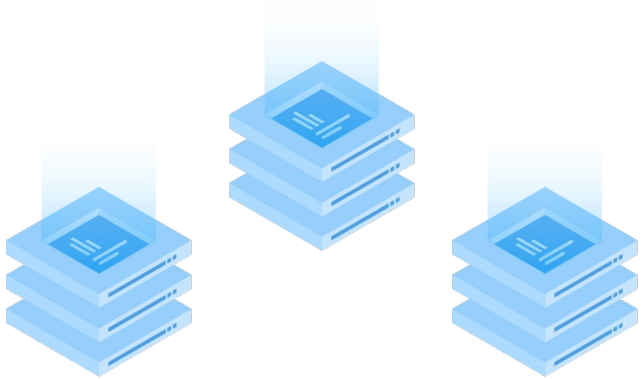
Using Rust to Build a Distributed Transactional Key-Value Database

LiuTang

PingCAP TiDB



About Me

- Chief Engineer of PingCAP
- Leader of TiKV
- Open source lovers: go-mysql, raft-rs, rust-prometheus, etc.
- tl@pingcap.com
- siddontang (Github, twitter)

Agenda

- Problems
- How do we build TiKV?
- Beyond TiKV

PingCAP

When we want to build a distributed transactional key-value database...

Problems

1. How to save the data in the machine?
2. How to support fault-tolerance?
3. How to provide ACID features?
4. How to communicate with the servers?
5. How to test the database?
6. …

PingCAP

A High Building,
A Low Foundation

PingCAP

Language

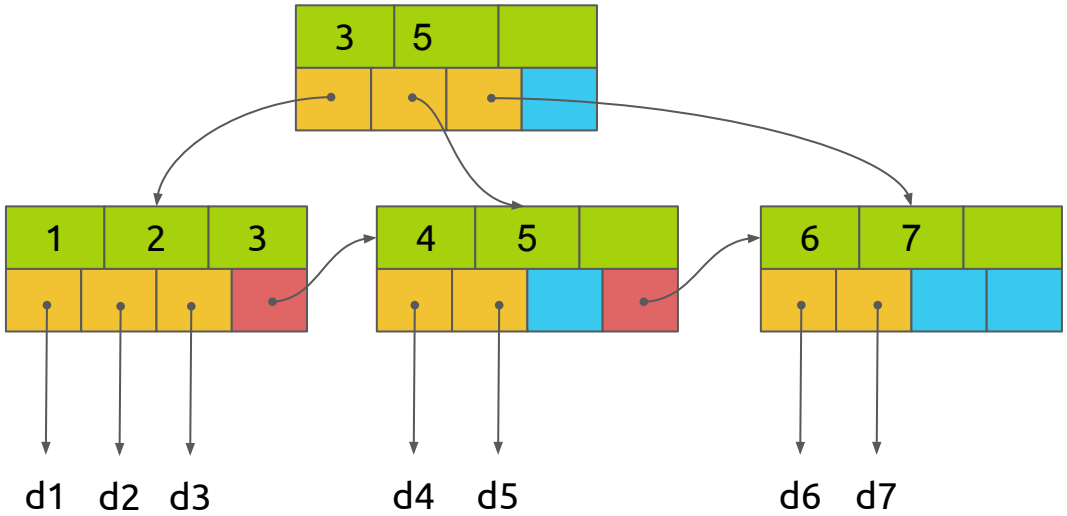
PingCAP

Let’s start from scratch!!!

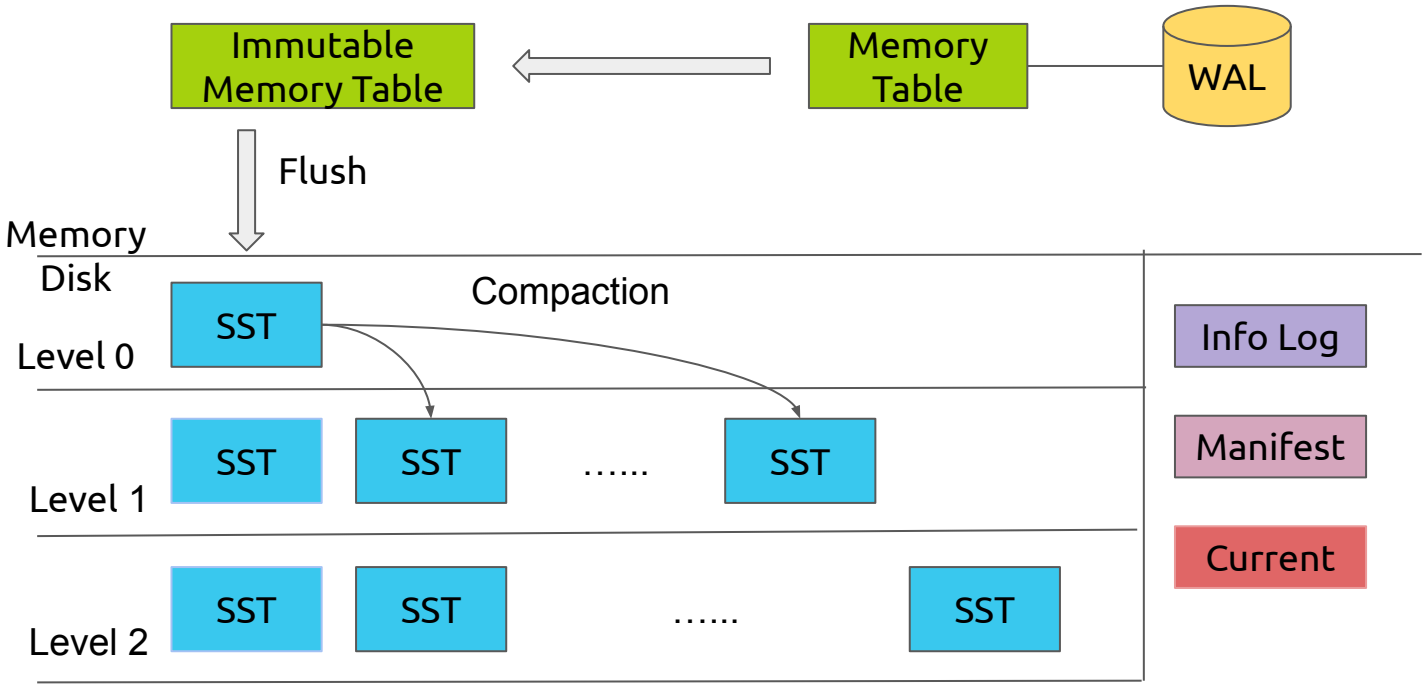
PingCAP

Storage

B+ Tree

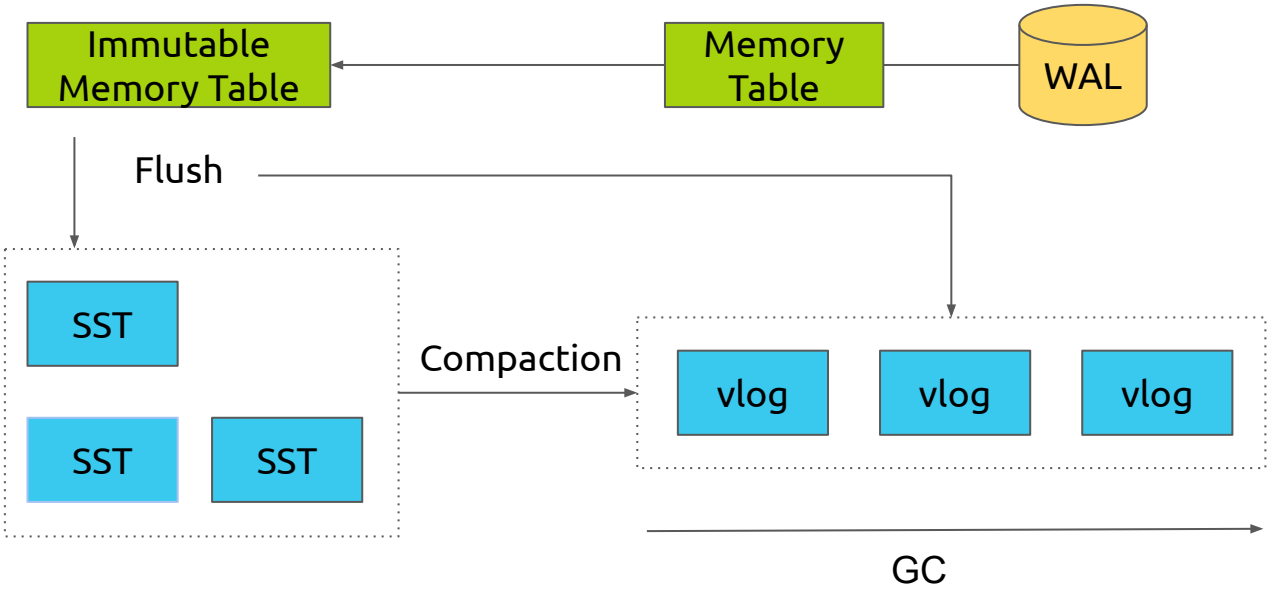


LSM


Immutable Memory Table
Memory Table
WAL
Flush
Memory
Disk
Compaction
Level 0
SST
Level 1
SST
SST
……
SST
Level 2
SST
SST
……
SST
Info Log
Manifest
Current


RocksDB

Titan


Immutable Memory Table
Memory Table
WAL
Flush
SST
SST
SST
Compaction
vlog
vlog
vlog
GC

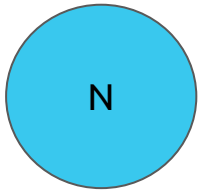
https://github.com/pingcap/rust-rocksdb

PingCAP TiDB

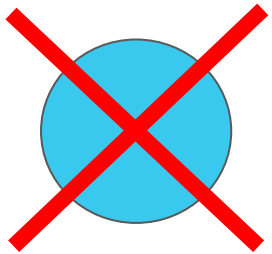
PingCAP

Raft

One Node

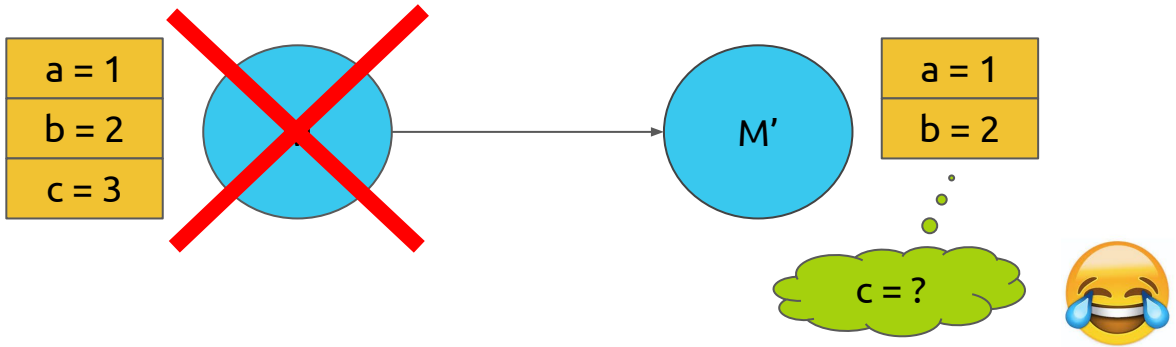


One Node



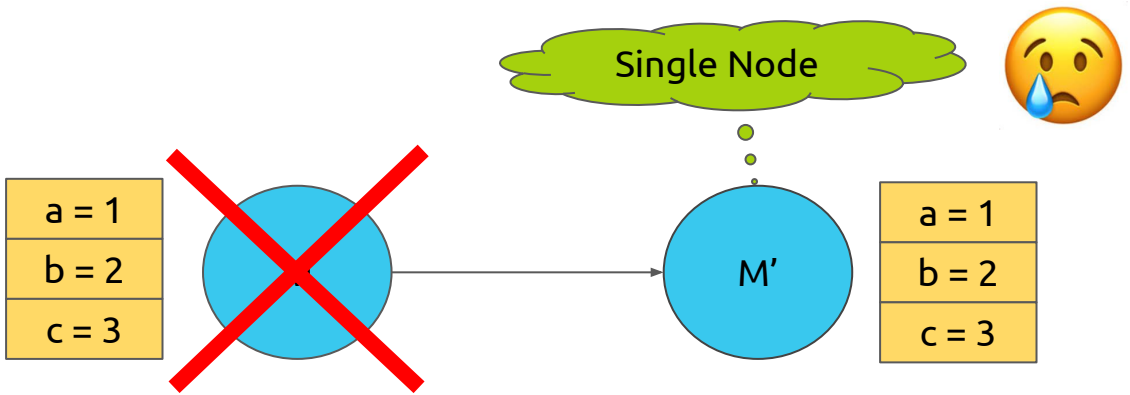
Async Replication

Async Replication

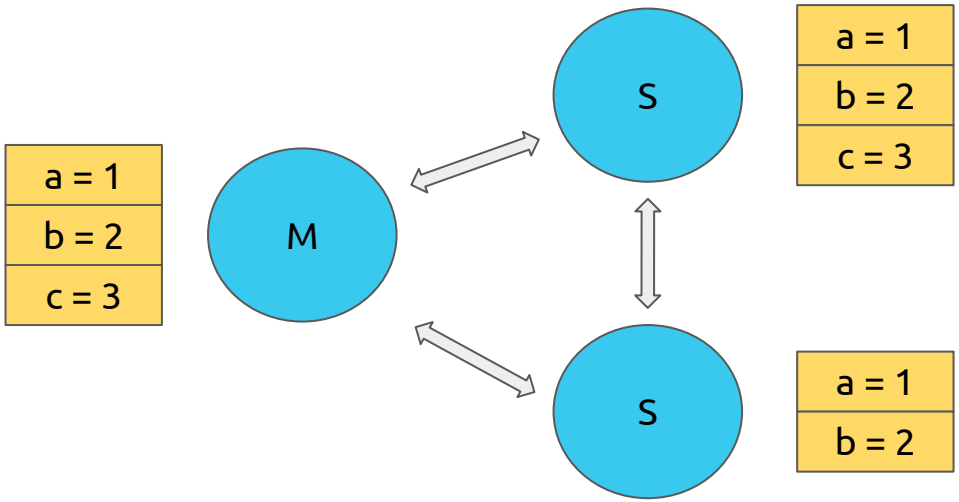


Sync Replication

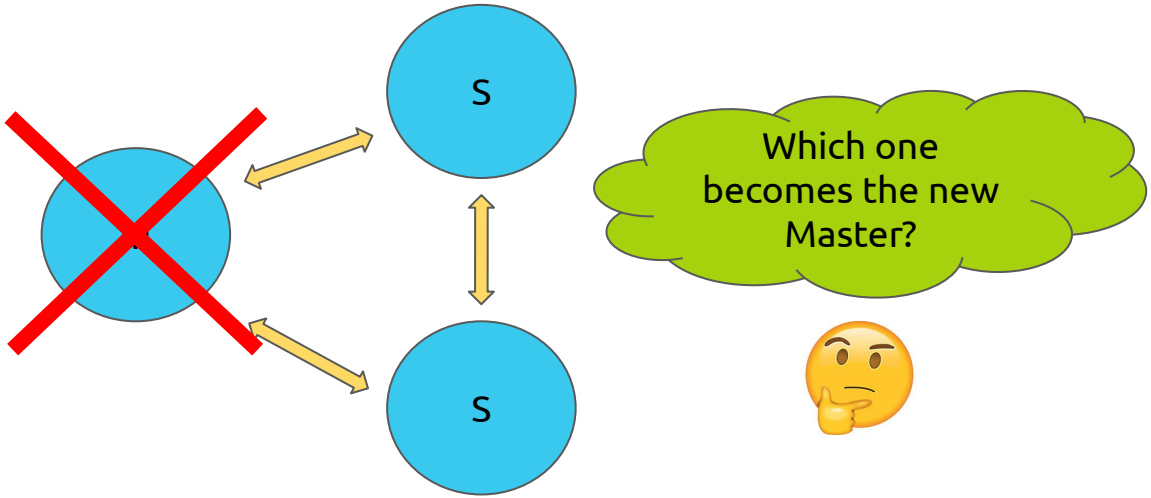
Sync Replication



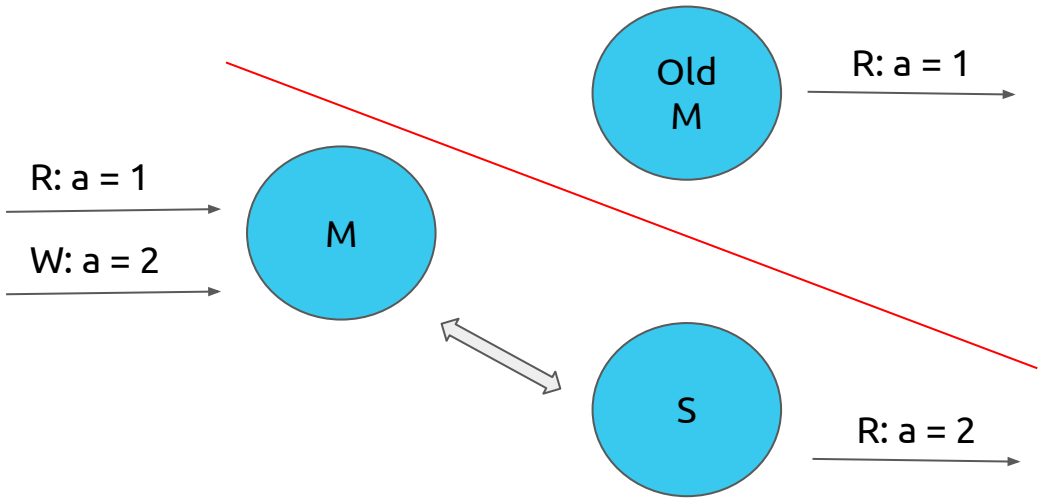
Quorum, 2 N + 1



Failover



Brain Split




raft.github.io
raft

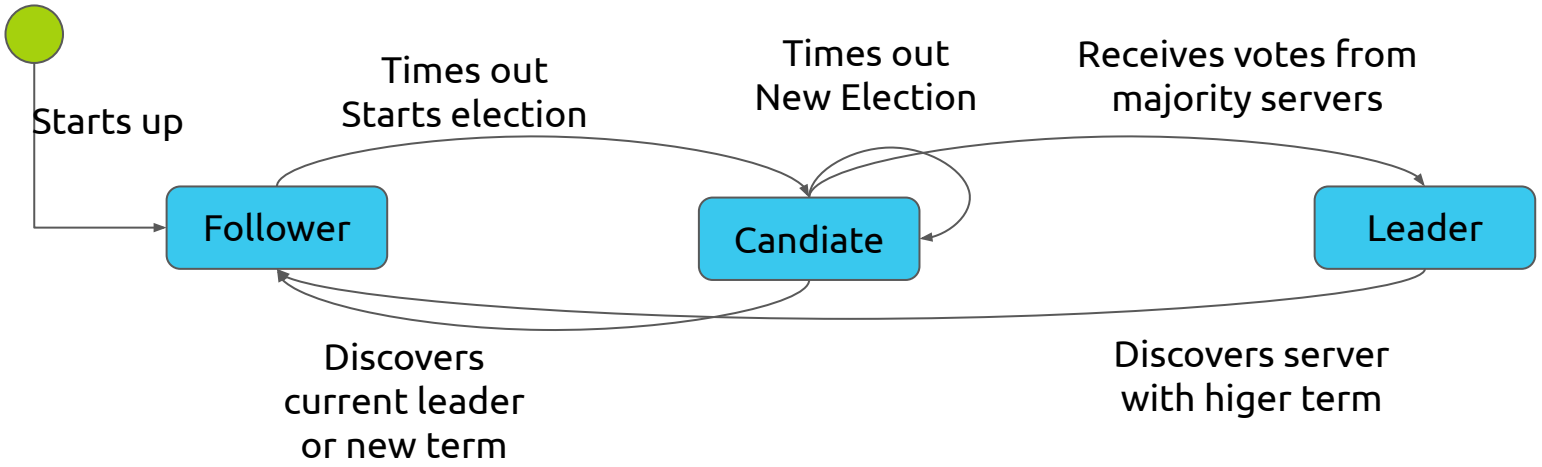
Key Points

- Election
- Replication

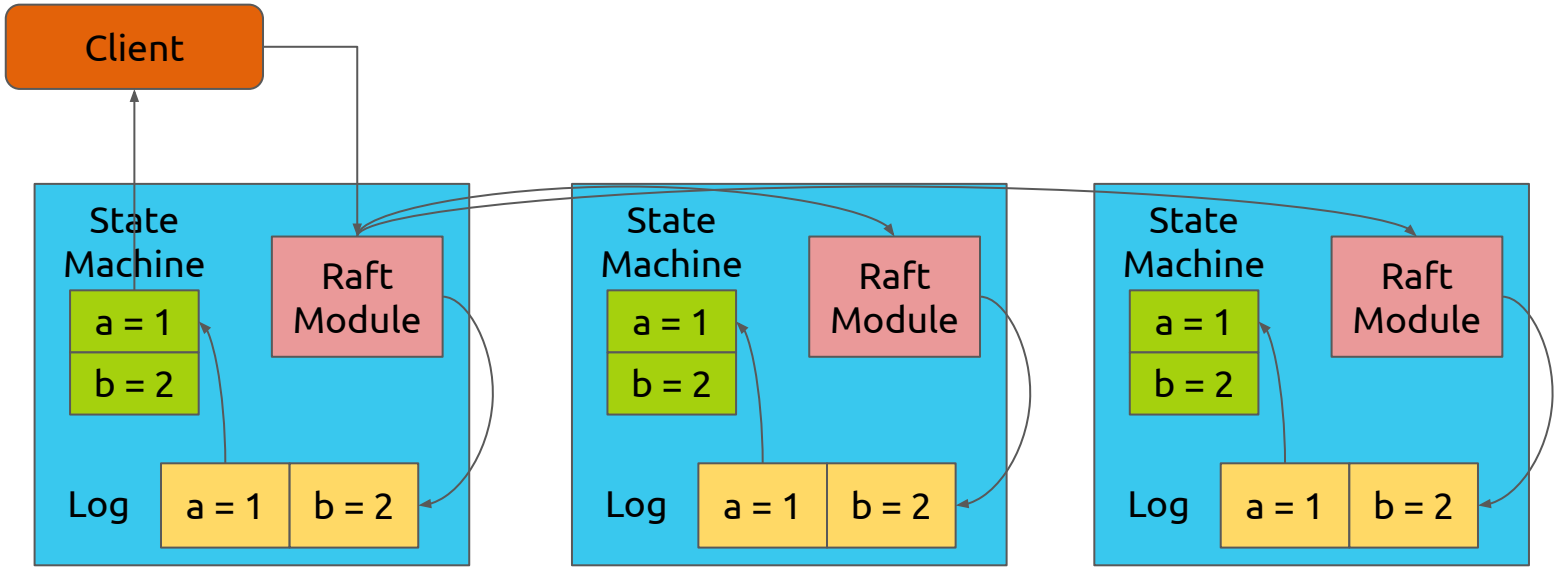
Role

- Leader
  - Only one Leader
  - Elected by the majority of the peers
  - Handles all the client requests
- Follower
  - Receives Replicated Logs from the Leader
- Candidate
  - Campaigns to become leader

Election


Starts up
Times out
Starts election
Times out
New Election
Receives votes from
majority servers
Follower
Candiate
Leader
Discovers
current leader
or new term
Discovers server
with higer term

Log Replication

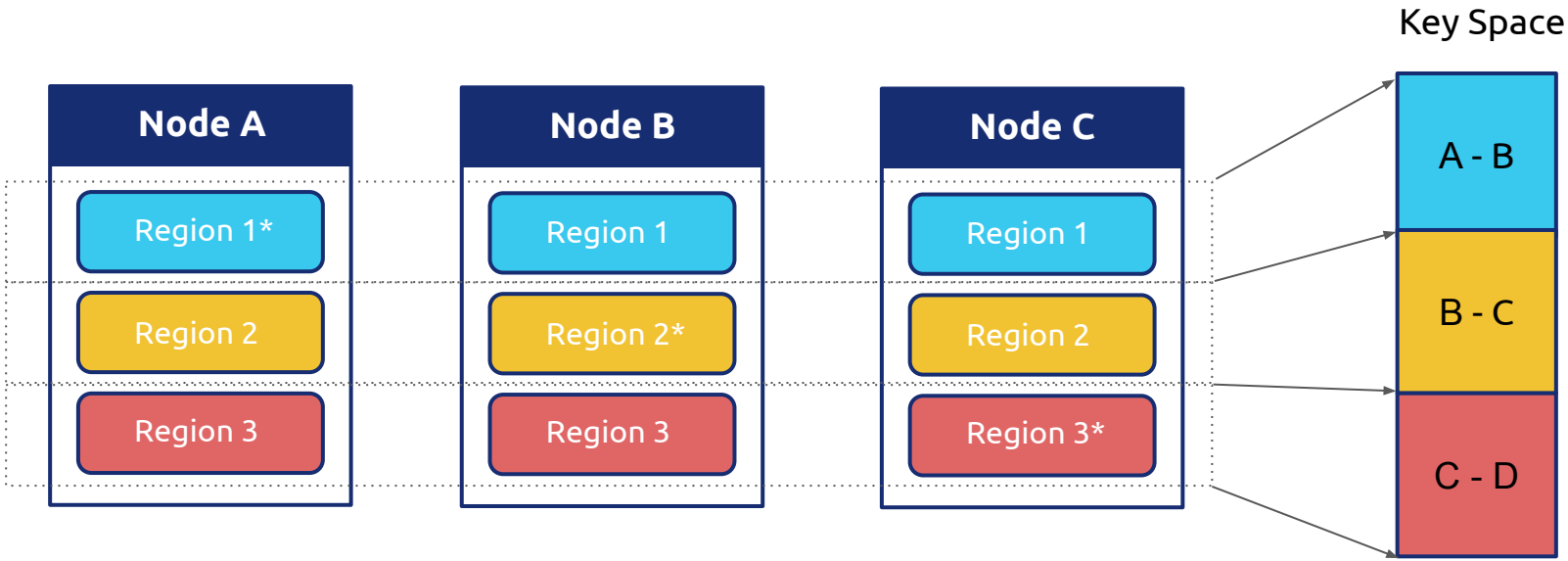


https://github.com/pingcap/raft-rs

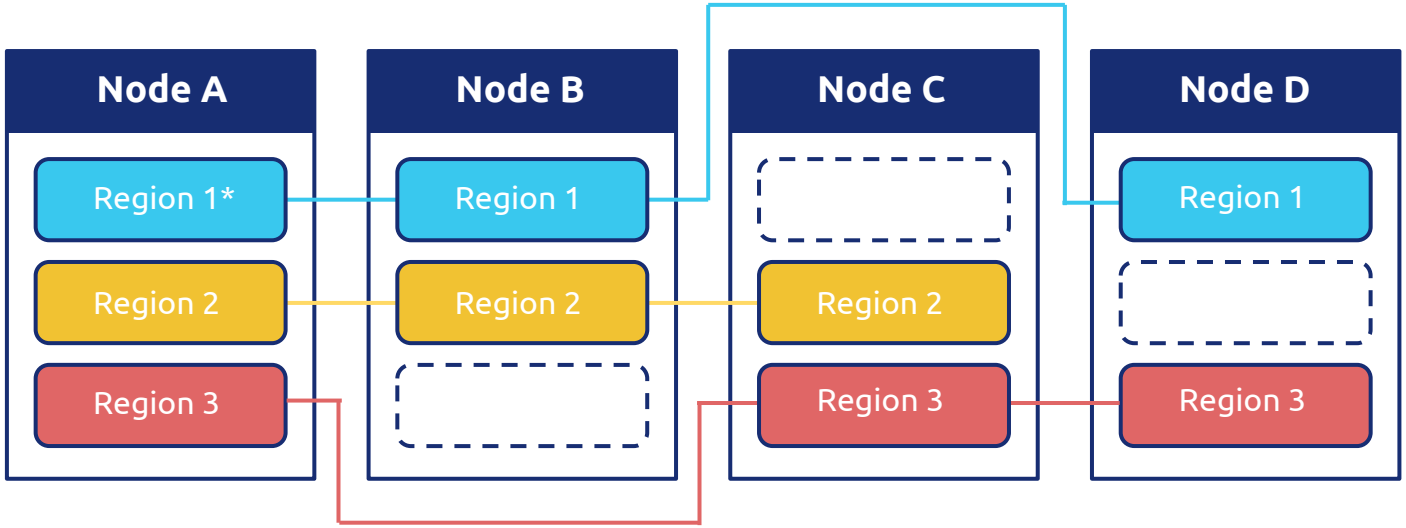
PingCAP

Multi Raft

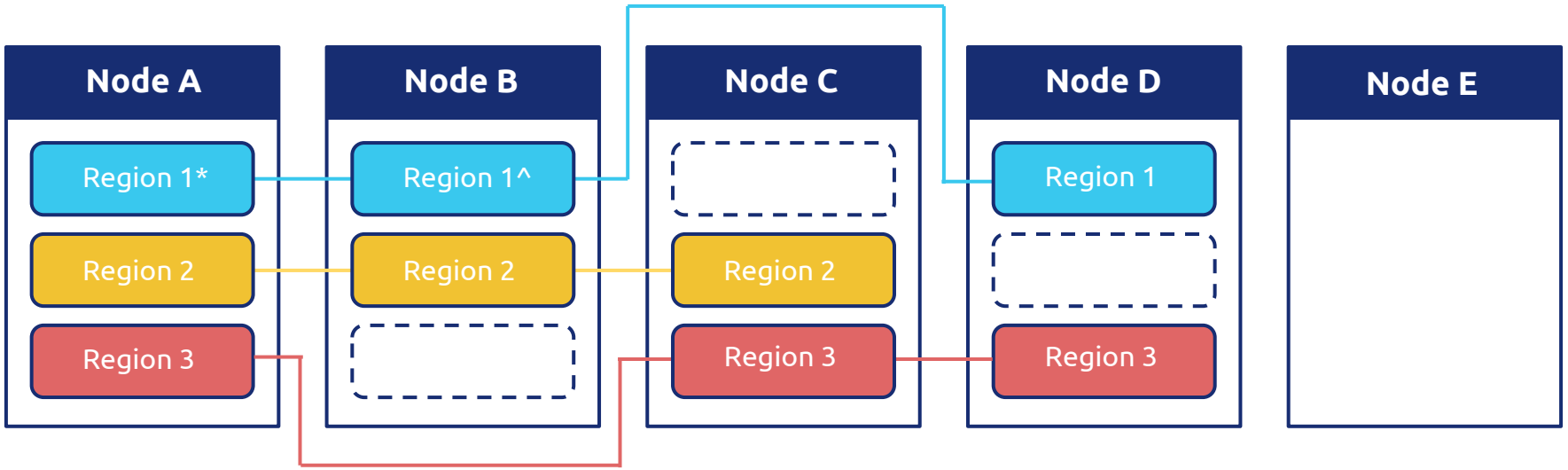
Multi-Raft


Key Space
Node A
Region 1*
Region 2
Region 3
Node B
Region 1
Region 2*
Region 3
Node C
Region 1
Region 2
Region 3*
A - B
B - C
C - D

Scale-out

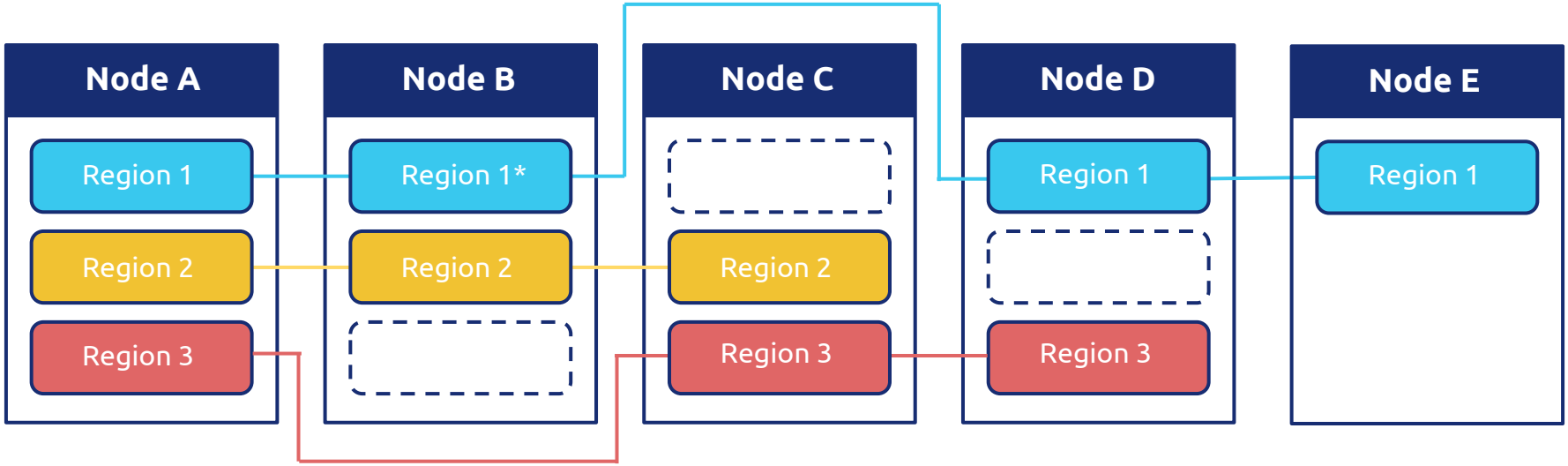

Node A
Node B
Node C
Node D
Region 1*
Region 2
Region 3
Region 1
Region 2
Region 1
Region 2
Region 3
Region 3

Scale-out



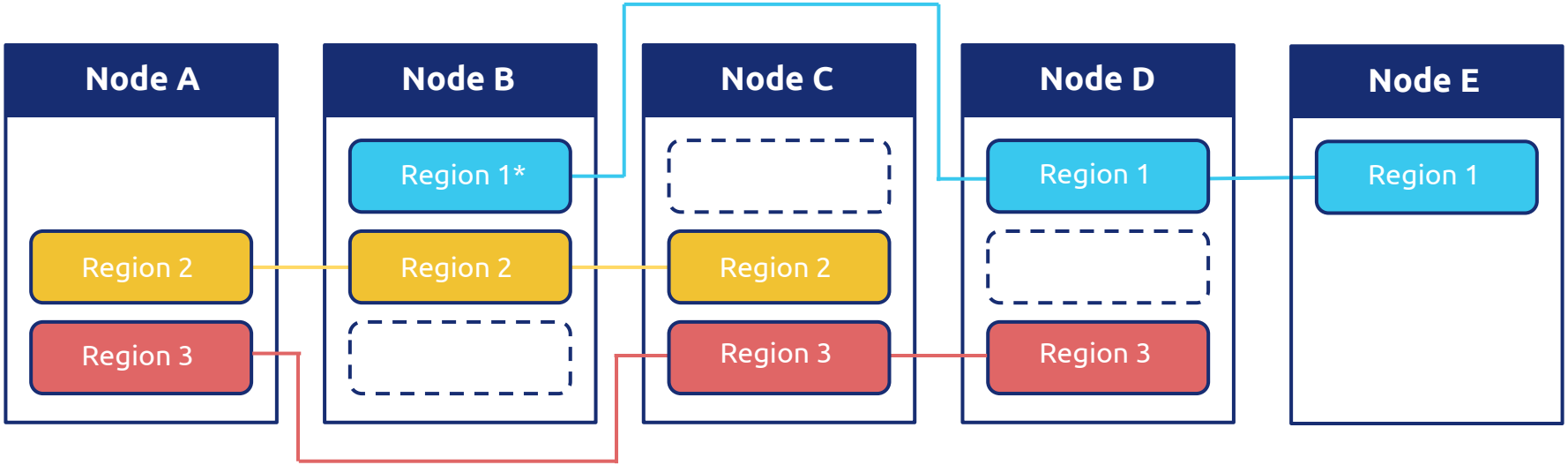
1. Transfer leadership of region 1 from Node A to Node B.

Scale-out



2. Add Replica on Node E.

Scale-out

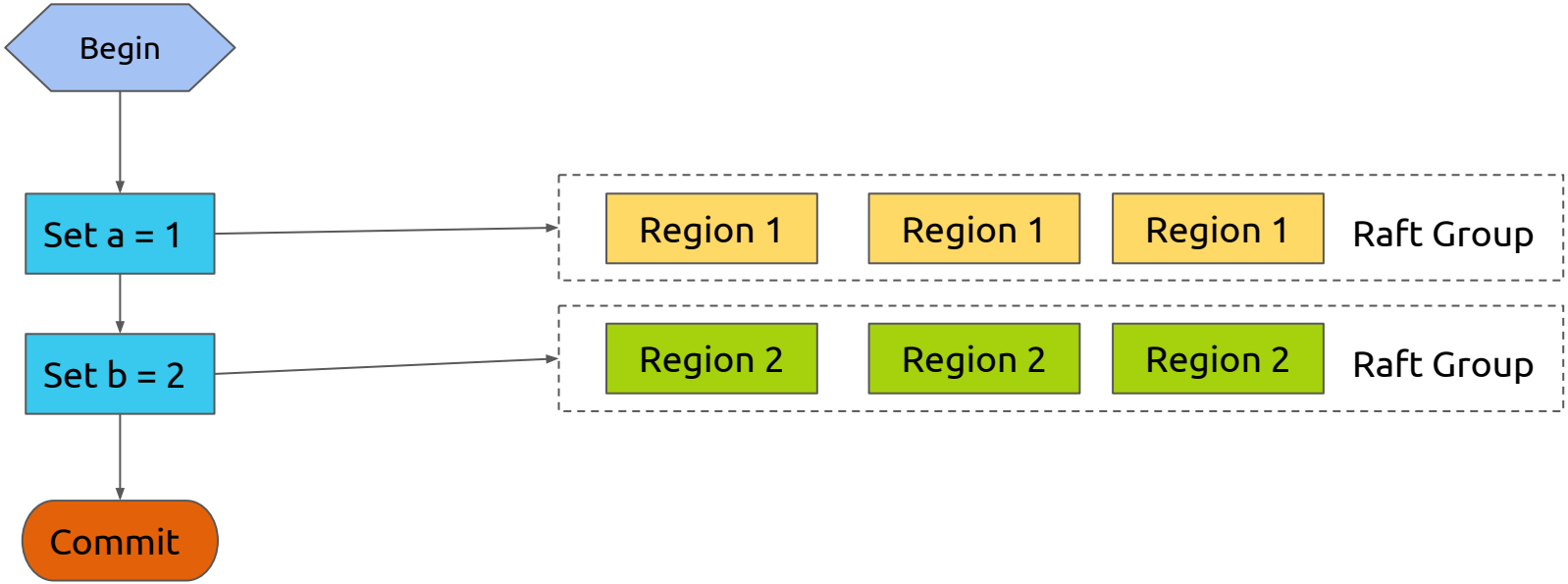


3. Remove replica from Node A.

PingCAP

Distributed Transaction

Why?


Begin
Set a = 1
Set b = 2
Commit
Region 1
Region 1
Region 1
Raft Group
Region 2
Region 2
Region 2
Raft Group

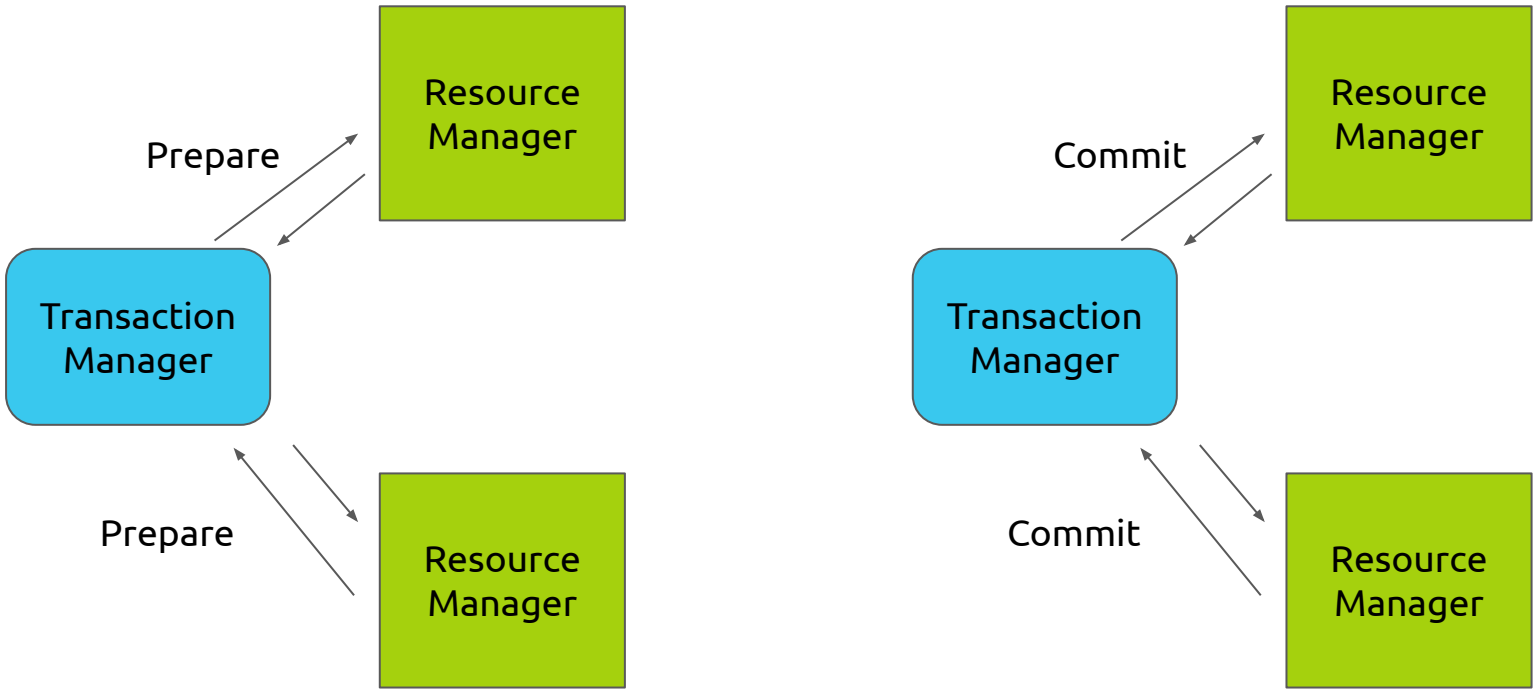
PingCAP

ACID

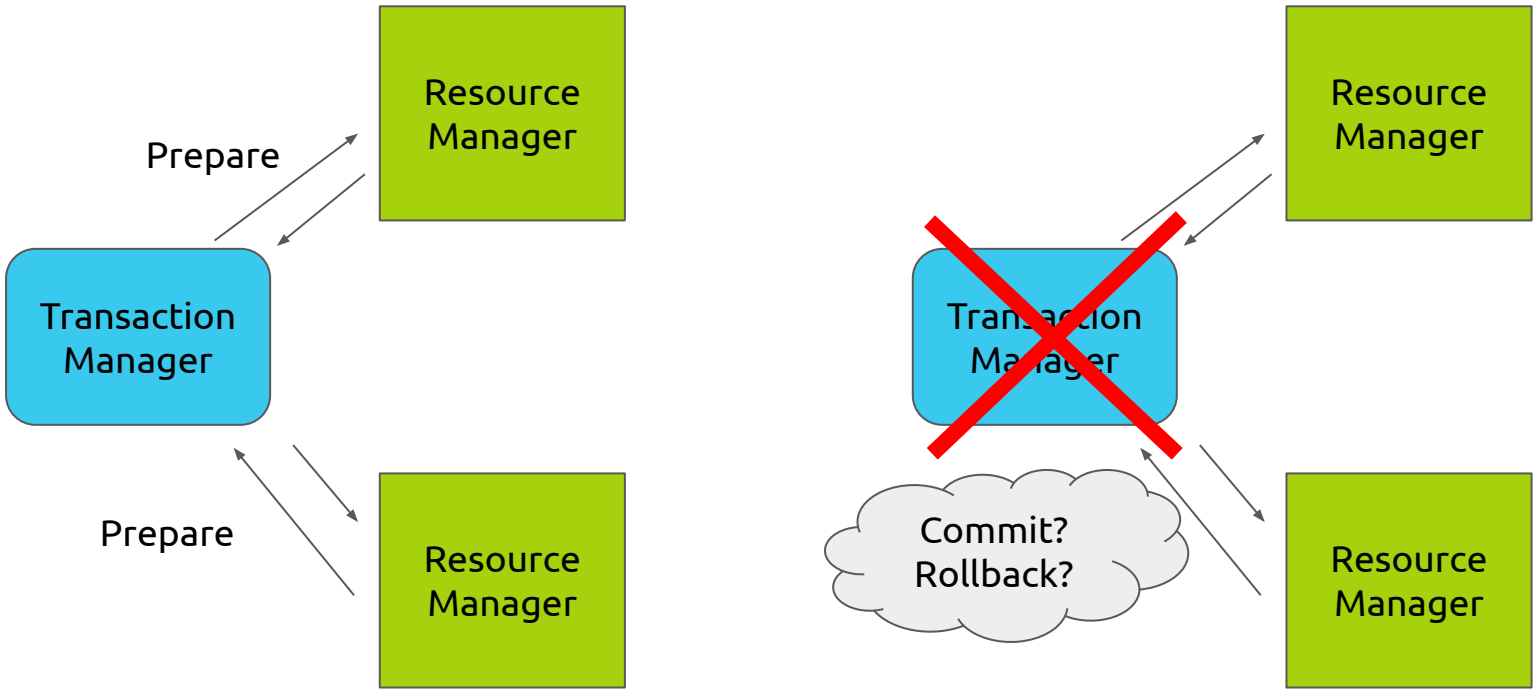
ACID

- Atomicity
- Concistency
- Isolation
- Durability

Two-Phase Commit



Two-Phase Commit



PingCAP

Communication


GRPC

gRPC

- Mode
  - Unary
  - Client streaming
  - Server streaming
  - Duplex streaming

- Using Futures to wrap the asynchronous C gRPC API

```
let f = unary(service, method, request);
let resp = f.wait();
```

https://github.com/pingcap/grpc-rs

PingCAP

Monitoring


Prometheus

Prometheus

- Type
  - Counter
  - Gauge
  - Histogram

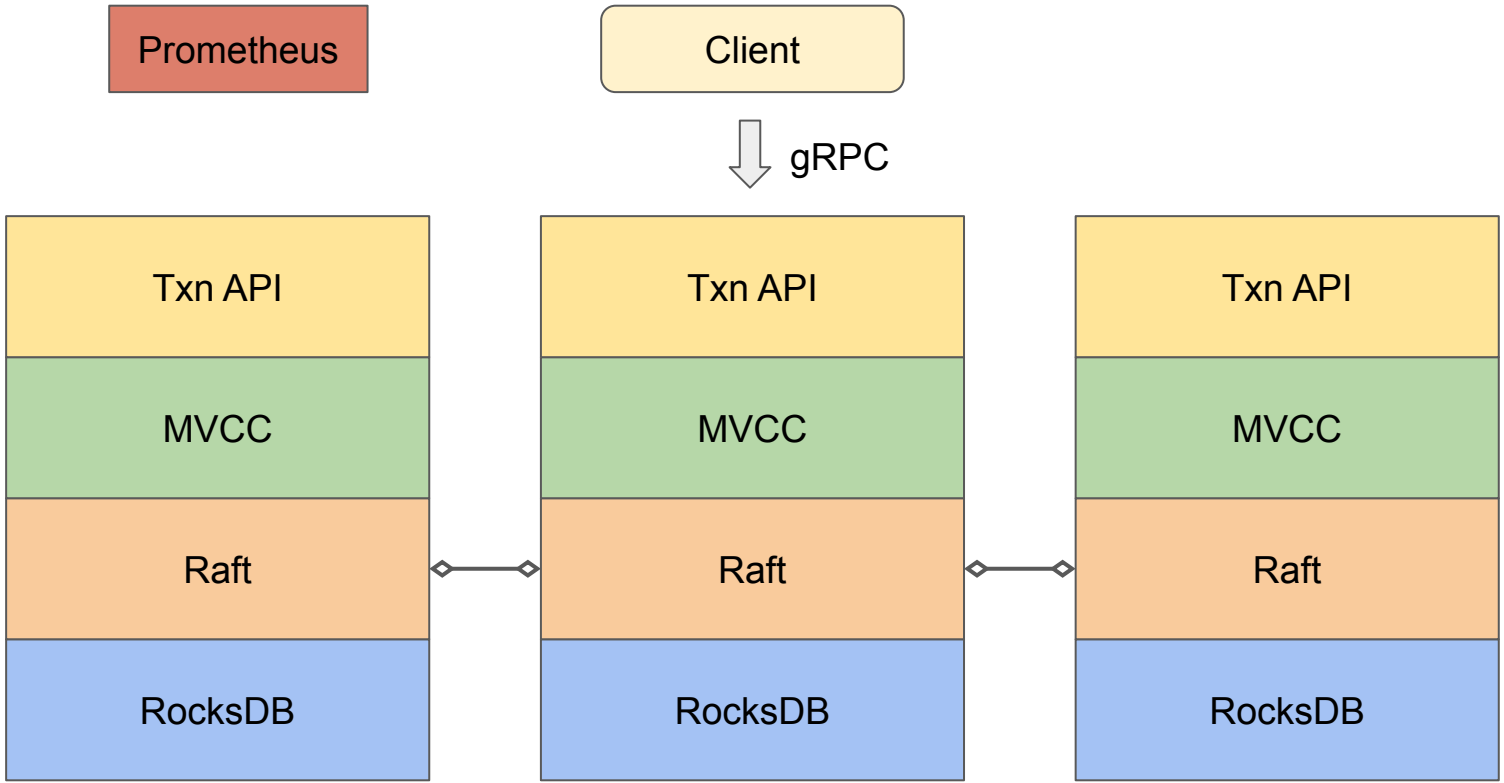
```
lazy_static! {
   static ref HTTP_COUNTER: Counter = register_counter!(
       "http_request_total",
       "Total number of HTTP request."
   ).unwrap();
}

HTTP_COUNTER.inc();
```

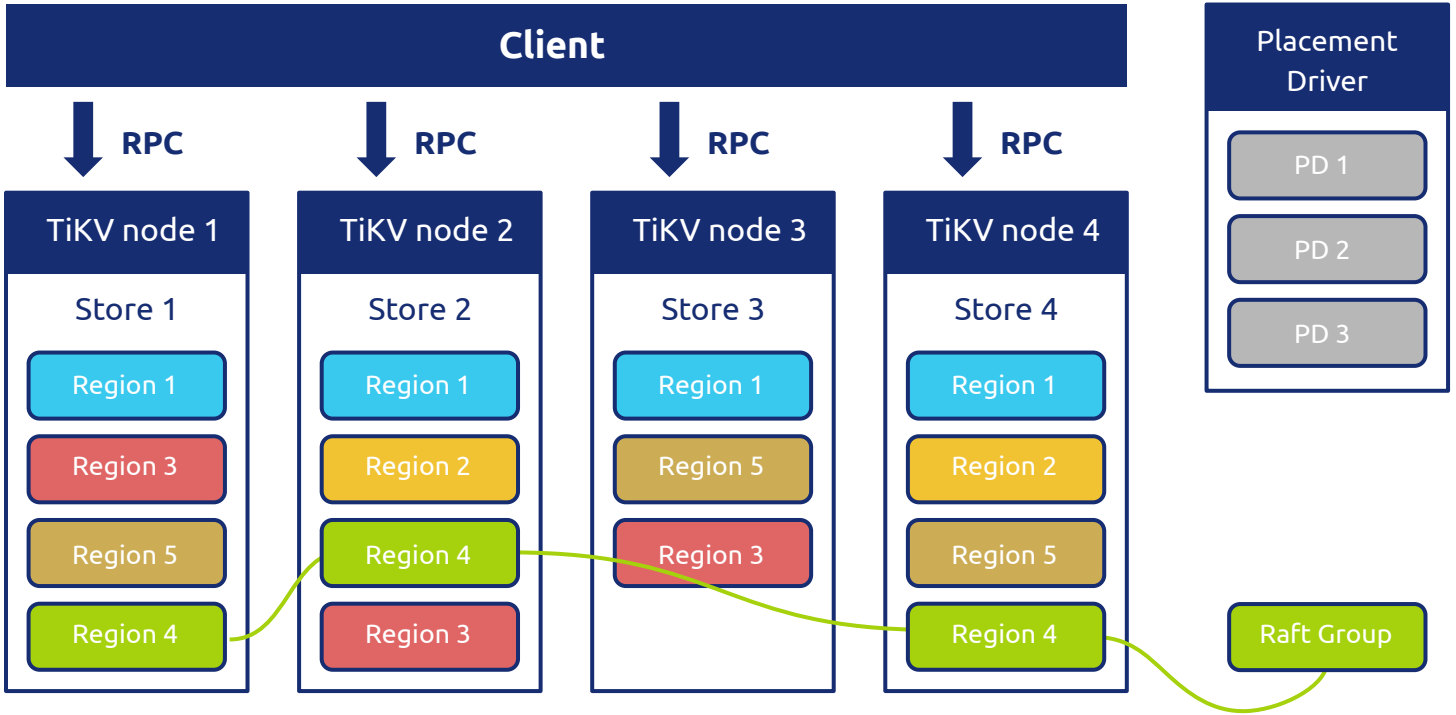
https://github.com/pingcap/rust-prometheus

PingCAP

TiKV


Prometheus
Client
gRPC
Txn API
MVCC
Raft
RocksDB
Txn API
MVCC
Raft
RocksDB
Txn API
MVCC
Raft
RocksDB

TiKV: The whole picture



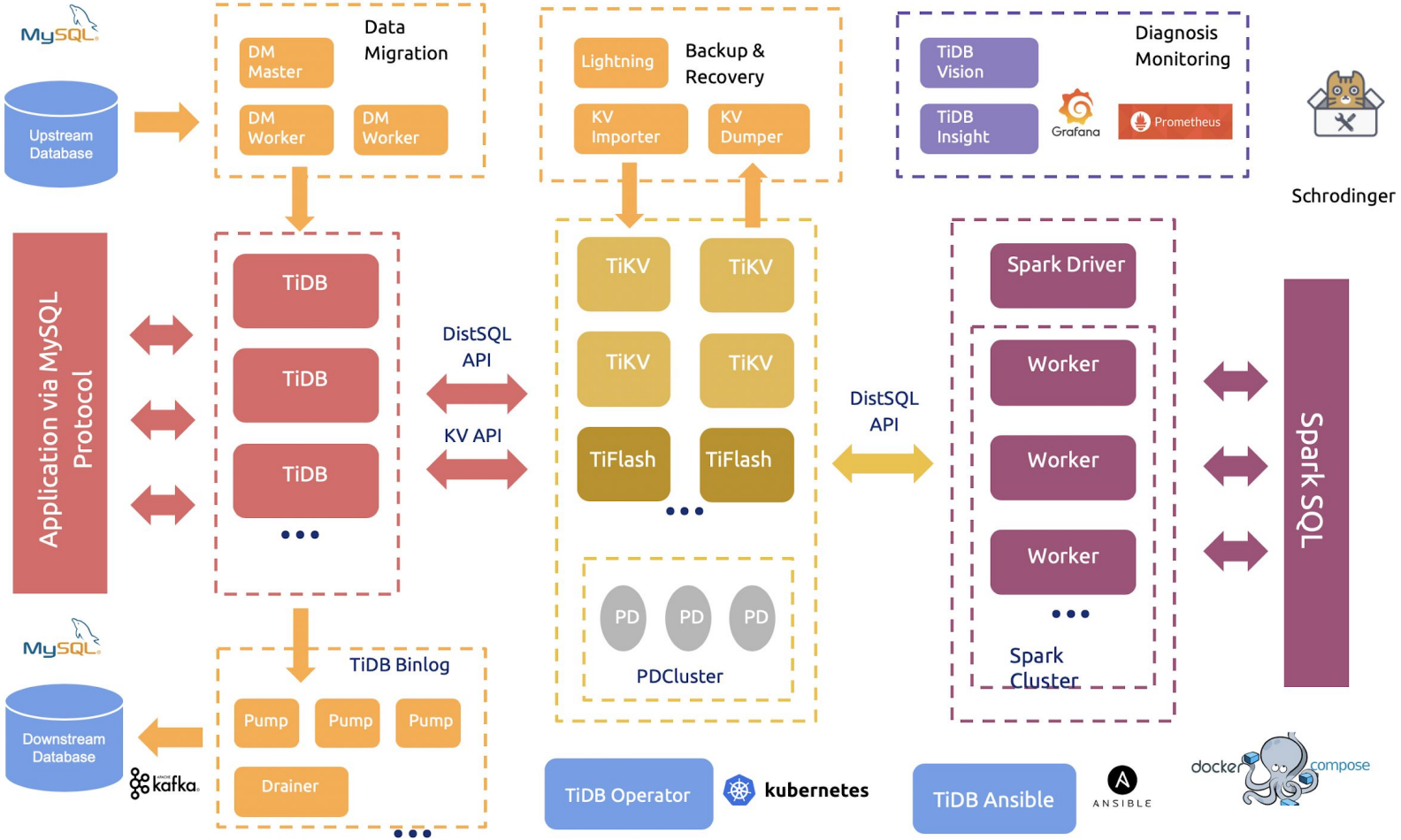

CLOUD NATIVE
COMPUTING FOUNDATION


KV

https://github.com/pingcap/tikv

PingCAP

Beyond TiKV


MySQL
Upstream Database
Data Migration
DM Master
DM Worker
DM Worker
Lightning
Backup & Recovery
KV Importer
KV Dumper
TiDB Vision
TiDB Insight
Grafana
Prometheus
Diagnosis Monitoring
Schrodinger
Application via MySQL Protocol
TiDB
TiDB
TiDB
DistSQL API
KV API
TiKV
TiKV
TiKV
TiKV
TiFlash
TiFlash
PD
PD
PD
PDCluster
DistSQL API
Spark Driver
Worker
Worker
Worker
Spark Cluster
Spark SQL
MySQL
Downstream Database
kafka
TiDB Binlog
Pump
Pump
Pump
Drainer
TiDB Operator
kubernetes
TiDB Ansible
ANSIBLE
docker compose

Thank You !

https://github.com/pingcap/tidb
https://github.com/pingcap/tikv