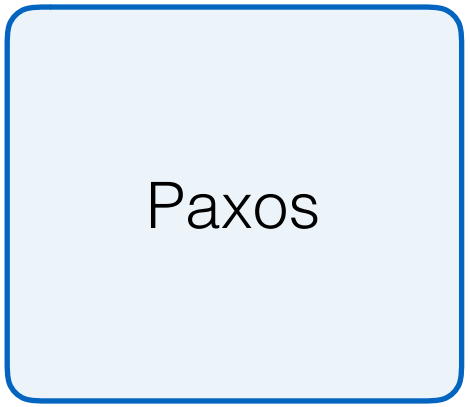
Paxos Made Moderately Complex

Arvind Krishnamurthy
University of Washington

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!)

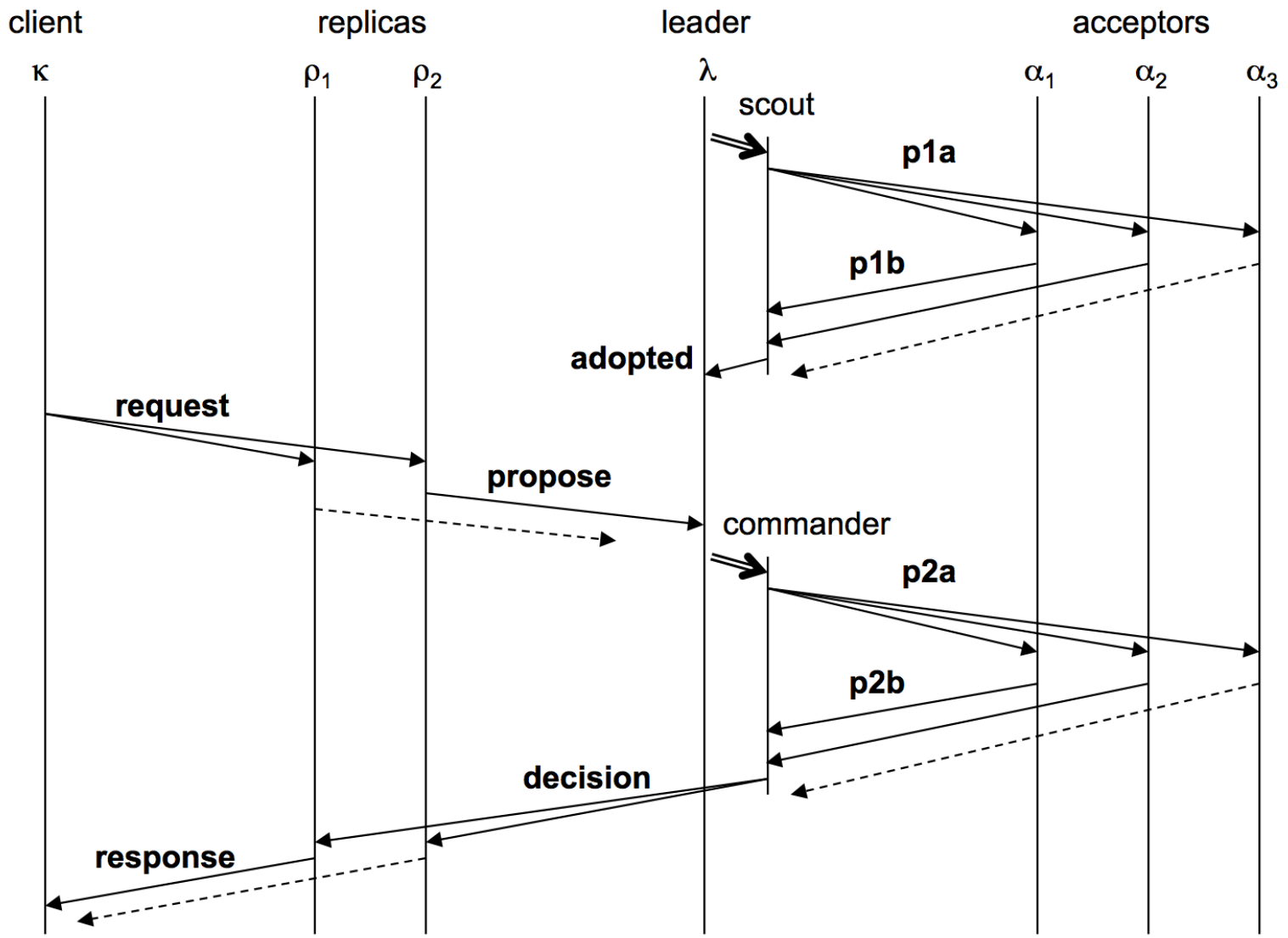
- “Vote” on leaders

Ballots (or proposal #s) in PMMC

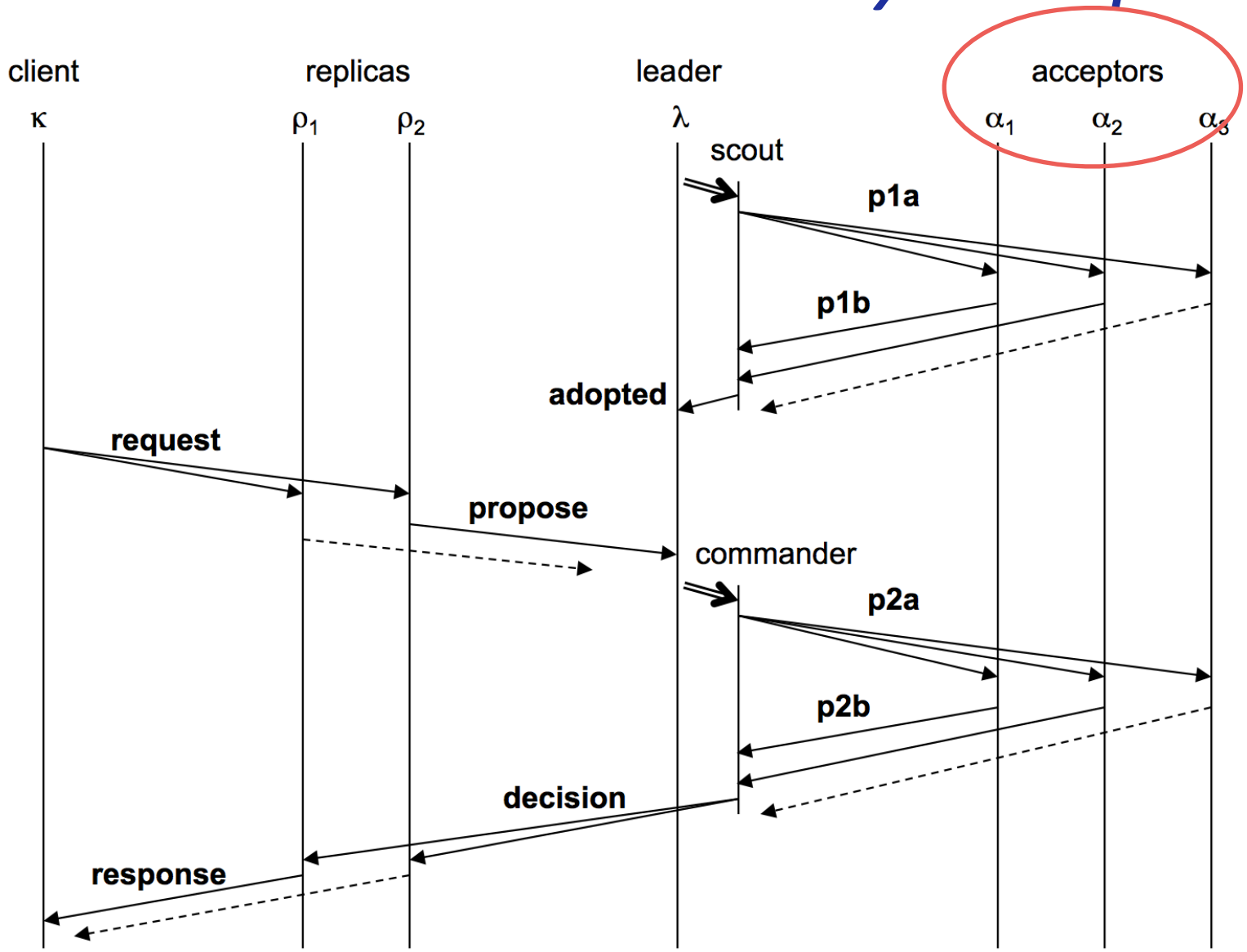
(leader, seqnum) pairs

- 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



Paxos Made Moderately Complex

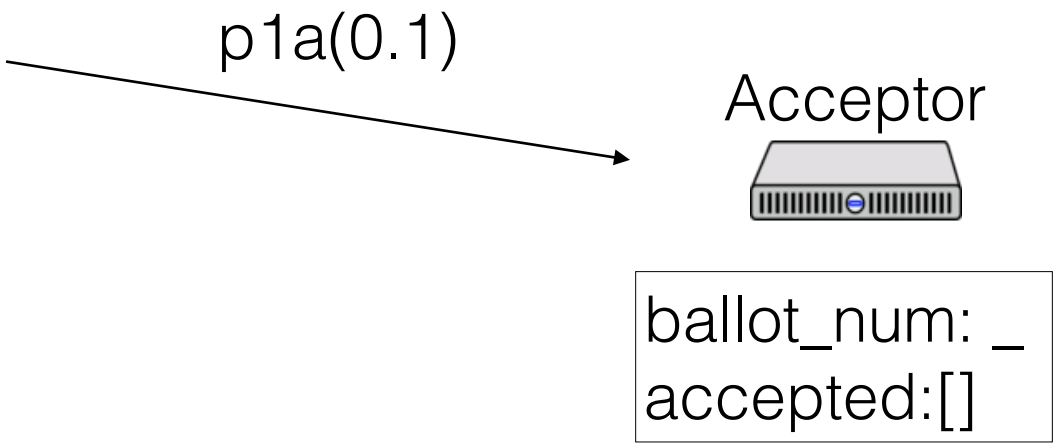


Acceptors

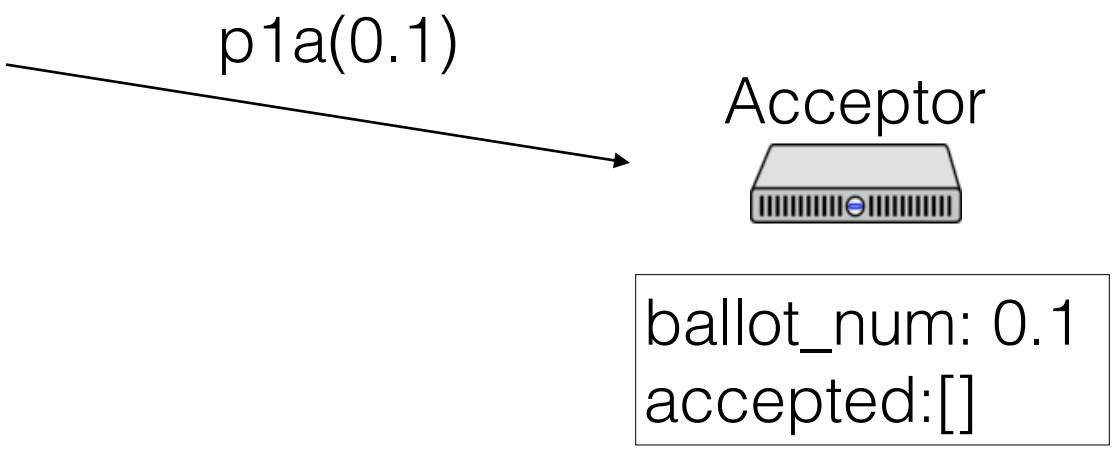
Acceptor

ballot_num: 0
accepted:[]

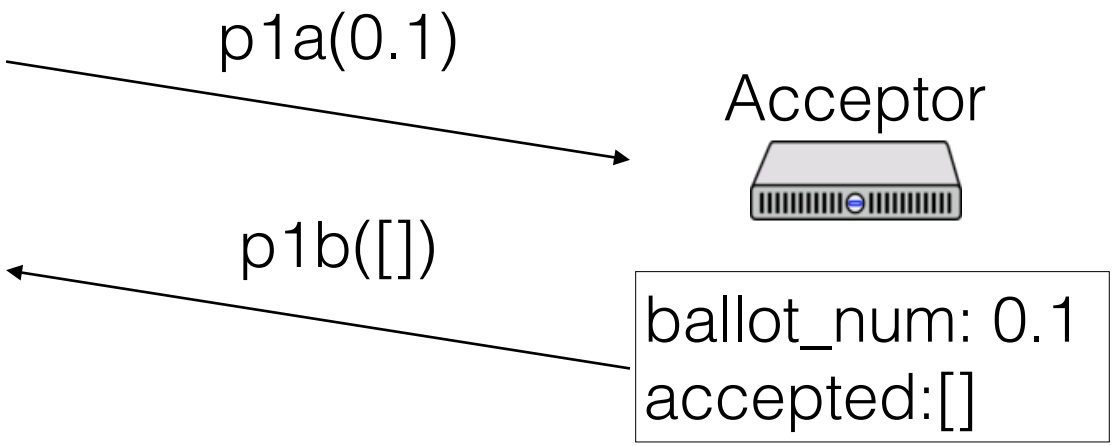
Acceptors



Acceptors



Acceptors

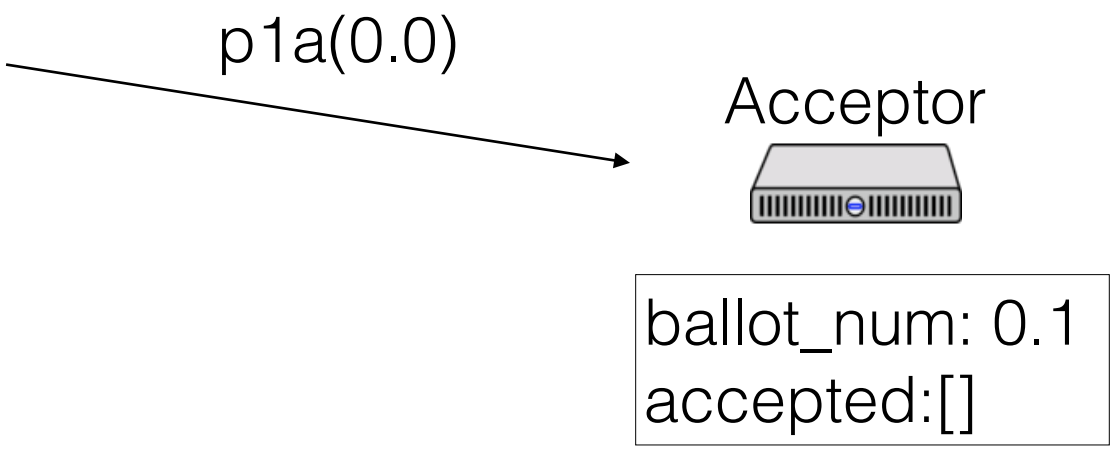


Acceptors

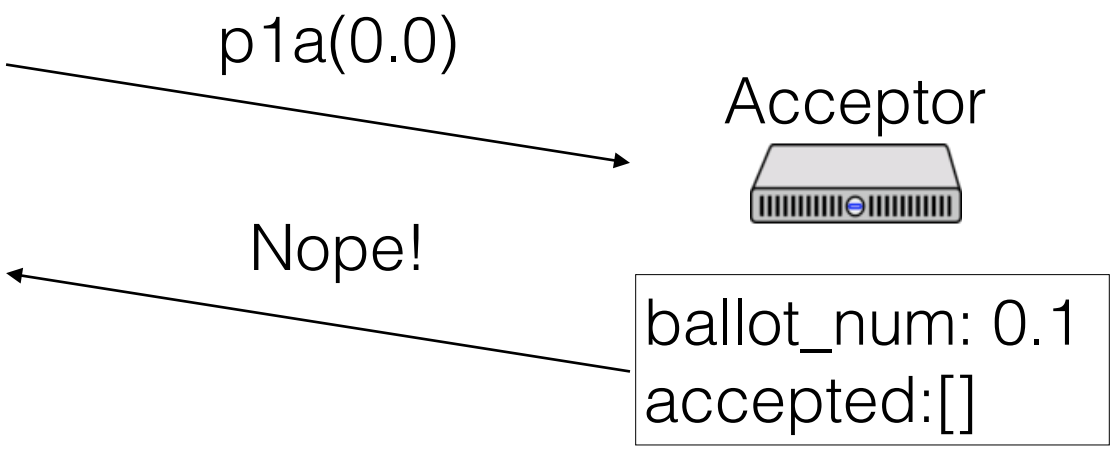
Acceptor

ballot_num: 0.1
accepted:[]

Acceptors



Acceptors

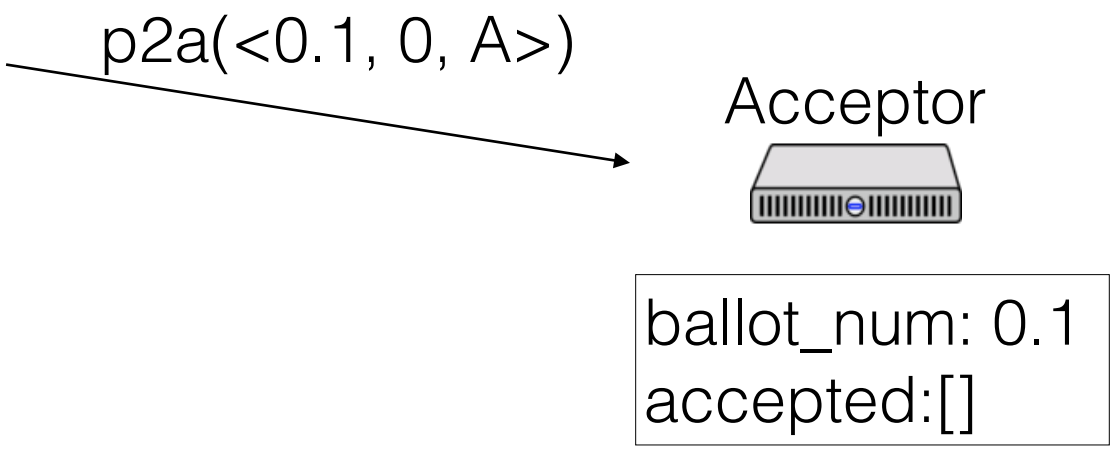


Acceptors

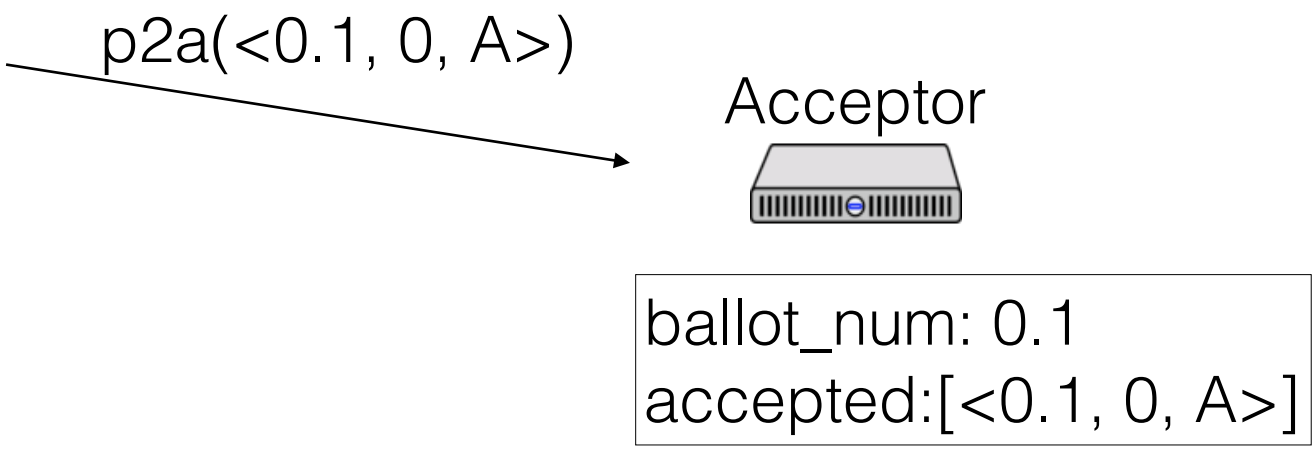
Acceptor

ballot_num: 0.1
accepted:[]

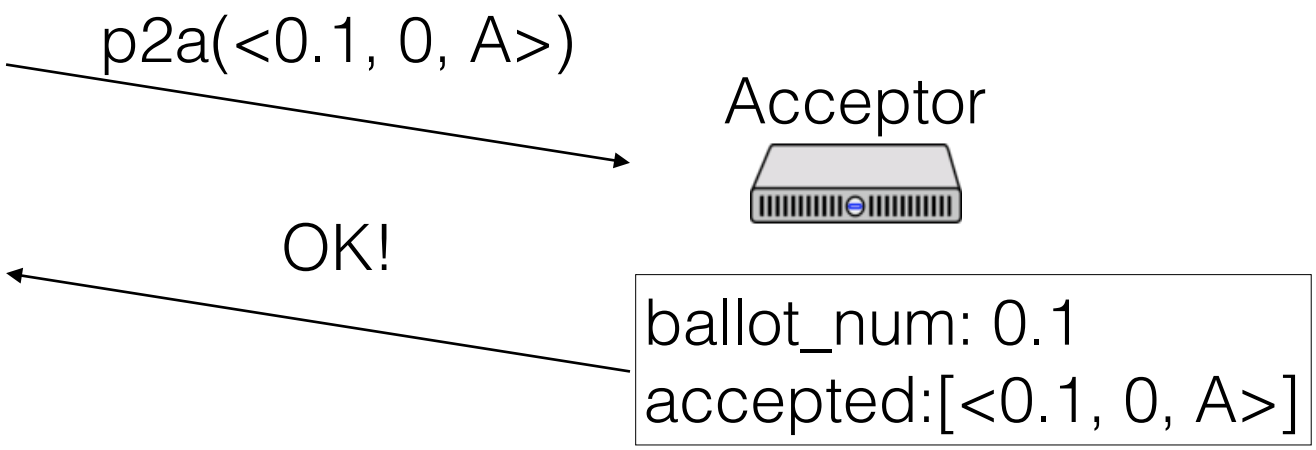
Acceptors



Acceptors



Acceptors

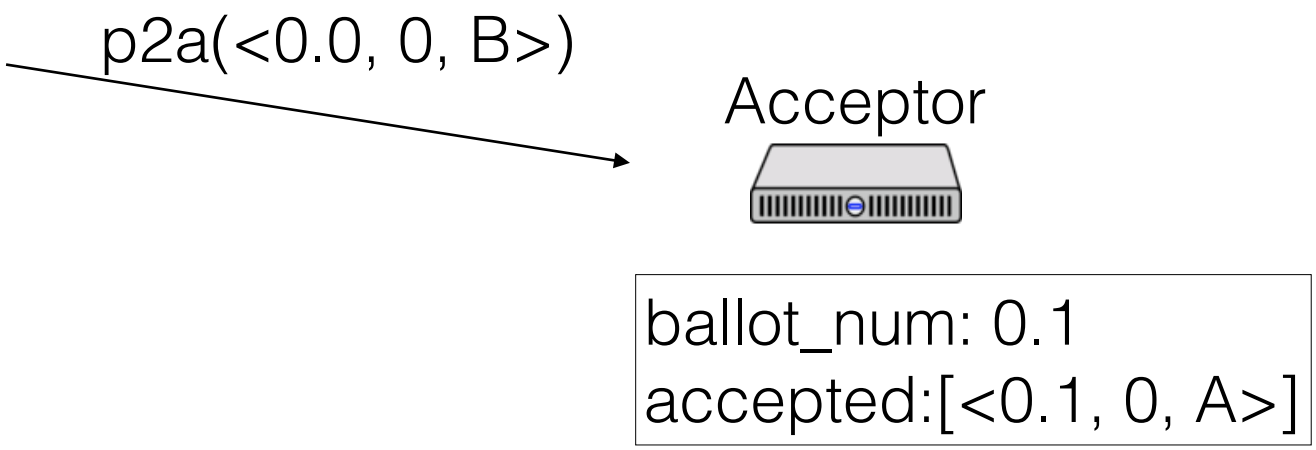


Acceptors

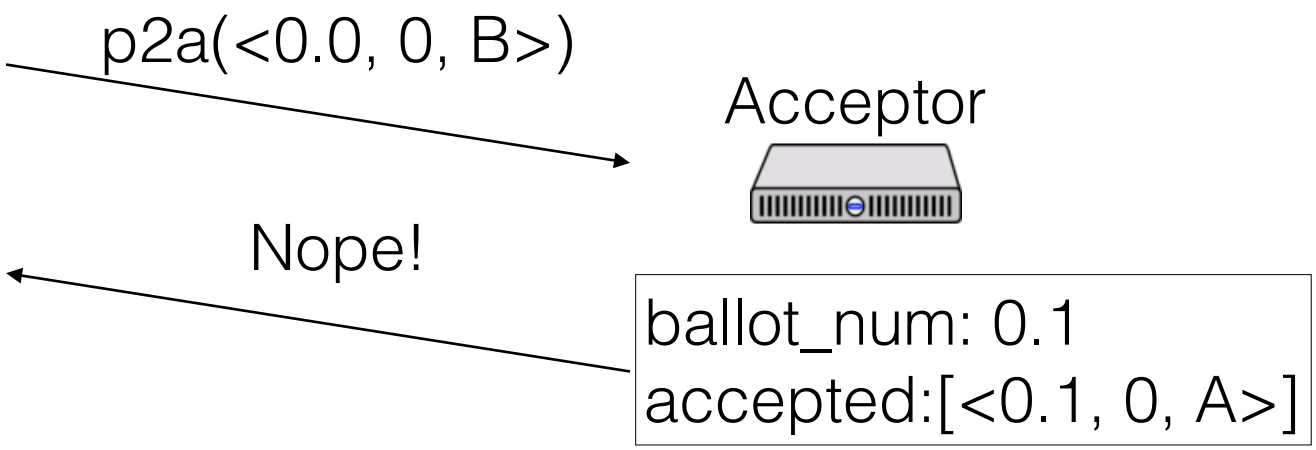
Acceptor

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

Acceptors



Acceptors



Acceptors

Acceptor

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

Acceptors

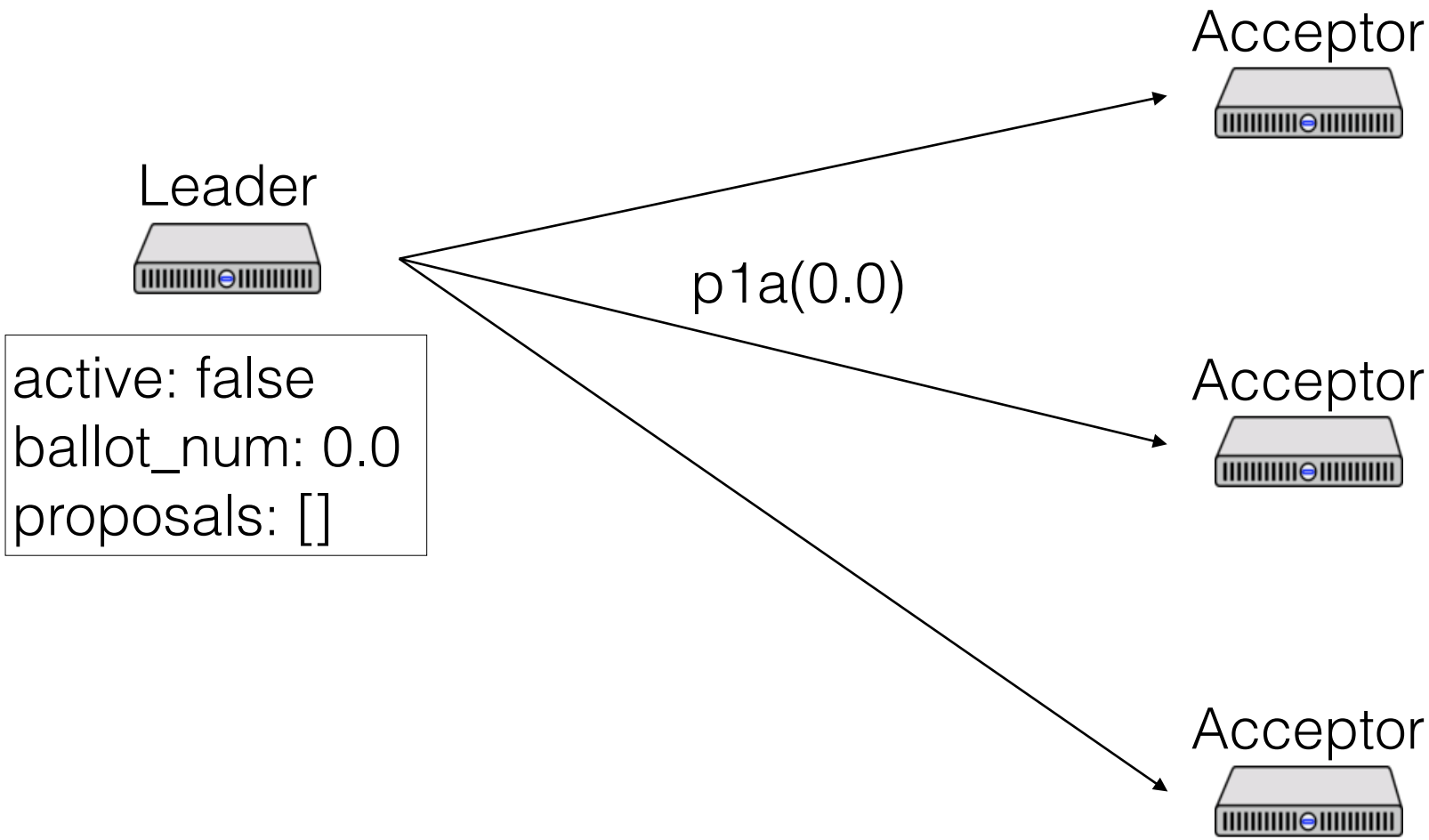
Ballot numbers increase

Only accept values from current ballot

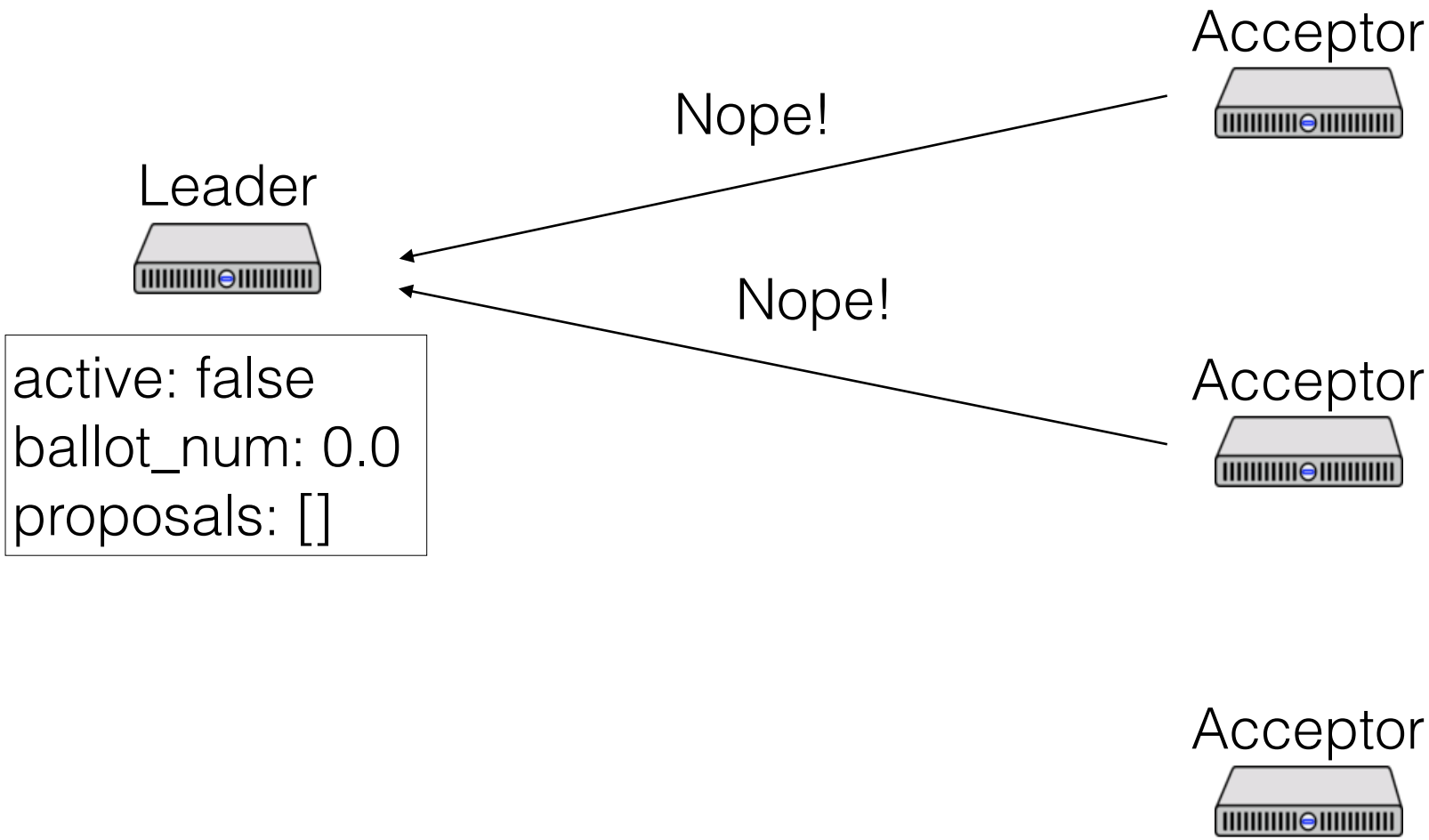
Never remove ballots

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

Leader: Getting Elected



Leader: Getting Elected



Leader: Getting Elected

Acceptor

Leader

active: false
ballot_num: 1.0
proposals: []

Acceptor

Acceptor

Leader: Getting Elected

Acceptor

Leader

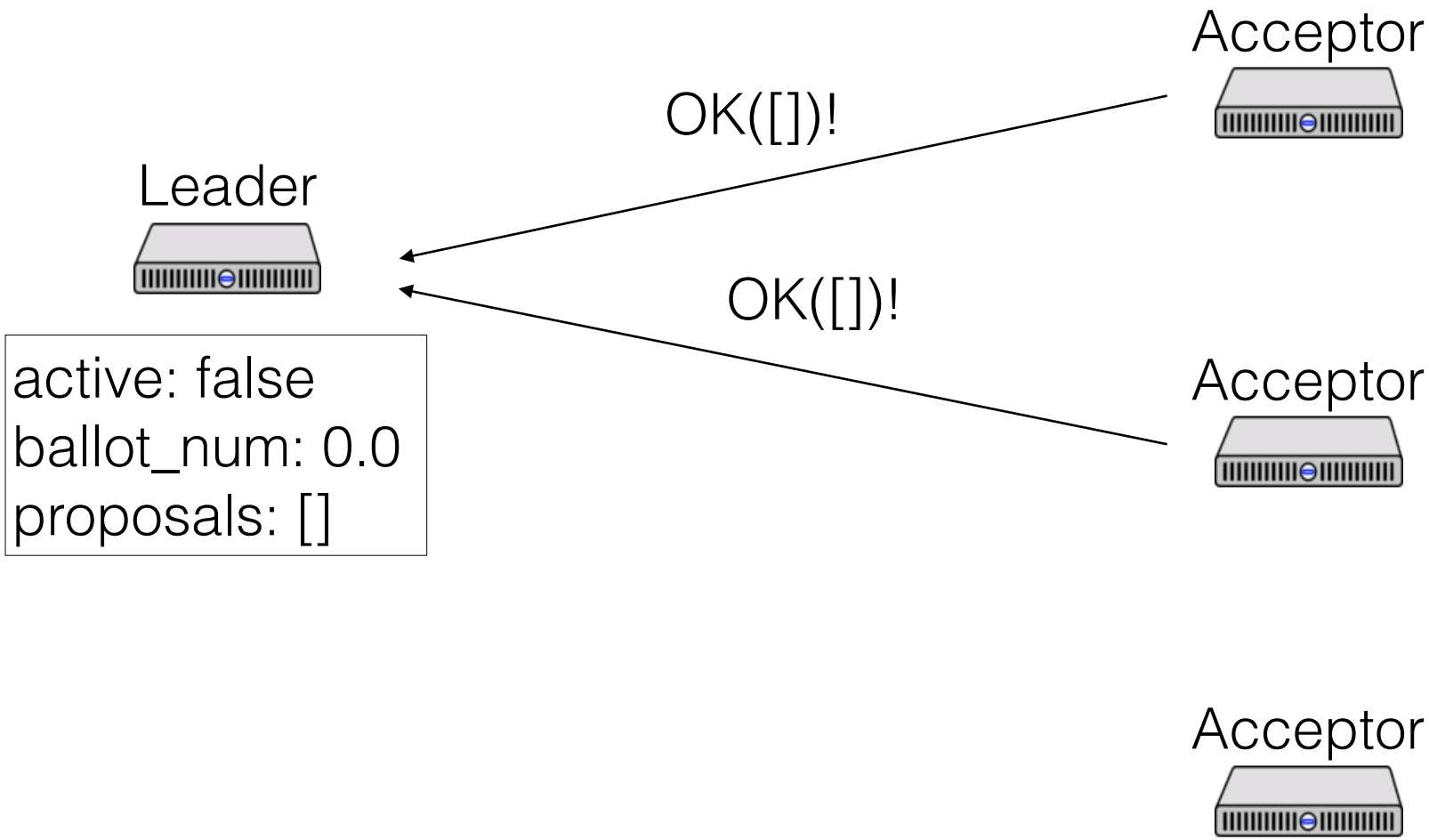
active: false
ballot_num: 1.0
proposals: []

Or…

Acceptor

Acceptor

Leader: Getting Elected



When to run for office

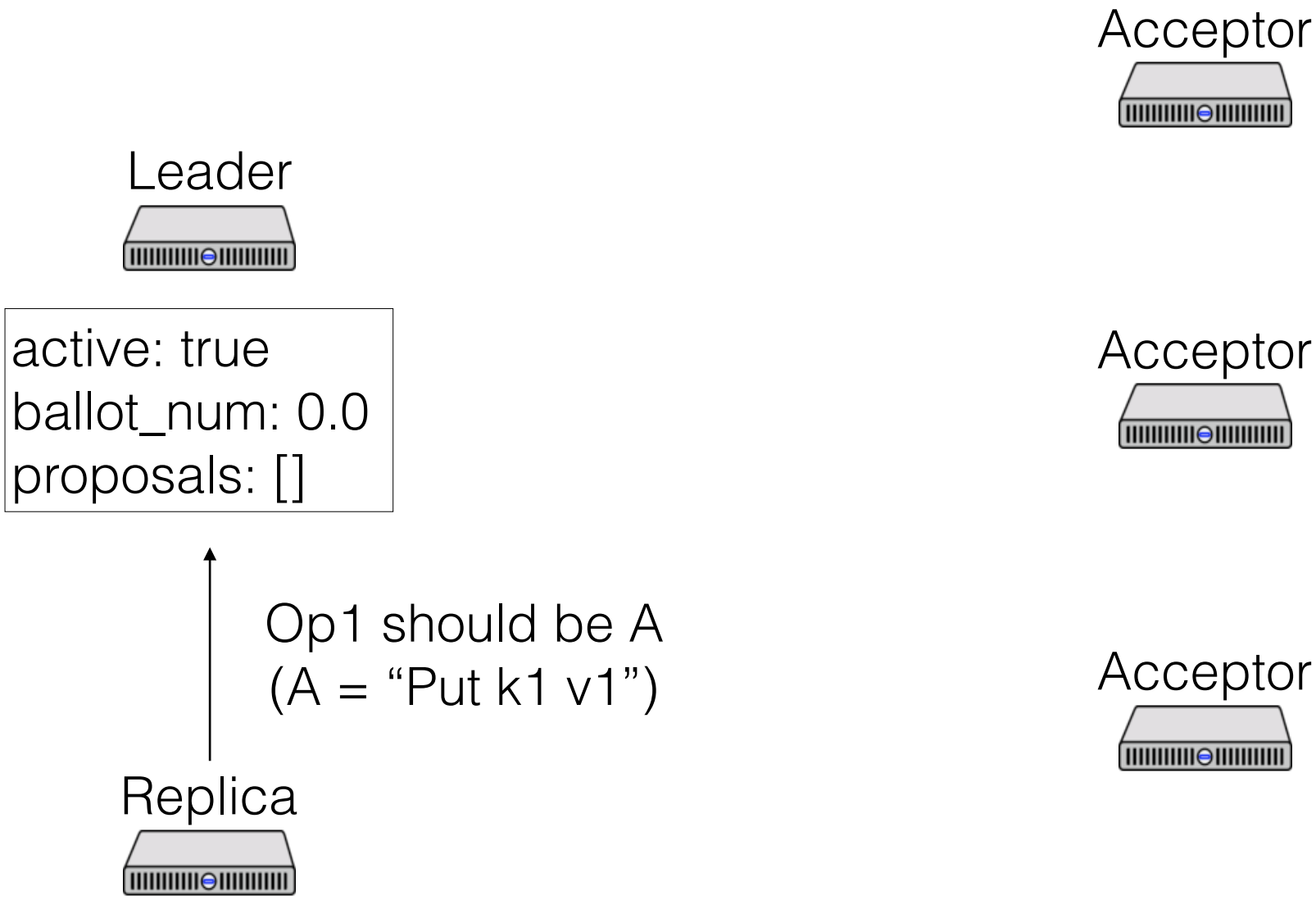
When should a leader try to get elected?

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

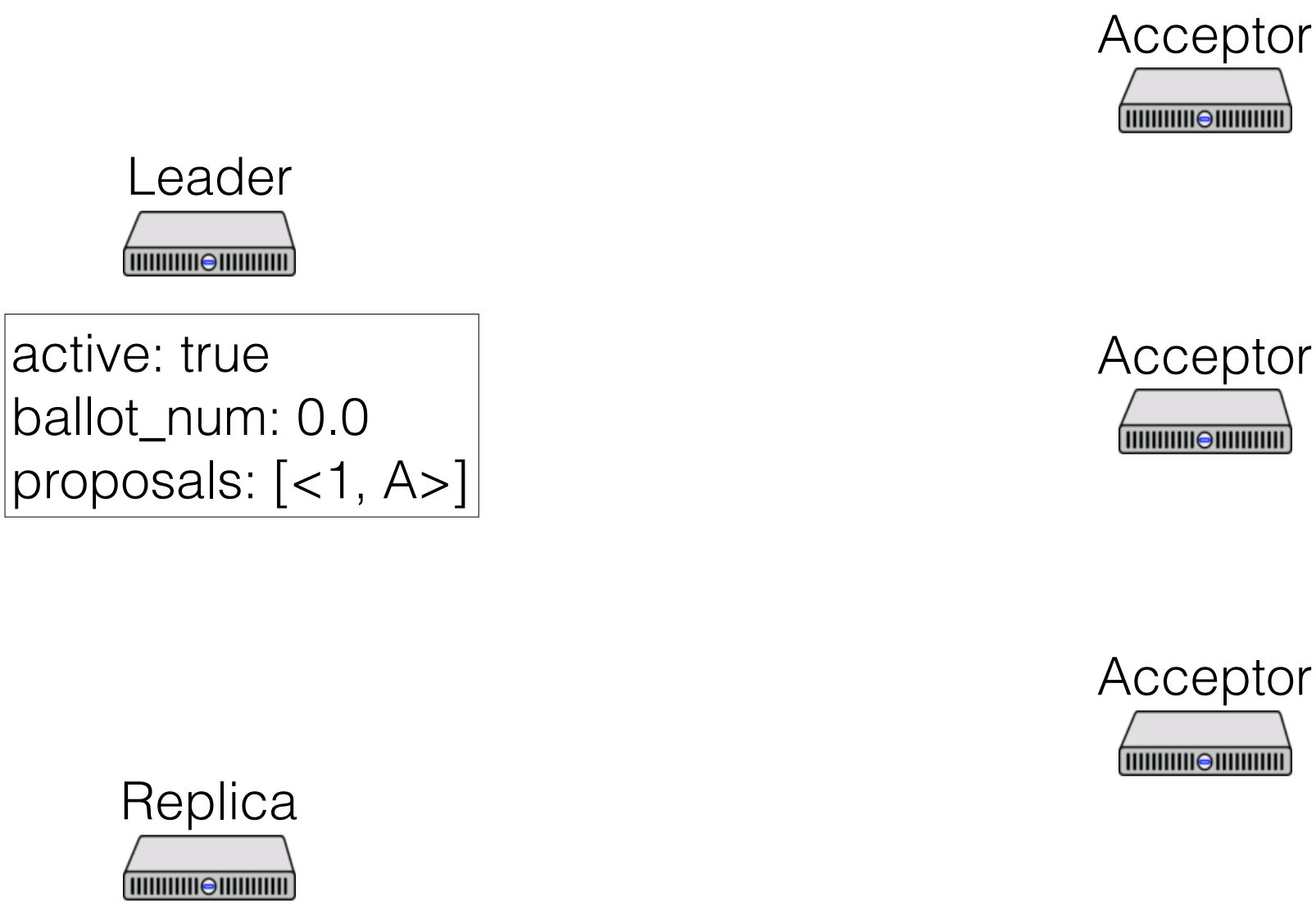
Paper describes an algorithm, based on pinging the leader and timing out

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

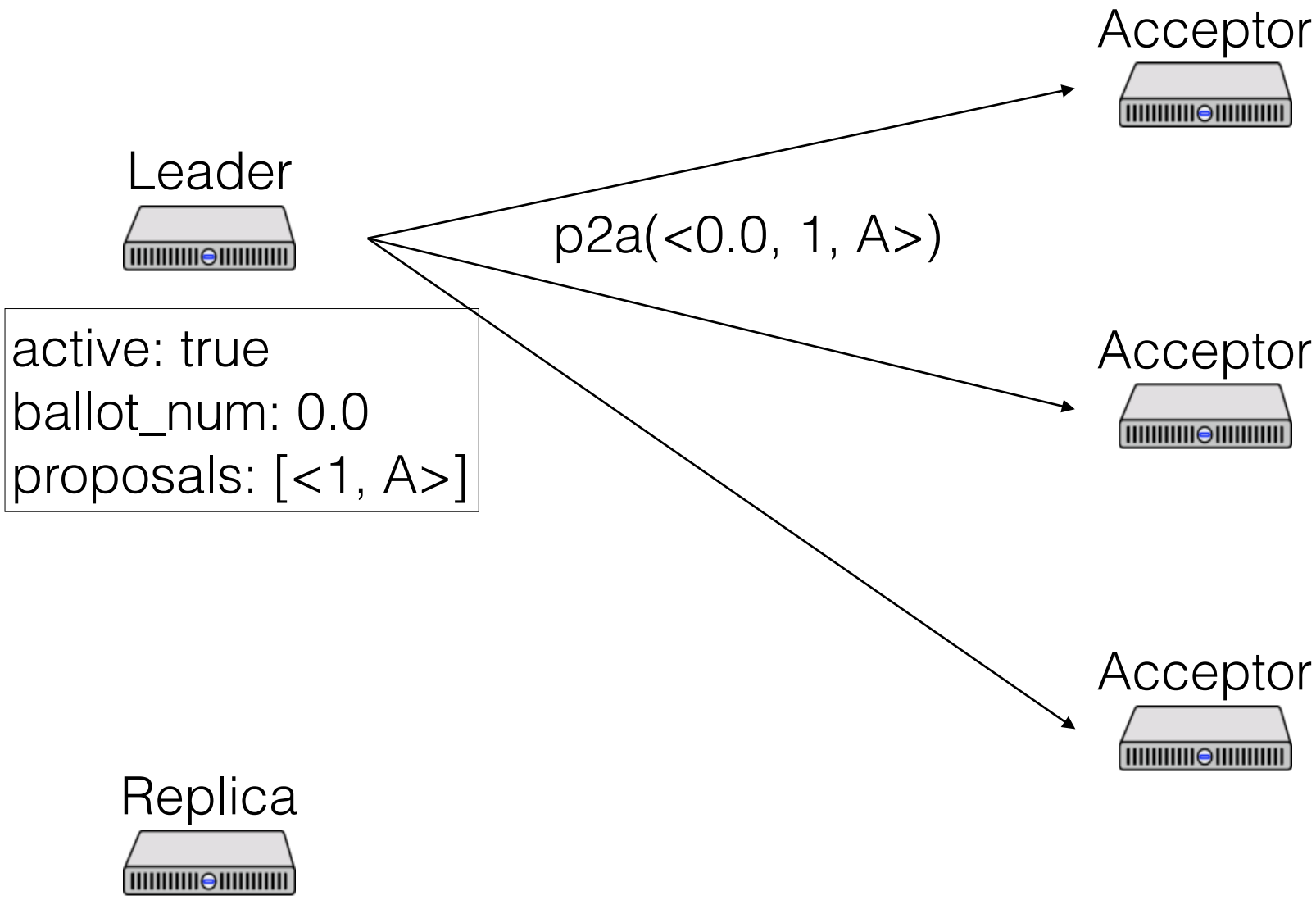
Leader: Handling proposals



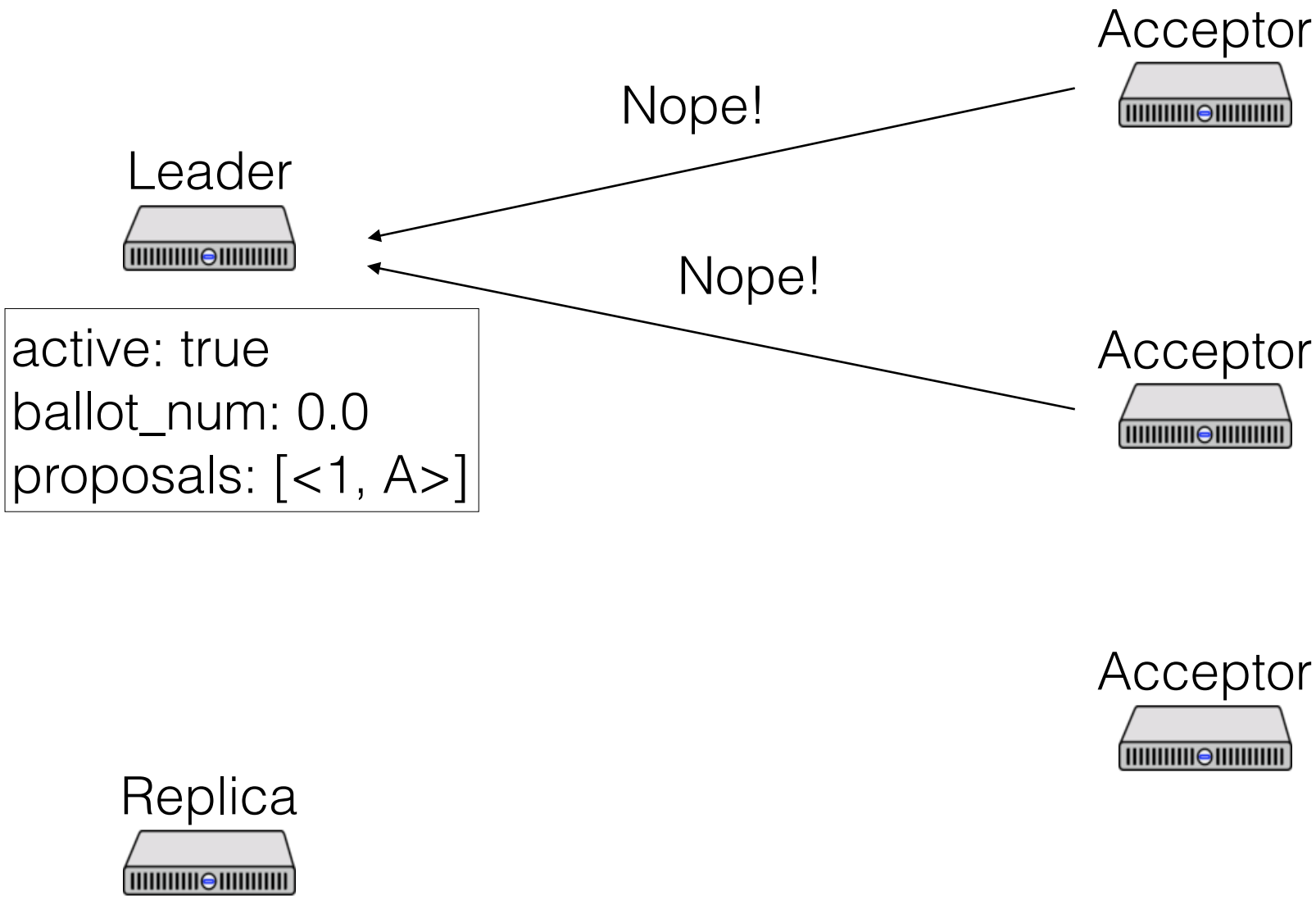
Leader: Handling proposals



Leader: Handling proposals



Leader: Handling proposals



Leader: Handling proposals

Acceptor

Leader

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

Acceptor

Acceptor

Replica

Leader: Handling proposals

Acceptor

Leader

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

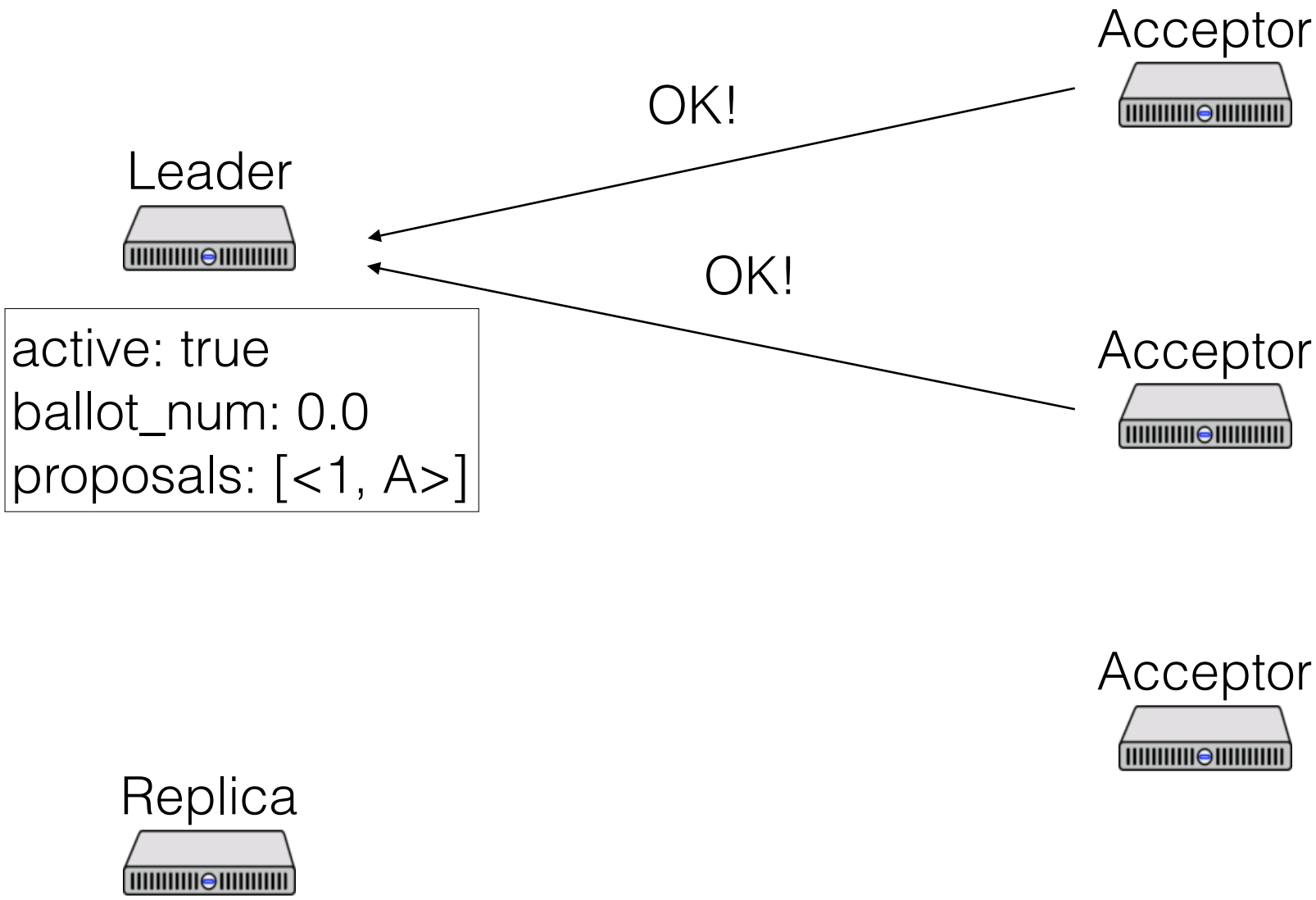
Or…

Acceptor

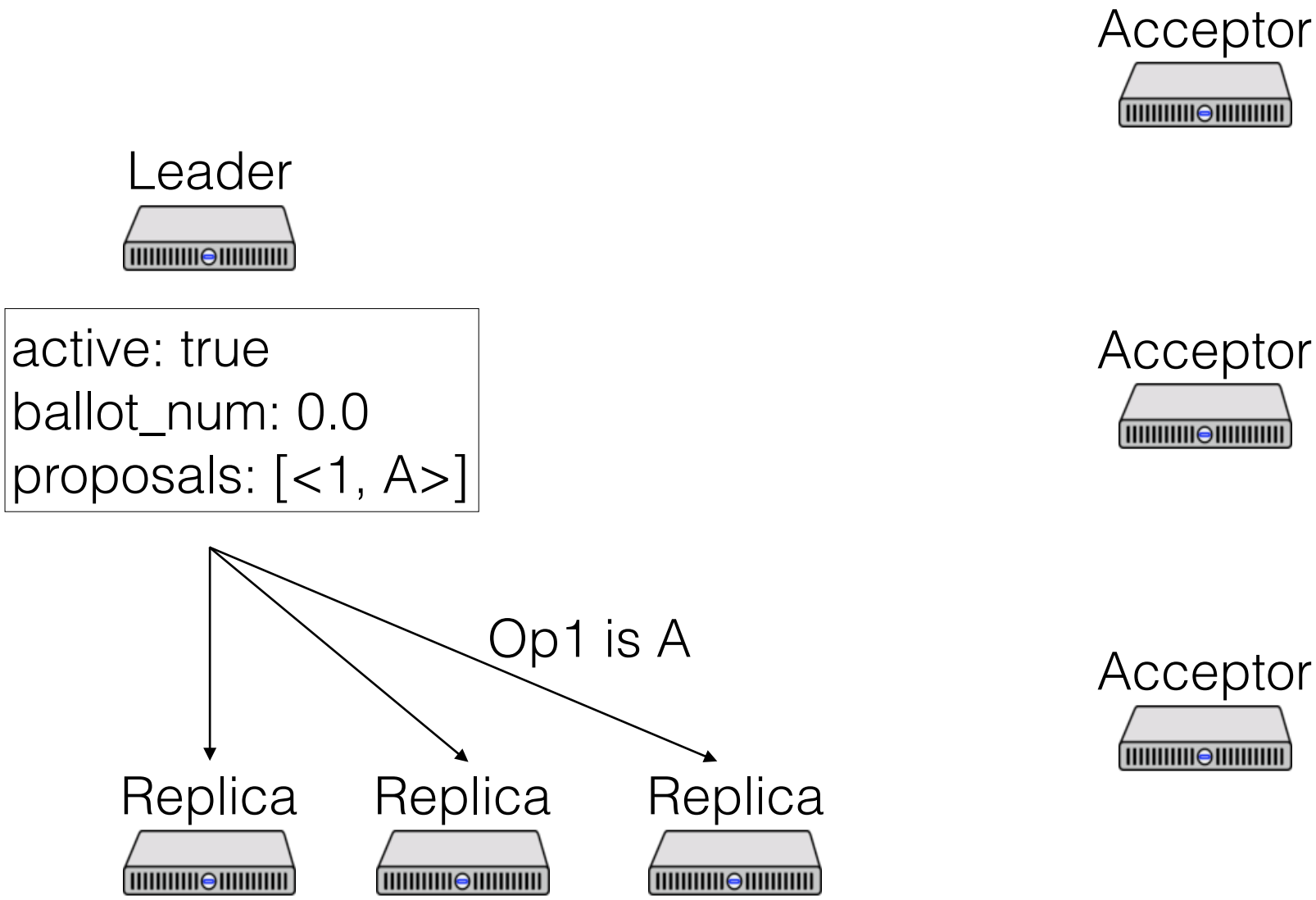
Acceptor

Replica

Leader: Handling proposals



Leader: Handling proposals



Questions

What should be in stable storage?

Question

What are the costs to using Paxos? Is it practical enough?