TiDB && TiKV 技术选型思考

刘奇

仅代表个人观点

Lease >= 2017.01.14

About me

- Qi Liu (刘奇)
- Co-founder & CEO of PingCAP
- JD / Wandoulabs / PingCAP
- Old programmer
  - asm/c/c++/go/rust
- Infrastructure software engineer / Open source hacker
- Codis / TiDB / TiKV

Why TiDB ?

一个老程序员夙愿
解决分布式系统的一些问题
分布式缓存：Codis
分布式数据库：TiDB

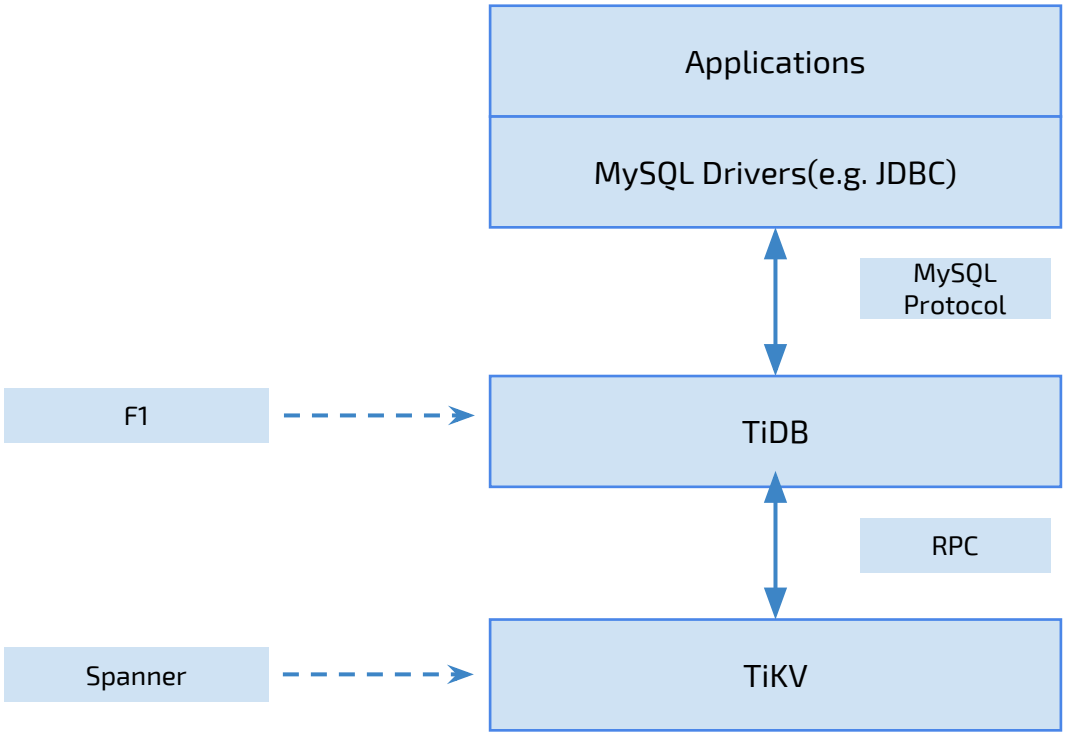
Why TiDB ?

- No sharding
- No inconsistent
- Keep transaction
- Scaling
- Make everything much faster

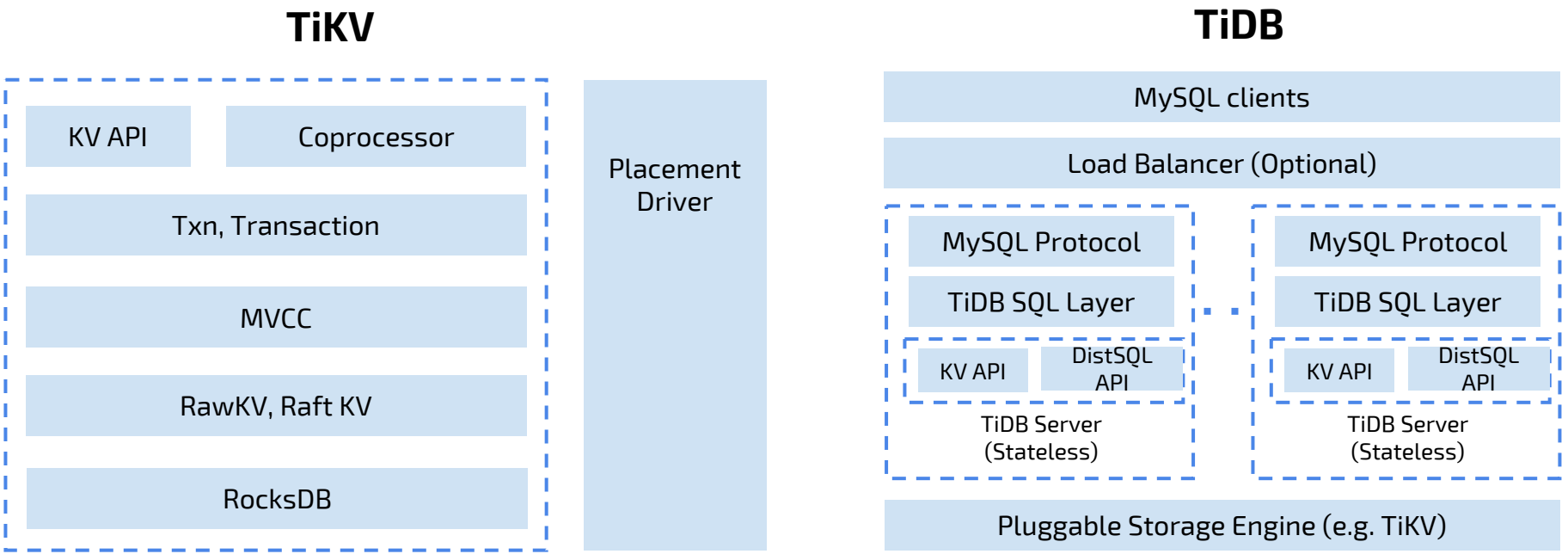
TiDB can be used as

- Unlimited MySQL for both OLTP/OLAP
- 一周无限回档/全球同服
- 异地多活
- MySQL warehouse, query is much faster

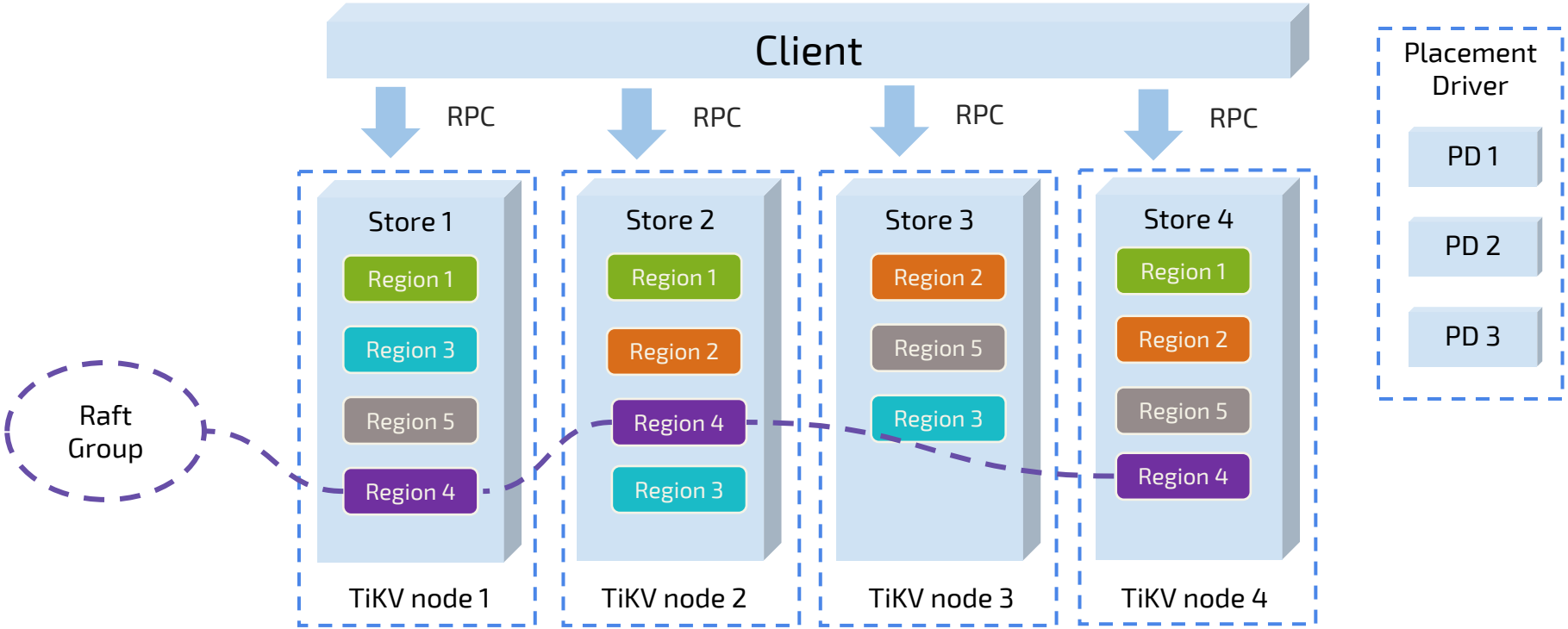
Why multiple layer ?



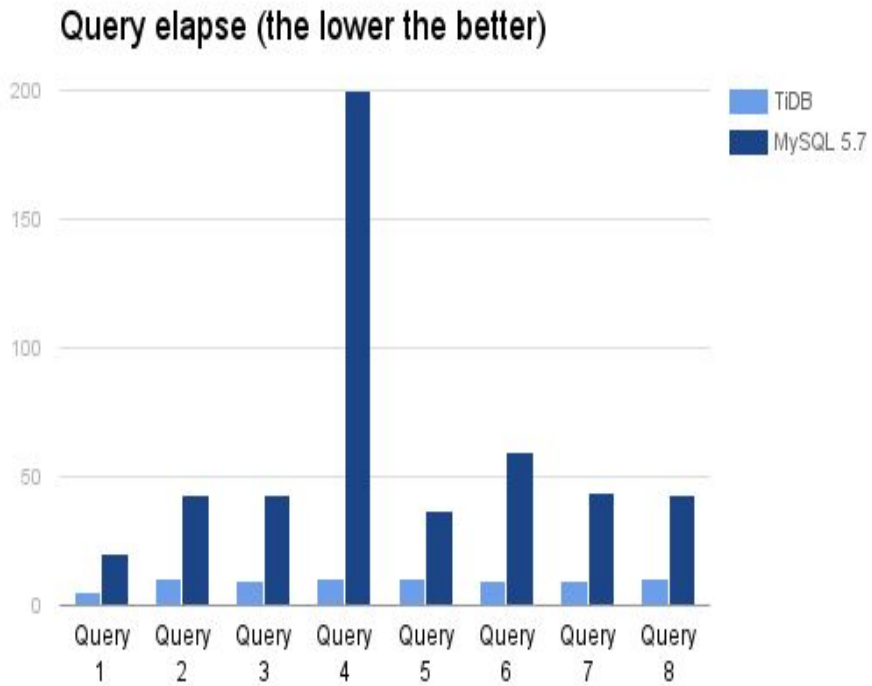
Architecture



TiKV: The whole picture



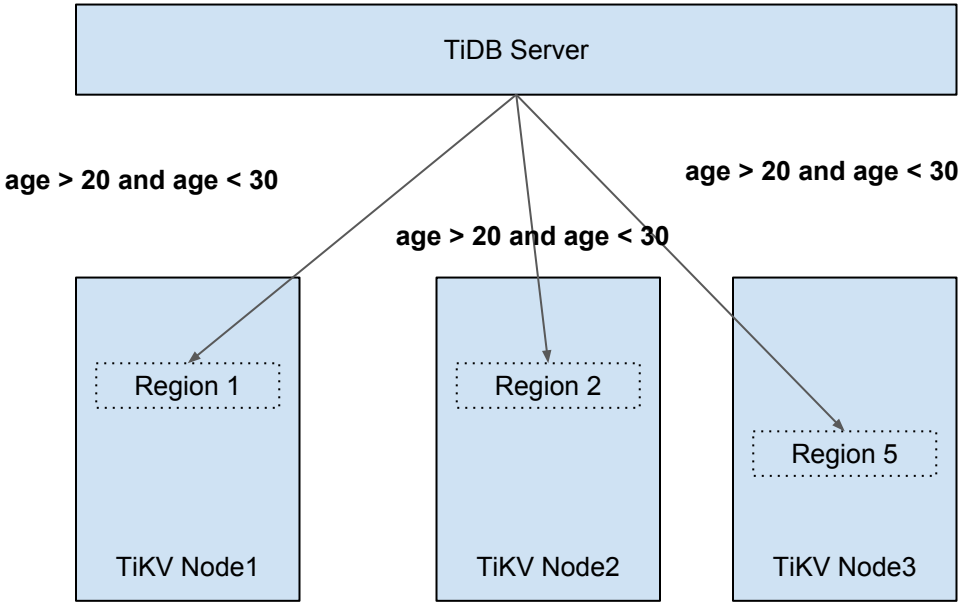
TiDB query performance



TiDB Elapse	MySQL Elapse
5.07699437s	19.93s
10.524703077s	43.23s
10.077812714s	43.33s
10.285957629s	>20 mins
10.462306097s	36.81s
9.968078965s	1 min 0.27 sec
9.998030375s	44.05s
10.866549284s	43.18s

Why TiDB is much faster? Predicate pushdown

TiDB knows that Region 1 / 2 / 5 stores the data of person table.



More data

- Scale to 200T
- Scale to 496 node

But how ?

- 长久以来的正确姿势
- make it run
- make it right
- make it fast

But how ?

然后这不是一个简单的程序
不是两三天就能搞定的
数据库是航母级别规模的东西
How to make it right ?
How to make it run ?
How to make it fast ?

How to make it right ?

- Tests first
- Mock layer by layer

Tech choices:**Paxos vs Raft**

XXX-Paxos is complicated
Difficult to implement
Needs to test the implementation for years

Programming language choices

go or rust or both?
FAST, SAFE AND PRODUCTIVE — PICK THREE
Go: productive
Rust: fast, safe

But why not c++ ?

- Hard to control for large project.
- Can’t detect data race when compiling the source
- No package management tools
- Personally programming cpp sucks

Is rust really wonderful?

- Sounds too good to be true
- Compiling speed makes you want to kill yourself
- Lacks of libraries, we are still waiting for grpc

Hold on, why MySQL protocol ?

- I heard that PostgreSQL is more beautiful and powerful.
- Users
  - MySQL >> MongoDB >> PostgreSQL

Embed storage: the war

- Lmdb
- Leveldb
- Innodb
- Rocksdb
- Wiredtiger

grpc VS others

Some rpc frameworks are really fast
But we want the rpc that widely used, work with protocol buffers
Came from ex-googler

prometheus VS others

Very active
Widely used
Works well with grafana

grafana VS others

Widely used
Came from ex-googler

Rebuilding all or use mysql with tikv ?

Everything should be distributed
Distributed query plan
Multiple writable servers
No read/write bottleneck

We are still on our way

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