“Paxos Made Moderately Complex”
Made Moderately Simple
State machine replication

Reminder: want to agree on order of ops

Can think of operations as a log
Put k1 v1

Put k2 v2

Paxos

Paxos

Phase 1
- Send prepare messages
- Pick value to accept

Phase 2
- Send accept messages
Can we do better?

Phase 1: “leader election”
- Deciding whose value we will use

Phase 2: “commit”
- Leader makes sure it’s still leader, commits value

What if we split these phases?
- Lets us do operations with one round-trip
Roles in PMMC

Replicas (like learners)
- Keep log of operations, state machine, configs

Leaders (like proposers)
- Get elected, drive the consensus protocol

Acceptors (*simpler* than in Paxos Made Simple!)
- “Vote” on leaders
A note about ballots in PMMC

(leader, seqnum) pairs

Isomorphic to the system we discussed earlier

0, 4, 8, 12, 16, ...

1, 5, 9, 13, 17, ...

2, 6, 10, 14, 18, ...

3, 7, 11, 15, 19, ...
A note about ballots in PMMC

(leader, seqnum) pairs

Isomorphic to the system we discussed earlier

0 0.0, 1.0, 2.0, 3.0, 4.0, ...

1 0.1, 1.1, 2.1, 3.1, 4.1, ...

2 0.2, 1.2, 2.2, 3.2, 4.2, ...

3 0.3, 1.3, 2.3, 3.3, 4.3, ...
Paxos Made Moderately Complex Made Simple

Diagram showing the protocol of Paxos involving a client, replicas, a leader, and acceptors. The process includes steps such as request, propose, adopted, decision, scout, p1a, p1b, commander, p2a, p2b, and response.
Paxos Made Moderately Complex Made Simple
Acceptors

Acceptors

**ballot_num**: 0
**accepted**: []
Acceptors

p1a(0.1) → Acceptor

ballot_num: _
accepted: []
Acceptors

p1a(0.1) -> Acceptor

ballot_num: 0.1
accepted: []
Acceptors

- \( p_1\alpha(0.1) \)
- \( p_1\beta([]) \)

Acceptors

ballot_num: 0.1
accepted: []
Acceptors

Acceptors

ballot_num: 0.1
accepted:[]
Acceptors

p1a(0.0) → Acceptor

ballot_num: 0.1
accepted: []
Acceptors

p1a(0.0)

Nope!

Acceptor

ballot_num: 0.1
accepted:[]
Acceptors

Acceptor

```
ballot_num: 0.1
accepted:[]
```
Acceptors

\[ \text{p2a}(\langle 0.1, 0, A \rangle) \]

Acceptor

\[
\begin{align*}
\text{ballot_num: } & 0.1 \\
\text{accepted: } & []
\end{align*}
\]
Acceptors

\[ p2a(<0.1, 0, A>) \]

**Acceptor**

- `ballot_num: 0.1`
- `accepted: [<0.1, 0, A>]`
Acceptors

p2a(<0.1, 0, A>)

OK!

Acceptors

ballot_num: 0.1
accepted:[<0.1, 0, A>]

OK!
Acceptors

Acceptor

ballot_num: 0.1
accepted:[<0.1, 0, A>]
Acceptors

\[ p2a(<0.0, 0, B>) \]

Acceptors

- `ballot_num: 0.1`
- `accepted:[<0.1, 0, A>]`
Acceptors

p2a(<0.0, 0, B>)

Nope!

Accepter

ballot_num: 0.1
accepted: [<0.1, 0, A>]
Acceptors

Acceptors

ballot_num: 0.1
accepted:[<0.1, 0, A>]
Acceptors

- Ballot numbers increase
- Only accept values from current ballot
- Never remove ballots
- If a value $v$ is chosen by a majority on ballot $b$, then any value accepted by any acceptor in the same slot on ballot $b' > b$ has the same value
Paxos Made Moderately Complex Made Simple

- Client \( \kappa \) to Replicas \( \rho_1, \rho_2 \)
- Leader \( \lambda \) to Acceptors \( \alpha_1, \alpha_2, \alpha_3 \)

- Request from client to replicas
- Propose from replicas to leader
- Adopted from leader to acceptors
- Command from leader to acceptors
- Decision from leader to replicas
- Response from replicas to client

- Path 1: \( p1a \) from \( \alpha_1 \) to \( \rho_1 \)
- Path 2: \( p1b \) from \( \alpha_2 \) to \( \rho_2 \)
- Path 3: \( p2a \) from \( \alpha_3 \) to \( \lambda \)
- Path 4: \( p2b \) from \( \lambda \) to \( \alpha_1 \)
Paxos Made Moderately Complex Made Simple

Diagram showing the process of Paxos, with arrows indicating request, propose, adopt, and decision phases between client, replicas, leader, and acceptors.
Leader: Getting Elected

Leader

active: false
ballot_num: 0.0
proposals: []
Leader: Getting Elected

Leader

active: false
ballot_num: 0.0
proposals: []
Leader: Getting Elected

Leader

active: false
ballot_num: 0.0
proposals: []

Nope!

Acceptor

Acceptor

Acceptor

Nope!
Leader: Getting Elected

Leader

active: false
ballot_num: 1.0
proposals: []

Acceptor

Acceptor

Acceptor
Leader: Getting Elected

Leader

- active: false
- ballot_num: 1.0
- proposals: []

Acceptor

Or...

Acceptor

Acceptor
Leader: Getting Elected

Leader

active: false
ballot_num: 0.0
proposals: []

Acceptors

OK([])!
Leader: Getting Elected

Leader

active: true
ballot_num: 0.0
proposals: []

Acceptor

Acceptor

Acceptor
Paxos Made Moderately Complex Made Simple
Paxos Made Moderately Complex Made Simple
Leader: Handling proposals

Leader

active: true
ballot_num: 0.0
proposals: []

Op1 should be A
(A = “Put k1 v1”)

Acceptor

Replica
Leader: Handling proposals

Leader

active: true
ballot_num: 0.0
proposals: [<1, A>]

Acceptors

Replica
Leader: Handling proposals

Leader

- active: true
- ballot_num: 0.0
- proposals: [(1, A)]

Acceptor

Replica
Leader: Handling proposals

Leader

active: true
ballot_num: 0.0
proposals: [<1, A>]

Acceptor

Nope!

Acceptor

Nope!

Acceptor

Nope!

Replica
Leader: Handling proposals

Leader

active: false
ballot_num: 0.0
proposals: [<1, A>]

Acceptor

Acceptor

Acceptor

Replica
Leader: Handling proposals

Leader

active: false
ballot_num: 0.0
proposals: [<1, A>]

Or...

Acceptor

Replica
Leader: Handling proposals

Leader

active: true
ballot_num: 0.0
proposals: [<1, A>]

Acceptor
Acceptor
Acceptor

OK!
OK!
OK!
Leader: Handling proposals

Leader

active: true
ballot_num: 0.0
proposals: [<1, A>]

Op1 is A

Replica

Replica

Replica

Acceptor

Acceptor

Acceptor
Paxos Made Moderately Complex Made Simple
Election revisited

Leader

active: false
ballot_num: 3.0
proposals: [<1, B>]

Acceptor

ballot_num: 2.1
accepted: [<2.1, 1, A>]
Election revisited

Leader

active: false
ballot_num: 3.0
proposals: [<1, B>]

Acceptor

ballot_num: 2.1
accepted: [<2.1, 1, A>]

p1a(3.0)
Election revisited

Leader

active: false
ballot_num: 3.0
proposals: [<1, B>]

Acceptor

ballot_num: 3.0
accepted: [<2.1, 1, A>]
Election revisited

Leader

OK([<2.1, 1, A>])

Acceptor

active: false
ballot_num: 3.0
proposals: [<1, B>]

ballot_num: 3.0
accepted:[<2.1, 1, A>]
Election revisited

Leader

active: true
ballot_num: 3.0
proposals: [<1, A>]

Acceptor

ballot_num: 3.0
accepted: [<2.1, 1, A>]
Leaders

- Only propose one value per ballot and slot.
- If a value $v$ is chosen by a majority on ballot $b$, then any value proposed by any leader in the same slot on ballot $b' > b$ has the same value.
Paxos Made Moderately Complex Made Simple
Paxos Made Moderately Complex Made Simple
Replicas

Replica

Put k1 v1  Put k2 v2

Replicas

Replica

slot_out

Put k1 v1  Put k2 v2  App k1 v1  App k2 v2

slot_in

Replicas

Replica

Leader

decision(3, “App k1 v1”)

slot_out

Put k1 v1
Put k2 v2
App k1 v1
App k2 v2

slot_in

Op1
Op2
Op3
Op4
Op5
Op6
Replicas

Put k1 v1
Put k2 v2
App k1 v1
App k2 v2

slot_out slot_in

Replicas

decision(4, “Put k3 v3”)

Replica

slot_out slot_in

Put k1 v1  Put k2 v2  App k1 v1  App k2 v2

Replicas

propose(5, "App k2 v2")

slot_out slot_in

Put k1 v1  Put k2 v2  App k1 v1  Put k3 v3  App k2 v2

Paxos Made Moderately Complex Made Simple
When to run for office

When should a leader try to get elected?

- At the beginning of time
- When the current leader seems to have failed

Paper describes an algorithm, based on pinging the leader and timing out

If you get preempted, don’t immediately try for election again!
Reconfiguration

All replicas *must* agree on who the leaders and acceptors are.

How do we do this?
Reconfiguration

All replicas *must* agree on who the leaders and acceptors are

How do we do this?

- Use the log!
- Commit a special reconfiguration command
- New config applies after WINDOW slots
Reconfiguration

What if we need to reconfigure *now* and client requests aren’t coming in?
Reconfiguration

What if we need to reconfigure now and client requests aren’t coming in?

- Commit no-ops until WINDOW is cleared
Other complications

State simplifications
- Can track much less information, esp. on replicas

Garbage collection
- Unbounded memory growth is bad
- Lab 3: track finished slots across all instances, garbage collect when everyone has learned result

Read-only commands
- Can’t just read from replica (why?)
- But, don’t need their own slot