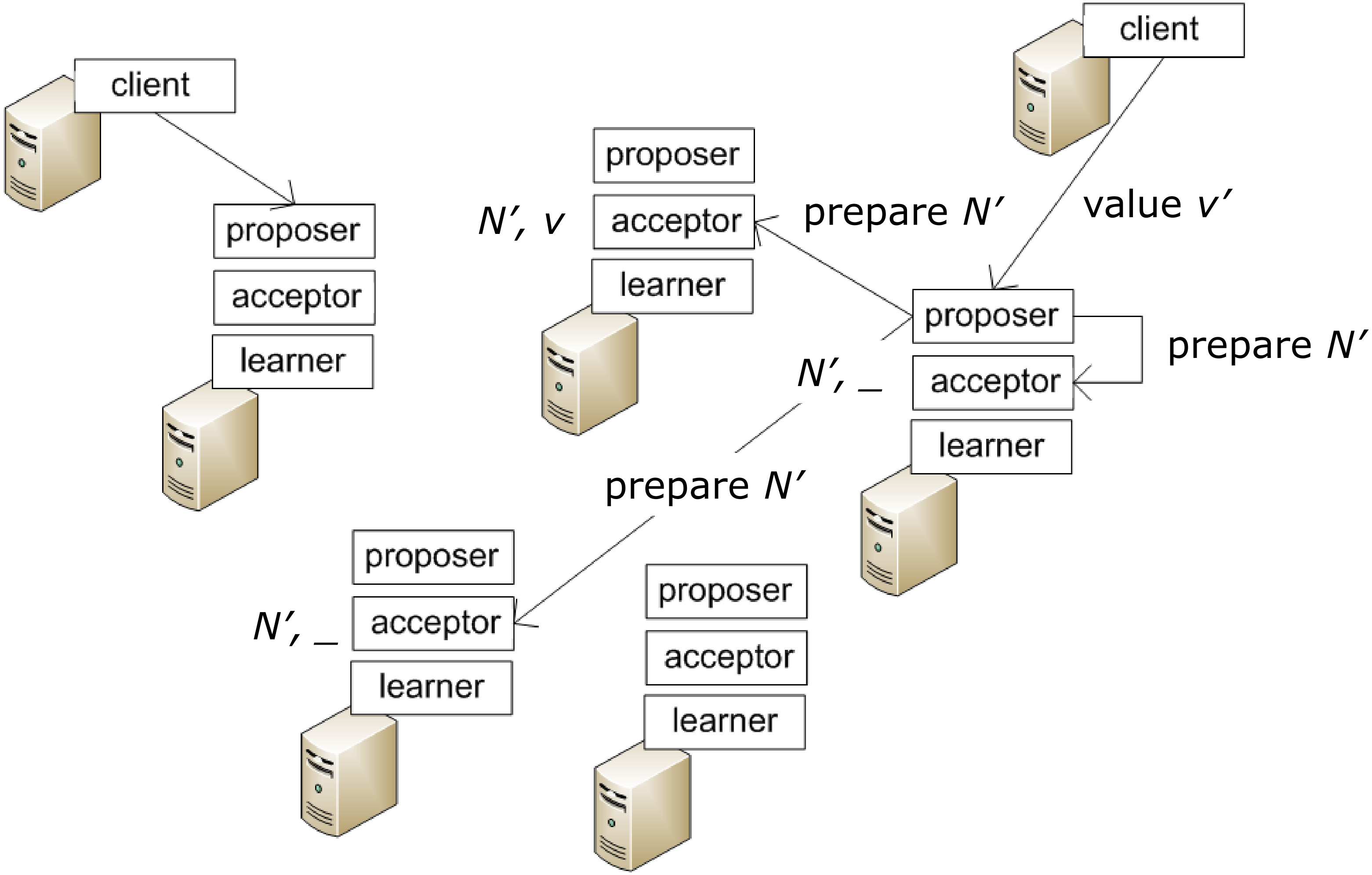
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

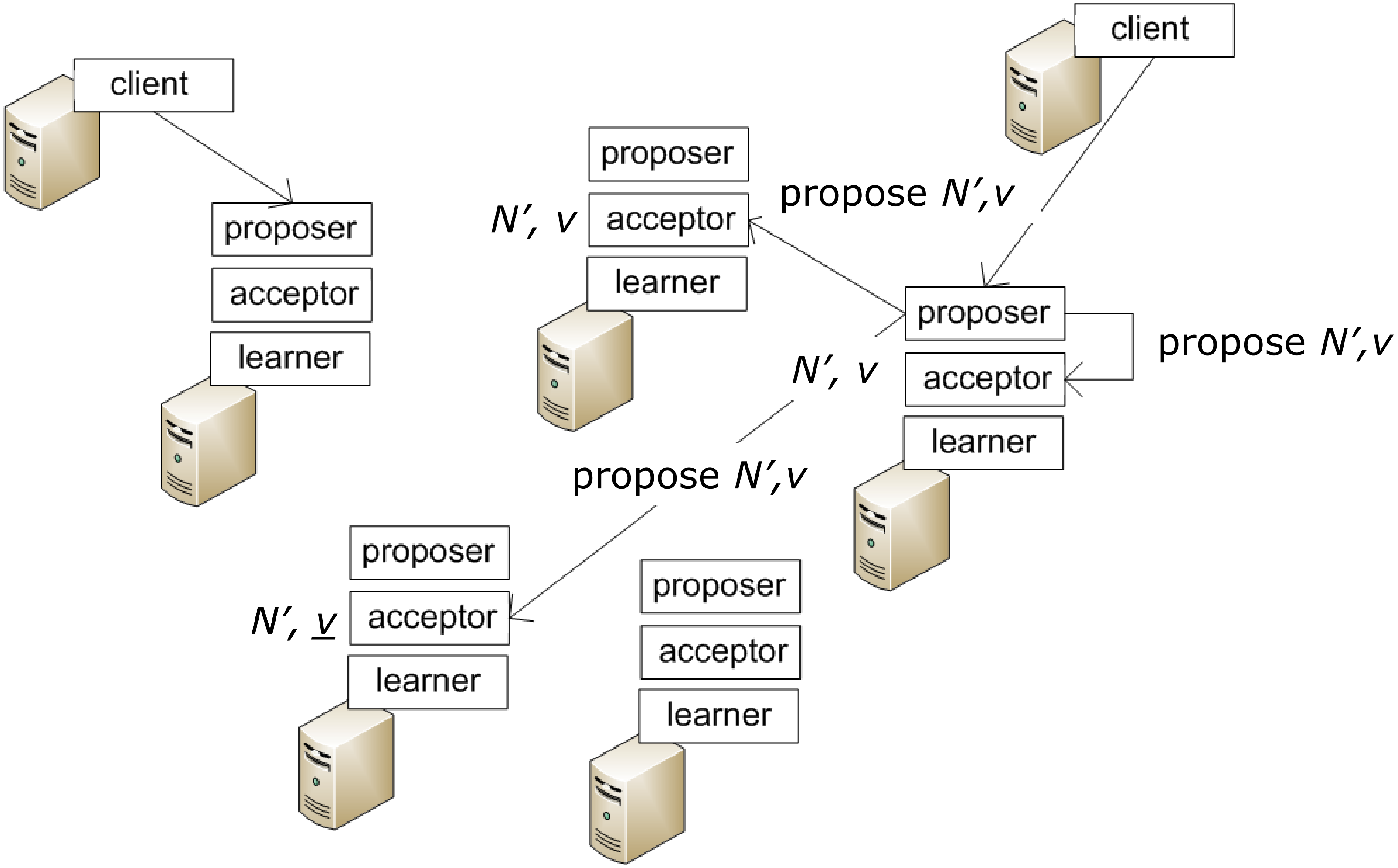
HY590.45
Modern Topics in Scalable Storage Systems

Kostas Magoutis
magoutis@csd.uoc.gr
http://www.csd.uoc.gr/~magoutis

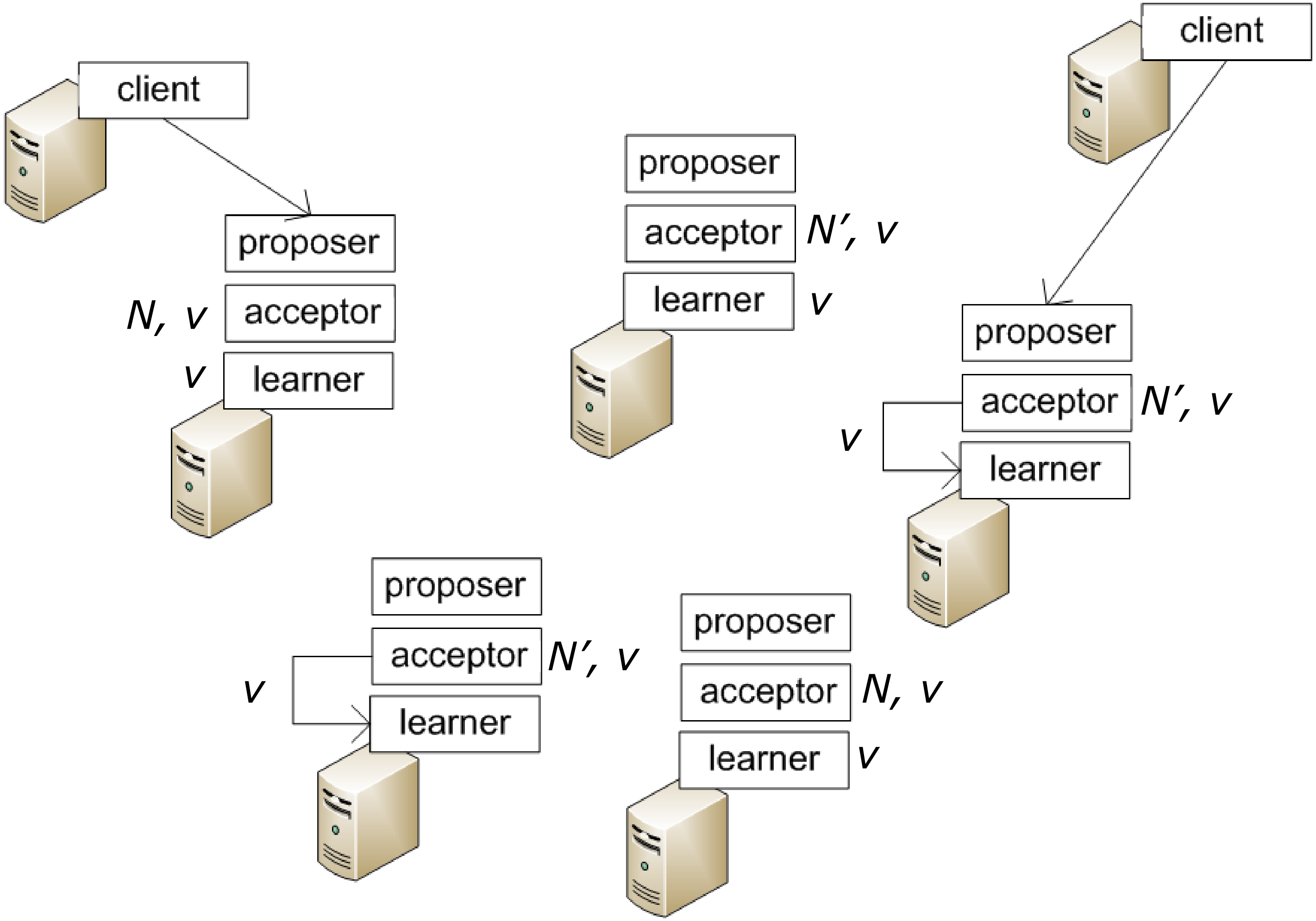
Paxos – learning chosen value



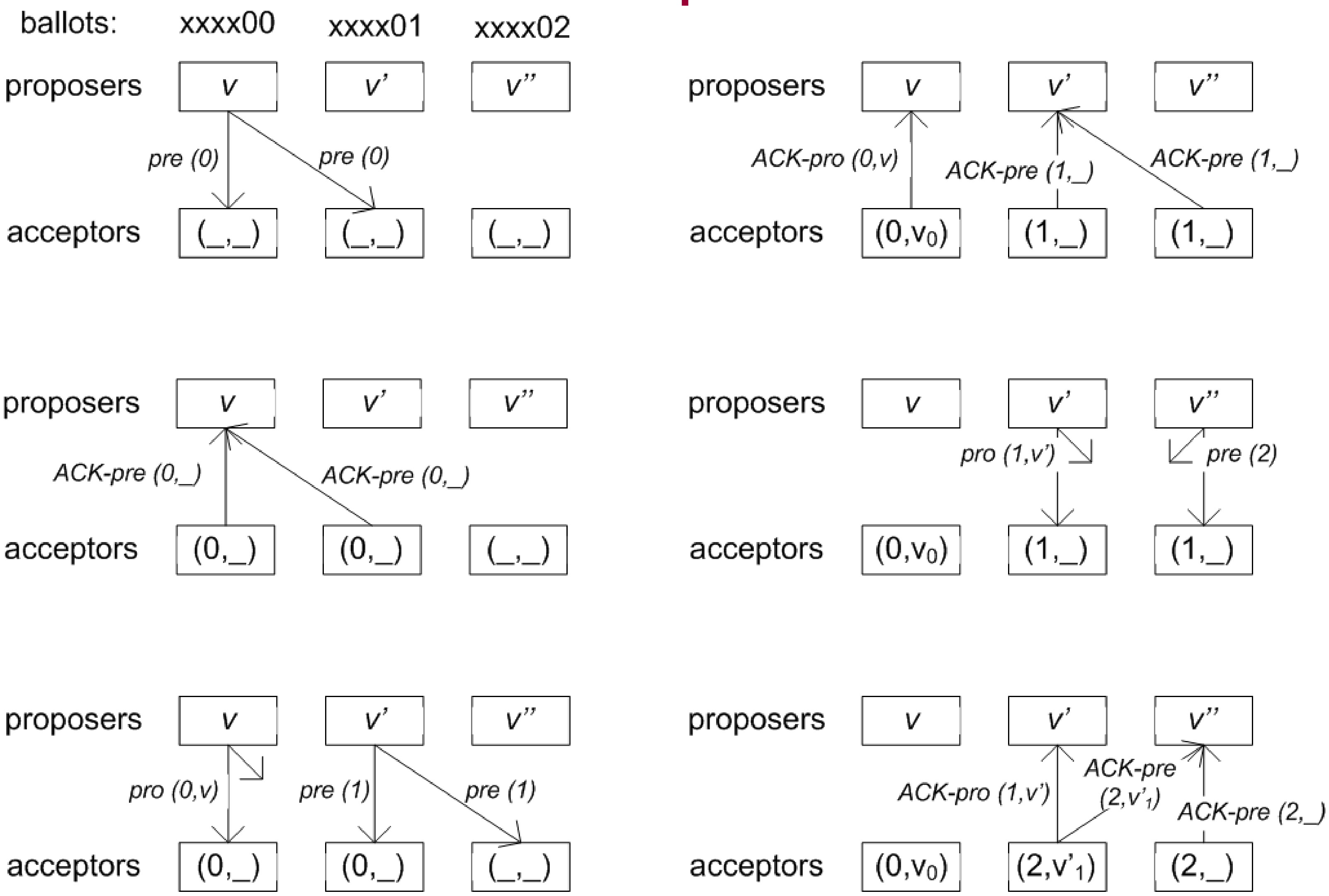
Paxos – propagate chosen value



Paxos – everyone learns outcome



Example



ballots: xxxx00 xxxx01 xxxx02

proposers v v' v''

pre (0) pre (0)

acceptors (_,_) (_,_) (_,_)

proposers v v' v''

ACK-pro (0,v) ACK-pre (1,_) ACK-pre (1,_)

acceptors $(0,v_0)$ (1,_) (1,_)

proposers v v' v''

ACK-pre (0,_) ACK-pre (0,_)

acceptors (0,_) (0,_) (_,_)

proposers v v' v''

pro (1,v') pre (2)

acceptors $(0,v_0)$ (1,_) (1,_)

proposers v v' v''

pro (0,v) pre (1) pre (1)

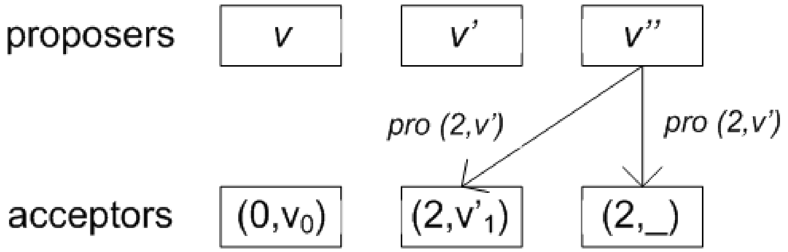
acceptors (0,_) (0,_) (_,_)

proposers v v' v''

ACK-pro (1,v') ACK-pre $(2,v'_1)$ ACK-pre (2,_)

acceptors $(0,v_0)$ $(2,v'_1)$ (2,_)

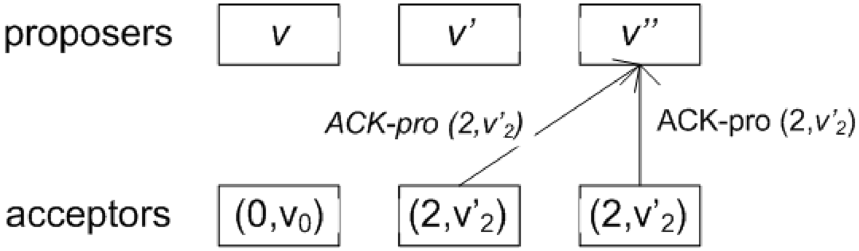
Example (*contd.*)



proposers v v' v''

pro (2,v') pro (2,v')

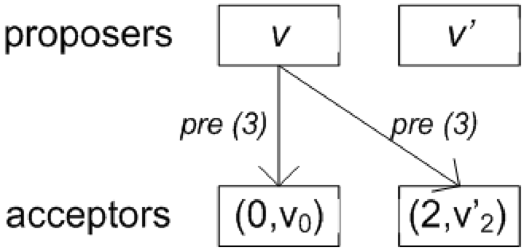
acceptors (0,v_0) (2,v'_1) (2,_)



proposers v v' v''

ACK-pro (2,v'_2) ACK-pro (2,v'_2)

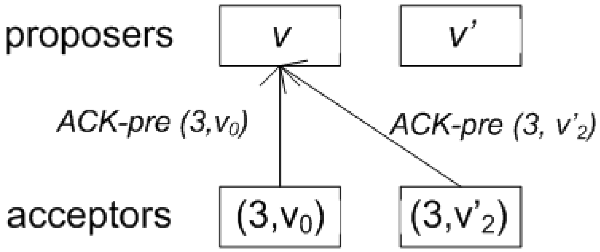
acceptors (0,v_0) (2,v'_2) (2,v'_2)



proposers v v'

pre (3) pre (3)

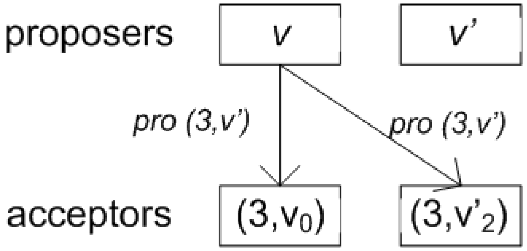
acceptors (0,v_0) (2,v'_2)



proposers v v'

ACK-pre (3,v_0) ACK-pre (3, v'_2)

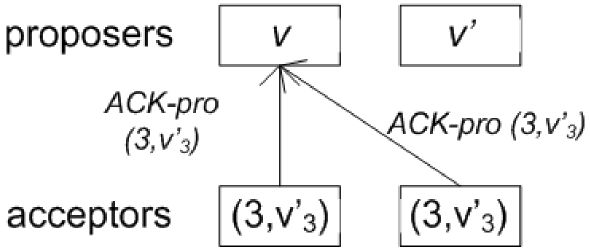
acceptors (3,v_0) (3,v'_2)



proposers v v'

pro (3,v') pro (3,v')

acceptors (3,v_0) (3,v'_2)



proposers v v'

ACK-pro (3,v'_3) ACK-pro (3,v'_3)

acceptors (3,v'_3) (3,v'_3)

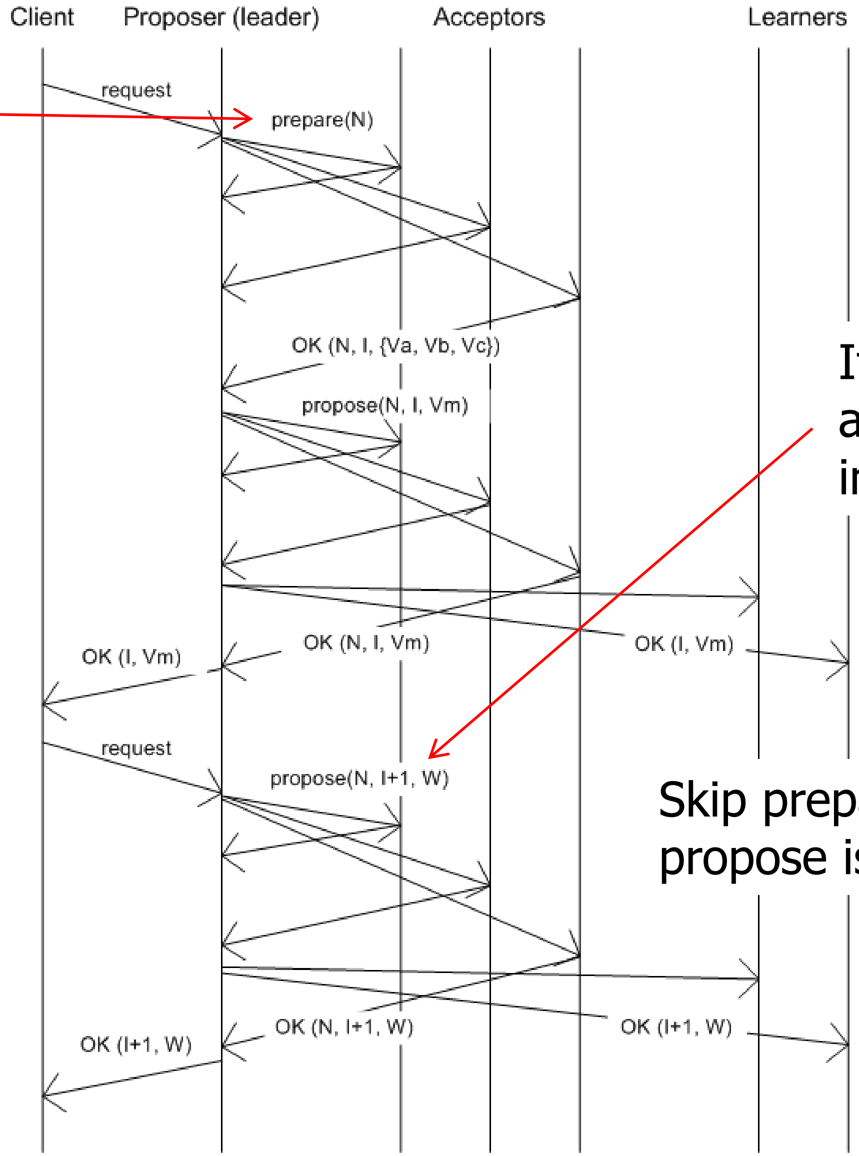
Lamport: implementing a state machine

- How to run multiple instances of Paxos
  - Assume the existence of a distinguished proposer (leader)
  - A leader will run Paxos for a number of instances
  - The leader may crash, at which point there may be gaps in the chosen instances (1-134, 138, ..)
  - A new leader will try to fill in those slots or propose *no-op*
  - As soon as gap fills, commands can be executed

- Multi-Paxos
  - New leader: execute phase 1 for infinitely many instances
  - Acceptors can respond with reasonably short messages
  - Cost of Paxos effectively the cost of executing phase 2

Multi-Paxos

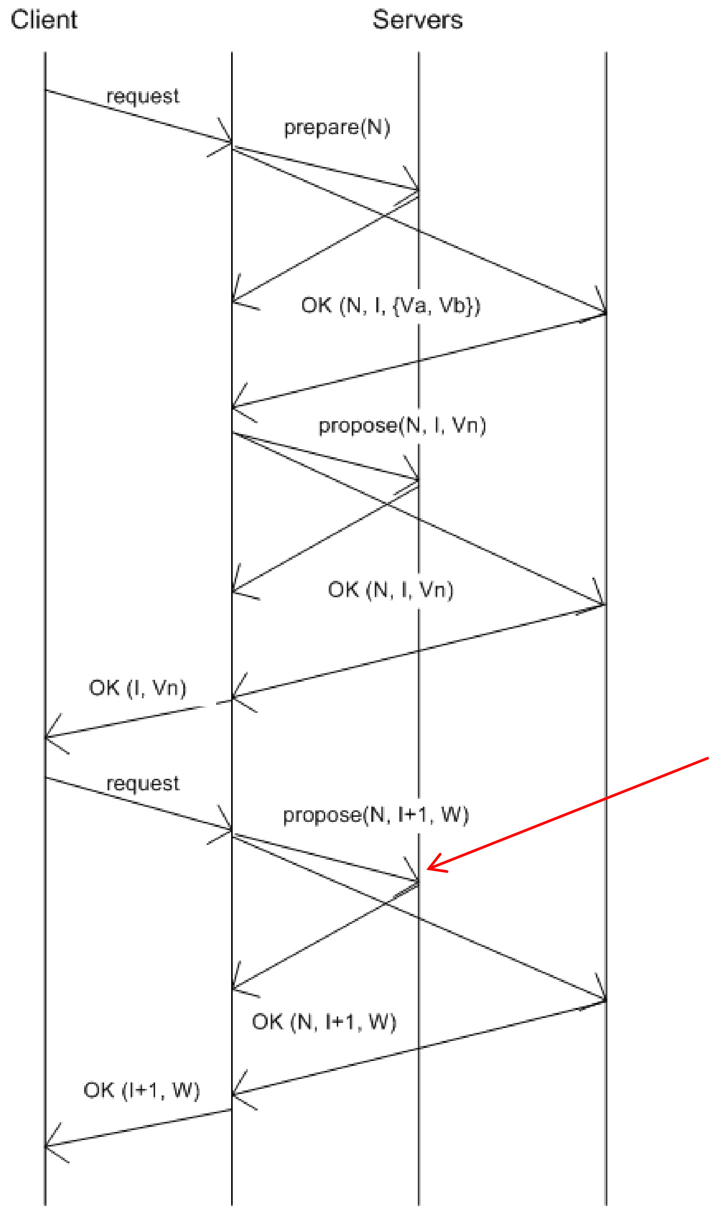
- New leader @N
- Learn accepted values for past instances



If a majority has not accepted anything for instances > I

Skip prepare phase until a propose is rejected!

Multi-Paxos



Servers play all roles

Replicas write to disk prior to sending ACK