PingCAP

Raft in Theory and Practice

Building a Distributed Consistent Database

关注公众号回复help,
可获取更多经典学习
资料和文档，电子书

PingCAP

About me

- Tang Liu
- Chief Architect of PingCAP
- TiDB/TiKV
- Open sources: LedisDB, go-mysql, etc...

PingCAP

Agenda

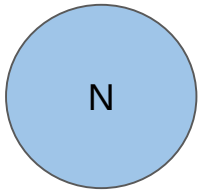
- Why Raft
- What is Raft
- The optimization
- Multi-Raft

PingCAP

Why

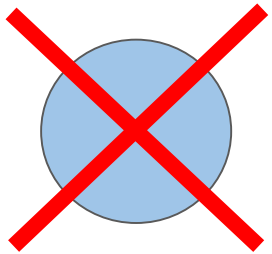
PingCAP

One Node



PingCAP

One Node

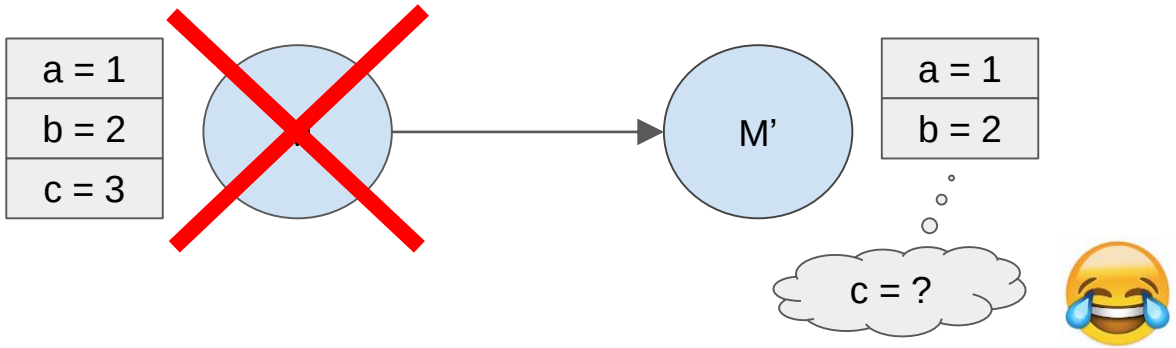


PingCAP

Replication

PingCAP

Async Replication

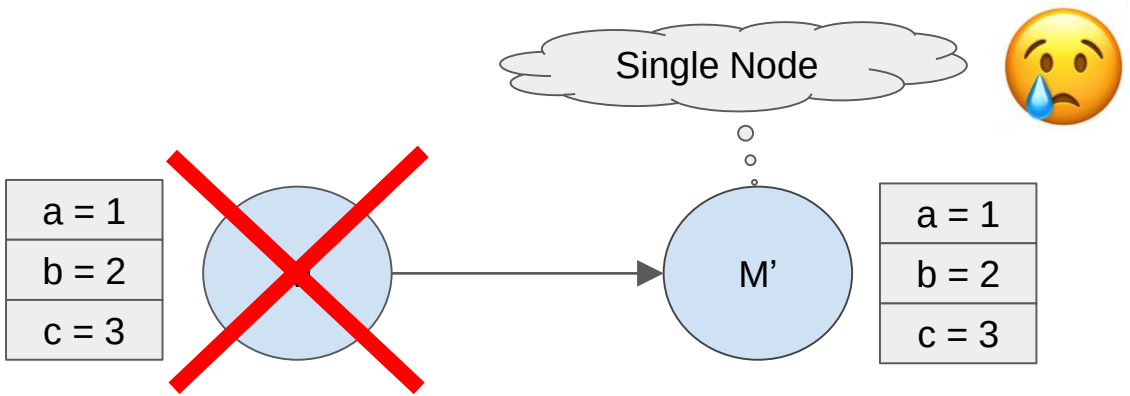


PingCAP

Sync Replication

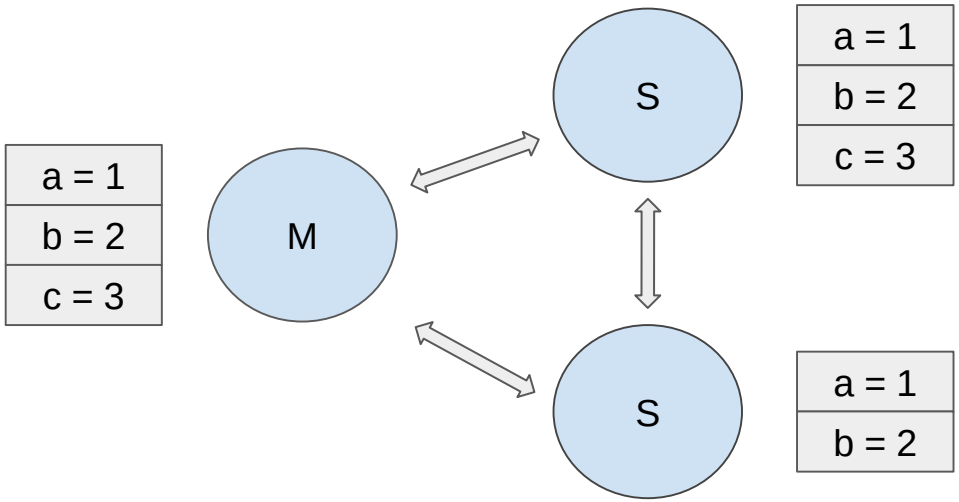
PingCAP

Sync Replication



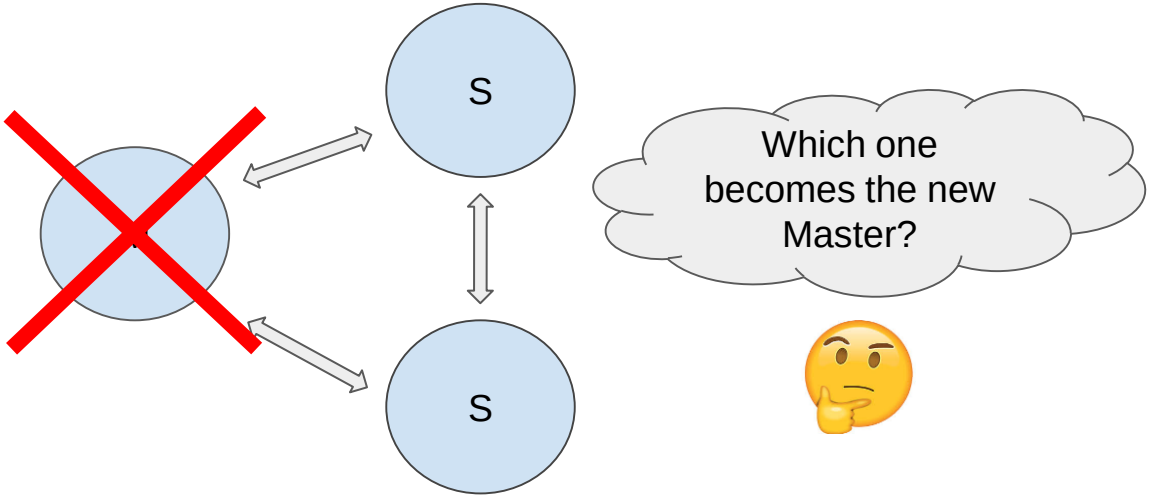
PingCAP

Quorum , 2 N + 1



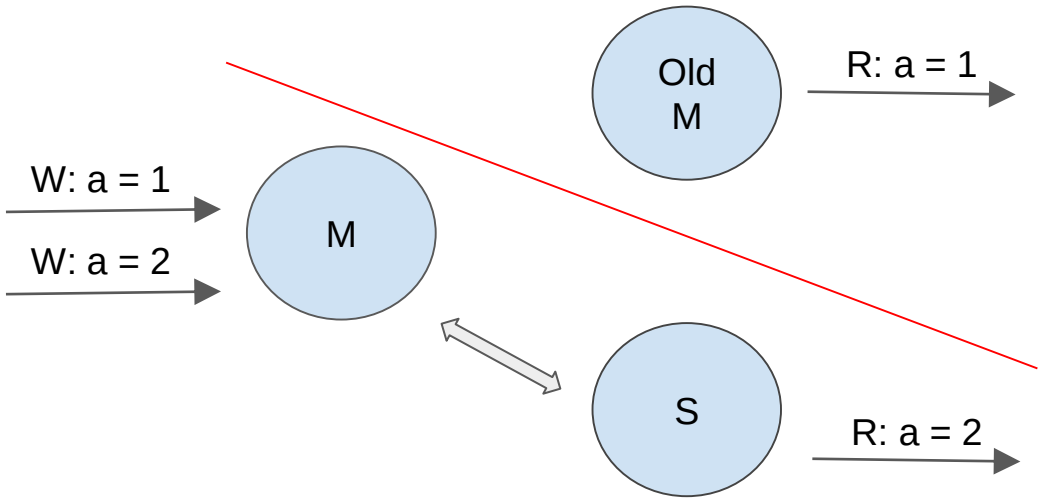
PingCAP

Failover



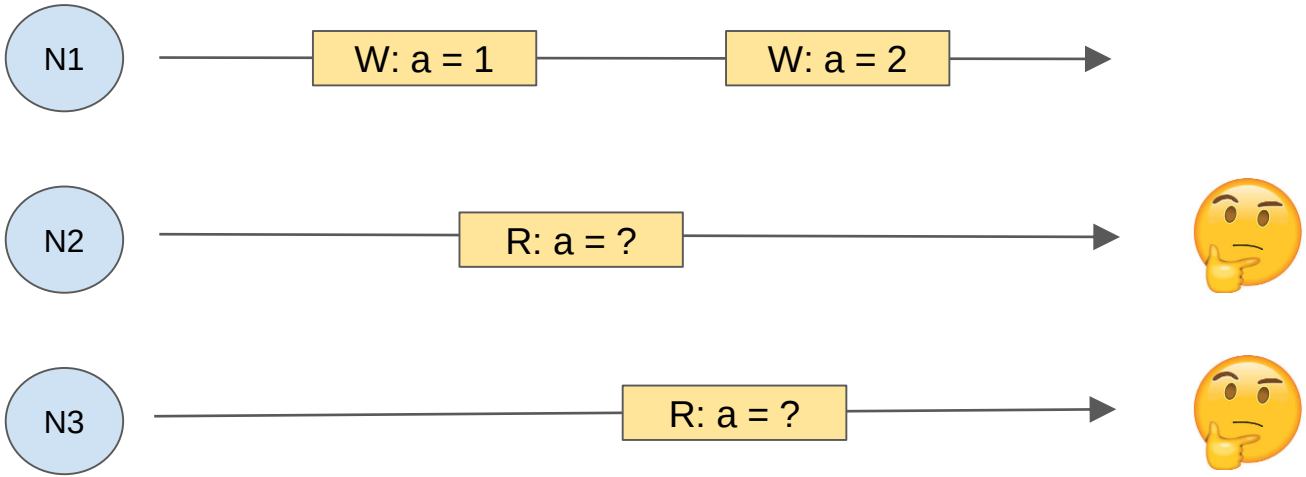
PingCAP

Brain Split



PingCAP

Linearizability



PingCAP

Consensus Algorithm

- Agreement on shared state
- Recovery from server failures autonomously
  - Minority failure: no problem
  - Majority failure: lose availability, retain consistency
- Strong consistency (linearizability)

PingCAP

Raft

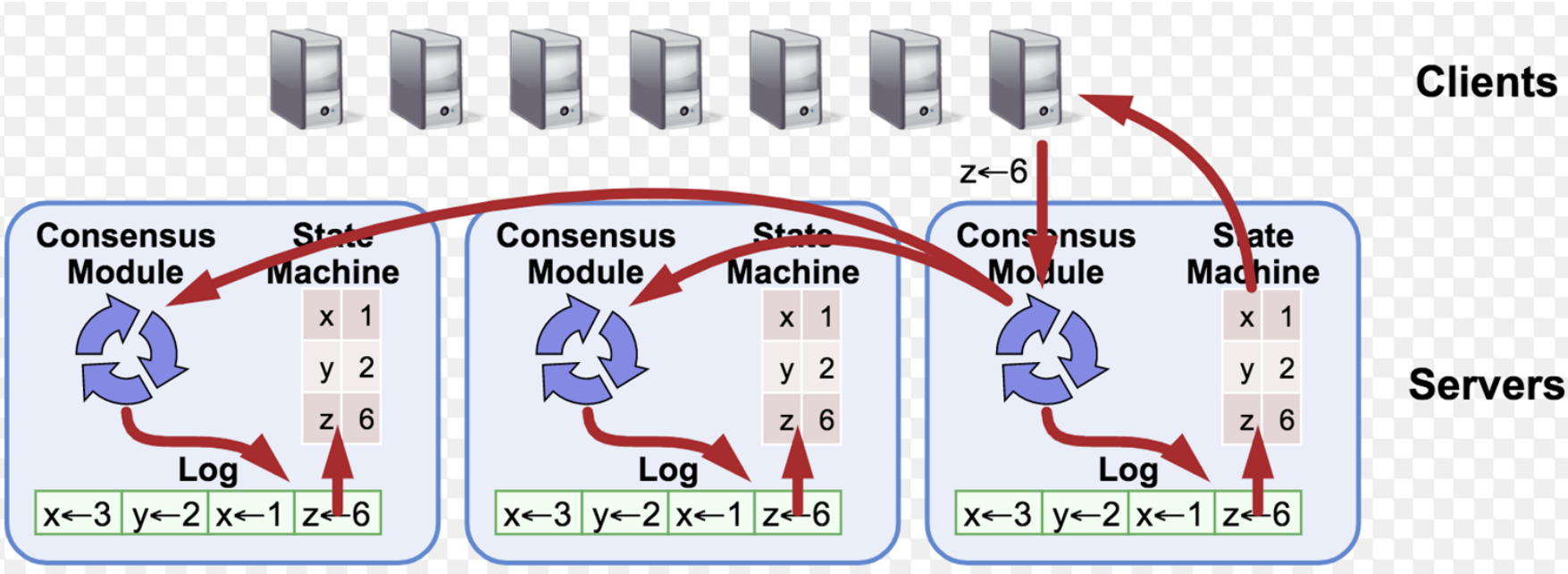
PingCAP

Keypoint

- Replicated Log for replicated state machine
- Leader, Follower and Candidate
- Term and Leader election
- Log Replication
- Membership Change

PingCAP

Replicated State Machine



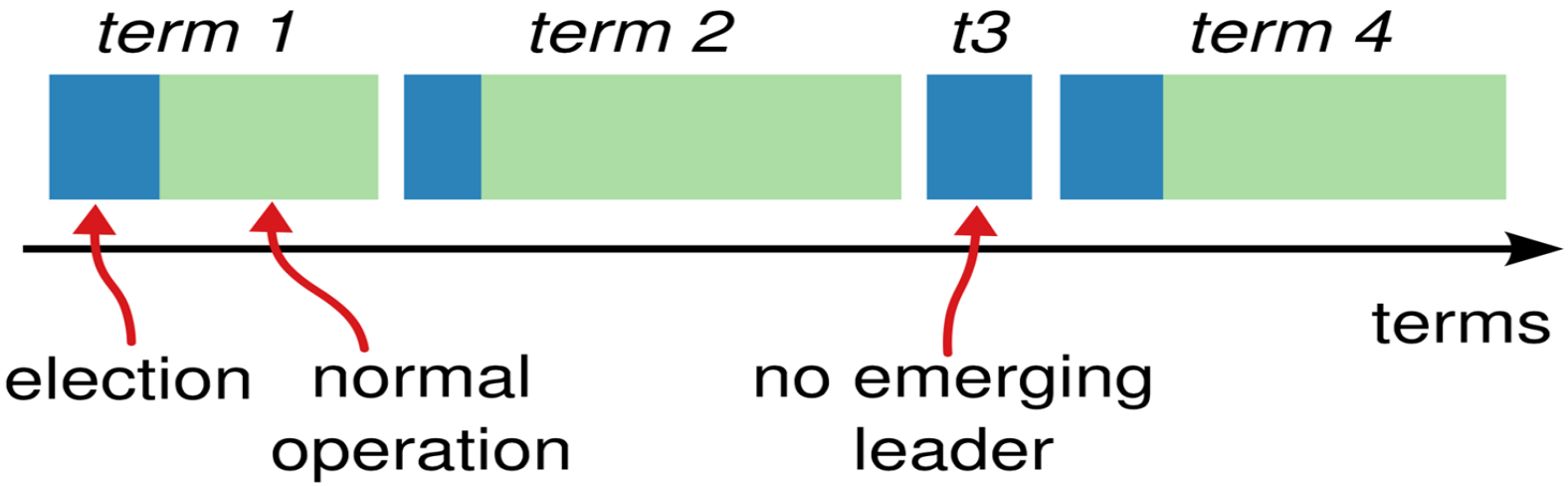
PingCAP

State

- Leader
  - Only one Leader
  - Elected by the majority of the peers
  - Handles all the client requests
- Follower
  - Receives Replicated Logs from the Leader
- Candidate
  - For campaign

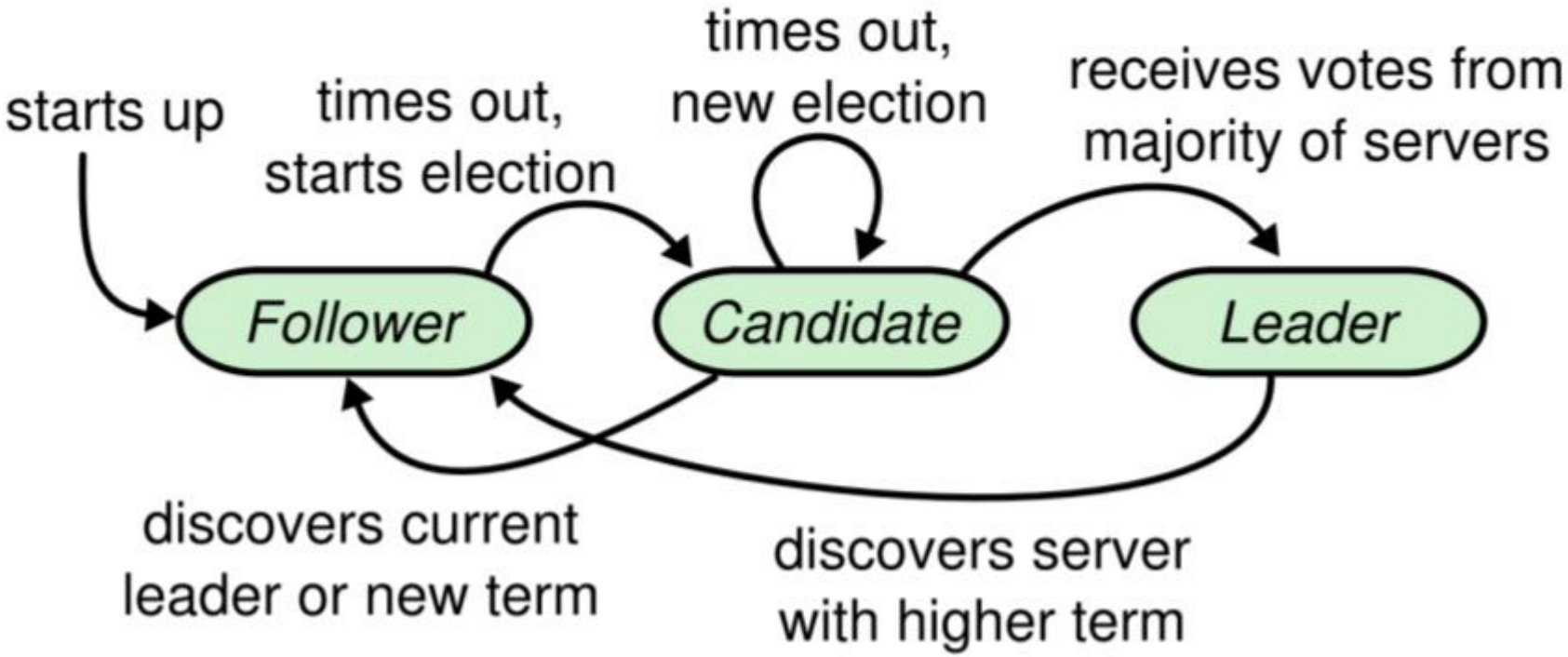
PingCAP

Term



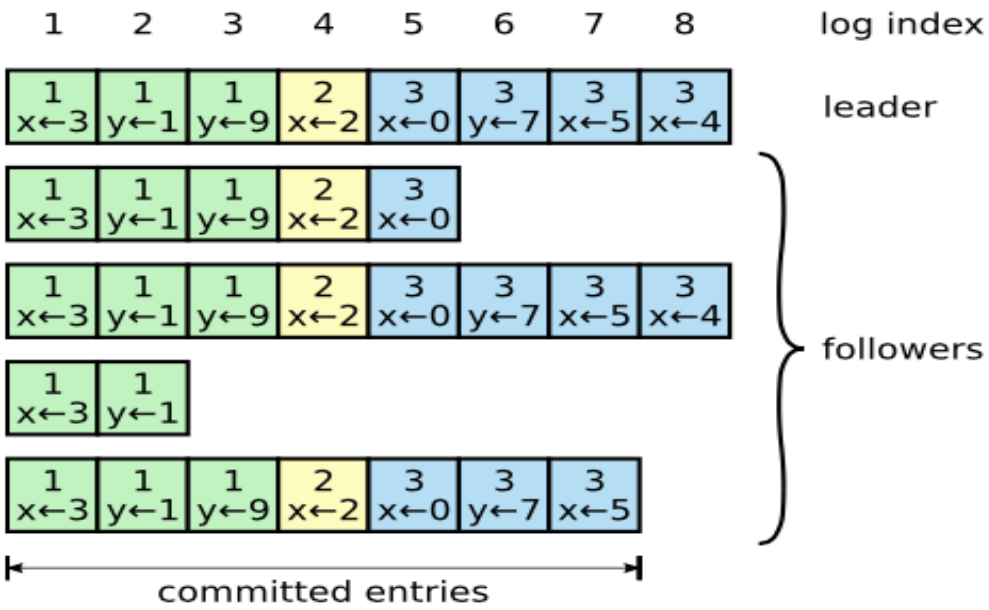
PingCAP

Election



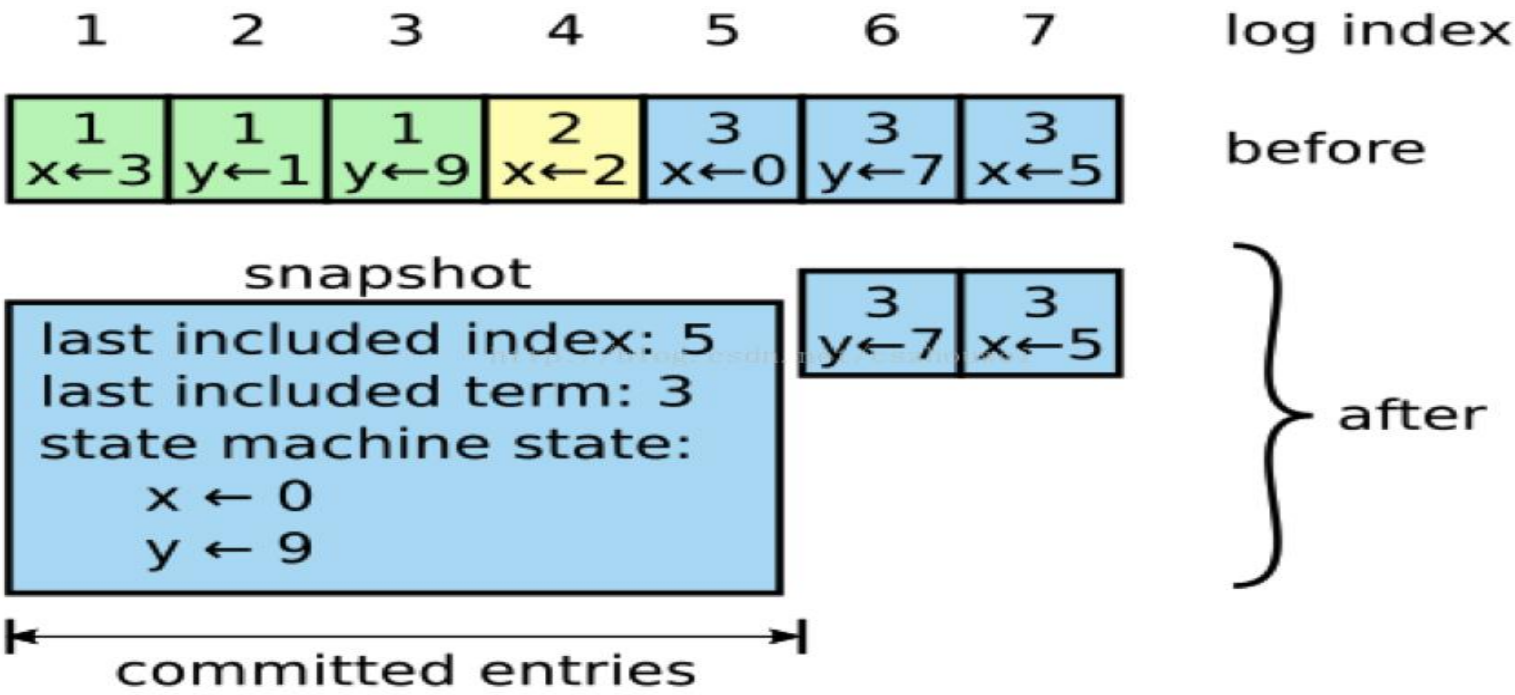
PingCAP

Log Replication



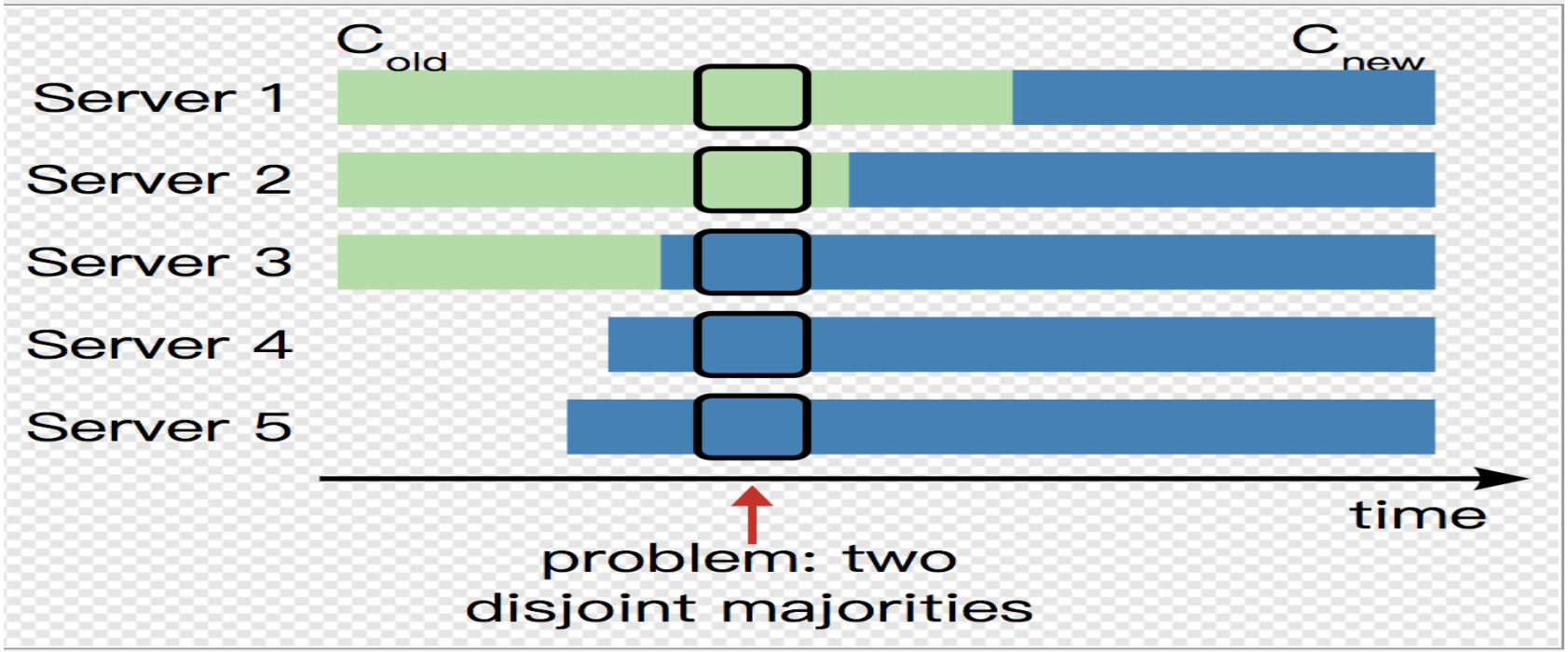
PingCAP

Snapshot



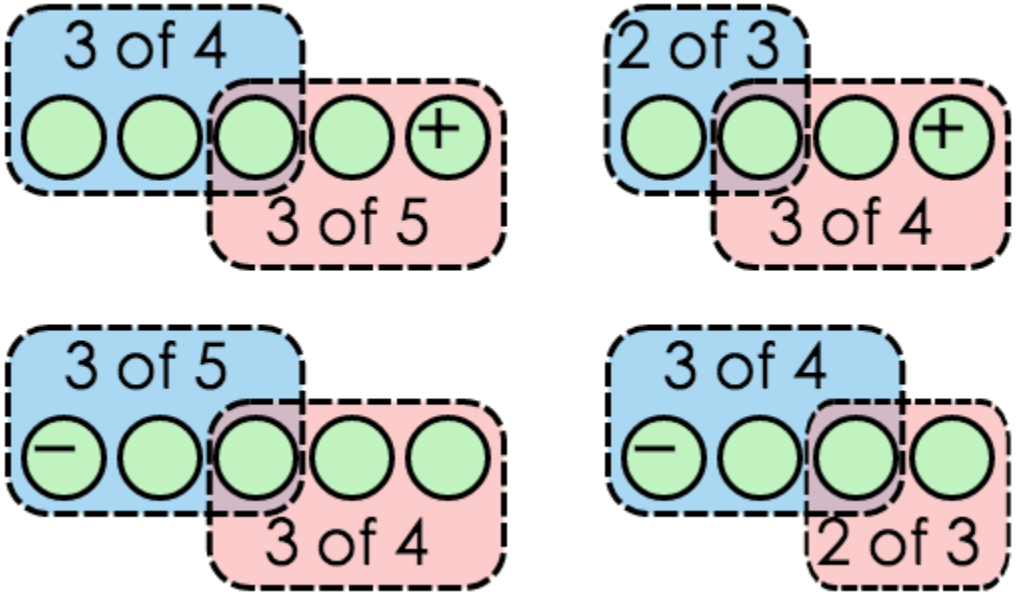
PingCAP

Membership Change



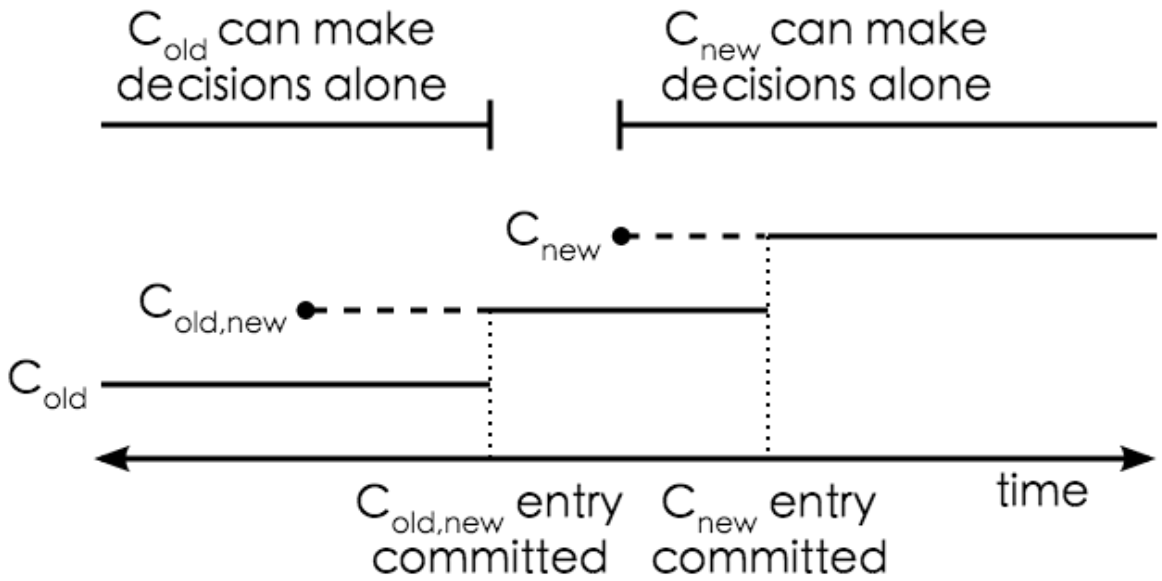
PingCAP

Membership Change



PingCAP

Membership Change

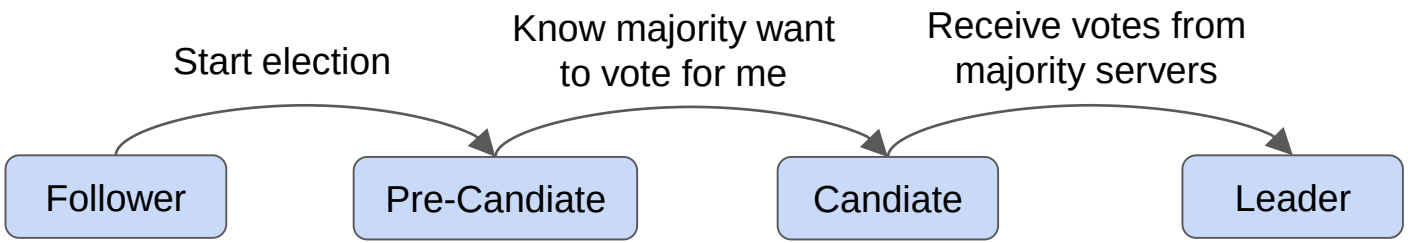


PingCAP

Optimization

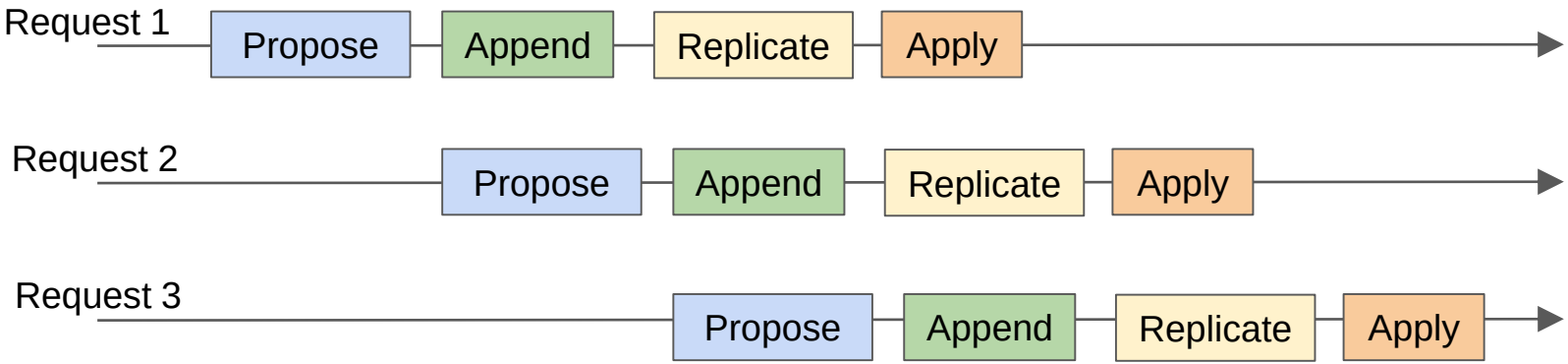
PingCAP

Pre-Vote



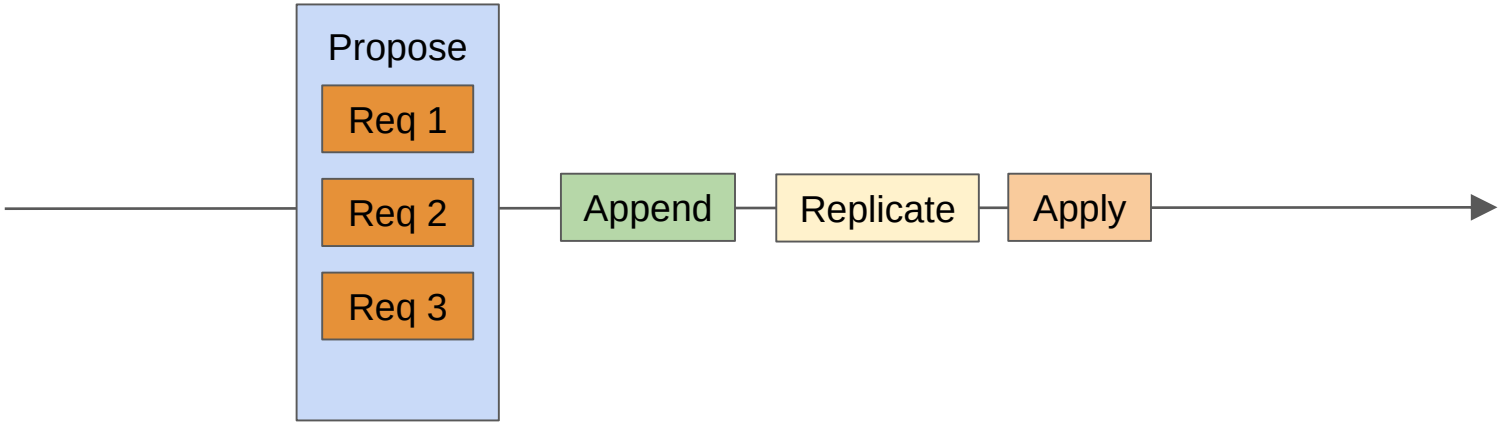
PingCAP

Pipeline



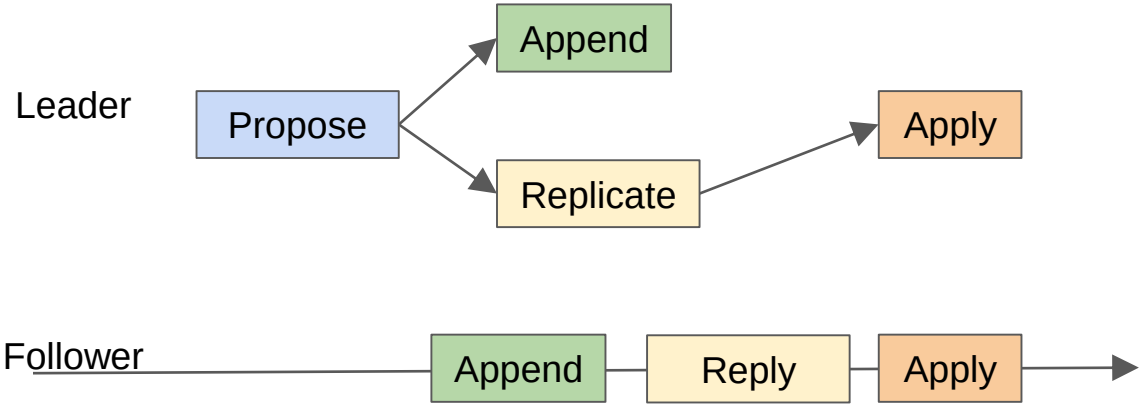
PingCAP

Batch



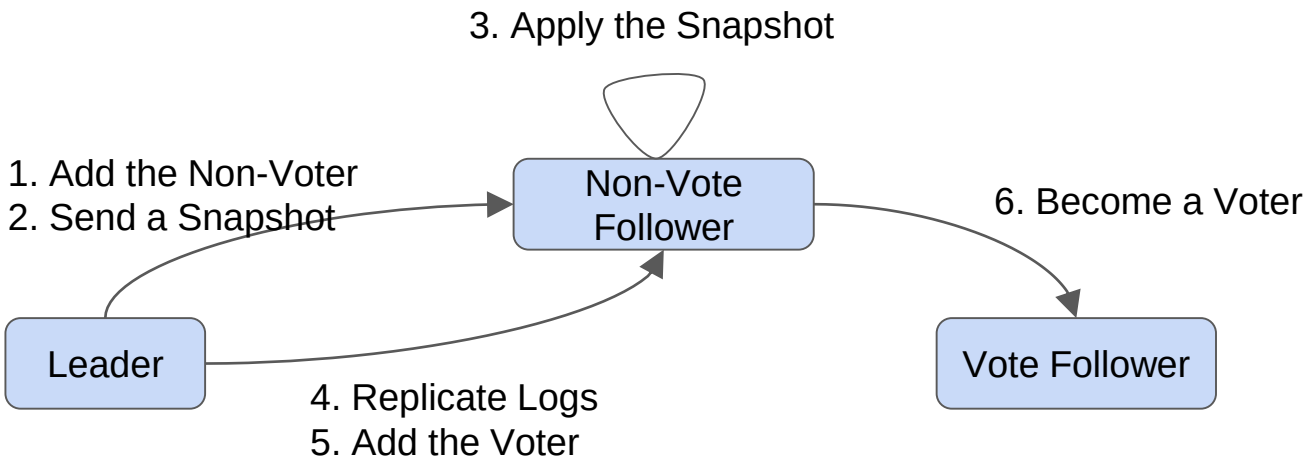
PingCAP

Writing to the Leader in parallel



PingCAP

Non-Voter



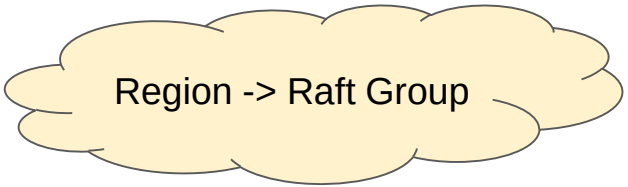
PingCAP

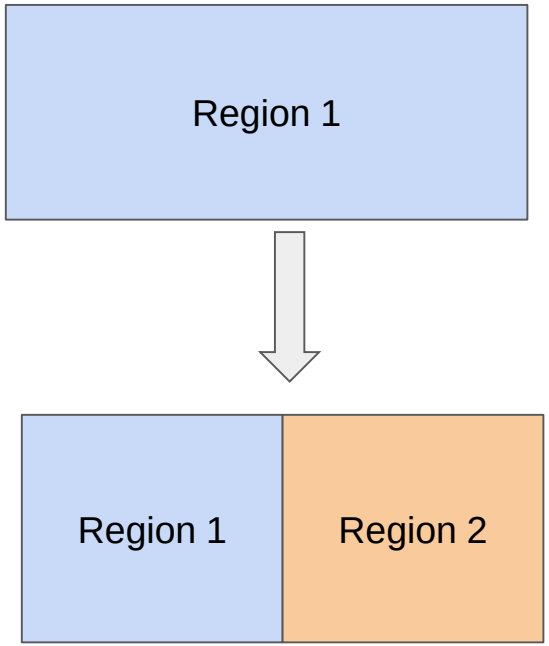
Multi-Raft

PingCAP

How to scale?

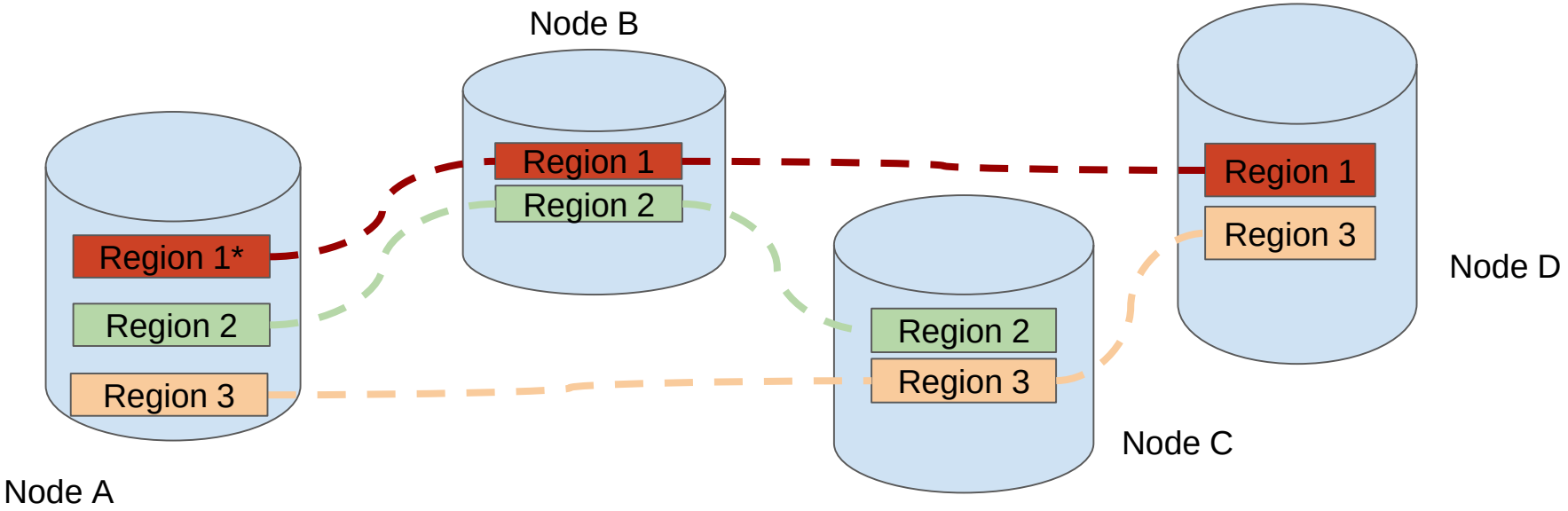
- **Logical split**
- Just Split && Move



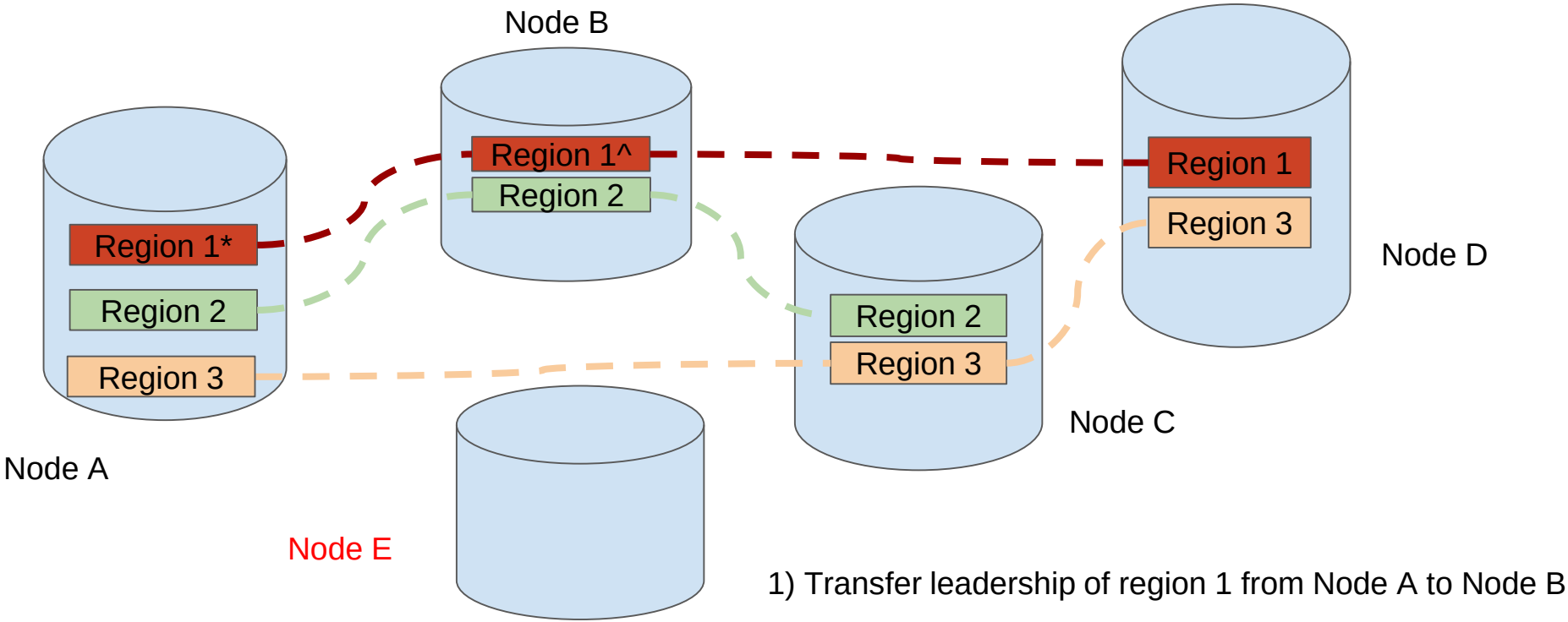


PingCAP

Scale-out (initial state)



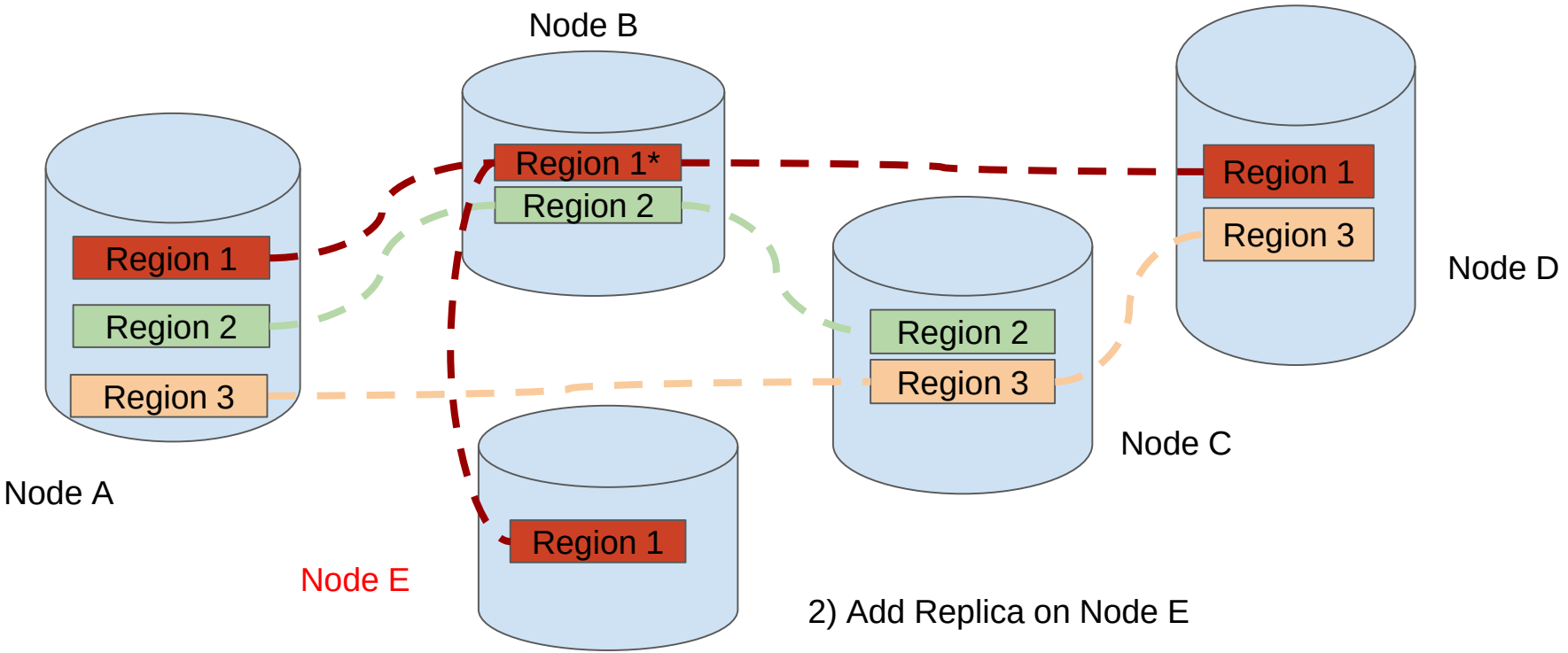
Scale-out (add new node)



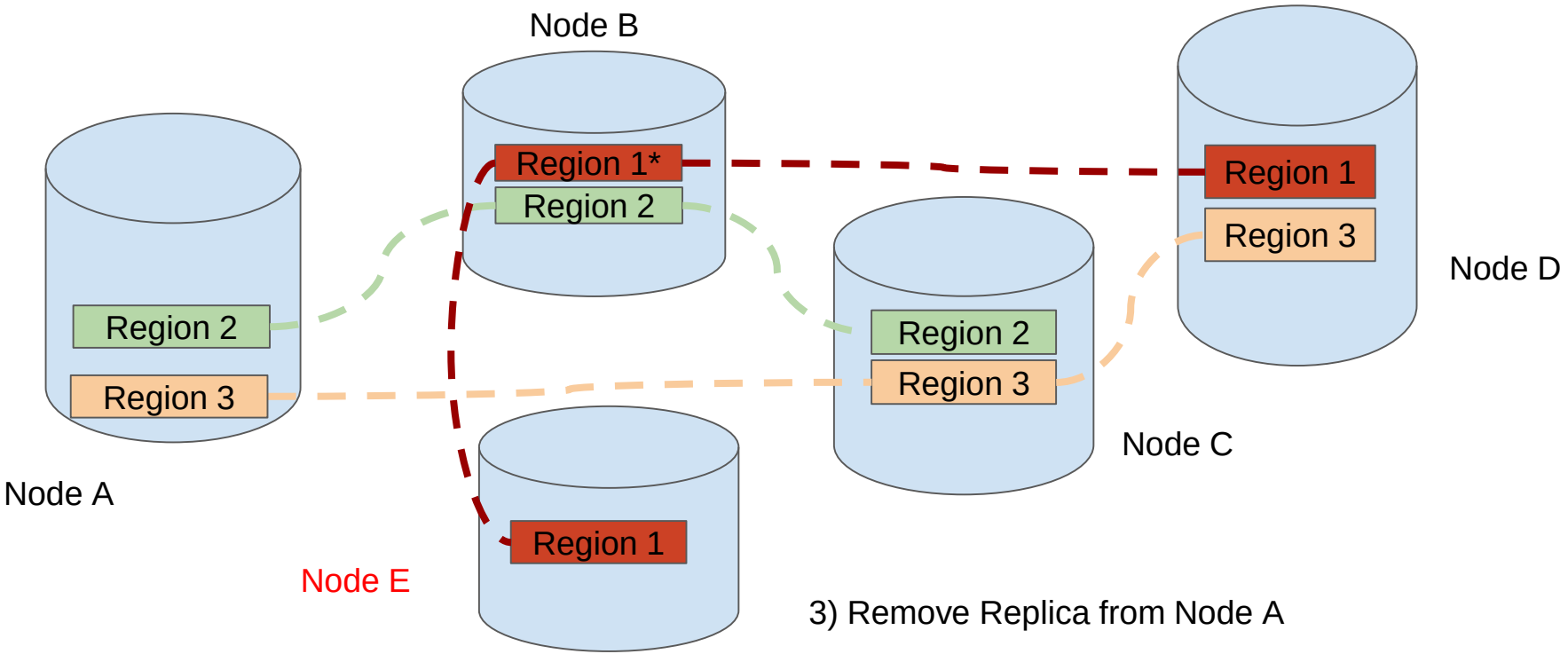
1) Transfer leadership of region 1 from Node A to Node B

PingCAP

Scale-out (balancing)



Scale-out (balancing)



PingCAP

一个分享交流的地方

微信号：eyygle

Long Press QR Code To
Identify The Concern

长按二维码识别关注

扫一扫，加入我们，分享更多知识

PingCAP

Questions?