Acknowledgement: Diego Ongaro and John Ousterhout

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
Table of contents

1 Consensus variants
2 Historical overview
   • Paxos
   • Raft
3 Raft protocol
   • Overview
   • Elections
   • Normal operation
   • Neutralizing old leaders
   • Client protocol
   • Configuration changes
### Atomic Broadcast

**Definition (RB1 – Validity)**

If a correct process broadcasts \( m \), then it eventually delivers \( m \)

**Definition (RB2 – Uniform Integrity)**

\( m \) is delivered by a process at most once, and only if it was previously broadcast

**Definition (RB3 – Agreement)**

If a correct process delivers \( m \), then all correct processes eventually deliver \( m \)

**Definition (Total Order)**

If correct processes \( p \) and \( q \) both deliver messages \( m, m' \), then \( p \) delivers \( m \) before \( m' \) if and only if \( q \) delivers \( m \) before \( m' \)

\[
delivery_p(m) \rightarrow delivery_p(m') \Rightarrow delivery_q(m) \rightarrow delivery_q(m')
\]
State machine replication

Definition (State machine)
A state machine consists of:
- State variables
- Commands which transforms its state
  - Implemented by deterministic programs
  - Atomic with respect to other commands

Specification
- Agreement: every correct replica receives the same set of commands
- Order: every non-faulty state machine processes the commands it receives in the same order
Consensus

In the (Uniform) Consensus problem, the processes propose values and need to decide (agree) on one of these values.

Definition (Uniform Validity)
Any value decided is a value proposed

Definition ((Uniform) Agreement)
No two correct (any) processes decide differently

Definition (Termination)
Every correct process eventually decides

Definition (Uniform Integrity)
Every process decides at most once
Consensus, Atomic Broadcast, State Machine Replication

Equivalence between Consensus and Atomic Broadcast

1. There is an algorithm $T_{\text{Consensus}} \rightarrow \text{AtomicBroadcast}$

2. There is an algorithm $T_{\text{AtomicBroadcast}} \rightarrow \text{Consensus}$
Transformation executed by process $p$

upon initialization do
  boolean decided ← false

upon propose($v$) do
  A-broadcast($v$)

upon A-deliver($v$) do
  if not decided then
    decided ← true
    decide($u$)
Transformation executed by process $p$

**upon** initialization **do**

- **Set** unordered $\leftarrow \emptyset$ \hspace{4cm} \% Messages to be ordered
- **Set** delivered $\leftarrow \emptyset$ \hspace{4cm} \% Messages already delivered
- **boolean** wait $\leftarrow$ false \hspace{4cm} \% true when Consensus is running
- **integer** $s \leftarrow 1$ \hspace{4cm} \% Consensus protocol identifier

**upon** A-broadcast($m$) **do**
- R-broadcast($m$)

**upon** R-deliver($m$) **do**
- if not $m \in$ delivered then
  - unordered $\leftarrow$ unordered $\cup \{m\}$
Transformation executed by process \( p \)

\[
\text{upon } \text{decide}_s(S) \text{ do}
\]

\[
\begin{align*}
\text{unordered} & \leftarrow \text{unordered} - S \\
\text{foreach } m \in S \text{ do} \\
& \quad \text{A-deliver}(m) \quad \text{\% In some deterministic order} \\
\text{delivered} & \leftarrow \text{delivered} \cup S \\
\text{s} & \leftarrow \text{s} + 1 \\
\text{wait} & \leftarrow \text{false}
\end{align*}
\]

\[
\text{upon } \text{unordered} \neq \emptyset \text{ and not wait do}
\]

\[
\begin{align*}
\text{wait} & \leftarrow \text{true} \\
\text{propose}_s(\text{unordered})
\end{align*}
\]
Consensus and total order broadcast are equivalent problems in an asynchronous system with crashes and Perfect Channels

- Consensus can be obtained from total order broadcast
- Total order broadcast can be obtained from Consensus

This means that the impossibility results of Consensus apply to Atomic Broadcast as well.
## Table of contents

1. **Consensus variants**
2. **Historical overview**
   - Paxos
   - Raft
3. **Raft protocol**
   - Overview
   - Elections
   - Normal operation
   - Neutralizing old leaders
   - Client protocol
   - Configuration changes
Paxos History

1989 Leslie Lamport developed a new consensus protocol called Paxos; it was published as DEC SRC Technical Report 49. 42 pages!

Abstract

Recent archaeological discoveries on the island of Paxos reveal that the parliament functioned despite the peripatetic propensity of its part-time legislators. The legislators maintained consistent copies of the parliamentary record, despite their frequent forays from the chamber and the forgetfulness of their messengers. The Paxon parliament’s protocol provides a new way of implementing the state-machine approach to the design of distributed systems — *an approach that has received limited attention because it leads to designs of insufficient complexity.*
Paxos History


1996  “How to Build a Highly Available System Using Consensus”, by B. Lampson was published in WDAG 1996, Bologna, Italy.

1997  “Revisiting the Paxos Algorithm”, by R. De Prisco, B. Lampson, N. Lynch was published in WDAG 1997, Saarbrücken, Germany.

1998  The original paper is resubmitted and accepted by TOCS.

2001  Lamport publishes “Paxos made simple” in ACM SIGACT News
       • Because Lamport “got tired of everyone saying how difficult it was to understand the Paxos algorithm”
       • Abstract: “The Paxos algorithm, when presented in plain English, is very simple”
       • Introduces the concept of Multi-Paxos
Paxos History

Paxos optimizations and extensions

2004 Leslie Lamport and Mike Massa. “Cheap Paxos”. DSN’04, Florence, Italy


An important milestone

Paxos implementations

- Google uses the Paxos algorithm in their Chubby distributed lock service in order to keep replicas consistent in case of failure. Chubby is used by Bigtable which is now in production in Google Analytics and other products.
- Google Spanner and Megastore use the Paxos algorithm internally.
- The OpenReplica replication service uses Paxos to maintain replicas for an open access system that enables users to create fault-tolerant objects. It provides high performance through concurrent rounds and flexibility through dynamic membership changes.
- IBM supposedly uses the Paxos algorithm in their IBM SAN Volume Controller product to implement a general purpose fault-tolerant virtual machine used to run the configuration and control components of the storage virtualization services offered by the cluster. (Original MIT & IBM research paper)
- Microsoft uses Paxos in the Autopilot cluster management service from Bing, and in Windows Server Failover Clustering.
- WANdisco have implemented Paxos within their DConE active-active replication technology.[26]
- XtremFS uses a Paxos-based lease negotiation algorithm for fault-tolerant and consistent replication of file data and metadata.[27]
- Heroku uses Doozerd which implements Paxos for its consistent distributed data store.
- Ceph uses Paxos as part of the monitor processes to agree which OSDs are up and in the cluster.
- The Clustrix distributed SQL database uses Paxos for distributed transaction resolution.
- Neo4j HA graph database implements Paxos, replacing Apache ZooKeeper from v1.9
- Apache Cassandra NoSQL database uses Paxos for Light Weight Transaction feature only.
- Amazon Elastic Container Services uses Paxos to maintain a consistent view of cluster state.
The sad state of Paxos

About publications...
“The dirty little secret of the NSDI community is that at most five people really, truly understand every part of Paxos ;).” – NSDI reviewer

About implementations...
“There are significant gaps between the description of the Paxos algorithm and the needs of a real-world system...the final system will be based on an unproven protocol.” – Chubby authors
Raft Consensus Protocol

An algorithm to build real systems

- Must be correct, complete, and perform well
- Must be understandable

Key design ideas

- What would be easier to understand or explain?
- Less complexity in state space
- Less mechanisms

Bibliography

Raft implementations

Actual deployments

- HydraBase by Facebook (replacement for Apache HBase)
- Consul by HashiCorp (datacenter management)
- Rafter by Basho (NOSQL key-value store called Riak)
- Apache Kudu (distributed database)
- Kubernetes and Docker Swarm (container management)
Raft implementations

Open-source projects: 80+ total (May 2017)

<table>
<thead>
<tr>
<th>Language</th>
<th>Numbers</th>
<th>Language</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Java</td>
<td>18</td>
<td>Javascript</td>
<td>6</td>
</tr>
<tr>
<td>Go</td>
<td>8</td>
<td>Clojure</td>
<td>4</td>
</tr>
<tr>
<td>Ruby</td>
<td>8</td>
<td>Erlang</td>
<td>4</td>
</tr>
<tr>
<td>C/C++</td>
<td>8</td>
<td>Rust</td>
<td>3</td>
</tr>
<tr>
<td>Scala</td>
<td>7</td>
<td>Bloom</td>
<td>3</td>
</tr>
<tr>
<td>Python</td>
<td>6</td>
<td>Others</td>
<td>9</td>
</tr>
</tbody>
</table>
Table of contents

1 Consensus variants
2 Historical overview
   • Paxos
   • Raft
3 Raft protocol
   • Overview
   • Elections
   • Normal operation
   • Neutralizing old leaders
   • Client protocol
   • Configuration changes
Two approaches to consensus / atomic broadcast / state replication:

- Symmetric, leader-less, active replication:
  - All servers have equal roles
  - Clients can contact any server

- Asymmetric, leader-based, passive replication:
  - At any given time, one server is in charge, others accept its decisions
  - Clients communicate with the leader

Raft is leader-based

- Decomposes the problem (normal operation, leader changes)
- Simplifies normal operation (no conflicts)
- More efficient than leader-less approaches
Raft overview

1. Leader election:
   - Select one of the servers to act as leader
   - Detect crashes, choose new leader

2. Normal operation
   - Basic log replication

3. Safety and consistency after leader changes

4. Neutralizing old leaders

5. Client interactions
   - Implementing linearizeable semantics

6. Configuration changes
   - Adding and removing servers
Server states

| LEADER                | Handles all client interactions, log replication  
|                      | At most 1 viable leader at a time               |
| FOLLOWER             | Completely passive (issues no RPCs, responds to  
|                      | incoming RPCs)                                  |
| CANDIDATE            | Used to elect a new leader                      
|                      | Normal operation: 1 leader, N-1 followers        |
Terms

- Time divided into terms:
  - Election
  - Normal operation under a single leader

- At most one leader per term

- Some terms have no leader (failed election)

- Each server maintains current term value

- Key role of terms: identify obsolete information
Server state

**Persistent state**

Each server persists the following variables to stable storage synchronously before responding to RPCs:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>currentTerm</strong></td>
<td>Latest term server has seen (initialized to 0 on first boot)</td>
</tr>
<tr>
<td><strong>votedFor</strong></td>
<td>ID of the candidate that received vote in current term (or null if none)</td>
</tr>
<tr>
<td><strong>log[]</strong></td>
<td>Log entries:</td>
</tr>
<tr>
<td></td>
<td><strong>term</strong></td>
</tr>
<tr>
<td></td>
<td><strong>command</strong></td>
</tr>
</tbody>
</table>
Server state

Non-persistent state

<table>
<thead>
<tr>
<th>state</th>
<th>Current state; could be LEADER, CANDIDATE, FOLLOWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>leader</td>
<td>ID of the leader</td>
</tr>
<tr>
<td>commitIndex</td>
<td>index of highest log entry known to be committed</td>
</tr>
<tr>
<td>nextIndex[]</td>
<td>index of next log entry to send to peer</td>
</tr>
<tr>
<td>matchIndex[]</td>
<td>index of highest log entry known to be replicated</td>
</tr>
</tbody>
</table>

Initialization

\[
\begin{align*}
\text{currentTerm} & \leftarrow 1 \\
\text{votedFor} & \leftarrow \text{nil} \\
\text{log} & \leftarrow \{\} \\
\text{state} & \leftarrow \text{FOLLOWER} \\
\text{leader} & \leftarrow \text{nil} \\
\text{commitIndex} & \leftarrow 0 \\
\text{nextIndex} & = \{1, 1, \ldots, 1\} \\
\text{matchIndex} & = \{0, 0, \ldots, 0\}
\end{align*}
\]
RPCs

Communication between leader and followers happen through two RPCs:

- **APPEND_ENTRIES**
  - Add an entry to the log, *or*
  - Empty messages used as heartbeats
  - Message tags: APPEND_REQ, APPENDREP

- **VOTE**
  - Message used by candidates to ask votes and win elections
  - Message tags: VOTE_REQ, VOTEREP
Hearthbeats and timeouts

- Servers start up as followers
- Followers expect to receive RPCs from leaders or candidates
- Leaders must send empty `APPENDENTRIES` RPCs to maintain authority
- If $\Delta_{election}$ time units elapse with no RPCs:
  - Follower assumes leader has crashed
  - Follower starts new election
  - Timeouts typically 100-500ms
Election basics - Election start

1. Set new timeout in range \([\Delta_{election}, 2 \cdot \Delta_{election}]\)

2. Increment current term

3. Change to Candidate state

4. Vote for self

5. Send VOTE RPCs to all other servers, retry until either:
   - Receive votes from majority of servers:
     - Become Leader
     - Send APPENDENTRIES heartbeats to all other servers
   - Receive APPENDENTRIES from valid leader:
     - Return to Follower state
   - No one wins election (election timeout elapses):
     - Start new election
Election - Pseudocode

Election code - executed by process $p$

upon timeout $\langle$ELECTIONTIMEOUT$\rangle$ do

  if $state \in \{$FOLLOWER, CANDIDATE$\}$ then
    $t \leftarrow$ random$(1.0, 2.0) \cdot \Delta_{election}$
    set timeout $\langle$ELECTIONTIMEOUT$\rangle$ at $\text{now}() + t$
    $currentTerm \leftarrow currentTerm + 1$
    $state \leftarrow$ CANDIDATE
    $votedFor \leftarrow p$
    $votes \leftarrow \{p\}$

  foreach $q \in \Pi$ do
    cancel timeout $\langle$RPCTIMEOUT, $q$}$
    set timeout $\langle$RPCTIMEOUT, $q$} at $\text{now}()$
Election - Pseudocode

RPC timeout code - executed by process $p$

upon timeout $\langle \text{RpcTimeout}, q \rangle$ do

    if state = CANDIDATE then
        set timeout $\langle \text{RpcTimeout}, q \rangle$ at now() + $\Delta_{vote}$
        send $\langle \text{VoteReq}, currentTerm \rangle$ to $q$
Election - Pseudocode

Election code - executed by process $p$

```plaintext
on receive ⟨VoteReq, term⟩ from q do
    if term > currentTerm then
        stepdown(term)
    if term = currentTerm and votedFor ∈ {q, nil} then
        votedFor ← q
        $t ← \text{random}(1.0, 2.0) \cdot \Delta_{\text{election}}$
        set timeout ⟨ELECTION_TIMEOUT⟩ at now() + $t$
        send ⟨VOTEREP, term, votedFor⟩ to q
```
Election - Pseudocode

Election code - executed by process $p$

```plaintext
on receive $\langle$VoteRep, term, vote$\rangle$ from $q$ do
    if term > currentTerm then
        stepdown(term)
    if term = currentTerm and state = CANDIDATE then
        if vote = p then
            votes ← votes $\cup \{q\}$
        cancel timeout $\langle$RpcTimeout, q$\rangle$
    if $|votes| > |\Pi|/2$ then
        state ← LEADER
        leader ← p
        foreach $q \in P - \{p\}$ do
            sendAppendEntries(q)
```

Alberto Montresor (UniTN)  
DS - Raft Consensus
procedure stepdown(term)
    currentTerm ← term
    state ← FOLLOWER
    votedFor ← nil
    \[ t ← \text{random}(1.0, 2.0) \cdot \Delta_{\text{election}} \]
    \[ \text{set timeout } \langle \text{ELECTIONTIMEOUT} \rangle \text{ at } \text{now()} + t \]
Election - Correctness

Safety: allow at most one winner per term

- Each server gives out only one vote per term (persist on disk)
- Two different candidates can’t accumulate majorities in same term

Liveness: some candidate must eventually win

- Choose election timeouts randomly in $[\Delta_{\text{election}}, 2 \cdot \Delta_{\text{election}}]$
- One server usually times out and wins election before others wake up
- Works well if $\Delta_{\text{election}} >>$ broadcast time
Randomize timeouts

- How much randomization is needed to avoid split votes?
- Conservatively, use random range $\approx 10 \times$ network latency
Log structure

- Log stored on stable storage (disk); survives crashes
- Entry committed if known to be stored on majority of servers
- Durable, will eventually be executed by state machines

Log entry = index, term, command

- Log stored on stable storage (disk); survives crashes
- Entry committed if known to be stored on majority of servers
- Durable, will eventually be executed by state machines
Normal operation

- Client sends command to leader
- Leader appends command to its log

Normal operation code executed by process $p$

\begin{verbatim}
upon receive ⟨REQUEST, command⟩ from client do
    if state = LEADER then
        log.append(currentTerm, command)
        foreach $q \in P - \{p\}$ do
            sendAppendEntries($q$)
\end{verbatim}
Normal operation

- Leader sends `APPENDENTRIES` RPCs to followers

- Once new entry committed:
  - Leader passes command to its state machine, returns result to client
  - Leader notifies followers of committed entries in subsequent AppendEntries RPCs
  - Followers pass committed commands to their state machines

- Crashed/slow followers?
  - Leader retries RPCs until they succeed
  - Performance is optimal in common case: one successful RPC to any majority of servers
Normal operation

RPC timeout code executed by process $p$

upon timeout $\langle\text{RpcTimeout}, q\rangle$ do

if $state = \text{CANDIDATE}$ then
  set timeout $\langle\text{RpcTimeout}, q\rangle$ at now() + $\Delta_{vote}$
  send $\langle\text{VoteReq}, currentTerm\rangle$ to $q$

if $state = \text{LEADER}$ then
  sendAppendEntries($q$)
How to send append entries

procedure sendAppendEntries(q)

set timeout ⟨RPC_TIMEOUT, q⟩ at now() + Δ_{election}/2

lastLogIndex ← choose in[nextIndex[q], log.len()]

nextIndex[q] = lastLogIndex

send

⟨APPENDREQ, term, lastLogIndex − 1, log[lastLogIndex[q] − 1].term,
log[lastLogIndex...log.len()], commitIndex⟩ to q
Log consistency

Consistency in logs

- If log entries on different servers have the same index and term:
  - They store the same command
  - The logs are identical in all preceding entries

- If a given entry is committed, all preceding entries are also committed
APPENDENTRIES Consistency Check

- Each APPENDENTRIES RPC contains index, term of entry preceding new ones
- Follower must contain matching entry; otherwise it rejects request
- Implements an induction step, ensures coherency

```
1 add T1
2 cmp T1
3 ret T2
4 mov
5 jmp T3
```

**leader**
- T1
- T1
- T1
- T2
- T3

**follower**
- T1
- T1
- T1
- T2
- T3

AppendEntries succeeds: matching entry

```
1 add T1
2 cmp T1
3 ret T1
4 mov
5 jmp T3
```

**leader**
- T1
- T1
- T1
- T2
- T3

**follower**
- T1
- T1
- T1
- T1
- T1

AppendEntries fails: mismatch
**APPENDENTRIES Consistency Check**

- Each **APPENDENTRIES** RPC contains index, term of entry preceding new ones
- Follower must contain matching entry; otherwise it rejects request
- Implements an **induction** step, ensures coherency

![Diagram showing the consistency check process for APPENDENTRIES.](image)

**AppendEntries succeeds:** matching entry
Normal operation - Pseudocode

Normal operation code - executed by process $p$

```
on receive ⟨APPENDREQ, term, prevIndex, prevTerm, entries, commitIndex⟩ from q do
  if term > currentTerm then
    stepdown(term)
  if term < currentTerm then
    send ⟨APPENDREP, currentTerm, false⟩ to q
  else
    index ← 0
    success ← prevIndex = 0 or (prevIndex ≤ log.len() and log[prevIndex].term = prevTerm)
    if success then
      index ← storeEntries(prevIndex, entries, commitIndex)
    send ⟨APPENDREP, currentTerm, success, index⟩
```
At beginning of new leader’s term

- Old leader may have left entries partially replicated
- No special steps by new leader: just start normal operation
- Leader’s log is “the truth”
- Will eventually make follower’s logs identical to leader’s
- Multiple crashes can leave many extraneous log entries
Safety Requirement

Once a log entry has been applied to a state machine, no other state machine must apply a different value for that log entry.

- Raft safety property:
  - If a leader has decided that a log entry is committed, that entry will be present in the logs of all future leaders.
  - This guarantees the safety requirement.

- Leaders never overwrite entries in their logs:
  - Only entries in the leader’s log can be committed.
  - Entries must be committed before applying to state machine.

\[\text{Committed} \rightarrow \text{Present in future leaders’ logs}\]
Picking the Best Leader

- Can’t tell which entries are committed!

- During elections, choose candidate with log most likely to contain all committed entries
  - Candidates include index & term of last log entry in VOTEREQ
  - Voting server \( V \) denies vote if its log is “more complete”:
    \[
    \begin{align*}
    (\text{lastLogTerm}_C < \text{lastLogTerm}_V) \text{ or } \\
    (\text{lastLogTerm}_C = \text{lastLogTerm}_V \text{ and } \text{lastLogIndex}_C < \text{lastLogIndex}_V)
    \end{align*}
    \]
  - Leader will have “most complete” log among electing majority

\[
\begin{array}{c}
T_1 \ T_1 \ T_1 \ T_2 \ T_2 \\
T_1 \ T_1 \ T_1 \ T_2 \\
T_1 \ T_1 \ T_1 \ T_2 \ T_2 \\
\end{array}
\]
Election - Modified pseudocode

RPC timeout code - executed by process $p$

\[
\text{upon timeout} \langle \text{RpcTimeout}, q \rangle \text{ do}
\]
\[
\quad \text{if } state = \text{CANDIDATE} \text{ then}
\]
\[
\quad \quad \text{set timeout} \langle \text{RpcTimeout}, q \rangle \text{ at now()} + \Delta_{vote}
\]
\[
\quad \quad \text{lastLogTerm} \leftarrow \log[\log\text{.len}()].\text{term}
\]
\[
\quad \quad \text{lastLogIndex} \leftarrow \log\text{.len}()
\]
\[
\quad \quad \text{send} \langle \text{VOTEREQ, currentTerm, lastLogTerm, lastLogIndex} \rangle \text{ to } q
\]
\[
\quad \text{if } state = \text{LEADER} \text{ then}
\]
\[
\quad \quad \text{set timeout} \langle \text{RpcTimeout}, q \rangle \text{ at now()} + \Delta_{election}/2
\]
\[
\quad \quad \text{sendAppendEntries}(q)
\]
Election - Modified pseudocode

Election code - executed by process \( p \)

```plaintext
on receive \( \langle \text{VOTEREQ}, \text{term}, \text{lastLogTerm}, \text{lastLogIndex} \rangle \) from \( q \) do

  if \( \text{term} > \text{currentTerm} \) then
    stepdown(\( \text{term} \))

  if \( \text{term} = \text{currentTerm} \) and \( \text{votedFor} \in \{ q, \text{nil} \} \) and
    (\( \text{lastLogTerm} > \text{log[log.len()].term} \) or
    (\( \text{lastLogTerm} = \text{log[log.len()].term} \) and \( \text{lastLogIndex} \geq \text{log.len()} \)) )
  then

    \( \text{votedFor} \leftarrow q \)
    \( t \leftarrow \text{random}(1.0, 2.0) \cdot \Delta_{\text{election}} \)
    set timeout \( \langle \text{ELECTIONTIMEOUT} \rangle \) at \( \text{now()} + t \)
    send \( \langle \text{VOTEREP}, \text{term}, \text{votedFor} \rangle \)
```

Alberto Montresor (UniTN)
Case 1/2: Leader decides entry in current term is committed

Safe: leader for term $T_3$ must contain entry 4
Committing Entry from Earlier Terms

Case 2/2: Leader is trying to commit entry from an earlier term

Unsafe: Entry 3 not safely committed

- \( s_5 \) can be elected as leader for term \( T_5 \)
- If elected, it will overwrite entry 3 on \( s_1, s_2, \) and \( s_3 \)!
New commitment rule

- For a leader to decide that an entry is committed:
  - Must be stored on a majority of servers
  - At least one new entry from leader’s term must also be stored on majority of servers

- Once entry 4 committed:
  - $s_5$ cannot be elected leader for term $T_5$
  - Entries 3 and 4 both safe

Combination of election and commitment rules makes Raft safe
Log inconsistencies

Leader changes can result in log inconsistencies:

<table>
<thead>
<tr>
<th>log index</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>leader for term 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(f)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Possible followers

- Missing Entries
- Extraneous Entries

March 3, 2013 Raft Consensus Algorithm Slide 22

Extraneous Entries

Missing Entries

Alberto Montresor (UniTN)  DS - Raft Consensus  2020/11/26  49 / 60
Repairing follower log

- New leader must make follower logs consistent with its own
  - Delete extraneous entries
  - Fill in missing entries

- Leader keeps $nextIndex$ for each follower:
  - Index of next log entry to send to that follower
  - Initialized to $(1 + \text{leader’s last index})$

- When APPENDENTRIES consistency check fails, decrement $nextIndex$ and try again
Repairing follower log – Pseudocode

Normal operation code - executed by process $p$

upon receive $\langle$APPENDREP, $term$, success, $index$$\rangle$ from $q$ do
  if $term > currentTerm$ then
    stepdown($term$)
  else if $state = LEADER$ and $term = currentTerm$ then
    if $success$ then
      $nextIndex[q] \leftarrow index + 1$
    else
      $nextIndex[q] \leftarrow \max(1, nextIndex[q] - 1)$
    fi
  fi
  if $nextIndex[q] \leq \log$ .len() then
    sendAppendEntries($q$)
When follower overwrites inconsistent entry, it deletes all subsequent entries.
Repairing follower log

```plaintext
procedure storeEntries (prevIndex, entries, c)
    index ← prevIndex
    for j ← 1 to entries.len() do
        index ← index + 1
        if log[index].term ≠ entries[j].term then
            log = log[1...index-1] + entries[j]
        commitIndex ← min(c, index)
    return index
```

Neutralizing Old Leaders

- Deposed leader may not be dead
  - Temporarily disconnected from network
  - Other servers elect a new leader
  - Old leader becomes reconnected, attempts to commit log entries

Terms used to detect stale leaders (and candidates)

- Every RPC contains term of sender
  - If sender’s term is older, RPC is rejected, sender reverts to follower and updates its term
  - If receiver’s term is older, it reverts to follower, updates its term, then processes RPC normally

Election updates terms of majority of servers

- Deposed server cannot commit new log entries
Neutralizing Old Leaders

Normal operation code - executed by process $p$

```plaintext
on receive ⟨APPENDREQ, term, prevIndex, prevTerm, . . .⟩ from q do
  if term > currentTerm then
    stepdown(term)
  if term < currentTerm then
    send ⟨APPENDREP, currentTerm, false⟩ to q
  else
    [. . .]
```
Client protocol

Clients sends commands to leader:

- If leader unknown, contact any server
- If contacted server not leader, it will redirect to leader

Leader responds when:

- command has been logged
- command has been committed
- command has been executed by leader’s state machine

If request times out (e.g., leader crash):

- Client re-issues command to some other server
- Eventually redirected to new leader
- Retry request with new leader
What if leader crashes after executing command, but before responding?

- Must not execute command twice

Solution: client embeds a unique id in each command

- Server includes id and response in log entry
- Before accepting command, leader checks its log for entry with that id
- If id found in log, ignore new command, return response from old command

Result: exactly-once semantics as long as client doesn’t crash
Configuration

System configuration

- ID, address for each server
- Determines what constitutes a majority

Consensus mechanism must support changes in the configuration

- Replace failed machine
- Change degree of replication
Configuration changes

Cannot switch directly from one configuration to another: conflicting majorities could arise.

Diagram:

- $C_{\text{old}}$ and $C_{\text{new}}$
- Server 1, Server 2, Server 3, Server 4, Server 5
- Majority of $C_{\text{old}}$
- Majority of $C_{\text{new}}$
Joint consensus

Raft uses a 2-phase approach

- Intermediate phase uses joint consensus (need majority of both old and new configurations for elections, commitment)
- Once joint consensus is committed, begin replicating log entry for final configuration
D. Ongaro and J. Ousterhout. *In search of an understandable consensus algorithm.*
