

Project 2: The Raft Consensus Algorithm

Please read all pages before writing code.

1 Overview

Important Dates:

Project release: **Monday, October 22, 2018 at 03:00pm**

Checkpoint due: **Monday, November 5, 2018 at 11:59pm**

Final Test due: **Monday, November 12, 2018 at 11:59pm**

Submission limits: **15 Autolab submissions per checkpoint**

What you will learn

- Real-world consensus protocols: replicated state machines (Lecture 12)
- How consensus fits into distributed applications (Lecture 14)
- Reasoning about the correctness of concurrent programs (Lecture 4)
- Implementing a protocol using its formal specification
- Writing tests to catch and debug **corner cases**

In this project, you'll implement the **Raft** consensus algorithm, a replicated state machine protocol. There will be one checkpoint for the implementation of your Raft. You will want to start early. For more information regarding what portion of the project is expected to be completed for the checkpoint and the final test, please refer to sections 4 and 5.

The starter code for this project is hosted as a read-only repository on **GitHub**. For instructions on how to build, run, test, and submit your server implementation, see the **README.md** file in the project's root directory. To clone a copy, execute the following **Git** command:

```
git clone https://github.com/CMU-440-F18/p2.git
```

You must work on this project **individually**. As per course policy, you may turn in the project up to **two** days late, and each day will incur a 10% deduction from your score (after two days, the maximum possible score is 80%). We will not accept your code after two days.

2 Raft

A replicated service (e.g., key/value database) achieves fault tolerance by storing copies of its data on multiple replica servers. Replication allows the service to continue operating even if some of its servers experience failures (crashes or a broken or flaky network). The challenge is that failures may cause the replicas to hold differing copies of the data.

Raft manages a service's state replicas and, in particular, it helps the service sort out what the correct state is after failures. Raft implements a replicated state machine. It organizes client requests into a sequence, called the log, and ensures that all the replicas agree on the contents and ordering of the entries in the log. Each replica asynchronously executes the client requests in the order they appear in the log, applying those requests to the replica's local copy of the service's state. Since all the live replicas see the same log contents, they all execute the same requests in the same order and thus continue to have identical service state. If a server fails but later recovers, Raft takes care of bringing the log of the recovered server up to date. Raft will continue to operate as long as at least a majority of the servers are alive and can talk to each other. If there is no such majority, Raft will make no progress but will pick up where it left off as soon as a majority can communicate again.

In this project, you'll implement Raft as a Go object type with associated methods, meant to be used as a module in a more substantial service. A set of Raft instances talk to each other with RPCs to maintain replicated logs. Your Raft interface will support an indefinite sequence of numbered commands, also called log entries. The entries are numbered with index numbers. The log entry with a given index will eventually be committed. At that point, your Raft should send the log entry to the more substantial service for it to execute.

Note: Only RPCs may be used for interaction between different Raft instances. For example, different peers in your Raft implementation are not allowed to share Go variables. Your implementation should not use files at all.

You'll implement a part of the Raft design described in the extended paper. You will not implement persistence, cluster membership changes (Section 6) or log compaction / snapshotting (Section 7).

You should consult the [extended Raft paper](#). You may find it useful to look at this [illustrated](#) guide to Raft. For a broader perspective, have a look at Paxos, Chubby, Paxos Made Live, Spanner, Zookeeper, Harp, Viewstamped Replication, and [Bolosky et al.](#)

3 Project Requirements

As you write code for this project, also keep in mind the following requirements:

- You must work on this project individually. You are free to discuss high-level design issues with other people in the class, but every aspect of your implementation must be entirely your work.
- You must format your code using `go fmt` and must follow Go's standard naming conventions. See the [Formatting](#) and [Names](#) sections of Effective Go for details.
- You may use any of the synchronization primitives in Go's `sync` package for this project.

4 Tips

Tip 1: Start early. Although the amount of code is not much (our reference solution is ≈ 700 lines), getting it to work correctly will be challenging. Both the algorithm and the code is tricky, and there are many corner cases to consider. When one of the tests fails, it may take a bit of puzzling to understand in what scenario your solution isn't correct, and how to fix your solution.

Tip 2: Read and understand the [extended Raft paper](#) before you start. Your implementation should follow the paper's description closely, particularly [Figure 2](#), since that's what the tests expect. You'll find a copy of that figure on the next page.

Tip 3: Your first implementation may not be clean enough that you can easily reason about its correctness. Give yourself enough time to rewrite your implementation so that you can easily reason about its correctness.

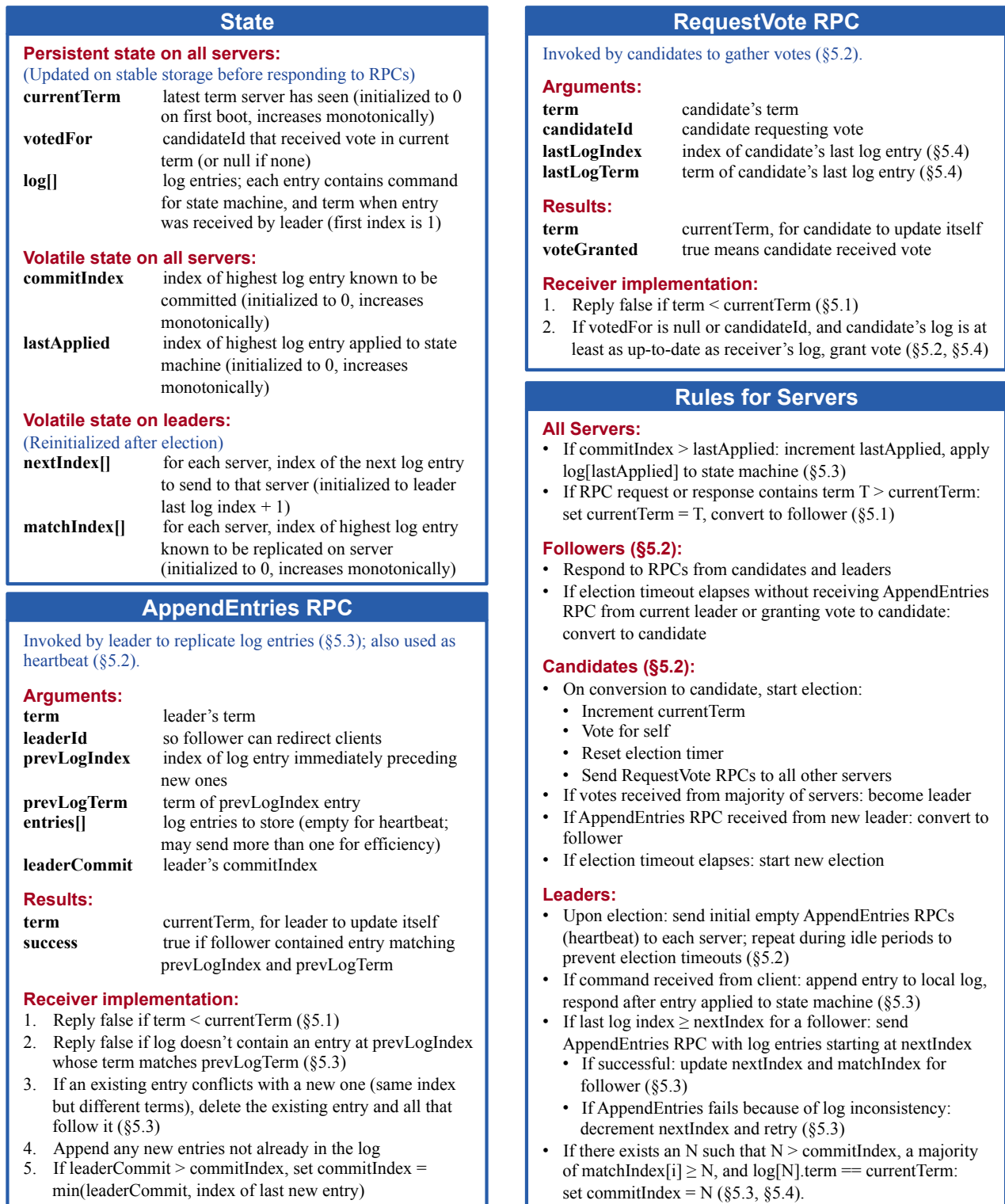


Figure 2: A condensed summary of the Raft consensus algorithm (excluding membership changes and log compaction). The server behavior in the upper-left box is described as a set of rules that trigger independently and repeatedly. Section numbers such as §5.2 indicate where particular features are discussed. A formal specification [31] describes the algorithm more precisely.

5 The code

Implement Raft by adding code to `raft/raft.go`. In that file, you'll find a bit of skeleton code and examples of how to send and receive RPCs.

Your implementation must support the following interface, which the tester will use. You'll find more details in comments in `raft.go`.

```
// rf = Make(...)
//   Create a new Raft server.
//
// rf.Start(command interface{}) (index, term, isleader)
//   Start agreement on a new log entry
//
// rf.GetState() (me, term, isLeader)
//   Ask a Raft peer for "me" (see line 58 in raft.go),
//   its current term, and whether it thinks it is a leader
//
// ApplyMsg
//   Each time a new entry is committed to the log, each Raft peer
//   should send an ApplyMsg to the service (e.g. tester) on the
//   same server, via the applyCh channel passed to Make()
```

A service calls `Make(peers, me, ...)` to create a Raft peer. The `peers` argument is an array of established RPC connections, one to each Raft peer (including this one). The `me` argument is the index of this peer in the `peers` array. `Start(command)` asks Raft to start the processing to append the command to the replicated log. `Start()` should return immediately, without waiting for this process to complete. The service expects your implementation to send an `ApplyMsg` for each new committed log entry to the `applyCh` argument to `Make()`.

Your Raft peers should exchange RPCs using the `rpc` Go package that we provide to you. It is modeled after Go's `rpc library`, but internally uses Go channels rather than sockets. `raft.go` contains some example code that sends an RPC (`sendRequestVote()`) and that handles an incoming RPC (`RequestVote()`). The reason you must use `rpc` instead of Go's RPC package is that the tester tells `rpc` to delay RPCs, re-order them, and delete them to simulate challenging network conditions under which your code should work correctly. Don't modify `rpc` because we will test your code with the `rpc` as handed out. **Note:** The `rpc` package only provides a subset of the functionality of Go's RPC system. For instance, asynchronous RPC calls are not provided by `rpc`.

6 Checkpoint

6.1 Task

Implement **leader election** and **heartbeats** (**AppendEntries** RPCs with no log entries). The goal for the checkpoint is for a single leader to be elected, for the leader to remain the leader if there are no failures, and for a new leader to take over if the old leader fails or if packets to/from the old leader are lost. Run `go test -race -run 2A` to test your checkpoint code.

Be sure you pass the checkpoint tests before submitting. Note that the checkpoint tests test the basic operation of leader election. The final tests will test leader election in more challenging settings and may expose bugs in your leader election code which the checkpoint tests miss.

6.2 General Guidelines

Add any state you need to the **Raft** struct in `raft.go`. You'll also need to define a struct to hold information about each log entry. Your code should follow Figure 2 in the paper as closely as possible.

Fill in the **RequestVoteArgs** and **RequestVoteReply** structs. Modify **Make()** to create a background goroutine that will kick off leader election periodically by sending out **RequestVote** RPCs when it hasn't heard from another peer for a while. This way a peer will learn who the leader is, if there is already a leader, or become the leader itself. Implement the **RequestVote()** RPC handler so that servers will vote for one another.

To implement heartbeats, define a **AppendEntries** RPC struct (though you may not need all the arguments yet), and have the leader send them out periodically. Write a **AppendEntries** RPC handler method that resets the election timeout so that other servers don't step forward as leaders when one has already been elected.

Make sure the election timeouts in different peers don't always fire at the same time, or else all peers will vote only for themselves, and no one will become the leader.

The tester requires that the leader send heartbeat RPCs no more than ten times per second.

The tester requires your Raft to elect a new leader within five seconds of the failure of the old leader (if a majority of peers can still communicate). Remember, however, that leader election may require multiple rounds in case of a split vote (which can happen if packets are lost or if candidates unluckily choose the same random backoff times). You must pick election timeouts (and thus heartbeat intervals) that are short enough that it's very likely

that an election will complete in less than five seconds even if it requires multiple rounds.

The paper's Section 5.2 mentions election timeouts in the range of 150 to 300 milliseconds. Such a range only makes sense if the leader sends heartbeats considerably more often than once per 150 milliseconds. Because the tester limits you to 10 heartbeats per second, you will have to use an election timeout larger than the paper's 150 to 300 milliseconds, but not too large, because then you may fail to elect a leader within five seconds.

6.3 Notes and Hints

There are some details and hints we want to emphasize here to help you pass our tests:

- Go RPC sends only those struct fields whose names start with capital letters. Substructures must also have capitalized field names (e.g., fields of log records in an array). Forgetting to capitalize field names sent by RPCs is the single most frequent source of bugs while using RPCs.
- You may find Go's `time.Sleep()` and `rand` useful.
- You'll need to write code that takes actions periodically or after delays in time. The easiest way to do this is to create a goroutine with a loop that calls `time.Sleep()`.
- If your code has trouble passing the tests, read the paper's Figure 2 again; the full logic for leader election is spread over multiple parts of the figure.
- A good way to debug your code is to insert print statements when a peer sends or receives a message, and collect the output in a file with `go test -race -run 2A > out`. Then, by studying the trace of messages in the `out` file, you can identify where your implementation deviates from the desired protocol.
- You should check your code with `go test -race`, and fix any races it reports.

7 Final Test

We want Raft to keep a consistent, replicated log of operations. A call to `Start()` at the leader starts the process of adding a new operation to the log; the leader sends the new operation to the other servers in `AppendEntries` RPCs.

7.1 Task

Implement the leader and follower code to append new log entries. This will involve implementing `Start()`, completing the `AppendEntries` RPC structs, sending them, fleshing out the `AppendEntry` RPC handler, and advancing the `commitIndex` at the leader. Your first goal should be to pass the `TestBasicAgree()` test (in `test.test.go`). Once you have that working, you should get all the final tests to pass (`go test -race -run 2B`).

7.2 General Guidelines

While the Raft leader is the only server that initiates appends of new entries to the log, all the servers need to independently give each newly committed entry to their local service replica (via their `applyCh`). You should try to keep the goroutines that implement the Raft protocol as separate as possible from the code that sends committed log entries on the `applyCh` (e.g., by using a separate goroutine for delivering committed messages). If you don't separate these activities cleanly, then it is easy to create deadlocks. Without a clean separation, a common deadlock scenario is as follows: an RPC handler sends on the `applyCh`, but it blocks because no goroutine is reading from the channel (e.g., perhaps because it called `Start()`). Now, the RPC handler is blocked while holding the mutex on the Raft structure. The reading goroutine is also blocked on the mutex because `Start()` needs to acquire it. Furthermore, no other RPC handler that needs the lock on the Raft structure can run.

You will need to implement the election restriction (section 5.4.1 in the paper).

8 Hand In

Create a tarball of the `raft/` folder with the `raft.go` file in it.

Use the following command:

```
tar cvf raft.tar raft/
```

9 Testing and Grading

We will use **Autolab** to automatically grade your implementation for correctness, and manual grading for style. **Please make sure that there are no unnecessary files (binary or other kinds) in your Autolab submission archive.**

In addition to the tests that are provided in `raft_test.go`, we will run additional, more

extensive tests on Autolab for both your checkpoint and the final submissions. We will not be able to provide you any details (except for the `stdout` you'll see on Autolab) about any of these new tests. We will run each test multiple times – you should pass every invocation of a test to pass that test.

You are free to write new tests on your own. You are, however, not required to do so and will not be graded on any new tests you write. Here are some tips to do this. Refer to `raft_test.go` as you read these tips.

- See `TestInitialElection2A`: you can use `cfg.checkOneLeader()` to check for a leader's election and to get the current leader's ID.
- See `TestFailAgree2B`: you can use `cfg.one(value, num_servers)` to start an agreement.
- See `TestFailNoAgree2B`: you can use `cfg.disconnect(server_id)` and `cfg.connect(server_id)` to disconnect and connect servers. You can also directly call `Start()` on one of the Raft peers by using `cfgrafts`.