强一致、高可用、高性能分布式Log存储系统的设计与实现

简怀兵

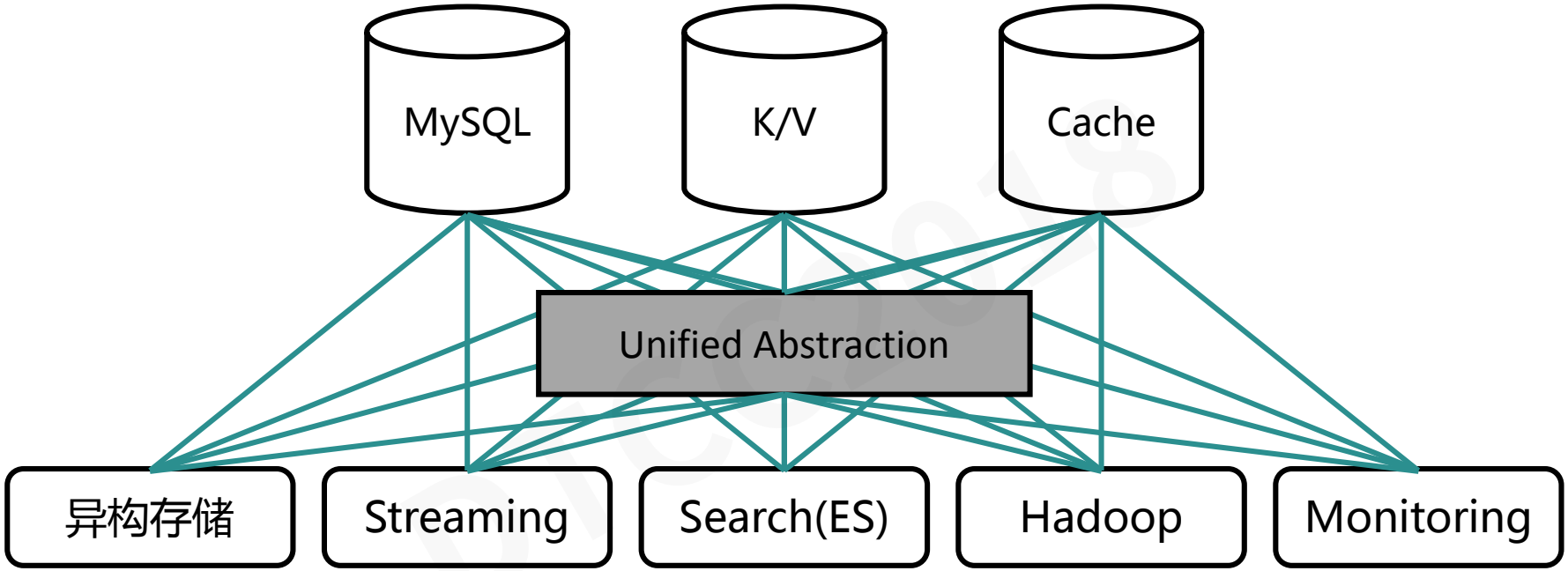
研发背景

设计思路

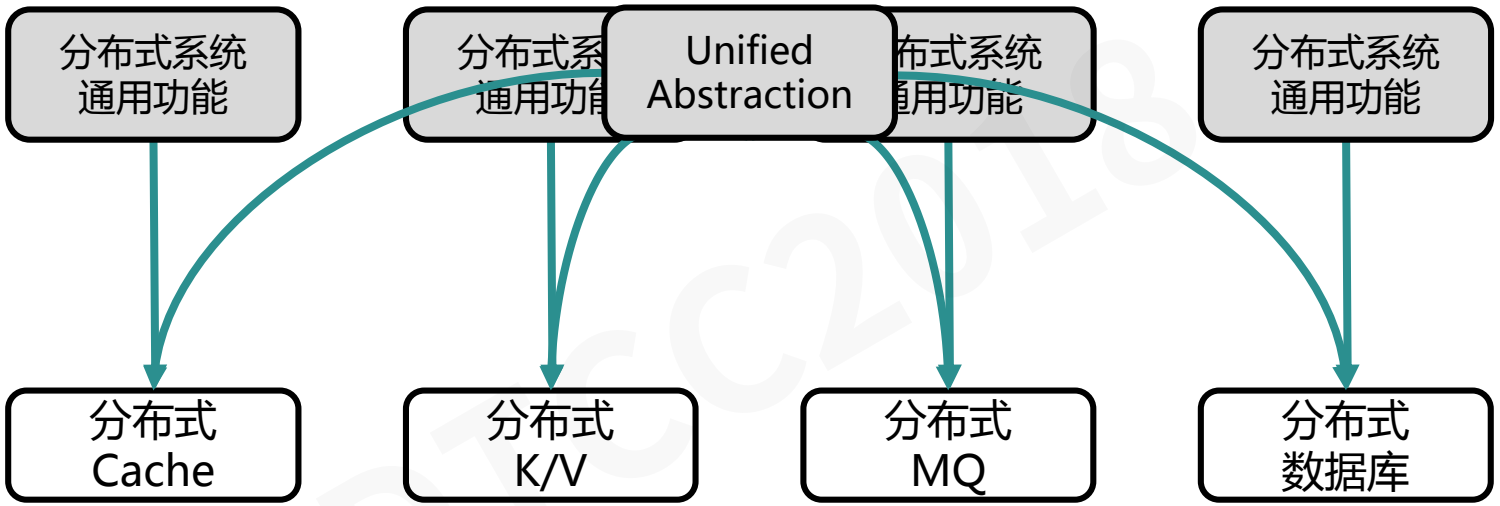
关键实现

应用形态

诉求1：“解耦”不同业务系统



诉求2：其它分布式系统的“积木”



思路起源

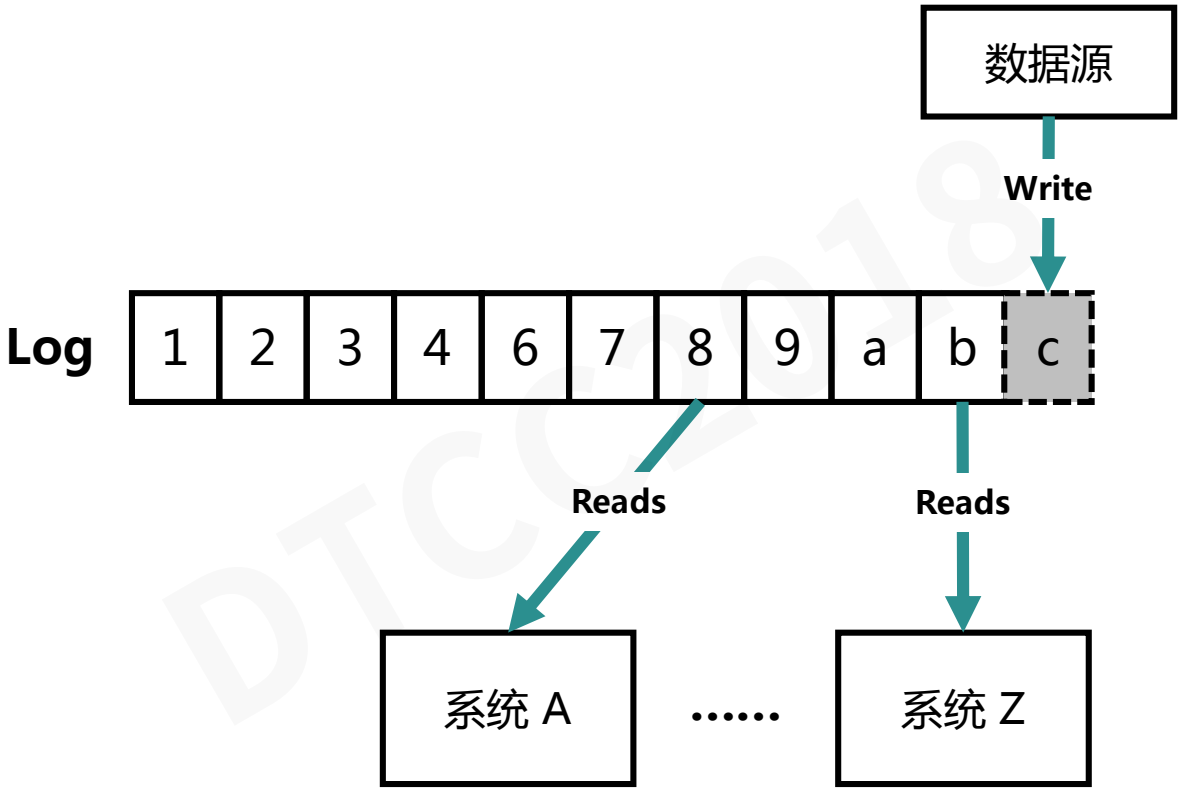
“

You can't fully understand databases, NoSQL stores, key value stores, replication, paxos, hadoop, version control, or almost any software system without understanding logs;

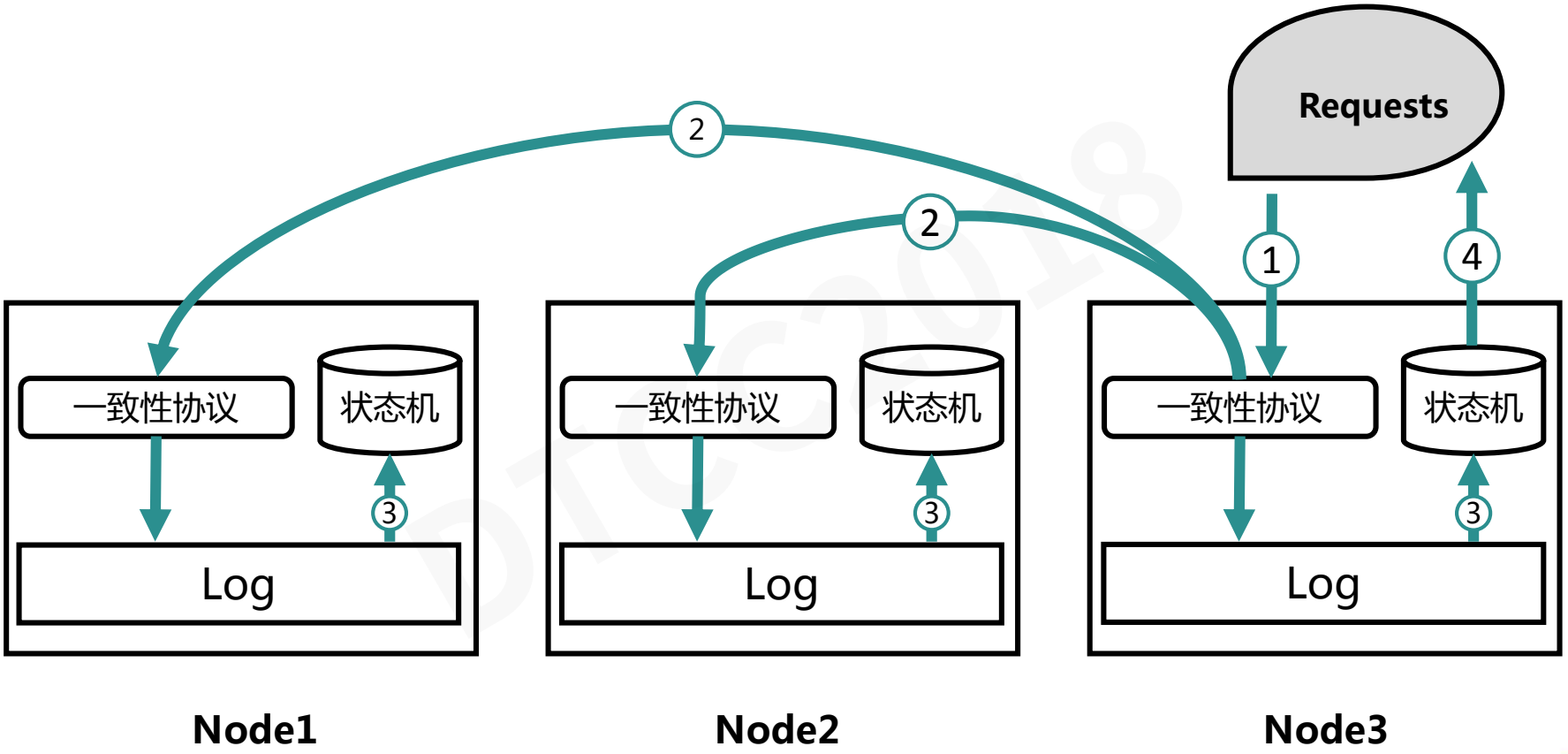
”

《The Log: What every software engineer should know about real-time data's unifying abstraction》

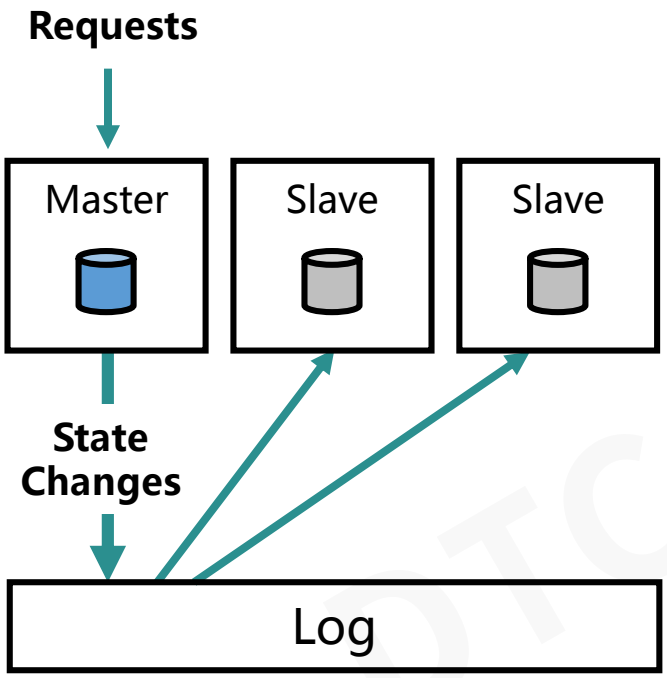
思路：统一的Log抽象



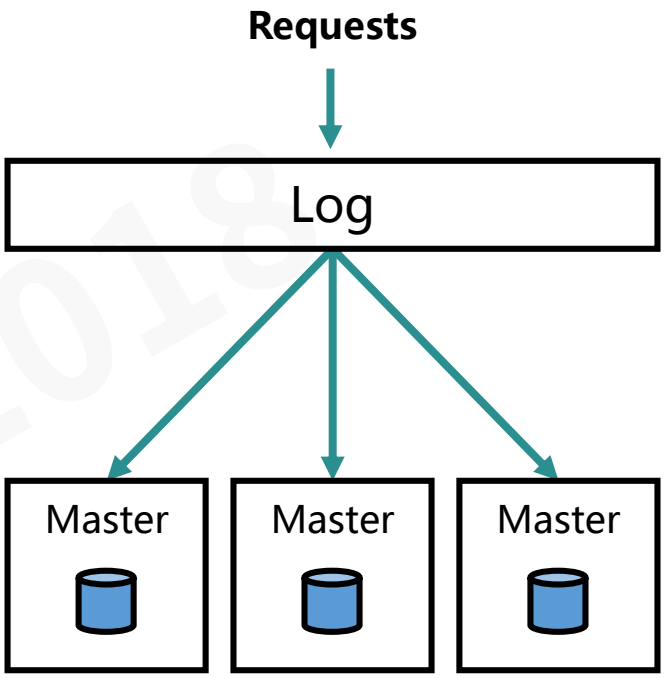
思路：RSM（Replicated State Machine）



思路：RSM延伸应用

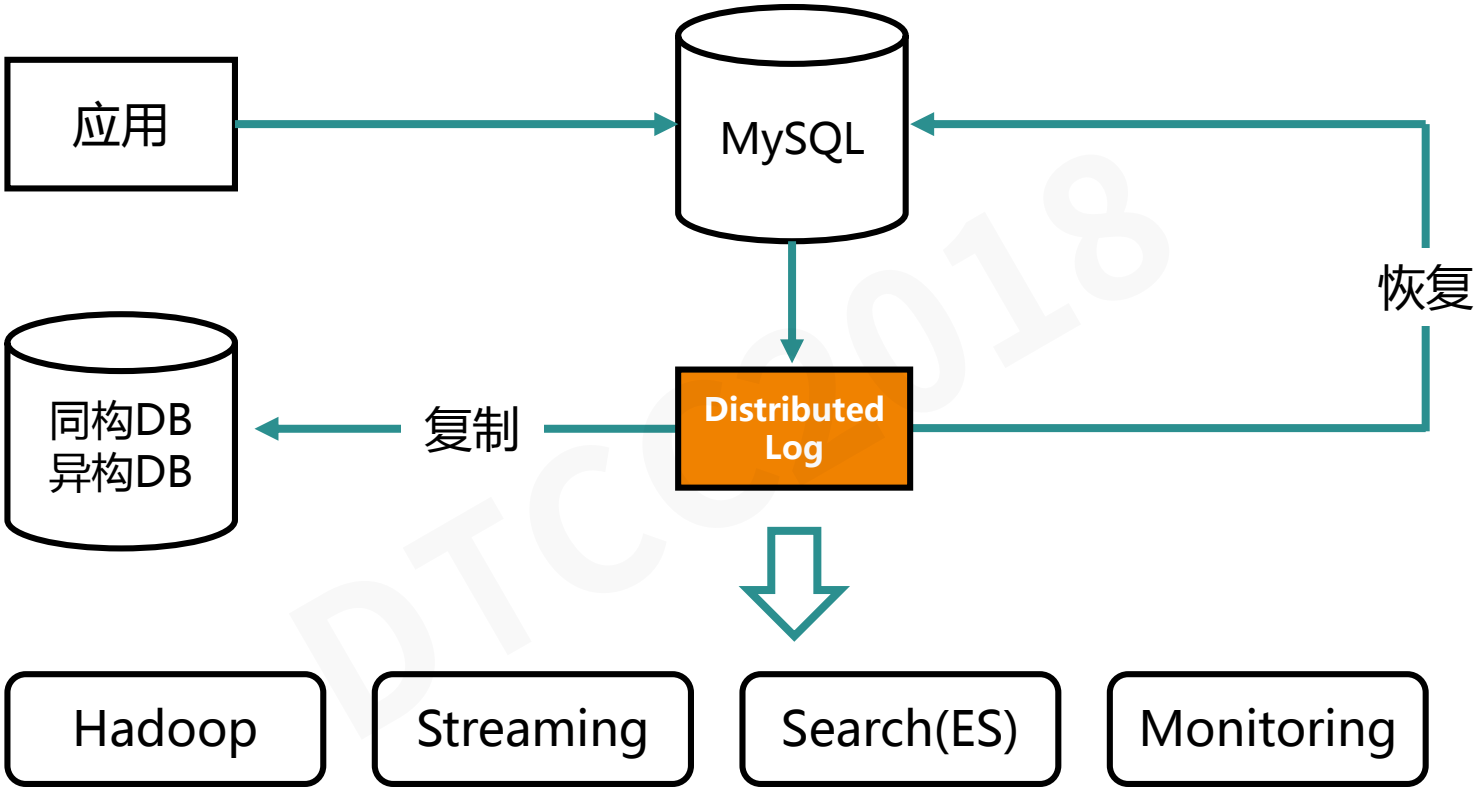


Passive Replication (Primary-Backup)

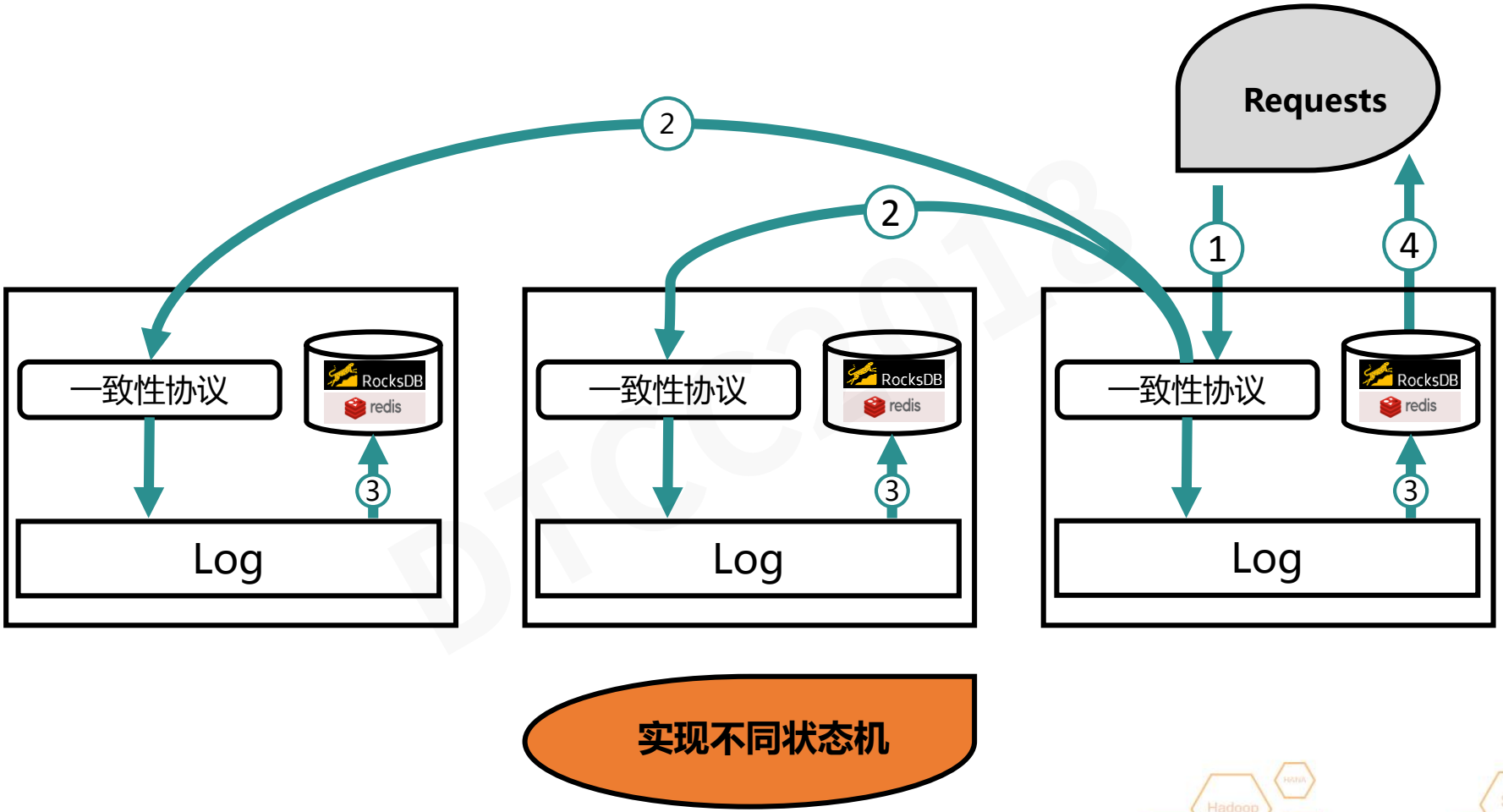


Active Replication (State Machine Replication)

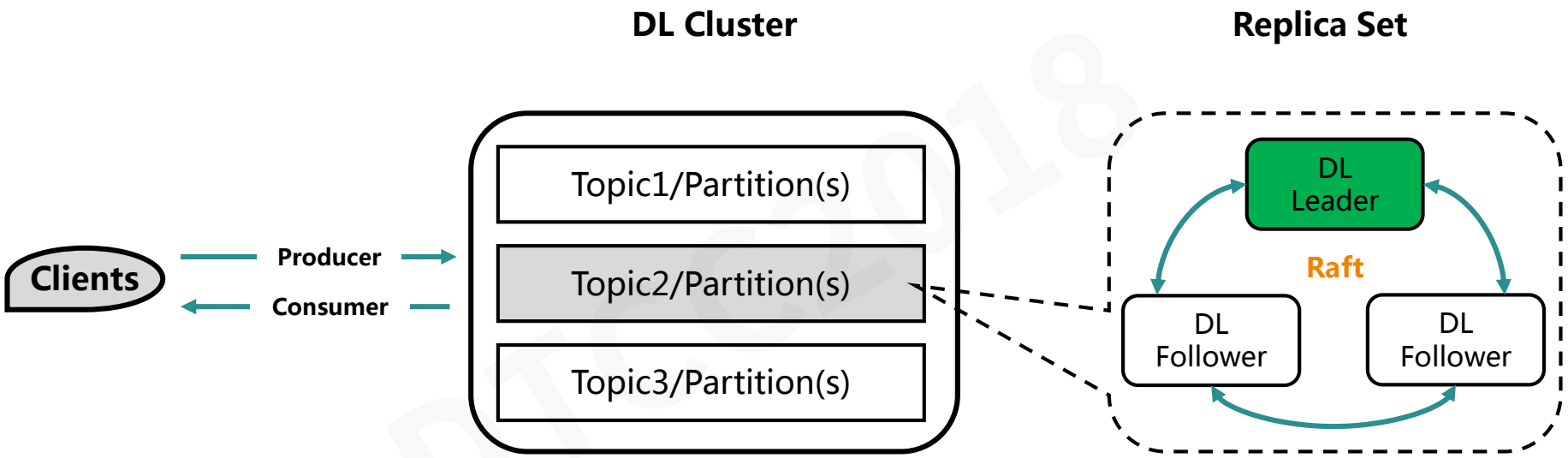
诉求1应对方案：Log-Oriented架构



诉求2应对方案：RSM


Requests
1
2
2
4
一致性协议
RocksDB
redis
3
Log
一致性协议
RocksDB
redis
3
Log
一致性协议
RocksDB
redis
3
Log
实现不同状态机

分布式Log存储架构抽象



关键实现1：分布式一致性

Viewstamped Replication

Paxos（Multi-Paxos, Fast-Paxos, Mencius, Epaxos…）

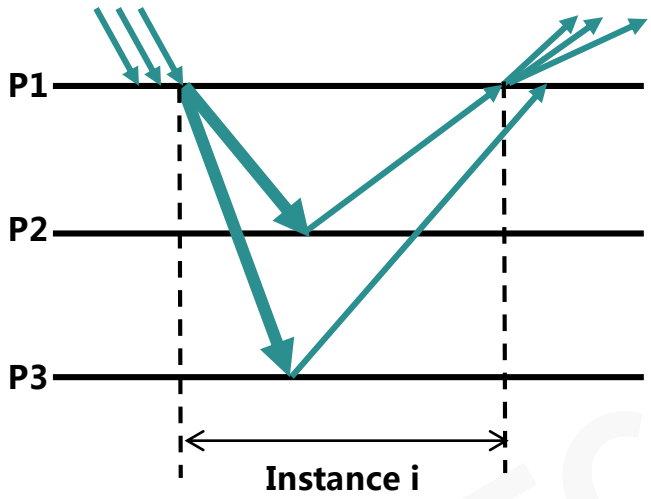
Zookeeper ZAB

Raft

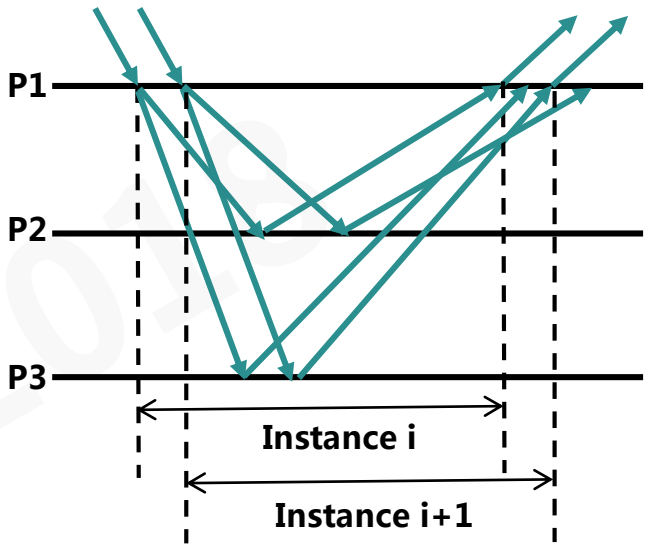
工程化完整程度高

贴近公司etcd等其它组件

关键实现2：强一致性能1/2

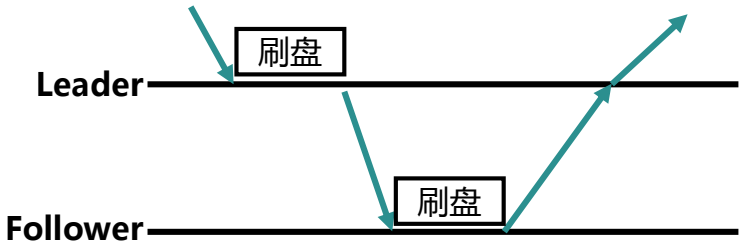


Batch

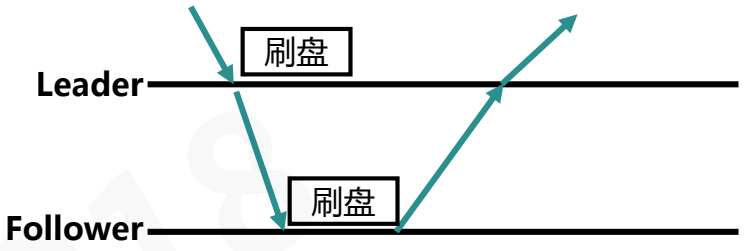


Pipeline

关键实现2：强一致性能2/2



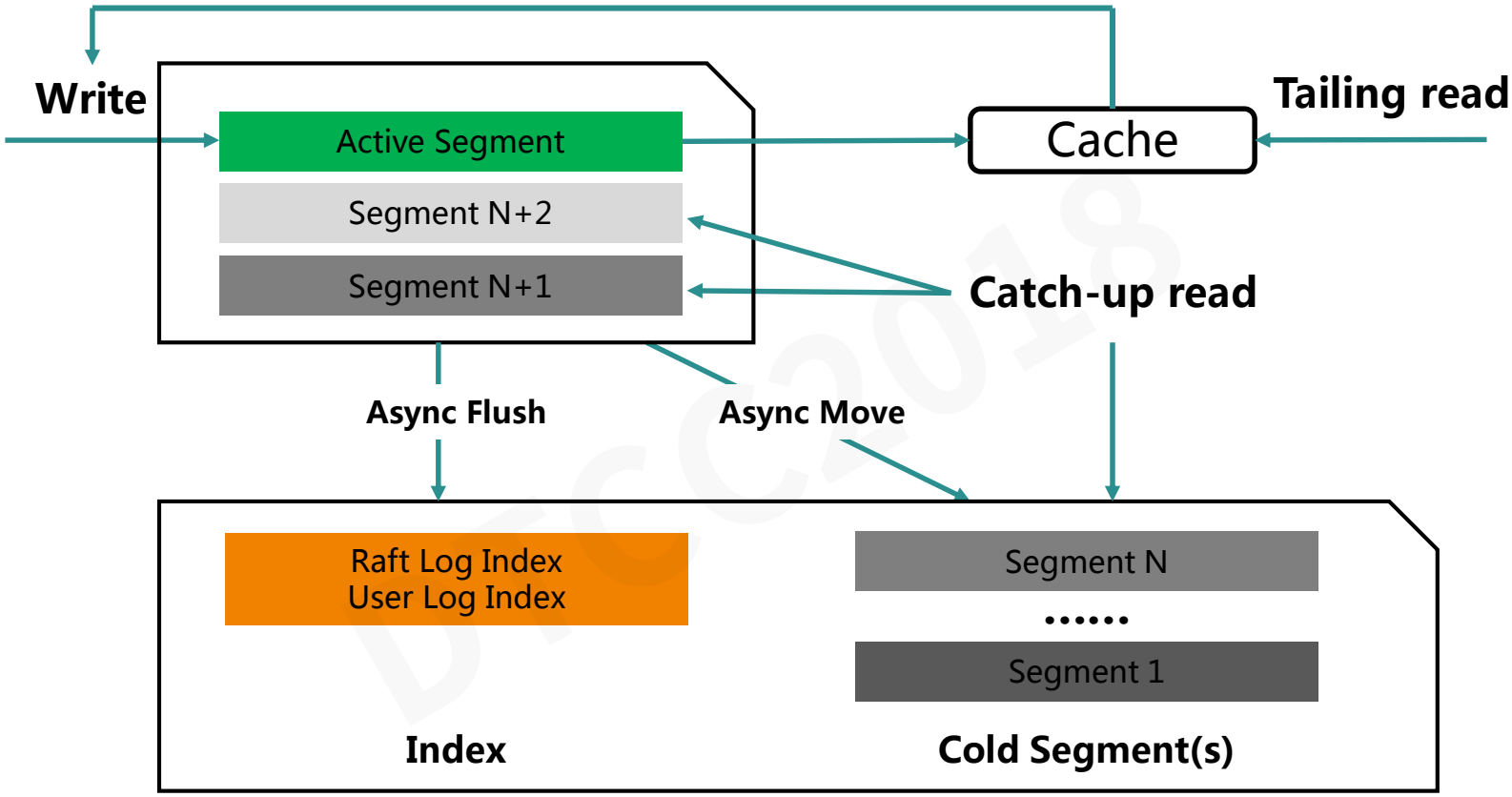
一次提交估算： 1RTT + 2刷盘



一次提交估算： 1RTT + 1刷盘

Pipeline优化

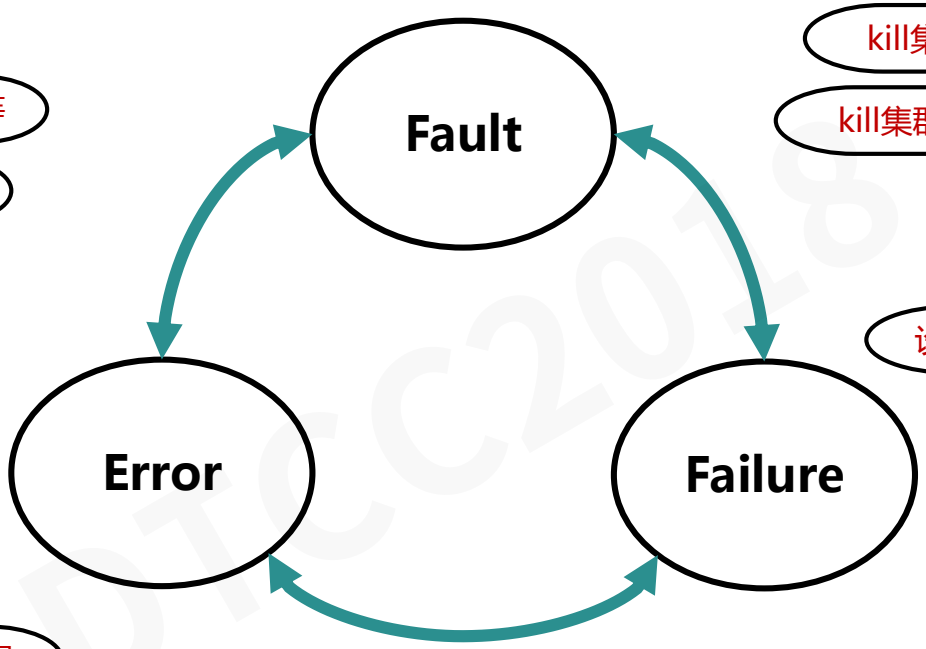
关键实现3：IO放大 & 隔离



写放大：Log as Data

隔离：读/写通过不同磁盘提供IO

关键实现4：分布式系统测试1/2



网络分区

随机 - 网络延时 / 丢包等

定点 - 网络延时 / 丢包等

随机kill一个APP进程

kill Leader APP进程

kill集群中多数派APP进程

kill集群中所有的APP进程

设置CPU亲缘性 / 模拟CPU压力

节点时钟设置

随机 - 磁盘I/O延时 / 错误

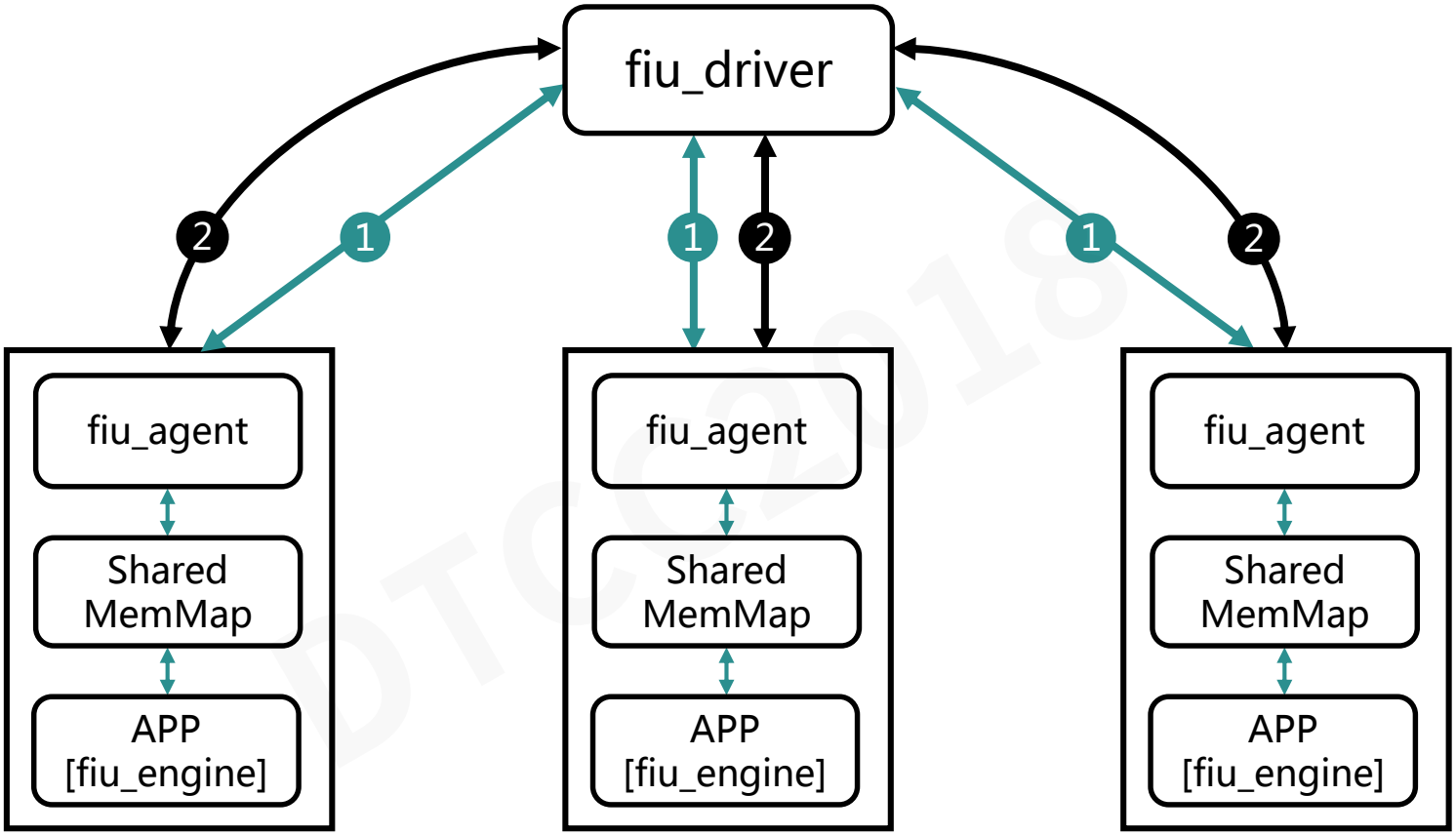
定点 - 磁盘I/O延时 / 错误

设置磁盘I/O调度的策略和优先级

异构节点（异构集群），shell脚本在线更改

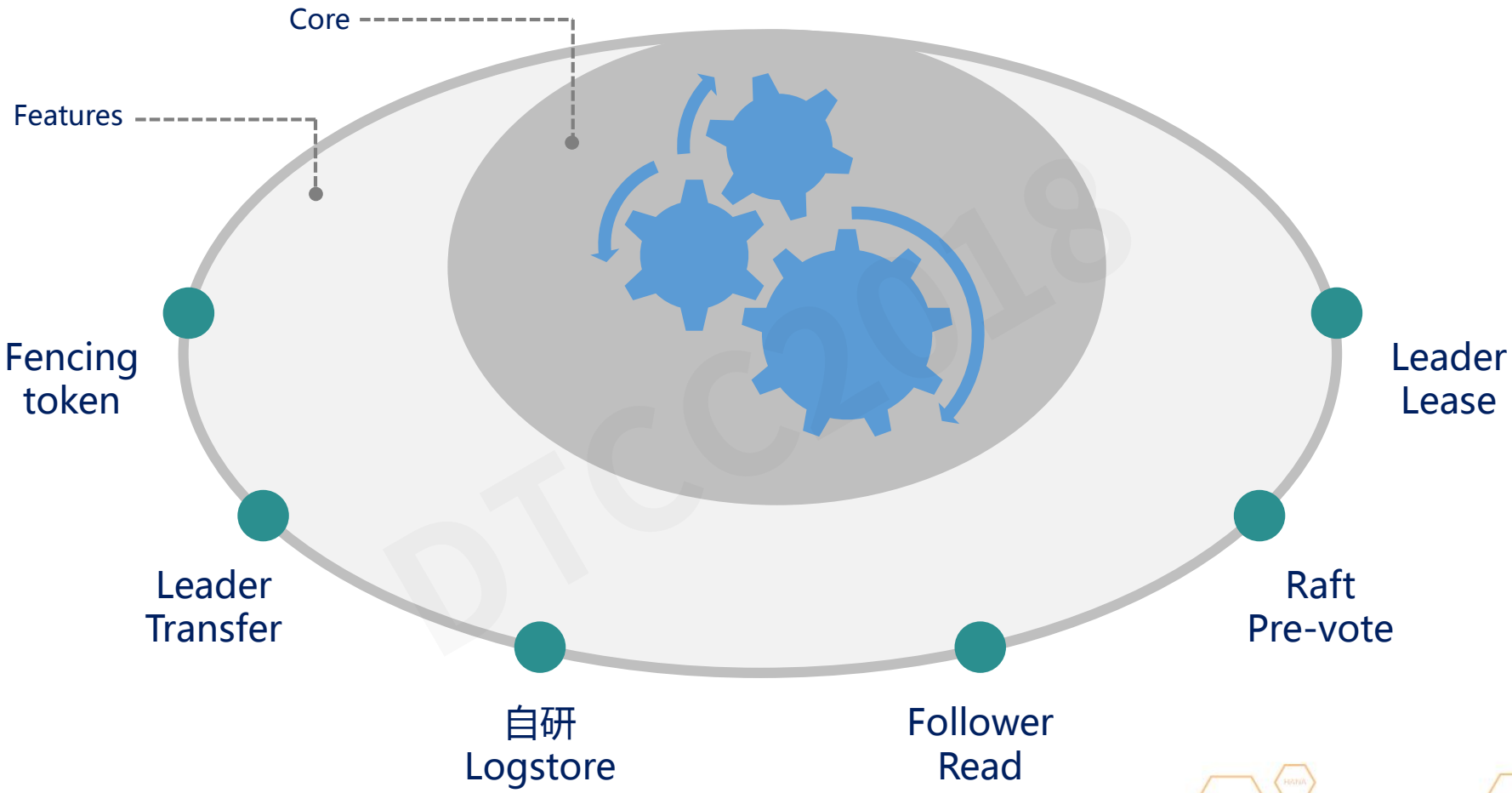
错误和失败注入：Enable代码中某些同步点（代码中已插入同步点）

关键实现4：分布式系统测试2/2



fiu = **F**ault **I**njection **U**tility

关键实现5：非常有意义的工程特性



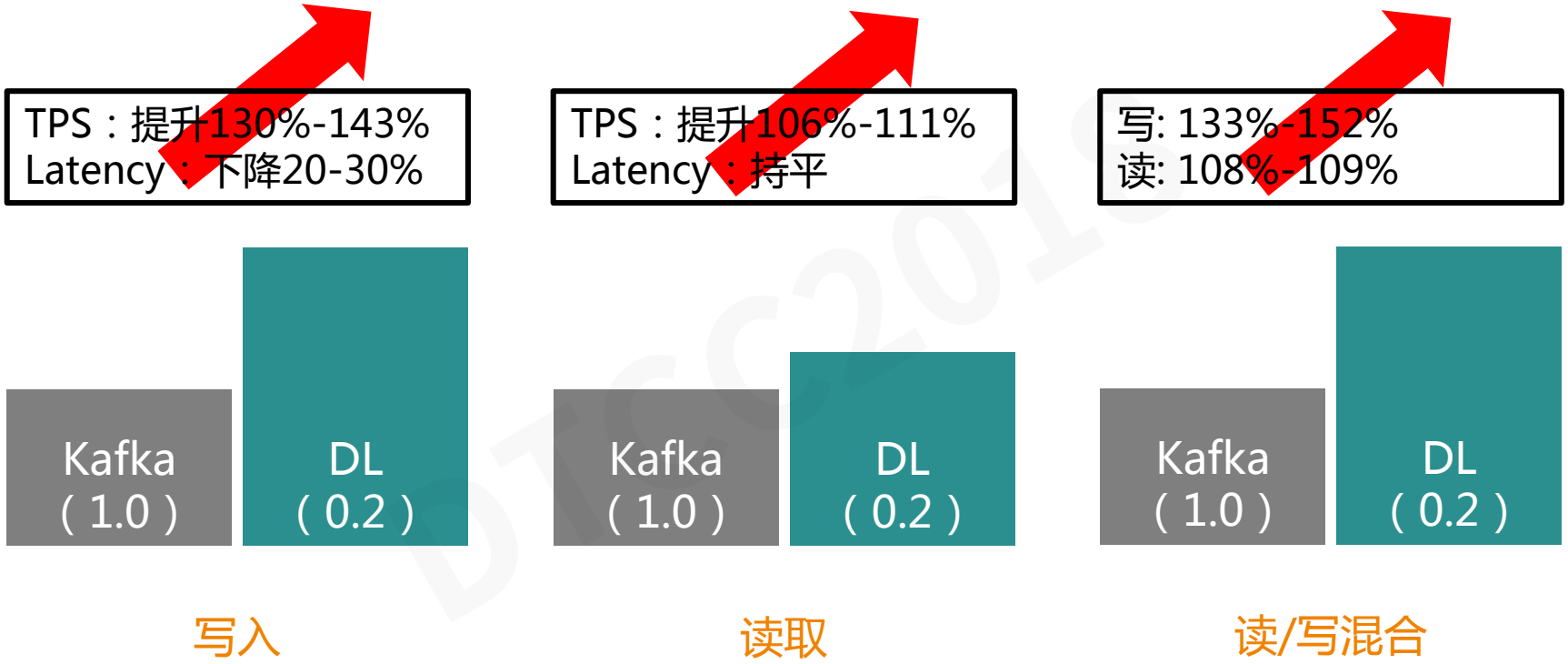
实测环境

CPU： 32 核 Intel (R) Xeon (R) CPU E5-2620 v4 @ 2.10GHz
Mem：32GB
磁盘： STAT 512GB
网卡： 1000Mb

兼容Kafka协议：
default.replication.factor=3
min.insync.replicas=2
log.flush.interval.messages=1

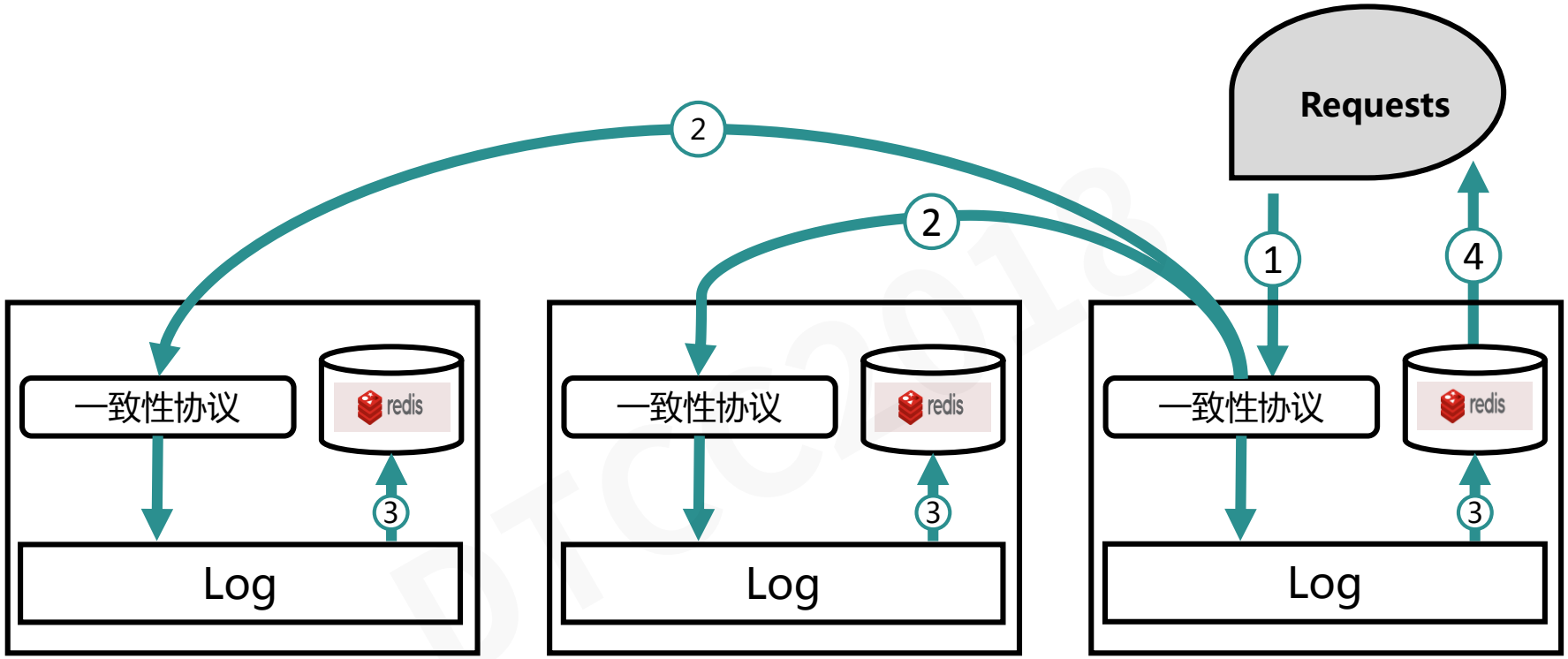
消息大小分为 500Byte 和 1024Byte，写消息的producer 的 batch.size 为16KB，同步发送，ack为all模式。读消息的consumer的fetch.message.max.bytes大小为 1MB。

实测性能



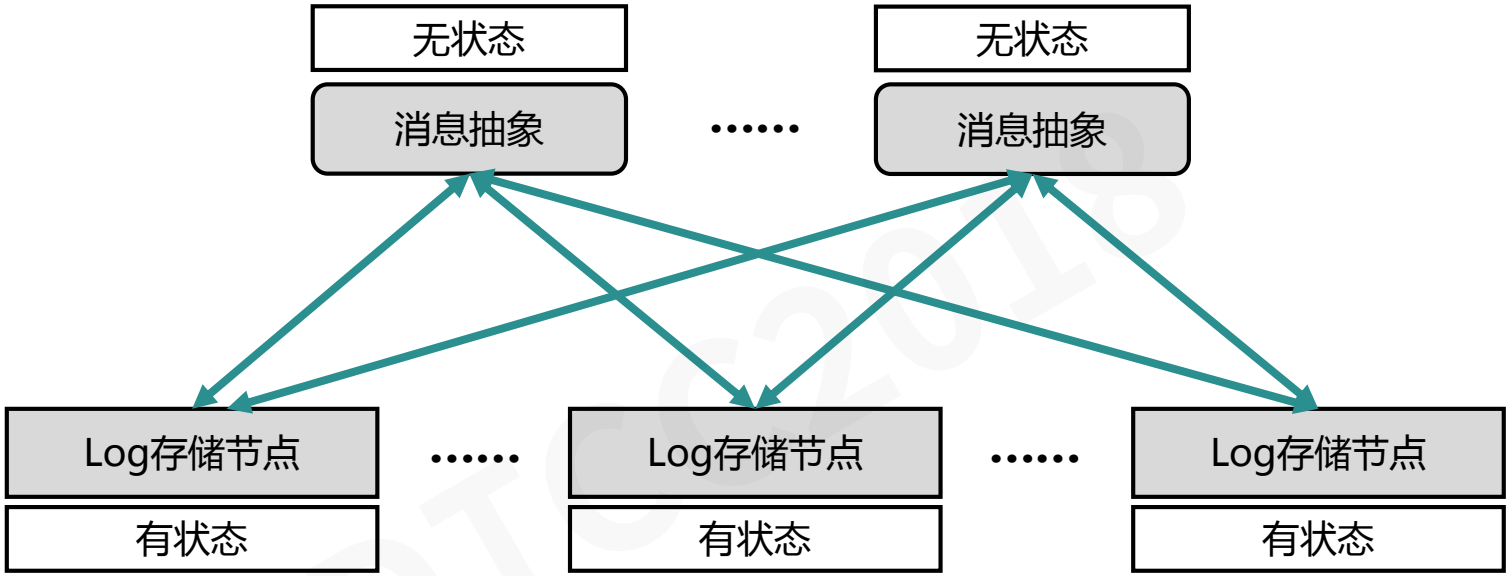
研发背景 设计思路 关键实现 应用形态

应用形态1：分布式Cache



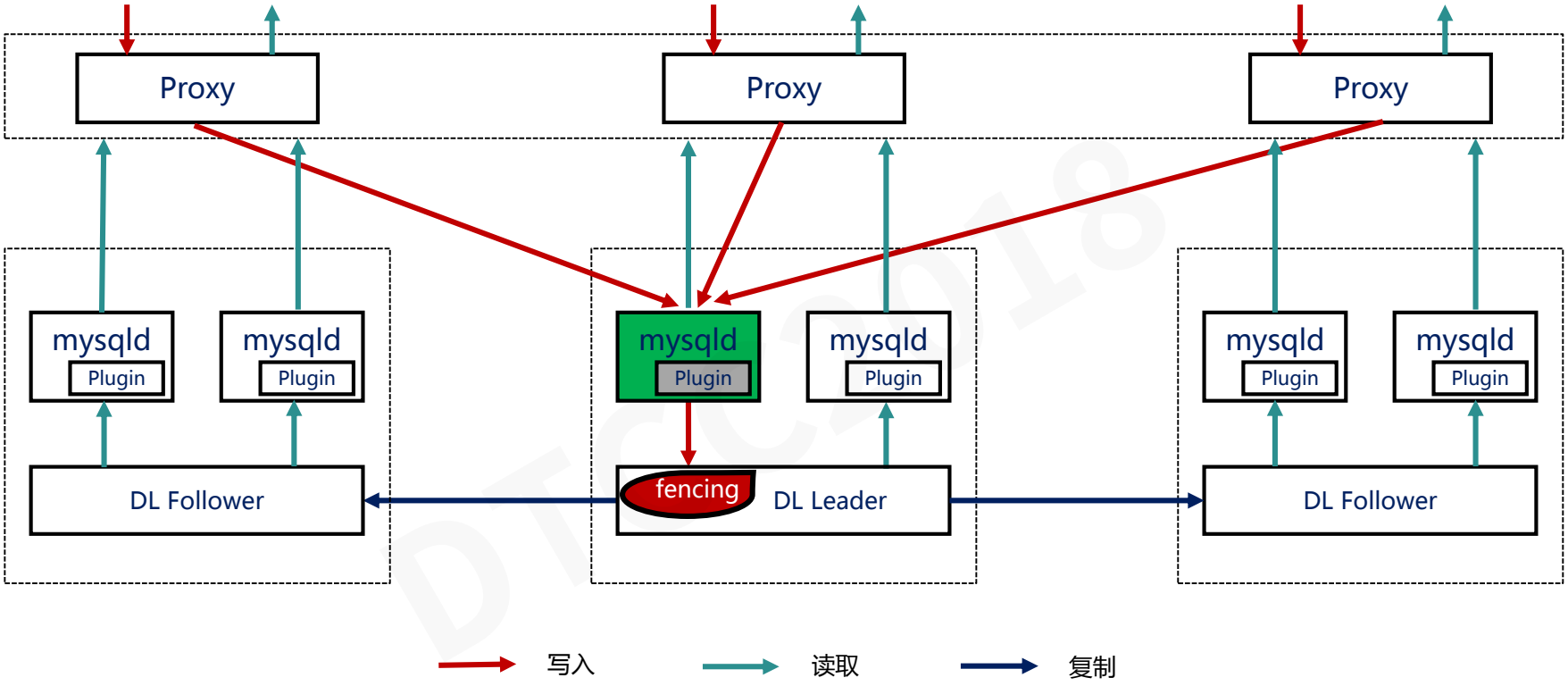
难点：借助于redis的bgsave实现snapshot功能
其他思路：多个state machine，其中一个专门产生snapshot

应用形态2：分布式MQ



计算（消息处理）和存储（Log存储）分离，不同模块可以独立Scale
难点：有状态的存储节点Scale out

应用形态3：Database State Machine



强一致的MySQL自治集群
难点：Binlog如何高效Replay到Slave；Binlog多点冗余；Proxy的无状态等

Q & A

邮箱: jianhuaibing@139.com
微信: huaibingjian

THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigDate
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168