Mechanised Verification of Paxos-like Consensus Protocols

Anirudh Pillai

February 28, 2018

Thesis advisor: Professor Dr. Ilya Sergey   Anirudh Pillai

Mechanised Verification of Paxos-like Consensus Protocols

Abstract

Distributed systems are hard to reason about.

Acknowledgments

Acknowledgements

Contents

1

Introduction

1.1 The Problem

1.2 Aims and Goals

I have highlighted the aims and goals separately. The aims are what I want to achieve out of undertaking this project and the goals are the things that this project tries to achieve.

1.2.1 Aims

1. Learn about distributed system protocols

2. Contribute to open source software

1.2.2 Goals

1. Read about and understand the classical Paxos-like consensus algorithms.

2. Develop state transition systems for the algorithms and identify the invariants that need to be preserved during the operation of the algorithm.

3. Implement a simulation of the protocols in Python.

4. Formulate the implemented protocols in Disel by using the developed state-transition systems.

5. Mechanise the proofs of the identified protocol invariants in Disel/Coq.

6. Add additional communication channels and prove composite invariants.

7. Provide an abstract specification of the protocol, usable by third-party clients.

8. Mechanise a client application of the protocol verified out of the abstract interface.

1.3   Project Overview

1.4   Report Overview

2

Background

This chapter lays down all the previous research which the project builds on. Before going over the design decisions on the project we first need to understand this background information and look at related work to see different approaches used to solve the problem.

2.1　Distributed Systems

A distributed system is a model in which processes running on running different computers, which are connected together in a network, exchange messages to coordinate their action, often resulting in the user thinking of the entire system as one single unified computer.

A computer in the distibuted system is also alternatively referred to as a processor or a node in the system. Each node in a distributed systems has its own memory.

We will now go over a few concepts of distributed systems which will help us understand the characteristics of the protocols that run on these systems. This will lay down the groundwork for us to understand the Paxos protocol on which this project is based.

2.1.1 Asynchronous Environment

An asynchronous distributed system is one where there are no guarantees about the timing and order in which events occur.

The clocks of each of the process in the system can be out of sync and may not be accurate. Therefore, there can be no guarantees about the order in which events occur.

Further, messages sent by one process to another can be delayed for an arbitary period of time.

A protocol running in an asynchronous enviroment has to account for these conditions in its design and try to achieve its goal without the guarantees of timed events. An asynchronous environment is very common for a real world distributed system but it also makes reasoning about the system harder because of the aforementioned properties.

2.1.2 Fault Tolerance

A fault tolerant distributed system is one which can continue to function correctly despite the failure of some of its components. A 'failure' of a node or 'fault' in a node means any unexpected behaviour from that node eg. not responding to messages, sending corrupted messages.

Fault tolerance is one of the main reasons for using a distributed system as it increases the chances of your application continuing to functioning correctly and makes it more dependable. As Netflix mention on their blog 'Fault Tolerance is a Requirement, Not a Feature'. With their Netflix API receiving more than 1 billion requests a day, they expect that it is guaranteed that some of the components of their distributed system will fail. Using a fault tolerant distributed system they are able to ensure that a small failure in some components doesn't hinder the performance of the overall system, hence, enabling them to achieve their uptime metrics.

Fault tolerant distributed system protocols are protocols which achieve their goals despite the failure of some of the components of the distributed system they run on. The protocol accounts for the failures and generally specifies the maximum number of failures and the types of failures it can handle before it stops functioning correctly.

2.1.3 State Machine Replication

For a client server model, the easiest way to implement it is to use one single server which handles all the client request. Obviously this isn't the most robust solution as if the single server fails, so does your service. To overcome the problem you use a collection of servers each of which is a replica of the original single server and ensure that each of these 'replicas' fails independantly, without effecting the other replicas. This adds more fault tolerance.

State Machine Replication is method for creating a fault tolerant distributed system by replicating servers and using protocols to coordinate the interactions of these replicated servers with the client.

A State Machine M can be defined as $M = \langle q_0, Q, I, O, \delta, \gamma \rangle$ where

q_0 is the starting state

Q is the set of all possible states. I is set of all valid inputs O is the set of all valide outputs δ is the state transition function, $\delta : IxQ-> Q$ γ is the output function, $\gamma : IxQ-> O$

The state machine begins in the start state and transitions to other states and produces outputs when it receives the inputs. The transition and output are found using the transition and output functions. A deterministic state machine is one whose state transition and output functions are injective, i.e. multiple copies of the machine when given the same input, pass through the same order of states and produce the same output in the same order.

The method of modelling a distributed system protocol as state transition system is very common and is a critical component of this project as we will see soon when we need to encode our protocol in Disel.

State machine replication involves modelling our single server, from the client server model, and using multiple copies (replicas) of the same deterministic state machine and providing all of them with the input from the client. As long as one of the replicas does not crash, while resolving the request, we can successfully return a response to the client.

2.1.4 Consensus Protocols

For handling faults in your distributed system you need to have replication. This leads to the problem of making all these replicas agree with each other to keep them consistent. Consensus protocols try to solve this problem.

Consensus protocols are the family of distibuted systems protocols which aim to make a distributed network of processes agree on one result.

These protocols are of interest because of their numerous real world applications. Let us take the example of a distributed database, which is a critical part of almost all large scale real world applications. This distributed database will run over a network of computers and everytime you use the database you aren't guaranteed to be served by the same computer.

Suppose you add a file to the database. This action is performed by the processor that was serving you 'add' request. Later when you want to retrieve the file from the database you might be served by a different computer that did not perform the 'add' request. Inorder for the new computer to know that the file exists in the database, you will need to use a consensus protocol which helps all the computers in the network (which handle user requests) agree upon the result that the file has been added to the database.

Popular consensus protocols include PageRank used by Google and the Blockchain consensus protocol. George and Ilya verified a subset of the protocol in Coq in their Toychain paper.

2.2 Paxos

Having understood the the main concepts behind distributed system protocols, we can now finally get to the protocol at the heart of this project. Paxos is a family of asynchronous, fault tolerant, consensus protocol which achieves consensus in a network of unrealiable processes as long as a majority of them don't fail.

Paxos is used for state machine replication. Once you have multiple replicas servicing client requests, how do you makes sure that all of these replicas agree on what action to take? The solution is simply to use a consensus protocol like Paxos to make all replicas agree on something.

Paxos has many variants but the one we will focus on is the one we actually prove in Disel, single decree Paxos, also know as simple paxos. Simple Paxos is an algorithm that helps a distributed network of processors to achieve consensus. Consensus is achieved when the network of processor agree on a common value.

For simple paxos, we assume the following assumptions hold about the processors and the environment, in order for the protocol to function correctly.

- Processors communicate between each other by exchanging asynchronous messages between each other.

- Processors run at an arbitrary speed and may fail or restart. Handling this relates to the fault tolerant nature of paxos. Also, we assume that Byzantine faults don't occur. This means that all processors actually work together to try to achieve consensus on a value. There are variants of paxos which can also handle Byzantine failure buy not simple paxos. (This can be linked to the 'PBFT' paper which states that any algorithm handling Byzantine faults must have three phases. Simple paxos only has two phases.)

As for fault tolerance of paxos, in order to handle a failure of upto f processors, we need to have a minimum $2f + 1$ processors participating in the algorithm. This means paxos

8

functions correctly as long as a majority of the processors in the network do not fail. We will see shortly why just a majority needs to function correctly.

A processor participating in simple paxos, may have one or more of these three different roles - proposer, acceptor or learner.

- **Proposer** - A process acting as a proposer listens for client request and proposes a value which the network of processes tries to agree upon.

- **Acceptor** - acceptors receive proposed values from the proposers and then respond to them stating whether they are in a position to accept the value or not. For a proposed value to be accepted, a majority of all the existing acceptors have to accept the proposed value.

- **Learner** - The learner has to be informed when an acceptor accepts a value. The learner can then figure out when consensus has been achieved by calculating when a majority of acceptors have accepted the same proposal. Once the acceptors agree on a value, the learner may act on the value eg. Send request to client informing them about the agreed value.

2.2.1 Choosing a Value

For passing around the value to be chosen from one processor to the other, a processor must send a 'proposal' to the other processor. You can think of a proposal as just a tuple $\langle n, v \rangle$. n is just a natural number associated with a proposal which makes it easy to keep track of all the different proposals.

A *quorum* of acceptors is a subset of the set of all acceptors with length greater than $N/2$ where N is the length of the set of acceptors. A *quorum* is just a set denoting a majority of all the available acceptors.

Consensus is achieved when a proposal is accepted by a majority of acceptors.

Simple paxos runs in rounds until consensus is achieved (a successful round has occurred, where a majority of acceptors have accepted a proposal). A successful round of the algorithm has two phases, each of which can be subdivided into parts a, b.

- **Phase 1a: Prepare Request.** A proposer sends a proposal $\langle n, v \rangle$ to each acceptor in any randomly chosen *quorum* of acceptors. This first message that the proposer sends out is called a *preparerequest*. As it the proposer tries to 'prepare' the acceptors to 'accept' a value in the future.

- **Phase 1b: Promise Response.** An acceptor on receiving a prepare request, responds with a *promiseresponse*, if and only if the acceptor has not already sent a promise response with a proposal containing a proposal number n' where $n' > n$.

  A promise response for proposal $\langle n, v \rangle$ is basically a guarantee (a 'promise') that this acceptor will not respond to any messages with proposals that have a proposal number n' where $n' < n$.

  Thus, if an incoming prepare request has proposal number less than what the acceptor has already promised earlier, then the acceptor can ignore this prepare request by not responding to it. Although, for speeding up the protocol, the acceptor can send out a *nackresponse* which tells the proposer to stop trying to achieve consensus with this proposal.

  If the acceptor has not sent any promise response before, then the body of the promise response can be empty, otherwise the acceptor must include the last proposal that it promised (before the current one) in the body of the message.

- **Phase 2b: Accept Request.** If the proposer successfully receives promise responses from a majority of acceptors, then it can send out an *acceptrequest*. A accept request is a message containing a proposal which tells an acceptor to accept this proposal if it can.

  The proposer creates a new proposal, $\langle n, v' \rangle$ where n is the same as in the proposal which the proposer sent in its prepare request. But, v' is the value from the highest

numbered proposal, selected from all the proposals that the proposer receives in the promise responses. If none of the promise responses received by the proposer contain a proposal, the proposer is free to set v' to any value it likes. The proposer then sends this accept request with proposal $\langle n, v' \rangle$ to another quorum of acceptors.

- **Phase 2b: Accepted Response.** Any acceptor that receives the accept request with proposal $\langle n, v \rangle$ responds with an accepted response if and only if it hasn't already promised not to respond to any proposals with proposal number n' where $n' > n$.

2.2.2 Informing learner

When consensus is achieved, a learner must be informed that a majority of acceptors have agreed on a value. There are various ways to do this.

1. Whenever an acceptor accepts a value, it should send the accepted proposal to all the learners. The learner will then know when a majority of acceptors have accepted the same value.

2. We can have a distinguished learner which informs other learners about the choose value. The acceptors only need to inform this particular learner when they accept a value. This reduces number of messages sent but the distinguished learner becomes the single point of failure and also requires an additional round of sending messages where the distinguished learner informs other learners that a value has been chosen.

3. We can use a set of distinguished learners. The acceptors inform these distinguished learners who then inform the other learners. This increases reliability but also increases the number of messages exchanged.

2.2.3 Inductive Invariant

A invariant in a program is a property of a program that always holds true, from the start through to the end of execution of the program. An invariant can be for something more

specific like a function or even a loop. The only requirement is the property described by the invariant should always hold before, during and after execution of the part of that code.

The problem with using just an invariant like $x > 2$ is that, you may assume that the invariant holds before the execution of the program, but you still do not have a guarantee that this invariant will hold during and after the execution of the program, i.e. in any state of the program.

Therefore, we need to make our invariants *inductive*. An inductive invariant of a program, is an invariant which if it holds in a particular state s of the program, it is guaranteed to hold in all states of the program reachable from s. Thus, having an inductive invariant that holds in the start state of a program is much more useful, as we can be sure that the invariant will continue to hold throughout and after the execution of the program.

This means that once you establish an inductive invariant from the given invariants of a protocol and a starting state for the protocol in which the invariant holds, in order to show that the protocol maintains the invariant, you only need to prove the inductive invariant maintains the induction property over any possible state transition in the protocol. This is exactly how you use an inductive invariant in Disel.

2.3   Disel

Disel is a verification framework, built on top of the Coq theorem prover, that enables one to prove the safety properties of a distributed protocol by breaking down the protocol into its state space invariants and its atomic properties.

2.3.1   Protocol Encoding

We will now look at how Disel requires the state space of a protocol to be encoded in it.

"Each global system state s is a finite partial mapping from protocol labels $l \in$ Lab to

statelets. Each statelet represents a protocol-specific component, consisting of a "message soup" MS and a per-node local state (`DistLocState`)."

As you can see from the image, a protocol *P* in Disel is defined as a tuple of the Coherence, the set of send transitions and the set of recieve transitions. Let us now look at each of these components in detail.

A `Coherence` is a function that takes in a statelet and returns a proposition indicating whether the statelet is valid or not. Thus, the coherence allows us to impose constraints on the local state of each node and on the message soup.

A transition is defined as a tuple of consisting of the following:

1. Tag - a unique natural number identifier for the message to be sent in the transition.

2. Precondition - The constraints that are imposed on identity of the sender of the message, identity of the reciever is, the message that is being sent and on the local state of the sender/receiver (depending on whether it is a send transition/receive transition).

3. Step function - Describes how the local state of the sender/reciever changes after making the transition.

You can see in the code example below how the step function and pre condition are encoded for sending the prepare request in Paxos.

```
(* Changes in the Node state triggered upon send *)
Definition step_send (s: StateT) (to : nid) (p: proposal): StateT :=
    let: (e, rs) := s in
    match rs with
    ...
    (* Step function for the sending prepare request *)
    | PInit p' =>
      if acceptors == [:: to] (* if only one acceptor *)
      then (e, PWaitPrepResp [::] p')
```

```
      else (e, PSentPrep [:: to] p')
    ...
    | _ => (e, rs)
    end.


(* Precondition for send prepare request transition *)
Definition send_prepare_req_prec (p: StateT) (m: payload) :=
  (exists n psal, p = (n, PInit psal)) \/
  (exists n tos psal, p = (n, PSentPrep tos psal)).
```

2.3.2 Protocol to Programs

2.3.3 Protocol Encoding

2.3.4 Approach to Mechanising Proofs in Disel

1. CODE THE CLIENT APPLICATION AND THE PROTOCOL

2. PROVE THE PROTOCOL AND CLIENT APPLICATION

3. IMPLEMENT AND PROVE THE INVARIANT

2.4 Related Work

3

Requirements and Analysis

3.1 Detailed Problem Statement

3.2 Requirements

1. Adapt Paxos for encoding in Disel and devise the state-transition system for this protocol.

2. Develop an inductive invariant for the adapted protocol that ensures the protocol functions correctly by imposing requirements on the global state of the system

3. Implement a simulation of the adapted protocol with the developed state transition system.

4. Mechanise the proof of the adapted protocol in Disel/Coq. Thereby, providing a library of reusable verified distributed components.

5. Mechanise a client application of the protocol verified out of the abstract interface.

3.3 Analysis

3.3.1 Requirement 1: Adapted Protocol and State Transition System

In order to mechanise the proof of Paxos in Disel, we had to first adapt the protocol in order to simplify the proof. There are many variants of Paxos and we had to study the we decided to focus on single decree paxos, the variant that was first proposed by Leslie Lamport

We studied the protocol in detail and also decided to focus on proving the part of the protocol that deals with achieving consensus, which is also the main function of the protocol. For this reason we did not include the learner in our client application or adapted protocol, nor did we focus on the part where the chosen value is learnt by all the nodes. We instead let our inductive invariant handle the case to detect when consensus had been achieved as the inductive invariant can impose requriements on the global system state.

Additionally, we also had to create a state transition system for the nodes in the protocol as Disel relies on this to impose pre and post conditions on the states of a node. In Paxos each node can have different roles but we had to split up each role into different states depending on the current data held by the node and the current function ofthe node in the protocol. We decided on the states each node could be in and how and when it transitions between them. This helped us come up with precondition and postcondition for the state of each node when it transitions on receiving or sending a message. We tried to minimise the number of transitions and the data held in each node's state in order to simplify the proof in Disel.

We look at these in more detail in the next chapter when we talk about modelling the protocol.

3.3.2   Requirement 2: Inductive Invariant

We also had to come up with an inductive invariant for the protocol such that if the inductive invariant holds in some state then in holds in every state reachable from that state. The inductive invariant was critical as it helped ensure that the protocol functions correctly by imposing requirements on the global state of the system. For proving the correctness of paxos we found that our invariant had to capture when consensus is achieved on a value and also that once consensus is achieved on a particular value, further rounds of the protocol don't change this value. We then also came up with a proof for how this inductive invariant holds in our adapted protocol.

3.3.3   Requirement 3: Simulation

I also needed to implement a simulation of our adapted protocol. The simulator must be based on a state-transition system like Disel and should be able to simulate different nodes in the distributed system. The main reason for implementing this simulation is to be sure that our adapted protocol, designed in Requirement 1, will actually be provable in Disel. The simulation will enable us to detect errors and fix them much faster than having to fixing them in the middle of the Disel proof which is a much more time consuming to implement. The simulator will be implemented with the same state transitions we decided upon in Requirement 1, thus, the corrent working of the simulator will give us confidence that our state transition system for Paxos will work correctly in Disel.

I decided to use Python to implement the simulation as that was the language I was strongest in. I studied how the simulation for Multi Paxos was implmented in the Paxos made moderately complex paper, which helped me learn how to create separate process for each node and also how to communicate by exchanging messages between the processes.

3.3.4   Requirement 4: Proof

Having designed the state transition diagram and the inductive invariant, the task of mechanising the proof in Disel becomes much easier. For implementing the proofs, I needed to learn more about Coq and SSReflect. I had to study examples of protocols proved in Disel, like the proof of the Two Phase Commit protocol.

3.3.5   Requirement 5: Client Application

After studying the Disel paper and looking at similar examples, I implemented the core of the adapted protocol in Disel. I also implemented a client application in Disel. The pre and post conditions from the state transition system helped me to implement the client application in such a way to adhere with the main protocol. Using the extraction feature in Disel and the shims runtime, I successfully extracted a working program of the client application in OCaml.

4

Design and Implementation

4.1 Modelling

4.1.1 The Protocol

For mechanising the proof of Paxos in Disel, we followed the approach highlighted in the Disel section of Chapter 2 by first writing and testing the code for the protocol before writing the proofs. Applying this approach meant we needed to come up with the state transition system and the inductive invariant for the protocol which we will encode in Disel. Before coming up with these things though, we decided to simplify the actual simple paxos protocol that we will prove.

The goal of this simplification was to reduce the amount of things that we will need prove in Disel by actually focusing on the 'core' parts of the protocol which lead to consensus being achieved. The 'convenience' parts of the protocol, like the sub phase of informing the learner, aren't necessary for consensus being achieved in the global state and can also be proved separately after we have finished proving the 'core' parts of the protocol.

In order to adapt the protocol and focus on the 'core' parts, the first thing we did was to do away with the 'Learning the Chosen Value' phase of simple paxos. This is the phase where once an acceptor has accepted a proposal, it then informs a leaner by sending it an

accepted request. We decided against proving this because we can use the inductive invariant to check the global state of the system. The inductive invariant allows to check when a majority of acceptors have accepted a proposal and what proposal each of them has accepted. Thus, we can know from the inductive invariant when consensus has been achieved by checking the values of each of the accepted proposals.

Doing away with the learning phase also meant that we did not have to implement a learner in our protocol as the role of the learner (detecting that consensus has been achieved) is performed by the inductive invariant. The learner is useful in actual paxos implementations as it can detect when consensus is achieved and can the inform the client, its presence does not change how consensus is achieved. Removing the learner simplified our state transition diagram for the protocol as we did not have to account for the states of the learner and also the *Phase2b* transition of the accepted request.

Additionally, we decided to remove some optimisation from the protocol. Optimisations are aspects of the protocol that improve its running time or resource consumption but improving upon the 'mundane' way of doing the same thing. Removing optimisations allowed us to simplify our proofs in Disel.

Firstly, we decided that a proposer, if it fails while trying to achieve consensus with a proposal number (i.e. it receives a nack), it then does not retry with a higher proposal number. Removing this optimisation meant we did not have to deal with the state of a proposer where it changes the proposal number it was initialised with. Removing this optimisation only effects the liveness of the protocol but not its safety, as if a proposer receives a nack, it is just an indication that it will not be able to achieve consensus with the proposal number that it is currently using. Thus, in order for consensus to be achieved, a proposer which is initialised with a higher proposal number will have to successfully get promises from a quorum of acceptors.

Additionally we removed the minor optimisation of choosing a majority for sending

requests from the proposer. The proposer instead of choosing a quorum of acceptors for sending a message, chooses the entire set of acceptors. This does not alter the protocol as the entire set of acceptors is a valid quorum of acceptors.

This adapated and simplified protocol allows us to focus on verifying the core logic (the part dealing with achieving consensus) of paxos. The optimisations and 'conveniece' that we removed can be verified separately after the core logic has been verified.

4.1.2   State Transitions

Having adapted the protocol, we then had to create a state transition system for the nodes in your protocol in order to encode it in Disel. For creating the states, we need to look at what the function of each node is in the protocol at a particular moment and what type of data it holds at that time.

A node should only be able to transition from one state to another when it either receives or sends a message. Therefore, the data held by the node in each state should be enough for it to be able to create the message it wants to send or to be able to correctly process the message it receives.

We tried to minimise the number of states and transitions between them, in order to simplify the proof in Disel. This was important because each state transition has to be shown to hold with the invariant so reducing the state transitions, reduces the number of proofs.

We decided that a node can either be initialised as an acceptor or a proposer. The state transitions of each node will depend upon this intial state, so below we will separately look at the state transition systems for the proposer and the acceptor.

The main difference between the state transition for the acceptor and the proposer is that the acceptor sends and receives messages from a single proposer while a proposer has to send and receive messages from all the acceptors.

The proposer starts off in the `PInit` state where it is initialised with a `proposal` (a custom defined data type which is tuple of two natural numbers), $\langle p, v \rangle$. The natural number p is the proposal number and v is the value that the proposer tries to achieve consensus on. This means that the first prepare request this proposer sends will this proposal $\langle p, v \rangle$.

The proposer then moves to the `PSentPrep` state when it starts to send prepare requests to the acceptors. In this state, the proposer still holds the proposal but additionally now also stores a list of natural numbers, `sent_to`. This list stores the natural number identifiers of the acceptors, this proposer has sent requests to. Whenever it sends a prepare request to an acceptor, it adds the identifier of the acceptor to this list. The proposer remains in the `PSentPrep` state and keeps sending prepare requests until the contents of `sent_to` become equal to the global list `acceptors` which holds the identifiers of all the acceptors in the system. This means the proposer stays in this state until it has sent a prepare request to every single acceptor in the system.

Once the proposer has sent the last prepare request, it them transitions to the `PWaitPrepResp` state. In this state the proposer again holds a proposal and another list `promises` which is defined as below to be a list of tuples each containing a `nid` (a natural number identifier for a node), a boolean and a `proposal`.

```
Definition promises := seq (nid * bool * proposal).
```

The proposer stays in this states and keeps receiving messages from the acceptor until one of the following two things happen:

1. It receives a nack response from the acceptor. This indicates that the acceptor might already have promised a proposal with a proposal number greater than p. This leads to the proposer to transition into the `PAbort` state. In this state the proposer basically gives up trying to achieve consensus using the proposal number p that it was
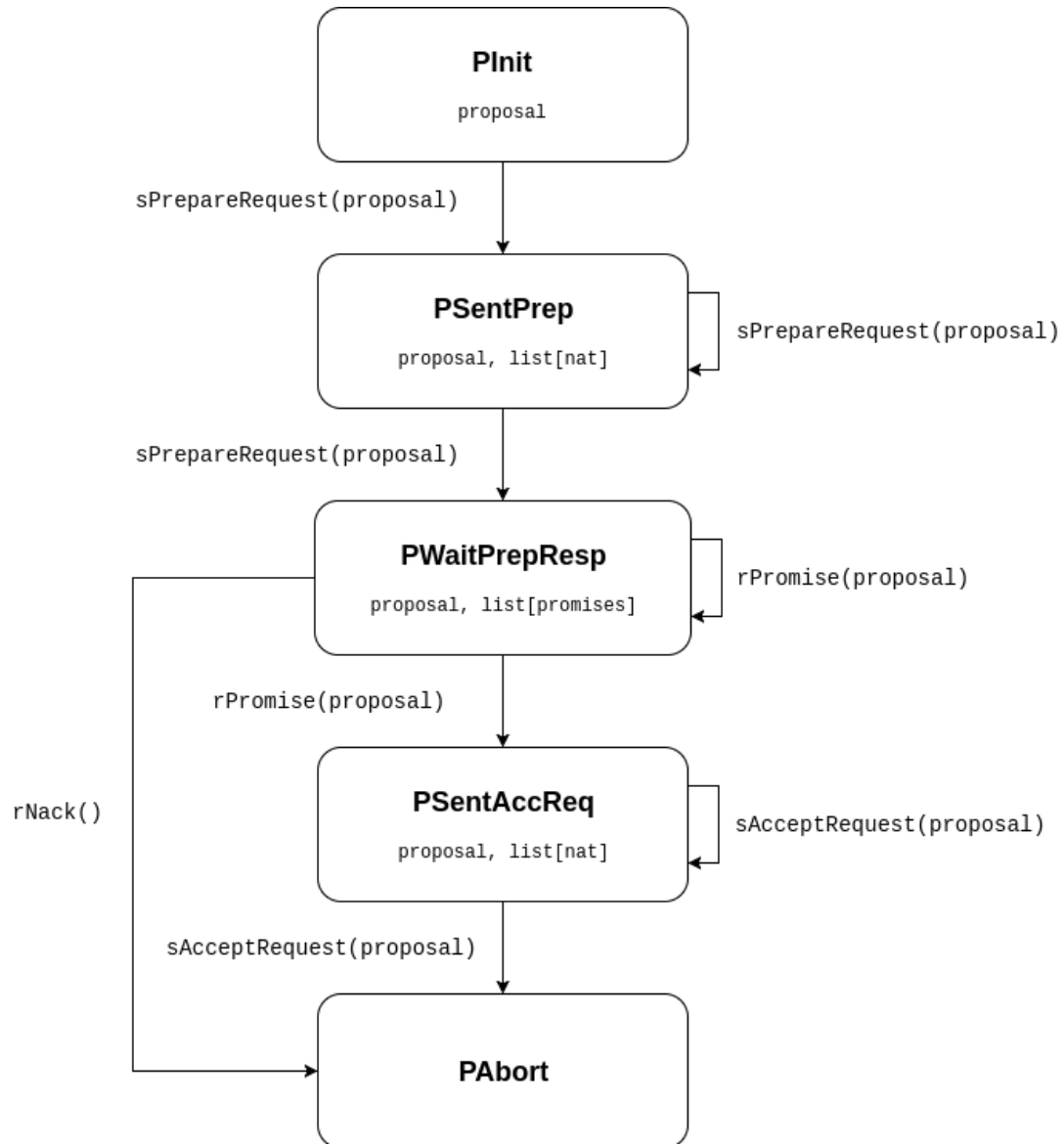
Figure 4.1: Proposer State Transition Diagram

initialised with and completely stops sending and receiving messages. Hence, the proposer doesn't need to hold any data in this state.

2. It receives a promise response from every single acceptor. When this happens, the proposer transitions to the `PSentAccReq` state.

When the proposer reaches the `PSentAccReq`, it means it has received a promise from every single acceptor and it can now start sending accept requests to each of the acceptors in the system. In the `PSentAccReq` the proposer again stores a list `sent_to` to keep track of every single acceptor it has already sent the accept request to. It also stores another `proposal` which is has the same proposal number p that the proposer was initialised with but the value v is the the value from the highest numbered proposal it received in a promise response. In the verification section, we will look at how it determines this value by looping over the `promises` list from the `PWaitPrepResp` state. The sending of accept requests works similar to sending prepare requests in the `PSentPrep` state. Finally, when the proposer finishes sending the accept requests to all the acceptors, it transitions to the `PAbort` state where it stops sending and receiving messages.

Acceptor

The Acceptor starts off in the `AInit` state. It doesn't hold any data in this state as it is not sending any messages. It keeps listening for messages and on receiving a prepare request message, it transitions to `APromised` state.

In the `APromised` state, the acceptor holds a `proposal`. This is the highest numbered proposal that it has received so far in a prepare request message. In this state, on receiving a prepare request, if the proposal number of the proposal in the prepare request is greater than the proposal number of the proposal it currently holds, it updates its current state to hold the new proposal but still remains in the `APromised` state. If the proposal number of
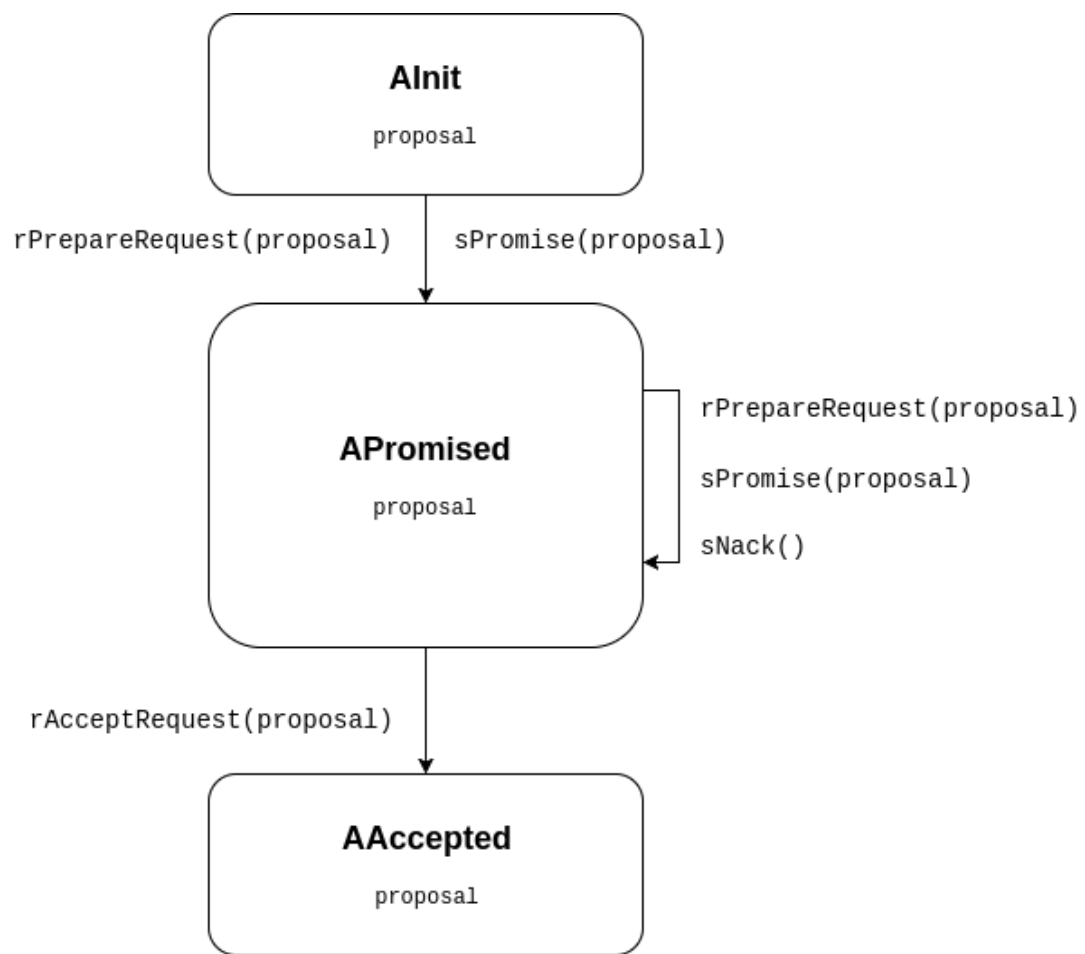
Figure 4.2: Acceptor State Transition Diagram

the proposal in the prepare request is not greater, the acceptor sends a nack response to the proposer who sent the prepare request and does not update its state.

In the `APromised`, on receiving an accept request, if and only if the value of the proposal number proposal in the accept request is greater than the proposal number of the proposal that it currently holds, it transitions to the `AAccepted` state where it now holds the new proposal with the greater proposal number. If the proposal number of the new proposal number is not greater, the acceptor remains in the same state.

In the `AAccepted` state, the acceptor stops listening for and responding to messages. This is similar to the `PAbort` state for the proposer.

4.1.3 Inductive Invariant

A critical part of proving our protocol in Disel was desiging an inductive invariant for our adapted protocol. The inductive invariant helps ensure the correctness of our adapted protocol enabling us to imposing requirements on the global state of the system. For proving the correctness of paxos we found that our invariant had to capture when consensus is achieved on a value and also that once consensus is achieved on a particular value, further rounds of the protocol don't change this value.

Inductive invariant means a property which when it holds for a state s, it will hold for any state s' reachable from s.

The crux of Paxos' correctness lies in the prepare phase where, before sending the accept request, the Proposer must first set the value of the proposal, that it wants to propose, to be the value of the highest numbered proposal it receives as a promise. This ensures that when consensus has been achieved on a value 'v', further rounds of the protocol also ensure that consensus will only be achieved on 'v'.

We established two invariants **I1** and **I2** which together form an inductive invariant for our protocol that also proves its safety.

- **I1** simply tries to say that there can only be one unique value associated with a particular proposal number for any proposal that has been accepted.

- **I2** states that once consensus has been achieved on a value v, every higher number proposal accepted by an acceptor also has the value v.

The mathematical representations for the invariants is given by.

- **I1** - $\forall a_i, a_j \in A, \langle p_i, v_i \rangle \in a_i.accepted \langle p_j, v_j \rangle \in toa_j.accepted \rightarrow v_i = v_j.$

- **I2** - $\forall \langle p_i, v_i \rangle, \forall a_j \in A, \exists \langle p_j, v_j \rangle \in a_j.accepted, p_j > p_i \rightarrow v_i = v_j$

I1 is preserved because if there are n proposers, they are initialised with a unique proposal numbers and throughout the running of our adapted protocol, the proposer always uses this unique proposal number for any value that it proposes. Hence, two different proposers never propose a proposal with the same proposal number. Additionally, each proposer only sends one round of accept requests with the same proposal. So as each proposer proposes only one value with a unique proposal number, we can deduce that each accepted proposal will have a unique value associated with a particular proposal number.

Once consensus has been achieved on a value, further runs of the algorithm don't change the value on which consensus has been achieved. We need to show that once consensus has been achieved on a proposal with value v then every other proposal, with a higher proposal number, on which consensus is achieved will also have proposal value set to v.

In order for consensus to be achieved on a new proposal, the new proposal first needs to be accepted by an acceptor.

$\Rightarrow$ If consensus has been achieved on a proposal $\langle p_1, v_1 \rangle$ then every other proposal $\langle p_2, v_2 \rangle$ accepted by any acceptor, where $p_2 > p_1$, has $v_2 = v_1$.

Further, acceptors can only accept a proposal which has been proposed by a proposer. So we can reduce the requirement as follows.

$\Rightarrow$ If consensus has been achieved on a proposal $\langle p_1, v_1 \rangle$ then every accept request $\langle p_2, v_2 \rangle$ sent by the proposer with $p_2 > p_1$, has $v_2 = v_1$.

In order to prove the above, let's assume that consensus has been achieved on a proposal $\langle p_1, v_2 \rangle$.

$$(4.1)$$

After that lets say that the systems achieves consensus on $\langle p_2, v_2 \rangle$ where $p_2 > p_1$ and there does not exist p_x such that consensus has been achieved on a proposal with proposal number p_x where $p_1 < p_x < p_2$.

So from our assumption (4.1), there must be a majority of acceptors such that they have accepted the proposal $\langle p_2, v_2 \rangle$. So we need to show that in $v_2 = v_1$. This is ensured in Paxos because of Phase 1 where the proposer must first get promises from a majority.

So any majority the proposer for p_2, v_2 gets in Phase 1, will have at least one acceptor a which has accepted $\langle p, v \rangle$. Paxos also ensures that before sending the accept request for p_2, v_2, the proposer must select the value of the highest numbered proposals which it receives in its promises.

So the Acceptor a will send $\langle p_1, v_1 \rangle$ in its promise message to the proposer. As $\langle p_2, v_2 \rangle$ is the only proposal number which has proposal number greater than p, the proposer must set $v_2 = v_1$ in its accept request message $\langle p_2, v_2 \rangle$ as v_1 is the value of the highest numbered proposal that it receives as a promise response. Thus, meeting our above requirement.

4.2   Verification

5

Client Application

As outlined in chapter 2, our approach to using Disel involved first coding the protocol and the client application, and then going on write the proofs. Having the client application working before writing the proofs gave us confidence that the design of our adapted protocol worked and could be used to achieve consensus.

In this chapter we look at how we implemented a simple client application that uses the proof of our adapted protocol to create an implementation of nodes that achieve consensus on a value. We first look at how the client application was designed and coded in Disel then we look into how we used Disel's `shims` runtime to extract the OCaml code for the client. Finally, we will look at how we verfied the client application.

5.1   Modelling

Disel programs can be extracted into their corresponding OCaml definitions. The extracted code contains modules to that define the various node states and transitions. This extracted code can then be used by a shim to create a client application.

In order for the extraction to work, we need to assign a program to run for each node participating in the protocol. Additionnaly, each node needs to be given an initial state that satisifed all the imposed invariants.

The design of our client application was very simple. We aimed to initialise two proposers, proposer1 with proposal $\langle 1, 1 \rangle$ and proposer2 with proposal $\langle 2, 2 \rangle$ along with three different acceptors.

Having decided on this, we needed to write the runnable implementation of a proposer and an acceptor which each of these nodes can run as programs. Here we will only look at the implementation of the proposer, the implementation of the acceptor follows from that and can be found in the PaxosAcceptor.v file.

As a proposer starts off in the PInit state, the runnable implementation of a proposer needs to take in a proposal as a parameter which it will use to initilise the proposer with. Below is the main function for the proposer. It first sends out the prepare requests and then starts receiving responses from the acceptors. check_promises function checks that none of the responses contain a nack request. If no nacks were received then the proposer sends accept requests to all the acceptors by choosing the value from the highest numbered proposal.

```
Program Definition proposer_round (psal: proposal):
  {(e : nat)}, DHT [p, W]
  (fun i => loc i = st :-> (e, PInit psal),
   fun res m => loc m = st :-> (e.+1, PAbort))
  :=
  Do (e <-- read_round;
      send_prepare_req_loop e psal;;
      recv_promises <-- receive_prepare_resp_loop e;
      check <-- check_promises recv_promises;
      if check
      then send_accept_reqs e (choose_highest_numbered_proposal psal recv_promises)
      else send_accept_reqs e [:: 0; 0]).
    (* If check fails then send an acc_req for (0, 0) which will never be
       accepted by any acceptor *)
```

Although, if a nack was recieved, the proposal still sends accept requests with the proposal $\langle 0, 0 \rangle$. This proposal will never be accepted by any acceptor as its proposal number is not greater than 0. We still need to send these accept requests as both branches of a `if` statement need to have the same type.

One can also see the distributed Hoare types and the pre and post conditions defined above. Once the proposer finished the round, it ends up in the `PAbort` state with its round number incremented. A proposer will only have one go at achieving consensus and will not try again if it fails. This corresponds to the decision we made while designing the adapted protocol.

5.2 Extraction

Before running the extraction, we need the `SimplePaxosApp.v` file which defines the code to instantiate the nodes. Our client implementation instantiates two proposers and three acceptors. Each proposer is instantiated with a unique proposol which is required in the `PInit` state.

```
(* [Run] Runnable proposer code *)
Program Definition run_proposer p psal:
  DHT [p, _] (
    fun i => network_rely W p init_state i,
    fun _ m => exists (r : nat),
    getLocal p (getStatelet m l) = st :-> (r, PInit psal))
  := Do (with_inv (PaxosInductiveInv.ii _ _ _) (proposer p psal)).
...


(* [Run] Runnable nodes *)
Program Definition run_proposer1 := run_proposer p1 psal_1.
```

```
...
(* [Run] Final programs to run with actual arguments supplied *)
Definition p_runner1 (u : unit) := run_proposer1 [:: 1; 1].
```

After extracting the Coq code, we then wrote a shim file (`PaxosMain.ml`) that parses
the arguement supplied to it and instantiates a program specified for the given node. Below
is the main function from the shim file.

```
let main () =
  parse_args (List.tl (Array.to_list Sys.argv));
  match !mode, !me with
  | Some mode, Some me ->
    begin
      Shim.setup { nodes = !nodes; me = me; st = SimplePaxosApp.init_state };
      match mode with
      | Acceptor ->
        begin match int_of_nat me with
        | 3 -> SimplePaxosApp.a_runner1 ()
        | 4 -> SimplePaxosApp.a_runner2 ()
        | 5 -> SimplePaxosApp.a_runner3 ()
        | n -> usage ("unknown acceptor name " ^ string_of_int n)
        end
      | Proposer ->
        begin match int_of_nat me with
        | 1 ->
          begin
            try
              SimplePaxosApp.p_runner1 ()
            with _ -> print_endline "A acceptor closed its connection, proposer
                exiting."
          end
        | 2 ->
```

```
      begin
        try
          SimplePaxosApp.p_runner2 ()
        with _ -> print_endline "A acceptor closed its connection, proposer
            exiting."
      end
    | n -> usage ("unknown proposer name " ^ string_of_int n)
    end
  end
| _, _ -> usage "-mode and -me must be given"
```

Finally, we had to write a shell script to call the shim file with the separate arguements to instantiate all the different nodes as different processes.

5.3   Verification

Once we could run the client implementation and see consensus being achieved, we then moved on to writing the proofs for the implementation.

6

Conclusion and Evaluation

References

A

Interim Report