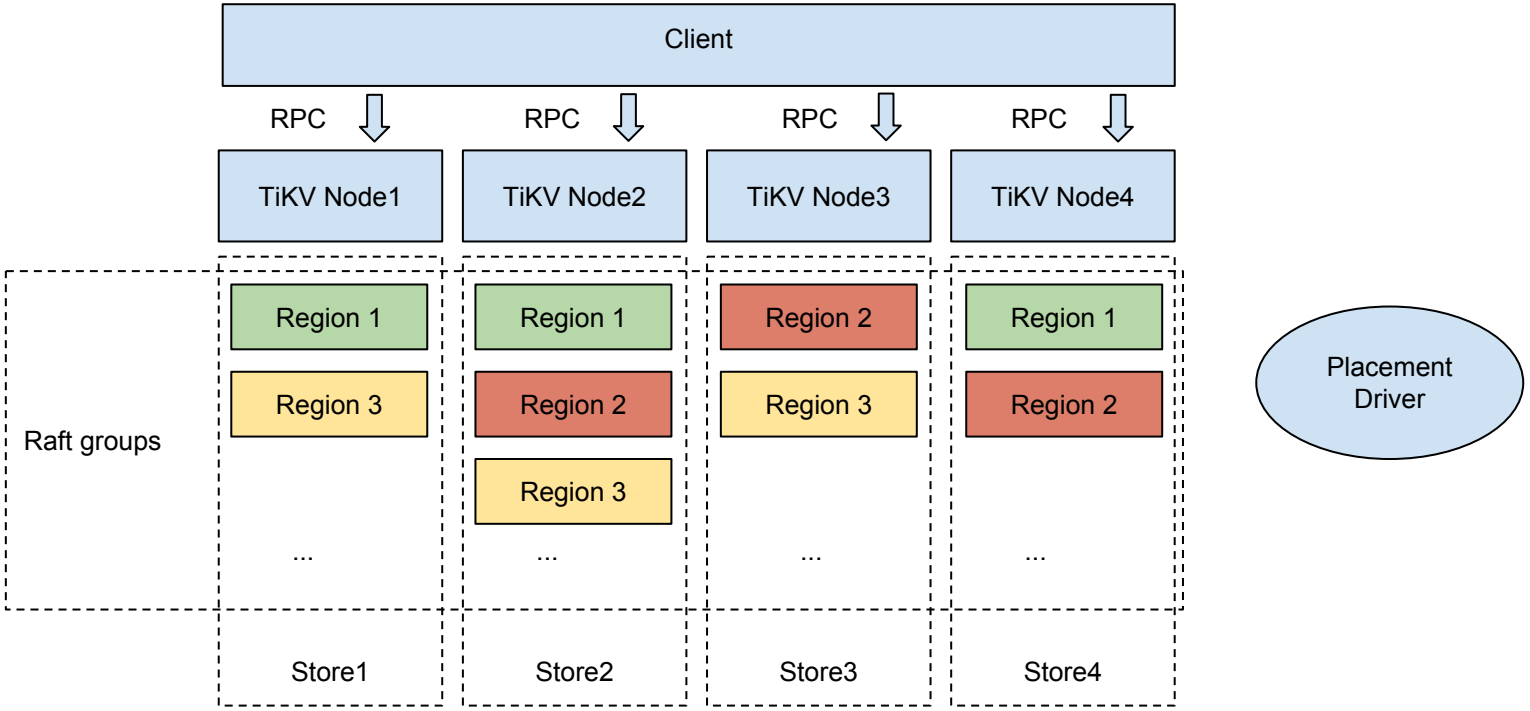
PingCAP

How does TiKV auto-balance work

cuiqiu@pingcap.com

PingCAP

TiKV Architecture



PingCAP

PD - Cluster Meta Router

--Btree

[a, b) -> Region{ID:1, Peers: [{11, N1},{12,N2},{13,N3}]}

[b, c) -> Region{ID:2, Peers: [{21, N1},{22,N2},{23,N4}]}

[c, d) -> Region{ID:3, Peers: [{31, N1},{32,N3},{33,N4}]}

{11, N1} -> {PeerID:11, Node1: 192.168.199.116:5551}

For key: bbbb -> Region2

PD - Balance Scheduler

For TiKV:

1) Store heartbeat -> report store status, including store region count, capacity...

PD cache the store status.

2) Region heartbeat -> report region status, including region peers, region key range...

PD cache the region status, check the region info and balance operator cache to decide whether to do balance.

PingCAP

PD - Balance Scheduler

For PD:

1) pd -> server -> raft cluster -> balance worker
2) balance loop -> do balance -> balance rules -> balance algorithm -> put into balance cache

PingCAP

PD - Balance Scheduler - Balance Rules

balance rules:

BalanceInterval

Balance loop interval time (seconds).

MaxBalanceCount

The max region count to balance at the same time.

MaxBalanceRetryPerLoop

The max retry count to balance in a balance schedule.

MaxBalanceCountPerLoop

The max region count to balance in a balance schedule.

ExpireRegionTime

The min balance interval time (seconds) for one region.

PingCAP

PD - Balance Scheduler

balance algorithm(old version)

-- do balance

1) Select a balance from store to remove peer.
2) Select a balance transfer store to transfer leader.
3) Select a balance to store to add peer.

-- do leader balance

1) Select a balance transfer from store.
2) Select a balance transfer to store.

Each step we call it balance operator.

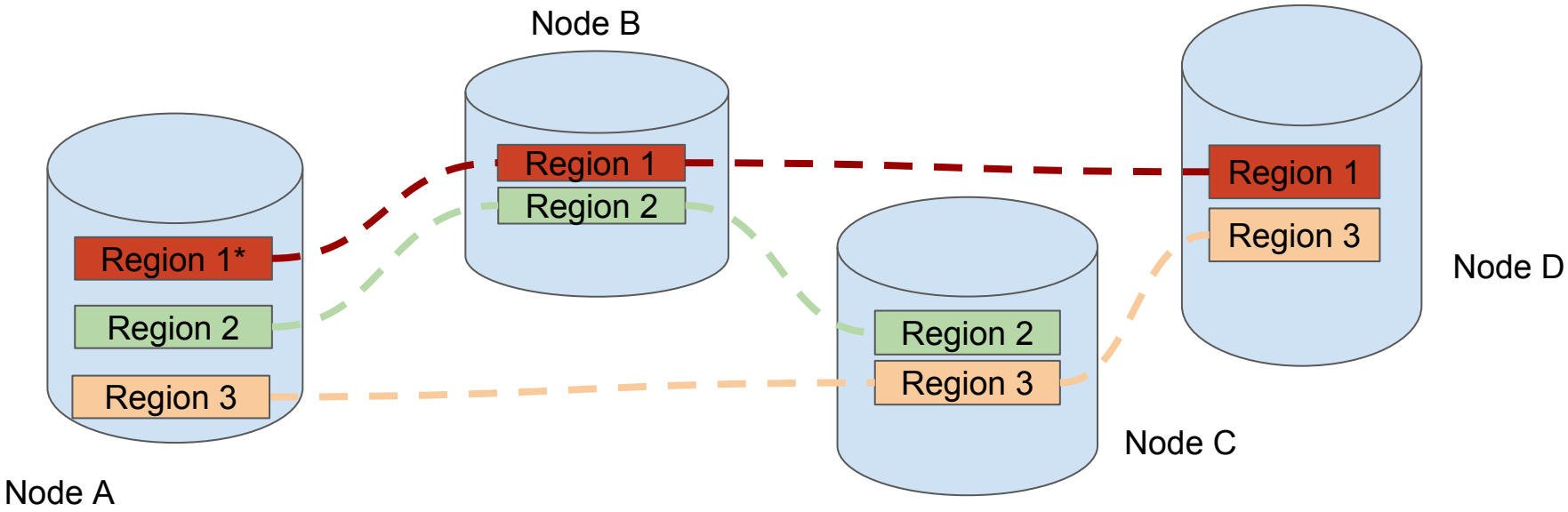
PingCAP

PD - Balance Scheduler

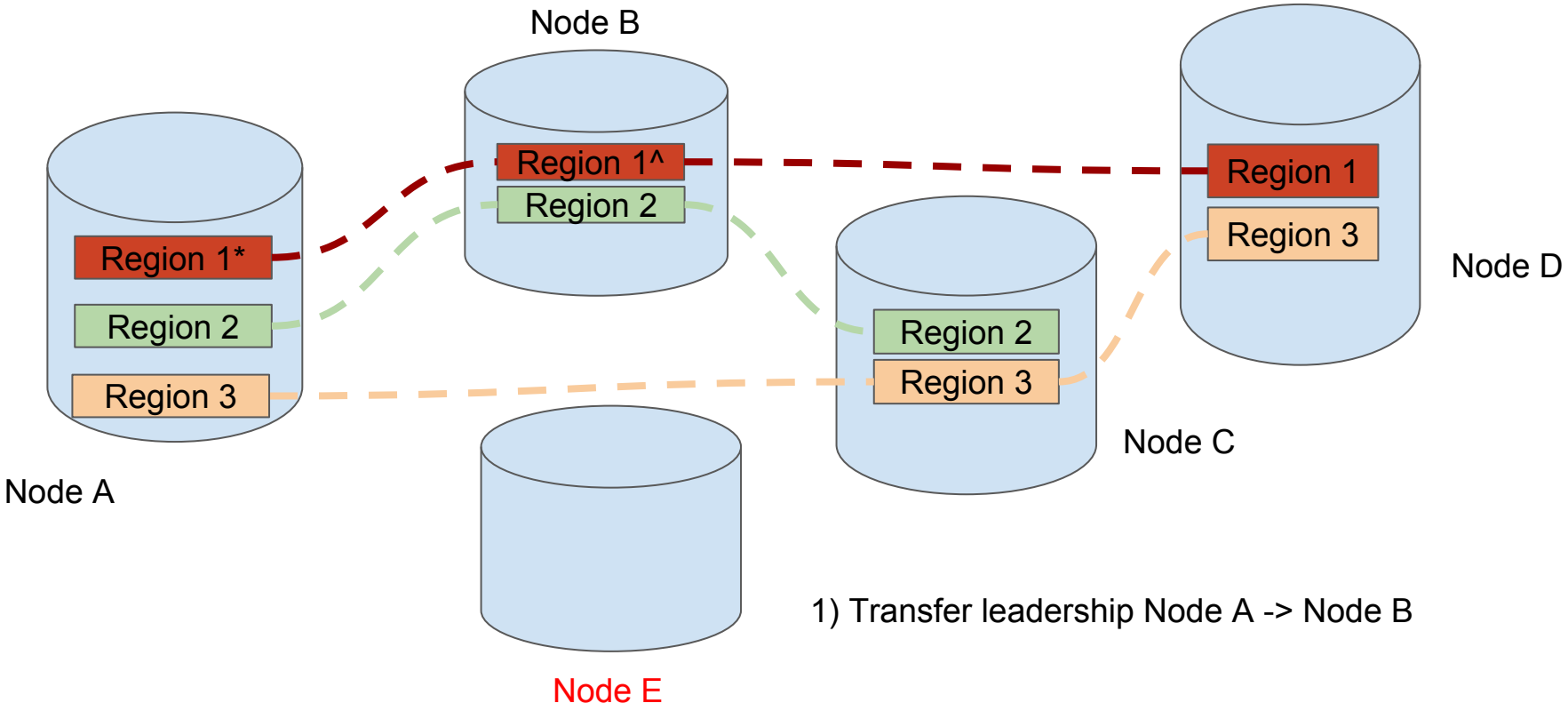
balance algorithm(old version)

Resource Balancer → do balance first, if not, try to do leader balance
resource score: capacity weight * capacity ratio + leader weight * leader ratio
--show an example
store capacity 6000000000
store avaiable 5373284429
leader count 10
region count 100
0.6 * (100 * (6000000000-5373284429) / 6000000000) + 0.4 *(100 * 10 / 100) = 10

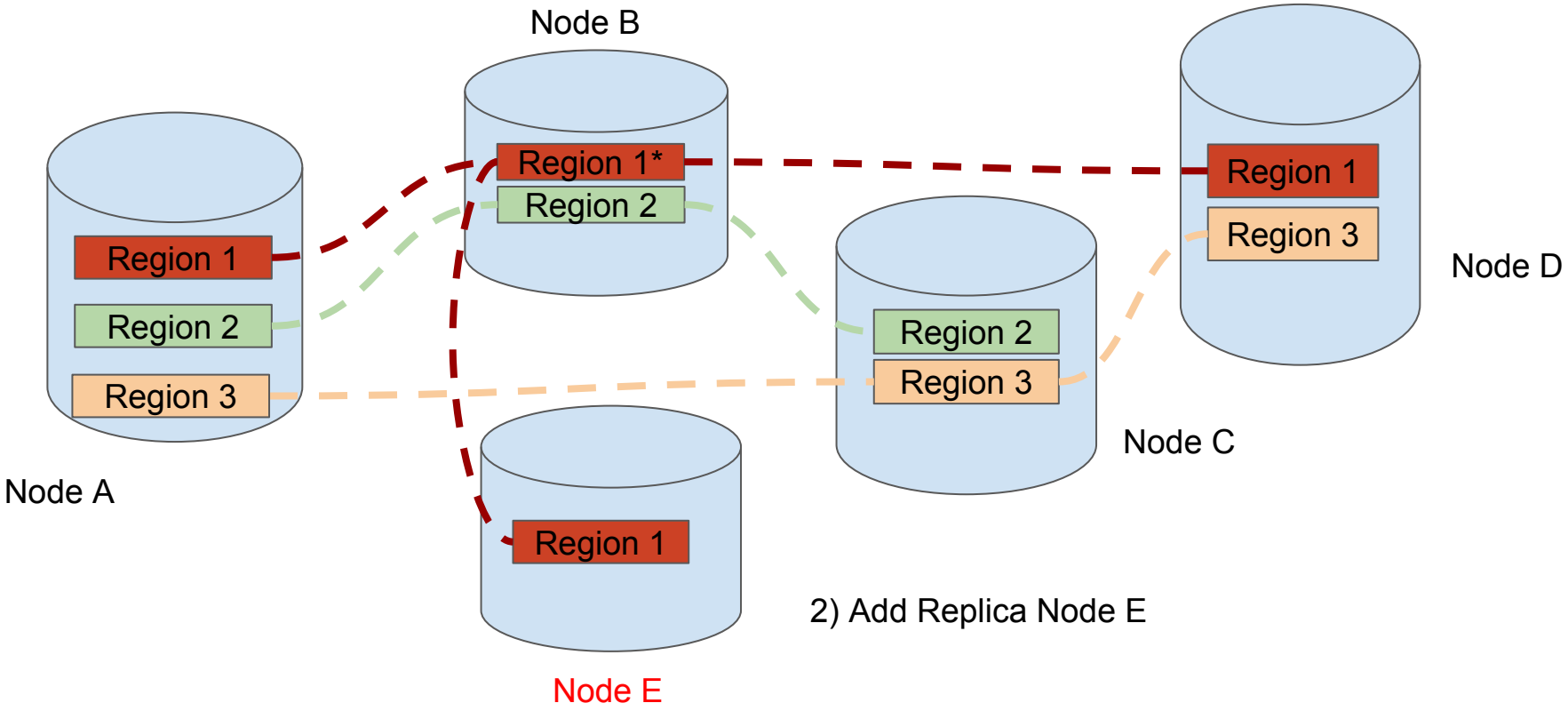
How does TiKV balance ?



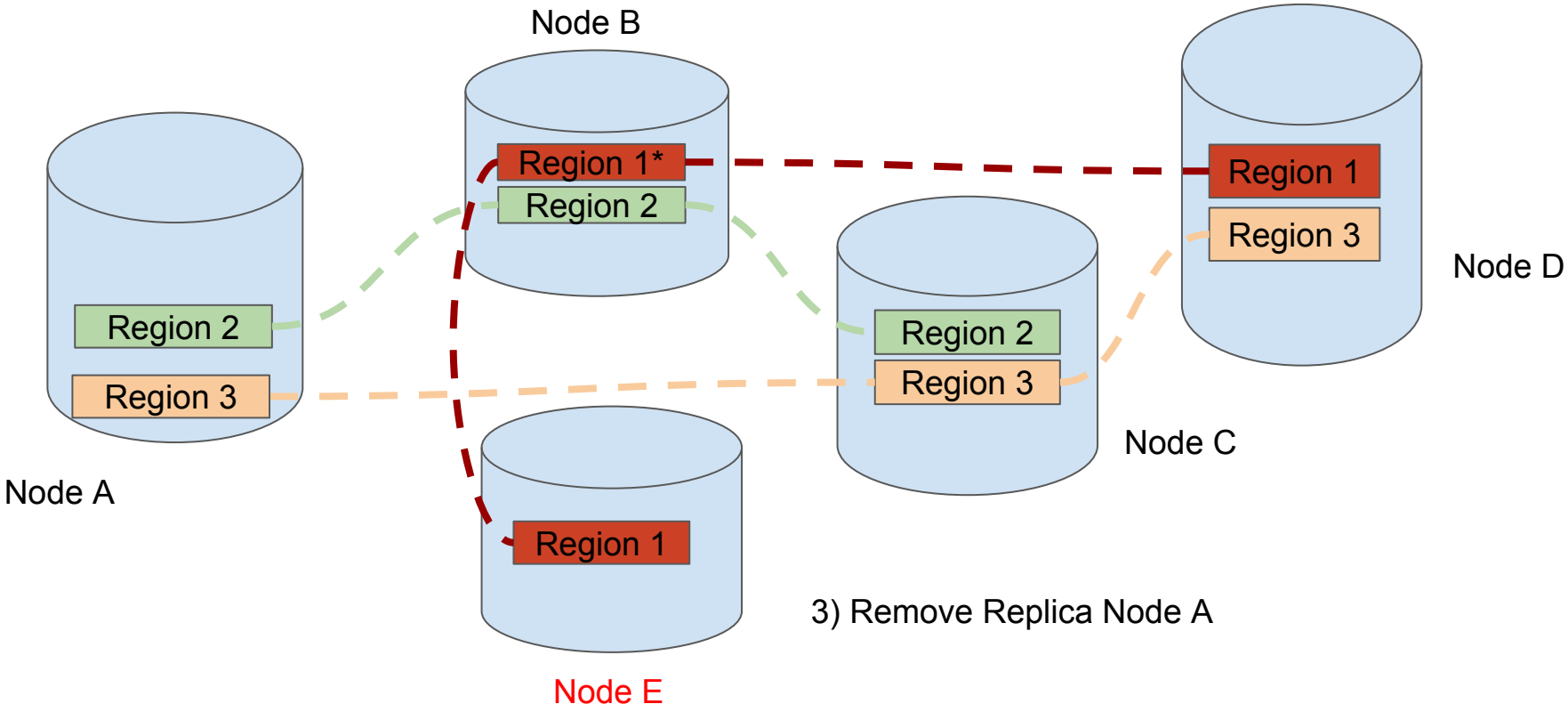
How does TiKV balance ?



How does TiKV balance ?



How does TiKV balance ?



PD - Balance Scheduler

Problems

1) resource imbalance, adjust capacity and leader weight cannot work

	stroe 1	store 2
capacity	483183820800	483183820800
available	312734025861	177911583506
region count	8000	8000
leader region count	3751	493
score	39	40

2) transfer leader overlap between two types of balances

3) none stable

store1: 30, store2: 40 -> store1: 38, store2: 35 -> store1: 32, store2: 38 ...

PingCAP

PD - Balance Scheduler

balance algorithm(new version)

resolve transfer leader overlap problem

-- do leader balance

1) Select a balance transfer from store.
2) Select a balance transfer to store.

-- do follower balance

1) Select a balance from store and a region follower peer.
2) Select a balance to store.

PingCAP

PD - Balance Scheduler

balance algorithm(new version)

resolve none stable problem

MinCapacityUsedRatio

// For capacity balance. If the used ratio of one store is less than this value, it will never be used as a from store.

MaxCapacityUsedRatio

// For capacity balance. If the used ratio of one store is greater than this value, it will never be used as a to store.

MaxSendingSnapCount

// For capacity balance.If the sending snapshot count of one storage is greater than this value, it will never be used as a from store.

MaxReceivingSnapCount

// For capacity balance. If the receiving snapshot count of one storage is greater than this value, it will never be used as a to store.

MaxLeaderCount

// For leader balance. If the leader region count of one store is less than this value, it will never be used as a from store.

MaxDiffScoreFraction

// If the new store and old store's diff scores are not beyond this value, the balancer will do nothing.

PingCAP

PD - Balance Scheduler

balance algorithm(new version)

resolve resource imbalance problem

1 calculate different factor scores

2 select the best balance candidate as result

	capacity	leader	other
store 1	90	20	50
store 2	70	70	60
store 3	60	60	80
store N	10	30	20
score	90-10=80	70-20=50	80-20=60

PingCAP

PD - Balance Scheduler

balance algorithm(new version)

What to do in futher?

1) add more rules to select best balance candidate

2) collect more status from TiKV, like cpu, memory, region key/value hot mark...

3) simulation system for balance algorithm better testing

4) ...