Distributed Storage

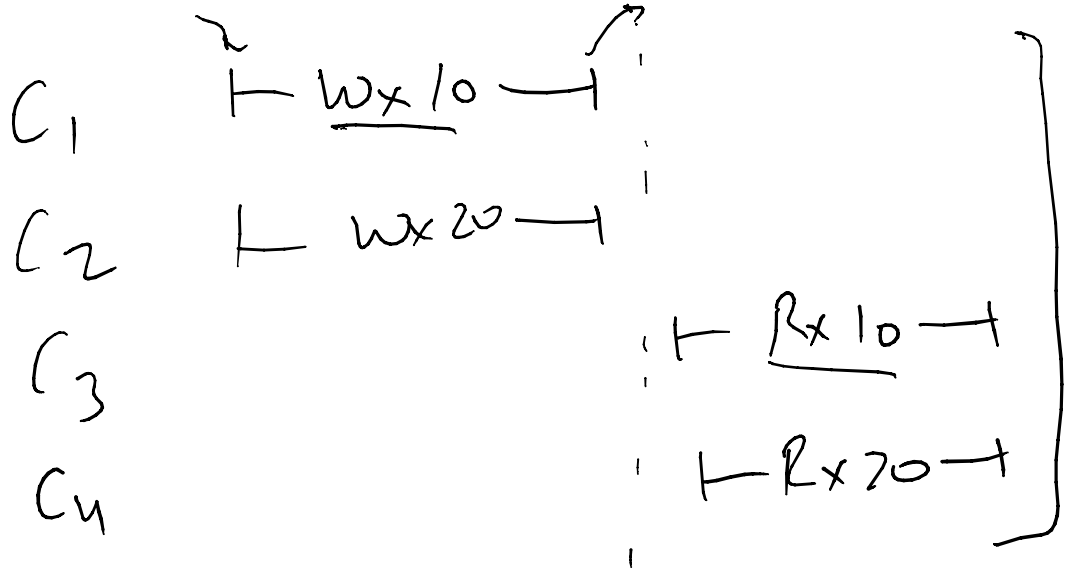
Note Title 08-Dec-22

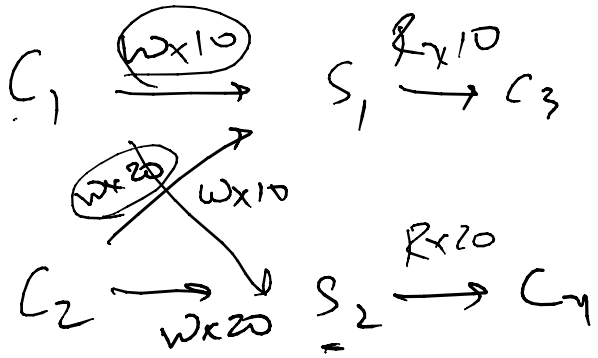
- Performance
- Recoverability

Availability: Hiding faults → Replication

- Consistency

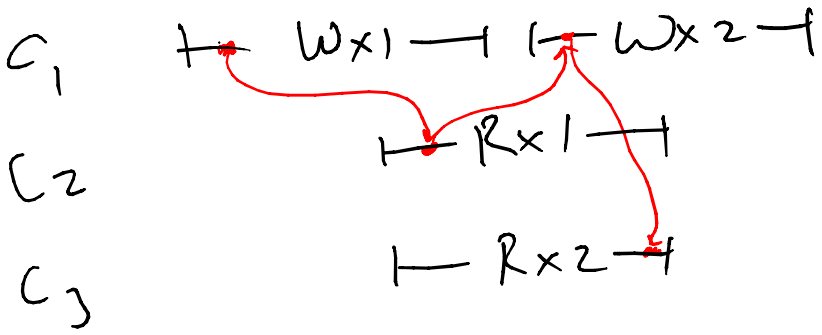
Scalability N disk

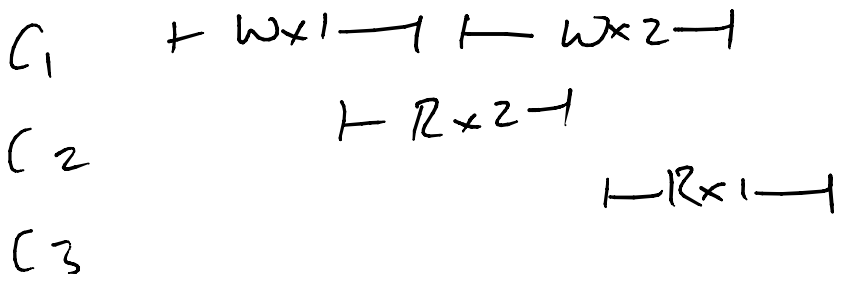

C1
C2
C3
C4
Wx10
Wx20
Rx10
Rx20

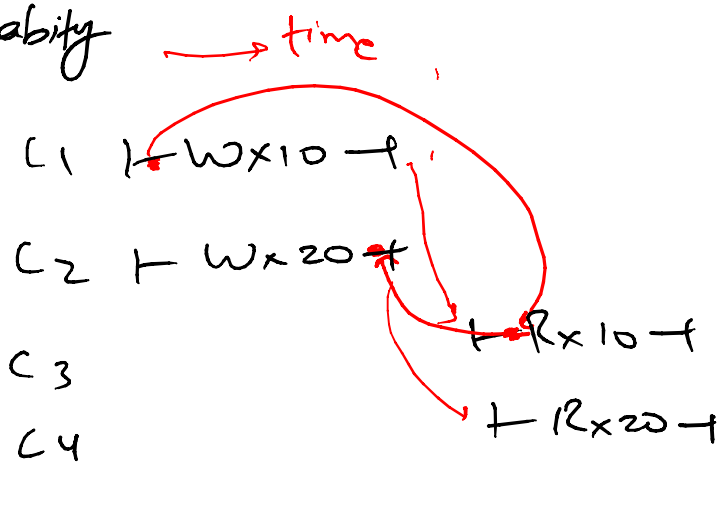

C1
Wx10
S1
Rx10
C3
Wx20
Wx10
C2
Wx20
S2
Rx20
C4

Raft :

Sequential Consistency : Linearizability







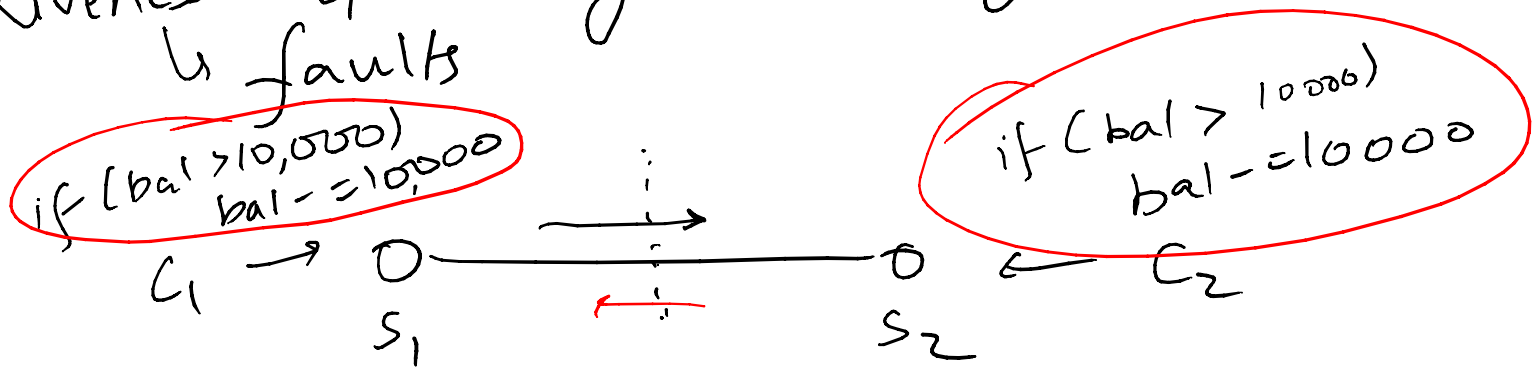
→ Safety: Bad things never happen

C_1 ⊢ wx10 - - - -

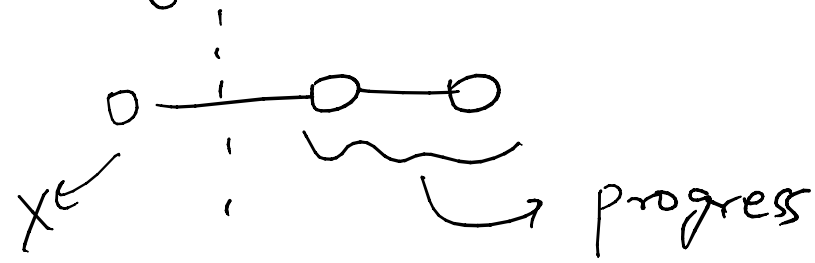
C_2 ⊢ Rx . - - -

Liveness: Good things eventually happen

↳ faults

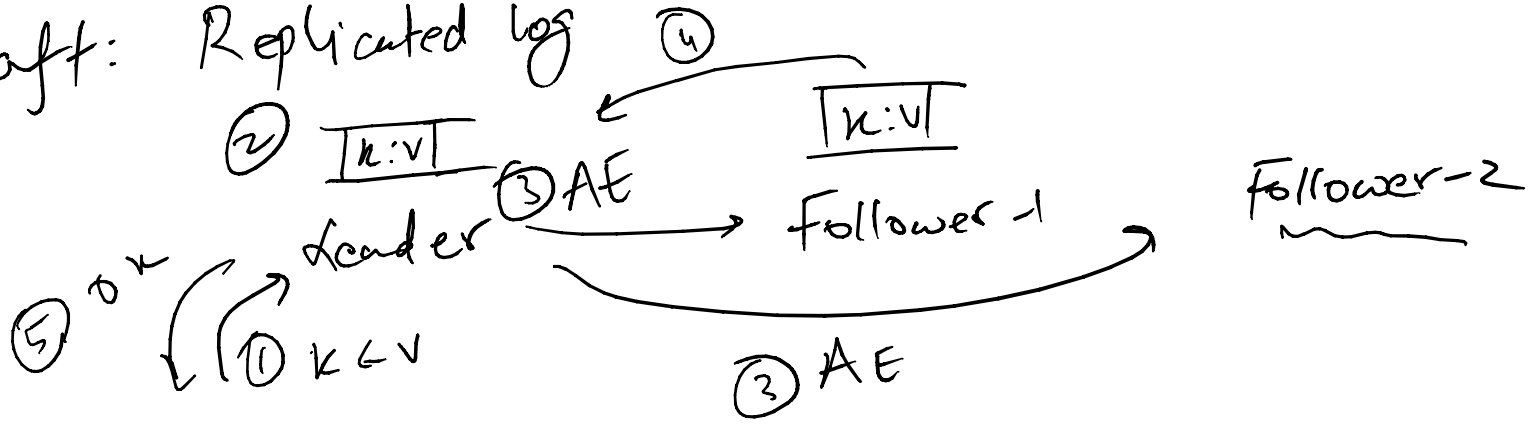


Use majority



$2f+1$ replicas → f faults

Raft: Replicated log



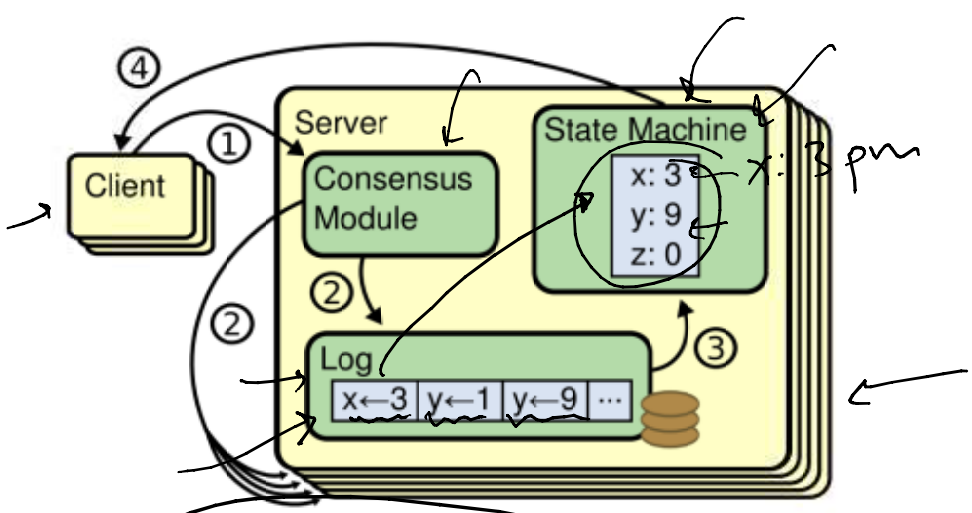


Figure 1: Replicated state machine architecture. The consensus algorithm manages a replicated log containing state machine commands from clients. The state machines process identical sequences of commands from the logs, so they produce the same outputs.

log => Linearizability

Wx1 Wx2

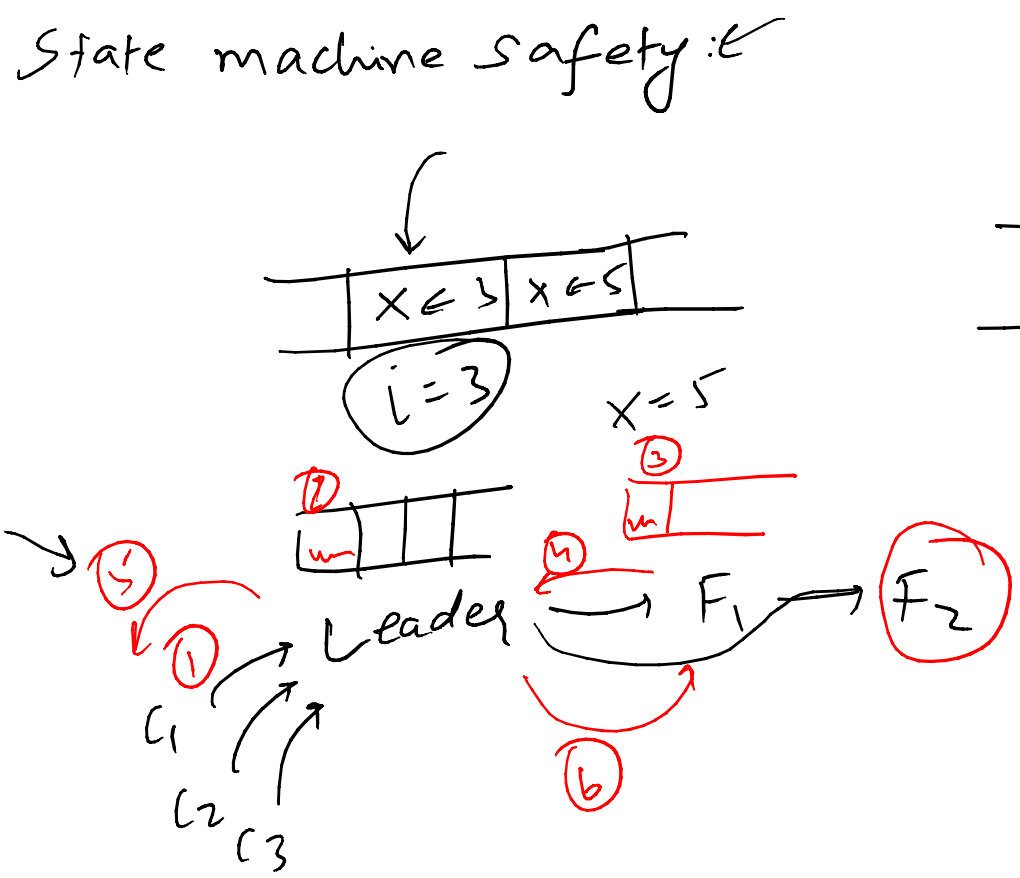
Rx1

Rx2

Wx1 | Rx1 | Wx2 | Rx2

Wx getTime

State machine safety:



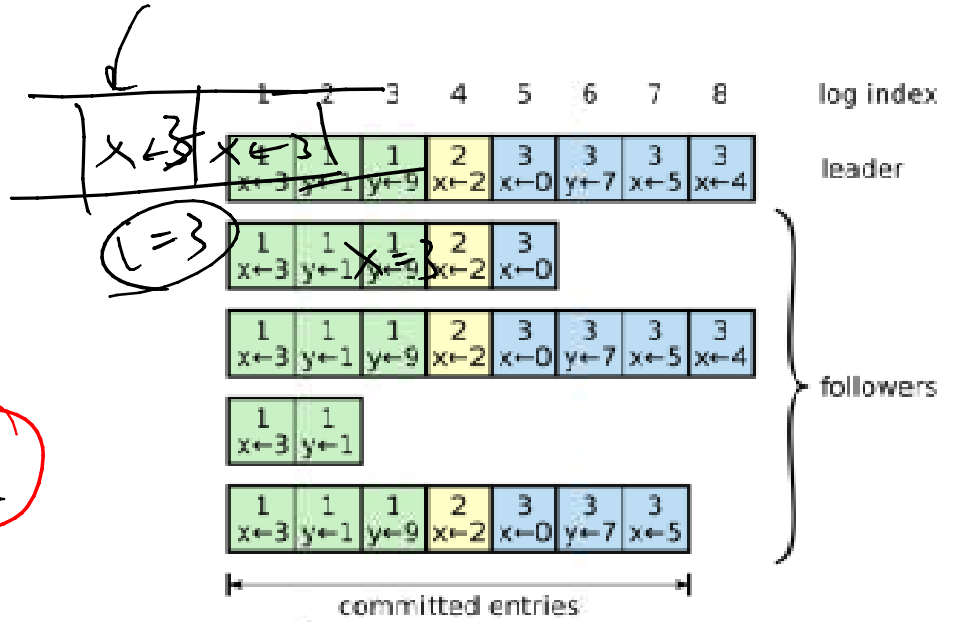


Figure 6: Logs are composed of entries, which are numbered sequentially. Each entry contains the term in which it was created (the number in each box) and a command for the state machine. An entry is considered *committed* if it is safe for that entry to be applied to state machines.

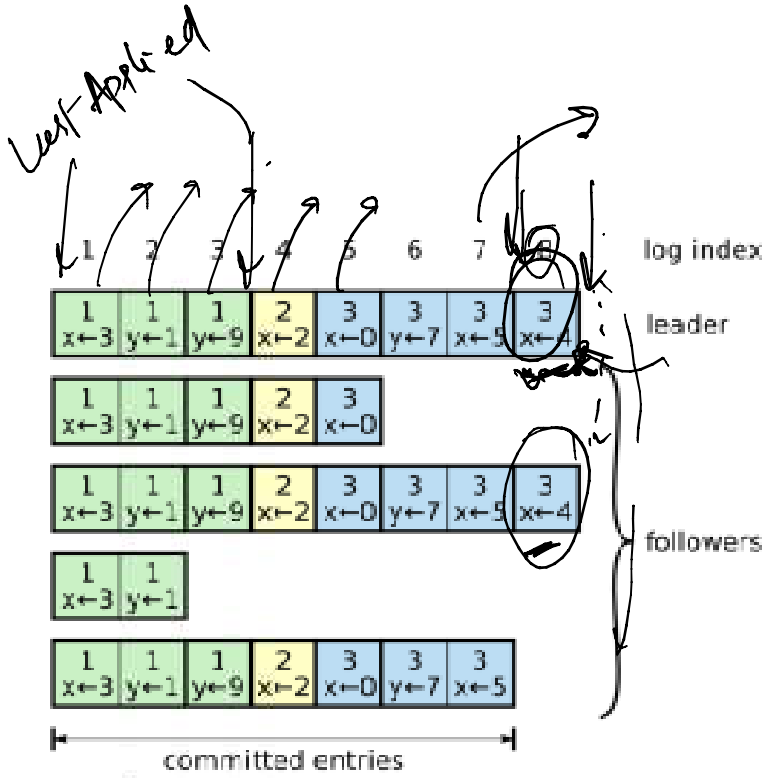
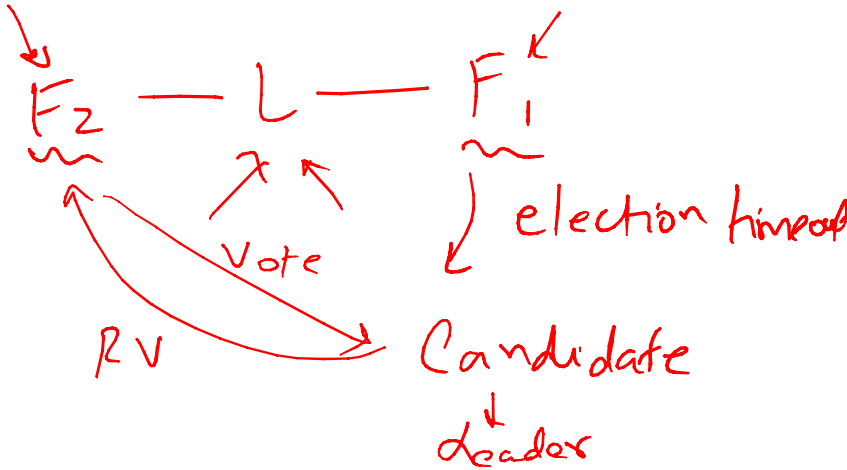


Figure 6: Logs are composed of entries, which are numbered sequentially. Each entry contains the term in which it was created (the number in each box) and a command for the state machine. An entry is considered *committed* if it is safe for that entry to be applied to state machines.



Election Safety: At most ONE leader (in a given term)

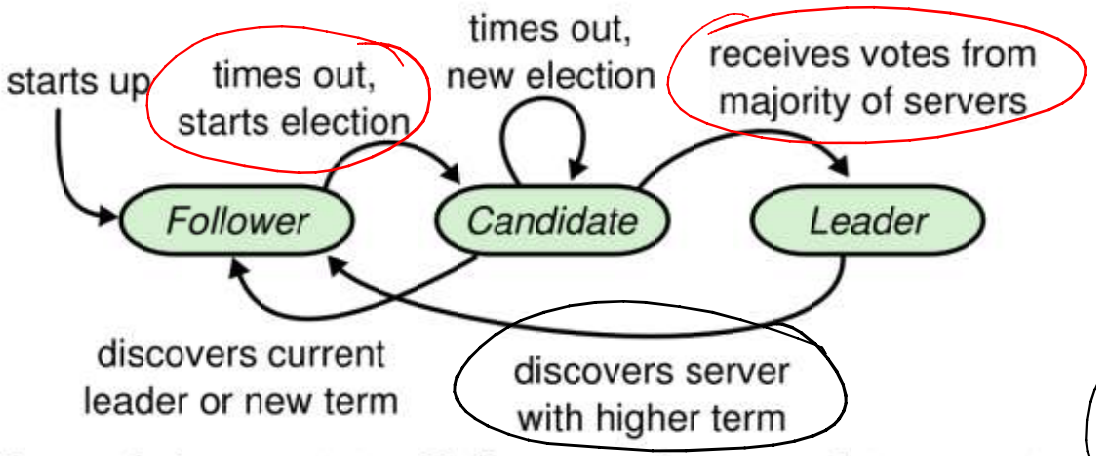
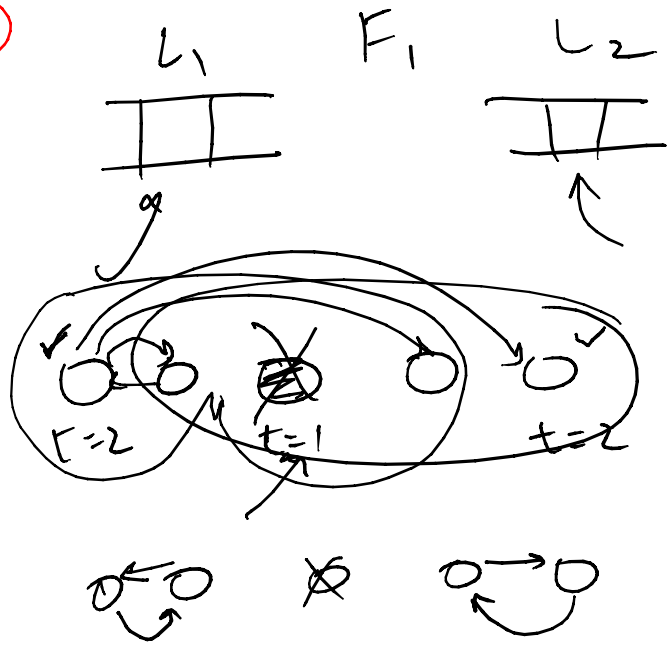
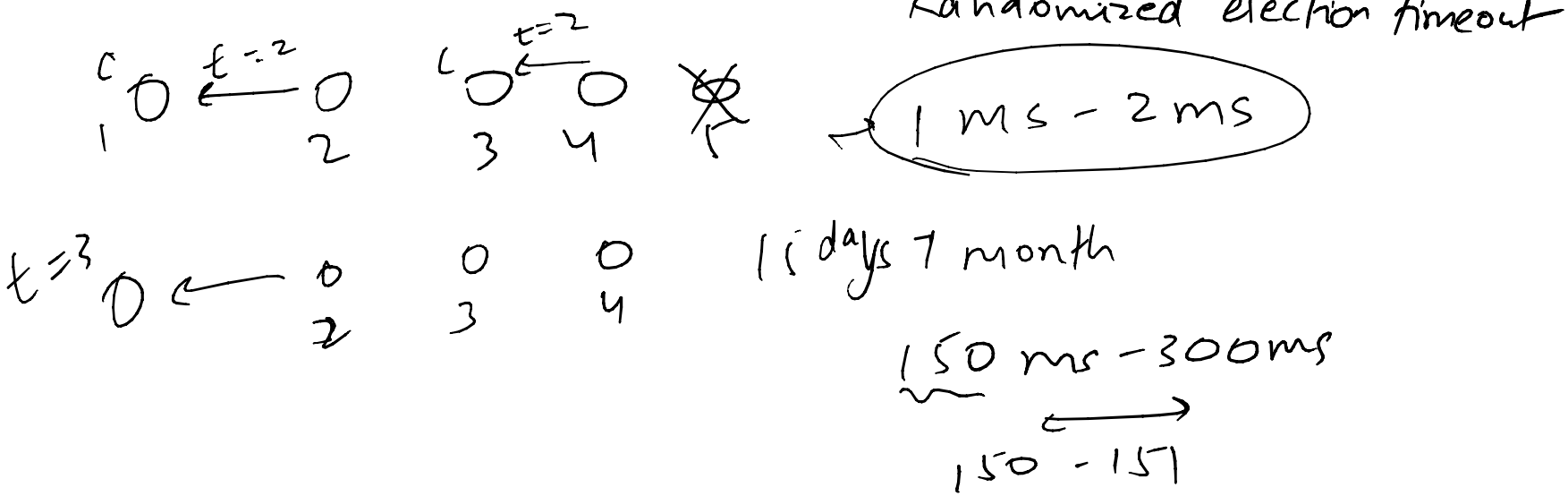
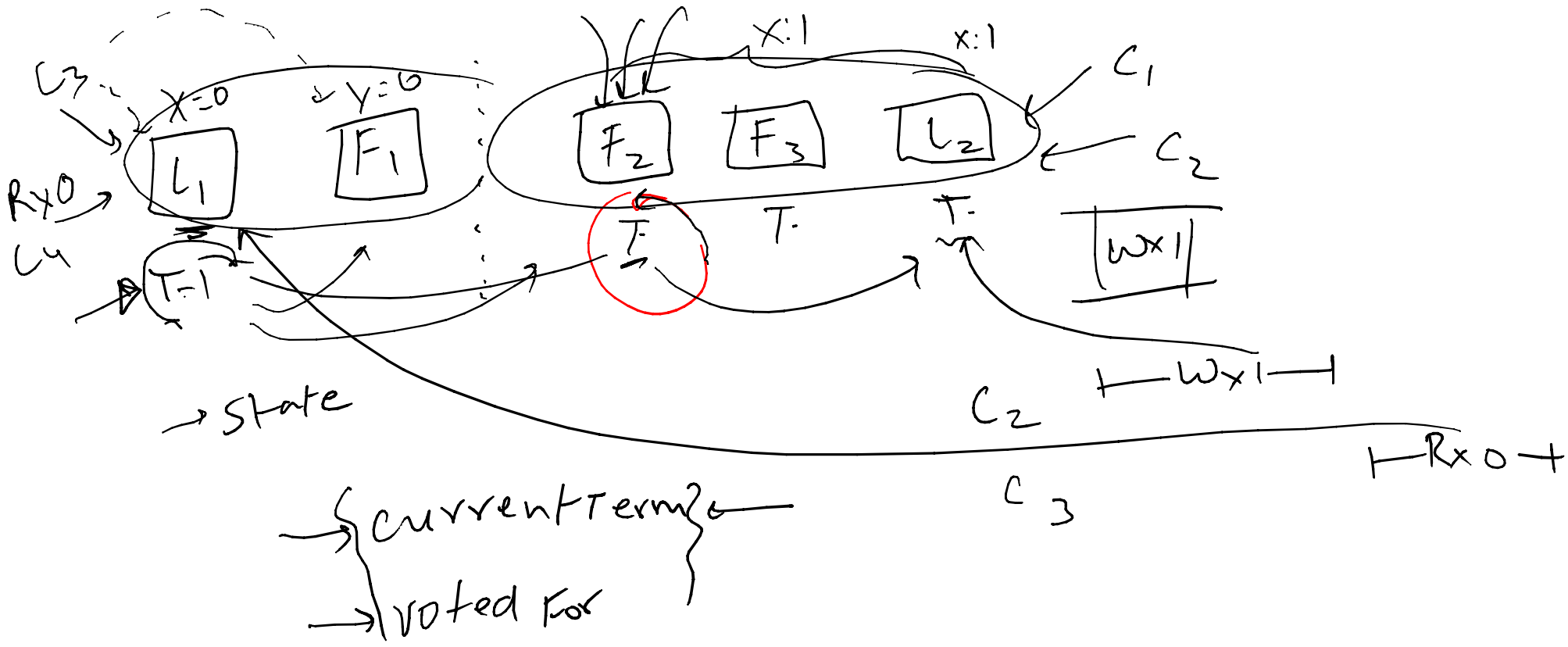
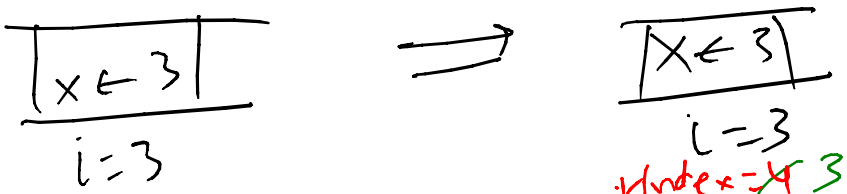


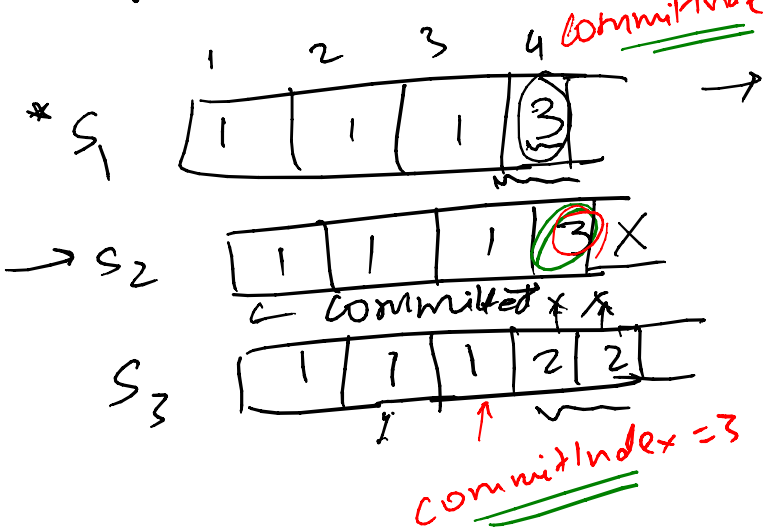
Figure 4: Server states. Followers only respond to requests from other servers. If a follower receives no communication, it becomes a candidate and initiates an election. A candidate that receives votes from a majority of the full cluster becomes the new leader. Leaders typically operate until they fail.



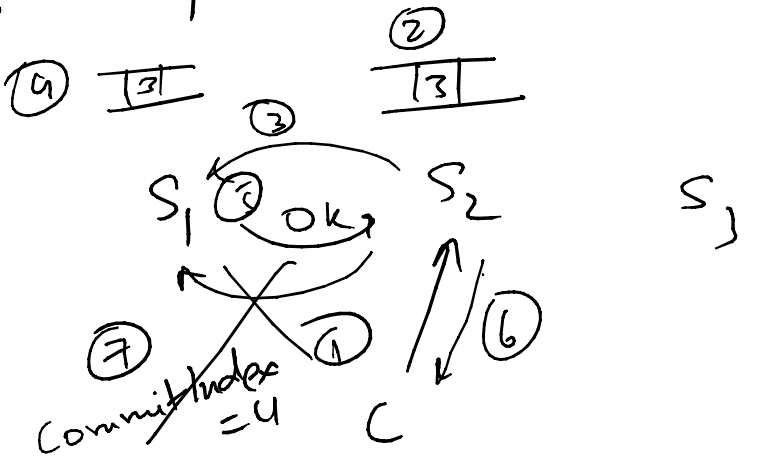

Randomized election timeout
t=2
t=2
1 ms - 2 ms
t=3
1 days 1 month
150 ms - 300ms
150 - 151


C3
X=0
Y=0
L1
F1
RxO
C4
T=1
X:1
X:1
C1
F2
F3
L2
C2
T.
T.
T.
WX1
WX1
C2
RxO
C3
State
currentTerm
votedFor





→ Leader Completeness

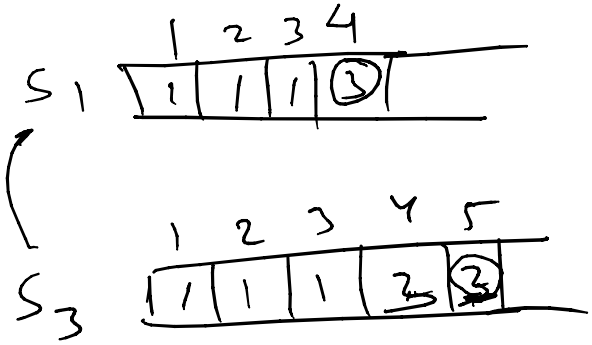


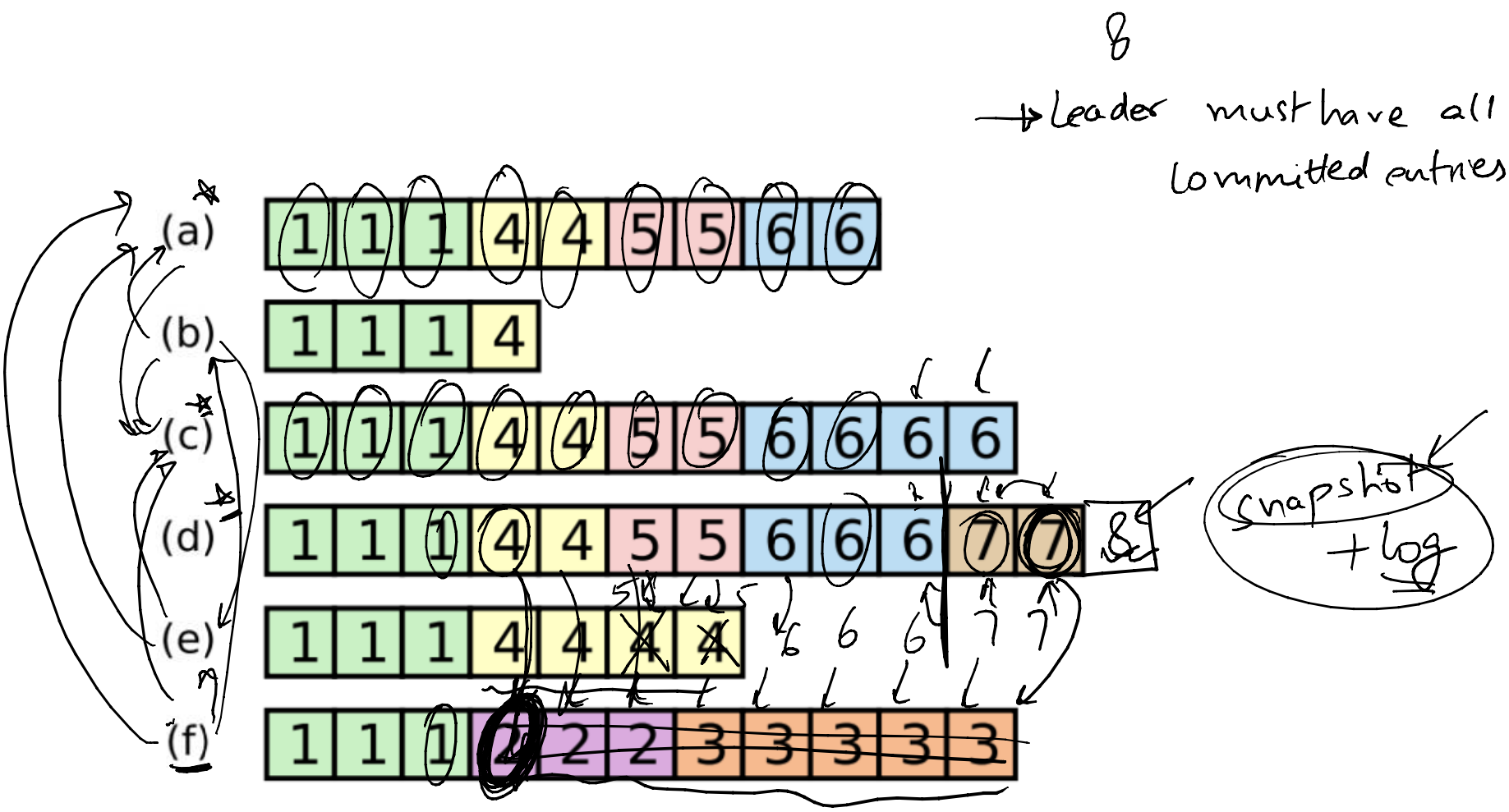
For my last log entry

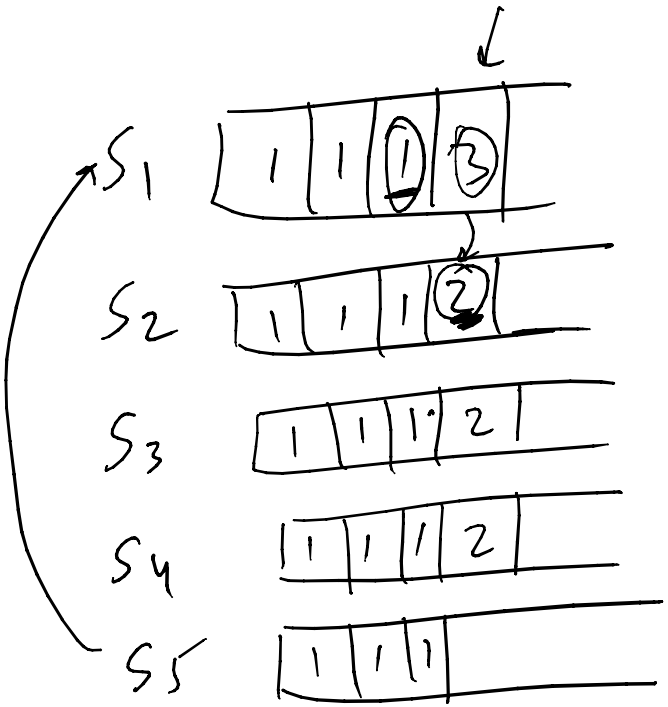
my term num > your term num

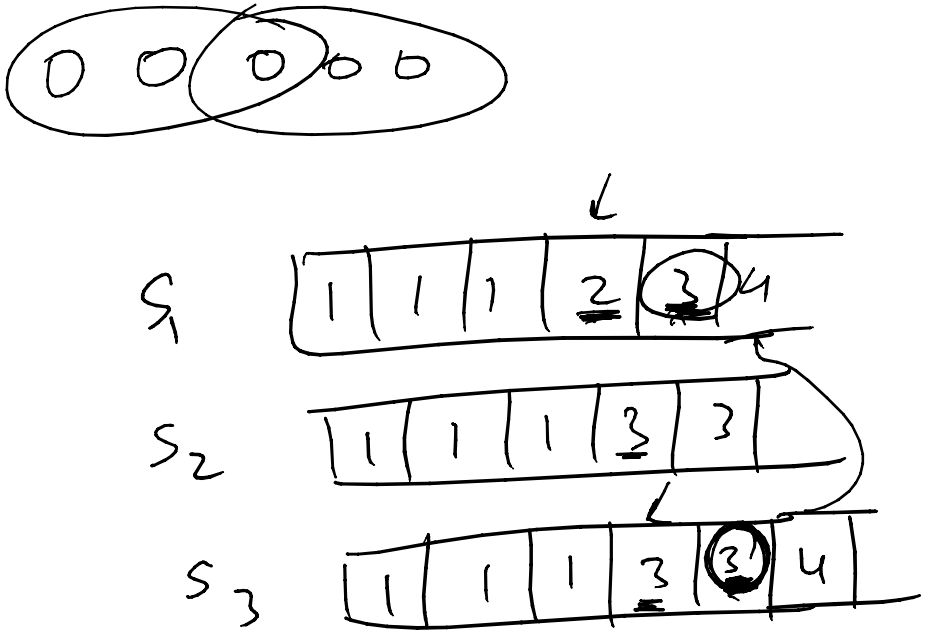
if =

my log length > your log length




8
→ Leader must have all committed entries
(a)
(b)
(c)
(d)
(e)
(f)
1 1 1 4 4 5 5 6 6
1 1 1 4
1 1 1 4 4 5 5 6 6 6 6
1 1 1 4 4 5 5 6 6 6 7 7
1 1 1 4 4 4 4
1 1 1 4 2 2 3 3 3 3 3
8
snapshot
+ log


S1
S2
S3
S4
S5


S1
S2
S3