PingCAP

基于 Raft 构建分布式系统 TiKV

LiuTang, tl@pingcap.com

PingCAP

About me

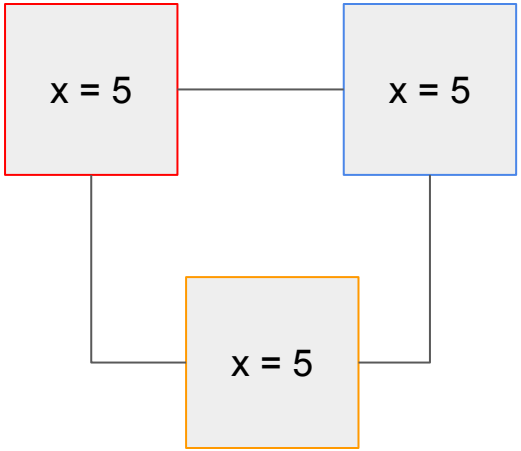
- 唐刘
- PingCAP 首席架构师
- 前金山软件资深系统工程师，分布式数据库专家，Hacker
- 知名开源软件 LedisDB / RebornDB / Mixer 作者
- 微信：siddontang
- Email: tl@pingcap.com

PingCAP

Raft

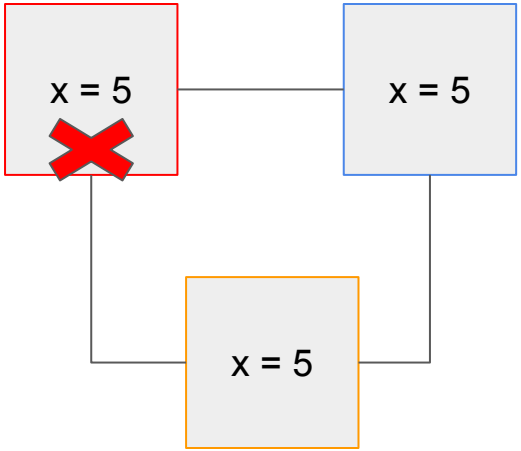
Consensus

- 多个服务决议一个值
- 如果这个值被确定，那么永远不会改变



Consensus

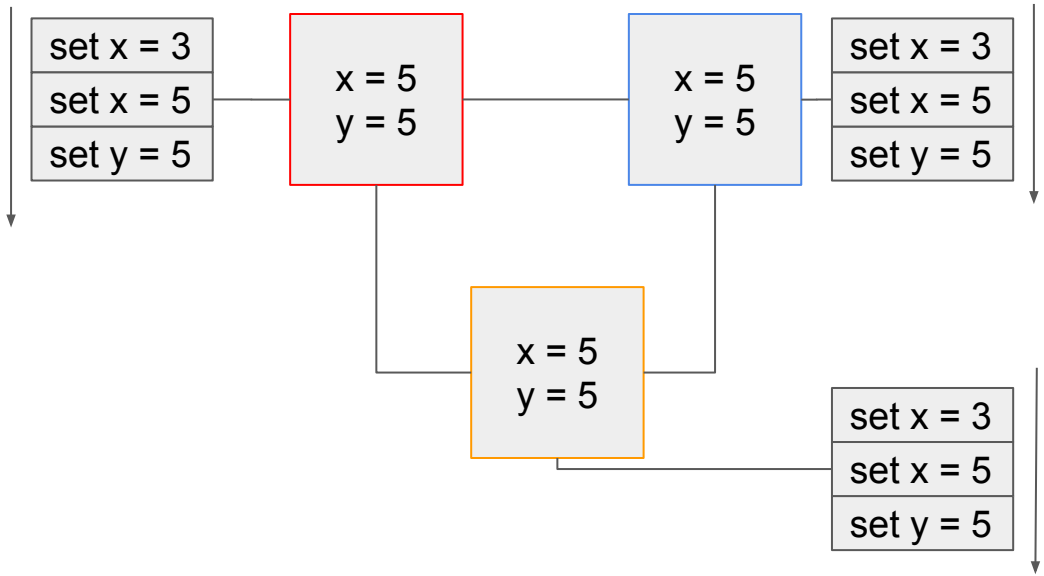
- 少数节点挂掉仍然能保证一致性
- 多数节点挂掉停止服务



PingCAP

Consensus

- 状态机
- 日志



PingCAP

关键点

- 节点的三个状态
- Term 和选举
- Log Replication
- Configuration Change

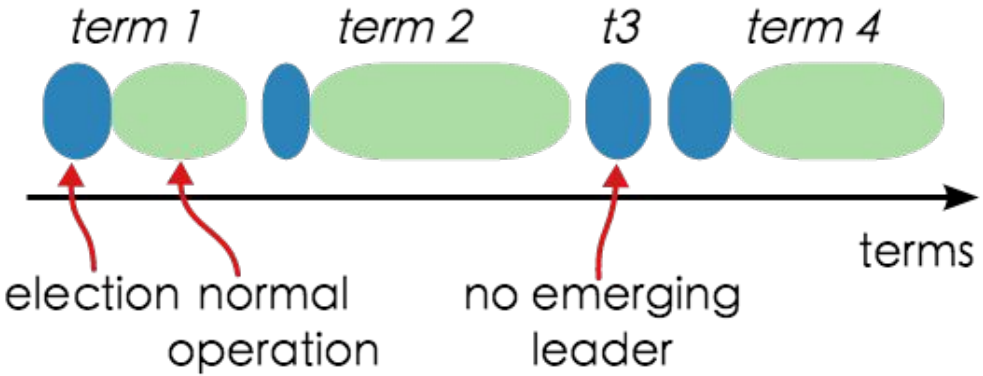
PingCAP

Role

- **Leader**
  - 集群大多数节点投票产生
  - 一个 Raft Group 只可能有一个 Leader 对外提供读写服务
- **Follower**
  - 只负责接受 Log，并 apply 已经 Committed 的 Log
- **Candidate**
  - Follower 长时间没收到 Leader 的消息，变成 Candidate 重新开始选举，Vote 新的 Leader

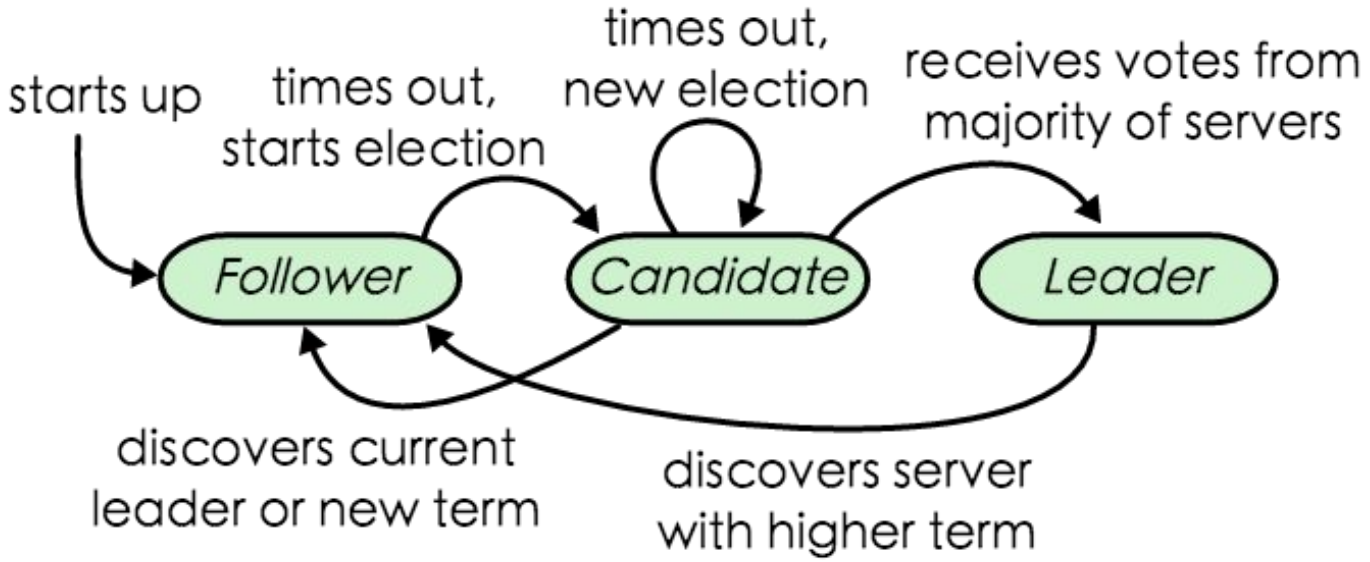
Term (任期)

- 任意时间长度
- 连续单调递增，类似 Logic Clock
- 每一个新的 Term 都开始于一次新的选举
- 如果 Candidate 成为 Leader，那么在剩下的 Term 时间，都会作为 Leader 服务
- 如果在 Term 里面，没有 Leader 选出，则下一个 Term 继续开始选举



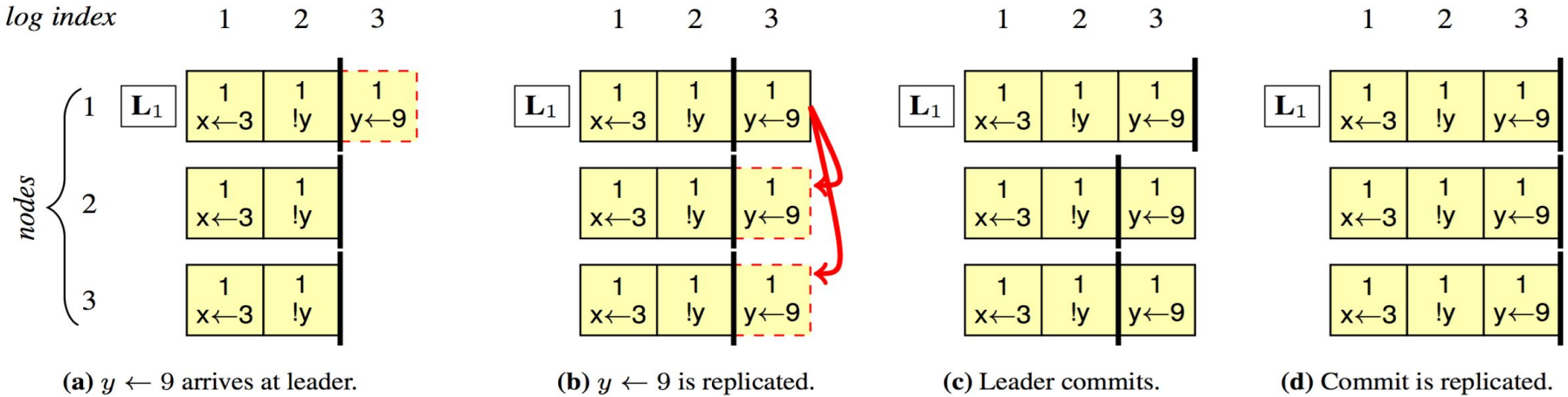
PingCAP

Election



Log Replication

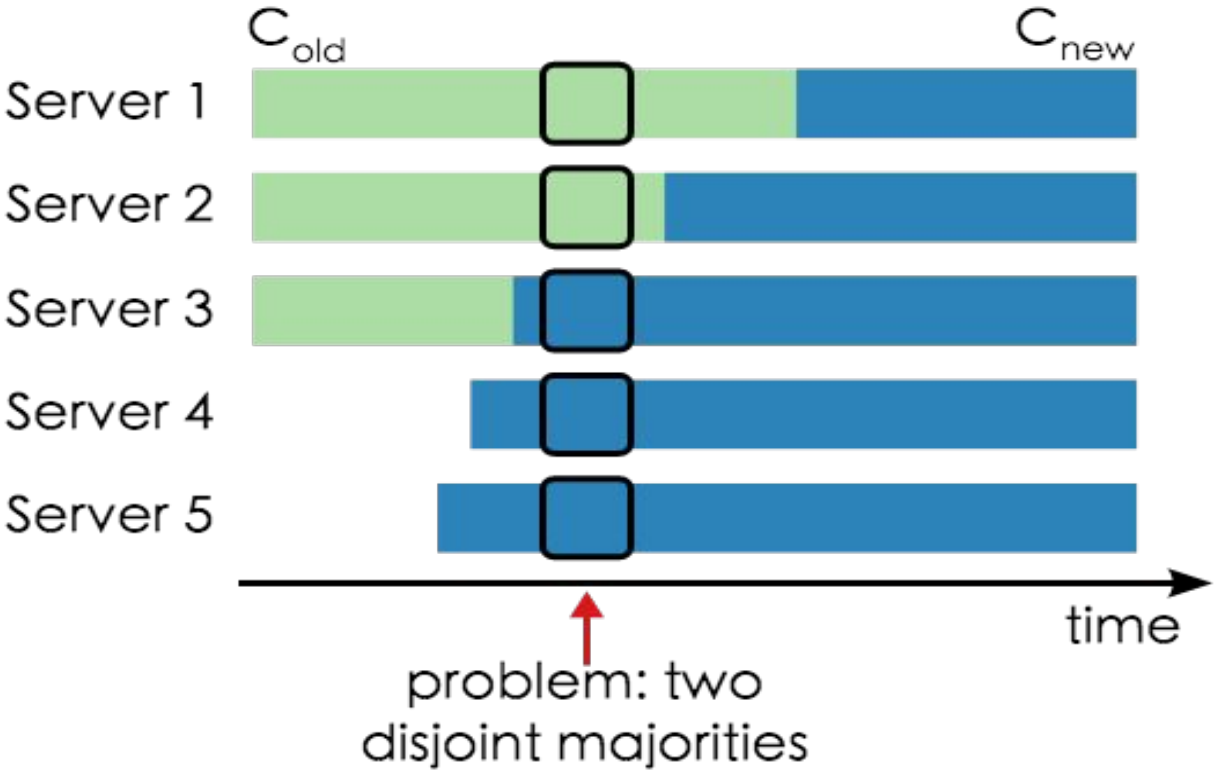
- Leader 处理 Client 的 Command
- Leader 负责将 Log 复制到其他节点
- 如果过半数节点接受到 Log，该 Log 就是 Committed
- Leader 和 Followers 各自去 Apply Log



(a) $y \leftarrow 9$ arrives at leader. **(b)** $y \leftarrow 9$ is replicated. **(c)** Leader commits. **(d)** Commit is replicated.

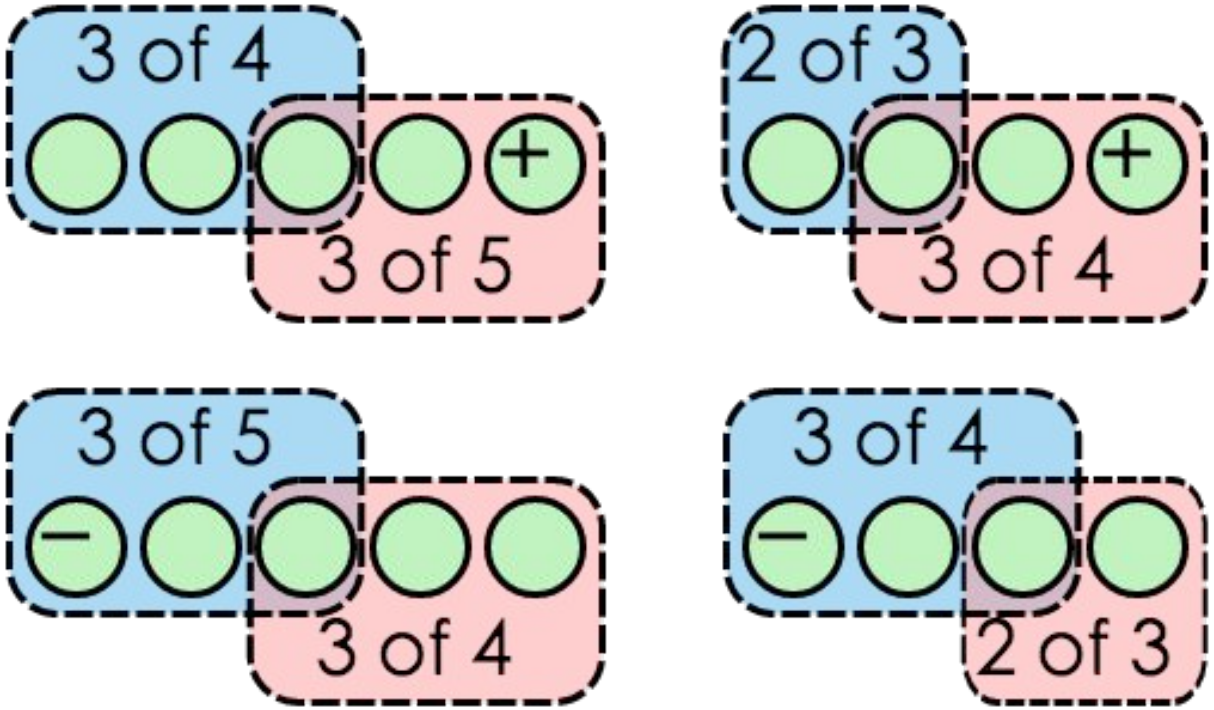
PingCAP

Configuration Change: Disjoint Problem



PingCAP

Configuration Change: Add/Remove One Node



PingCAP

TiKV

PingCAP

TiKV

- Distributed Key-Value Store
- 基于 Raft，强一致，高可用
- Auto Balance
- ACID 事务
- 使用 Rust 编写
- https://github.com/pingcap/tikv

TiKV - Raft 官方排名前三的实现

PingCAP

The best place to ask questions about Raft and its implementations is the raft-dev Google group. Some of the implementations also have their own mailing lists; check their READMEs.

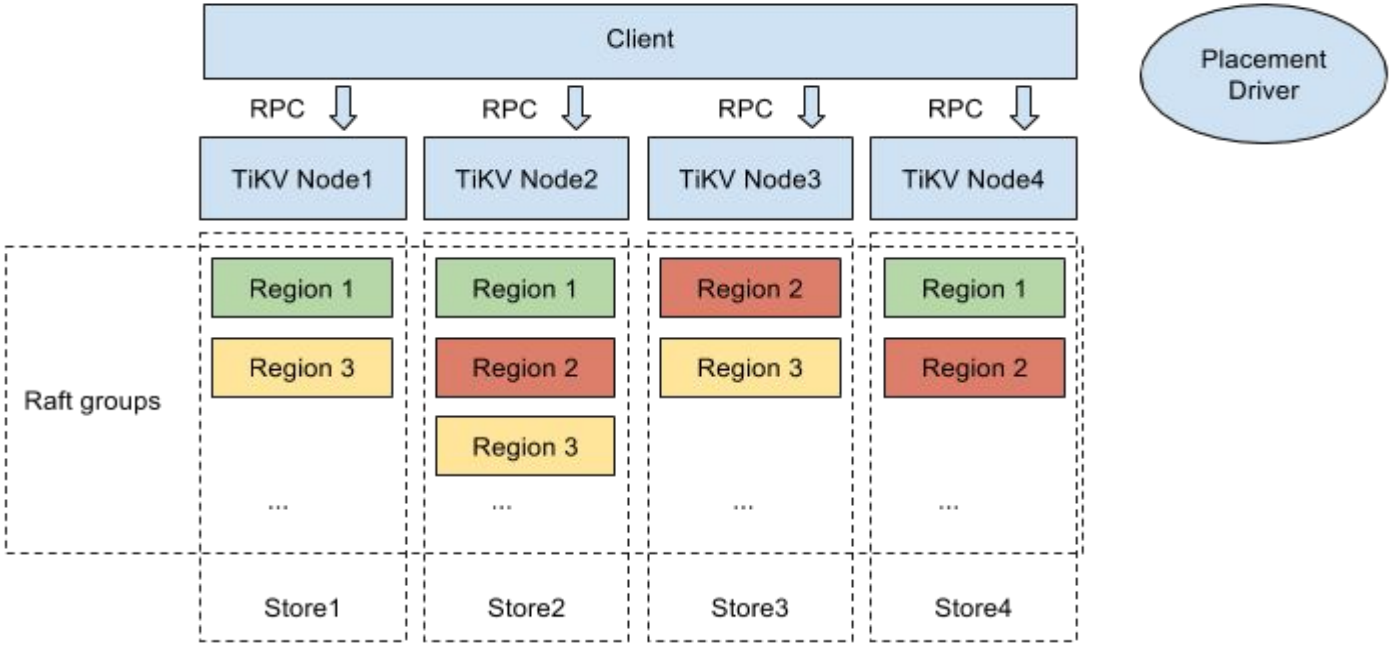
Where can I get Raft?

There are many implementations of Raft available in various stages of development. This table lists the implementations we know about with source code available. The most popular and/or recently updated implementations are towards the top. This information will inevitably get out of date; please submit a pull request or an issue to update it.

Name	Primary Authors	Language	License	Leader Election + Log Replication?	Membership Changes?	Log Compaction?	Row Last Updated
RethinkDB/clustering		C++	AGPL	Yes	Yes	Yes	2015-09-15
etcd/raft	Blake Mizerany, Xiang Li and Yicheng Qin	Go	Apache 2.0	Yes	Yes	Yes	2014-10-27
TiKV	Jay, ngaut, siddontang, tiancaiamao.	Rust	Apache2	Yes	Yes	Yes	2016-06-02
go-raft	Ben Johnson (Sky) and Xiang Li (CMU, CoreOS)	Go	MIT	Yes	Partial?	Yes	2013-07-05
hashicorp/raft	Armon Dadgar (hashicorp)	Go	MPL-2.0	Yes	Yes	Yes	2014-04-21
verdi/raft	James Wilcox, Doug Woos, Pavel Panchekha, Zach Tatlock, Xi Wang, Mike Ernst, and Tom Anderson (University of Washington)	Coq	BSD	Yes	No	No	2015-09-15

PingCAP

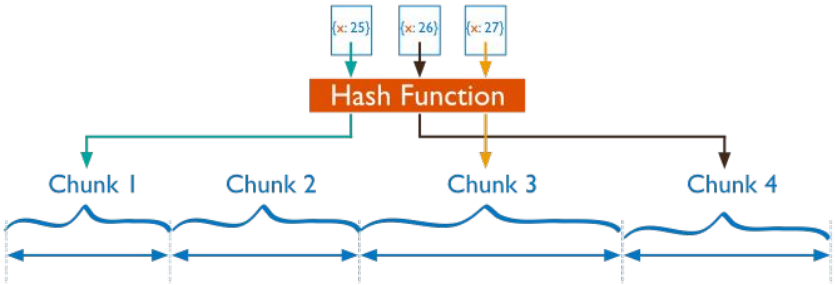
Software Stack



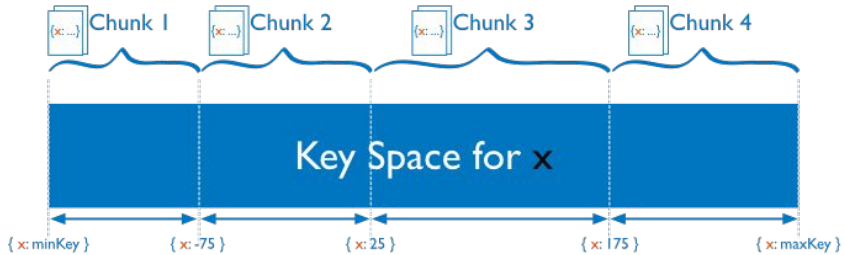
PingCAP

Shard

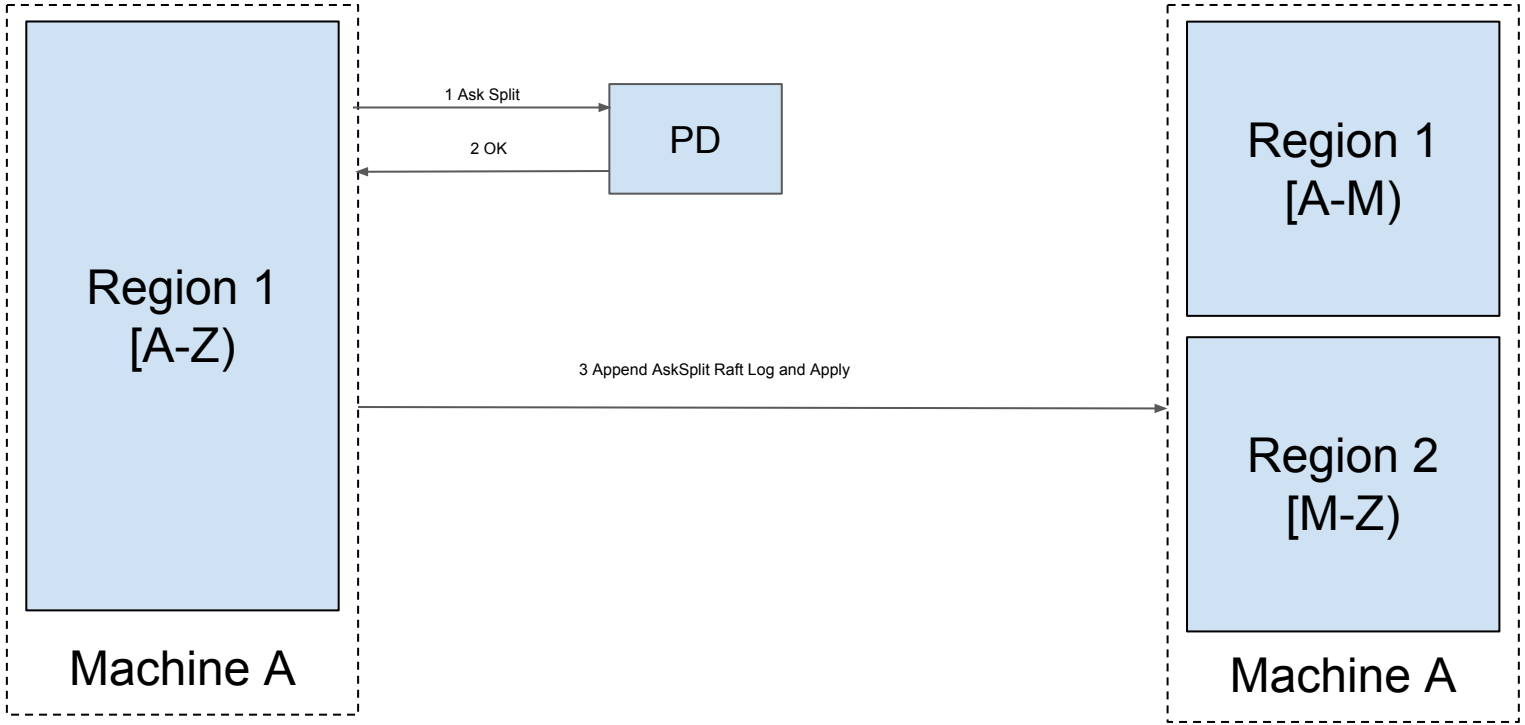
Hash



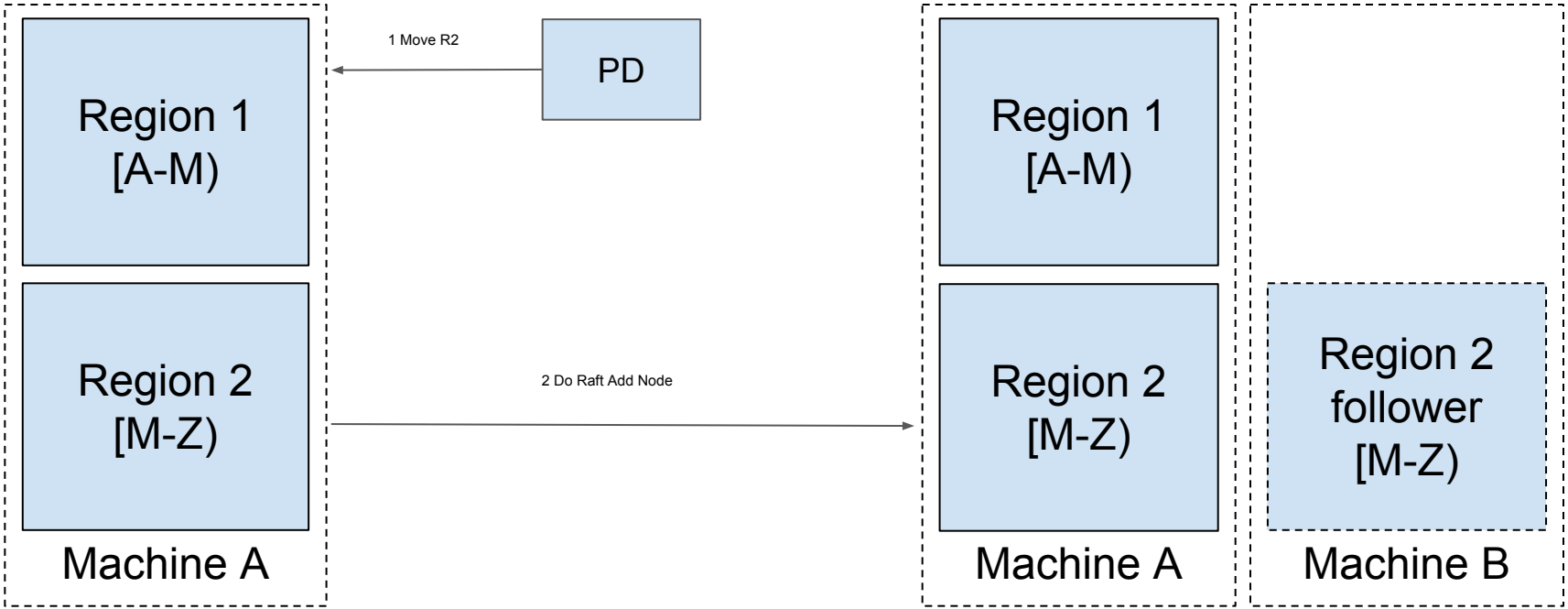
Range



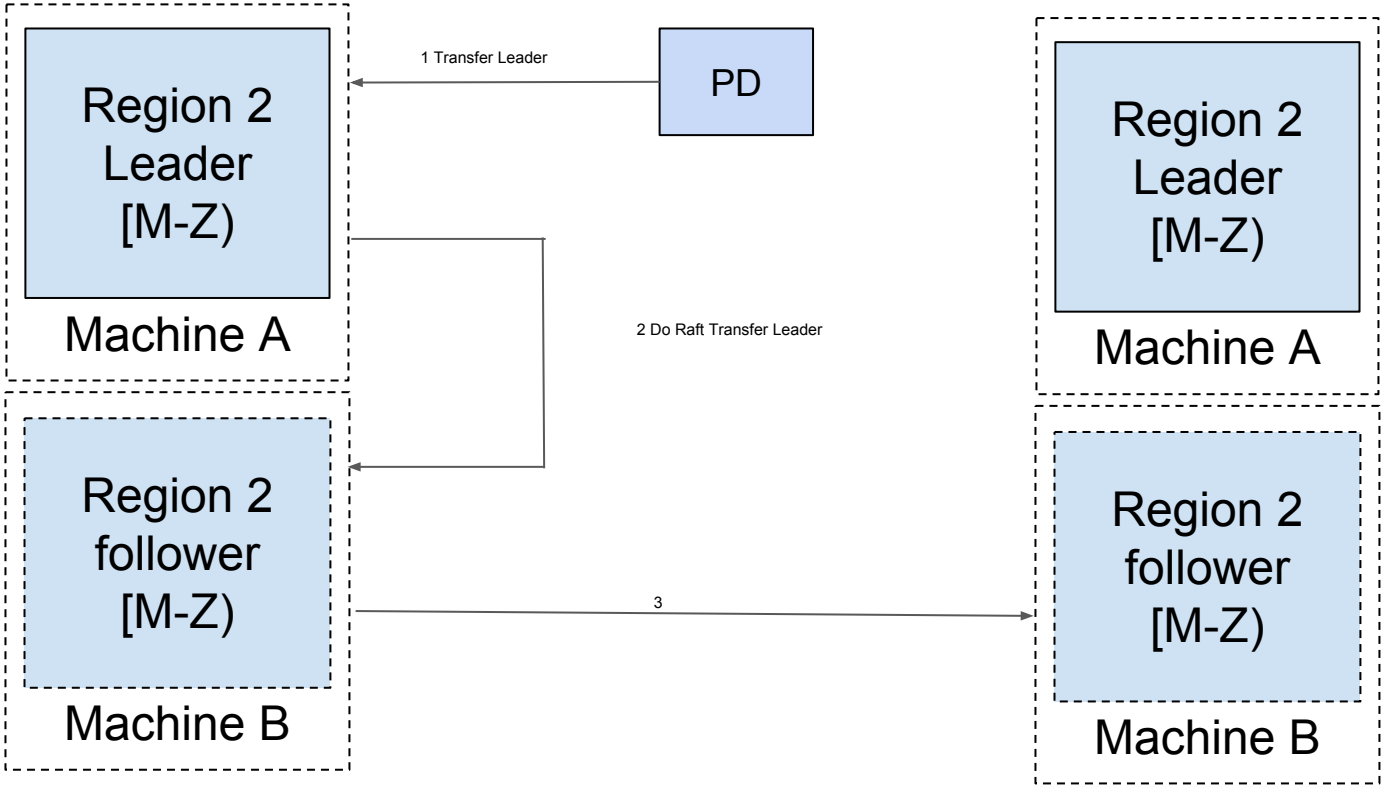
Auto Balance: Split



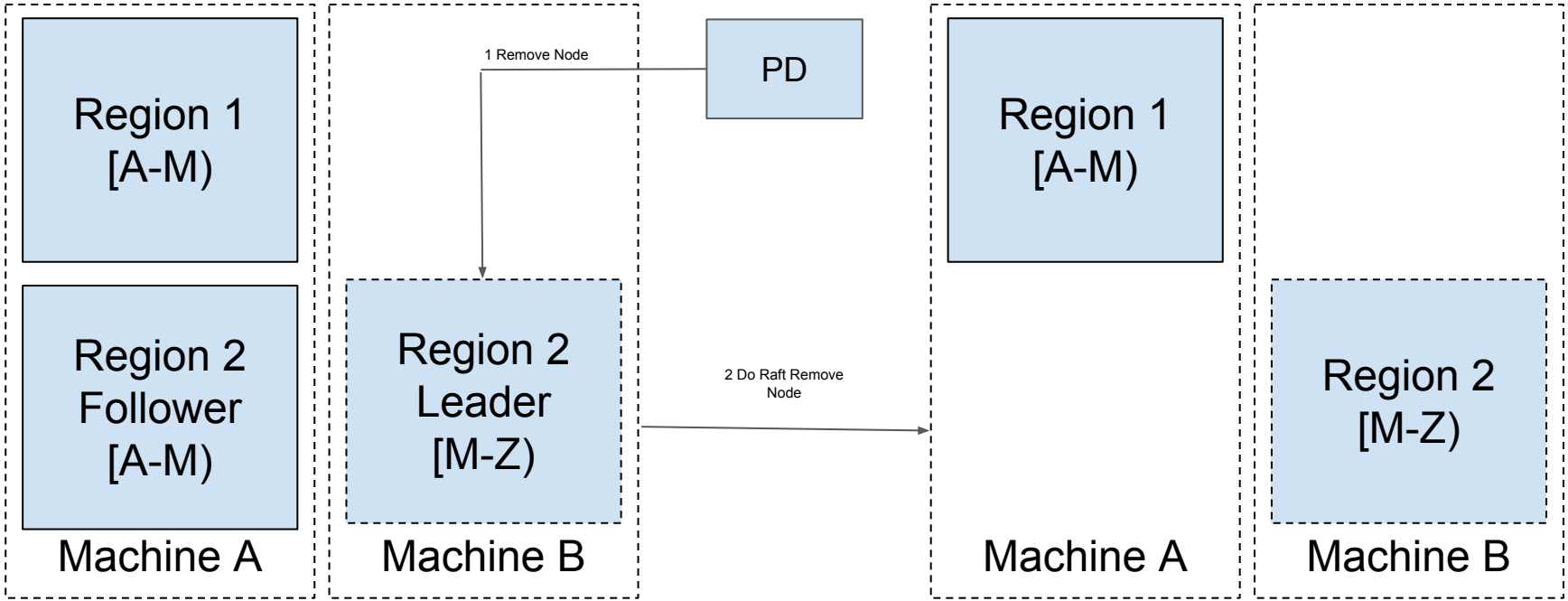
Auto Balance: Move



Auto Balance: Move



Auto Balance: Move



PingCAP

Transaction

- Snapshot Isolation + Lock
- Optimistic Lock
- Percolator
  - TSO (全局时间戳)
  - Primary and Secondary Lock

PingCAP

Route

- PD 知道所有 Region 信息
- Client 本地缓存 Route Table

```
{startKey1, endKey1}   ->    {Region1, NodeA}
{startKey2, endKey2}   ->    {Region2, NodeB}
{startKey3, endKey3}   ->    {Region3, NodeC}
```

- Cache Miss，从 PD 重新获取

Q&A：

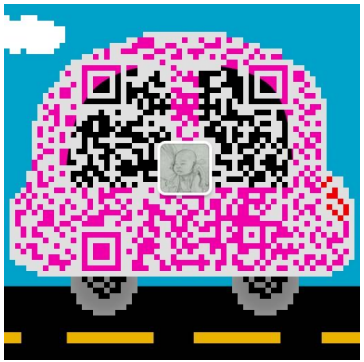
PingCAP

项目地址

TiDB: http://github.com/pingcap/tidb （4600+ stars）

TiKV: https://github.com/pingcap/tikv （1000+ stars）

我的微信号



PingCAP公众号