In Search of an Understandable Consensus Algorithm

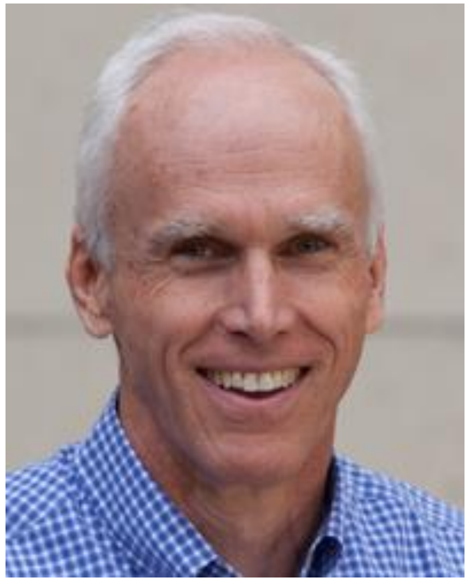
Diego Ongaro and John Ousterhout

Stanford University

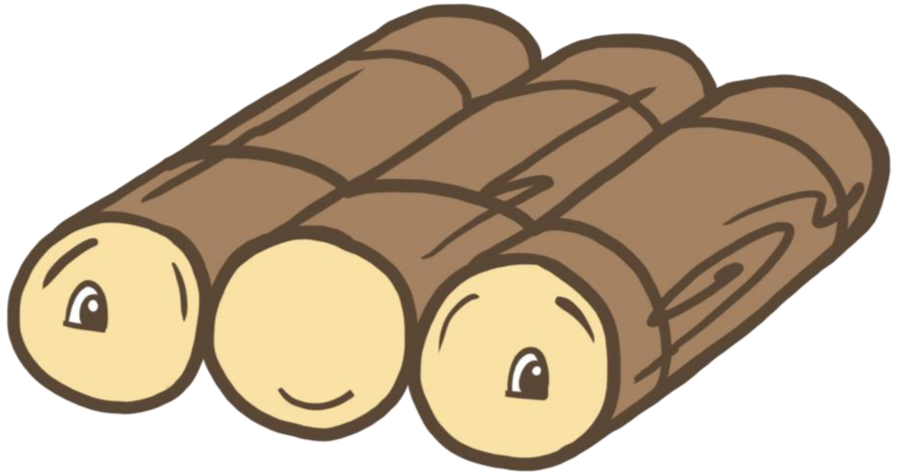
Distributed Computing (1986) 1: 5-6

Introducción

Diego Ongaro


John Ousterhout
Premio Grace Hooper(1987)

Introducción

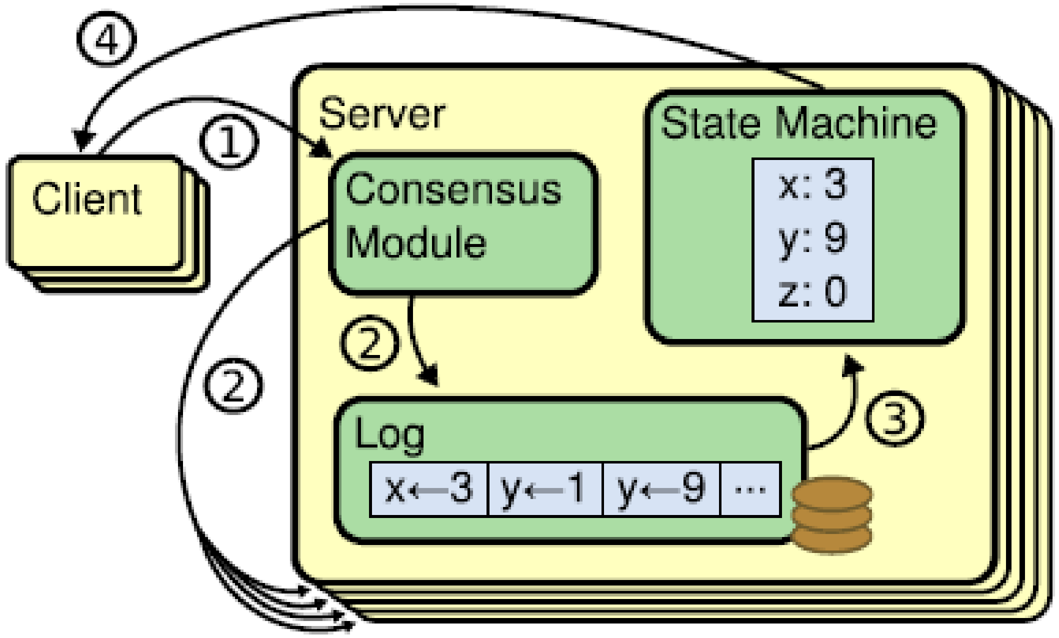


The Raft Consensus Algorithm

Objetivos del algoritmo

- Eficiencia
- Seguridad
- Disponibilidad
- Base completa para la construcción de un sistema
- Comprensibilidad

Introducción



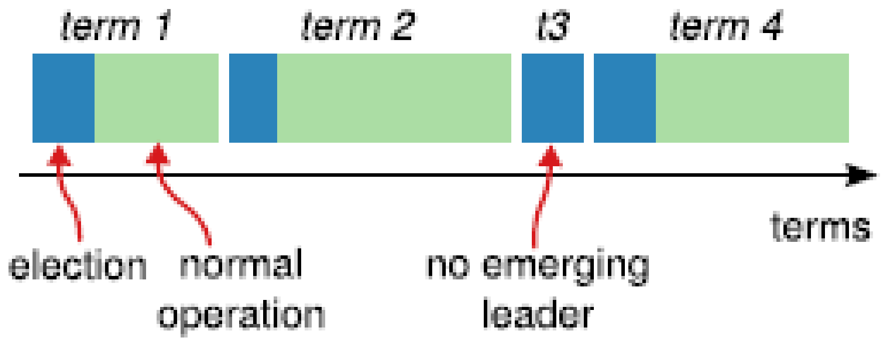
Introducción

- Elección de un Líder
- Replicación de registros
- Seguridad

Elección de un líder

Estado de un servidor:

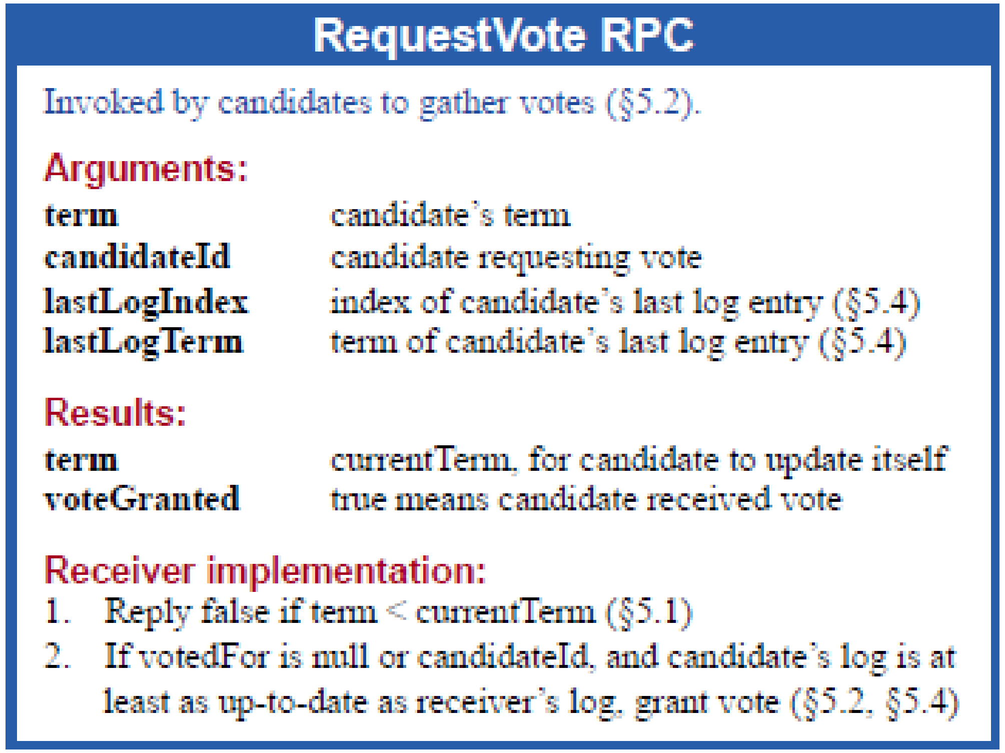
- Líder
- Seguidor
- Candidato



Elección de un líder

Estado de un servidor:

- Líder
- Seguidor
- Candidato



RequestVote RPC

Invoked by candidates to gather votes (§5.2).

Arguments:

term	candidate's term
candidateId	candidate requesting vote
lastLogIndex	index of candidate's last log entry (§5.4)
lastLogTerm	term of candidate's last log entry (§5.4)

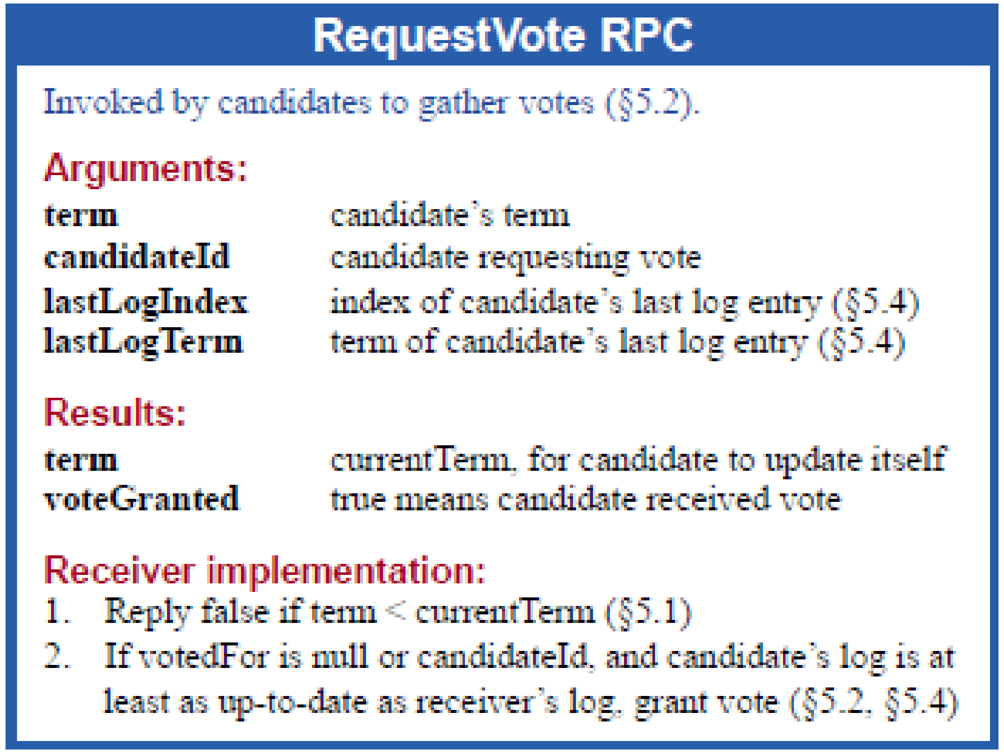
Results:

term	currentTerm, for candidate to update itself
voteGranted	true means candidate received vote

Receiver implementation:

1. Reply false if term < currentTerm (§5.1)
2. If votedFor is null or candidateId, and candidate's log is at least as up-to-date as receiver's log, grant vote (§5.2, §5.4)

Elección de un líder



RequestVote RPC

Invoked by candidates to gather votes (§5.2).

Arguments:

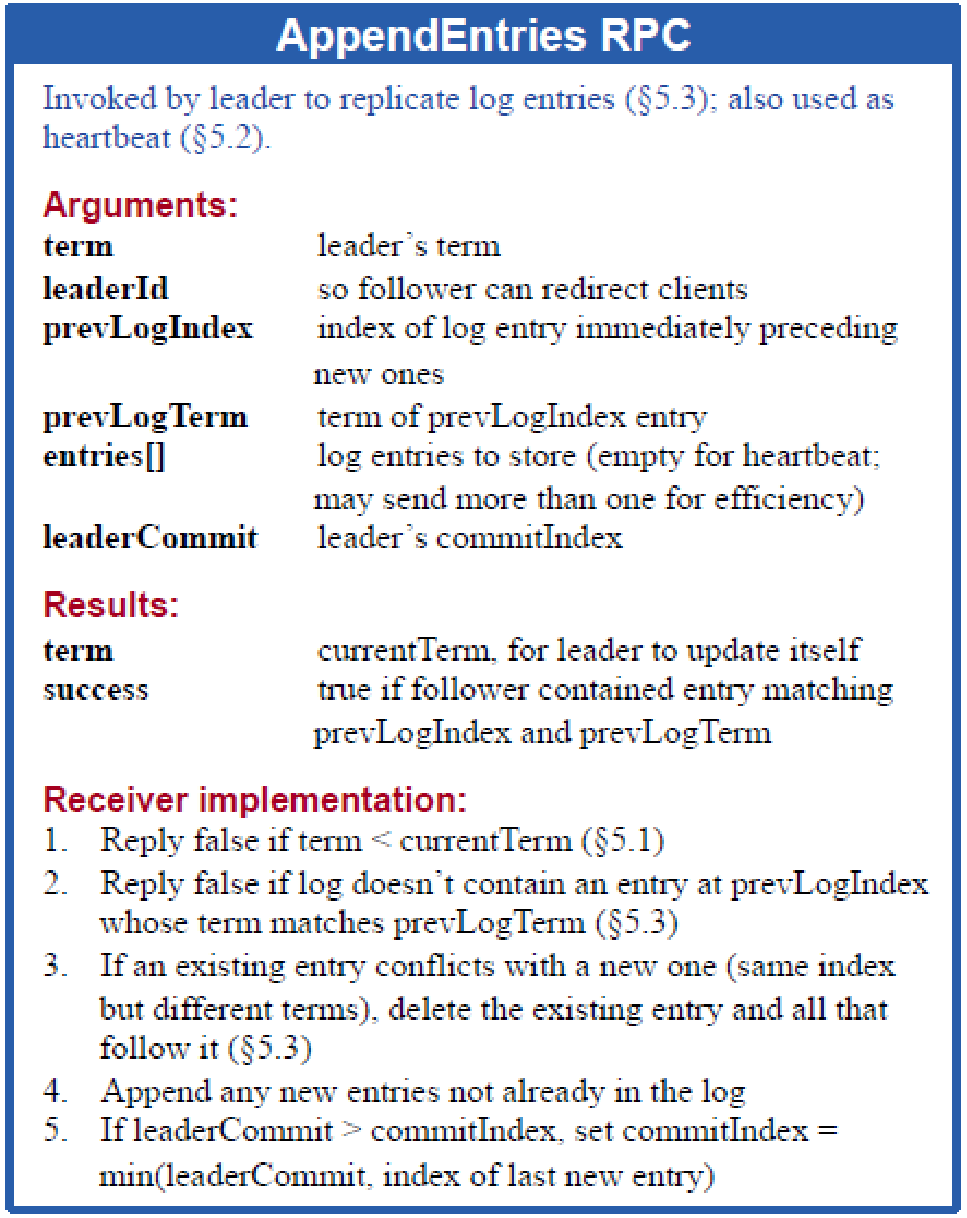
term	candidate's term
candidateId	candidate requesting vote
lastLogIndex	index of candidate's last log entry (§5.4)
lastLogTerm	term of candidate's last log entry (§5.4)

Results:

term	currentTerm, for candidate to update itself
voteGranted	true means candidate received vote

Receiver implementation:

1. Reply false if term < currentTerm (§5.1)
2. If votedFor is null or candidateId, and candidate's log is at least as up-to-date as receiver's log, grant vote (§5.2, §5.4)



AppendEntries RPC

Invoked by leader to replicate log entries (§5.3); also used as heartbeat (§5.2).

Arguments:

term	leader's term
leaderId	so follower can redirect clients
prevLogIndex	index of log entry immediately preceding new ones
prevLogTerm	term of prevLogIndex entry
entries[]	log entries to store (empty for heartbeat; may send more than one for efficiency)
leaderCommit	leader's commitIndex

Results:

term	currentTerm, for leader to update itself
success	true if follower contained entry matching prevLogIndex and prevLogTerm

Receiver implementation:

1. Reply false if term < currentTerm (§5.1)
2. Reply false if log doesn't contain an entry at prevLogIndex whose term matches prevLogTerm (§5.3)
3. If an existing entry conflicts with a new one (same index but different terms), delete the existing entry and all that follow it (§5.3)
4. Append any new entries not already in the log
5. If leaderCommit > commitIndex, set commitIndex = min(leaderCommit, index of last new entry)

LiveSlides web content

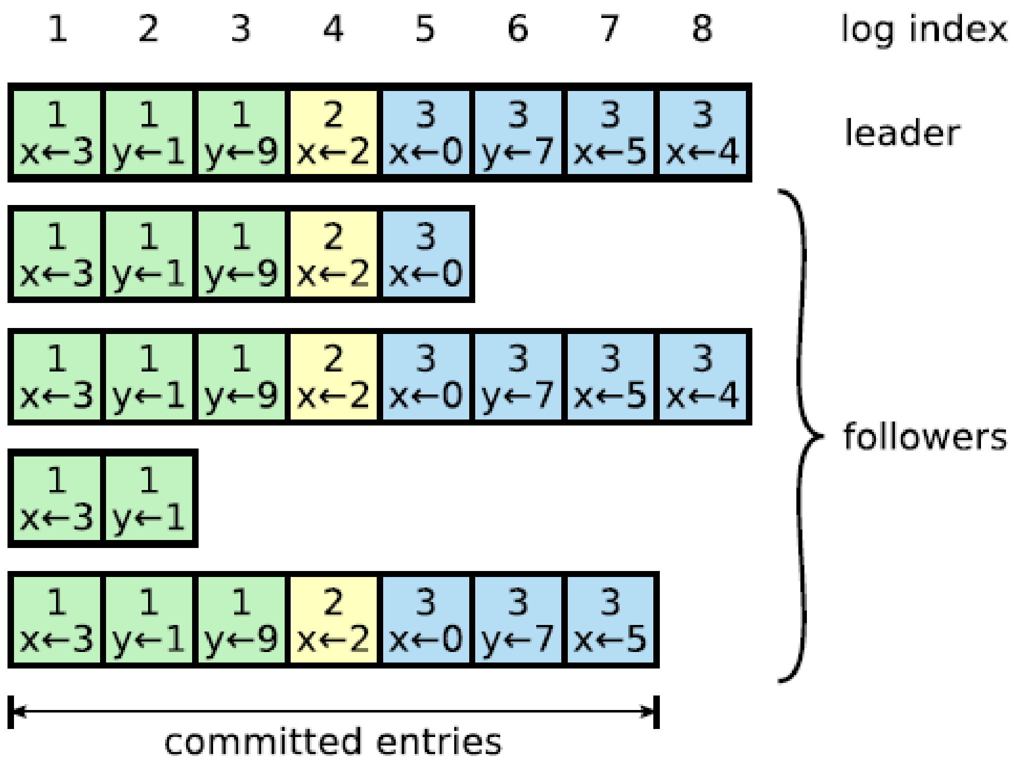
To view

Download the add-in.

liveslides.com/download

Start the presentation.

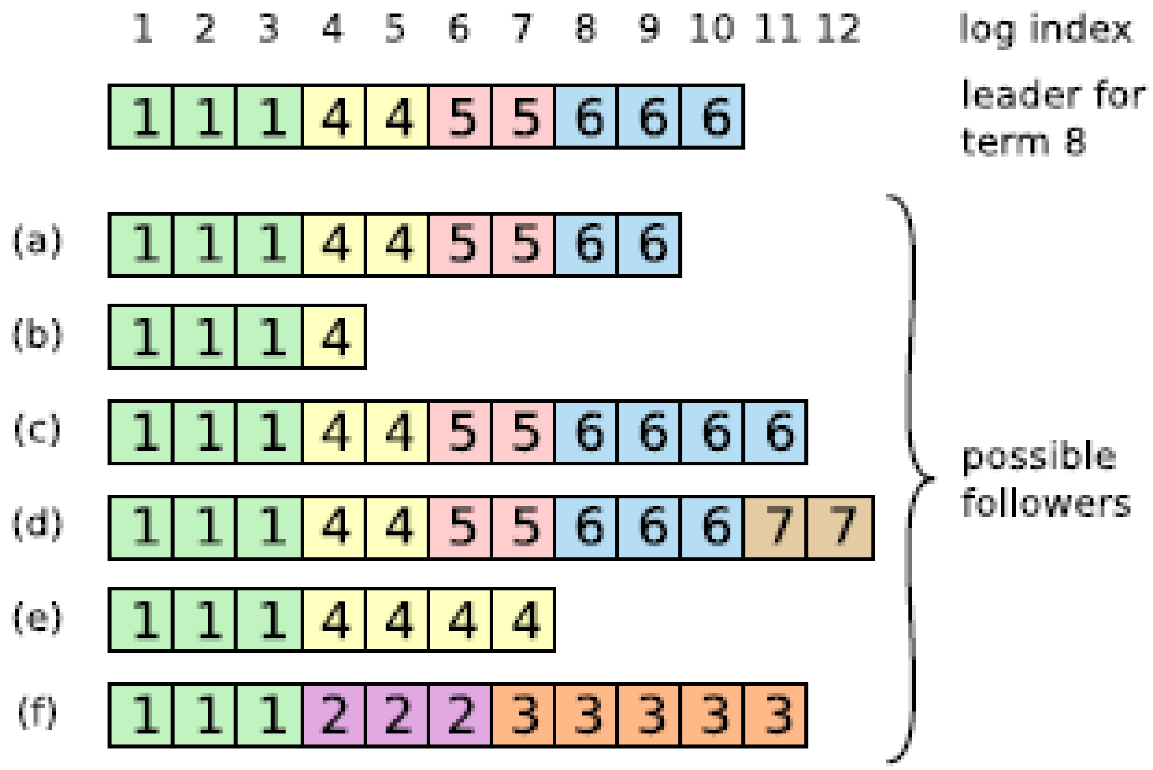
Replicación de registros



1 2 3 4 5 6 7 8
log index
1 x←3 | 1 y←1 | 1 y←9 | 2 x←2 | 3 x←0 | 3 y←7 | 3 x←5 | 3 x←4
leader
1 x←3 | 1 y←1 | 1 y←9 | 2 x←2 | 3 x←0
1 x←3 | 1 y←1 | 1 y←9 | 2 x←2 | 3 x←0 | 3 y←7 | 3 x←5 | 3 x←4
1 x←3 | 1 y←1
1 x←3 | 1 y←1 | 1 y←9 | 2 x←2 | 3 x←0 | 3 y←7 | 3 x←5
followers
committed entries

Leader Append-Only: Un líder nunca sobrescribe o elimina entradas en su registro; solo agrega nuevas entradas
Log Matching: Si dos registros contienen una entrada con el mismo índice y término, entonces los registros son idénticos en todas las entradas hasta el índice dado.

Inconsistencias en registros



- A un seguidor le pueden faltar entradas (a – b)
- Puede tener entradas extra no comprometidas (c – d), o ambas (e – f).

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

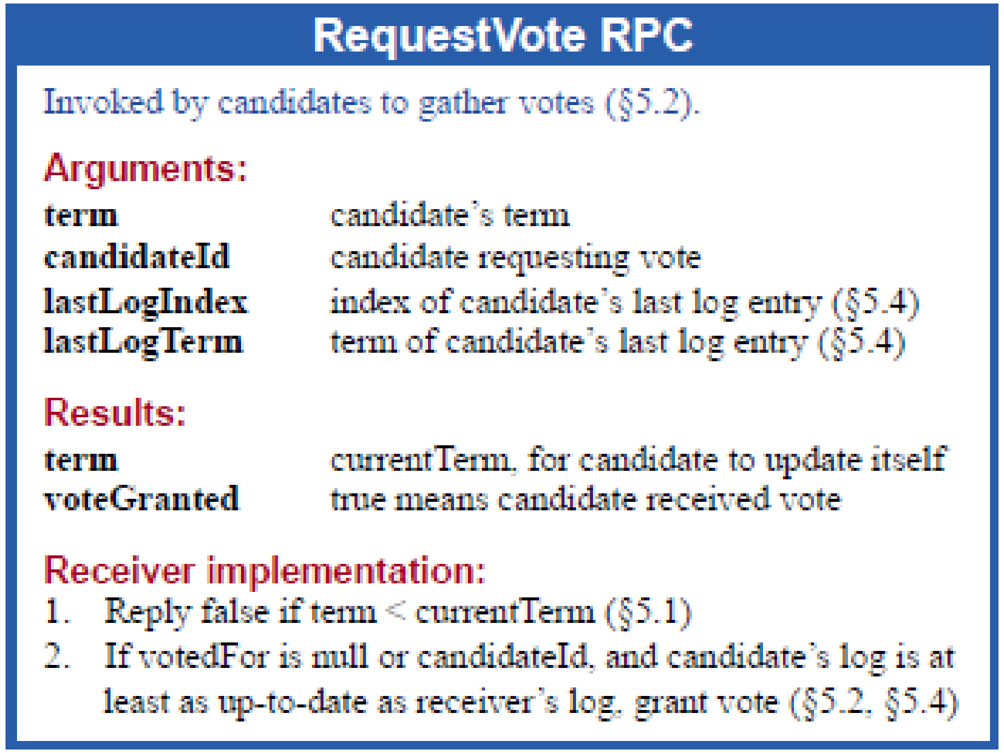
Start the presentation.

Seguridad

Completes de Líder: Si una entrada de registro se confirma en un término dado, entonces esa entrada estará presente en los registros de los líderes para todos los términos de mayor número.

Estado Maquina Seguro: Si un servidor ha aplicado una entrada de registro en un índice dado a su máquina de estado, ningún otro servidor aplicará una entrada de registro diferente para el mismo índice

Restricción de la elección



RequestVote RPC

Invoked by candidates to gather votes (§5.2).

Arguments:

term	candidate's term
candidateId	candidate requesting vote
lastLogIndex	index of candidate's last log entry (§5.4)
lastLogTerm	term of candidate's last log entry (§5.4)

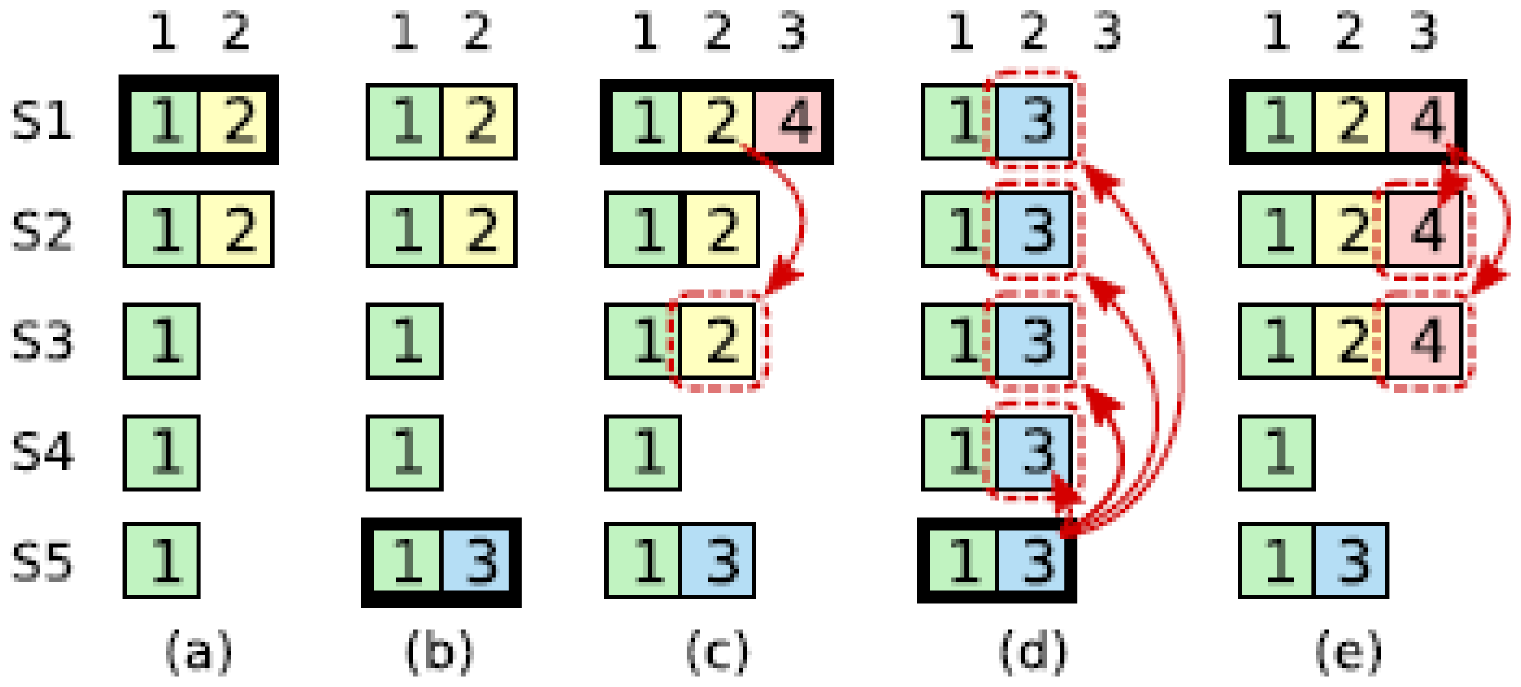
Results:

term	currentTerm, for candidate to update itself
voteGranted	true means candidate received vote

Receiver implementation:

1. Reply false if term < currentTerm (§5.1)
2. If votedFor is null or candidateId, and candidate's log is at least as up-to-date as receiver's log, grant vote (§5.2, §5.4)

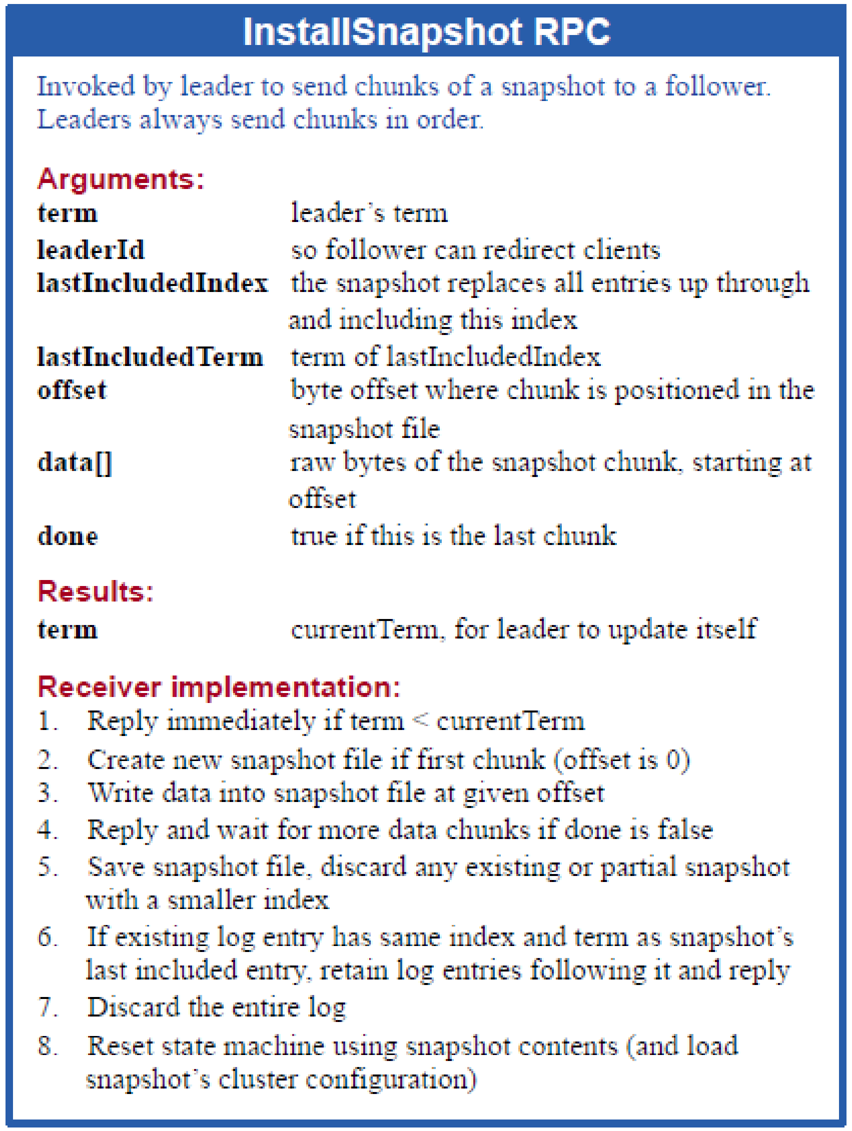
Confirmando entradas de términos anteriores



Choques de seguidores y candidatos

Si un seguidor o candidato falla, los futuros RPC RequestVote y AppendEntries que se le envíen fallarán. Raft maneja estas fallas reintentando indefinidamente; Si el servidor bloqueado se reinicia, la RPC se completará con éxito. Si un servidor falla después de completar un RPC pero antes de responder, recibirá el mismo RPC nuevamente después de reiniciarse

Instantánea del líder



InstallSnapshot RPC

Invoked by leader to send chunks of a snapshot to a follower. Leaders always send chunks in order.

Arguments:

term	leader's term
leaderId	so follower can redirect clients
lastIncludedIndex	the snapshot replaces all entries up through and including this index
lastIncludedTerm	term of lastIncludedIndex
offset	byte offset where chunk is positioned in the snapshot file
data[]	raw bytes of the snapshot chunk, starting at offset
done	true if this is the last chunk

Results:

term	currentTerm, for leader to update itself

Receiver implementation:

1. Reply immediately if term < currentTerm
2. Create new snapshot file if first chunk (offset is 0)
3. Write data into snapshot file at given offset
4. Reply and wait for more data chunks if done is false
5. Save snapshot file, discard any existing or partial snapshot with a smaller index
6. If existing log entry has same index and term as snapshot's last included entry, retain log entries following it and reply
7. Discard the entire log
8. Reset state machine using snapshot contents (and load snapshot's cluster configuration)

Reglas de reemplazo

- Si la instantanea contiene toda la informacion del registro, se elimina todo el registro
- Si hay entradas nuevas en el registrose elimina solo lo que abarca la instatanea y se conservan otras entradas