DISTRIBUTED CONSENSUS PROTOCOLS

Mr. Sanjeev Kumar Dwivedi | Place: DSPM IIIT-NR

Date: 01.01.2021

Dr. Shyama Prasad Mukherjee International Institute of Information Technology, Naya Raipur

Supervised by-
Dr. Ruhul Amin
Dr. Satyanarayana Vollala
Department of CSE

Formal Definition

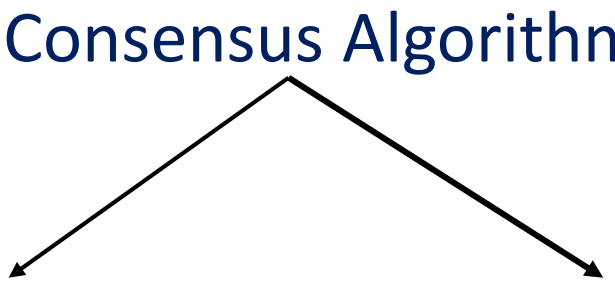
There are “n” nodes, each have an input value. Some nodes are faulty or malicious. A distributed consensus protocol[1ℓ] has the following two properties:

- ☞ The protocol terminates and all honest nodes are in agreement on the same value.

- ☞ This value must have been proposed by some honest node.

1ℓ : In the context of blockchain, consensus is the valid agreement for adding the new blocks in the blockchain network.

Division of Consensus Algorithm



Probabilistic Finality Algorithm (Chain-based)
E.g. Proof-of-work and its variants

Absolute Finality Algorithm (exact Finality)
E.g. Practical byzantine fault tolerance

Proof-of-work (PoW) Consensus

- Proof-of-work (PoW) is a combination of cryptography and computational power which ensure consensus and authenticity of the data recorded in the blockchain framework.
- The core idea of PoW is a solution that difficult to find but, very easy to verify.
- Bitcoin network use (PoW) as consensus protocol in implementation of the bitcoin network in 2009 by Satoshi Nakamoto.
- All participants of the blockchain network keep on calculating hash values using different nonce every time until, the target is achieved 2ℓ.
- When a peer is successful in computing the required hash value, all other participants must mutually agree on the correctness of the hash value.

2ℓ : prefix of current hash value is equivalent or lesser than the specific target value.

Cont…

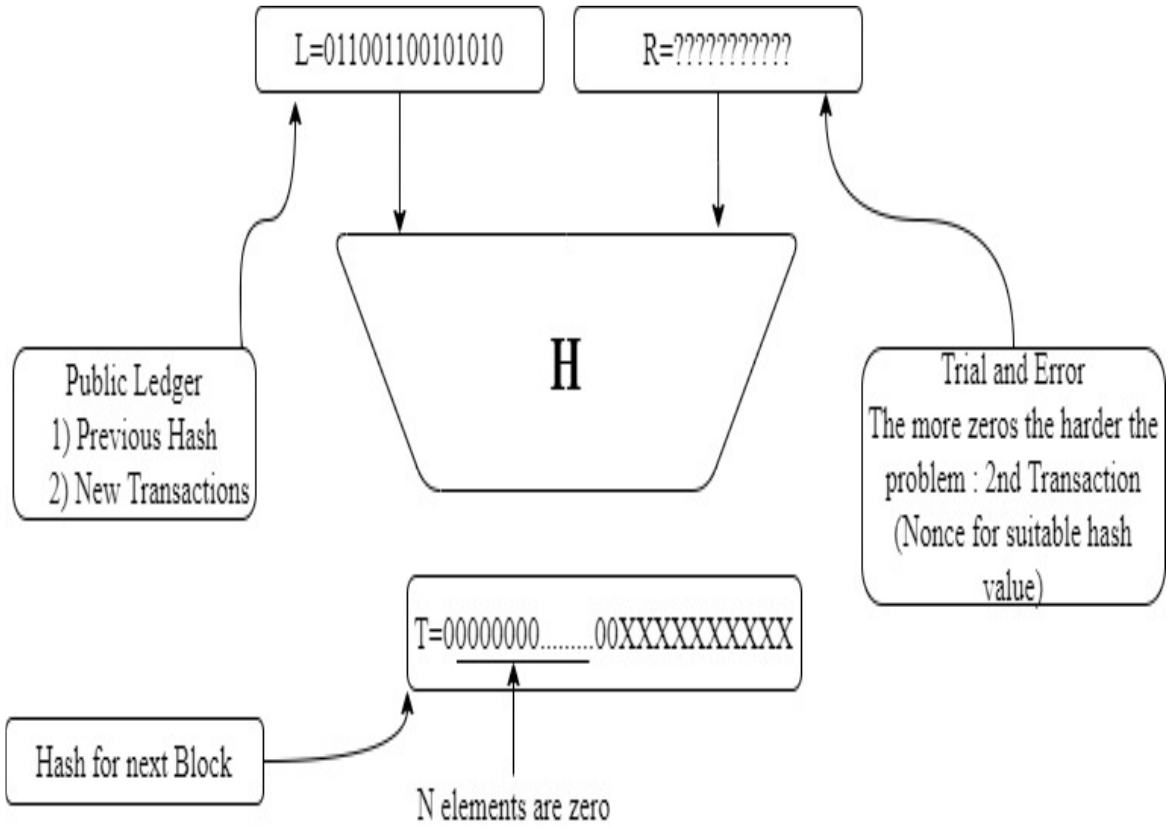


Fig. PoW Puzzle 3ℓ

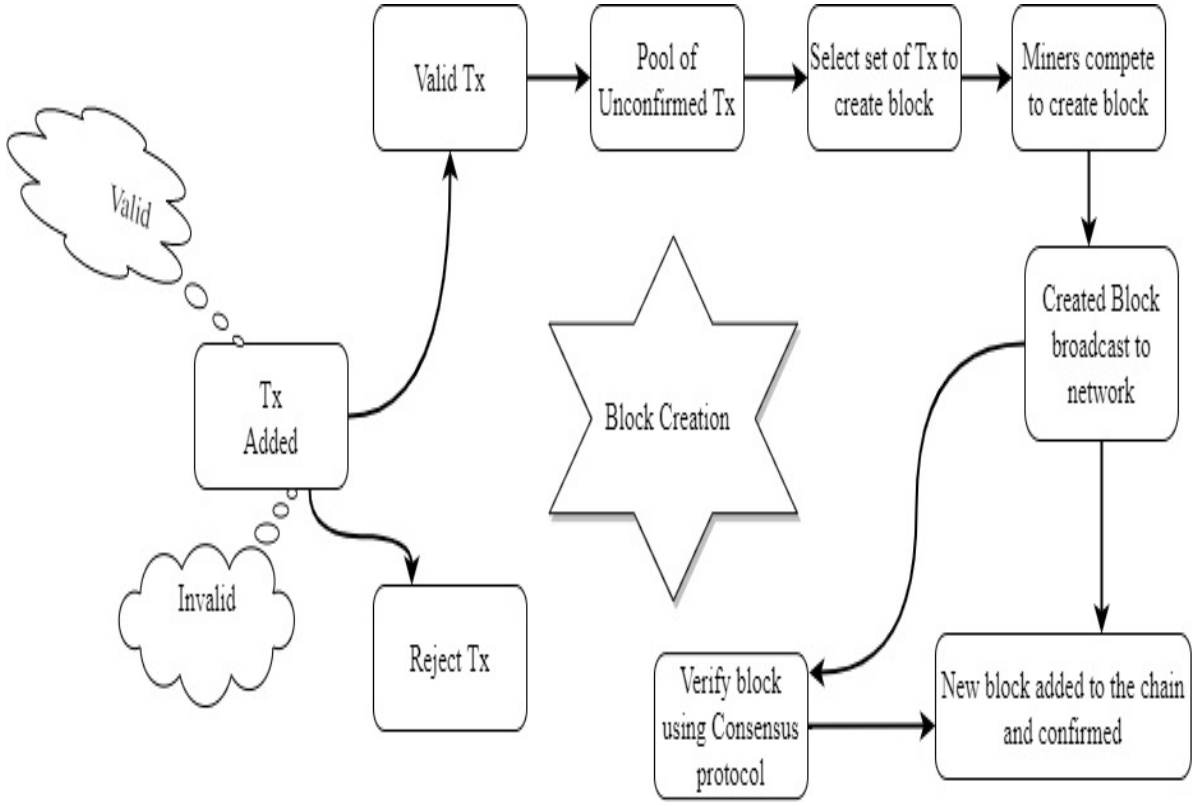


Fig. PoW Working Flow

3ℓ : $H(nonce \mid\mid prev\_hash \mid\mid t_x \mid\mid t_x \mid\mid \ldots\ldots t_x) < target.$

Cont…

- The collection of transactions used for the calculation of hash is considered as authenticated transactions, the nodes that compete to mine the blocks are called miners, and the PoW process is known as the mining algorithm.
- Calculation of the hash is a time-consuming process. Therefore to motivate the miners, an incentive mechanism is proposed.
- There is a possibility that two competing nodes may compute the hash and create a new block at the same instant. However, it is impossible that two contending forks will produce the next block at the same time. In such a case, the Longest chain becomes an authentic one.
- Example: Bitcoin.

Cont…

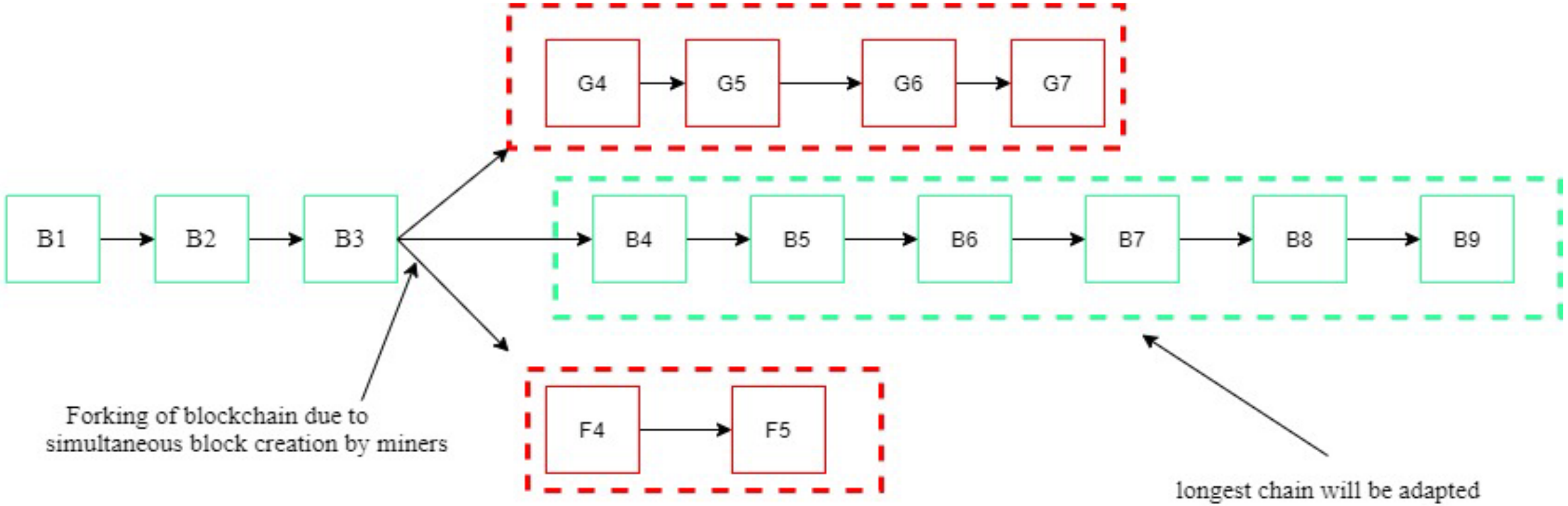


Fig. Blockchain Forking[4ℓ]

4ℓ : In Bitcoin blockchain, when around six blocks are axed, the applicable blockchain is seen as unchanging and credible, and each block is generated every 10 minutes.

PoW simplified:

- ☞ New transactions are broadcast to all nodes.
- ☞ Each node collects new transactions and verify this transaction (unspent and valid signature) .
- ☞ A random node collects the transaction (from pool of unconfirmed but valid transaction) and create a new block (solve hash puzzle).
- ☞ A random node broadacst this new block to all the peers nodes.
- ☞ Nodes express the acceptance of the block (verify the hash using nonce) by including its hash in the next block.

Proof-of-stake (PoS) Consensus

- PoW requires huge amount of energy (in terms of computational power), and has left the researchers to think for alternative of PoW to attain consensus in the blockchain network.
- Proof-of-stake (PoS) may be one of candidate to solve the energy requirement problem in the blockchain network.
- In PoS protocol, ownership of currency allows peer to participate in mining process (to validate the transaction and generate new blocks).
- In PoW, random node (no leader selection method is present) is create a new block. Whereas, in PoS, leader is selected based on the amount of stake which the miner currently holds proportion to the network capacity.

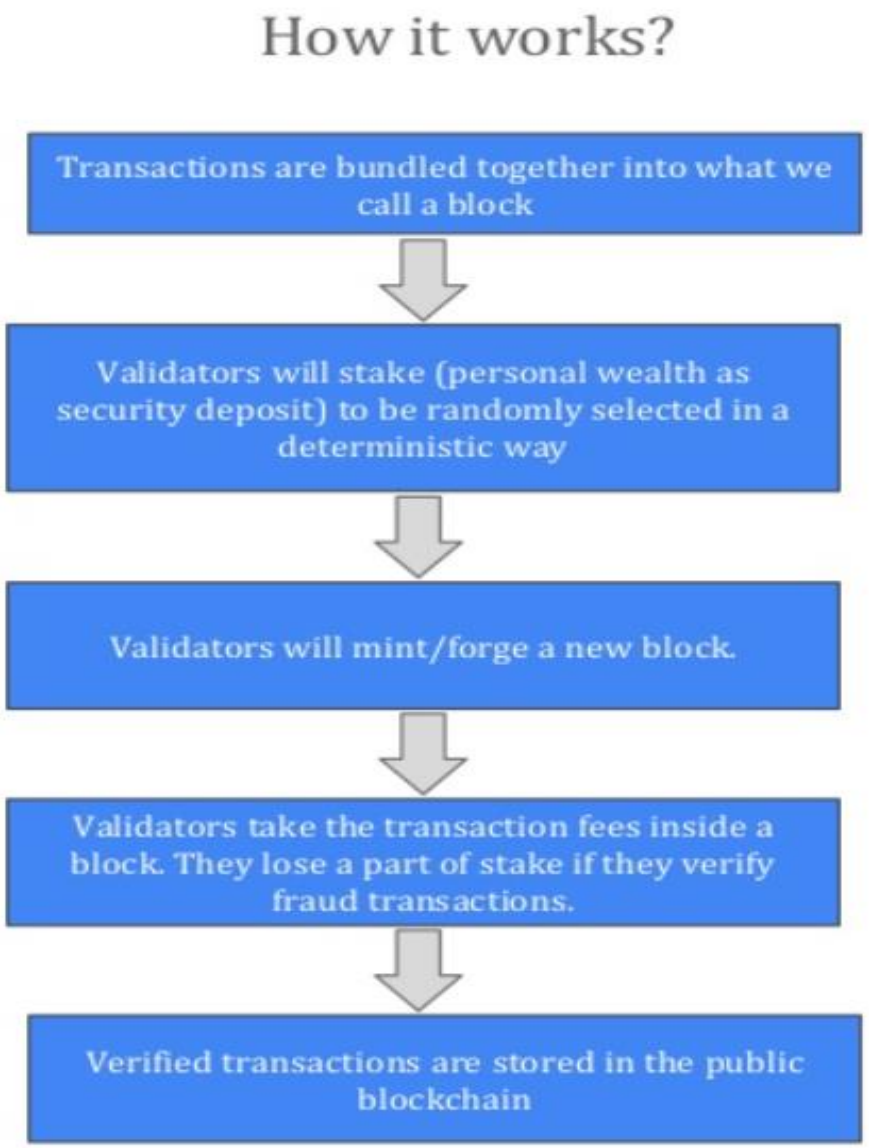


Fig. Working of PoS consensus

Points to be remember

- ☞ Creator of new block is chosen in a deterministic way depending on its wealth (stake).
- ☞ No block reward, so the validator takes the transaction fees.
- ☞ If a node stop being a validator, validator stake plus transaction fees will be released after a certain period of time.
- ☞ Example: Peercoin.

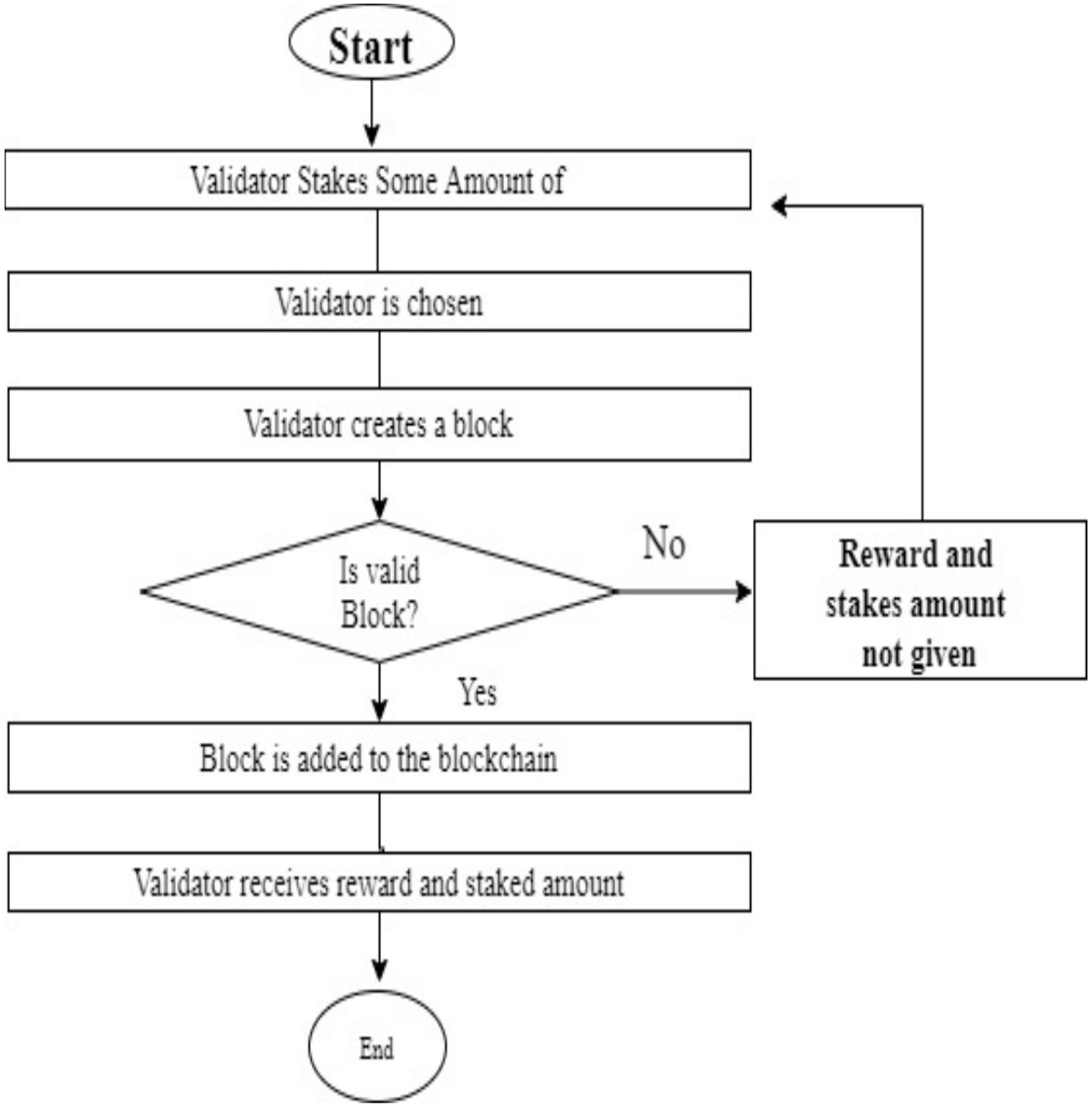


Fig. Flow of PoS consensus

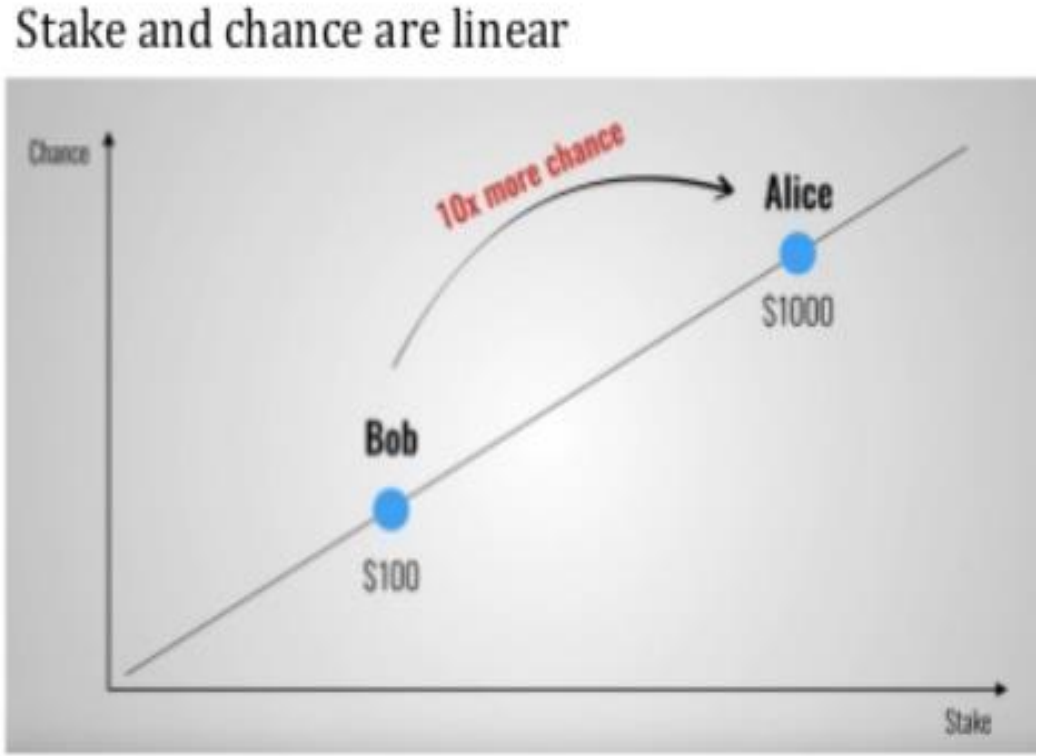


Fig. Relation between stake and its winning chance

Byzantine Fault Tolerance (BFT) Consensus

Faulty Component

- Crash Failure
- Byzantine Failure

To achieve the consensus in the presence of faulty component, the following goals must be satisfied by the system:

- Validity: Any value decided upon must be proposed by one of the process (proposer).
- Agreement: All non-faulty processes must agree on the same value.
- Termination: All non-faulty node eventually decide on the output value.

☞ In a message-passing system with n components, if f components are Byzantine and n ≤ 3f, then it is impossible for the system to reach the consensus goal.

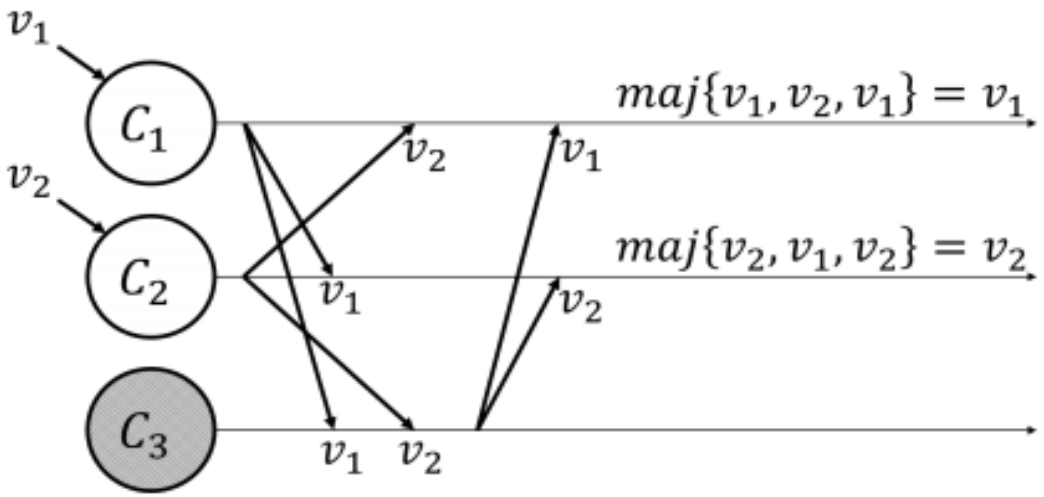


Fig. Three-component message-passing system with one component being Byzantine [5ℓ]

Conclusion: In the general case, for any distributed system with N components and f being Byzantine, N ≥ 3f +1 is required to ensure consensus.

5ℓ : Majority rule is applied to achieve the consensus.

Cont…

- The BFT consensus did not perform well for permissionless blockchain network. the reason for this is that in the permissionless system, we did not memorize the identity of peers node. Therefore, the number of nodes currently present in the system is an open question?
- Whereas, to achieve the consensus using BFT, it is desirable that every peers should know the identity of others and also number of nodes present in the system.

Conclusion: BFT consensus works well for permissioned blockchain system.

Practical Byzantine Fault Tolerance (PBFT) Consensus

- The practical Byzantine fault tolerance algorithm is an example of Byzantine fault tolerance (BFT), published by Miguel Castro and Barbra Lisko in 1999.
- It is a replication algorithm to deal with byzantine faults in a distributed network.
- To decide the faulty node, the honest nodes of the system reach a consensus and a system that can conclude is not affected by a malicious/faulty node.
- The communication overhead is more in the PBFT consensus.
- In the PoW Algorithm block is created by the winning miner node. In PoS, the block creator is the richest miner. Unlike PoW and PoS, in PBFT block is not generated by any special node, rather the most agreed block is committed to the chain.

Cont…

- PBFT has pre-prepared, prepared, and commit stages to complete the block creation process.
- PBFT system can tolerate up to n faulty node out of 3n+1 node.
- To make any decision, PBFT needs approval of [(3n + 1) – n] = 2n + 1 node from the network which has 3n + 1 node.
- In PBFT, the block is not created by a special miner but is the most agreed block from the network. PBFT protocol will append the most agreed block in the network.
- PBFT is an energy-efficient algorithm because the consensus is achieved without solving complex cryptographical mathematical puzzle and transactions do not require multiple confirmation

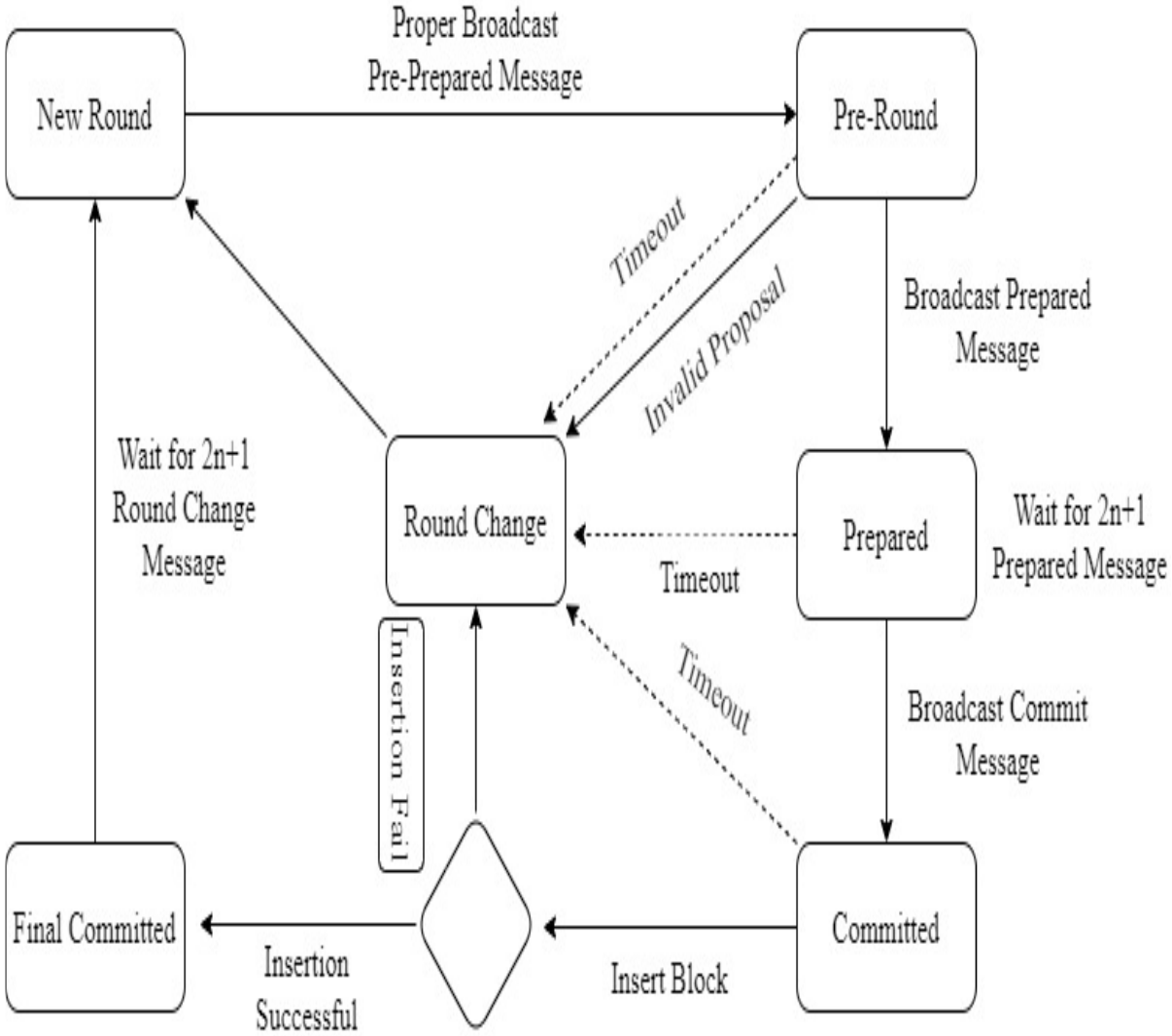


Fig. Process of PBFT mechanism

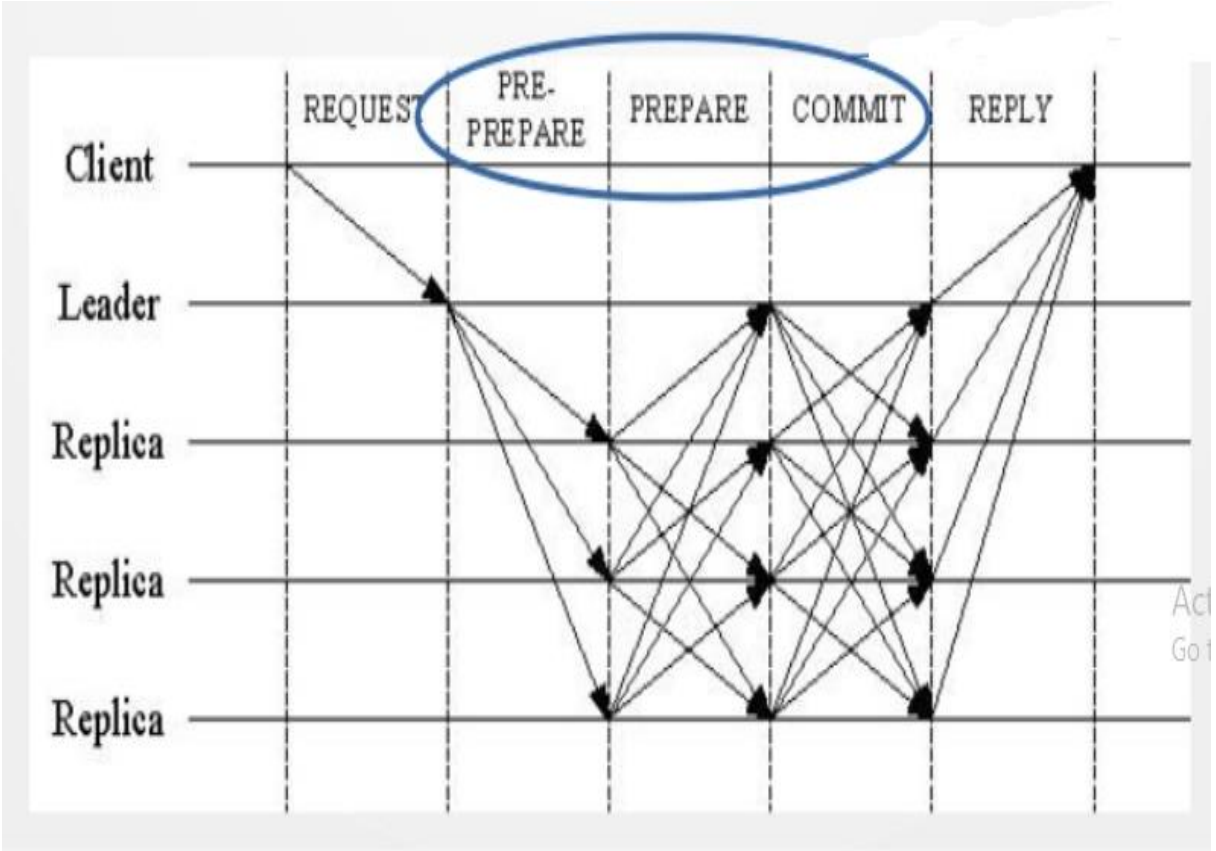


Fig. 3-phase protocol of PBFT

Example: Hyperledger Fabric

Paxos Consensus

☞ Using paxos consensus, one or more client can propose a value, and a majority of systems running Paxos agrees on one of the proposed values.

☞ Paxos has three roles of the nodes present in the network, proposer, acceptor, and learner.

- Proposer: the node who proposes value to reach on consensus (leader-node).
- Acceptor: the nodes who contribute to reach in consensus.
- Learner: the nodes who agree upon a consensus single value.

☞ The aim of the Paxos algorithm is to reach a single consensus. Once a consensus is achieved; it cannot switch to another consensus. If the network wants to reach another consensus, a different Paxos execution is required.

Assumptions for the PAXOS algorithm

- Concurrent proposals (One or more systems may propose a value concurrently).
- Validity (chosen value that is agreed upon must be one of the proposed values).
- Majority rule (To survive *m* failures, we will need *2m+1* systems).
- Asynchronous network (messages may get lost or arbitrarily delayed).
- Fail-stop faults.
- Unicasts (no mechanism to multicast a message).
- Announcement (Once consensus is reached, the results can be made known to everyone).

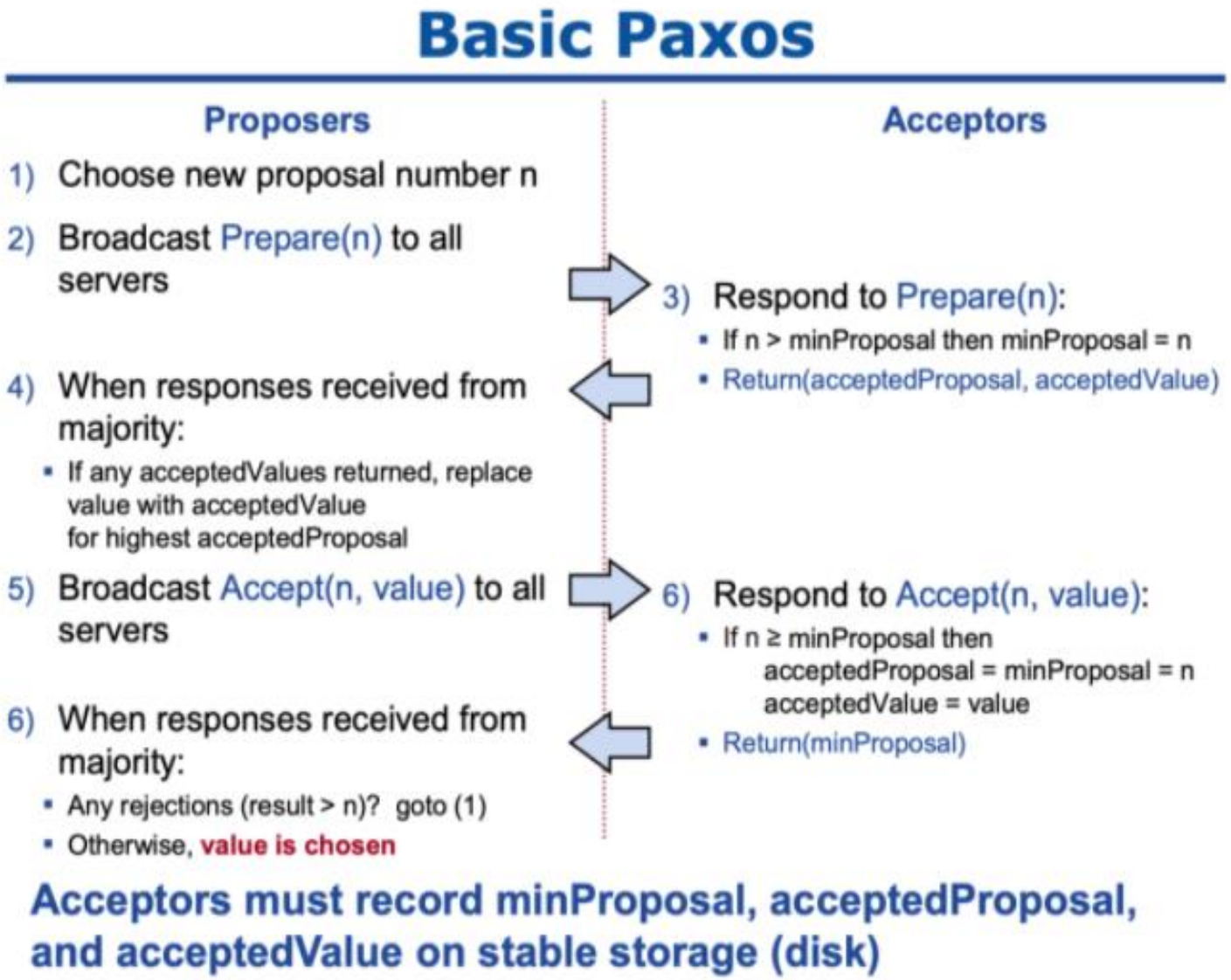
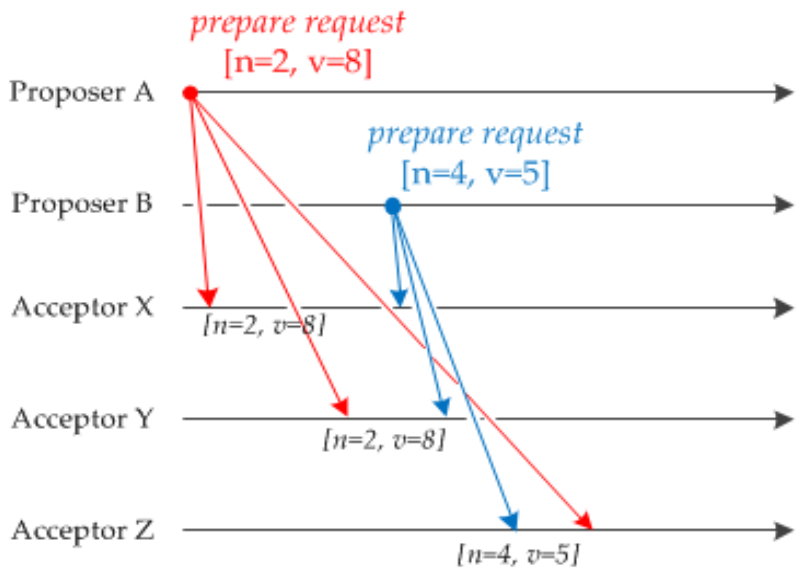
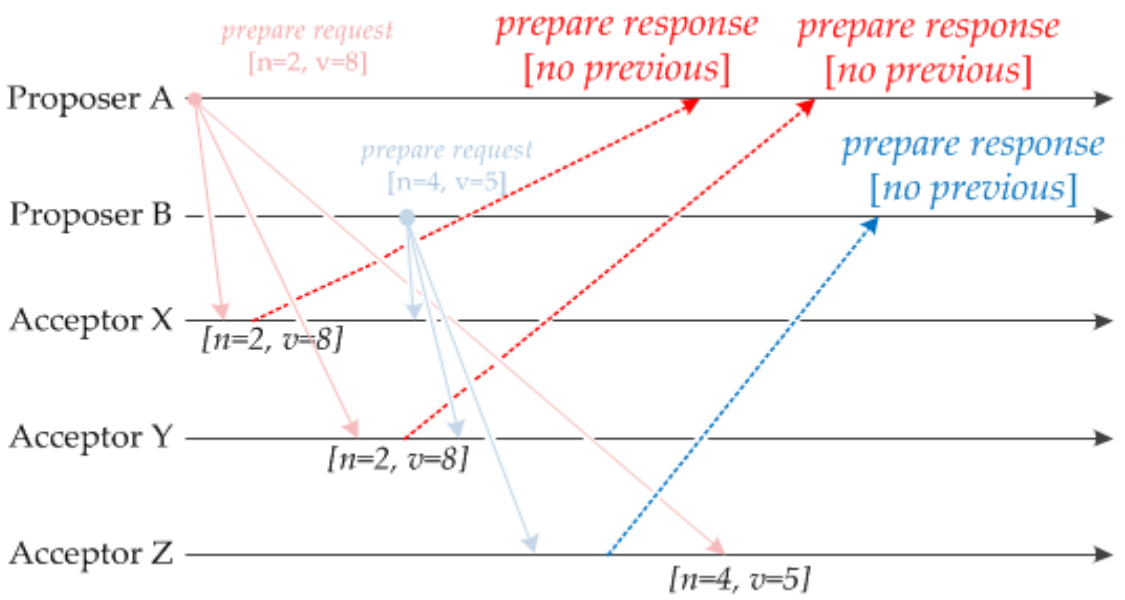


Fig. Process of PAXOS mechanism

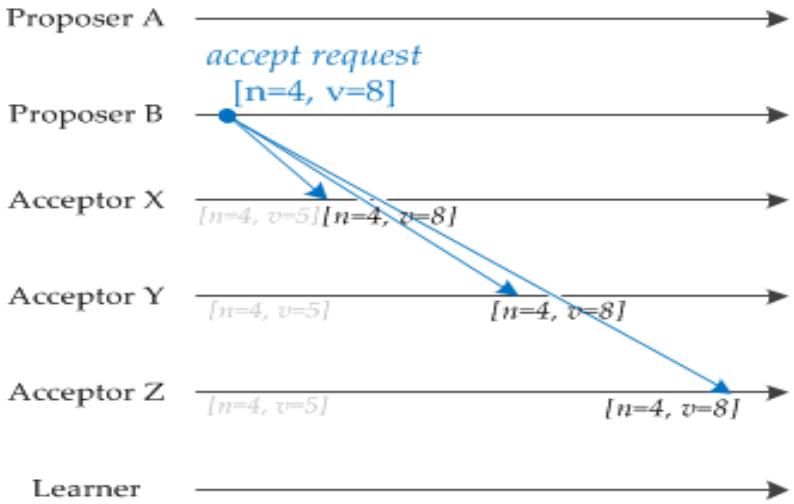
(1)



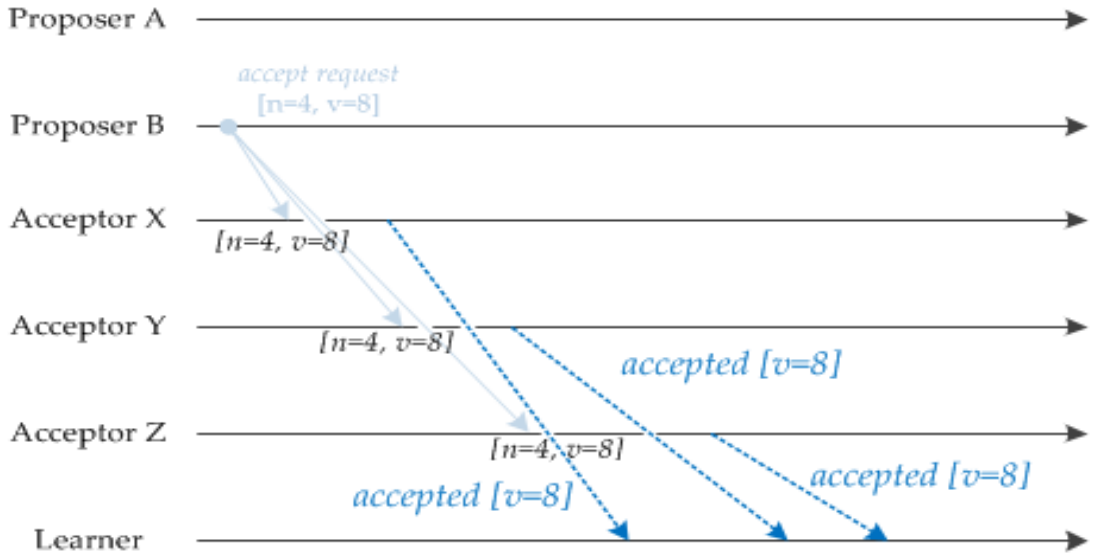
(2)



(3)



(4)



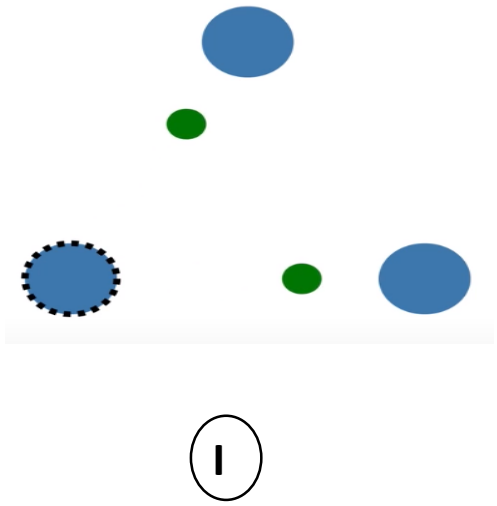
Note: Here, accepted value of “v” is 5 , not 8.

Replicated and Fault Tolerant (RAFT) Consensus

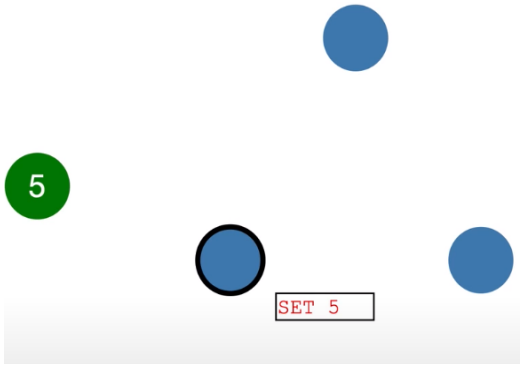
- In the RAFT consensus, a node can be in three state: follower, candidate, and leader.
- All the node start with the follower state and if the follower do not hear from leader, than it can be candidate node.
- The candidate node sends the request vote to the other followers and followers reply with their vote.
- If the candidate node gets the majority of votes from the followers, then it becomes a leader node. This process is called a leader election process.
- Now, all the changes in the system is performed by the leader node.

Cont…

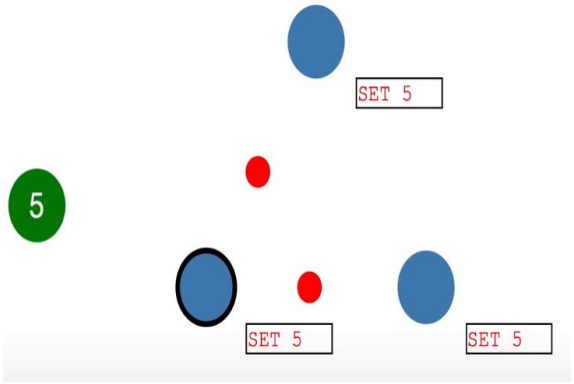
- The client first sends the log entry (for e.g. set 5) to the leader node and currently this log entry is in the uncommitted state.
- To commit the log entry, the leader node first replicated the log entry to the other followers, and leader node waits until the majority of followers have written the log entry.
- If the majority of the followers node reply positively, the log entry is now committed by the leader node. (the leader is in state 5 as in example).
- The leader notifies the followers for node entry which is committed earlier and cluster of nodes reach into consensus for system state and leader notifies the client as well. This process is called as log replication.



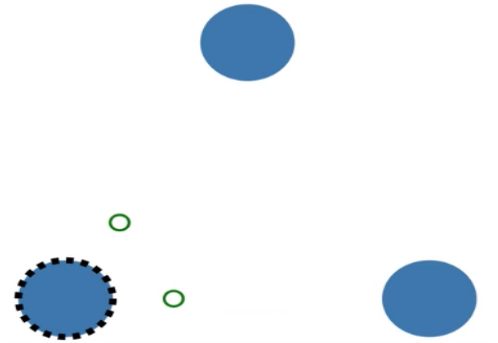
I



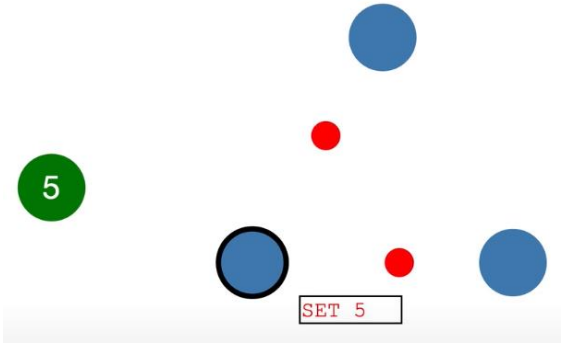
III



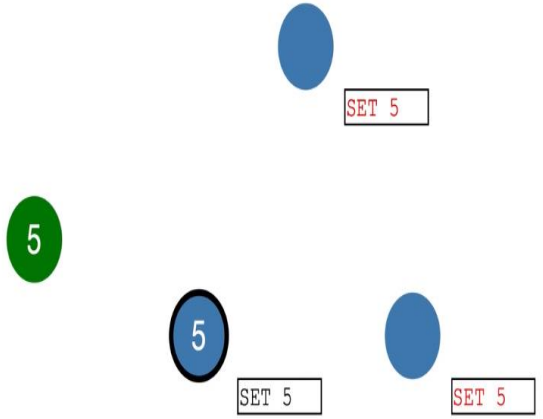
V



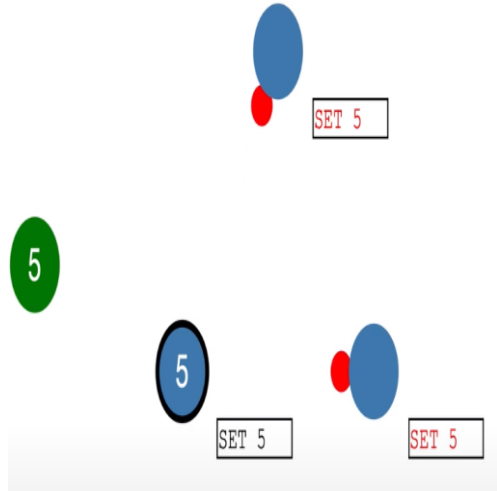
II



IV



VI



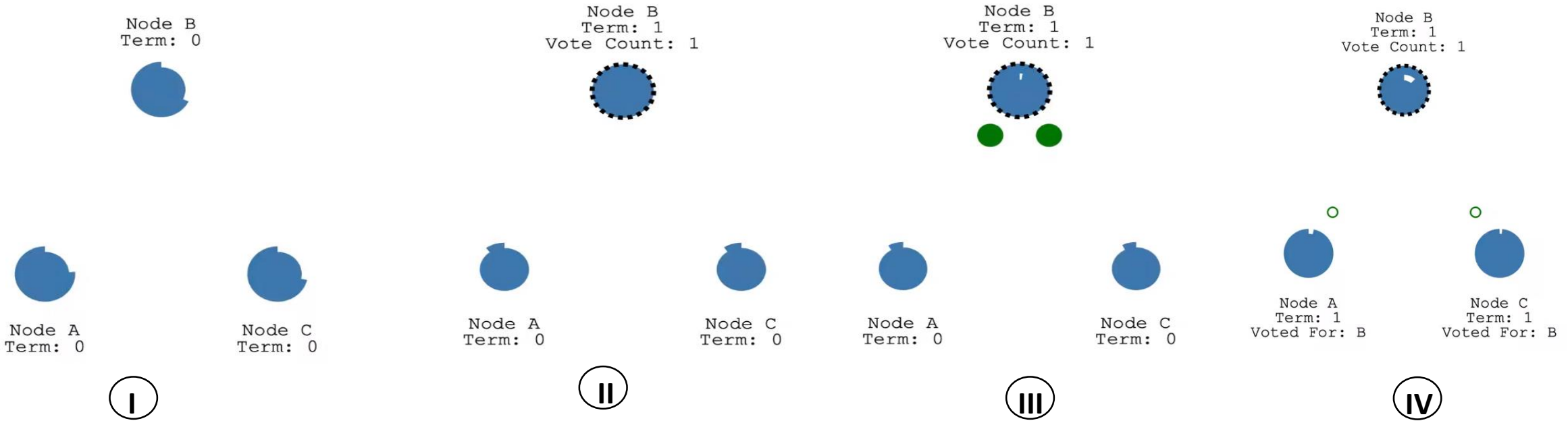
VII

Cont…(Leader-election)

- In Raft consensus, election timeout timer present which controls the election of leader node.
- The election timeout is the amount of time, the followers wait, till it becomes a candidate node. The election timeout is randomized between 150msec to 300msec.
- After the election timeout (who complete first), one of the follower becomes a candidate node and it starts a new election term.
- Now, the candidate node sends the request vote message to other followers node. If the receiving nodes are not voted yet in the current term than it vote for candidate node and all the node reset its election timeout.

Cont…

☞ Once a candidate node gets a majority of votes from other followers node, than the candidate become a leader.



Thank you

Dr. Shyama Prasad Mukherjee International Institute of Information Technology, Naya Raipur