

# Modelling the Raft Distributed Consensus Protocol in mCRL2

Parth Bora

Pham Duc Minh

Tim A.C. Willemse

Department of Mathematics and Computer Science  
Eindhoven University of Technology  
Eindhoven, The Netherlands

p.bora@student.tue.nl

minh.pham@ximuis.eu

T.A.C.Willemse@tue.nl

The consensus problem is a fundamental problem in distributed systems. It involves a set of actors, or entities, that need to agree on some values or decisions. The Raft algorithm is a solution to the consensus problem that has gained widespread popularity as an easy-to-understand and implement alternative to Lamport’s Paxos algorithm. In this paper we discuss a formalisation of the Raft algorithm and its associated correctness properties in the mCRL2 specification language.

## 1 Introduction

Consensus is the process of reaching an agreement on a particular issue or decision among a group of entities or individuals. In the context of distributed systems, reaching consensus is challenging, in particular because the entities are scattered across the network and need to use communication to reach agreement on decisions. Naive solutions to consensus may then lead to faulty decisions, mainly due to communication being asynchronous and potentially unreliable, or entities that may disappear and reappear. Consensus is a fundamental ingredient for guaranteeing security and reliability of, *e.g.*, blockchains and distributed ledgers.

The Paxos algorithm [8], devised by Lamport, and its variations, is one of the most well-known solutions to the consensus protocol. While the algorithm has been studied widely, it is considered to be rather involved and hard to understand and implement. Consequently, there have been many attempts to find alternative, simpler solutions to the consensus problem. The Raft algorithm [12, 11], proposed by Ongaro and Ousterhout in 2014, is one such alternative. It is generally regarded to be simpler because it breaks down the process of reaching consensus in smaller subproblems. Raft is used in, *e.g.*, *etcd*, a popular key-value store for coordinating distributed systems, facilitating service discovery, *etc.* The Raft algorithm is based on a leader-follower model, where a leader is elected among the entities to make decisions and propagate them to other entities. The other entities follow the leader’s decisions and thereby reach consensus.

The Raft algorithm achieves fault tolerance using state machine replication. This is a technique for implementing a fault-tolerant service, which uses replication of servers and which coordinates the interactions between clients and server replicas. Each server hosts a state machine that generates an identical copy of a particular state [13], thus ensuring that in the event of (a limited number of) server failures, the system remains operational. Typically, state machine replication involves log replication. In the Raft algorithm, the log is simply a sequence of commands with some minimal additional information, which it keeps consistent. The logs are maintained by every server in the network and executed sequentially by these, and their uniformity guarantees that servers processes the same commands in the same order.

Given the practical significance of the consensus problem and the complexity of the solutions to the problem, found in the literature, formalising and analysing these solutions is highly relevant. The Raft algorithm has been modelled and verified in TLA+ [9] by Ongaro<sup>1</sup>, one of the authors proposing the Raft algorithm. This specification contained a couple of minor mistakes which have been fixed, as pointed out by Evrard in [4], where an LNT [5] model of Raft is discussed. An earlier version of the LNT model has been used in the Model Checking contest [7] in 2015, where a few generic requirements were analysed. Another model of the Raft algorithm was presented in [14]. They used the Verdi framework to formally prove the State Machine Safety property, *i.e.*, the property that logs that appear in each node must provide a uniform, consistent view on the state of the servers.

In this paper, we discuss a model written in the mCRL2 language [6] and the formalisation of several properties coined in [12, 11]. The mCRL2 language is a process algebra with data; its process language is based on the **algebra of communicating processes (ACP)**, whereas its data language is based on the theory of abstract data types. The language is supported by the mCRL2 tool set [1], which allows for generating and visualising state spaces, and which can be used to verify properties expressed in the modal  $\mu$ -calculus with data. While both mCRL2 and LNT are process algebras, their syntax is quite different, and modelling in both languages requires quite a different style. We discuss the design decisions underlying our model of the Raft algorithm, and present modal  $\mu$ -calculus formalisations of the properties.

**Outline.** We discuss our mCRL2 model of the Raft algorithm in Section 2. The mCRL2 language is introduced and explained using snippets of our model. For a full explanation of the language, we refer to [6]. In Section 3, we describe the properties that we formalised in the modal  $\mu$ -calculus. We discuss some of our findings in Section 4, and end with conclusions and future work in Section 5. Full details of the model and the properties can be found in the accompanying artefact in the Mars repository<sup>2</sup>.

## 2 Modelling RAFT in mCRL2

Our mCRL2 model of the Raft Algorithm focusses on the behaviour of the nodes in the network. For our models, we draw inspiration from the TLA+ [11] and LNT [4] specifications of the protocol and, like the LNT and TLA+ specifications, focus on leader election and log replication as these form the core of the protocol. Features such as cluster membership changes and log compaction have not been modelled for simplicity's sake and in the interest of keeping the state space minimal. We additionally model a communication infrastructure that facilitates reliable communication between nodes. **Our network model can be modified easily to also capture unreliable communication, but this is not our initial focus.** The nodes process commands that can be sent by clients; in our model, the latter is a simple process that has no other purpose than to send commands.

All actors are modelled as dedicated (parameterised) processes in mCRL2: we have `Node` processes, a `Network` process and a `Client` process. The actors run in parallel and can synchronise and exchange data by executing communicating actions. In mCRL2, this is defined by a top-level process such as:

```
init allow( {sendRPC, receiveRPC, clientCommand, advanceCommitIndex, timeout, sendRPCset ...},
  comm ( { sendClientRequest | recvClientRequest -> clientCommand,
          sendToNetwork | receiveFromServer -> sendRPC,
          sendToServer | receiveFromNetwork -> receiveRPC,
          sendToNetworkSet | receiveFromServerSet -> sendRPCset },
```

<sup>1</sup><https://github.com/ongardie/raft.tla>

<sup>2</sup><http://mars-workshop.org/repository.html>

```

    Client(1) || Node(...) || Node(...) || ... || HealthyNetwork(...)
  )
);

```

Parallelism is modelled by means of the parallel operator ‘`||`’; which actions communicate is declared using the communication operator ‘`comm`’, by specifying which pairs of action labels can engage in a communication. For instance, `sendToNetwork | receiveFromServer -> sendRPC` specifies that when a parameterised action with action label `sendToNetwork` and a parameterised action with action label `receiveFromServer` can happen simultaneously (provided their parameters match), this results in a `sendRPC` action carrying the parameters of the individual actions. By disallowing actions that are meant to communicate, synchronisation is enforced. This is achieved by means of the ‘`allow`’ operator, which blocks any action other than the ones for which an action label is specified in the set of allowed action labels. For instance, by including the `sendRPC` action label, every action with an `sendToNetwork` action label is blocked and only actions with an `sendRPC` action label are allowed.

A Raft cluster may have any number of Nodes. Analysing our model using simulation (*i.e.*, stepping through the model interactively) or verification (*e.g.*, computing the validity of requirements fully automatically) to assess the correctness of (our model of) the algorithm, however, requires a fixed, concrete number of servers. Since the behaviour of the Raft algorithm crucially depends on the number of Nodes in the network, we model this number by means of a constant that all our processes can refer to. This is done as follows:

```

map NumberOfServers: Nat;
eqn NumberOfServers = 3;

```

This declares a constant `NumberOfServers` and sets it to 3; this constant should be the same as the number of `Node` processes running in parallel in the top-level process. Our model contains a few other constants which can be set similarly.

In the remainder of this section, we describe the `Client` process and the `Network` process (Section 2.1) and the `Node` process (Section 2.2).

## 2.1 The Raft Environment

Clients of the Raft algorithm can use it to store data and request commands that are to be executed on multiple interconnected Nodes. These Nodes operate independently and may hold different copies of the same data, with the consistency thereof being guaranteed by the Raft algorithm. For the purpose of analysing the algorithm, we introduce a simple client model: only a single client and, since we are not interested in the actual data or the commands issued by this client, we use unique ID’s, modelled by natural numbers `Nat`, to abstract from the different messages of the client:

```

proc Client(clientCommandID: Nat) =
  (clientCommandID <= NumberOfClientRequests) ->
    sendClientRequest(clientCommandID) . Client(clientCommandID+1);

```

This defines a process `Client` that can be instantiated by passing a positive number as argument. Assuming that the constant `NumberOfClientRequests` is 3, process `Client(1)` then executes the action `sendClientRequest(1)`, followed by `sendClientRequest(2)` and finally `sendClientRequest(3)`, after which the process is unable to perform any further actions. This behaviour is described compactly using the sequential composition operator ‘`.`’ of mCRL2, and through the use of recursion.

We assume that communication between the client and the Raft cluster is synchronous, unlike the communication among the different nodes in a Raft cluster, which proceeds asynchronously. Raft claims

to be correct even when network communication between nodes is unreliable, including delays, partitions, and packet loss, duplication, and reordering. As mentioned earlier, communication in our model of Raft happens via the communication of actions between the various `Node` processes and the `Network` process. If node *A* wants to send a message to node *B*, it sends the message to the network, which then sends it to node *B*. The network layer is introduced as an intermediary in message exchange between nodes to model message reordering. While we do not analyse the Raft algorithm in the presence of message loss or duplication, our network model can easily be modified to accommodate for these. Messages exchanged between nodes are essentially Remote Procedure Calls (RPCs). Raft utilises two distinct types of RPCs: *vote request/response* RPCs and *append entries request/response* RPCs.

```
sort RPC = struct RequestVoteRequest(currentTermRPC: Nat, endLogIndex: Nat, endLogTerm: Nat)
              ?isRequestVoteRequest
            | RequestVoteResponse(currentTermRPC: Nat, isVoteGranted: Bool)
              ?isRequestVoteResponse
            | ...
```

We model the messages exchanged between a node and a network using the data type `NetworkPayload`, which is a triple consisting of the ID of the sending node, a command of type `RPC` and the ID of the receiving node:

```
sort NetworkPayload = struct Message(senderID: Nat, rpc: RPC, receiverID: Nat);
```

Our network model allows a node to send a message to another nodes using a `SendToNetwork` action, which can then communicate with a `receiveFromServer` action offered by the network. Alternatively, a set of nodes can be addressed in one go, using a `sendToNetworkSet` action and which can communicate with a `receiveFromServerSet` action offered by the network. The network then takes care of dispatching the messages to these nodes using a `sendToServer` action. This is achieved by the following process:

```
proc Network(messageCollection: FSet(NetworkPayload)) =
  (# messageCollection < NetworkSize) ->
    sum msg: NetworkPayload . receiveFromServer(msg)
                          . Network(messageCollection = messageCollection + {msg})
  +
  (# messageCollection + NumberOfServers < NetworkSize + 1) ->
    sum msg: FSet(NetworkPayload) . receiveFromServerSet(msgs)
                          . Network(messageCollection = messageCollection + msgs)
  +
  sum msg: NetworkPayload .
    (msg in messageCollection) ->
      sendToServer(msg) . Network(messageCollection = messageCollection - {msg});
```

Informally, this process can, execute a `receiveFromServer` action carrying a (non-deterministically chosen) message of type `NetworkPayload`, as long as the network is not yet full, indicated by the condition `# messageCollection < NetworkSize`. Once the action was executed, the process again behaves as `Network`, but the parameter `messageCollection` has been updated to also contain the message `msg`. Alternatively, as indicated by the binary non-deterministic choice operator ‘+’, the process may receive a message that is to be sent to all other nodes (second summand), or send a message that is in the set of messages `messageCollection`.

Note that the model depicted above models a ‘perfect’ network; however, using a minor adjustment, it can be turned into an unreliable network. For instance, by extending the third summand to include an option to lose the message instead of sending the message, message loss can be modelled as follows:

```
...
+
  sum msg: NetworkPayload .
```

```

(msg in messageCollection) ->
(
  sendToServer(msg) . Network(messageCollection = messageCollection - {msg})
+
  lose . Network(messageCollection = messageCollection - {msg})
);

```

Here, `lose` is a new action indicating a message is lost, not revealing which message this is. By including this action in the set of actions that are in the `allow` set of the top-level process, the network can non-deterministically decide to drop messages.

## 2.2 Node

The core logic of the Raft algorithm is described by the `Node` processes. This process needs to deal with messages received from other `Node` processes, and, send messages (potentially received from a `Client` process) to other `Node` processes. Logical decisions are based on the local state of the process; this state is reflected in the parameters of the `Node` process:

```

proc Node(id: Nat, currentState: State, currentTerm: Nat, log: LogType,
  commitIndex: Nat, votedFor: Int, voterLog: FSet(Nat), nextIndex: List(Nat),
  matchIndex: List(Nat), replyToBeSent: replyHelper) =
  currentState != Crashed ->
  ( (IsNone(replyToBeSent)) ->
    Node_process_receiveFromNetwork(id, currentState, currentTerm, log,
      commitIndex, votedFor, voterLog, nextIndex,
      matchIndex, replyToBeSent)
+
    Node_process_sendToNetwork(id, currentState, currentTerm, log, commitIndex,
      votedFor, voterLog, nextIndex, matchIndex, replyToBeSent)
+
    (currentState != Leader && currentTerm < MaxTerm) ->
      timeout . Node(currentState = Candidate, currentTerm = currentTerm + 1,
        votedFor = id, voterLog = {id}, replyToBeSent = none)
  )
+
  currentState == Crashed -> ...;

```

As can be seen, for the sake of readability we have split part of the `Node` process in two subprocesses, viz., the process `Node_process_receiveFromNetwork` and `Node_process_sendToNetwork`. So long as the node has not crashed, it offers a non-deterministic choice between the behaviour described by these two processes and (conditionally) timing out (as described by the third summand). Subprocess `Node_process_receiveFromNetwork` handles all messages the node receives through the network, whereas `Node_process_sendToNetwork` takes care of sending messages, received from the client, or replies to previous messages, to other nodes. If the node has crashed, a recovery mechanism can be initiated (not depicted here).

The Raft algorithm divides time into terms of arbitrary length. The current term number is represented by the `currentTerm` parameter of type `Nat` of the `Node` process. Each node in a Raft cluster can be in one of three states: *Leader*, *Follower*, or *Candidate*; see also Fig. 1. This state is maintained by parameter `currentState` in process `Node`. In addition to these three possible states, we have introduced a fourth state to indicate that the node has crashed. The data type `State` thus is as follows:

```
sort State = struct Leader | Candidate | Follower | Crashed;
```

The way Raft (and algorithms like Raft) implements replicated state machines is by means of a replicated log. Each node in the Raft cluster stores a log consisting of entries that contain a state machine command and the term number that indicates when the entry was received by the *Leader*. In our log entries, the

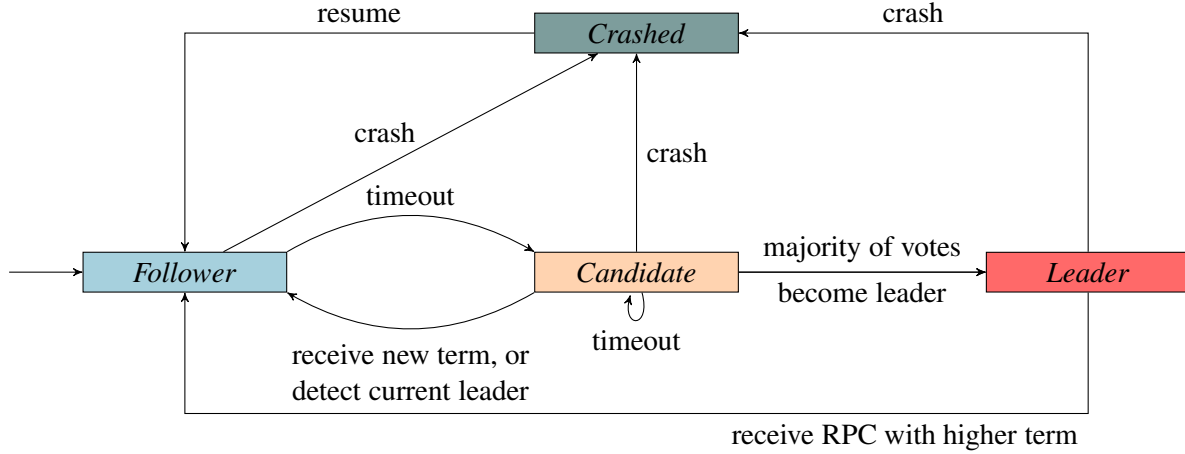


Figure 1: State transitions for a node in Raft.

state machine commands are represented by the command ID, since we are not interested in the actual command itself. The parameter `log` of process `Node` is thus basically a list of entries that contain a command ID and term:

```

sort logEntry = struct Command(term: Nat, commandID: Nat);
sort LogType = List(logEntry);

```

A practical complication with our formalisation is that mCRL2 lists are zero-indexed, unlike the logs described in the original Raft paper [12, 11], which are one-indexed. While converting from one representation to the other is straightforward, it is equally easy to make mistakes. We have circumvented this by utilising helper functions that make the conversion less error-prone.

**The Raft State Machine.** All nodes start out as *Followers*. Depending on the state of the node, certain actions are permitted. Only when a node is in state *Leader*, it can start accepting messages from a client. A node that is *Leader* sends periodic heartbeats to followers to assert its presence. During an election, nodes that are *Candidates* send *vote request* RPCs to other nodes to garner votes. Subprocess `Node_process_sendToNetwork` takes care of these events. If a *Follower* does not receive either of these messages over a period of time, called the election timeout and modelled by means of the `timeout` action, it starts a new election by changing into a *Candidate* state and increments its term. Subsequently, it will vote for itself. We remark that we here closely follow the LNT model, allowing the *Candidate* to vote for itself rather than by sending a *vote request* RPC to itself, and by modelling the timeout by means of non-determinism rather than by imposing hard real-time requirements. Safety requirements should not be affected by modelling timeouts using non-determinism. However, due to this abstraction, we cannot analyse real-time requirements, nor the real-time performance of the algorithm. Also, when phrasing liveness requirements, the abstraction may require one to be explicit about the absence or occurrence of these timeouts.

After a node becomes a *Candidate*, it sends a *vote request* RPC to all other servers in the cluster. In our model, this is achieved using a `sendToNetworkSet` action, carrying a set of messages consisting of the RPC and a target node as its parameter; the `Network` process then relays the request to all targeted nodes. The set of messages is created using a recursive function `CreateRequestVoteSet` that builds the set by iterating over all possible node IDs that have not voted for the *Candidate* node yet; the latter is

specified in an auxiliary function `CreateRequestVoteSetHelper`:

```
map CreateRequestVoteSet: Nat # Nat # Nat # Nat # FSet(Nat) -> FSet(NetworkPayload);
var sender, termNode, lengthLog, lastTermLog: Nat;
    voterLog: FSet(Nat);
eqn CreateRequestVoteSet(sender, termNode, lengthLog, lastTermLog, voterLog)
    =
    CreateRequestVoteSetHelper(sender, RequestVoteRequest(termNode, lengthLog, lastTermLog),
        voterLog, 0);

map CreateRequestVoteSetHelper: Nat # RPC # FSet(Nat) # Nat -> FSet(NetworkPayload);
var sender, receiver: Nat;
    rvr: RPC;
    voterLog: FSet(Nat);
eqn (receiver==NumberOfServers) ->
    CreateRequestVoteSetHelper(sender, rvr, voterLog, receiver) = {};
(receiver<NumberOfServers) ->
    CreateRequestVoteSetHelper(sender, rvr, voterLog, receiver) =
    CreateRequestVoteSetHelper(sender, rvr, voterLog, receiver + 1 )
    + if(receiver!=sender && !(receiver in voterLog), {Message(sender, rvr, receiver)},{});
```

If a node receives a stale message, *i.e.*, a message with a term smaller than `currentTerm`, it immediately discards it. When it receives a message with a term greater than `currentTerm`, the node steps down to the *Follower* state and resets the `votedFor` parameter to `-1`, to indicate it has not voted for anyone in that term, and it sends a reply. The type of message received determines the type of reply sent by the node. This reply is then stored in the `replyToBeSent` parameter so that it can be sent out before the node engages in other interactions but potentially only after the node has updated its state. This allows for analysing the effects (in any) of nodes crashing random moments. In particular, when nodes crash, part of their state information is saved and restored, and, hence, the order of events might matter.

When a server receives a *vote request* RPC from a *Candidate*, it votes for them if it has not yet voted for any other node in that term previously. Additionally, to prevent a *Candidate* with an out-of-date log from becoming *Leader*, the node compares the index and term of the last entries in the logs of the voter and the *Candidate*. The Raft algorithm uses an ingenious scheme—taking the type of RPC, the current term of the node and the message and the log of the *Candidate* into account—to decide whether the vote is granted to the *Candidate* or not; it then informs the *Candidate* of its decision. On receiving a reply from the node, the *Candidate* evaluates the number of votes it has received. If it successfully acquires votes from a majority of the nodes in the cluster, it becomes a *Leader* by changing to state `Leader`. While waiting for votes, if the *Candidate* receives a valid heartbeat from a *Leader*, it steps down to become a *Follower*. In case of a split vote, the *Candidate* can timeout again and start a new election.

**Log Replication.** When a *Leader* receives a request from the client, the request is appended to the log and all other nodes are informed using *append entries* request RPCs. In our model, this is achieved in a way that is similar to how *Candidate* nodes deal with *vote request* RPCs. **A *Leader* sends only one log entry at a time; this is in line with the TLA+ specification, although the Raft algorithm supports sending multiple log entries at once.**

In an *append entries* request RPC, the *Leader* includes the index and term of the log entry immediately preceding the new entries. If a *Follower* does not find a matching entry in its log with the same index and term, then it refuses the entry and sends back a negative response. The *Leader*, upon receiving a negative response, decrements the *Follower*'s `nextIndex`, which is a list where each index corresponds to the same `serverID` and which stores the index of the log entry the *Leader* will send to that node. The *Leader*, when first elected, initialises all `nextIndex` values to the index just after the last one in its log. After decrementing the `nextIndex`, the *append entries* RPC is retried. Eventually `nextIndex` will



reach the point where the *Leader* and *Follower* logs match. When this happens, any subsequent conflicting entries in the *Follower*'s log are removed and entries from the *Leader*'s log are appended (if any). Consequently, a positive response is sent back to the *Leader* and log replication is successful.

Once the leader has successfully replicated a log entry on majority of the servers, the entry is deemed committed. We use the action `advanceCommitIndex` to model this. Moreover, we have introduced a function `MaxAgreeIndex` to find the highest possible index that can be committed. Once an entry has been committed, the *Leader* applies it to its state machine. The *Leader* keeps track of the highest index it knows to be committed, in parameter `commitIndex`, and includes this in *append entries* RPCs (heartbeat messages included) so other nodes can commit the entries, too. This method of counting successful replication on a majority is not used to commit entries from previous terms: only log entries from the *Leader*'s current term are committed by counting replicas. Once an entry from the current term has been committed in this way, then all prior entries are committed indirectly. The function `isAdvanceCommitIndexOk` is used to keep this in check.

**Model Statistics.** We have generated state spaces of various instances of the model as described (see also the Mars repository for the full model). The base case is a configuration in which there are 3 nodes, 2 commands from clients, 1 term, a network capacity of 3 messages, and no crashes and recovery of nodes. This basic configuration already leads to a rather large state space of over 200k states, which can be generated in slightly under a minute on a 2017 Macbook Pro. We do note that there is some redundancy in the model, since strong bisimilarity reduction manages to compress the state space with almost a factor of 5. Table 1 shows the statistics of all configurations we explored, including a configuration that shows the effect of using a lossy network. To give a rough indication of the time required to generate these state spaces, we have included the time it takes for a 2017 Macbook Pro with 16GB memory to generate these state spaces. This clearly shows the dramatic effect of nodes crashing and of increasing the number of possible terms.

#Nodes	#Commands	#Terms	#Network Capacity	Lossy Network	Crashing	Size	Time
3	2	1	3	no	no	$2.14 \cdot 10^5$	~ 1 min.
3	1	2	3	no	no	$1.17 \cdot 10^6$	~ 2 min.
3	1	3	3	no	no	$1.32 \cdot 10^7$	~ 13 min.
3	2	1	3	no	yes	$1.79 \cdot 10^7$	~ 5 min.
3	2	2	3	no	no	$2.25 \cdot 10^7$	~ 19 min.
3	2	1	3	yes	yes	$5.97 \cdot 10^7$	~ 105 min.
3	1	2	3	no	yes	$1.48 \cdot 10^8$	~ 24 min.
3	1	3	3	no	yes	$2.38 \cdot 10^9$	~ 820 min. <sup>†</sup>

Table 1: Some statistics for the state space sizes for various configurations of the Raft algorithm; (†) this experiment was conducted on a compute server so the runtime is not directly comparable to the other runtimes reported.

### 3 Raft Properties

The Raft algorithm is quite involved, and it is easy to **make** small mistakes when formalising the algorithm. A simple example of the subtleties include the aforementioned difference between the zero-indexed lists of mCRL2 and the one-indexed arrays used in the original description of the algorithm. As



part of the original description of the algorithm, the authors also list several properties that the algorithm guarantees; such properties can be seen as partial specifications of the algorithm. We have taken some of these properties and formalised them as modal  $\mu$ -calculus formulas. These formulas have been used throughout the model development to hunt for bugs in our formalisation, and provide an extra layer of validation in addition to manual simulation of the model, increasing our confidence in the model.

One central complication in formalising the original properties in the modal  $\mu$ -calculus is the fact that the properties refer to the variables that span the state of each node; in our case, those are, for instance, the parameters of the `Node` process. Since the mCRL2 language is action-based rather than state-based, these parameters cannot be referred to in the modal  $\mu$ -calculus formulas. We have sidestepped this issue by extending our model with auxiliary actions that expose the relevant information. For instance, for the purpose of verification, we have introduced self-loops labelled by actions such as `exposeLeader`, `exposeLogLeader` and `exposeLog`.

In what follows, we will briefly discuss four main properties that have been formalised and their modal  $\mu$ -calculus formalisation next to it. **The *Append Entries* property that is also mentioned in [12, 11] is omitted since it follows immediately from the operations a *Leader* can carry out in our model.** All formulas happen to fall in a category of formulas that can be represented in a PDL-style language, called *regular formulae*, that abstracts from the fixed points one typically expects in modal  $\mu$ -calculus formulas. We will explain the meaning of the formulas as we proceed. All main properties hold true for all configurations of Table 1; verification of each property takes roughly 2-3 times that of generating the state space using the symbolic technique described in [10]. We additionally verify a number of simple liveness properties that demonstrate that the main properties we verify do not hold true vacuously. Also these can be expressed in terms of regular formulae, save one.

**Election Safety.** One of the fundamental properties on which the Raft algorithm relies for its correct functioning is the property that at most one *Leader* can be elected in a term. This is called the *Election Safety* property in [12, 11]. **Since the state of each node cannot be read directly, we use the `exposeLeader` action to expose that the state of a node is *Leader*.** Then, the correctness property can be phrased as the inability for two distinct nodes to execute `exposeLeader` actions in the same term. The formula we use to express this is as follows:

```
[true*] forall id1, termx: Nat . [exposeLeader(id1, termx)]
[true*] forall id2: Nat . val(id1!=id2) => [exposeLeader(id2, termx)] false
```

This formula should be read as follows. Invariantly (captured by the first occurrence of `[true*]`), executing any `exposeLeader` action, carrying some ID (here represented by `id1`) of a node and a term, will lead to a state from which invariantly (captured by the second `[true*]` formula) it is impossible to execute another `exposeLeader` action carrying an ID (represented by `id2`) of a different node and the same term. The latter part is captured by the `[exposeLeader(id2, termx)] false` subformula.

Note that this property can hold true **trivially** if no *Leader* is ever elected. To exclude this scenario, we have additionally phrased a liveness property that verifies that such actions can indeed take place:

```
<true*> exists id1, termx:Nat . <exposeLeader(id1, termx)>true
```

This property, which should be read as follows: following zero or more actions, it is possible that an `exposeLeader` action can be executed, carrying some ID (represented by `id1`) and some term (represented by `termx`). Also this property holds true for all the configurations of Table 1. To assess whether there is a scenario in which two different nodes can be a leader, we can check the following formula:

```
<true*> exists id1, termx:Nat . <exposeLeader(id1, termx)>
<true*> exists id2, termy:Nat. val(id1 != id2) && <exposeLeader(id2, termy)>true
```

This formula only holds true in configurations in which there are at least two terms, as can be expected. Taking this into account, we can verify a formula that states that there is a sequence of events in which we see `MaxTerm` times a different leader announce itself. Such a formula requires us to explicitly use a least fixed point and keep track of the number of times we have witnessed the `exposeLeader` action and the ID of the leader that announced itself most recently.

```
mu X(id:Nat = 1, n:Nat = 0).
  ( val (n >= MaxTerm) || <true>X(id,n)
    || exists id2, termy:Nat. (val (id != id2) && <exposeLeader(id2,termy)>X(id2,n+1))
  )
```

This formula holds true in all configurations.

**Log Matching.** The log replication mechanism ensures that each node in the Raft cluster has the same view on the state of the cluster. In particular, if, for two nodes, their logs contain an entry with the same index and term, then these logs are identical in all entries up to (and including) the given index. This is called the *Log Matching* property in [12, 11]. Log information of nodes, which is tracked in the `log` parameter of the `Node` process cannot be inspected using the modal  $\mu$ -calculus formula without exposing the information through actions. This is achieved by extending the model with self loops of `exposeLog` actions; the property can then be formalised as follows:

```
[true*] forall id1, term1, commitIndex1: Nat, log1: LogType .
  val (log1 != []) => [exposeLog(id1, term1, commitIndex1, log1)]
  [true*] forall id2, term2, index, commitIndex2: Nat, log2: LogType .
    val (index <#log1 && index <#log2 && id1 != id2 && log2 != [] && log1.index == log2.index)
    =>
      ([exposeLog(id2, term2, commitIndex2, log2)]
        val (slice(log1, 1, index+1) == slice(log2, 1, index+1)))
```

Again, the `[true*]` should be read as ‘invariantly’. The formula can then be understood to state the following: invariantly, for any state in which there is a node (the ID of which is `id1`) with a non-empty log `log1` it is the case that invariantly from that moment onwards, any other state in which there is another node (the ID of which is `id2`) with a non-empty log `log2` that has an entry at position `index` in common, the slices of `log1` and `log2` coincide up to, and including position `index`.

**Leader Completeness.** Another aspect of the log replication mechanism is that log entries, committed in a given term, will persist in the logs of the *Leaders* in future terms; in [12, 11] this is referred to as the *Leader Completeness* property. This ensures that the logs are indeed a proper reflection of what has happened in the past. We modify the model to include `exposeLogLeader` self-loops that expose the log information of the node that is currently in state `Leader`. The `advanceCommitIndex` actions, already present in the model, are used as signals that information has been committed in the log up to, and including entry `currentCommitIndex`. Note that in our model, the `advanceCommitIndex` action also exposes the log of the leader through the `log1` parameter.

```
[true*] forall currentCommitIndex, nextCommitIndex, term1: Nat, log1: LogType .
  [advanceCommitIndex(currentCommitIndex, nextCommitIndex, term1, log1)]
  [true*] forall term2, index: Nat, log2: LogType .
    val (term2 > term1 && index > currentCommitIndex && index <= nextCommitIndex)
    =>
      [exposeLogLeader(term2, log2)] val ((log1 . index) in log2)
```

This formula should be read as follows: invariantly, whenever a `advanceCommitIndex` action happens, exposing the current commit index `currentCommitIndex`, the next commit index `nextCommitIndex`, the

term `term1` and the *Leader*'s log `log1`, then whenever we subsequently inspect the log of a *Leader* in a future term `term2`, then those log entries in `log1` that can be found at indices beyond `currentCommitIndex` and `nextCommitIndex` are contained in the log entries of `log2`.

**State Machine Safety.** The logs that appear in each node furthermore must provide a uniform, consistent view on the state of the cluster. That means that after a node has applied a log entry at a given index to its state machine, no other node will ever apply a different log entry for the same index. This property is referred to as the *State Machine Safety* property in [12, 11]. In order to express this property, we again assume that the model has been extended with self-loops labelled with `exposeLog` actions. The property can then be formalised as follows:

```
[true*] forall id1, term1, commitIndex1: Nat, log1: LogType .
  val(commitIndex1 > 0)
=>
[exposeLog(id1, term1, commitIndex1, log1)]
[true*] forall id2, term2, commitIndex2: Nat, log2: LogType .
  val(id1!=id2 && commitIndex2>=commitIndex1)
=>
[exposeLog(id2, term2, commitIndex2, log2)]
  val(slice(log1, 1, commitIndex1) == slice(log2, 1, commitIndex1))
```

This formula can be understood as follows: invariantly, whatever the log of a node is, given the commit index `commitIndex1` of the node at that time, the log will overlap up-to and including this index in any future moment in which the commit index `commitIndex2` of a node is equal or larger. This ensures consistency of the logs over time, meaning that the same entries have been applied to the state machine. Note that the condition `commitIndex1 > 0` ensures that an entry has been committed.

In order to assess whether or not the property holds true vacuously, we have phrased the following simple liveness requirement:

```
<true*> exists id1, term1, commitIndex1: Nat, log1: LogType .
  val(commitIndex1 > 0) && <exposeLog(id1, term1, commitIndex1, log1)>
  <true*> exists id2, term2, commitIndex2: Nat, log2: LogType .
    val(id1 != id2 && commitIndex2 >= commitIndex1)
    &&
    <exposeLog(id2, term2, commitIndex2, log2)> true
```

This property holds true for every model we have analysed.

## 4 Discussion

We briefly touch on a few observations related to modelling in mCRL2, but also related to how our model compares to existing formalisations of the Raft algorithm.

**Modelling in mCRL2.** The mCRL2 language has all the features that allow one to concisely describe the workings of complex distributed algorithms. Parallelism and message passing, both key ingredients in the Raft algorithm, are key concepts that allow the model to stay close to reality. Furthermore, parameterisation of processes and actions allows for reusing parts of the specifications, avoiding copy-paste mistakes and improving readability. Finally, the rich and expressive data language of mCRL2 is essential when describing the more complex operations on arrays that are part of the Raft algorithm. Built-in types such as lists, natural numbers, and sets, and the facility to specify custom data types and operations on these turned out to be essential for keeping the specification readable and its size to a minimum.

Modelling in the mCRL2 language does require experience, and there is not really a practical guide-book that explains how to use the language effectively. This can lead to sub-optimal ways of modelling. For instance, the initial model, which was created by the first two authors and who had no prior experience of using mCRL2, used a modelling style that is perfectly valid but that led to state spaces that were orders of magnitude larger than needed. To illustrate, consider the following mCRL2 specification:

```
act a:Bool;
proc X(c:Bool) = sum b:Bool. a(b). ( (b -> X(true) + !b -> X(false)));
init X(true);
```

The transition system that is generated for this specification has 3 states and 6 transitions. Based on the specification, one would expect at most 2 states: one representing  $X(\text{true})$  and one representing  $X(\text{false})$ . The third state is introduced by the pre-processing of the mCRL2 toolset, which rewrites the above specification to normal form and which may introduce extra process parameters. Now consider the following mCRL2 specification:

```
act a:Bool;
proc X(c:Bool) = sum b:Bool. ( b -> a(b).X(true) + !b -> a(b).X(false));
init X(true);
```

The state space generated for this specification has only 1 state and 2 transitions. It is strongly bisimilar to the state space of the previous example, and therefore, for all intents and purposes, equivalent to it. In this case, the pre-processing conducted by the mCRL2 toolset does not lead to an additional state because the process is already in a shape it wishes to produce. Even better, during the pre-processing it detects that parameter  $c$  of process  $X$  is irrelevant, which can easily be seen because it does not appear in the right-hand side.

**Comparison to Other Formalisations.** As we have indicated earlier, in constructing the mCRL2 model for the Raft algorithm, we have drawn inspiration from both the TLA+ and the LNT specifications. However, there are cases where the TLA+ and LNT specifications make different modelling choices, and, consequently, we have had to make a choice between the two. **A case in point is the way the TLA+ specification deals with stale RPC messages: it drops stale responses but stale requests are replied to, to alert the sending party of the newer term. The LNT specification, on the other hand, discards stale requests as well.** In our model, we chose to here follow the LNT specification. An example where we followed the TLA+ specification is where we send the minimum between `commitIndex` of the current node and the `nextIndex` of the receiver when sending the *append entries* request, rather than LNT's choice to send the `commitIndex` of the *Leader*. There are other places where our model deviates subtly from the LNT model, for instance in dealing with crashed nodes. In particular, the LNT model does not appear to allow for nodes to reboot, and the inclusion of rebooting nodes has had some implications on how we dealt with sending replies to requests.

Concerning the TLA+, LNT and mCRL2 modelling languages, we remark that due to LNT having many traits of an imperative language, unlike mCRL2 and TLA+, the LNT specification is in all likelihood more appealing to the average software engineer than the TLA+ or mCRL2 specifications. Also, the ability to specify crashing of nodes using LNT's *disrupt* operator is rather elegant; in our mCRL2 model, this requires hard-coding the option to crash. **While in our model, the difference turns out to be minimal, this would not have been the case if our model had used large numbers of actions that could be executed sequentially.** For instance, specifying that a *crash* action can interrupt the process  $a.b.c$  would require a specification of the form  $\text{crash} + a. (\text{crash} + b. (\text{crash} + c)).$

## 5 Conclusions and Future Work

In this paper, we have highlighted and discussed several parts of our mCRL2 model of the Raft algorithm and the modal  $\mu$ -calculus formulas capturing its properties. The full details of the model and the formulas can be found in the Mars repository. While our model shares many aspects with the TLA+ and LNT specification that have been published before, the formalisation of some of the key properties of the algorithm using a modal logic appear to be new. Note that only the simplest configurations can be verified in reasonable time, but it may still be interesting to verify the more complex configurations as well, including non-perfect network behaviour. We consider this part of future work.

Furthermore, it would be interesting to verify stronger liveness requirements. We have only covered a few very basic, weak liveness properties, asserting that it is possible to, e.g., (repeatedly) become a leader. Stronger liveness requirements, asserting that always inevitably a leader *must* be elected, are simply not true in our model because in some configurations, messages are lost or nodes crash, but also due to us imposing limits on the maximum number of terms we consider. Phrasing the exact properties while taking all exceptions into account is non-trivial: for some properties, a counterexample may not simply be a run of the system but it can consist of an entire subgraph of the transition system [2, 3], consisting of a 1 000 or more states. In such cases, understanding the root cause of the violation can be virtually impossible. Proving liveness properties for the unrestricted model (*i.e.*, when not limiting the number of terms) can be even more challenging.

Furthermore, in the model, when a node receives a message, it computes the reply atomically. This simplifies the model but does not accurately reflect real-world scenarios where the computation of a reply would involve multiple steps and could be interrupted by other events. Refining these aspects would increase the applicability of the model to real-life scenarios but a careful tradeoff must be made between the level of abstraction and the granularity of the model to keep the state space from exploding.

Finally, like the TLA+ and LNT specifications, our model lacks real-time, even though the algorithm suggests typical timing intervals. For instance, Raft chooses election timeouts arbitrarily from a fixed interval (e.g., 150–300ms), whereas in our model a timeout can happen non-deterministically. While mCRL2 has facilities to model real-time aspects, the current status of the tooling is not sufficiently powerful to deal with real-time systems with state spaces of this size. A real-time extension of our model could therefore serve as a challenging benchmark for real-time model checking tools.

**Acknowledgements.** We would like to thank Myrthe Spronck (TU/e) for discussions on the problem of consensus and the solution provided by Paxos.

## References

- [1] Olav Bunte, Jan Friso Groote, Jeroen J. A. Keiren, Maurice Laveaux, Thomas Neele, Erik P. de Vink, Wieger Wesselink, Anton Wijs & Tim A. C. Willemse (2019): *The mCRL2 Toolset for Analysing Concurrent Systems - Improvements in Expressivity and Usability*. In Tomás Vojnar & Lijun Zhang, editors: *Tools and Algorithms for the Construction and Analysis of Systems - 25th International Conference, TACAS 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6-11, 2019, Proceedings, Part II, Lecture Notes in Computer Science 11428*, Springer, pp. 21–39, doi:10.1007/978-3-030-17465-1\_2.
- [2] Sjoerd Cranen, Bas Luttik & Tim A. C. Willemse (2013): *Proof Graphs for Parameterised Boolean Equation Systems*. In Pedro R. D’Argenio & Hernán C. Melgratti, editors: *CONCUR 2013 - Concurrency Theory - 24th International Conference, CONCUR 2013, Buenos Aires, Argentina, August 27-30, 2013. Proceedings*,

- Lecture Notes in Computer Science* 8052, Springer, pp. 470–484, doi:10.1007/978-3-642-40184-8\_33.
- [3] Sjoerd Cranen, Bas Luttik & Tim A. C. Willemse (2015): *Evidence for Fixpoint Logic*. In Stephan Kreutzer, editor: *24th EACSL Annual Conference on Computer Science Logic, CSL 2015, September 7-10, 2015, Berlin, Germany, LIPIcs* 41, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 78–93, doi:10.4230/LIPICS.CSL.2015.78.
  - [4] Hugues Evrard (2020): *Modeling the Raft Distributed Consensus Protocol in LNT*. In Ansgar Fehnker & Hubert Garavel, editors: *Proceedings of the 4th Workshop on Models for Formal Analysis of Real Systems, MARS@ETAPS 2020, Dublin, Ireland, April 26, 2020, EPTCS* 316, pp. 15–39, doi:10.4204/EPTCS.316.2.
  - [5] Hubert Garavel, Frédéric Lang & Wendelin Serwe (2017): *From LOTOS to LNT*. In Joost-Pieter Katoen, Rom Langerak & Arend Rensink, editors: *ModelEd, TestEd, TrustEd - Essays Dedicated to Ed Brinksma on the Occasion of His 60th Birthday, Lecture Notes in Computer Science* 10500, Springer, pp. 3–26, doi:10.1007/978-3-319-68270-9\_1.
  - [6] Jan Friso Groote & Mohammad Reza Mousavi (2014): *Modeling and Analysis of Communicating Systems*. MIT Press, doi:10.5555/2628007.
  - [7] Fabrice Kordon, Hubert Garavel, Lom-Messan Hillah, Emmanuel Paviot-Adet, Loïc Jezequel, César Rodríguez & Francis Hulin-Hubard (2016): *MCC'2015 - The Fifth Model Checking Contest*. *Trans. Petri Nets Other Model. Concurr.* 11, pp. 262–273, doi:10.1007/978-3-662-53401-4\_12.
  - [8] Leslie Lamport (1998): *The Part-Time Parliament*. *ACM Trans. Comput. Syst.* 16(2), pp. 133–169, doi:10.1145/279227.279229.
  - [9] Leslie Lamport (2002): *Specifying Systems, The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley, doi:10.5555/579617. Available at <http://research.microsoft.com/users/lamport/tla/book.html>.
  - [10] Maurice Laveaux, Wieger Wesselink & Tim A. C. Willemse (2022): *On-The-Fly Solving for Symbolic Parity Games*. In Dana Fisman & Grigore Rosu, editors: *Tools and Algorithms for the Construction and Analysis of Systems - 28th International Conference, TACAS 2022, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2022, Munich, Germany, April 2-7, 2022, Proceedings, Part II, Lecture Notes in Computer Science* 13244, Springer, pp. 137–155, doi:10.1007/978-3-030-99527-0\_8.
  - [11] Diego Ongaro (2014): *Consensus: bridging theory and practice*. Ph.D. thesis, Stanford University, USA, doi:10.5555/AAI28121474.
  - [12] Diego Ongaro & John K. Ousterhout (2014): *In Search of an Understandable Consensus Algorithm*. In Garth Gibson & Nickolai Zeldovich, editors: *2014 USENIX Annual Technical Conference, USENIX ATC '14, Philadelphia, PA, USA, June 19-20, 2014*, USENIX Association, pp. 305–319, doi:10.5555/2643634.2643666.
  - [13] Fred B. Schneider (1990): *Implementing Fault-Tolerant Services Using the State Machine Approach: A Tutorial*. *ACM Comput. Surv.* 22(4), pp. 299–319, doi:10.1145/98163.98167.
  - [14] Doug Woos, James R. Wilcox, Steve Anton, Zachary Tatlock, Michael D. Ernst & Thomas Anderson (2016): *Planning for change in a formal verification of the raft consensus protocol*. *CPP 2016 - Proceedings of the 5th ACM SIGPLAN Conference on Certified Programs and Proofs, co-located with POPL 2016*, pp. 154–165, doi:10.1145/2854065.2854081.