SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

一个架构的演进和开发哲学

黄东旭 PingCAP

关于我

- PingCAP Cofounder & CTO
- MSRA / Netease / PingCAP
- 基础软件工程师 / 架构师 / 开源狂热分子
  - Codis
  - TiDB
  - TiKV
- Weibo: @Dongxu_Huang
- Email: huang@pingcap.com

一个基础软件的开发者的创业和技术故事

The Crazy Idea

- 被 MySQL 分库分表和中间件折磨得不行了
- Codis 解决分布式缓存问题，然后呢？
- Google 的 Spanner 和 F1 好帅呀！

So, 经过一晚上的深(tou)思(nao)熟(fa)虑(re)，决定：

从头做一个数据库，从根本上解决 MySQL 的扩展性问题

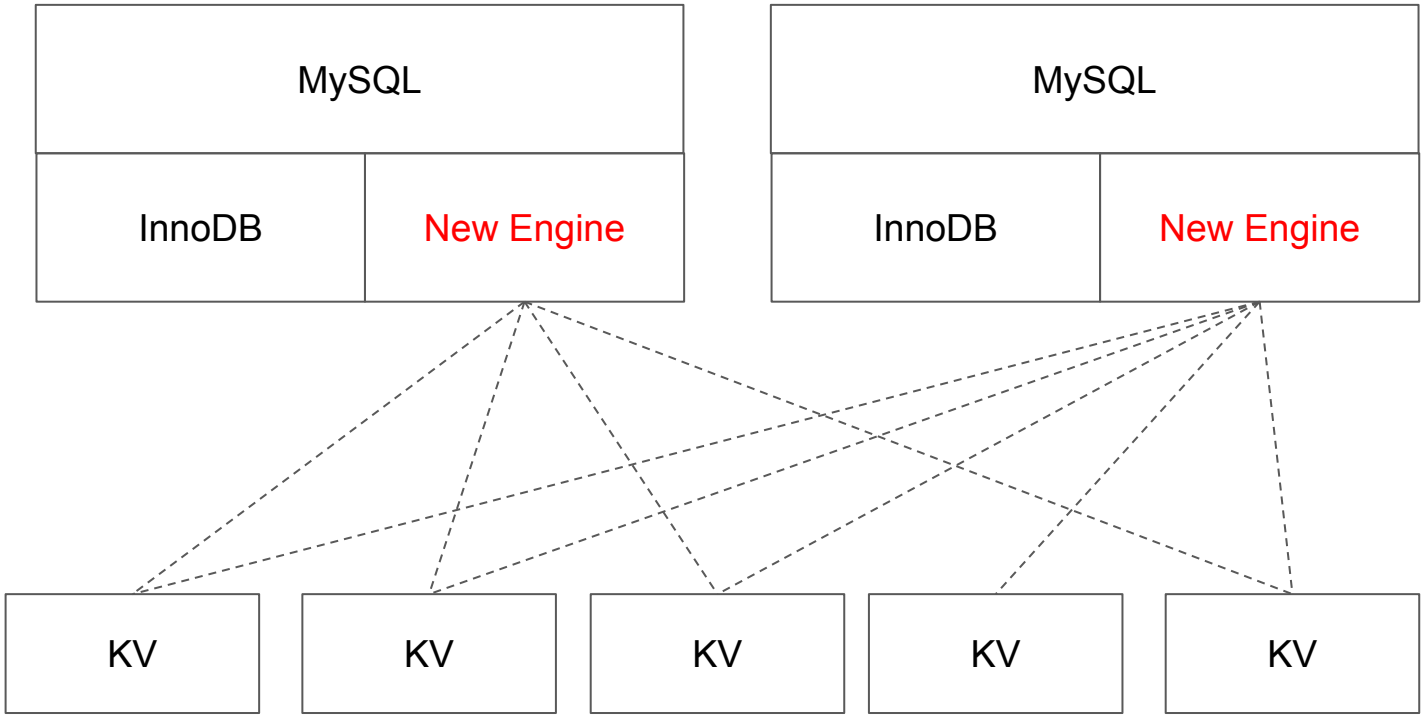
过了一天...

- 选择什么语言
- 架构怎么做
- 怎么保证写对了
- 开源还是不开源
- 人从哪来
- 钱从哪来
- 目标是个分布式系统怎么测试啊
- ...

冷静一下，先确定一个小目标

- 完整的 MySQL 兼容
  - 海量的测试
  - 极小的用户迁移成本
  - 明确的开发目标
- 等等，兼容 MySQL 的最简单的方案是？

架构 0.1


MySQL
InnoDB
New Engine
MySQL
InnoDB
New Engine
KV
KV
KV
KV
KV

pingcap / mysql-5.6.24

Unwatch 3 | Star 1 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Pulse | Graphs | Settings

add cocdb impl

master

Browse files

c4pt0r committed on 23 Apr 2015

1 parent df98aff commit d088a19ddc3ef13e750b8d4aaf0bc8d130296df1

Showing 13 changed files with 5,727 additions and 126 deletions.

Unified | Split

8 Docs/INFO_SRC

View

```
@@ -1,7 +1,7 @@
-commit: bc821a1ab786bb304782fbed4cfcad6e7947b76a
-date: 2015-04-19 12:51:26 +0800
-build-date: 2015-04-22 10:09:28 +0800
-short: bc821a1
 branch: master
```

```
+commit: df98aff37c3607cac03ebf7c28a764ef1866b2d8
+date: 2015-04-22 10:18:47 +0800
+build-date: 2015-04-23 21:39:37 +0800
+short: df98aff
 branch: master
```

6 storage/rocksdb/CMakeLists.txt

View

```
@@ -30,7 +30,6 @@ ENDIF()
 INCLUDE_DIRECTORIES(AFTER ${Rocksdb_INCLUDE_DIRS})
 SET(CMAKE_REQUIRED_INCLUDES ${Rocksdb_INCLUDE_DIRS})
-
 SET(CMAKE_CXX_FLAGS "${CMAKE_CXX_FLAGS} --std=c++11")

 # TODO is there a better way to do this?
@@ -47,10 +46,13 @@ IF(ROCKSDB_OK)
       ha_rocksdb.cc ha_rocksdb.h
       rdb_datadic.cc rdb_datadic.h
       rdb_locks.cc rdb_locks.h
+      cocdb.grpc.pb.cc cocdb.grpc.pb.h
+      cocdb.pb.cc cocdb.pb.h
+      cocdb_client.cc cocdb_client.h
       rdb_rowmods.cc rdb_rowmods.h
       rdb_applyiter.cc rdb_applyiter.h)
```

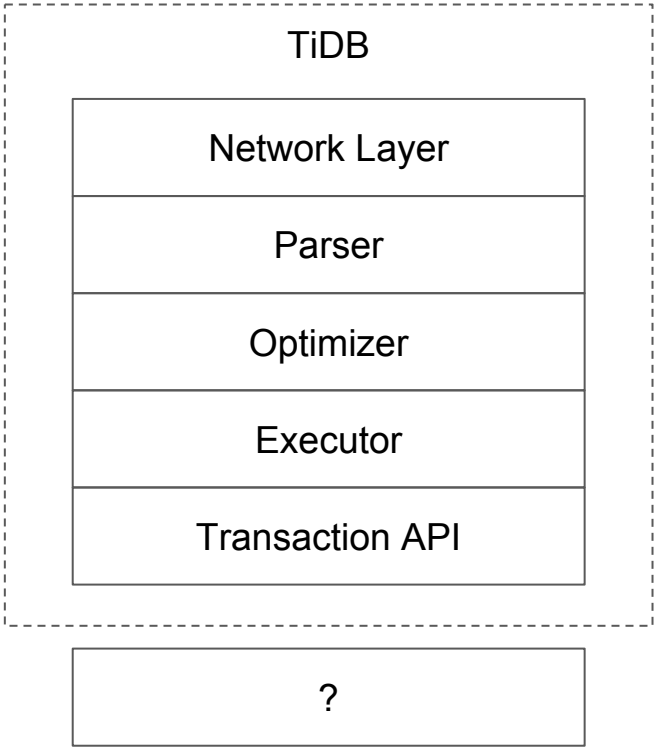
架构 0.1

- 为什么那么慢。。。
- 我想改改 SQL 优化器。。。
  - No way
- 我想改改事务模型。。。
  - No way

架构 0.2

- 好吧，为了灵活性，看来是绕不过去了
- 重写 SQL Layer
  - Parser
  - Optimizer
  - Executor
- 不过，终于可以用最喜欢的编程语言了

架构 0.2

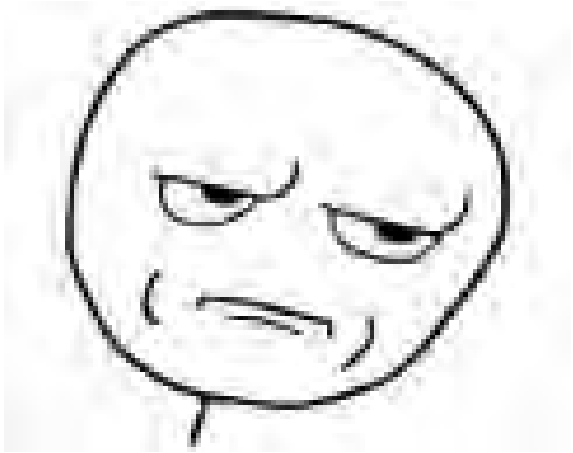


CTO 说了要保证这货长得和 MySQL 一模一样

- 网络协议
- 语法

架构 0.2

- 3 个月过去了，一切进展顺利
- 不过。。。



?

Rule #1:

"All problems in computer science can be solved by another level of indirection"

--- David Wheeler

架构 0.2

在苦逼的写 Parser 的同时。。。我们还做了：

- 收集了 MySQL 社区所有我们能找到的集成测试，到现在大约累计了 1000w 个
  - MySQL unittests
  - SQL logic tests
  - ORM tests
  - ...
- 将存储引擎的行为抽象成很薄的几个接口，使得可以无缝的接入各种嵌入式 kv engine
  - BoltDB
  - LevelDB，RocksDB
  - LMDB
  - Mem
  - …
- 团队大约十号人了。。。还好每层都拆分的比较彻底，否则没法并行了
- 开源了，顺便上了把 HackerNews 的首页。。。

Rule #2:

Talk is cheap, show me the tests

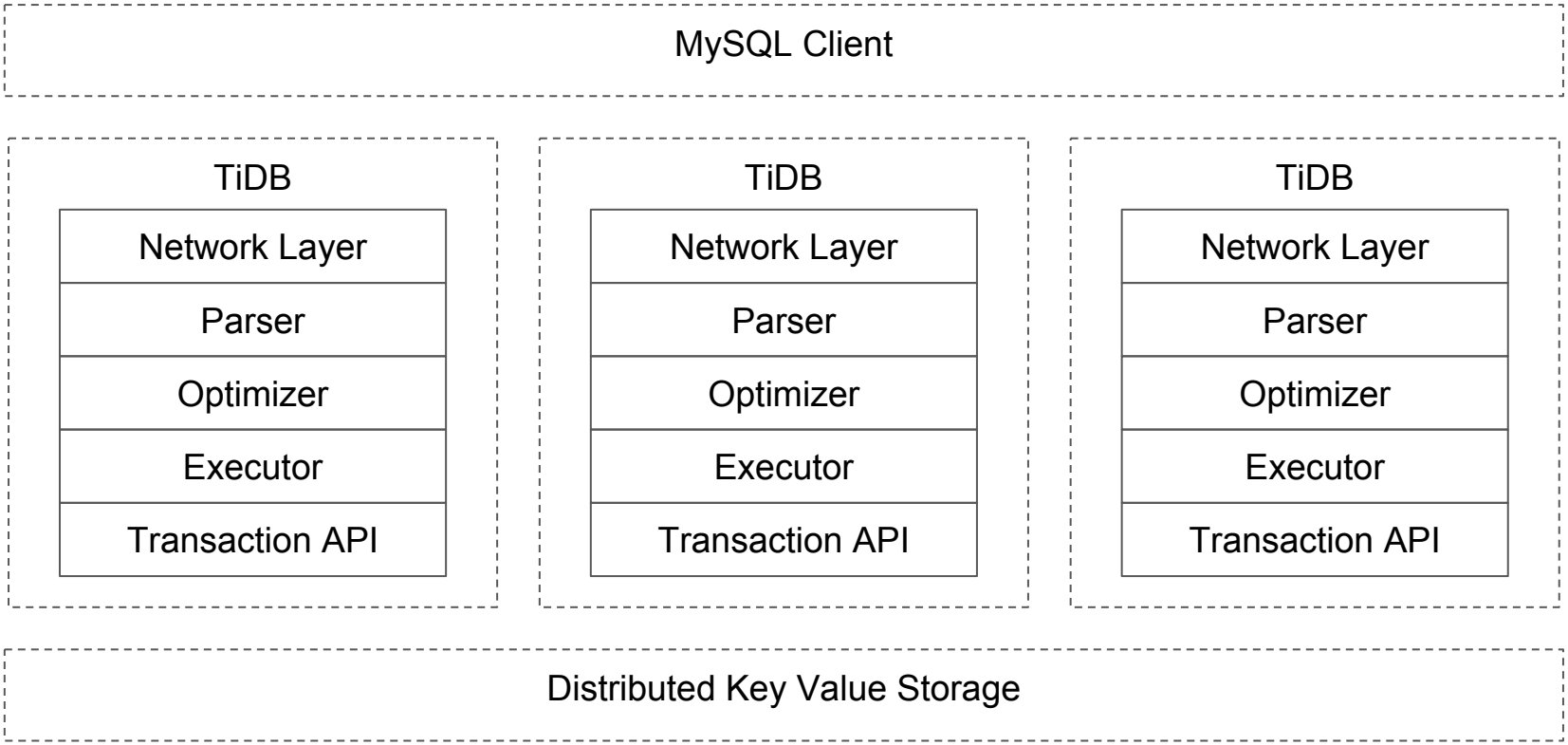
Test matters and it's complex

- 构建一个数据库最难不是写出来，而是证明它是对的
- 对于一个分布式数据库来说更加困难
- 新的开发者，新的模块加入，是否会带来问题？怎么证明？
  - 只有测试

架构 0.5

- 由于 SQL 层已经和存储层完全分离，那么一个很自然的想法：接一个分布式存储上去啊！
- 有了一开始的前车之鉴，这次我们比较谨慎
  - 选择一个比较稳定的分布式引擎，否则你也不知道是谁的锅。。。
  - HBase
  - Why

架构 0.5



距此创业已半年。。。

架构 1.0：TiKV

接口层有了海量的 Test 保证，让设计工作没有太过困难

- 架构选型
  - 语言
  - 模型
  - 核心算法
- 当有极大的自由度的前提下，你是否能控制住膨胀的野心？因为复杂度才是你最大的敌人。

Why Rust

- 我们的团队背景
- 高性能
- 安全
- 更现代的 C++
- Why not Go？你不怕吗？

Why Rust

btw...一不小心成了 Rust 社区最大的开源项目之一

Other Weeklies from Rust Community

- This week in Rust docs 29. Updates from the Rust docur
- These weeks in Servo 82. Servo is a prototype web brow
- This week in Ruru 4. Ruru lets you write native Ruby exter
- What's coming up in imag 19. imag is a text based perso suite
- This week in TiKV 2016-11-07. TiKV is a distributed Key-
- PlanetKit week 3: Hexagons! PlanetKit generates colorful resemble planets. (Week 1 introduces PlanetKit and week terrain).

navitia　OneSignal　OVH.com Innovation is Freedom

Friends of Rust
(Organizations running Rust in production)

POSTMATES　PingCAP

你司

Sandstorm.io　SENTRY　snips

架构 1.0: TiKV

- 弹性扩展
- *真正的*高可用
- 高性能
- 强一致

另外时间不太多。。

2015 年的冬天。。。我们就在纠结中度过

Rule #3:

Embrace the community

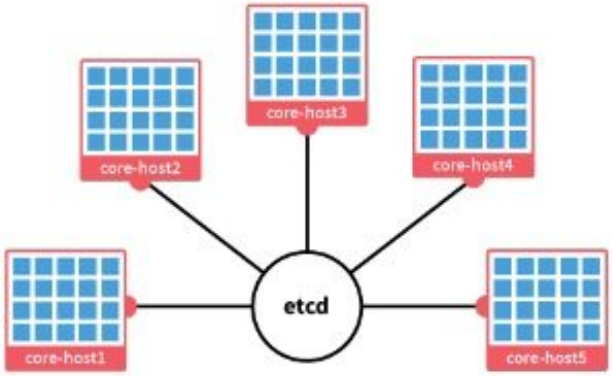
you don't need to do everything

(主要还是懒)

Raft: Etcd

ETCD

- Distributed key/value store
- Like a directory tree
- JSON/REST API
- Uses a Discovery URL



images: coreos.com

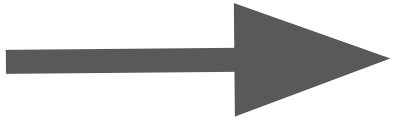
Why Etcd?

- 唯一在生产环境中大量验证过的 Raft 实现
- 架构设计合理，状态机抽象彻底，测试极其充分
- 在工程上对论文的原始算法做了大量优化
- 还有。。。

我们和 CoreOS 的 Etcd team 是好盆友。。。。

但是 etcd 是 go 写的啊。。。

- Port etcd’s Raft implementation line by line.
- Port etcd’s tests line by line.



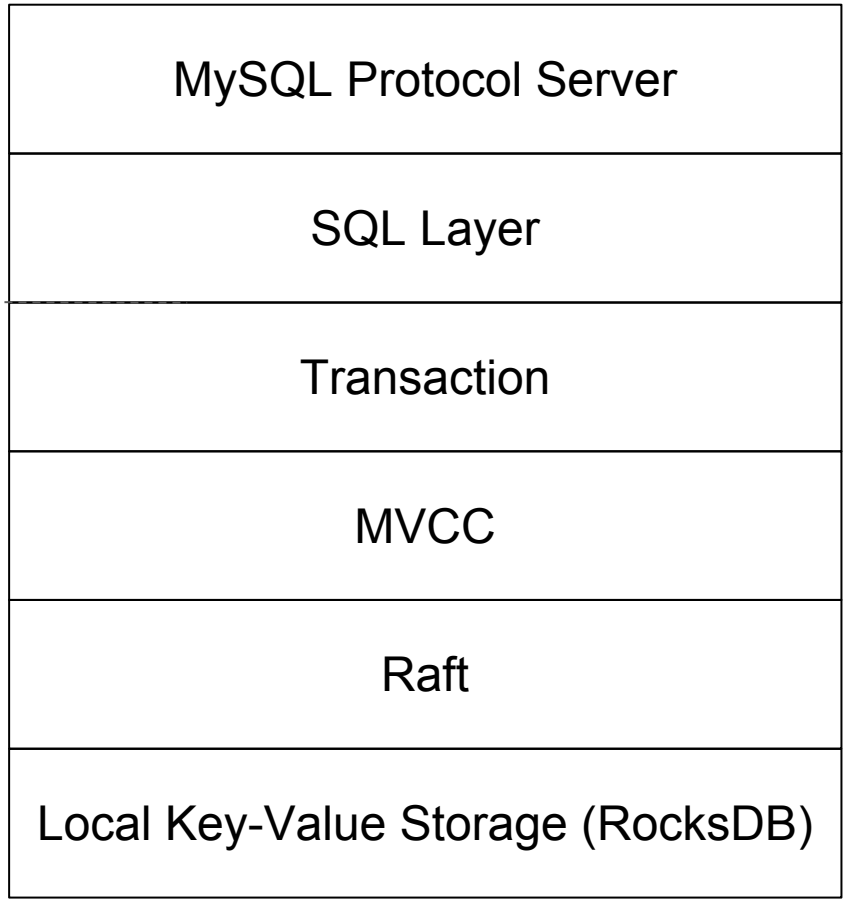
Storage engine: RocksDB

- Codebase: LevelDB
- 活跃的社区
- 多线程 Compaction
- 各种贴心的 API
- 无数个 Tuning Point
- 还有...

2016 年 4 月 1 日。。。TiKV 开源了

TiDB
The MySQL Protocol layer

TiKV
The Key-Value layer

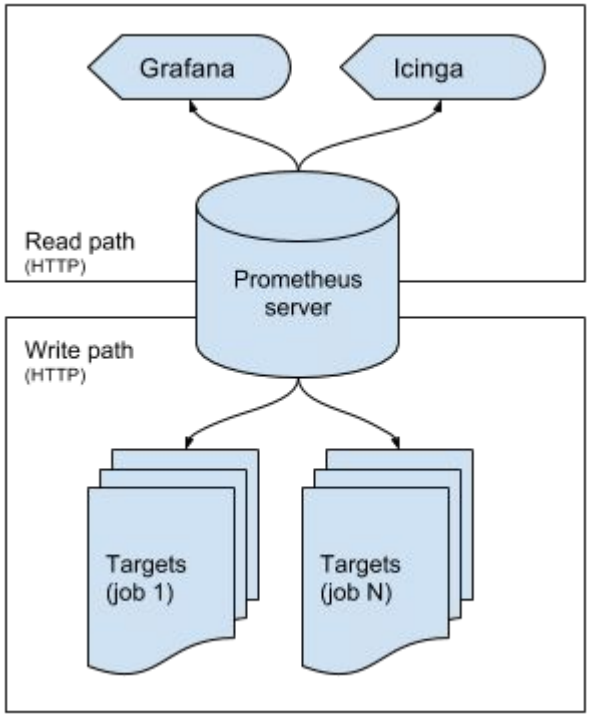


What's more: Metrics 监控是个大问题

- Metrics 监控是最容易被忽视的重要组件
- 自己写 Metrics 监控工具是个体力活，对于小团队来说得不偿失
- Where there's a metric there's a way.

Prometheus + Grafana

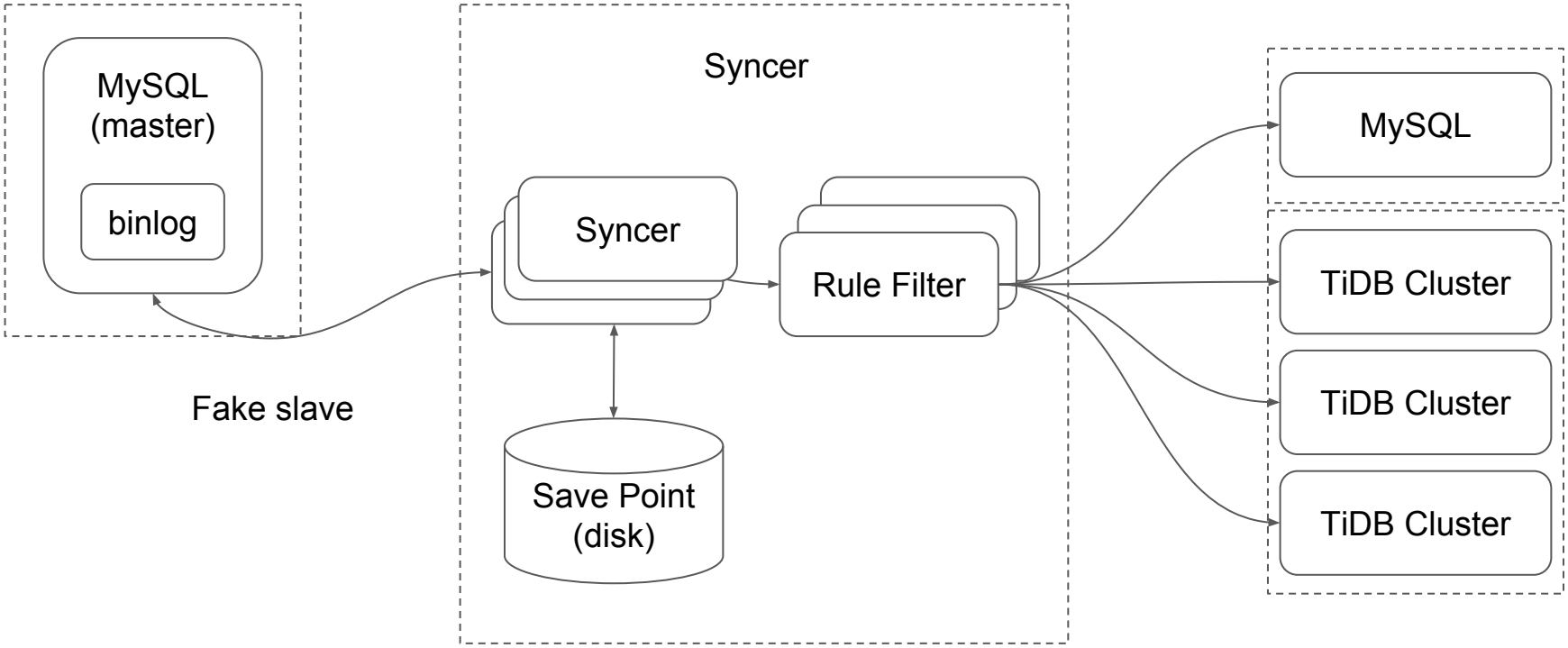
- Prometheus 负责收集 Metrics，并提供一个灵活的 DSL 提供查询，提供监控报警的机制
- Grafana 负责可视化，自定义 Dashboard



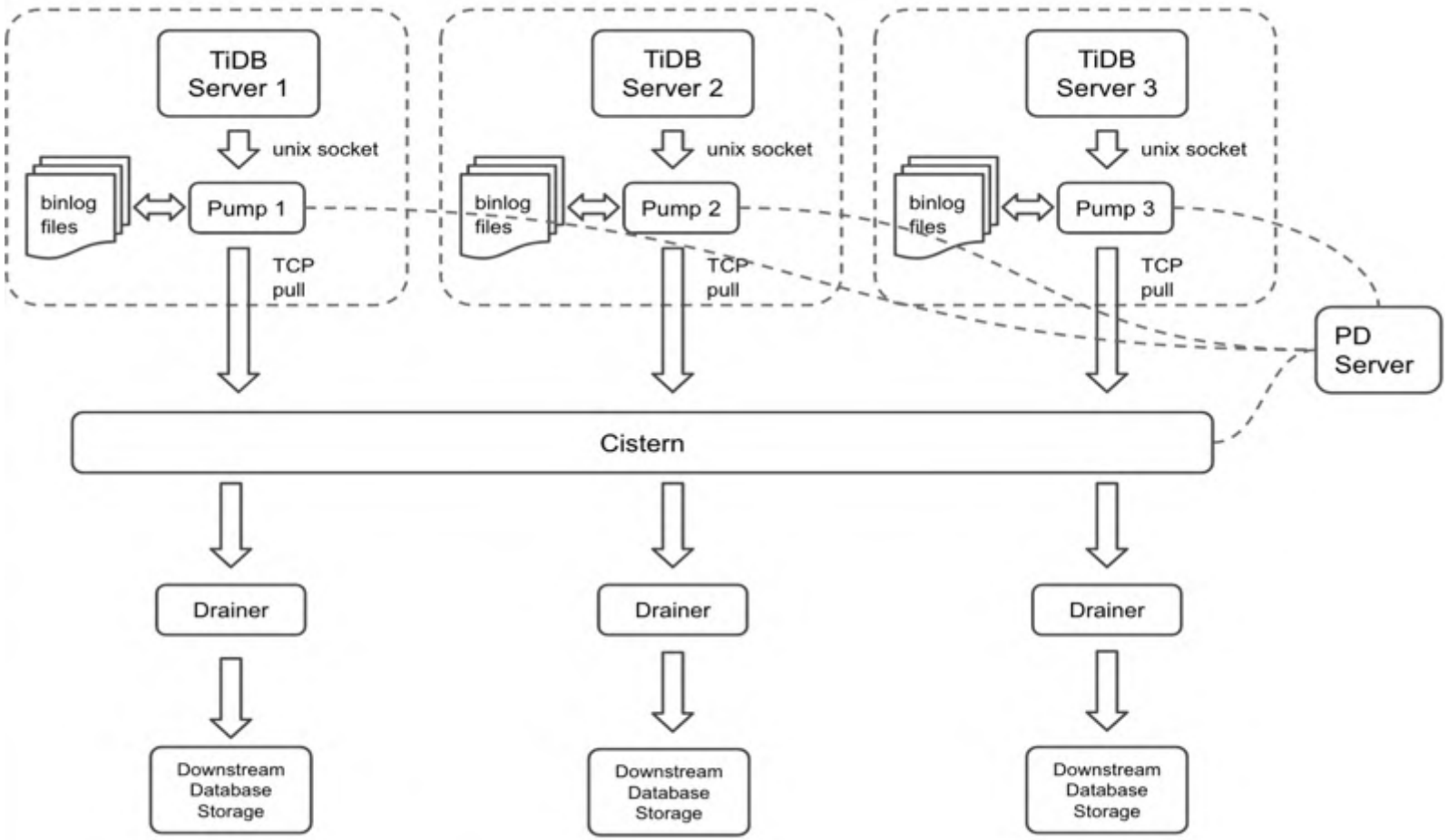
Tools matter

- 最小化业务方的迁移成本 --- 基础设施团队的自我修养
- 「无意外」原则
- 「Always believe shit is about to happen」
  - 业务数据重如泰山
  - 基础设施需要假设随时会挂
  - 如何保护自己，如何保护业务

syncer: MySQL to *



TiDB binlog：TiDB to *



展望 : 基础软件架构的趋势 - Cloud-Native

- Google 和 CNCF
- 被忽略的运维成本
  - 人的价值到底是什么
- 面向未来的基础架构是什么？
  - 对架构师的挑战

Thanks

Q&A

PingCAP - SDCC 2016 交流群