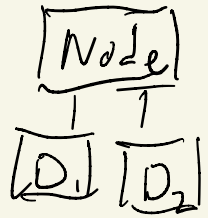
CS-438 Decentralized Systems Engineering

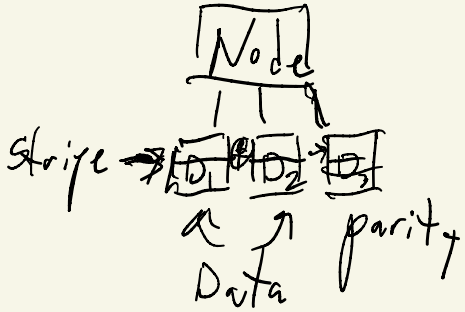
Week 8

Replication

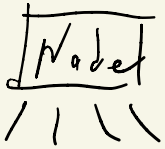
- Storage or Nodes
- Storage: eg. disk - RAID (Redundant Array of Inexpensive Disks)



RAID 1: mirroring
(2 copies on 2 disks)



RAID 5: striping with parity
tolerates any single disk failure



RAID 6: double parity
tolerate 2 disk failures

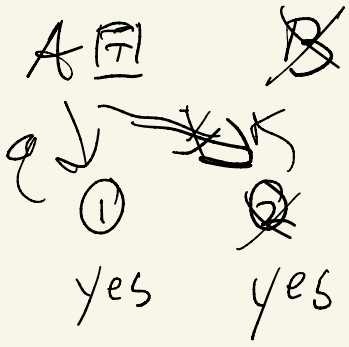
Exists an "authority" who knows/governs "state of universe"

- Replication of distributed Nodes

Agreement or Consensus

- Several (n) nodes agreeing on one state (value)
- Permissioned or permissionless (since 2008)
  - Permissioned: assumes well-defined, closed group of n nodes
  - Permissionless: (Bitcoin) proof-of-work - invest
- Crash failures vs Byzantine failures
- Many consistency models - "How" consistent?
  - Serialization: all nodes agree on single history
  - Eventual consistency: nodes disagree for a while, but eventually come to agreement

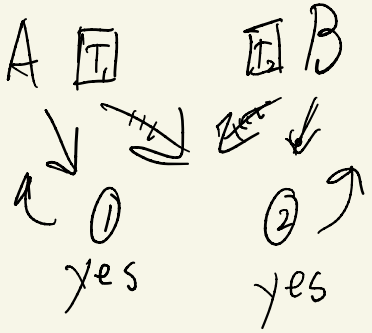
Paxos - Leslie Lamport



clients - drive agreement - Proposers

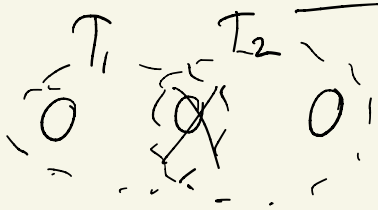
state nodes - Acceptors

if A wants to tolerate f failures ($f < n$),
must wait for at most $n-f$ answers

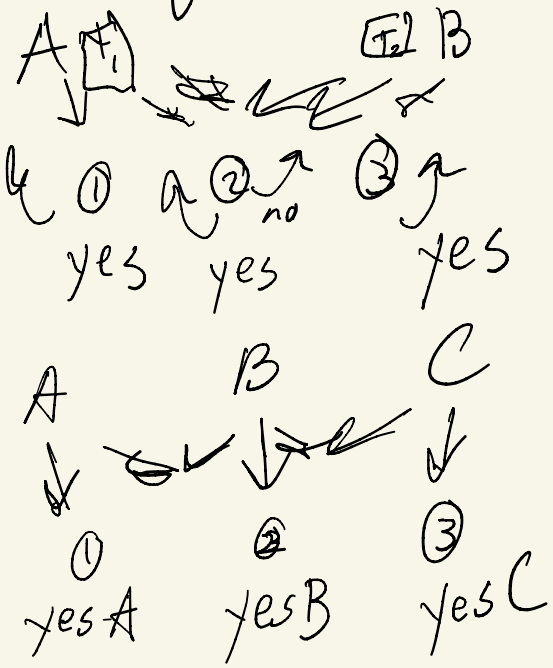


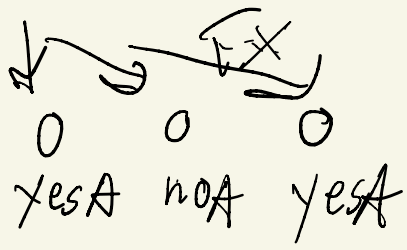
to reach agreement of only 1 value;
need a majority: $n > 2f$ $(n \geq 2f+1)$

$n-f > f$



Majoritarian consensus





Alice & Bob propose T_1, T_2

Alice wins

no majority — no winner

how do you know when conses reached?

each client has $\leq n-f$ observations

eg. if I got 1 yes, might have succeeded

if I got f yeses, might have failed

$\Rightarrow$ "I dont know"

if got 0 yes, definitely failed

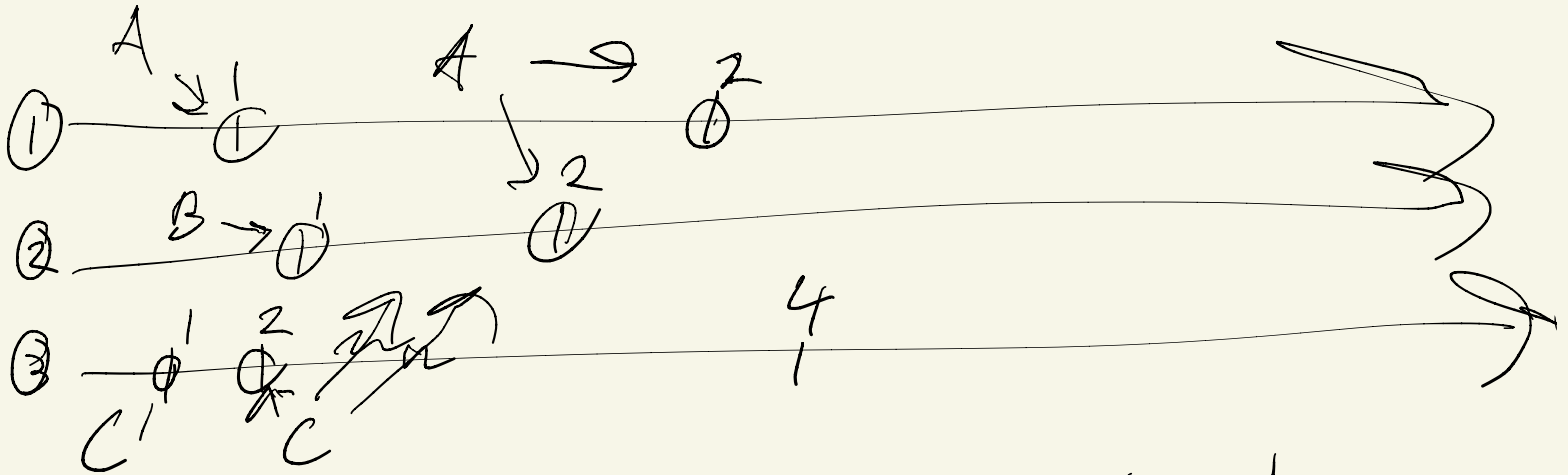
got $f+1$ yes, "looks like" succeed

Fundamental challenges

1- Be able to try again on failure

2- Be consistent with past attempts that might have agreed

1. Lock-step virtual time - steps, "ballot numbers"



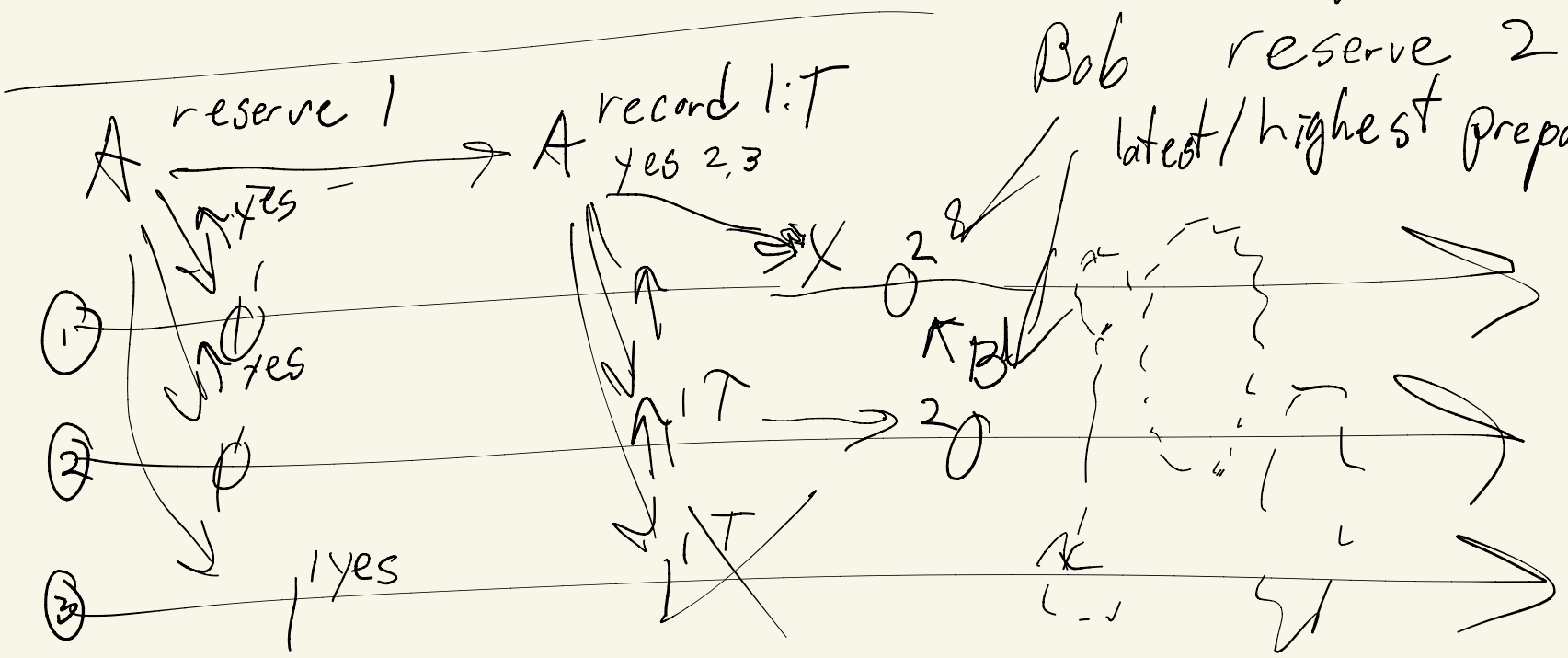
2. Need 2-phase protocol

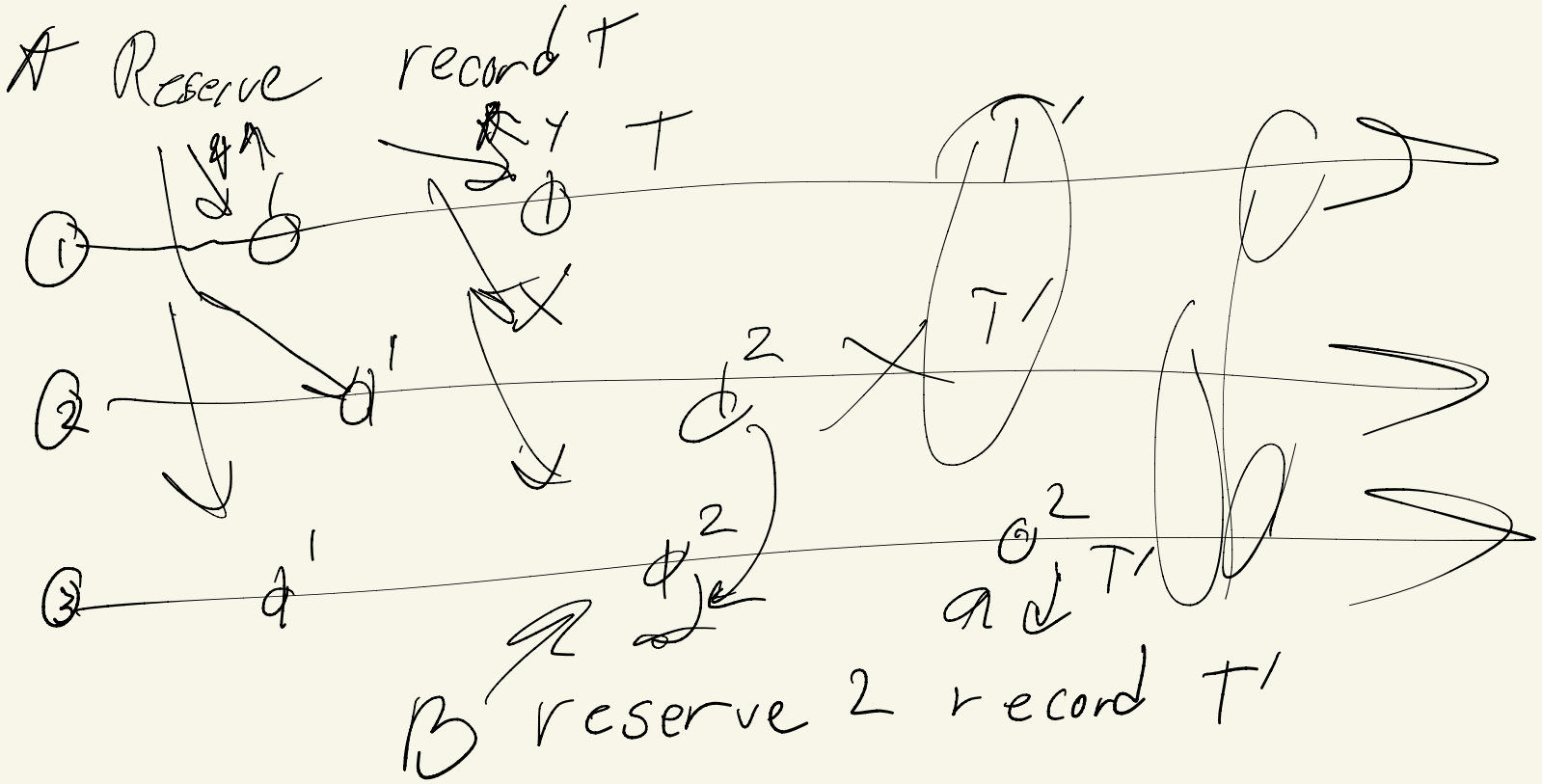
2 phases (majority for each)

1. figure out if I can succeed
2. figure out if I did succeed

time-step reservation:

1. ask each server to reserve ("prepare") step for me
2. (if succeeds)
   ask each server to record proposal




A Reserve
record T
T
T'
T'
B reserve 2 record T'