

Learning TLA+ by Examples

Richard Tang

January 27, 2025

Contents

1	Introduction	4
1.1	Catching Problems Early	4
1.2	The Generalized Problem	5
1.3	TLA+	6
1.4	About This Book	6
1.5	Book Layout	7
2	TLA+ Primer	8
2.1	Design Intent	8
2.2	Requirement	9
2.3	Spec	9
2.4	Safety	10
2.5	Liveness	10
2.6	Model Checker	11
I	Examples	13
3	Simple Gossip Protocol	14
3.1	Requirement	14
3.2	Spec	15
3.3	Safety	17
3.4	Liveness	17
4	Simple Scheduler	18
4.1	Requirement	18
4.2	Spec	18
4.3	Safety	20
4.4	Liveness	20
5	Selective Retransmit	22
5.1	Requirement	24
5.2	Spec	25

6	Raft Consensus Protocol	28
6.1	Spec	29
6.2	Simplify Model	33
6.2.1	Modeling Messages as a Set	33
6.2.2	Limit Term Divergence	33
6.2.3	Normalize Cluster Term	34
6.2.4	Sending Request as a Batch	35
6.2.5	Prune Messages with Stale Terms	35
6.2.6	Enable Symmetry	36
6.3	Safety	36
6.4	Liveness	36
6.5	Requirement	38
6.6	Spec	38
II	Examples with PlusCal	39
7	SPSC Lockless Queue	40
7.1	Requirement	40
7.2	Spec	41
7.3	Safety	42
7.4	Liveness	42
7.5	Configuration	42
8	SPMC Lockless Queue	43
III	Language Reference	44
9	Data Structure	45
9.1	Set	45
9.2	Tuple	45
10	Idiom	47
11	Fairness and Liveness	48
11.1	Liveness	48
11.2	Weak Fairness	49
11.3	Strong Fairness	50
12	Abstraction Guideline	52
13	Reference	53
14	Nano	55
14.1	Requirement	55
14.2	Spec	55

Contents	3
----------	---

15 Debugging	57
---------------------	-----------

Chapter 1

Introduction

1.1 Catching Problems Early

Years ago, I worked on a proprietary low power processor in an embedded system. The processor ran microcode featuring a custom instruction set. To enter a low power state, a set of (possibly hundreds) instructions were executed. These instructions progressively puts the system in lower power state. For example: turn off IP A, then turn off IP B, then turn off the power island to the IPs. To save cost and power, the low power processor had very limited debuggability support.

An experienced reader may start to notice some redflags.

If the microcode attempts to access the memory interface when the power island has been shut off, processor would hang. Since the power island has been shut off, the physical hardware debug port is also unavailable, leaving the developer with *no way* of live debugging related problem. At this point the developer needs to siphon through (possibly hundreds) of instructions to catch invariant violation *manually*.

As one can imagine, maintaining the microcode was very expensive. Fortunately, the proprietary low power processor only had a handful of instructions, I created an emulator for this proprietary processor to verify the microcode prior to deploying it on target. The emulator models the processor states as a state graph, with executed instruction transitions the state machine to the next state. At every state all the invariants are evaluated to ensure none have been violated. Some of the invariants included:

- Accessing memory interface after power off leads to a hang
- Accessing certain register in certain chip revision leads to a hang
- Verify IPs are shut off in the allowed order

The verification algorithm was implemented using a *depth-first search* algorithm, providing *100%* microcode coverage before deploy on target.

To generalize, we can model arbitrary system as a set of states with a collection of invariant that must be upheld at all times. The complexity of the such arbitrarily system generally grows quadratically as the number of states grow linearly (eg. in a N state system, adding state $N+1$ may introduce N transitions into the new state). There are many engineering problems that exhibit a large number of states, such as lockless or waitfree data structure, distributed algorithms, OS scheduler, and more.

So, how do we build a solution that is correct by design?

1.2 The Generalized Problem

Fast forward to now: I stumbled across TLA+, a formalized solution of what I was looking for.

Leslie Lamport invented the TLA+ 1999, but TLA+ didn't appear to have caught on until the 2010's. My personal opinion is TLA+ was invented ahead of its time, and the problem complexity finally caught up in the past decade or so to allow TLA+ to visibly demonstrate its strength.

We are also at a point in the technology curve where vertical scaling is no longer practical, with CPU speed plateau'd in the past decade or so. The industry is exploring horizontal scaling solution, such as hardware vendor focusing adding more CPU cores, or software vendors buying many low end hardware instead of a few high end hardware. This shifts the technology complexity from hardware to software, demanding software solution to maximize concurrent hardware resource utilization.

One slight problem: *The cognitive load of of a person is between 5 to 9 items at most.*

Humans are good at high level reasoning, but not so good at keeping track of many things happening at the same time. It is hard to enumerate all possible scenario in one's mind to ensure the design accommodates all the edge cases.

Consider a distributed system. The system is a cluster of independently operating entities and need to somehow collectively offer the correct system behaviour, while any one of the machines may receive instructions out of order, crash, recover, etc.

Consider a single producer multiple consumer lockless queue. The consumers may reserve an index in the queue in certain order, but may release them in

different order. What if one reader is really slow, and another reader is super fast and possibly lapse the slow reader?

Consider an OS scheduler with locks. Assume all the processes have the same priority. Can a process starve the other processes by repeatedly acquire and release the lock? How do we ensure scheduling is fair?

The usual *anti-pattern* is to keep bandaiding the design until bugs stop comming. This is never ideal. Per Murphy's law, anything that can go wrong will go wrong, and a hard to reproduce bug will come in at the most inconvenient time. How do we make sure the solution is actually *correct by design*? To solve this problem, we must rely on tools to do the reasoning *for us*.

1.3 TLA+

TLA+ is a *system specification language*, with the intent to describe the system with implementation details removed. TLA+ allows designer to describe the system as a sequence of states. The designer can expresses transition condition from one state to another, describe invariants that must hold true in every state and liveness properties that the overall system should converge to. The key innovation of TLA+ is once the system is modeled as a finite state machine, the states can be *exhaustively* explored (via breath-first-search) to ensure certain properties are held through out the entire state space (either per state or a sequence of states).

1.4 About This Book

This book was initially a set of notes I took as I was learning TLA+. To my surprise, there's not as much material on TLA+ as I had assumed for such a critical tool in the designer toolbox. I decided to formalize the notes into this short book, which I hope the readers find this helpful in their journey to master TLA+.

The intent for the book is to teach reader how to write TLA+ spec for their design to provide confidence in *design correctness*. The content to this book is appropriate for software designer, hardware designer, system architect, and such.

As for the readers: some computing science knowledge is required. One doesn't need to be expert at a particular language to understand this book; TLA+ is effectively its own language. This book is example driven and will go through designs such as lockless queue, simple task scheduler, consensus algorithm, etc. Reader will likely enjoy a deeper insight if she has some familiarity

with these topics.

1.5 Book Layout

This book was motivated by the intent to solve problems. The book is designed to be example heavy with many chapter each representing a problem that can be modelled using TLA+.

Examples are split into two categories: A set of examples written using native TLA+ syntax, and another set of examples written using PlusCal (C-like syntax). I believe they are useful under different use cases. The differences will be highlighted in their respective sections. All examples will follow a similar layout, covering the expected design process (eg. requirement, spec, safety and liveness properties).

Finally, there will be a part that language reference portion that discusses a few topics deserving extra attention. The intent is to use this section of the book as a *reference*.

Chapter 2

TLA+ Primer

2.1 Design Intent

The key insight into TLA+ is modelling a system as a state machine. A simple digital clock can be represented by two variables, hour and minute and the number of possible states in a digital clock is $24 * 60 = 1440$. For example, 10:01 is the next state 10:00 can transition to. Extrapolating further, Assume an arbitrarily system described by N variables, each variable having K possible values such arbitrary system can have up to N^K state.

For every specification, designer can specify *safety* property (or invariants) that must be true in *every* states. For example, in any state of the digital clock hour *must* be between 0 to 23, or formally described as $hour \in 0..23$. Similarly, minute must have value between 0 to 59, or $minute \in 0..59$. Examples invariants of a system include: Only one thread has exclusive access to a critical region, all variables in the system are within allowable value, resource allocation manager never allocates more than available resources.

Designer can also specify *liveness* property. These are properties to be satisfied by a *sequence of state*. One liveness property for the digital clock could be when the clock is 10 : 00, it will eventually become 11 : 00 (10 : 00 *leads to* 11 : 00). Example liveness property include: a distributed system eventually converges, the scheduler eventually schedules every tasks in the task queue, the resource allocation manager fairly allocates resources.

A TLA+ Spec can be checked by TLC, the model checker. TLC uses *breadth-first search* algorithm to explore *all* states in the state machine and ensure safety and liveness properties are upheld.

A TLA+ Spec describes the system using *temporal logic*. The syntax may appear unfamiliar if one hasn't seen it before, but like any other programming

language an initiated reader should become familiarized quickly. In this book I will use

2.2 Requirement

In this example, we will specify a *digital clock*. The digital clock has a few simple requirements:

- Two variables to represent state: hour and minute
- The clock increment one minute at a time
- The clock wraps around at midnight (ie. 23:59 transitions to 00:00)

2.3 Spec

The *Init* state of such system can be described as:

$$\begin{aligned} Init &\triangleq \\ &\wedge hour = 0 \\ &\wedge minute = 0 \end{aligned}$$

\triangleq is the *defines equal* symbol and \wedge is the *logical and* symbol. The above TLA+ syntax can be read as *Init* state is defined as both hour and minute are both 0.

The spec also always include a *Next* definition, an *action formula* describing how the system transition from one state to another. Action formula contains *primed* variables what happens to the variable in its next state. The *Next* action for the digital clock can be defined as:

$$\begin{aligned} NextHour &\triangleq \\ &\wedge minute = 59 \\ &\wedge hour' = (hour + 1) \% 24 \\ &\wedge minute' = 0 \\ NextMinute &\triangleq \\ &\wedge minute \neq 59 \\ &\wedge hour' = hour \\ &\wedge minute' = minute + 1 \\ Next &\triangleq \\ &\vee NextMinute \\ &\vee NextHour \end{aligned}$$

Here's a breakdown of what the spec does:

- *Next* can take *NextMinute* or *NextHour*

- *Next* takes *NextMinute* when *minute* is not 59, next hour is hour, next minute is minute + 1.
- *Next* takes *NextHour* when *minute* is 59, next hour is (hour + 1) modulus 24, next minute set to 0

Technically it's possible for *Next* to take both *NextMinute* and *NextHour*. This is not possible in this definition as *NextHour* and *NextMinute* are defined in a *mutually exclusively* fashion.

Finally, the spec itself is formally defined as:

$$\begin{aligned} vars &\triangleq \langle hour, minute \rangle \\ Spec &\triangleq \\ &\quad \wedge Init \\ &\quad \wedge \Box [Next]_{vars} \end{aligned}$$

$\Box [Next]_{vars}$ deserves some special attention:

- *vars* is defined to be *all* variables in the spec. Different combination of these variables constitute the states of the system (eg. 23:59 and 00:00 are both states in the system).
- $\Box [Next]_{vars}$ is a *box-action formula*, where *Next* is an action and *vars* is a state function.
- \Box operator asserts the formula is always true for every step in the behaviour.
- And steps in the behaviour is defined as $[Next]_{vars}$, where *Next* describe the action and *vars* capturing all variables representing the state.

2.4 Safety

Safety property describes invariant that must hold true in every state of system. A common invariant is *type safety* checks. In a digital clock, hour can only be in value between 0 to 23, and minute can only be value of 0 to 59:

$$\begin{aligned} Type_OK &\triangleq \\ &\quad \wedge hour \in 0 \dots 23 \\ &\quad \wedge minute \in 0 \dots 59 \end{aligned}$$

2.5 Liveness

Liveness property verifies certain behavioural across a sequence of state. One liveness property can be confirming the clock wraps around correctly at midnight (which involves multiple states):

$$\begin{aligned} \text{Liveness} &\triangleq \\ &\wedge \text{hour} = 23 \wedge \text{minute} = 59 \leadsto \text{hour} = 0 \wedge \text{minute} = 0 \end{aligned}$$

\leadsto is the *leads to* operator, suggesting something is eventually true. TLA+ provides a set of formulas that can be used to describe liveness property.

To verify liveness, we need to modify the spec slightly to enable *fairness* to prevent *stuttering*. In plain terms, fairness ensure *something* always happen in every step, allowing the states to transition. Without fairness the spec is allowed to *do nothing* as next step, this means liveness condition may fail because the spec permits the system to do nothing in perpetuity as next state. Fairness will be covered in more detailed in later chapter.

$$\begin{aligned} \text{Spec} &\triangleq \\ &\wedge \text{Init} \\ &\wedge \Box[\text{Next}]_{\text{vars}} \\ &\wedge \text{WF}_{\text{vars}}(\text{Next}) \end{aligned}$$

$\text{WF}_{\text{vars}}(\text{Next})$ is the fairness qualifier.

2.6 Model Checker

The TLA+ spec can be verified using TLC model checker. The TLC model checker runs the spec and verifies all configured safety and liveness properties are satisfied during execution. To run TLC, we need two things:

- clock.tla - the spec itself
- clock.cfg - the corresponding configuration file

For reference, clock.tla spec is listed below:

<pre> EXTENDS <i>Naturals</i> VARIABLES <i>hour, minute</i> <i>vars</i> \triangleq $\langle \text{hour}, \text{minute} \rangle$ <i>Type_OK</i> \triangleq $\wedge \text{hour} \in 0 \dots 23$ $\wedge \text{minute} \in 0 \dots 59$ <i>Liveness</i> \triangleq $\wedge \text{hour} = 23 \wedge \text{minute} = 59 \leadsto \text{hour} = 0 \wedge \text{minute} = 0$ <i>Init</i> \triangleq $\wedge \text{hour} = 0$ $\wedge \text{minute} = 0$ <i>NextMinute</i> \triangleq $\wedge \text{minute} = 59$ $\wedge \text{hour}' = (\text{hour} + 1) \% 24$ </pre>	MODULE <i>clock</i>
--	---------------------

$$\begin{aligned}
& \wedge \text{minute}' = 0 \\
\text{NextHour} & \triangleq \\
& \wedge \text{minute} \neq 59 \\
& \wedge \text{hour}' = \text{hour} \\
& \wedge \text{minute}' = \text{minute} + 1 \\
\text{Next} & \triangleq \\
& \vee \text{NextMinute} \\
& \vee \text{NextHour} \\
\text{Spec} & \triangleq \\
& \wedge \text{Init} \\
& \wedge \Box[\text{Next}]_{\text{vars}} \\
& \wedge \text{WF}_{\text{vars}}(\text{Next})
\end{aligned}$$

The corresponding clock.cfg is listed below:

```

SPECIFICATION Spec
INVARIANTS Type_OK
PROPERTIES Liveness

```

Now run TLC and one should see something like this:

```

Model checking completed. No error has been found.
...
The depth of the complete state graph search is 1440.

```

Part I

Examples

Chapter 3

Simple Gossip Protocol

This section the author's notes on a simple gossip protocol by Andrew Hewler:
<https://ahelwer.ca/post/2023-11-01-tla-finite-monotonic/>

3.1 Requirement

In a distributed system, a cluster of nodes collectively provide a service. A distributed database may have a collection of 10s to 100s of nodes working together to offer the service in a geo diverse fashion to be immune to partial outage. The nodes often have requirements to know about each other. In the context of distributed database, a node may need to know the key range another of its peers. The cluster needs a way to communicate this information. One such mechanism is the gossip protocol.

Gossip protocols are used to communicate cluster information in a distributed fashion, (unsurprisingly) in a distributed system. Without gossip protocol, nodes in a cluster learn about its neighbours by contacting a centralized server. This introduces a single failure point in the system. As the name suggests, gossip protocol relies on nodes to gossip with each other. The nodes in the cluster periodically select a set of neighbors to exchange what it knows about the cluster. The recency information is part of the gossip message itself, allowing the node and the peer it's talking to quickly decide who has the latest information on a node, and converge to it. Assume a N node cluster and each internal a node selects k neighbours to gossip with, the total amount of gossip propagation time is described logarithmically below:

$$propagation_time = \log_k N * gossip_interval \quad (3.1)$$

With the total number of messages exchanged:

$$messages_exchanged = \log_k N * k \quad (3.2)$$

Now let's look at how a simple gossip protocol can be described by TLA+.

3.2 Spec

In gossip protocol, every node needs to remember all its peers current version. This can be represented as a two dimension array:

$$Init \triangleq counter = [n \in Node \mapsto [o \in Node \mapsto 0]]$$

This defines counter a collection of nodes, where each node also contains a collection of nodes initialized to 0. The nodes can bump to a new version:

$$Increment(n) \triangleq counter' = [counter \text{ EXCEPT } ![n][n] = @ + 1]$$

Notice increment only bumps node's version. This change needs to be gossiped across the cluster:

$$\begin{aligned} Gossip(n, o) &\triangleq \\ &LET \ Max(a, b) \triangleq IF \ a > b \ THEN \ a \ ELSE \ b \\ &IN \ counter' = [\\ &\quad counter \text{ EXCEPT } ![o] = [\\ &\quad \quad nn \in Node \mapsto \\ &\quad \quad \quad Max(counter[n][nn], counter[o][nn]) \\ &\quad] \\ &] \end{aligned}$$

A few things to unpack here:

- n, o are the two nodes exchanging gossip. o is the node to be updated and n is the neighbor o gossips with.
- *LET..IN* allows local definition under *LET* used under *IN*. In this case *Max* is a local macro defined to return maximum between a and b .
- $counter'$ (or referred to as counter *prime*) is what the variable will be in the next state. TLA+ doesn't provide a way to update a variable in a collection, so the convention is to assign a new array to the variable.
- $counter \text{ EXCEPT } ![o] = [...]$ return *counter* with $counter[o]$ defined in the bracket.
- where $[...]$ is a collection of nodes with with counter set to the max between the current node and neighbour.

Finally, the actual spec:

$$Next \triangleq \vee \exists n \in Node : Increment(n)$$

$$\forall \exists n, o \in \text{Node} : \text{Gossip}(n, o)$$

Next supports two possible next steps describe using disjunctions. The first is bumping the version of a random node, the second is select a pair of nodes to gossip. Note the *existential qualifier* on both, which basically states there exists a node n in nodes, or there exists a pair of nodes n, o in nodes, respectively.

There's a minor problem with the definition above. Gossip protocol, like many converging protocols, have a *monotonic increasing* requirement. On failures, the protocol bumps the version, which increases monotonically. Since TLA+ spec models the system as a graph, a monotonic increasing version number means the state graph is *infinitely large*. To put the specification back into finite space, we can normalize the state:

$$\begin{aligned} \text{GarbageCollect} &\triangleq \\ &\text{LET } \text{SetMin}(s) \triangleq \text{CHOOSE } e \in s : \forall o \in s : e \leq o \text{ IN} \\ &\text{LET } \text{Transpose} \triangleq \text{SetMin}(\{\text{counter}[n][o] : n, o \in \text{Node}\}) \text{ IN} \\ &\quad \wedge \text{counter}' = [\\ &\quad \quad n \in \text{Node} \mapsto [\\ &\quad \quad \quad o \in \text{Node} \mapsto \text{counter}[n][o] - \text{Transpose} \\ &\quad \quad] \\ &\quad] \\ &\quad \wedge \text{UNCHANGED } \text{converge} \end{aligned}$$

GarbageCollect subtracts every version value with set minimum. To limit range of version value, the increment function is now updated to:

$$\begin{aligned} \text{Increment}(n) &\triangleq \\ &\quad \wedge \neg \text{converge} \\ &\quad \wedge \text{counter}[n][n] < \text{Divergence} \\ &\quad \wedge S! \text{Increment}(n) \\ &\quad \wedge \text{UNCHANGED } \text{converge} \end{aligned}$$

Finally, the *Next* is updated to the follow:

$$\begin{aligned} \text{Next} &\triangleq \\ &\quad \vee \exists n \in \text{Node} : \text{Increment}(n) \\ &\quad \vee \exists n, o \in \text{Node} : \text{Gossip}(n, o) \\ &\quad \vee \text{Converge} \\ &\quad \vee \text{GarbageCollect} \end{aligned}$$

Note *GarbageCollect* is a now part of possible state transition. We will discuss *Converge* later, as it is related to liveness check. Lastly:

$$\begin{aligned} \text{Fairness} &\triangleq \forall n, o \in \text{Node} : \text{WF}_{\text{vars}}(\text{Gossip}(n, o)) \\ \text{Spec} &\triangleq \\ &\quad \wedge \text{Init} \end{aligned}$$

$$\begin{aligned} & \wedge \Box[Next]_{vars} \\ & \wedge Fairness \end{aligned}$$

The *Fairness* definition ensures Gossip runs between every pair of nodes gossip.

3.3 Safety

In every state, counter[n][o] next must be larger than counter[n][o] current:

$$\begin{aligned} Monotonic & \triangleq \forall n, o \in Node : counter'[n][o] \geq counter[n][o] \\ Monotonicity & \triangleq \Box[\\ & \quad \vee S!Monotonic \\ & \quad \vee \forall a, b, c, d \in Node : \\ & \quad \quad (counter'[a][b] - counter[a][b]) = (counter'[c][d] - counter[c][d]) \\ &]_{vars} \end{aligned}$$

3.4 Liveness

For liveness we want to check the version value across all nodes eventually converge. *Next* is updated to set *Converge* to true, which triggers the liveness condition and ensure all pair of nodes eventually have the same information.

$$\begin{aligned} Convergence & \triangleq \forall n, o \in Node : counter[n] = counter[o] \\ Liveness & \triangleq converge \leadsto S!Convergence \\ Converge & \triangleq \\ & \quad \wedge converge' = TRUE \\ & \quad \wedge UNCHANGED counter \\ Next & \triangleq \\ & \quad \vee \exists n \in Node : Increment(n) \\ & \quad \vee \exists n, o \in Node : Gossip(n, o) \\ & \quad \vee Converge \\ & \quad \vee GarbageCollect \end{aligned}$$

Chapter 4

Simple Scheduler

4.1 Requirement

In this section we will define a spec for a simple task scheduler. The task scheduler has the following requirements:

- Supporting N execution context (ie. CPUs)
- Supporting T number of tasks
- Tasks have identical priority and are scheduled cooperatively
- System has a single global lock
- Any task can attempt to acquire the lock, Any task attempting to acquire the lock are guaranteed to be scheduled.
- If multiple tasks attempt to grab the lock, the tasks will be scheduled in lock request order.

4.2 Spec

We will model scheduler using the following variables:

$$\begin{aligned} Init &\triangleq \\ &\wedge \textit{cpus} = [i \in 0 \dots N - 1 \mapsto \textit{""}] \\ &\wedge \textit{ready}_q = S2T(\textit{Tasks}) \\ &\wedge \textit{blocked}_q = \langle \rangle \\ &\wedge \textit{lock_owner} = \textit{""} \end{aligned}$$

A few things to note:

- The system has N executing context, represented as number of CPUs. When a task is running, $cpus[k]$ is set to *taskName*. When CPU is idle, $cpus[k]$ is set to an empty string.
- *ready_q* and *blocked_q* are initialized as *ordered tuple*, due to the cooperative scheduling requirement.
- *S2T* is a macro that converts a set into a ordered tuple. This is to accommodate the fact it appears I cannot define tuple in .cfg file.
- Finally, the single system lock is represented as *lock_owner*.

A task can be in three possible state: Ready, Blocked and Running. The *Next* box-action fomula will define a Ready and Running action, and the implementation will include related lock contention handling.

MODULE *scheduler*

$$\begin{aligned}
 & MoveToReady(k) \triangleq \\
 & \quad \wedge cpus[k] \neq "" \\
 & \quad \wedge lock_owner \neq cpus[k] \\
 & \quad \wedge ready_q' = Append(ready_q, cpus[k]) \\
 & \quad \wedge cpus' = [cpus \text{ EXCEPT } ![k] = ""] \\
 & \quad \wedge UNCHANGED \langle lock_owner, blocked_q \rangle \\
 & Lock(k) \triangleq \\
 & \quad lock \text{ is empty} \\
 & \quad \vee \wedge cpus[k] \neq "" \\
 & \quad \quad \wedge lock_owner = "" \\
 & \quad \quad \wedge lock_owner' = cpus[k] \\
 & \quad \quad \wedge UNCHANGED \langle ready_q, cpus, blocked_q \rangle \\
 & \quad \text{someone else has the lock} \\
 & \quad \vee \wedge cpus[k] \neq "" \\
 & \quad \quad \wedge lock_owner \neq "" \\
 & \quad \quad \wedge lock_owner \neq cpus[k] \text{ cannot double lock} \\
 & \quad \quad \wedge blocked_q' = Append(blocked_q, cpus[k]) \\
 & \quad \quad \wedge cpus' = [cpus \text{ EXCEPT } ![k] = ""] \\
 & \quad \quad \wedge UNCHANGED \langle ready_q, lock_owner \rangle \\
 & Unlock(k) \triangleq \\
 & \quad \wedge cpus[k] \neq "" \\
 & \quad \wedge lock_owner = cpus[k] \\
 & \quad \wedge lock_owner' = "" \\
 & \quad \wedge cpus' = [cpus \text{ EXCEPT } ![k] = ""] \\
 & \quad \wedge ready_q' = ready_q \circ blocked_q \circ \langle cpus[k] \rangle \\
 & \quad \wedge blocked_q' = \langle \rangle \\
 & Running \triangleq \\
 & \quad \exists k \in \text{DOMAIN } cpus : \\
 & \quad \quad \wedge cpus[k] \neq ""
 \end{aligned}$$

$$\begin{aligned}
& \wedge \vee \text{MoveToReady}(k) \\
& \vee \text{Lock}(k) \\
& \vee \text{Unlock}(k)
\end{aligned}$$

4.3 Safety

4.4 Liveness

I believe this is the most important part of cooperative scheduler design. While the scheduler can't *force* a task to relinquish a lock (the scheduler doesn't dictate when the task is *done*), the scheduler can ensure scheduling fairness by scheduling the next lock requester instead of the task that just relinquished the lock.

$$\begin{aligned}
& \text{MODULE scheduler} \\
& \text{Liveness} \triangleq \\
& \quad \forall t \in \text{Tasks} : \\
& \quad \text{LET} \\
& \quad \quad b \triangleq \{x \in \text{DOMAIN blocked_q} : \text{blocked_q}[x] = t\} \\
& \quad \text{IN} \\
& \quad \quad \wedge b \neq \{\} \leadsto b = \{\}
\end{aligned}$$

The formula defines set b to be either an empty set or a set of one task. Assume a set of {"p0", "p1", "p2"}. Possible value of b include: {}, {"p0"}, {"p1"} and {"p2"}. The formula then states a non empty set of b leads to an *empty set* of b . In other words:

If a task ever becomes blocked, it will eventually become unblocked.

However, when we actually run the model checker, we will find the liveness property is *violated*. The failure scenario is basically one task holding onto the lock in one CPU, while the scheduler repeatedly schedule/deschedule a separate task in another CPU. While this is perfectly allowed, the model checker detects a possible path for the the system to trap in a local state and fail the liveness property.

Perhaps not surprisingly, if you construct similar liveness property to verify a task is *eventually* always scheduled, it will also fail. The model checker will provide a counter case where a task is never scheduled because another task is repeatedly acquire/release the global lock.

We need *Strong Fairness* to solve this problem:

$$\text{MODULE scheduler}$$

$$\begin{aligned}
L &\triangleq \\
&\quad \forall t \in Tasks : \\
&\quad \quad \forall n \in 0 \dots (N-1) : \\
&\quad \quad \quad \mathbf{WF}_{vars}(HoldingLock(t) \wedge Unlock(n)) \\
Spec &\triangleq \\
&\quad \wedge Init \\
&\quad \wedge \Box[Next]_{vars} \\
&\quad \wedge \mathbf{WF}_{vars}(Next) \\
&\quad \wedge L
\end{aligned}$$

Fairness ensures that we are never stuck in a repeated state.

Chapter 5

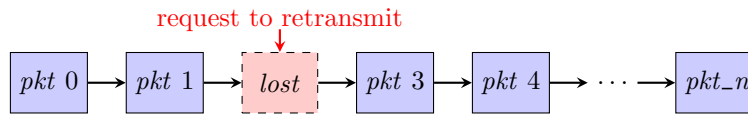
Selective Retransmit

Assume client device that playback a video stream. Structurally, a video is composed of frames, frames are then segmented into packets to stream across a network. The client device recombines the packets into a frame, then sequence the frame to playback the video.

However, network is not-deterministic. Depending on the route the packets take to get to the client, they may arrive out-of-order. The client may need to maintain a receive buffer for the packets, re-order the packets back into sequence before pushing the packets down to decoding engine.

The network may also drop packets if any of the switches gets too busy. In the case of a packet drop, the client has a few options. The client can either discard the frame and let the decoding engine downstream to deal with it (which may result in visible artifacts during play back). The client can request the whole frame to be re-sent, which results in additional bandwidth consumption. The client can selectively request the missing packet to be retransmitted, which will minimize additional bandwidth consumption, but increases implementation complexity.

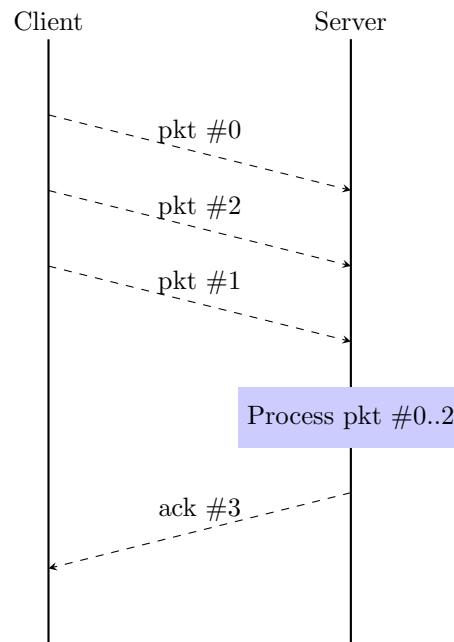
In this chapter, we will implement a simple selective retransmit algorithm.



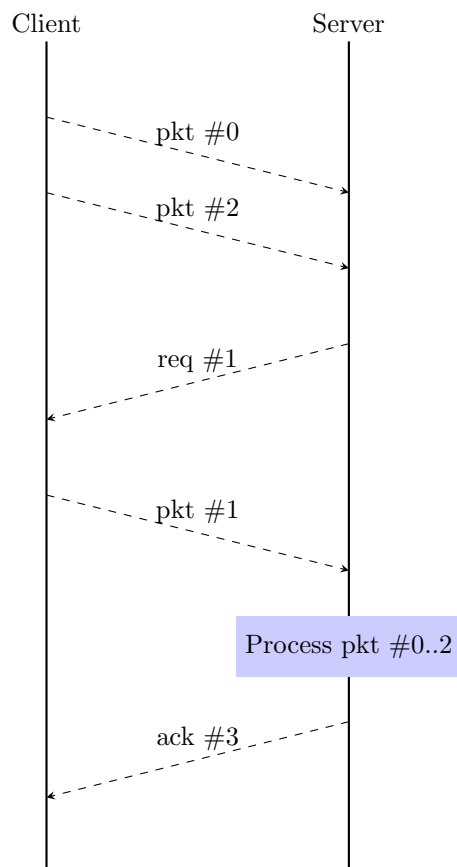
Since packets may arrive out-of-order, the server stamps the packets with sequence number to allow the client to order the packets as they arrive. Once the client has a set of ordered packets, it moves the packets from the receive buffer into the decoding engine to be displayed.

The video packets are often sent via unreliable channel to minimize network overhead and latency. The client sends acknowledgement back to the server to acknowledge the received packet. This indicates to the server it can send more video data to the client. Acknowledgements are not latency sensitive in nature, and take up a very small proportion of bandwidth, so they are transported through reliable channel.

The following illustrates packet reorder handling:



The following illustrates packet loss handling:



There are other design considerations. The server is allowed to send up to W packets before getting an acknowledgement, this reduces latency perceived by the user. The client also doesn't need to acknowledge all the packets, since the server assumes once an acknowledgement of packet N is received, then all packets prior to N have also been received.

5.1 Requirement

With the above description, we are now ready to provide a more formal description for our requirement:

- Client is the receiver that displays the video stream
- Server is the sender that sends the video stream
- Server always sends the packets in-order
- Client may receive the packets out-of-order

- Client may never receive some packet due to loss
- Server can send up to W packets before an acknowledgement is received
- Per packet sequence number is represented by fixed number of bytes in the network header, the sequence number will eventually wrap around once it hits the maximum representable value. This upper limit will be represented by the constant N
- Client will remove the packets from the receive buffer once the sequence number of the received packets are contiguous. The client will also send an acknowledgement back to the Server with the most recent sequence number
- Server to Client data packets are assumed unreliable due to bulk bandwidth
- Client to Server control packets are assumed to be reliable

We are now ready to implement the *Spec*.

5.2 Spec

The following is the skeleton of the *Spec*:

$$\begin{aligned}
 Init &\triangleq \\
 &\wedge network = \{\} \\
 &\wedge server_tx = 0 \\
 &\wedge server_tx_limit = W \\
 &\wedge server_tx_ack = 0 \\
 &\wedge client_rx = 0 \\
 &\wedge client_buffer = \{\} \\
 &\wedge lost = 0 \\
 \\
 Next &\triangleq \\
 &\vee Send \\
 &\vee \exists p \in network : \\
 &\quad Receive(p) \\
 &\vee ClientRetransmitRequest \\
 &\vee ClientAcknowledgement \\
 &\vee \exists p \in network : \\
 &\quad \wedge p.dst = \text{"client"} \\
 &\quad \wedge Drop(p)
 \end{aligned}$$

The server is represented by three variables:

- $tx+1$ represents the sequence number to be used in the next packet

- `tx_limit` represents the highest sequence number server can send without waiting for an acknowledgement
- `tx_ack` represents the most recent acknowledged sequence number

The client is represented by two variables:

- `client_rx` is the most recently acknowledged sequence number
- `client_buffer` is the receive buffer holding all the packets waiting to be re-ordered prior to being acknowledged

The allowed actions include packet *Receive*, which the existential quantifier also has the side effect of re-ordering. *ClientRetransmitRequest* detects and sends retransmit request. *ClientAcknowledgement* sends acknowledgement. Finally, data packets may be dropped.

Before we start defining the actions, let us define some helper functions:

$$\begin{aligned} \text{MinS}(s) &\triangleq \\ &\text{CHOOSE } x \in s : \forall y \in s : x \leq y \\ \text{MaxS}(s) &\triangleq \\ &\text{CHOOSE } x \in s : \forall y \in s : x \geq y \\ \text{MaxIndex} &\triangleq \\ &\text{LET} \\ &\quad \text{upper} \triangleq \{x \in \text{client_buffer} : x > N - W\} \\ &\quad \text{lower} \triangleq \{x \in \text{client_buffer} : x < W\} \\ &\quad \text{maxv} \triangleq \text{IF } \text{upper} \neq \{\} \wedge \text{lower} \neq \{\} \\ &\quad \text{THEN} \\ &\quad \quad \text{MaxS}(\text{lower}) \\ &\quad \text{ELSE} \\ &\quad \quad \text{MaxS}(\text{client_buffer}) \\ &\text{IN} \\ &\quad \text{maxv} \\ \text{MinIndex} &\triangleq \\ &\text{LET} \\ &\quad \text{upper} \triangleq \{x \in \text{client_buffer} : x > N - W\} \\ &\quad \text{lower} \triangleq \{x \in \text{client_buffer} : x < W\} \\ &\quad \text{minv} \triangleq \text{IF } \text{upper} \neq \{\} \wedge \text{lower} \neq \{\} \\ &\quad \text{THEN} \\ &\quad \quad \text{MinS}(\text{upper}) \\ &\quad \text{ELSE} \\ &\quad \quad \text{MinS}(\text{client_buffer}) \\ &\text{IN} \\ &\quad \text{minv} \end{aligned}$$

$$Range \triangleq$$

IF $MaxIndex \geq MinIndex$
THEN
 $MaxIndex - MinIndex + 1$
ELSE
 $MaxIndex + 1 + N - MinIndex$

Chapter 6

Raft Consensus Protocol

Raft is a consensus algorithm that enables a cluster of nodes to agree on a collective state even in the presence of failures. An application of Raft is database replication protocol. With a replication factor of 3 (which means the data is replicated across 3 nodes) and hard drive failure rate of 0.81% per year, the possibility of the total failure where the entire replication group goes down is $1 - 0.0081^3 = 99.9999\%$ uptime [3].

Since the purpose of the book is to learn TLA+, this chapter will only implement a portion of the Raft protocol, namely leader election. For a full description of the the Raft protocol, please refer to the original paper [4].

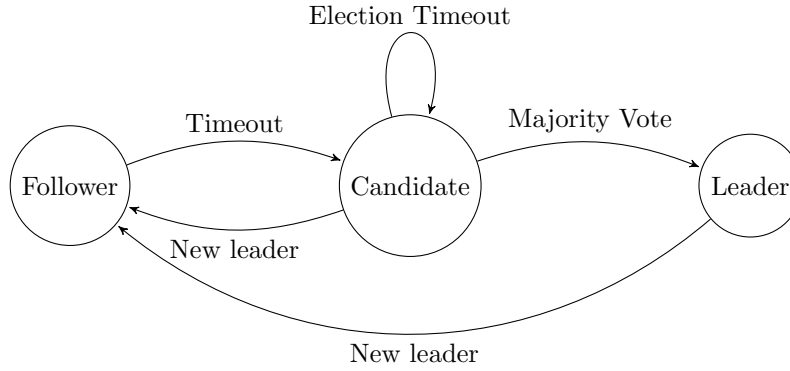
We will briefly describe Raft and its leader election process below:

- A Raft cluster have N nodes, the cluster work collective as a *system* to offer some service
- Each node can be in one of three possible states: Follower, Candidate, Leader
- During normal operations, a cluster of N nodes have a single leader and N-1 followers
- The leader handles all the client interactions. Requests sent to followers will be redirected to the leader
- The leader regularly sends heartbeat to the follower, indicate its alive
- If a follower fails to receive heartbeat from the leader after timeout, it will become a candidate, vote for itself, and campaign to be leader
- A candidate collect the majority of the vote becomes the leader
- If multiple candidates are campaigning and a split vote happen, candidates will eventually declare election timeout and start new round of election

- The cluster can have multiple leaders due to unfavourable network conditions, but the leaders must be on different terms
- A newly elected leader will send a heartbeat to other nodes to establish leadership
- All request and responses include the sender's term, allowing the receiver to react accordingly

The protocol also included description about log synchronization, state recovery, and more. Many details are omitted in this chapter to reduce modeling cost. The N nodes in the cluster operate *independently* following the above heuristics. Hopefully this highlights the complexity around verifying the correctness of the protocol.

The following illustrate the state diagram of one node in the cluster:



6.1 Spec

The following implements the skeleton portion of the leader election protocol:

$$\begin{aligned}
 Init &\triangleq \\
 &\wedge state = [s \in Servers \mapsto \text{"Follower"}] \\
 &\wedge messages = \{\} \\
 &\wedge voted_for = [s \in Servers \mapsto \text{""}] \\
 &\wedge vote_granted = [s \in Servers \mapsto \{\}] \\
 &\wedge vote_requested = [s \in Servers \mapsto 0] \\
 &\wedge term = [s \in Servers \mapsto 0] \\
 \\
 RequestVoteSet(i) &\triangleq \{ \\
 &[fSrc \mapsto i, fDst \mapsto s, fType \mapsto \text{"RequestVoteReq"}, fTerm \mapsto term[i]] \\
 &\quad : s \in Servers \setminus \{i\} \\
 &\}
 \end{aligned}$$

$$\begin{aligned}
\text{Campaign}(i) &\triangleq \\
&\wedge \text{vote_requested}[i] = 0 \\
&\wedge \text{vote_requested}' = [\text{vote_requested} \text{ EXCEPT } ![i] = 1] \\
&\wedge \text{messages}' = \text{messages} \cup \text{RequestVoteSet}(i) \\
&\wedge \text{UNCHANGED } \langle \text{state}, \text{term}, \text{vote_granted}, \text{voted_for} \rangle \\
\\
\text{KeepAliveSet}(i) &\triangleq \{ \\
&[f\text{Src} \mapsto i, f\text{Dst} \mapsto s, f\text{Type} \mapsto \text{"AppendEntryReq"}, f\text{Term} \mapsto \text{term}[i]] \\
&\quad : s \in \text{Servers} \setminus \{i\} \\
&\} \\
\\
\text{Leader}(i) &\triangleq \\
&\wedge \text{state}[i] = \text{"Leader"} \\
&\wedge \text{messages}' = \text{messages} \cup \text{KeepAliveSet}(i) \\
&\wedge \text{UNCHANGED } \langle \text{state}, \text{voted_for}, \text{term}, \text{vote_granted}, \text{vote_requested} \rangle \\
\\
\text{BecomeLeader}(i) &\triangleq \\
&\wedge \text{Cardinality}(\text{vote_granted}[i]) > \text{Cardinality}(\text{Servers}) \div 2 \\
&\wedge \text{state}' = [\text{state} \text{ EXCEPT } ![i] = \text{"Leader"}] \\
&\wedge \text{UNCHANGED } \langle \text{messages}, \text{voted_for}, \text{term}, \text{vote_granted}, \text{vote_requested} \rangle \\
\\
\text{Candidate}(i) &\triangleq \\
&\wedge \text{state}[i] = \text{"Candidate"} \\
&\wedge \vee \text{Campaign}(i) \\
&\quad \vee \text{BecomeLeader}(i) \\
&\quad \vee \text{Timeout}(i) \\
\\
\text{Follower}(i) &\triangleq \\
&\wedge \text{state}[i] = \text{"Follower"} \\
&\wedge \text{Timeout}(i) \\
\\
\text{Receive}(msg) &\triangleq \\
&\vee \wedge msg.f\text{Type} = \text{"AppendEntryReq"} \\
&\quad \wedge \text{AppendEntryReq}(msg) \\
&\vee \wedge msg.f\text{Type} = \text{"AppendEntryResp"} \\
&\quad \wedge \text{AppendEntryResp}(msg) \\
&\vee \wedge msg.f\text{Type} = \text{"RequestVoteReq"} \\
&\quad \wedge \text{RequestVoteReq}(msg) \\
&\vee \wedge msg.f\text{Type} = \text{"RequestVoteResp"} \\
&\quad \wedge \text{RequestVoteResp}(msg) \\
\\
\text{Next} &\triangleq \\
&\vee \exists i \in \text{Servers} : \\
&\quad \vee \text{Leader}(i) \\
&\quad \vee \text{Candidate}(i) \\
&\quad \vee \text{Follower}(i) \\
&\vee \exists msg \in \text{messages} : \text{Receive}(msg)
\end{aligned}$$

- *Next* either picks a server to make progress, or picks a message in the message pool to process. Message processing is done by *Receive*, handling is state agnostic
- *message* is defined to be a set that holds a collection of functions, where each function is a message with source, destination, type, and more specified
- *voted_for* tracks who a given node previously voted for. This prevents a node from voting more than once
- *vote_granted* tracks how many votes a candidate has received
- *vote_requested* tracks if a node has already issued request vote to its peers
- *Follower* either Receive or Timeout and campaign to be a leader
- *Candidate* campaigns to be a leader, and becomes one if it has enough vote. Failing to collect enough votes, *Candidate* start a new election on a new term. It can also receive a request with a higher term and transition to be a *Follower*.
- *Leader* will establish its leadership by sending *AppendEntryReq* to all its peers

The Spec implements four messages *AppendEntry* request/response, *RequestVote* request/response. Handling for all messages are fairly similar in structure. In this chapter we will look at *RequestVoteReq* only. Readers are encouraged to check the remaining definition as an exercise:

$$\begin{aligned}
 \text{RequestVoteReq}(msg) &\triangleq \\
 &\text{LET} \\
 &\quad i \triangleq msg.fDst \\
 &\quad j \triangleq msg.fSrc \\
 &\quad type \triangleq msg.fType \\
 &\quad t \triangleq msg.fTerm \\
 &\text{IN} \\
 &\quad \text{haven't voted, or whom we voted re-requested} \\
 &\quad \vee \wedge t = term[i] \\
 &\quad \wedge \vee voted_for[i] = j \\
 &\quad \quad \vee voted_for[i] = "" \\
 &\quad \wedge voted_for' = [voted_for \text{ EXCEPT } ![i] = j] \\
 &\quad \wedge messages' = AddMessage([fSrc \mapsto i, \\
 &\quad \quad \quad fDst \mapsto j, \\
 &\quad \quad \quad fType \mapsto "RequestVoteResp", \\
 &\quad \quad \quad fTerm \mapsto t, \\
 &\quad \quad \quad fSuccess \mapsto 1], \\
 &\quad \quad \quad RemoveMessage(msg, messages))
 \end{aligned}$$

$$\begin{aligned}
& \wedge \text{UNCHANGED } \langle \text{state}, \text{term}, \text{vote_granted}, \text{vote_requested}, \text{establish_leadership} \rangle \\
& \text{already voted someone else} \\
\vee & \wedge t = \text{term}[i] \\
& \wedge \text{voted_for}[i] \neq j \\
& \wedge \text{voted_for}[i] \neq "" \\
& \wedge \text{messages}' = \text{AddMessage}([f\text{Src} \mapsto i, \\
& \quad f\text{Dst} \mapsto j, \\
& \quad f\text{Type} \mapsto \text{"RequestVoteResp"}, \\
& \quad f\text{Term} \mapsto t, \\
& \quad f\text{Success} \mapsto 0], \\
& \quad \text{RemoveMessage}(\text{msg}, \text{messages})) \\
& \wedge \text{UNCHANGED } \langle \text{state}, \text{voted_for}, \text{term}, \text{vote_granted}, \text{vote_requested}, \text{establish_leadership} \rangle \\
\vee & \wedge t < \text{term}[i] \\
& \wedge \text{messages}' = \text{AddMessage}([f\text{Src} \mapsto i, \\
& \quad f\text{Dst} \mapsto j, \\
& \quad f\text{Type} \mapsto \text{"RequestVoteResp"}, \\
& \quad f\text{Term} \mapsto \text{term}[i], \\
& \quad f\text{Success} \mapsto 0], \\
& \quad \text{RemoveMessage}(\text{msg}, \text{messages})) \\
& \wedge \text{UNCHANGED } \langle \text{state}, \text{voted_for}, \text{term}, \text{vote_granted}, \text{vote_requested}, \text{establish_leadership} \rangle \\
& \text{revert back to follower} \\
\vee & \wedge t > \text{term}[i] \\
& \wedge \text{state}' = [\text{state} \text{ EXCEPT } ![i] = \text{"Follower"}] \\
& \wedge \text{term}' = [\text{term} \text{ EXCEPT } ![i] = t] \\
& \wedge \text{voted_for}' = [\text{voted_for} \text{ EXCEPT } ![i] = j] \\
& \wedge \text{vote_granted}' = [\text{vote_granted} \text{ EXCEPT } ![i] = \{\}] \\
& \wedge \text{vote_requested}' = [\text{vote_requested} \text{ EXCEPT } ![i] = 0] \\
& \wedge \text{establish_leadership}' = [\text{establish_leadership} \text{ EXCEPT } ![i] = 0] \\
& \wedge \text{messages}' = \text{AddMessage}([f\text{Src} \mapsto i, \\
& \quad f\text{Dst} \mapsto j, \\
& \quad f\text{Type} \mapsto \text{"RequestVoteResp"}, \\
& \quad f\text{Term} \mapsto t, \\
& \quad f\text{Success} \mapsto 1], \\
& \quad \text{RemoveMessage}(\text{msg}, \text{messages}))
\end{aligned}$$

The handling is split into three cases:

- If received request is on a higher term, processing node grants vote and becomes a Follower
- If received request is on a lower term, processing node ignores request
- If received request is on the same term, processing node only grants vote if it hasn't voted, or has had voted for the same requester prior

6.2 Simplify Model

The model checker will run the spec as defined, but due to the exponential growth of states it is unlikely to complete in a reasonable amount of time. We need to simplify the model and possibly trades off some correctness. Careful consideration must go into finding the right balance between maximizing model correctness and minimizing model checker runtime.

The main strategy is to *bound* the state graph. The following describe a set of optimization implemented for this example.

6.2.1 Modeling Messages as a Set

In the original Raft TLA+ Spec [5], messages are modeled as an *unordered map* to track the count of each message. It is possible for a sender to repeatedly send the same message (eg. keepalive), and grow the message count in an unbounded fashion.

messages in this example has been implemented as a set, which effectively limits message instance count to one. It is still possible for messages to grow unboundedly because of the monotonically increasing term value. Further changes are described below.

6.2.2 Limit Term Divergence

It is possible for a node to *never* make progress. Such case can occur when a node is partitioned off while the rest of the cluster elects new leader and move onto newer terms. Many of the interesting behaviours of Raft are how it addresses these cases. In a cluster of nodes with mixed terms, the nodes with older term will eventually converge onto newer terms when they are contacted by new leader. This converging behaviour will happen whether the stale node is either 1 or N terms away from the current leader, and the former is much less costly to simulate than the latter because the reduced number of states.

We can include *LimitDivergence* as a conjunction in *Timeout*:

$$\begin{aligned}
 \text{LimitDivergence}(i) &\triangleq \\
 &\text{LET} \\
 &\quad \text{values} \triangleq \{ \text{term}[s] : s \in \text{Servers} \} \\
 &\quad \text{max_v} \triangleq \text{CHOOSE } x \in \text{values} : \forall y \in \text{values} : x \geq y \\
 &\quad \text{min_v} \triangleq \text{CHOOSE } x \in \text{values} : \forall y \in \text{values} : x \leq y \\
 &\text{IN} \\
 &\quad \vee \wedge \text{term}[i] \neq \text{max_v} \\
 &\quad \vee \wedge \text{term}[i] = \text{max_v} \\
 &\quad \quad \wedge \text{term}[i] - \text{min_v} < \text{MaxDiff}
 \end{aligned}$$

$$\begin{aligned}
\text{Timeout}(i) &\triangleq \\
&\wedge \text{LimitDivergence}(i) \\
&\wedge \text{state}' = [\text{state} \text{ EXCEPT } ![i] = \text{"Candidate"}] \\
&\wedge \text{voted_for}' = [\text{voted_for} \text{ EXCEPT } ![i] = i] \quad \text{voted for myself} \\
&\wedge \text{vote_granted}' = [\text{vote_granted} \text{ EXCEPT } ![i] = \{i\}] \\
&\wedge \text{vote_requested}' = [\text{vote_requested} \text{ EXCEPT } ![i] = 0] \\
&\wedge \text{term}' = [\text{term} \text{ EXCEPT } ![i] = @ + 1] \quad \text{bump term} \\
&\wedge \text{establish_leadership}' = [\text{establish_leadership} \text{ EXCEPT } ![i] = 0] \\
&\wedge \text{UNCHANGED } \langle \text{messages} \rangle \\
&/ \text{PrintT}(\text{state}')
\end{aligned}$$

6.2.3 Normalize Cluster Term

However, term *itself* can grow unbounded. This is a key tenet converging protocols rely on, an monotonically increasing counter. We want to *normalize* the range of terms in the cluster so the minimum value resets back to 0. This provides an upper bound to the state graph.

$$\begin{aligned}
\text{Normalize} &\triangleq \\
&\text{LET} \\
&\quad \text{values} \triangleq \{ \text{term}[s] : s \in \text{Servers} \} \\
&\quad \text{max_v} \triangleq \text{CHOOSE } x \in \text{values} : \forall y \in \text{values} : x \geq y \\
&\quad \text{min_v} \triangleq \text{CHOOSE } x \in \text{values} : \forall y \in \text{values} : x \leq y \\
&\text{IN} \\
&\quad \wedge \text{max_v} = \text{MaxTerm} \\
&\quad \wedge \text{term}' = [s \in \text{Servers} \mapsto \text{term}[s] - \text{min_v}] \\
&\quad \wedge \text{messages}' = \{ \} \\
&\quad \wedge \text{UNCHANGED } \langle \text{state}, \text{voted_for}, \text{vote_granted}, \text{vote_requested}, \text{establish_leadership} \rangle \\
\text{Next} &\triangleq \\
&\vee \wedge \forall i \in \text{Servers} : \text{term}[i] \neq \text{MaxTerm} \\
&\quad \wedge \vee \exists i \in \text{Servers} : \\
&\quad \quad \vee \text{Leader}(i) \\
&\quad \quad \vee \text{Candidate}(i) \\
&\quad \quad \vee \text{Follower}(i) \\
&\quad \vee \exists \text{msg} \in \text{messages} : \text{Receive}(\text{msg}) \\
&\vee \wedge \exists i \in \text{Servers} : \text{term}[i] = \text{MaxTerm} \\
&\quad \wedge \text{Normalize}
\end{aligned}$$

The implementation ensures only the state machine only moves forward when none of the nodes is on *MaxTerm*. If any of the node is on *MaxTerm*, the cluster terms are normalized.

Another caveat here is in the initial implementation I didn't update messages. This led to liveness property violation as the messages had terms disagree-

ing with the system state. To simplify the spec I simply cleared all messages. This indirectly verifies a portion of the packet loss handling in the spec as well.

6.2.4 Sending Request as a Batch

The send requests were initially implemented using the existential quantifier. This introduces many interleaving states. This was replaced with a universal quantifier so the set of messages are only sent once. The implementation no longer tracks if the responses were received, since the spec should handle packet loss scenarios as well.

$$\begin{aligned}
\text{RequestVoteSet}(i) &\triangleq \{ \\
&\quad [fSrc \mapsto i, fDst \mapsto s, fType \mapsto \text{"RequestVoteReq"}, fTerm \mapsto term[i]] \\
&\quad : s \in Servers \setminus \{i\} \\
&\} \\
\text{Campaign}(i) &\triangleq \\
&\quad \wedge vote_requested[i] = 0 \\
&\quad \wedge vote_requested' = [vote_requested \text{ EXCEPT } ![i] = 1] \\
&\quad \wedge messages' = messages \cup \text{RequestVoteSet}(i) \\
&\quad \wedge \text{UNCHANGED } \langle state, term, vote_granted, voted_for, establish_leadership \rangle \\
\text{KeepAliveSet}(i) &\triangleq \{ \\
&\quad [fSrc \mapsto i, fDst \mapsto s, fType \mapsto \text{"AppendEntryReq"}, fTerm \mapsto term[i]] \\
&\quad : s \in Servers \setminus \{i\} \\
&\} \\
\text{Leader}(i) &\triangleq \\
&\quad \wedge state[i] = \text{"Leader"} \\
&\quad \wedge establish_leadership[i] = 0 \\
&\quad \wedge establish_leadership' = [establish_leadership \text{ EXCEPT } ![i] = 1] \\
&\quad \wedge messages' = messages \cup \text{KeepAliveSet}(i) \\
&\quad \wedge \text{UNCHANGED } \langle state, voted_for, term, vote_granted, vote_requested \rangle
\end{aligned}$$

6.2.5 Prune Messages with Stale Terms

When a node's term advances, all messages targeted to this node with older terms are discarded. Keeping messages with stale terms allows the model checker to verify the node correctly discards them, but can exponentially grow the state machine. To simplify the model, we can prune stale messages as we add a new message:

$$\begin{aligned}
\text{AddMessage}(to_add, msgs) &\triangleq \\
&\quad \text{LET} \\
&\quad \quad pruned \triangleq \{msg \in msgs : \\
&\quad \quad \quad \neg(msg.fDst = to_add.fDst \wedge msg.fTerm < to_add.fTerm)\} \\
&\quad \text{IN}
\end{aligned}$$

$$pruned \cup \{to_add\}$$

$$RemoveMessage(to_remove, msgs) \triangleq$$

6.2.6 Enable Symmetry

Since the behaviour is symmetric between nodes, we can enable symmetry to speed up model checker runtime:

$$Perms \triangleq Permutations(Servers)$$

6.3 Safety

One of the goals for the protocol is to ensure the cluster only have one leader. It is possible for the clusters to have multiple leaders due to unfavourable network connections. For example, a leader node is partitioned off and a new leader is elected. However, even when the cluster have multiple leaders, they *must* be on different terms. The leader with the highest term is effectively the *true leader*. This invariant can be implemented like so:

$$\begin{aligned} LeaderUniqueTerm &\triangleq \\ &\forall s1, s2 \in Servers : \\ &\quad (state[s1] = \text{"Leader"} \wedge state[s2] = \text{"Leader"} \wedge s1 \neq s2) \Rightarrow (term[s1] \neq term[s2]) \end{aligned}$$

For every pair of nodes, they cannot both be Leaders and have the same term.

6.4 Liveness

In any failure recovery scenario, the nodes in the cluster converges to a higher term value either voluntary or involuntarily. For example:

- A node timed out and starts a new election on a new term
- A partitioned follower receives heartbeat from a new leader on a new term
- A candidate receiving a request vote from another candidate on a higher term

In any case, a node's term number always increase. This can be described as below:

$$\begin{aligned} Converge &\triangleq \\ &\forall s \in Servers : \\ &\quad term[s] = 0 \rightsquigarrow term[s] = MaxTerm - MaxDiff \end{aligned}$$

Instead of *MaxTerm*, we use *MaxTerm-MaxDiff* to ensure the liveness property is always upheld even after *Normalization*. However, running the spec against TLC now will encounter a set of stuttering issues. We also need to update the fairness description to ensure all possible actions are called when the enabling conditions are *eventually always* true:

$$\begin{aligned}
\textit{Liveness} &\triangleq \\
&\wedge \forall i \in \textit{Servers} : \\
&\quad \wedge \textit{WF}_{\textit{vars}}(\textit{Leader}(i)) \\
&\quad \wedge \textit{WF}_{\textit{vars}}(\textit{Candidate}(i)) \\
&\quad \wedge \textit{WF}_{\textit{vars}}(\textit{Follower}(i)) \\
&\wedge \textit{WF}_{\textit{vars}}(\exists \textit{msg} \in \textit{messages} : \textit{Receive}(\textit{msg}))
\end{aligned}$$

6.5 Requirement

6.6 Spec

$Init \triangleq$
 $\wedge lastHash = NoHash$
 $\wedge distributedLedger = [n \in Node \mapsto [h \in Hash \mapsto NoBlock]]$
 $\wedge received = [n \in Node \mapsto \{\}]$

- Every node is a ledger in this system, initialized to NoBlock
- Every node's received set is initialized to nothing

Part II

Examples with PlusCal

Chapter 7

SPSC Lockless Queue

A single producer single consumer (SPSC) lockless queue is a data exchange queue between a producer and a consumer. The SPSC lockfree queue enables data exchange between producer and consumer without the use of a lock, allowing both producer and consumer to make progress in all scenarios.

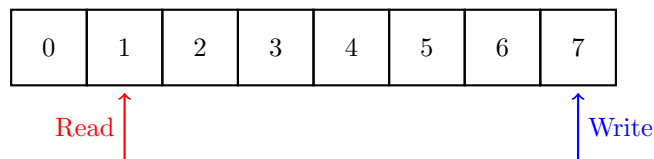
A real implementation need to account for memory ordering effects specific to the architecture. For example, ARM has weak memory ordering model where read/write may appear out-of-order between CPUs. In this chapter we will assume *logical* execution order where each command is perceived issued sequentially (even across CPUs) to focus the discussion on describing the system using TLA+.

7.1 Requirement

The following describes the SPSC queue requirements:

- Two executing context, reader and writer
- Writer advances $wptr$ after writes
- Reader advances $rptr$ after reads
- If $rptr$ equals $wptr$, queue is empty
- If $(wptr + 1) \% N$ equals $rptr$, queue is full

The following is an example of a SPSC queue:



Since reader and writer execute in different context, the instructions in read and write can interleave in *any* way imaginable:

- Reader empty check can happen just as writer is writing data
- Writer full check can happen just as reader is reading data
- Reading and writing can occur concurrently

The key observations is the index held by write pointer is reserved by the writer. Similarly, index held by the read pointer is reserved by the reader. The only exception is when read pointer equals to write pointer, then the queue is empty. Given the possible ways the reader and writer execution can interleave, we can use TLA+ to verify the design.

7.2 Spec

TLA+ specification can be written using its native formal specification language, or a C-like syntax called PlusCal (which transpiles down to its native form). In this example, I chose to implement the specification using PlusCal, since the content to be verified is psuedo implementation. While it is possible specify SPSC in native TLA+, I find the approach more error prone as each line is effectively an individual state to be modeled.

The following is a snippet of the *Spec* written in PlusCal:

```

procedure reader()
begin
  r_chk_empty:      if rptr = wptr then
    r_early_ret:    return ;
  end if ;
  r_read_buf:      assert buffer[rptr]  $\neq$  0 ;
  r_cs:            buffer[rptr] := 0 ;
  r_upd_rptr:      rptr := (rptr + 1)%N ;
  return ;
end procedure ;

procedure writer()
begin
  w_chk_full:      if (wptr + 1)%N = rptr then
    w_early_ret:    return ;
  end if ;
  w_write_buf:     assert buffer[wptr] = 0 ;
  w_cs:            buffer[wptr] := wptr + 1000 ;
  w_upd_wptr:      wptr := (wptr + 1)%N ;
  return ;
end procedure ;

```

Note some lines start with *label* (eg. `r_chk_empty`). All the actions associated with the label is assumed executed atomically. This is reflected in the generated TLA+ code:

```

r_chk_empty(self)  $\triangleq$   $\wedge$  pc[self] = "r_chk_empty"
 $\wedge$  IF rptr = wptr
  THEN  $\wedge$  pc' = [pc EXCEPT ![self] = "r_early_ret"]
  ELSE  $\wedge$  pc' = [pc EXCEPT ![self] = "r_read_buf"]
 $\wedge$  UNCHANGED (rptr, wptr, buffer, stack)

```

7.3 Safety

As mentioned before, safety properties need to hold true in every single state. Some safety requirement we can enforce, for example:

Reader and writer cannot access the same index at the same time:

$$\sim ((pc[100] = "w_cs") \wedge (pc[101] = "r_cs") \wedge rp\textit{tr} = w\textit{ptr}) \quad (7.1)$$

All unused index should be set to 0:

$$\forall kk \in \textit{unused} : \textit{buffer}[kk] = 0 \quad (7.2)$$

At any given moment, `buffer[wptr]` may be unused or written. `buffer[rptr]` may be unused or read:

$$\begin{aligned}
&\vee \textit{Cardinality}(\textit{to_be_read}) + 1 = \textit{Cardinality}(\textit{reading}) \\
&\vee \textit{Cardinality}(\textit{to_be_read}) = \textit{Cardinality}(\textit{reading}) + 1 \\
&\vee \textit{Cardinality}(\textit{to_be_read}) = \textit{Cardinality}(\textit{reading})
\end{aligned}$$

7.4 Liveness

All indices are eventually used:

$$\begin{aligned}
&\textit{Liveness} \triangleq \\
&\forall k \in 0 \dots N - 1 : \\
&\diamond (\textit{buffer}[k] \neq 0) \\
&\text{Unused index 0 becomes used, used index 0 becomes unused.} \\
&\textit{Liveness2} \triangleq \\
&\wedge (\textit{buffer}[0] = 0) \rightsquigarrow \textit{buffer}[0] = 1000 \\
&\wedge (\textit{buffer}[0] = 1000) \rightsquigarrow \textit{buffer}[0] = 0
\end{aligned}$$

7.5 Configuration

Chapter 8

SPMC Lockless Queue

Part III

Language Reference

Chapter 9

Data Structure

Like other languages, TLA+ provides its data structure. I assume the readers are already familiar with common data structure, and this chapter will only focus on the TLA+ language semantics.

9.1 Set

This is the most common data structure used in TLA+ spec. The following is a few examples on how a set can be used:

$a \triangleq$	$\{0, 1, 2\}$	
$b \triangleq$	$\{2, 3, 4\}$	
$c \triangleq$	$a \cup b$	$\{0, 1, 2, 3, 4\}$
$d \triangleq$	$a \cap b$	$\{2\}$
$e \triangleq$	$\exists x \in c : x > 3$	TRUE - because 4 in c is bigger than 3
$f \triangleq$	$\exists x \in c : x > 5$	FALSE - nothing in c is bigger than 5
$g \triangleq$	$\forall x \in c : x < 3$	FALSE - not all elements in c are smaller than 3
$h \triangleq$	$\forall x \in c : x < 5$	TRUE - all elements in c are smaller than 5
$i \triangleq$	$\{x \in c : x < 3\}$	$\{0, 1, 2\}$ - all elements less than 3
$j \triangleq$	$Cardinality(c)$	5 - the number of elements in c
$k \triangleq$	$c \setminus d$	$\{0, 1, 3, 4\}$ - c subtracts d

9.2 Tuple

$A \triangleq$	$\langle 0, 1, 2 \rangle$	
$B \triangleq$	$\langle 2, 3, 4 \rangle$	
$C \triangleq$	$A \circ B$	tuple: 0, 1, 2, 2, 3, 4
$D \triangleq$	$Len(C)$	6
$E \triangleq$	$\forall x \in 1 \dots Len(C) : C[x] \neq 10$	TRUE - every C[x] is not 10
		First tuple element is at index 1 (not 0)
$F \triangleq$	$\exists x \in 1 \dots Len(C) : C[x] = 2$	TRUE - there exists a C[x] that is 2

$$G \triangleq \{x \in 1 \dots \text{Len}(C) : C[x] = 2\} \quad \{3, 4\} \text{ - when index is 3 or 4, } C[x] = 2$$

Chapter 10

Idiom

Choose a x in set S such that for every y in S x is smaller than y . Finding minimum in set:

$$\text{Min}(S) \triangleq \text{CHOOSE } x \in S : \forall y \in S : x \leq y$$

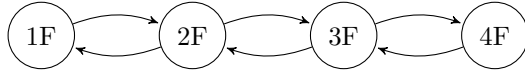
`messages` is an unordered map with untyped key and integer value:

$$\text{messages} = [m \in \{\} \mapsto 0]$$

Chapter 11

Fairness and Liveness

For rigorous definition and proof, please refer to (TODO: citations). This chapter focus on the application aspect of liveness and fairness and define an elevator spec that goes up and down.



11.1 Liveness

Consider the following elevator *Spec*:

	MODULE <i>elevator</i>	
EXTENDS	<i>Integers</i>	
VARIABLES	<i>a</i>	
<i>vars</i>	$\triangleq \langle a \rangle$	
<i>TOP</i>	$\triangleq 4$	
<i>BOTTOM</i>	$\triangleq 1$	
<i>Init</i>	\triangleq	
	$\wedge a = BOTTOM$	
<i>Up</i>	\triangleq	
	$\wedge a \neq TOP$	
	$\wedge a' = a + 1$	
<i>Down</i>	\triangleq	
	$\wedge a \neq BOTTOM$	
	$\wedge a' = a - 1$	
<i>Spec</i>	\triangleq	
	$\wedge Init$	
	$\wedge \Box[Up \vee Down]_a$	

The building has a set of floors and the elevator can go either up or down. The elevator keeps going up until it's the top floor, or keep going down until

it's the bottom floor. TLC will pass the *Spec* as is.

Let's introduce a liveness property. The elevator should always at least go to the second floor:

$$\begin{aligned} \text{Liveness} &\triangleq \\ &\wedge a = 1 \leadsto a = 2 \end{aligned}$$

Running the *Spec* against TLC will report a violation:

```
Error: Temporal properties were violated.
Error: The following behavior constitutes a counter-example:
State 1: <Initial predicate>
a = 1
State 2: Stuttering
```

Since the *Spec* permits *stuttering*, the state machine is allowed to perpetually stay on 1F and *never* go to 2F. This can be fixed by introducing fairness description.

11.2 Weak Fairness

Weak fairness is defined as:

$$\Diamond \Box (ENABLED \langle A \rangle_v) \Rightarrow \Box \Diamond \langle A \rangle_v \quad (11.1)$$

$ENABLED \langle A \rangle$ represents *conditions required* for action A. The above translates to: if conditions required for action A to occur is *eventually always* true, then action A will *always eventually* happen.

Without weak fairness defined, the elevator may *stutter* at 1F and never go to 2F. Weak fairness states that if the conditions of an action is *eventually always* true (ie. elevator decides to stay on 1F but *can* go up), the elevator *always eventually* go up.

$$\begin{aligned} \text{Spec} &\triangleq \\ &\wedge \text{Init} \\ &\wedge \Box [\text{Down} \vee \text{Up}]_a \\ &\wedge \text{WF}_a(\text{Down}) \\ &\wedge \text{WF}_a(\text{Up}) \end{aligned}$$

Running the spec against TLC passes again. What if we want to verify the elevator eventually always goes to the top, not just to 2F? Let's modify the Liveness property again:

$$\text{Liveness} \triangleq$$

$$\wedge a = BOTTOM \leadsto a = TOP$$

TLC now reports the following violation:

Error: Temporal properties were violated.

Error: The following behavior constitutes a counter-example:

State 1: <Initial predicate>

a = 1

State 2: <Up line 10, col 5 to line 11, col 17 of module elevator>

a = 2

Back to state 1: <Down line 13, col 5 to line 14, col 17 of module elevator>

TLC identified a case where the elevator is perpetually stuck going between 1F and 2F, but never go to 3F. Weak fairness is no longer enough, because the the elevator is not stuck on 2F repeatedly, but stuck going between 1F and 2F. This is where we need strong fairness.

11.3 Strong Fairness

Strong fairness is defined as:

$$\Box\Diamond(ENABLED\langle A \rangle_v) \Rightarrow \Box\Diamond\langle A \rangle_v \quad (11.2)$$

The difference between weak and strong fairness is the *eventually always* vs. *always eventually*.

In weak fairness, once the state machine is stuck in a state forever, the state machine always transition to a possible next state permitted by the *Spec* (eg. if the elevator is stuck on 1F but can go to 2F, it will). With strong fairness, the elevator doesn't need to be stuck on 2F to go to 3F. If the elevator *always eventually* makes it to 2F, it *always eventually* go to 3F.

Intuitively we are tempted to enable strong fairness like so:

$$\begin{aligned} Spec &\triangleq \\ &\wedge Init \\ &\wedge \Box[Up \vee Down]_a \\ &\wedge WF_a(Down) \\ &\wedge SF_a(UP) \end{aligned}$$

However, TLC *still* reports the same violation. What's going on?

If we take a closer look at the enabling condition for *Up*, it only requires current floor to be not the *top floor*. When the elevator is stuck in a loop going Up and Down between 1F and 2F indefinitely, strong fairness for Up is *already satisfied*. What we really want is strong fairness on *Up* for *every floor*, instead

of *any floor except top floor*. So if elevator makes to 2F once, it will *always eventually* go to 3F. If elevator makes to 3F once, it will *always eventually* go to 4F, so on and so forth. The following is the change required:

$$\begin{aligned}
 Spec &\triangleq \\
 &\wedge Init \\
 &\wedge \Box[Up \vee Down]_a \\
 &\wedge WF_a(Down) \\
 &\wedge \forall f \in BOTTOM \dots TOP - 1 : \\
 &\quad \wedge WF_a(Up \wedge f = a)
 \end{aligned}$$

Once again with this change TLC will pass.

Chapter 12

Abstraction Guideline

Chapter 13

Reference

Bibliography

- [1] Srikumar Subramanian <https://sriku.org/posts/fairness-in-tlaplus/>, 2015
- [2] Richard M. Murray, Nok Wongpiromsarn *Linear Temporal Logic, Lecture 3*, 2012
- [3] <https://www.backblaze.com/blog/cloud-storage-durability/>
- [4] <https://raft.github.io/raft.pdf>
- [5] <https://github.com/ongardie/raft.tla>

Chapter 14

Nano

TODO: add this to reference

https://content.nano.org/whitepaper/Nano_Whitepaper_en.pdf

14.1 Requirement

14.2 Spec

$Init \triangleq$
 $\wedge lastHash = NoHash$
 $\wedge distributedLedger = [n \in Node \mapsto [h \in Hash \mapsto NoBlock]]$
 $\wedge received = [n \in Node \mapsto \{\}]$

- Every node is a ledger in this system, initialized to NoBlock
- Every node's received set is initialized to empty set

$Next \triangleq$
 $\vee \exists account \in PrivateKey : CreateGenesisBlock(account)$
 $\vee \exists node \in Node : CreateBlock(node)$
 $\vee \exists node \in Node : ProcessBlock(node)$

PrivateKey represents the identity of the account, create the genesis block for every account. Let us look at how a genesis block is created:

$HashOf(block) \triangleq$
IF $\exists hash \in Hash : hashFunction[hash] = block$
THEN CHOOSE $hash \in Hash : hashFunction[hash] = block$
ELSE CHOOSE $hash \in Hash : hashFunction[hash] = N!NoBlock$

$CalculateHashImpl(block, oldLastHash, newLastHash) \triangleq$
LET $hash \triangleq HashOf(block)$ IN

$$\begin{aligned}
& \wedge \text{newLastHash} = \text{hash} \\
& \wedge \text{hashFunction}' = [\text{hashFunction} \text{ EXCEPT } ![\text{hash}] = \text{block}] \\
\text{CreateGenesisBlock}(\text{privateKey}) & \triangleq \\
\text{LET} & \\
& \text{publicKey} \triangleq \text{KeyPair}[\text{privateKey}] \\
& \text{genesisBlock} \triangleq \\
& \quad [\text{type} \mapsto \text{"genesis"}, \\
& \quad \text{account} \mapsto \text{publicKey}, \\
& \quad \text{balance} \mapsto \text{GenesisBalance}] \\
\text{IN} & \\
& \wedge \neg \text{GenesisBlockExists} \\
& \wedge \text{CalculateHash}(\text{genesisBlock}, \text{lastHash}, \text{lastHash}') \\
& \wedge \text{distributedLedger}' = \\
& \quad \text{LET } \text{signedGenesisBlock} \triangleq \\
& \quad \quad [\text{block} \mapsto \text{genesisBlock}, \\
& \quad \quad \text{signature} \mapsto \text{SignHash}(\text{lastHash}', \text{privateKey})] \\
& \text{IN} \\
& \quad [n \in \text{Node} \mapsto \\
& \quad \quad [\text{distributedLedger}[n] \text{ EXCEPT } \\
& \quad \quad \quad ![\text{lastHash}'] = \text{signedGenesisBlock}]] \\
& \wedge \text{UNCHANGED } \text{received}
\end{aligned}$$

Every account maintains its own chain of blocks. The first block in the account chain is the genesis block. The genesis block contains the type, account name, and genesis balance. The genesis block is then hashed and signed.

TODO: add this to reference

https://content.nano.org/whitepaper/Nano_Whitepaper-en.pdf

Chapter 15

Debugging

TODO: debug using -dump TODO: check for run away state TODO: look for low value states to trim