and a round $r$ can conclude that \textit{all}
participants $pt ∈ pts$ have also reached the current round $r$ and have logs equal to its own.

\textit{Putting the inductive invariant to work}. We can freely use the elaborated invariant in
proofs of programs. Fig. 14 shows a coordinator program that executes a series of rounds
based on a given list \texttt{data\_seq} of data elements. Its postcondition asserts that all participants
have finished the round and have logs agreeing with the one of the coordinator. The proof
of this specification is by a straightforward application of the \texttt{WITHINV} rule, making use
of the elaborated invariant \texttt{TPCInv} as well as the lemma \texttt{cn_log_agreement}. Importantly,
the postcondition is stable, because each round of the Two-Phase Commit begins with a
coordinator’s move, hence no participant can change its state from the “initial” one while
the coordinator’s status is $CInit$. 

Fig. 13. Spec and code of a coordinator round. Fig. 14. Coordinator spec elaborated with \texttt{TPCInv}. 

Programming and Proving with Distributed Protocols

\begin{verbatim}
Definition coordinator\_round (d : data) :
  \{r log\}, DHT [cn, TPC]
(fun s ⇒ loc cn s = st → (r, CInit) ∪ lg → log,
  fun res s' ⇒ loc cn s' = st → (r+1, CInit) ∪ lg → (log + +(res, d)))
:= Do (r ← read\_round;
  send\_prep\_loop r d;;
  res ← receive\_prep\_loop r;
  b ← read\_resp\_result;
  (if b then send\_commits r d;;
    receive\_commit\_loop r
    else send\_aborts r d;;
    receive\_abort\_loop r););
  return b).
\end{verbatim}

\begin{verbatim}
Definition run\_coordinator (data\_seq : seq data) :
  DHT [cn, _]
(fun s ⇒ s = loc cn s = st → (0, CInit) ∪ lg → [::]
  fun _ s' ⇒ ∃ (choices : seq bool),
    let r := size data\_seq in
    let lg := zip choices data\_seq in
    loc cn s' = st → (r, CInit) ∪ lg → log ∧
    ∀ pt, pt ∈ pts →
    loc pt s' = st → (r, PInit) ∪ lg → log)
:= Do (with\_inv TPCInv (coordinator data\_seq)).
\end{verbatim}

4.4 Composing Two-Phase Commit with a Querying Application using Hooks

Even though core consensus protocols, such as TPC, are not designed to exist in isolation, but rather to be used in a context of larger applications (e.g., for crash recovery), formal reasoning about client-specific properties (i.e., properties of applications relying on certain characteristics of a “core” distributed protocol) is only barely covered in classical textbooks [60] and, with a rare exception [31], almost never a focus of major verification efforts [16, 46, 63], which, therefore cannot be reused in any larger verified context.

We now demonstrate how to employ DISEL’s logical mechanisms for restricted composition of protocols in order to prove, in a modular fashion, properties of client code from a core protocol’s invariants. To do so, we verify a composite application, which uses TPC for building a replicated log of data elements, and a side-channel protocol for sending independent queries about the state of TPC participants (e.g., for the purpose of implementing recovery after a coordinator’s failure). Fig. 15 shows a program that first calls the coordinator program \texttt{run/coordinator}, and then uses the side protocol to query the local state of a participant \texttt{pt}, which the program then returns as its final result \texttt{res}. Ignoring the \texttt{query/init} part in the pre/postcondition for now, notice that the postcondition asserts that \texttt{res} is equal to the pair \texttt{d} (round, log) stored in the local state of the coordinator (which did not crash this time)!

Establishing such validity of the query \texttt{wrt.} TPC-related state is, however, not trivial at all, given how the querying protocol is defined. The protocol \texttt{Query} is very similar to the calculator from Section 2: any node \texttt{n1} in it can send a request to any other node \texttt{n2}, to which \texttt{n2} may respond with any arbitrary message (the details of the formal protocol definition can be found in our Coq code). This protocol definition is intentionally made very weak: while it allows one to prove some interesting inductive invariants (e.g., no request is answered twice), it leaves all other interaction aspects for the final client to specify. In particular, it does not enforce any specific shape of data being sent in a response to a request.

Thus, without imposing the additional restriction that the protocol \texttt{Query can only transmit the local state of a node wrt.} TPC, we will not be able to prove the spec in Fig. 15. The necessary restriction is provided by a send-hook entry \texttt{QHook} that is used when composing the protocols TPC and \texttt{Query in the spec of run_and_query}, and is defined in Fig. 16.

In order to make the client verification effort reusable in the context of any consensus protocol, not just TPC, we formulate the hook statement in terms of an abstract type \texttt{Data} and an abstract predicate \texttt{core_state}, which will be later instantiate specifically for TPC, both afforded by Coq’s higher-order programming capabilities. The hook enforces that any
message \( m \) containing a request id \( \text{rid} \) and serialized \( \text{data} \) adequately encodes the current local state (storing \( \text{data} \)) of the sender node, at the moment of sending \( m \), with respect to the protocol with label \( \text{lab}_c \). The abstract predicate \( \text{core\_state\_d\_lc} \), capturing precisely this “adequacy of the encoding”, is supplied with the injectivity hypothesis \( \text{core\_state\_inj} \) (to be proved by each consensus implementation), which ensures that the abstract data representation is unambiguous.

We also declare an abstract predicate \( \text{local\_indicator} \) and the corresponding hypothesis \( \text{core\_state\_step} \), which essentially corresponds to irrevocability of consensus and should be proved for each consensus implementation (in particular, for TPC), ensuring that if a local state of a node \( n_1 \) is of certain shape \( \text{data} \), the local state of \( n_2 \), captured by \( \text{core\_state\_data} \) will be remaining the same under interference (network\_step) wrt. the core \( \text{lab}_c \)-labelled protocol \( \text{pc} \)—precisely what is ensured by the lemma \( \text{cn\_log\_agreement} \) of TPC.

Finally, we can use the abstract predicates from Fig. 16 to provide specifications for querying procedures from Fig. 15, stating \( \text{query\_init} \) in terms of assertions involving \( \text{local\_indicator} \) and \( \text{query\_state} \), in the context parameterized over a “core” consensus protocol \( \text{pc} \) and restricted with \( \text{QHook} \). To verify the program in Fig. 15 against the desired spec we only need to instantiate the predicates as follows and prove the corresponding hypotheses for TPC, which follow from the invariant \( \text{TPCInv} \) and Lemma \( \text{cn\_log\_agreement} \):

\[
\text{(* For TPC, abstract Data type is instantiated with a round number (nat) and Log. *)}
\]

\[
\text{Definition Data := nat * Log.}
\]

\[
\text{Definition local\_indicator (d : Data) l := l = st \rightarrow (d.1, CInit) \sqcup \text{log} \rightarrow d.2.}
\]

\[
\text{Definition core\_state (d : Data) l := l = st \rightarrow (d.1, PInit) \sqcup \text{log} \rightarrow d.2.}
\]

The rest of the proof is via the FRAME rule with \( W = \langle \text{TPC}, \emptyset \rangle \), \( C = \text{Query} \) and \( H = \text{QHook} \). Since \( \text{QHook} \) does not restrict the transitions of TPC, \( \text{NotHooked} \) holds. Thanks to the parametrization of querying programs with abstract predicates and hypotheses from Fig. 16, we can compose them with any other instance of a consensus protocol, e.g., Paxos [29] or Raft [43], thus, reusing the proofs of their core invariants.

5 IMPLEMENTATION AND EXPERIENCE

DISEL combines two traits that rarely occur in a single tool for reasoning about programs. First, thanks to the representation of Hoare types by means of Coq’s dependent types, the soundness result of DISEL scales not just to a toy core calculus, but to the entirety of Gallina, the programming language of Coq, enhanced with general recursion and message-passing primitives. Second, DISEL programs are immediately executable by means of extracting them into OCaml, which provides the features that Gallina lacks: general fixpoints, mutable state, and networking constructs, enabled by our trusted shim implementation.

Extraction and execution. DISEL’s logic reasons about programs in terms of their denotational semantics as traces, but each primitive also has a straightforward operational meaning. For example, executing a wrapped send transition should actually send the corresponding network message. Thus it is relatively straightforward to extract DISEL programs by providing OCaml implementations of the primitive operations in a trusted shim. Our shim consists of about 250 lines of OCaml, including primitives for sending and receiving messages and general recursion. The local state of each node is implemented as a map from protocol labels to heaps, where a heap is implemented as a map from locations to values. Since DISEL does not draw a distinction between real and auxiliary state so far, both are manifested at run time. In the future, we plan to allow users mark state as auxiliary to improve performance. Due to artifacts of the extraction process, a DISEL program that appears tail-recursive at

the Coq source level does not extract to a tail-recursive OCaml program. This causes long running loops (such as those typically used to implement blocking receive) to quickly blow the OCaml stack. To circumvent this issue, we added a while-loop combinator to DISEL, which is encoded using the general fixpoint combinator, but is extracted to an efficient OCaml procedure that uses constant stack space. Our implementations of the calculator and TPC use this while-loop combinator to implement blocking receive.

In this work, our goal was not to extract high-performance code for DISEL programs, but rather show that, with a careful choice of low-level primitives with precise operational meaning, such extraction is feasible and requires a very small trusted codebase.

### Formal development and proof sizes.

The size of our formalization of the metatheory, inference rules and soundness proofs is about 4500 LOC. Our development builds on well-established Ssreflect/MathComp libraries \[13, 36, 48\] as well as on the implementation of partial finite maps and heap theory by Nanevski et al. \[40\]. Table 1 summarizes the proof effort for the calculator, TPC/Query systems. TheDefs/Specs column measures all specification components, including, e.g., auxiliary predicates, whereas Impl reports the sizes of actual DISEL programs. Due to the high degree of code reuse, it is difficult to provide separate metrics in some cases; for those parts we only report the joint numbers. Although DISEL is not yet a production-quality verification tool, safety proofs of interesting systems can be obtained in it in a reasonably short period of time and with moderate verification effort (e.g., the full development of the core TPC system took nine person-days of work). Given that the current version of DISEL employs no advanced proof automation, beyond what is offered by Coq/Ssreflect, for discharging program-level verification conditions \[5\] or inductive invariant proofs \[44\], we consider these results encouraging.

### 6 RELATED WORK

Program logics for concurrency. DISEL builds on many ideas from modern program logics for compositional concurrency reasoning. The notion of protocols (often called regions) in shared-memory concurrency logics \[8, 37, 45, 52, 55, 56\] provides a “localized” version of more traditional Rely/Guarantee obligations \[20\], which, in their original formulation, are not modular \[11, 12, 57\]. The two closest to DISEL logics employing protocols to reason about interference are FCSL \[37\] and GPS \[55\]. Besides those being logics for shared-memory, rather than message-passing concurrency, protocols in FCSL and GPS are tailored for the notion of ownership transfer \[42\], as a way to express exclusivity of access to shared resources. Due to the lack of immediate synchronization between nodes in a message-passing setting, we consider the notion of ownership to be of less use for most of the systems of interest.

<table>
<thead>
<tr>
<th>Component</th>
<th>Defs/Specs</th>
<th>Impl</th>
<th>Proofs</th>
<th>Build</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculator (§2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>protocol (§2.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INV1 (§2.3)</td>
<td>239</td>
<td>-</td>
<td>243</td>
<td>4.8</td>
</tr>
<tr>
<td>INV2 (§2.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>simple_server (§2.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>batch_server (§2.4)</td>
<td></td>
<td>192</td>
<td>43</td>
<td>153</td>
</tr>
<tr>
<td>memo_server (§2.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>compute (§2.4)</td>
<td>120</td>
<td>24</td>
<td>99</td>
<td>4.8</td>
</tr>
<tr>
<td>deleg_server (§2.4)</td>
<td>75</td>
<td>7</td>
<td>49</td>
<td>2.4</td>
</tr>
<tr>
<td>Two-Phase Commit (§4.1–§4.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>protocol (§4.1)</td>
<td>465</td>
<td>-</td>
<td>231</td>
<td>3.9</td>
</tr>
<tr>
<td>coordinator (§4.2)</td>
<td>236</td>
<td>35</td>
<td>440</td>
<td>18</td>
</tr>
<tr>
<td>participant (§4.2)</td>
<td>163</td>
<td>24</td>
<td>198</td>
<td>10</td>
</tr>
<tr>
<td>TPCInv (§4.3)</td>
<td>997</td>
<td>-</td>
<td>2113</td>
<td>25</td>
</tr>
<tr>
<td>Query/TPC (§4.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>protocol</td>
<td>169</td>
<td>-</td>
<td>115</td>
<td>2.1</td>
</tr>
<tr>
<td>querying procedures</td>
<td>326</td>
<td>18</td>
<td>707</td>
<td>19</td>
</tr>
<tr>
<td>run_and_query</td>
<td>76</td>
<td>5</td>
<td>89</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Table 1. Statistics for implemented systems: sizes of protocol definitions/specs, programs, proofs of protocol axioms/invariants/specs (LOC), and build times (sec).
That said, even though DISEL does not feature explicit ownership transfer, it can be easily encoded on a per-protocol basis, by defining a suitable local state and transitions.

Composition of modular proofs about protocols is a problem that has not received much attention in modern concurrency logics. In FCSL, which tackles a similar challenge, in order to constrain inter-protocol interaction, a user must set up her protocols with a very specific foresight of how they are going to be composed with other protocols, defining "intrinsic \"ownership communication channels\" for all involved components, thus, effectively prohibiting unforeseen interaction scenarios. This is not the case in DISEL: as we have shown in Section 4, \"core\" and \"client\" protocols (e.g., TPC and Query) can be developed and verified independently and then composed in joint applications via \textit{extrinsic} client-specified send-hooks.

The recent logical framework Iris [21, 22] suggests to express protocols as a specific case of resources, represented, in general, by partial commutative monoids, viewing state reachability as a specific instance of framing [47]. This generality does not buy much for verifying distributed applications, as the resulting proof obligations are the same as when proving inductive invariants. Having an \textit{explicit} notion of protocols in the logic, though, allowed us to provide the novel protocol-tailored rules \texttt{WithInv} and \texttt{Frame} (cf. Fig. 8), which enabled modular invariant proofs and distributed systems composition.

A related logic by Villard et al. [59] only considers protocols associated with specific message-passing channels, rather than entire distributed systems. In Villard et al.'s logic, messages do not carry any payload: they are simply \textit{tags}, indicating ownership transfer of a certain heap portion in the \textit{same} shared memory space. It is not immediately obvious how to use Villard et al.'s specifications for \textit{locally} asserting \textit{global} properties of stateful distributed systems (e.g., the agreement of TPC in Fig. 14) without considering all involved processes. In addition to that, Villard et al.'s logic does not provide a mechanism for establishing inductive contract invariants. A recent framework \textit{Actor Services} by Summers and Müller [51] provides abstractions similar to our protocol transitions, but only allows to state \textit{local} actor invariants, and lacks a formal metatheory and soundness proof.

To the best of our knowledge, none of the existing concurrency logics features \textit{both} foundational soundness proof (i.e., the proof that the entire logic, not just its toy subset, is sound as a verification tool), \textit{and} a mechanism to extract and run verified applications.

\textbf{Types for distributed systems}. Session Types [18] are traditionally used to ensure that distributed parties follow a predefined communication protocol \textit{wrt.} a specific channel. While the \textit{multiparty} [19] and \textit{multirole} [7] Session Types enable a form of system composition and role-play, and \textit{dependent session types} allow one to quantify over messages [54], session types do not allow quantification over the global system state and reasoning out of inductive invariants. We believe that DISEL’s combination of Hoare types and protocols provides the necessary level of expressivity to capture rich safety properties of distributed applications. A similar approach has been previously explored in F* by Swamy et al. [53], although that work did not consider reasoning about inductive invariants separately from implementations, neither did it address composition of systems with inter-protocol dependencies.

\textbf{Verification of large systems}. Recent work has verified implementations of core pieces of distributed systems infrastructure, both by using specialized models and DSLs.

IronFleet [16] supports proving \textit{liveness} in addition to safety, all embedded in Dafny [30]. IronFleet focuses on layered verification of standalone monolithic systems. In those systems, each layer is a state-transition system (STS) specifying the system’s behavior at a certain abstraction level, with the top-most layer expressing how a collection of nodes together implement a high-level (e.g., shared-memory) specification, and the actual implementation,
run by the nodes, at the bottom. Adjacent layers are connected by establishing refinement between their STSs via reduction [33], which often involves proving inductive invariants, similar to what we have proven in DISEL. In our understanding, such specifications do not allow for horizontal composition, i.e., reasoning about interaction with separately verified systems in a client code. Such an interaction has been, however, explored wrt. shared-memory concurrency by Gu et al. [15], who built a series of abstraction layers in a verified concurrent OS kernel. That work has shown that establishing a refinement between a specific STS and a family of interacting lower-level STSs is possible, although the proofs are usually quite complex, as they involve reasoning about semantics of a restricted product of STSs. In contrast with those systems, DISEL’s logic does not provide machinery to establish STS refinement, but rather explicitly identifies valid linearization points [17] in the implementations, as they correspond precisely to taken protocol transitions. Abstract specifications and the corresponding system properties, usable by client code, such as consensus, are encoded in DISEL via parametrized Hoare types and abstract predicates, as shown in Section 4.4.

Verdi [62, 63] provides a form of vertical compositionality by means of verified system transformers, which allow systems to be decomposed into layers of functionality (e.g., sequence numbers or state machine replication). The Chapar framework [31] is tailored to causally consistent key-value stores, and also provides verified model checking for client programs using the verified KV stores. Ivy is a tool to assist users in iteratively discovering inductive invariants by finding counterexamples to induction [44]. PSYNC [9] is a DSL allowing one to prove inductive invariants of consensus algorithms in networks with potential faults, operating in a synchronous round-based model [10]. This assumption enables efficient proof automation, but prohibits low-level optimizations, such as, e.g., batching. Mace [23] and DistAlgo [34] adopt an asynchronous protocol model, similar to ours. Mace provides a suite of tools for generating and model checking distributed systems, while DistAlgo allows extraction of efficient implementation from a high-level protocol description. EventML is another DSL for verifying monolithic distributed systems, based on compiling to the Logic of Events in Nuprl [46]. None of these frameworks tackles the challenges of modular reasoning about horizontally composed system implementations (2) and elaborated protocols (3), stated in the introduction of this paper.

Arguably, our Two-Phase Commit implementation is a relatively small case study when compared to the systems verified in IronFleet, Verdi, and EventML. Nevertheless, we are sure that, given enough time and manpower, we can conduct safety proofs of Raft [43] and MultiPaxos [58] in DISEL, as their implementations and invariants are based on the same semantic primitives and reasoning principles that were employed for TPC. We believe, though, that compositional reasoning, afforded by DISEL’s logical mechanisms, is a key to make the results of future verification efforts reusable for building even larger verified distributed ecosystems.

7 CONCLUSION

Almost two decades ago, Lamport propounded the thesis Composition: a way to make proofs harder [28], favoring mathematical models over program logics for real system verification: “in 1997, the unfortunate reality is that engineers rarely specify and reason formally about the systems they build. [...] It seems unlikely that reasoning about the composition of open-system specifications will be a practical concern within the next 15 years”. He was right: it took two decades of active research in rigorous program verification, combining the strengths of mathematical models (protocols) and program logics, to make compositional verification of open-world distributed systems today’s reality.
ACKNOWLEDGMENTS

We thank Philippa Gardner, Alexey Gotsman, Yoichi Hirai, Ranjit Jhala, Luke Nelson, Karl Palmskog, Daniel Ricketts, Doug Woos, and Nobuko Yoshida for their comments on earlier drafts of this paper. We are grateful to the PLDI’17 reviewers, especially Reviewers C and E, for their feedback regarding insufficient support for modularity in an earlier version of Disel, which forced us to revise the approach and introduce the notion of send-hooks. We wish to acknowledge the feedback by the OOPSLA’17 reviewers on the presentation. Finally, we thank the POPL’18 reviewers for the careful reading and many constructive suggestions.

Sergey’s research was supported by EPSRC grant EP/P009271/1 “Program Logics for Compositional Specification and Verification of Distributed Systems”.

REFERENCES

1. Programming and Proving with Distributed Protocols


