MODULE *ceph*

This is a specification of the paxos algorithm implemented in Ceph. The specification is based on the following source file: https://github.com/ceph/ceph/blob/master/src/mon/Paxos.cc

For a more detailed overview of the specification: https://github.com/afonsonf/ceph-consensus-spec

EXTENDS *Integers*, *FiniteSets*, *Sequences*, *TLC*

Utils

Max element from a set.
@type: *Set*(*Int*) ⇒ *Int*;
$Max(S) \triangleq \text{CHOOSE } x \in S : \forall\, y \in S : x \geq y$

Min element from a set.
@type: *Set*(*Int*) ⇒ *Int*;
$Min(S) \triangleq \text{CHOOSE } x \in S : \forall\, y \in S : x \leq y$

Set of monitors to a sequence.
RECURSIVE $SetToSeq(\_)$
@type: *Set*(*MONITOR*) ⇒ *Seq*(*MONITOR*);

$$
\begin{aligned}
&SetToSeq(S) \triangleq \\
&\quad \text{IF } S = \{\} \text{ THEN } \langle\rangle \\
&\qquad \text{ELSE LET } x \triangleq \text{CHOOSE } x \in S : \text{TRUE} \\
&\qquad\quad \text{IN } \langle x \rangle \circ SetToSeq(S \setminus \{x\})
\end{aligned}
$$

Constants

Set of *Monitors*.
CONSTANTS @type: *Set*(*MONITOR*); *Monitors*

Sequence of monitors.
@type: *Seq*(*MONITOR*);
$MonitorsSeq \triangleq TLCEval(SetToSeq(Monitors))$

Number of monitors.
@type: *Int*;
$MonitorsLen \triangleq TLCEval(Len(MonitorsSeq))$

Rank predicate, used as a monitor identifier.
@type: *MONITOR* ⇒ *Int*;
$rank(mon) \triangleq \text{CHOOSE } i \in 1 \,..\, MonitorsLen : MonitorsSeq[i] = mon$

Set of possible values.
CONSTANTS @type: *Set*(*VALUE*); *Value_set*

Predicate used in the cfg file to define the symmetry set.

Workaround for typechecker:
@*typeAlias*: $MONITOR = T$;
@*typeAlias*: $VALUE = T$;
$SYMM \triangleq Permutations(Monitors) \cup Permutations(Value\_set)$

Reserved value.
CONSTANTS @type: VALUE; Nil

Paxos states.
CONSTANTS @type: *STATE_NAME*; $STATE\_RECOVERING$, @type: *STATE_NAME*; $STATE\_ACTIVE$,
@type: *STATE_NAME*; $STATE\_UPDATING$,
@type: *STATE_NAME*; $STATE\_UPDATING\_PREVIOUS$,
@type: *STATE_NAME*; $STATE\_WRITING$,
@type: *STATE_NAME*; $STATE\_WRITING\_PREVIOUS$,
@type: *STATE_NAME*; $STATE\_REFRESH$, @type: *STATE_NAME*; $STATE\_SHUTDOWN$

Paxos auxiliary phase states.
They are used to force some sequence of steps.
CONSTANTS @type: *PHASE_NAME*; $PHASE\_ELECTION$,
@type: *PHASE_NAME*; $PHASE\_SEND\_COLLECT$,
@type: *PHASE_NAME*; $PHASE\_COLLECT$,
@type: *PHASE_NAME*; $PHASE\_LEASE$, @type: *PHASE_NAME*; $PHASE\_LEASE\_DONE$,
@type: *PHASE_NAME*; $PHASE\_BEGIN$, @type: *PHASE_NAME*; $PHASE\_COMMIT$

Paxos message types.
CONSTANTS @type: *MESSAGE_OP*; $OP\_COLLECT$, @type: *MESSAGE_OP*; OP_LAST,
@type: *MESSAGE_OP*; OP_BEGIN, @type: *MESSAGE_OP*; OP_ACCEPT,
@type: *MESSAGE_OP*; OP_COMMIT,
@type: *MESSAGE_OP*; OP_LEASE, @type: *MESSAGE_OP*; OP_LEASE_ACK

Global variables

Integer representing the current epoch. When epoch is odd trigger an election.
VARIABLE @type: *Int*; $epoch$

Store messages waiting to be handled.
VARIABLE @type: $MONITOR \to (MONITOR \to Seq(MESSAGE))$; $messages$

Stores history of messages. Can be useful to find specific states.
VARIABLE @type: $Set(MESSAGE)$; $message\_history$

Stores if a monitor is up or down. All available monitors, in a given epoch, are part of the quorum.
VARIABLE @type: $MONITOR \to Bool$; $quorum$

Size of the current quorum.
VARIABLE @type: *Int*; $quorum\_sz$

State variables

A function that stores the current leader. $isLeader[mon]$ is True iff mon is a leader, else False.
VARIABLE @type: $MONITOR \to Bool$; $isLeader$

A function that stores the state of each monitor.
VARIABLE @type: $MONITOR \to STATE\_NAME$; $state$

A function that stores the phase of each monitor.
VARIABLE @type: $MONITOR \to PHASE\_NAME$; $phase$

Restart variables

A function that stores, for each monitor, a proposal number when the commit phase starts.
This proposal number can be retrieved after a monitor crashes and restarts.
VARIABLE @type: $MONITOR \to PN$; $pending\_pn$

A function that stores, for each monitor, a value version when the commit phase starts.
This value version can be retrieved after a monitor crashes and restarts.
VARIABLE @type: $MONITOR \to VALUE\_VERSION$; $pending\_v$

A function that stores, for each monitor, the best uncommitted pn received in the collect phase.
VARIABLE @type: $MONITOR \to PN$; $uncommitted\_pn$

A function that stores, for each monitor, the best uncommitted value version received in the collect phase.
VARIABLE @type: $MONITOR \to VALUE\_VERSION$; $uncommitted\_v$

A function that stores, for each monitor, the best uncommitted value received in the collect phase.
VARIABLE @type: $MONITOR \to VALUE$; $uncommitted\_value$

Data variables

A function that stores, for each monitor, the store where the transactions are applied.
In this specification, a transaction represents changing the value in the store.
VARIABLE @type: $MONITOR \to VALUE$; $monitor\_store$

A function that stores the transaction log of each monitor.
VARIABLE @type: $MONITOR \to (VALUE\_VERSION \to VALUE)$; $values$

A function that stores the last proposal number accepted by each monitor.
VARIABLE @type: $MONITOR \to PN$; $accepted\_pn$

A function that stores the first value version committed by each monitor.
VARIABLE @type: $MONITOR \to VALUE\_VERSION$; $first\_committed$

A function that stores the last value version committed by each monitor.
VARIABLE @type: $MONITOR \to VALUE\_VERSION$; $last\_committed$

Collect phase variables

A function that stores the number of peers that accepted a collect request.

VARIABLE @type: $MONITOR \to Int$; num_last

Used by leader when receiving responses in collect phase.
VARIABLE @type: $MONITOR \to (MONITOR \to VALUE\_VERSION)$; $peer\_first\_committed$
VARIABLE @type: $MONITOR \to (MONITOR \to VALUE\_VERSION)$; $peer\_last\_committed$

Lease phase variables

A function that stores, for each monitor, which of the peers have acked the lease request.
VARIABLE @type: $MONITOR \to (MONITOR \to Bool)$; $acked\_lease$

Commit phase variables

A function that stores, for each monitor, the value proposed by a client.
VARIABLE @type: $MONITOR \to VALUE$; $pending\_proposal$

A function that stores, for each monitor, the value to be committed in the begin phase.
VARIABLE @type: $MONITOR \to VALUE$; new_value

A function that stores, for each monitor, which of the peers have acked the begin request.
VARIABLE @type: $MONITOR \to (MONITOR \to Bool)$; $accepted$

Debug variables

Variables to help debug a behavior.
step is the diameter of a behavior/path.
step_name the current predicate being called.
VARIABLE @type: Str; $step\_name$

Variables initialization

@*typeAlias*: $VALUE\_VERSION = Int$;
@*typeAlias*: $PN = Int$;

$$
\begin{aligned}
global\_vars &\triangleq \langle epoch, messages, message\_history, quorum, quorum\_sz\rangle \\
state\_vars &\triangleq \langle isLeader, state, phase\rangle \\
restart\_vars &\triangleq \langle pending\_pn, pending\_v, uncommitted\_pn, uncommitted\_v, uncommitted\_value\rangle \\
data\_vars &\triangleq \langle monitor\_store, values, accepted\_pn, first\_committed, last\_committed\rangle \\
collect\_vars &\triangleq \langle num\_last, peer\_first\_committed, peer\_last\_committed\rangle \\
lease\_vars &\triangleq acked\_lease \\
commit\_vars &\triangleq \langle pending\_proposal, new\_value, accepted\rangle
\end{aligned}
$$

$$
\begin{aligned}
vars \triangleq \langle &global\_vars, state\_vars, restart\_vars, data\_vars, collect\_vars, \\
&lease\_vars, commit\_vars\rangle
\end{aligned}
$$

$$
\begin{aligned}
&Init\_global\_vars \triangleq \\
&\quad \land epoch = 1 \\
&\quad \land messages = [mon1 \in Monitors \mapsto [mon2 \in Monitors \mapsto \langle\rangle]]
\end{aligned}
$$

$\land\ message\_history = \{\}$
$\land\ quorum = [mon \in Monitors \mapsto \text{TRUE}]$
$\land\ quorum\_sz = MonitorsLen$

$Init\_state\_vars \triangleq$
$\land\ isLeader = [mon \in Monitors \mapsto \text{FALSE}]$
$\land\ state\ = [mon \in Monitors \mapsto STATE\_RECOVERING]$
$\land\ phase = [mon \in Monitors \mapsto PHASE\_ELECTION]$

$Init\_restart\_vars \triangleq$
$\land\ pending\_pn = [mon \in Monitors \mapsto 0]$
$\land\ pending\_v = [mon \in Monitors \mapsto 0]$
$\land\ uncommitted\_pn = [mon \in Monitors \mapsto 0]$
$\land\ uncommitted\_v = [mon \in Monitors \mapsto 0]$
$\land\ uncommitted\_value = [mon \in Monitors \mapsto Nil]$

$Init\_data\_vars \triangleq$
$\land\ monitor\_store\ = [mon \in Monitors \mapsto Nil]$
$\land\ values = [mon \in Monitors \mapsto [version \in \{\} \mapsto Nil]]$
$\land\ accepted\_pn = [mon \in Monitors \mapsto 0]$
$\land\ first\_committed = [mon \in Monitors \mapsto 0]$
$\land\ last\_committed = [mon \in Monitors \mapsto 0]$

$Init\_collect\_vars \triangleq$
$\land\ num\_last = [mon \in Monitors \mapsto 0]$
$\land\ peer\_first\_committed = [mon1 \in Monitors \mapsto [mon2 \in Monitors \mapsto\ -1]]$
$\land\ peer\_last\_committed = [mon1 \in Monitors \mapsto [mon2 \in Monitors \mapsto\ -1]]$

$Init\_lease\_vars \triangleq$
$\land\ acked\_lease = [mon1 \in Monitors \mapsto [mon2 \in Monitors \mapsto \text{FALSE}]]$

$Init\_commit\_vars \triangleq$
$\land\ pending\_proposal\ \ = [mon \in Monitors \mapsto Nil]$
$\land\ new\_value = [mon \in Monitors \mapsto Nil]$
$\land\ accepted = [mon1\ \ \in Monitors \mapsto [mon2 \in Monitors \mapsto \text{FALSE}]]$

$Init \triangleq$
$\land\ Init\_global\_vars$
$\land\ Init\_state\_vars$
$\land\ Init\_restart\_vars$
$\land\ Init\_data\_vars$
$\land\ Init\_collect\_vars$
$\land\ Init\_lease\_vars$
$\land\ Init\_commit\_vars$
$\land\ step\_name =$ "init"

Message manipulation

@*typeAlias*: *MESSAGE* = [type: *MESSAGE_OP*, from: *MONITOR*, *dest*: *MONITOR*, *first_committed*: *VALUE_VERSION*, *last_committed*: *VALUE_VERSION*, values: (*VALUE_VERSION* → *VALUE*), *uncommitted_pn*: *PN*, *pn*: *PN*];
@*typeAlias*: *MESSAGE_QUEUE* = *MONITOR* → (*MONITOR* → *Seq*(*MESSAGE*));

Note: Variable *message_history* has impact in performace, update only when debugging.

Converts a set with at most one element to a sequence.
@type: *Set*(*MESSAGE*) ⇒ *Seq*(*MESSAGE*);
$SingleMessageSetToSeq(S) \triangleq$
IF $\exists\, elem \in S$: TRUE THEN LET $elem \triangleq$ CHOOSE $x \in S$: TRUE
IN $\langle elem \rangle$
ELSE $\langle\rangle$

Add message m to the network *msgs*.
@type: (MESSAGE, *MESSAGE_QUEUE*) ⇒ *MESSAGE_QUEUE*;
$WithMessage(m,\ msgs) \triangleq$
$[msgs$ EXCEPT $![m.from] =$
$[msgs[m.from]$ EXCEPT $![m.dest] = Append(msgs[m.from][m.dest],\ m)]]$

Remove message m from the network *msgs*.
@type: (MESSAGE, *MESSAGE_QUEUE*) ⇒ *MESSAGE_QUEUE*;
$WithoutMessage(m,\ msgs) \triangleq$
$[msgs$ EXCEPT $![m.from] =$
$[msgs[m.from]$ EXCEPT $![m.dest] = Tail(msgs[m.from][m.dest])]]$

Adds the message m to the network.
Variables changed: messages, *message_history*.
@type: *MESSAGE* ⇒ *Bool*;
$Send(m) \triangleq$
$\wedge\ messages' = WithMessage(m,\ messages)$
$\wedge\ message\_history' = message\_history \cup \{m\}$
$\wedge$ UNCHANGED *message_history*

Adds a set of messages to the network.
Variables changed: messages, *message_history*.
@type: (MONITOR, *Set*(*MESSAGE*)) ⇒ *Bool*;
$Send\_set(from,\ m\_set) \triangleq$
$\wedge\ messages' = [messages$ EXCEPT $![from] =$
$[mon \in Monitors \mapsto$
$messages[from][mon] \circ SingleMessageSetToSeq(\{m \in m\_set : m.dest = mon\})]]$
$\wedge\ message\_history' = message\_history \cup m\_set$
$\wedge$ UNCHANGED *message_history*

Removes the request from network and adds the response.
Variables changed: messages, *message_history*.
@type: (MESSAGE, MESSAGE) ⇒ *Bool*;

$Reply(response,\ request) \triangleq$
$\land\ messages' = WithoutMessage(request,\ WithMessage(response,\ messages))$
$\land\ message\_history' = message\_history \cup \{response\}$
$\land$ UNCHANGED $message\_history$

Removes the request from network and adds a set of messages.
Variables changed: messages, *message_history*.
@type: (MONITOR, $Set(MESSAGE)$, $MESSAGE$) $\Rightarrow Bool$;
$Reply\_set(from,\ response\_set,\ request) \triangleq$
$\land$ LET $msgs \triangleq WithoutMessage(request,\ messages)$
IN $messages' = [msgs$ EXCEPT $![from] =$
$[mon \in Monitors \mapsto$
$msgs[from][mon] \circ SingleMessageSetToSeq(\{m \in response\_set : m.dest = mon\})]]$
$\land\ message\_history' = message\_history \cup response\_set$
$\land$ UNCHANGED $message\_history$

Removes message m from the network.
Variables changed: messages, *message_history*.
@type: $MESSAGE \Rightarrow Bool$;
$Discard(m) \triangleq$
$\land\ messages' = WithoutMessage(m,\ messages)$
$\land$ UNCHANGED $message\_history$

Helper predicates

Computes a new unique proposal number for a given monitor.

Version A - Equal to the one in the source.
This version breaks the symmetry of the monitor set.
Example: $oldpn = 305$, $rank(mon) = 5$, $newpn = 405$.
@type: (MONITOR, Int) $\Rightarrow Int$;
$get\_new\_proposal\_number(mon,\ oldpn) \triangleq ((oldpn \div 100) + 1) * 100 + rank(mon)$

Version $B-$ Adapted to not break symmetry.
Example: $oldpn = 302$, $epoch = 4$, $newpn = 404$.
@type: (MONITOR, Int) $\Rightarrow Int$;
$get\_new\_proposal\_number(mon,\ oldpn) \triangleq ((oldpn \div 100) + 1) * 100 + epoch$

Clear the variable *peer_first_committed*.
Variables changed: *peer_first_committed*.
@type: $MONITOR \Rightarrow Bool$;
$clear\_peer\_first\_committed(mon) \triangleq$
$peer\_first\_committed' = [peer\_first\_committed$ EXCEPT $![mon] = [m \in Monitors \mapsto -1]]$

Clear the variable *peer_last_committed*.
Variables changed: *peer_last_committed*.

```
@type: MONITOR ⇒ Bool;
clear_peer_last_committed(mon) ≜
    peer_last_committed' = [peer_last_committed EXCEPT ![mon] = [m ∈ Monitors ↦ −1]]

Store peer values and update first_committed, last_committed and monitor_store accordingly.
Variables changed: values, first_committed, last_committed, monitor_store.
@type: (MONITOR, MESSAGE) ⇒ Bool;
store_state(mon, msg) ≜
    Choose peer values from mon last committed +1 to peer last committed.
    ∧ LET logs ≜ (DOMAIN msg.values) ∩ (last_committed[mon] + 1 .. msg.last_committed)
      IN  ∧ values' = [values EXCEPT ![mon] =
                [i ∈ DOMAIN values[mon] ∪ logs ↦
                  IF i ∈ logs  priority to peer committed values over mon uncommitted
                   THEN msg.values[i]
                   ELSE values[mon][i]]]
          Update last committed and first committed.
          ∧ last_committed' = [last_committed EXCEPT ![mon] = Max(logs ∪ {last_committed[mon]})]
          ∧ IF logs ≠ {} ∧ first_committed[mon] = 0
            THEN first_committed' =
                        [first_committed EXCEPT ![mon] = Min(logs)]
            ELSE first_committed' =
                        [first_committed EXCEPT ![mon] = Min(logs ∪ {first_committed[mon]})]
    Update monitor store.
    ∧ IF last_committed'[mon] = 0
      THEN UNCHANGED monitor_store
      ELSE monitor_store' = [monitor_store EXCEPT ![mon] = values'[mon][last_committed'[mon]]]

Check if uncommitted value version is still valid, else reset it.
Variables changed: uncommitted_pn, uncommitted_v, uncommitted_value.
@type: MONITOR ⇒ Bool;
check_and_correct_uncommitted(mon) ≜
    IF uncommitted_v[mon] ≤ last_committed'[mon]
     THEN ∧ uncommitted_v' = [uncommitted_v EXCEPT ![mon] = 0]
          ∧ uncommitted_pn' = [uncommitted_pn EXCEPT ![mon] = 0]
          ∧ uncommitted_value' = [uncommitted_value EXCEPT ![mon] = Nil]
     ELSE UNCHANGED ⟨uncommitted_pn, uncommitted_v, uncommitted_value⟩

Trigger new election by incrementing epoch.
Variables changed: epoch.
@type: Bool;
bootstrap ≜
    ∧ epoch' = epoch + 1
```

Lease phase predicates

Changes *mon* state to *STATE_ACTIVE*.
Variables changed: state.
@type: *MONITOR* ⇒ *Bool*;

$$
\begin{array}{l}
finish\_round(mon) \triangleq \\
\quad \land\ isLeader[mon] = \text{TRUE} \\
\quad \land\ state' = [state \text{ EXCEPT } ![mon] = STATE\_ACTIVE]
\end{array}
$$

Resets the variable acked lease and send lease messages to peers.
Variables changed: *acked_lease*, messages, *message_history*, phase.
@type: *MONITOR* ⇒ *Bool*;

$$
\begin{array}{l}
extend\_lease(mon) \triangleq \\
\quad \land\ isLeader[mon] = \text{TRUE} \\
\quad \land\ acked\_lease' = [acked\_lease \text{ EXCEPT } ![mon] = \\
\qquad [m \in Monitors \mapsto \text{IF } m = mon \text{ THEN TRUE ELSE FALSE}]] \\
\quad \land\ Send\_set(mon, \\
\qquad \{[type \mapsto OP\_LEASE, \\
\qquad\ from \mapsto mon, \\
\qquad\ dest \mapsto dest, \\
\qquad\ last\_committed \mapsto last\_committed[mon]] : dest \in \{m \in Monitors \setminus \{mon\} : quorum[m]\} \\
\qquad \}) \\
\quad \land\ phase' = [phase \text{ EXCEPT } ![mon] = PHASE\_LEASE]
\end{array}
$$

Handle a lease message. The peon changes his state and replies with a lease ack message.
The reply is commented because the lease ack is only used to check if all peers are up.
In the model this is done by "randomly" triggering the predicate *Timeout*. In this way, the search space is reduced.
Variables changed: messages, *message_history*, state.
@type: (MONITOR, MESSAGE) ⇒ *Bool*;

$$
\begin{array}{l}
handle\_lease(mon,\ msg) \triangleq \\
\quad \land\ \text{discard if not peon or peon is behind} \\
\qquad \text{IF } \lor\ isLeader[mon] = \text{TRUE} \\
\qquad\quad \lor\ last\_committed[mon] \neq msg.last\_committed \\
\qquad \text{THEN } \land\ Discard(msg) \\
\qquad\quad\ \ \land \text{ UNCHANGED } state \\
\qquad \text{ELSE } \land\ state' = [state \text{ EXCEPT } ![mon] = STATE\_ACTIVE] \\
\qquad\quad \land\ Reply([type \mapsto OP\_LEASE\_ACK, \\
\qquad\qquad from \mapsto mon, \\
\qquad\qquad dest \mapsto msg.from, \\
\qquad\qquad first\_committed \mapsto first\_committed[mon], \\
\qquad\qquad last\_committed \mapsto last\_committed[mon]], msg) \\
\qquad\quad\ \ \land\ Discard(msg) \\
\quad \land \text{ UNCHANGED } \langle epoch,\ quorum,\ quorum\_sz,\ isLeader,\ phase \rangle \\
\quad \land \text{ UNCHANGED } \langle restart\_vars,\ data\_vars,\ collect\_vars,\ lease\_vars,\ commit\_vars \rangle
\end{array}
$$

Handle a lease ack message. The leader updates the *acked_lease* variable.
Because the *lease_ack* messages are not sent, this predicate is never called.
The reasoning for this is given in *handle_lease* comment.

Variables changed: *acked_lease*, messages, *message_history*.
@type: (MONITOR, MESSAGE) ⇒ *Bool*;

$handle\_lease\_ack(mon,\ msg) \triangleq$

$\land\ phase[mon] = PHASE\_LEASE$
$\land\ acked\_lease' = [acked\_lease \text{ EXCEPT } ![mon] =$
$\quad [acked\_lease[mon] \text{ EXCEPT } ![msg.from] = \text{TRUE}]]$
$\land\ Discard(msg)$
$\land \text{ UNCHANGED } \langle epoch,\ quorum,\ quorum\_sz \rangle$
$\land \text{ UNCHANGED } \langle state\_vars,\ restart\_vars,\ data\_vars,\ collect\_vars,\ commit\_vars \rangle$

Predicate that is called when all peers ack the lease. The phase is changed to prevent loops.
Because the *lease_ack* messages are not sent, this predicate is never called.
The reasoning for this is given in *handle_lease* comment.
Variables changed: phase.
@type: *MONITOR* ⇒ *Bool*;

$post\_lease\_ack(mon) \triangleq$

$\land\ phase[mon] = PHASE\_LEASE$
$\land\ phase' = [phase \text{ EXCEPT } ![mon] = PHASE\_LEASE\_DONE]$
$\land\ \forall\, m \in Monitors : quorum[m] \Rightarrow acked\_lease[mon][m] = \text{TRUE}$
$\land \text{ UNCHANGED } \langle isLeader,\ state \rangle$
$\land \text{ UNCHANGED } \langle global\_vars,\ restart\_vars,\ data\_vars,\ collect\_vars,$
$\qquad\qquad lease\_vars,\ commit\_vars \rangle$

Commit phase predicates

Start a commit phase by the leader. The variable *new_value* is assigned. Send begin messages to the peers.
The new value is stored in values and *pending_pn* is assigned in order for the leader to be
able to recover from a crash.
Variables changed: accepted, *new_value*, phase, messages, *message_history*, values, *pending_pn*, *pending_v*.
@type: (MONITOR, VALUE) ⇒ *Bool*;

$begin(mon,\ v) \triangleq$

$\land\ isLeader[mon] = \text{TRUE}$
$\land\ \lor\ state'[mon] = STATE\_UPDATING$
$\quad \lor\ state'[mon] = STATE\_UPDATING\_PREVIOUS$
$\land\ quorum\_sz = 1 \lor num\_last[mon] > MonitorsLen \div 2$
$\land\ new\_value[mon] = Nil$
$\land\ accepted' = [accepted \text{ EXCEPT } ![mon] =$
$\quad [m \in Monitors \mapsto \text{IF } m = mon \text{ THEN TRUE ELSE FALSE}]]$
$\land\ new\_value' = [new\_value \text{ EXCEPT } ![mon] = v]$
$\land\ phase' = [phase \text{ EXCEPT } ![mon] = PHASE\_BEGIN]$
$\land\ values' = [values \text{ EXCEPT } ![mon] =$
$\quad ((last\_committed[mon] + 1) :> new\_value'[mon]) \mathbin{@@} values[mon]]$
$\land\ Send\_set(mon,$
$\quad \{[type \mapsto OP\_BEGIN,$
$\quad\ from \mapsto mon,$

$dest \mapsto dest,$
$last\_committed \mapsto last\_committed[mon],$
$values \mapsto values'[mon],$
$pn \mapsto accepted\_pn[mon]] : dest \in \{m \in Monitors \setminus \{mon\} : quorum[m]\}$
$\})$
$\land\ pending\_pn' = [pending\_pn \text{ EXCEPT } ![mon] = accepted\_pn[mon]]$
$\land\ pending\_v' = [pending\_v \text{ EXCEPT } ![mon] = last\_committed[mon] + 1]$

Handle a begin message. The monitor will accept if the proposal number in the message is greater or equal than the one he accepted.
Similar to what happens in begin, values and *pending_pn* are assigned in order for the monitor to recover in case of a crash.
Variables changed: messages, *message_history*, state, values, *pending_pn*, *pending_v*.
@type: (MONITOR, MESSAGE) ⇒ *Bool*;

$handle\_begin(mon,\ msg) \triangleq$
$\land\ isLeader[mon] = \text{FALSE}$
$\land\ \text{IF } msg.pn < accepted\_pn[mon]$
THEN
$\land\ Discard(msg)$
$\land\ \text{UNCHANGED } \langle state,\ values,\ pending\_pn,\ pending\_v\rangle$
ELSE
$\land\ msg.pn = accepted\_pn[mon]$
$\land\ msg.last\_committed = last\_committed[mon]$

assign $values[mon][last\_committed[mon] + 1]$
$\land\ values' = [values \text{ EXCEPT } ![mon] =$
$((last\_committed[mon] + 1) :> msg.values[last\_committed[mon] + 1]) @@ values[mon]]$

$\land\ state' = [state \text{ EXCEPT } ![mon] = STATE\_UPDATING]$
$\land\ pending\_pn' = [pending\_pn \text{ EXCEPT } ![mon] = accepted\_pn[mon]]$
$\land\ pending\_v' = [pending\_v \text{ EXCEPT } ![mon] = last\_committed[mon] + 1]$
$\land\ Reply([type \mapsto OP\_ACCEPT,$
$from \mapsto mon,$
$dest \mapsto msg.from,$
$last\_committed \mapsto last\_committed[mon],$
$pn \mapsto accepted\_pn[mon]],\ msg)$
$\land\ \text{UNCHANGED } \langle epoch,\ quorum,\ quorum\_sz,\ isLeader,\ phase,\ monitor\_store,$
$accepted\_pn,\ first\_committed,\ last\_committed,\ uncommitted\_pn,$
$uncommitted\_v,\ uncommitted\_value\rangle$
$\land\ \text{UNCHANGED } \langle collect\_vars,\ lease\_vars,\ commit\_vars\rangle$

Handle an accept message. If the leader receives a positive response from the peer, it will add it to the variable accepted.
Variables changed: messages, *message_history*, accepted
@type: (MONITOR, MESSAGE) ⇒ *Bool*;

$handle\_accept(mon,\ msg) \triangleq$

$\land\ isLeader[mon] = \text{TRUE}$
$\land\ \lor\ state[mon] = STATE\_UPDATING\_PREVIOUS$
$\quad \lor\ state[mon] = STATE\_UPDATING$
$\land\ phase[mon] = PHASE\_BEGIN$
$\land\ new\_value[mon] \neq Nil$
$\land$ IF $\lor\ msg.pn \neq accepted\_pn[mon]$
$\quad \lor\ \land\ last\_committed[mon] > 0$
$\quad\quad \land\ msg.last\_committed < last\_committed[mon] - 1$
$\quad$ THEN UNCHANGED $accepted$
$\quad$ ELSE $accepted' = [accepted$ EXCEPT $![mon] =$
$\quad\quad [accepted[mon]$ EXCEPT $![msg.from] = \text{TRUE}]]$
$\land\ Discard(msg)$
$\land$ UNCHANGED $\langle epoch,\ quorum,\ quorum\_sz,\ pending\_proposal,\ new\_value\rangle$
$\land$ UNCHANGED $\langle restart\_vars,\ state\_vars,\ data\_vars,\ collect\_vars,\ lease\_vars\rangle$

Predicate that is enabled and called when all peers in the quorum accept begin request from leader.
The leader commits the transaction in *new_value* and sends commit messages to his peers.
Variables changed: *first_committed*, *last_committed*, *monitor_store*, *new_value*, messages, *message_history*, state, phase
@type: $MONITOR \Rightarrow Bool$;

$post\_accept(mon) \triangleq$
$\land\ phase[mon] = PHASE\_BEGIN$
$\land\ \forall\ m \in Monitors : quorum[m] \Rightarrow accepted[mon][m] = \text{TRUE}$
$\land\ new\_value[mon] \neq Nil$
$\land\ \lor\ state[mon] = STATE\_UPDATING\_PREVIOUS$
$\quad \lor\ state[mon] = STATE\_UPDATING$
$\land\ last\_committed' = [last\_committed$ EXCEPT $![mon] = last\_committed[mon] + 1]$

$\land$ IF $first\_committed[mon] = 0$
$\quad$ THEN $first\_committed' = [first\_committed$ EXCEPT $![mon] = first\_committed[mon] + 1]$
$\quad$ ELSE UNCHANGED $first\_committed$

$\land\ monitor\_store' = [monitor\_store$ EXCEPT $![mon] = values[mon][last\_committed[mon] + 1]]$
$\land\ new\_value' = [new\_value$ EXCEPT $![mon] = Nil]$
$\land\ Send\_set(mon,$
$\quad \{[type \mapsto OP\_COMMIT,$
$\quad\ from \mapsto mon,$
$\quad\ dest \mapsto dest,$
$\quad\ last\_committed \mapsto last\_committed'[mon],$
$\quad\ pn \mapsto accepted\_pn[mon],$
$\quad\ values \mapsto values[mon]] : dest \in \{m \in Monitors \setminus \{mon\} : quorum[m]\}$
$\quad \})$
$\land\ state' = [state$ EXCEPT $![mon] = STATE\_REFRESH]$
$\land\ phase' = [phase$ EXCEPT $![mon] = PHASE\_COMMIT]$
$\land$ UNCHANGED $\langle isLeader,\ values,\ accepted\_pn,\ pending\_proposal,\ accepted\rangle$
$\land$ UNCHANGED $\langle epoch,\ quorum,\ quorum\_sz,\ restart\_vars,\ collect\_vars,\ lease\_vars\rangle$

Predicate that is called after *post_accept*. The leader finishes the commit phase by updating his state to *STATE_ACTIVE* and by extending the lease to his peers.
Variables changed: state, phase, *acked_lease*, messages, *message_history*.
@type: *MONITOR* ⇒ *Bool*;

$$
\begin{aligned}
&finish\_commit(mon) \triangleq \\
&\quad \land\ state[mon] = STATE\_REFRESH \\
&\quad \land\ phase[mon] = PHASE\_COMMIT \\
&\quad \land\ finish\_round(mon) \\
&\quad \land\ extend\_lease(mon) \\
&\quad \land\ \text{UNCHANGED}\ \langle epoch,\ quorum,\ quorum\_sz,\ isLeader\rangle \\
&\quad \land\ \text{UNCHANGED}\ \langle restart\_vars,\ data\_vars,\ collect\_vars,\ commit\_vars\rangle
\end{aligned}
$$

Handle a commit message. The monitor stores the values sent by the leader commit message.
Variables changed: messages, *message_history*, values, *first_committed*, *last_committed*, *monitor_store*, *uncommitted_v*, *uncommitted_pn*, *uncommitted_value*.
@type: (MONITOR, MESSAGE) ⇒ *Bool*;

$$
\begin{aligned}
&handle\_commit(mon,\ msg) \triangleq \\
&\quad \land\ isLeader[mon] = \text{FALSE} \\
&\quad \land\ store\_state(mon,\ msg) \\
&\quad \land\ check\_and\_correct\_uncommitted(mon) \\
&\quad \land\ Discard(msg) \\
&\quad \land\ \text{UNCHANGED}\ \langle epoch,\ quorum,\ quorum\_sz,\ accepted\_pn,\ pending\_pn,\ pending\_v\rangle \\
&\quad \land\ \text{UNCHANGED}\ \langle state\_vars,\ collect\_vars,\ lease\_vars,\ commit\_vars\rangle
\end{aligned}
$$

Client Request

Request a transaction v to the monitor. The transaction is saved on pending proposal to be committed in the next available commit phase.
Variables changed: *pending_proposal*.
@type: (MONITOR, VALUE) ⇒ *Bool*;

$$
\begin{aligned}
&client\_request(mon,\ v) \triangleq \\
&\quad \land\ isLeader[mon] = \text{TRUE} \\
&\quad \land\ state[mon] = STATE\_ACTIVE \\
&\quad \land\ pending\_proposal[mon] = Nil \\
&\quad \land\ pending\_proposal' = [pending\_proposal\ \text{EXCEPT}\ ![mon] = v] \\
&\quad \land\ \text{UNCHANGED}\ \langle new\_value,\ accepted\rangle \\
&\quad \land\ \text{UNCHANGED}\ \langle global\_vars,\ state\_vars,\ restart\_vars,\ data\_vars,\ collect\_vars,\ lease\_vars\rangle
\end{aligned}
$$

Start a commit phase with the value on pending proposal.
Variables changed: state, *pending_proposal*, accepted, *new_value*, phase, messages, *message_history*, values, *pending_pn*, *pending_v*.
@type: *MONITOR* ⇒ *Bool*;

$$
\begin{aligned}
&propose\_pending(mon) \triangleq \\
&\quad \land\ phase[mon] = PHASE\_LEASE \lor phase[mon] = PHASE\_ELECTION \\
&\quad \land\ state[mon] = STATE\_ACTIVE
\end{aligned}
$$

$\land$ $pending\_proposal[mon] \neq Nil$
$\land$ $pending\_proposal' = [pending\_proposal$ EXCEPT $![mon] = Nil]$
$\land$ $state' = [state$ EXCEPT $![mon] = STATE\_UPDATING]$
$\land$ $begin(mon, pending\_proposal[mon])$
$\land$ UNCHANGED $\langle isLeader, monitor\_store, accepted\_pn, first\_committed, last\_committed,$
$epoch, quorum, quorum\_sz, uncommitted\_v, uncommitted\_pn, uncommitted\_value\rangle$
$\land$ UNCHANGED $\langle collect\_vars, lease\_vars\rangle$

Collect phase predicates

Start collect phase. This first part of the collect phase is divided in two parts (collect and *send_collect*) in order to simplify variable changes (when collect is triggered from *handle_last*).
Variables changed: *accepted_pn*, phase.
@type: (MONITOR, *Int*) $\Rightarrow$ *Bool*;

$collect(mon, oldpn) \triangleq$
$\land$ $state[mon] = STATE\_RECOVERING$
$\land$ $isLeader[mon] =$ TRUE
$\land$ LET $new\_pn \triangleq get\_new\_proposal\_number(mon, Max(\{oldpn, accepted\_pn[mon]\}))$
IN $\land$ $accepted\_pn' = [accepted\_pn$ EXCEPT $![mon] = new\_pn]$
$\land$ $phase' = [phase$ EXCEPT $![mon] = PHASE\_SEND\_COLLECT]$

Continue the start of the collect phase. Initialize the number of peers that accepted the proposal (*num_last*) and the variables with peers version numbers. Check if there is an uncommitted value.
Send collect messages to the peers.
Variables changed: *peer_first_committed*, *peer_last_committed*, *uncommitted_pn*, *uncommitted_v*, *uncommitted_value*, *num_last*, messages, *message_history*, phase.
@type: *MONITOR* $\Rightarrow$ *Bool*;

$send\_collect(mon) \triangleq$
$\land$ $state[mon] = STATE\_RECOVERING$
$\land$ $isLeader[mon] =$ TRUE
$\land$ $phase[mon] = PHASE\_SEND\_COLLECT$
$\land$ $clear\_peer\_first\_committed(mon)$
$\land$ $clear\_peer\_last\_committed(mon)$

$\land$ IF $last\_committed[mon] + 1 \in$ DOMAIN $values[mon]$
THEN $\land$ $uncommitted\_v' =$
$[uncommitted\_v$ EXCEPT $![mon] = last\_committed[mon] + 1]$
$\land$ $uncommitted\_value' =$
$[uncommitted\_value$ EXCEPT $![mon] = values[mon][last\_committed[mon] + 1]]$
$\land$ $uncommitted\_pn' = [uncommitted\_pn$ EXCEPT $![mon] = pending\_pn[mon]]$
$\land$ UNCHANGED $\langle pending\_pn, pending\_v\rangle$
ELSE UNCHANGED $\langle restart\_vars\rangle$

$\land$ $num\_last' = [num\_last$ EXCEPT $![mon] = 1]$
$\land$ $Send\_set(mon,$

$\{[type \mapsto OP\_COLLECT,$
$from \mapsto mon,$
$dest \mapsto dest,$
$first\_committed \mapsto first\_committed[mon],$
$last\_committed \mapsto last\_committed[mon],$
$pn \mapsto accepted\_pn[mon]] : dest \in \{m \in Monitors \setminus \{mon\} : quorum[m]\}$
$\})$
$\land\ phase' = [phase \text{ EXCEPT } ![mon] = PHASE\_COLLECT]$
$\land$ UNCHANGED $\langle isLeader, state\rangle$
$\land$ UNCHANGED $\langle epoch, quorum, quorum\_sz, data\_vars, lease\_vars, commit\_vars\rangle$

Handle a collect message. The peer will accept the proposal number from the leader if it is bigger than the last proposal number he accepted.
Variables changed: messages, *message_history*, epoch, state, *accepted_pn*.
@type: (MONITOR, MESSAGE) $\Rightarrow$ *Bool*;

$handle\_collect(mon, msg) \triangleq$
$\land\ isLeader[mon] = \text{FALSE}$
$\land\ state' = [state \text{ EXCEPT } ![mon] = STATE\_RECOVERING]$
$\land\ \lor\ \land\ msg.first\_committed > last\_committed[mon] + 1$
$\land\ bootstrap$
$\land\ Discard(msg)$
$\land$ UNCHANGED $\langle accepted\_pn\rangle$
$\lor\ \land\ msg.first\_committed \leq last\_committed[mon] + 1$
$\land$ IF $msg.pn > accepted\_pn[mon]$
THEN $accepted\_pn' = [accepted\_pn \text{ EXCEPT } ![mon] = msg.pn]$
ELSE UNCHANGED $accepted\_pn$
$\land\ Reply([type \mapsto OP\_LAST,$
$from \mapsto mon,$
$dest \mapsto msg.from,$
$first\_committed \mapsto first\_committed[mon],$
$last\_committed \mapsto last\_committed[mon],$
$values \mapsto values[mon],$
$uncommitted\_pn \mapsto pending\_pn[mon],$
$pn \mapsto accepted\_pn'[mon]], msg)$
$\land$ UNCHANGED $epoch$
$\land$ UNCHANGED $\langle isLeader, phase, values, first\_committed, last\_committed, monitor\_store\rangle$
$\land$ UNCHANGED $\langle quorum, quorum\_sz, restart\_vars, collect\_vars, lease\_vars, commit\_vars\rangle$

Handle a last message (response from a peer to the leader collect message).
The peers first and last committed version are stored. If the leader is behind, bootstraps. Stores any value that the peer may have committed (*store_state*). If peer is behind send commit message with leader values.
If peer accepted proposal number increase num last, if he sent a bigger proposal number start a new collect phase.
Variables changed: messages, *message_history*, epoch, phase, *uncommitted_pn*, *uncommitted_v*, *uncommitted_value*, *monitor_store*, values, *accepted_pn*, *first_committed*, *last_committed*, *num_last*, *peer_first_committed*, *peer_last_committed*.
@type: (MONITOR, MESSAGE) $\Rightarrow$ *Bool*;

```
handle_last(mon, msg) ≜
    ∧ isLeader[mon] = TRUE

    ∧ peer_first_committed' = [peer_first_committed EXCEPT ![mon] =
        [peer_first_committed[mon] EXCEPT ![msg.from] = msg.first_committed]]
    ∧ peer_last_committed' = [peer_last_committed EXCEPT ![mon] =
        [peer_last_committed[mon] EXCEPT ![msg.from] = msg.last_committed]]

    ∧ IF msg.first_committed > last_committed[mon] + 1
      THEN
       ∧ bootstrap
       ∧ Discard(msg)
       ∧ UNCHANGED ⟨first_committed, last_committed, num_last, accepted_pn, values, phase⟩
       ∧ UNCHANGED ⟨monitor_store, uncommitted_pn, uncommitted_v, uncommitted_value⟩
      ELSE
       ∧ store_state(mon, msg)
       ∧ IF ∃ peer ∈ Monitors :
              ∧ peer ≠ mon
              ∧ peer_last_committed'[mon][peer] ≠ −1
              ∧ peer_last_committed'[mon][peer] + 1 < first_committed[mon]
              ∧ first_committed[mon] > 1
         THEN
          ∧ bootstrap
          ∧ check_and_correct_uncommitted(mon)
          ∧ Discard(msg)
          ∧ UNCHANGED ⟨phase, accepted_pn, num_last⟩
         ELSE
          Send commit messages with my values to peers that are behind.
          ∧ LET monitors_behind ≜ {peer ∈ Monitors :
                  ∧ peer ≠ mon
                  ∧ peer_last_committed'[mon][peer] ≠ −1
                  ∧ peer_last_committed'[mon][peer] < last_committed[mon]
                  ∧ quorum[peer]}
            IN  Reply_set(mon,
                  {[type           ↦ OP_COMMIT,
                    from           ↦ mon,
                    dest           ↦ dest,
                    last_committed ↦ last_committed'[mon],
                    pn             ↦ accepted_pn[mon],
                    values         ↦ values[mon]] : dest ∈ monitors_behind
                  }, msg)
          ∧ ∨  Start new collect phase with a bigger proposal number.
               ∧ msg.pn > accepted_pn[mon]
               ∧ collect(mon, msg.pn)
               ∧ check_and_correct_uncommitted(mon)
```

```
                    ∧ UNCHANGED num_last

                ∨  Peer accepted the request, count his vote.
                    ∧ msg.pn = accepted_pn[mon]
                    ∧ num_last' = [num_last EXCEPT ![mon] = num_last[mon] + 1]
                    ∧ IF  ∧ msg.last_committed + 1 ∈ DOMAIN msg.values
                          ∧ msg.last_committed ≥ last_committed'[mon]
                          ∧ msg.last_committed + 1 ≥ uncommitted_v[mon]
                          ∧ msg.uncommitted_pn ≥ uncommitted_pn[mon]
                      THEN  ∧ uncommitted_v' =
                                  [uncommitted_v EXCEPT ![mon] = msg.last_committed + 1]
                            ∧ uncommitted_pn' =
                                  [uncommitted_pn EXCEPT ![mon] = msg.uncommitted_pn]
                            ∧ uncommitted_value' =
                                  [uncommitted_value EXCEPT ![mon] = msg.values[msg.last_committed + 1]]
                      ELSE  check_and_correct_uncommitted(mon)
                    ∧ UNCHANGED ⟨phase, accepted_pn⟩

                ∨ ∧ msg.pn < accepted_pn[mon]
                    ∧ check_and_correct_uncommitted(mon)
                    ∧ UNCHANGED ⟨phase, accepted_pn, num_last⟩
            ∧ UNCHANGED epoch
        ∧ UNCHANGED epoch

    ∧ UNCHANGED ⟨quorum, quorum_sz, isLeader, state, pending_pn, pending_v⟩
    ∧ UNCHANGED ⟨lease_vars, commit_vars⟩

Predicate that is enabled and called when all peers in quorum accept collect request from leader. If there is an
uncommitted value, a commit phase is started with that value, else the leader changes to ACTIVE_STATE and extends
the lease to his peers.
Variables changed: peer_first_committed, peer_last_committed, state, accepted, new_value, phase, messages,
message_history, values, pending_pn, pending_v, acked_lease.
@type: MONITOR ⇒ Bool;
post_last(mon) ≜
    ∧ isLeader[mon]  = TRUE
    ∧ num_last[mon] = quorum_sz
    ∧ phase[mon] = PHASE_COLLECT

    ∧ clear_peer_first_committed(mon)
    ∧ clear_peer_last_committed(mon)

    ∧ IF  ∧ uncommitted_v[mon] = last_committed[mon] + 1
          ∧ uncommitted_value[mon] ≠ Nil
      THEN  ∧ state' = [state EXCEPT ![mon] = STATE_UPDATING_PREVIOUS]
            ∧ begin(mon, uncommitted_value[mon])
            ∧ UNCHANGED ⟨acked_lease, uncommitted_v, uncommitted_pn, uncommitted_value⟩
      ELSE  ∧ finish_round(mon)
```

$\land$ *extend_lease*(*mon*)
$\land$ UNCHANGED $\langle$*accepted*, *new_value*, *values*, *restart_vars*$\rangle$

$\land$ UNCHANGED $\langle$*isLeader*, *monitor_store*, *accepted_pn*, *first_committed*, *last_committed*$\rangle$
$\land$ UNCHANGED $\langle$*epoch*, *quorum*, *quorum_sz*, *num_last*, *pending_proposal*$\rangle$

Leader election

Elect one monitor as a leader and initialize the remaining ones as peons.
Variables changed: *isLeader*, state, phase, *new_value*, *pending_proposal*, epoch.
@type: *Bool*;

leader_election $\triangleq$
$\land \exists\, mon \in Monitors$:
$\land$ *quorum*[*mon*]
$\land$ *isLeader*$'$ = [$m \in Monitors \mapsto$ IF $m = mon$ THEN TRUE ELSE FALSE]
$\land$ *state*$'$ = [$m \in Monitors \mapsto$
IF *quorum_sz* = 1 THEN *STATE_ACTIVE* ELSE *STATE_RECOVERING*]
$\land$ *phase*$'$ = [$m \in Monitors \mapsto$ *PHASE_ELECTION*]
$\land$ *new_value*$'$ = [$m \in Monitors \mapsto$ *Nil*]
$\land$ *pending_proposal*$'$ = [$m \in Monitors \mapsto$ *Nil*]
$\land$ *epoch*$'$ = *epoch* + 1
$\land$ *messages*$'$ = [*mon*1 $\in$ *Monitors* $\mapsto$ [*mon*2 $\in$ *Monitors* $\mapsto \langle\rangle$]]
$\land$ UNCHANGED $\langle$*quorum*, *quorum_sz*, *accepted*, *message_history*$\rangle$
$\land$ UNCHANGED $\langle$*data_vars*, *restart_vars*, *collect_vars*, *lease_vars*$\rangle$

Start recovery phase if number of monitors in quorum is greater than 1.
Variables changed: *accepted_pn*, phase.
@type: *MONITOR* $\Rightarrow$ *Bool*;

election_recover(*mon*) $\triangleq$
$\land$ *quorum_sz* > 1
$\land$ *phase*[*mon*] = *PHASE_ELECTION*
$\land$ *collect*(*mon*, 0)
$\land$ UNCHANGED $\langle$*isLeader*, *state*, *values*, *first_committed*, *last_committed*, *monitor_store*$\rangle$
$\land$ UNCHANGED $\langle$*global_vars*, *restart_vars*, *collect_vars*, *lease_vars*, *commit_vars*$\rangle$

Timeouts and restart

Remove monitor from quorum, if there are enough monitors in the quorum.
@type: *MONITOR* $\Rightarrow$ *Bool*;

crash_mon(*mon*) $\triangleq$
$\land$ *quorum_sz* > (*MonitorsLen* $\div$ 2) + 1
$\land$ *quorum*[*mon*] = TRUE
$\land$ *quorum*$'$ = [*quorum* EXCEPT ![*mon*] = FALSE]
$\land$ *quorum_sz*$'$ = *quorum_sz* $-$ 1
$\land$ *bootstrap*

$\land$ UNCHANGED $\langle messages, message\_history\rangle$
$\land$ UNCHANGED $\langle state\_vars, restart\_vars, data\_vars, collect\_vars, lease\_vars, commit\_vars\rangle$

Add monitor to the quorum.
@type: $MONITOR \Rightarrow Bool$;
$restore\_mon(mon) \triangleq$
$\land\ quorum[mon] = \text{FALSE}$
$\land\ quorum' = [quorum \text{ EXCEPT } ![mon] = \text{TRUE}]$
$\land\ quorum\_sz' = quorum\_sz + 1$
$\land\ bootstrap$
$\land$ UNCHANGED $\langle messages, message\_history\rangle$
$\land$ UNCHANGED $\langle state\_vars, restart\_vars, data\_vars, collect\_vars, lease\_vars, commit\_vars\rangle$

Monitor timeout (simulate the various timeouts that can occur). Triggers new elections.
Variables changed: epoch.
@type: $MONITOR \Rightarrow Bool$;
$Timeout(mon) \triangleq$
$\land\quad bootstrap$
$\land\quad$ UNCHANGED $\langle messages, quorum, quorum\_sz, message\_history, state\_vars, restart\_vars,$
$data\_vars, collect\_vars, lease\_vars, commit\_vars\rangle$

Dispatchers and next statement

Handle a message.
@type: $MESSAGE \Rightarrow Bool$;
$Receive(msg) \triangleq$
$\land\quad \lor \land msg.type = OP\_COLLECT$
$\land\ handle\_collect(msg.dest, msg)$
$\land\ step\_name' =$ "receive collect"

$\lor \land msg.type = OP\_LAST$
$\land\ handle\_last(msg.dest, msg)$
$\land\ step\_name' =$ "receive last"

$\lor \land msg.type = OP\_LEASE$
$\land\ handle\_lease(msg.dest, msg)$
$\land\ step\_name' =$ "receive lease"

$\lor \land msg.type = OP\_LEASE\_ACK$
$\land\ handle\_lease\_ack(msg.dest, msg)$
$\land\ step\_name' =$ "receive lease_ack"

$\lor \land msg.type = OP\_BEGIN$
$\land\ handle\_begin(msg.dest, msg)$
$\land\ step\_name' =$ "receive begin"

$\lor \land msg.type = OP\_ACCEPT$

$\land\ handle\_accept(msg.dest,\ msg)$
$\land\ step\_name' = \text{“receive accept”}$

$\lor\ \land\ msg.type = OP\_COMMIT$
$\land\ handle\_commit(msg.dest,\ msg)$
$\land\ step\_name' = \text{“receive commit”}$

Limit some variables to reduce search space.
@type: *Bool*;
$reduce\_search\_space\ \triangleq$
$\land\ epoch \neq 14$
$\land\ \forall\, mon \in Monitors : last\_committed[mon] < 2$
$\land\ \forall\, mon \in Monitors : accepted\_pn[mon] < 400$

State transitions.
@type: *Bool*;
$Next\ \triangleq$
$\land$ *reduce_search_space*
$\land$ IF $epoch\%2 = 1$ THEN
$\land$ *leader_election*
$\land\ step\_name' = \text{“election”}$
ELSE
$\lor\ \land\ \exists\, mon \in Monitors : election\_recover(mon)$
$\land\ step\_name' = \text{“election\_recover”}$

$\lor\ \land\ \exists\, mon \in Monitors : send\_collect(mon)$
$\land\ step\_name' = \text{“send\_collect”}$

$\lor\ \land\ \exists\, mon \in Monitors : post\_last(mon)$
$\land\ step\_name' = \text{“post\_last”}$

$\lor\ \land\ \exists\, mon \in Monitors : post\_lease\_ack(mon)$
$\land\ step\_name' = \text{“post\_lease\_ack”}$

$\lor\ \land\ \exists\, mon \in Monitors : post\_accept(mon)$
$\land\ step\_name' = \text{“post\_accept”}$

$\lor\ \land\ \exists\, mon \in Monitors : finish\_commit(mon)$
$\land\ step\_name' = \text{“finish\_commit”}$

$\lor\ \land\ \exists\, mon \in Monitors : \exists\, v \in Value\_set : client\_request(mon,\ v)$
$\land\ step\_name' = \text{“client\_request”}$

$\lor\ \land\ \exists\, mon \in Monitors : propose\_pending(mon)$
$\land\ step\_name' = \text{“propose\_pending”}$

$\lor\ \land\ \exists\, mon1,\ mon2 \in Monitors :$
$\land\ mon1 \neq mon2$

$\land\ Len(messages[mon1][mon2]) > 0$
$\land\ Receive(messages[mon1][mon2][1])$

$\lor\ \land \exists\, mon \in Monitors : crash\_mon(mon)$
$\land\ step\_name' = \text{“crash\_mon”}$

$\lor\ \land \exists\, mon \in Monitors : restore\_mon(mon)$
$\land\ step\_name' = \text{“restore\_mon”}$

$\lor\ \land \exists\, mon \in Monitors : Timeout(mon)$
$\land\ step\_name' = \text{“timeout\_and\_restart”}$

Safety invariants

If two monitors are in state active then their *monitor_store* must have the same value.
@type: *Bool*;

$same\_monitor\_store \triangleq$
$\forall\, mon1,\, mon2 \in Monitors :$
$state[mon1] = STATE\_ACTIVE \land state[mon2] = STATE\_ACTIVE$
$\Rightarrow monitor\_store[mon1] = monitor\_store[mon2]$

In each version, every monitor chooses the same value.
@type: *Bool*;

$same\_monitor\_values \triangleq$
$\forall\, version \in 1\,..\,Max(\{last\_committed[mon] : mon \in Monitors\}) :$
$\exists\, val \in Value\_set :$
$\forall\, mon \in Monitors :$
$last\_committed[mon] < version \lor values[mon][version] = val$

Invariant.
@type: *Bool*;

$Inv \triangleq \land\ same\_monitor\_store$
$\land\ same\_monitor\_values$

Test/Debug invariants

Invariant used to search for a state where 'x' happens.

$Inv\_find\_state(x) \triangleq \neg x$

Invariant used to search for a behavior of diameter equal to 'size'.
$TLCGet(\text{“}level\text{”})$ not supported by snowcat typechecker.

$Inv\_diam(size) \triangleq TLCGet(\text{“}level\text{”}) \neq \text{size} - 1$

Invariants to test in model check

$DEBUG\_Inv \triangleq \land\ \text{TRUE}$
$\land\ Inv\_diam(20)$

Examples:

Find a behavior with a diameter of size 60.
$Inv\_diam(60)$

Find a behavior where two different monitors assume the role of a leader (in different epochs). Note: In the message manipulation predicates, the variable *message_history* needs to be updated.

$$
\begin{array}{l}
Inv\_find\_state(\\
\quad \exists\, msg1,\, msg2 \in message\_history : \\
\qquad \land\, msg1.type = OP\_COLLECT \land msg2.type = OP\_COLLECT \\
\qquad \land\, msg1.from \neq msg2.from \\
)
\end{array}
$$

Find a state where a monitor crashed during the collect phase and fails to send a OP_LAST message.

$$
\begin{array}{l}
Inv\_find\_state(\\
\quad \land\, step\_name = \text{“crash\_mon”} \\
\quad \backslash * \textit{ The system is in collect phase and no } OP\_LAST \textit{ message has been received.} \\
\quad \backslash * \textit{ } isLeader[mon] = \text{TRUE } \textit{and } quorum[mon] \textit{assures that the leader was not the one that crashed.} \\
\quad \land\, \exists\, mon \in Monitors : \\
\qquad \land\, isLeader[mon] = \text{TRUE} \land quorum[mon] \\
\qquad \land\, phase[mon] = PHASE\_COLLECT \\
\qquad \land\, num\_last[mon] = 1 \\
\quad \backslash * \textit{ All the collect requests have been handled by the peers.} \\
\quad \land\, \forall\, mon1,\, mon2 \in Monitors : \\
\qquad \forall\, i \in 1 \,..\, Len(messages[mon1][mon2]) : messages[mon1][mon2][i].type \neq OP\_COLLECT \\
)
\end{array}
$$

Find a state where the leader crashes during the commit phase, failing to complete the commit.

$$
\begin{array}{l}
Inv\_find\_state(\\
\quad \land\, step\_name = \text{“crash\_mon”} \\
\quad \land\, \exists\, mon1,\, mon2 \in Monitors : \\
\qquad \exists\, i \in 1 \,..\, Len(messages[mon1][mon2]) : messages[mon1][mon2][i].type = OP\_ACCEPT \\
\quad \land\, \forall\, mon \in Monitors : \\
\qquad isLeader[mon] = \text{FALSE} \lor quorum[mon] = \text{FALSE} \\
)
\end{array}
$$

Note: After finding a state, that complete state can be used as an initial state to analyze behaviors from there.

* Modification History
* Last modified *Thu Jun* 10 18:34:59 WEST 2021 by *afonsonf*
* Created *Mon Jan* 11 16:15:26 WET 2021 by *afonsonf*