A Brief Introduction to Paxos

by

Dion Harmon and Jason Burns

Paxos (Synod) Algorithm

- Distributed consensus
- Assumption: stopping failures only.
- Basic idea: Processors take ballots until at least one succeeds.

Ballots

- Ballots have the following properties
  - Number: An identifier from a totally ordered set.
  - Value: Some value that the nodes are to agree upon.
  - Quorum: Some set of nodes.
  - Approved: Set of nodes that have approved the decree.
- A ballot B is "passed" if $B_{qrm} \subseteq B_{approved}$. Nodes decide on the value of a passed ballot.

Safety

Safety is enforced by ensuring that for any set $\mathfrak{B}$ of ballots, the following properties hold:

- $B1(\mathfrak{B})$: Every ballot has a unique number.
- $B2(\mathfrak{B})$: The intersection of any two quorums is never empty.
- $B3(\mathfrak{B})$: If any processor in B_{qrm} approved an earlier (smaller numbered) ballot then B_{val} is the value of the latest (largest numbered) of these ballots.

We need to show that no two passed ballots have different values.

- Suppose B is the earliest ballot in $\mathfrak{B}$ that is passed. If none exist, safety holds.
- Let B' be the earliest ballot after B with a different value than B. If none exists, safety holds.
- As B passed, $B_{qrm} \subseteq B_{approved}$. This implies $B_{approved} \cap B'_{qrm} \neq \emptyset$. Thus some processor in B'_{qrm} approved B.
- Let C be the latest ballot before B' which a processor from B'_{qrm} approved. Such a C exists and $C_{num} \geq B_{num}$ as some processor in B'_{qrm} approved B. $B3$ ensures $B'_{val} = C_{val}$. This implies $C_{num} > B_{num}$.
- Since C is later than B and earlier than B', but $C_{val} = B'_{val} \neq B_{val}$, we have a contradiction: B' is not the earliest ballot after B to have a different value than B.

The Paxos (Synod) Implementation Summarized

Phase 1: Proposer sends *request* for ballot n to choosers.

- Choosers *accept* request IF no higher request accepted.
- (Choosers also send last ballot approved, if any.)

If a quorum of choosers accept:

Phase 2: Proposer sends *ballot* n, with value v to this quorum.

If any choosers have approved a ballot, v is the value of the highest numbered ballot approved.

- Choosers *approve* ballot IF no higher request accepted.
- When proposer learns all choosers in the quorum have approved the ballot, it decides on the value of the ballot and tells everyone of the success.

Liveness Issues

- Problem: Multiple Proposers
- Solution: None
- Partial Solution: Real-Time Elections
- Leadership Achieves Liveness

Multiple Decrees

- Electing a Leader
- Passing Decrees in Order
- Paxos is efficient!

Making a Better State Machine

- Fast Reads vs. Slow Reads
- Specialists
  - We can assign functions to individual nodes for efficiency
  - Multiple specialists
  - Leases guarantee fault-tolerance
- Self-Stabilization

Extensions to Paxos

- Byzantine failures
- Dynamic addition/deletion of nodes