China · Beijing

ArchSummit全球架构师峰会北京站

姓名 许家滔

Title 微服务在微信的架构实践

ArchSummit 全球架构师峰会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ

China · Beijing

微服务在微信后台的架构实践

姓名 许家滔

Title 微服务在微信的架构实践

ArchSummit
全 球 架 构 师 峰 会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ

自我介绍

- 许家滔（sunnyxu@tencent.com）
- 微信技术架构部，后台开发中心
- 基础微信业务 / 基础服务 / 基础组件
- 负责微信后台系统，一直专注于布式存储平台和后台服务框架等
- 开源 libco
  github.com/tencent/libco

大纲

第一部分 背景与概览

第二部分 基础架构

第三部分 监控与测试

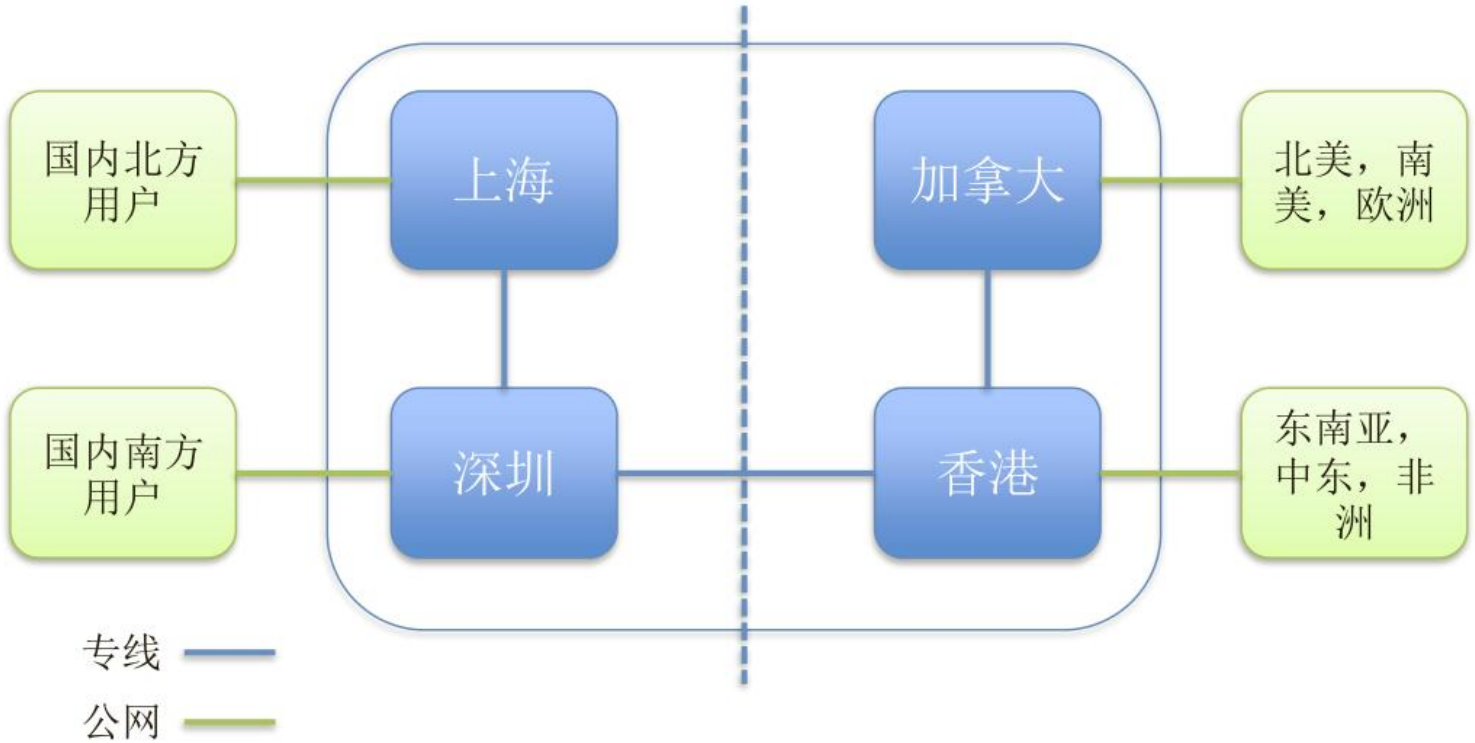
第四部分 跨业务服务关系

背景与概览

- 敏捷开发
- 统一的基础架构与运营管理
- 千级模块的庞大系统
  基础微信/公众平台/微信支付等等
  数万服务器

微信的后台架构与微服务理念有很多相通之处，尝试从微信后台架构出发描述互联网后台的微服务实践

背景与概览

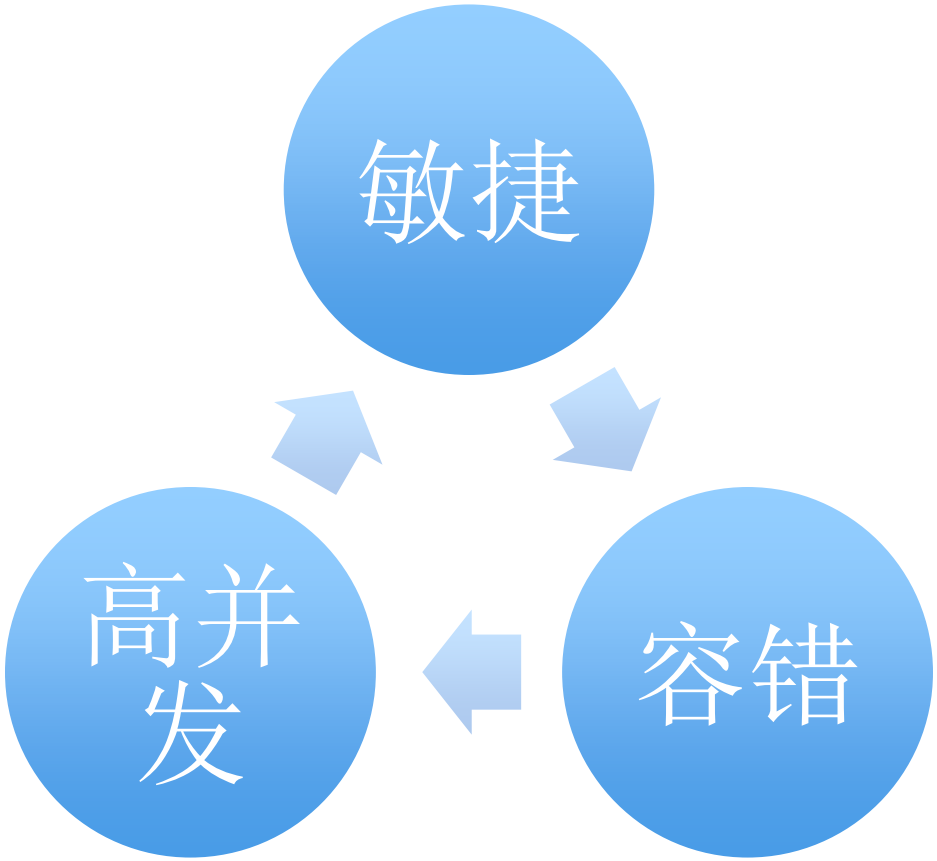


背景与概览


产品
帐号/联系人/消息
朋友圈
公众平台/支付/广告...
接入
长连接
短连接
逻辑
逻辑服务（推送/在线/具体逻辑）
异步任务队列
RPC / 自动分SET / 过载保护
数据
数据操作层
异地数据队列
PaxosStore存储平台
Seq
冷备
监控
统计
配置
OSS
发布

微信的海量场景下

微服务架构需要怎样的能力？


敏捷
容错
高并发

微服务基础架构

服务框架

- 服务布局（多城市/园区管理）
- 远程调用（c++大规模协程使用，protobuf，异步队列 ）
  https://github.com/tencent/libco
- 容错 （ 重试策略，过载保护，负载均衡，柔性可用 ）
- 部署管理，发布系统

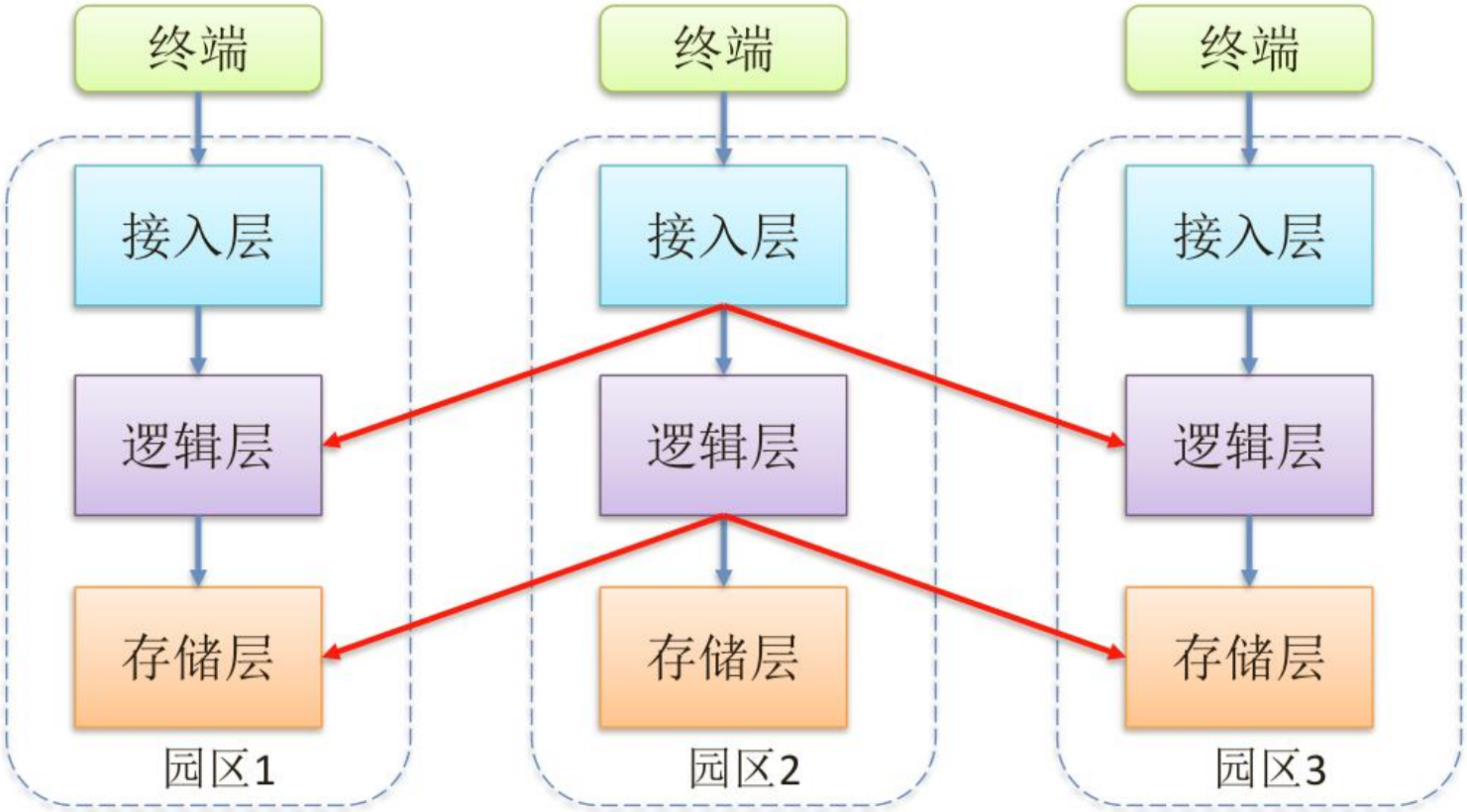
服务布局

多地自治，园区互备

城市间后备的公网udp通道

服务布局

多地自治，园区互备



远程调用

当你掏出手机微信支付，背后会引发多少次RPC？

细腻的业务逻辑，高扇出的RPC调用

事件驱动的异步编程

callback / stat machine / future promise

如何重用，如何敏捷？

https://github.com/tencent/libco

远程调用 / libco

极简

一行代码10块钱的故事

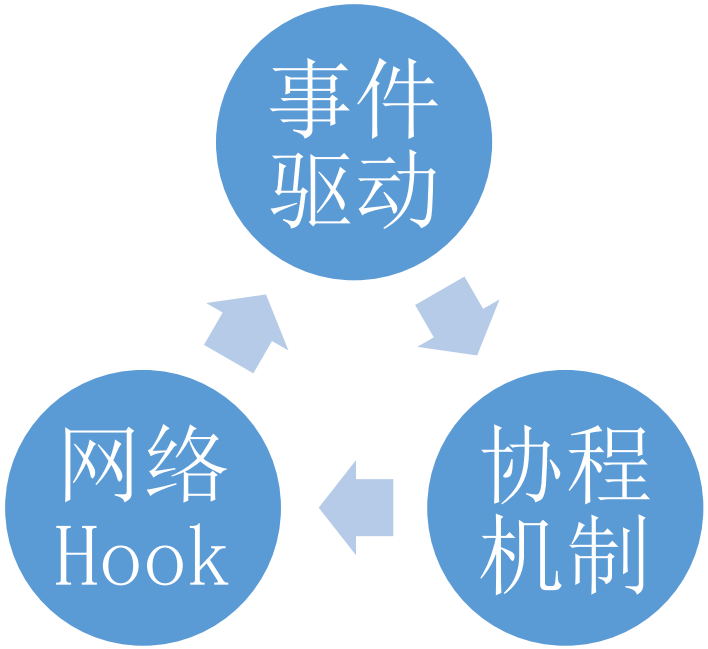
完备

时间轮盘与事件处理链

常用网络编程模式

同步原语

…



https://github.com/tencent/libco

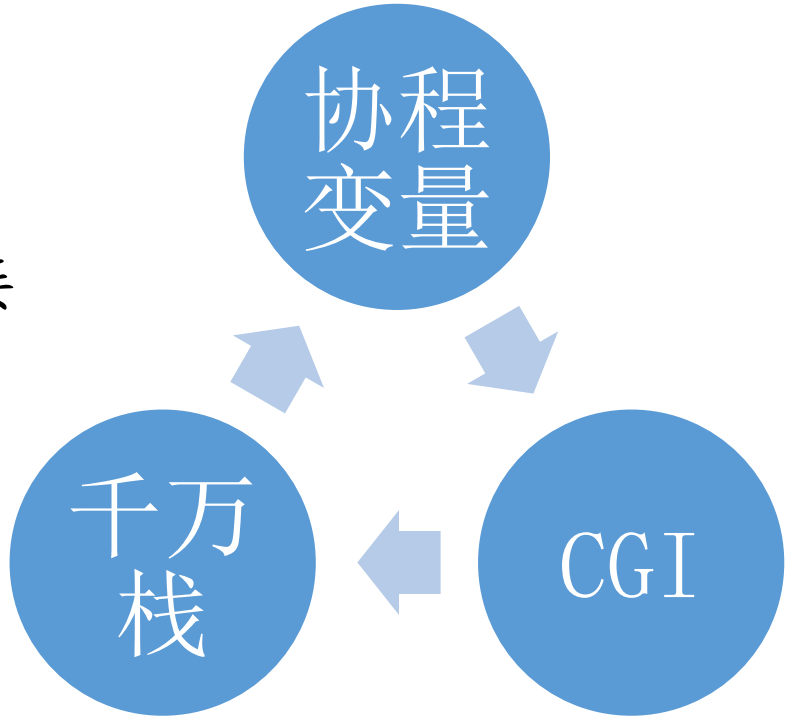
远程调用 / libco

协程变量

支持标准CGI开发

支持gethostbyname

高级用户可以开启千万连接

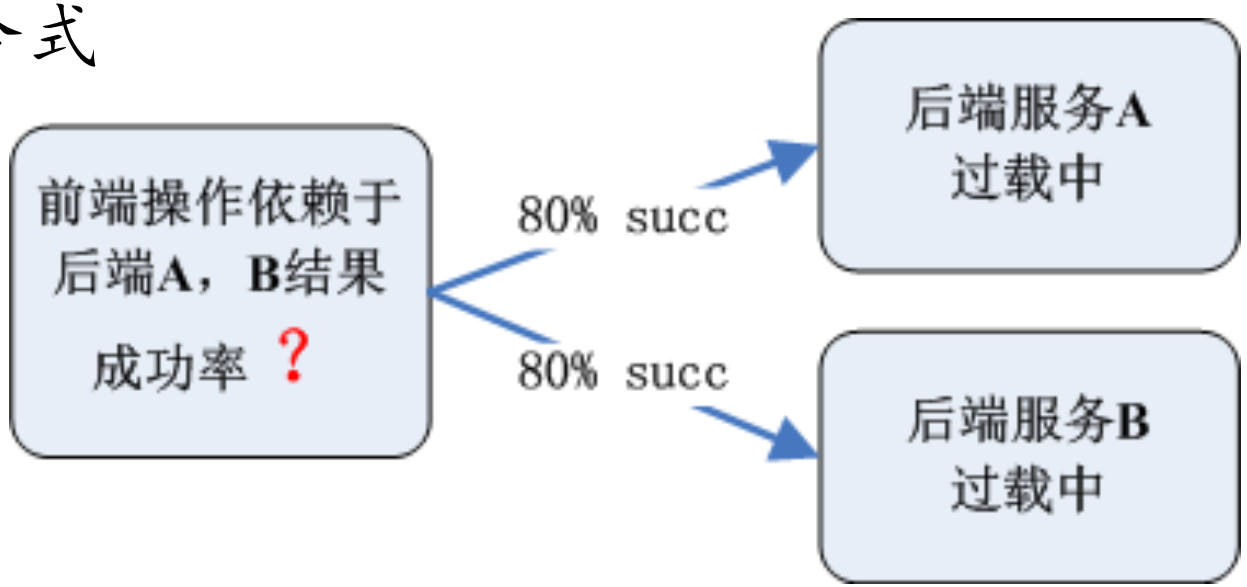


https://github.com/tencent/libco

微服务基础架构

过载保护

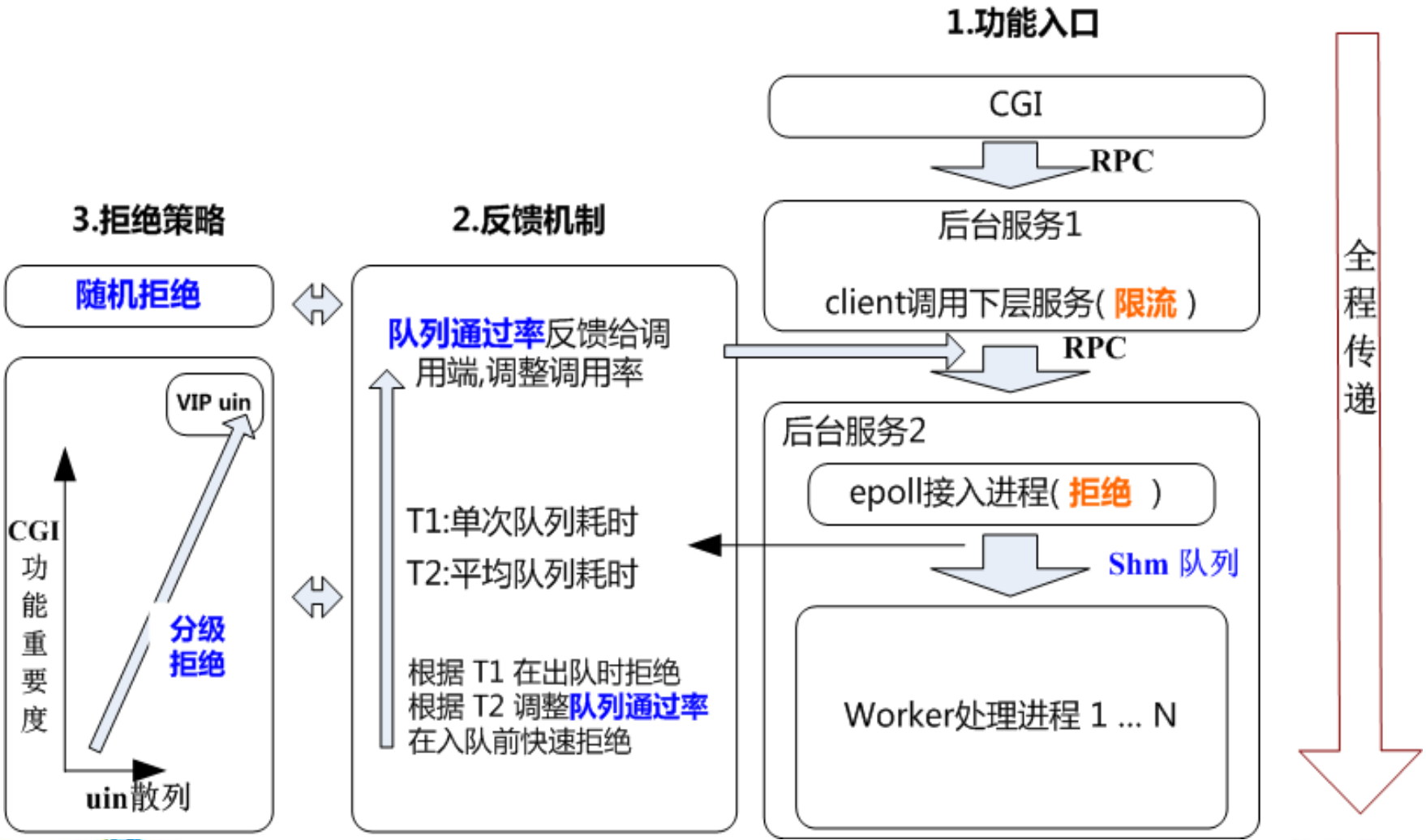
- 轻重分离
- 队列式
- 组合命令式



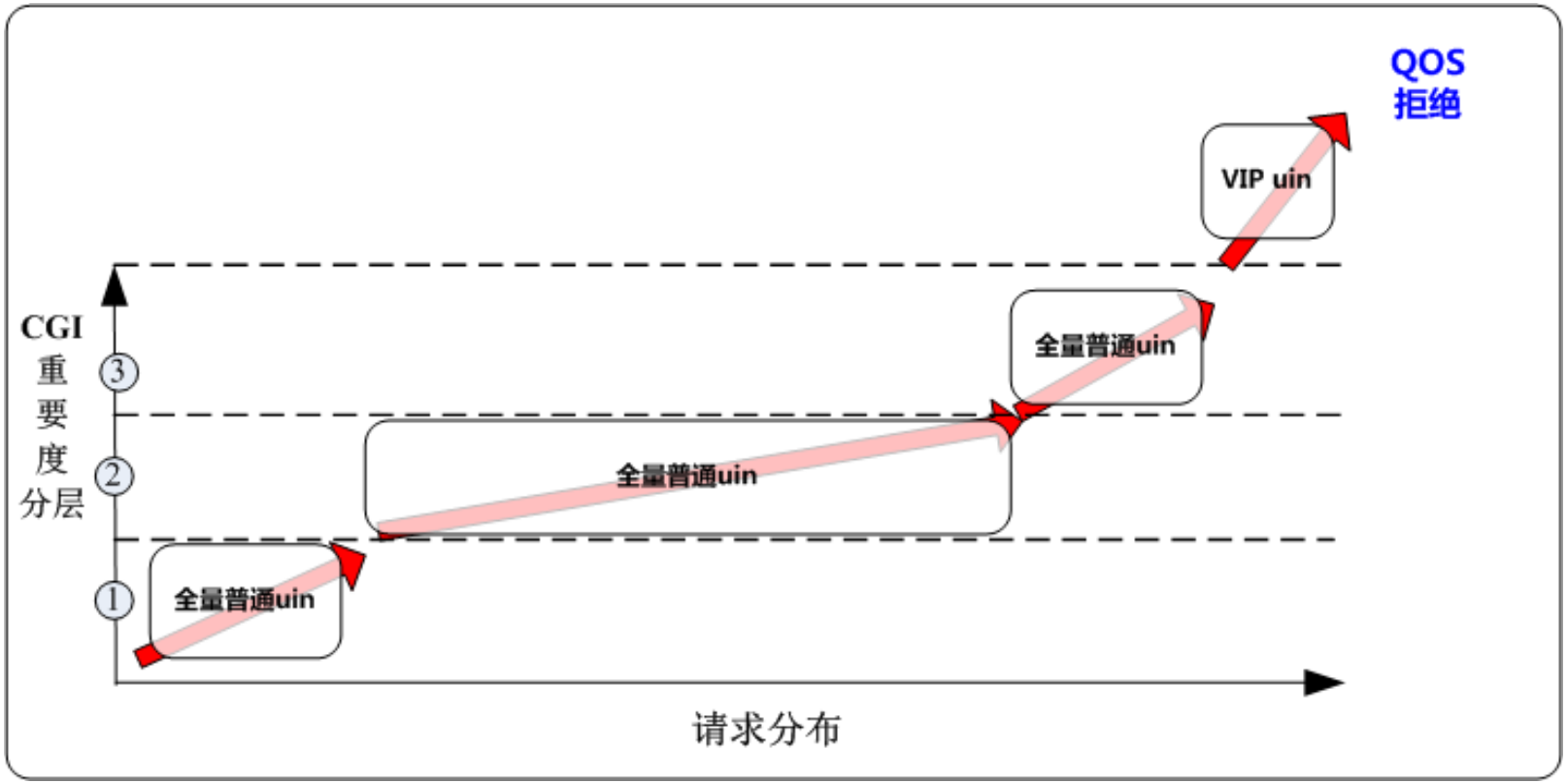
过载保护的核心是反馈

推荐《失控》

微服务基础架构 / 过载保护



微服务基础架构 / 过载保护



微服务基础架构

数据存储

- 一致性的权衡
- PaxosStore
- 异步队列
- 其他

微信PaxosStore

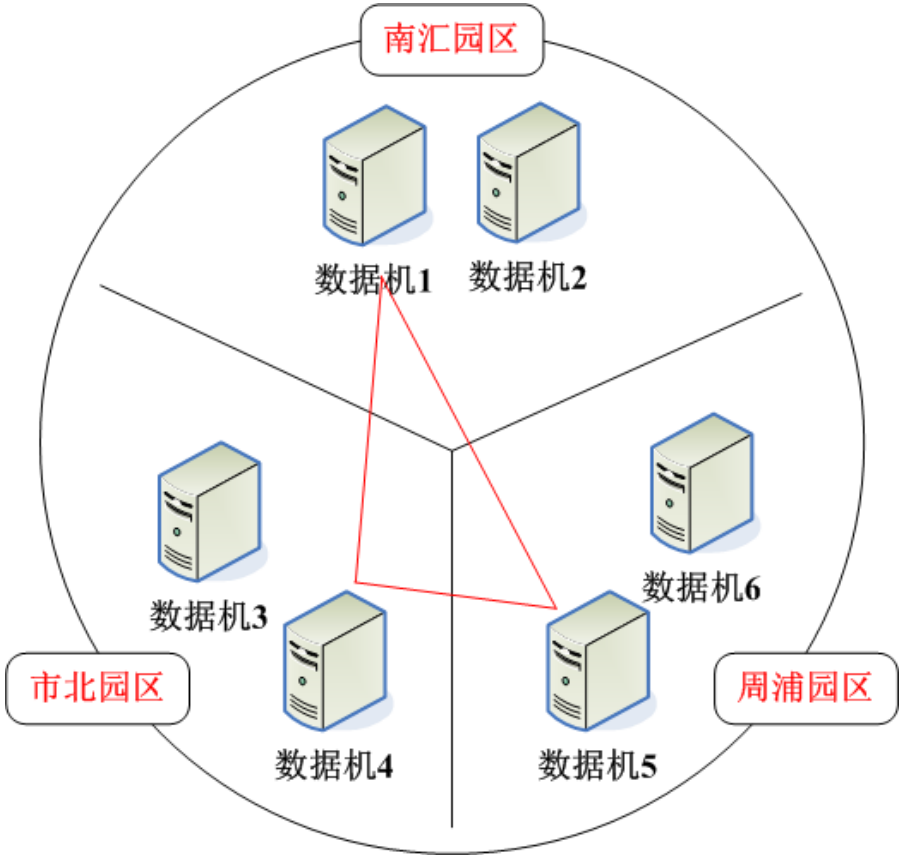
- 同步复制，多主多写（园区间数据一致）
- 支持键值/列表等数据结构
- 支持表格，单表亿行，并可执行类SQL事务
- 插件化设计，统一的分布框架下面定制的存储引擎
- 低成本，灵活组合不同的资源（内存/ssd/sata）
- 数千台机器组成的集群

微信PaxosStore

如何设计数据容灾以获得自由切换的能力
- 以同城为大前提的低延迟环境
- 非租约Paxos

负载均衡
- 单机/园区分级容灾，kv64架构
- 单机故障分摊25%流量
- 园区故障分摊33%流量

SET部署,一致性哈希均匀分布
- 具备细粒度的伸缩性



为何需要这么重的方式？

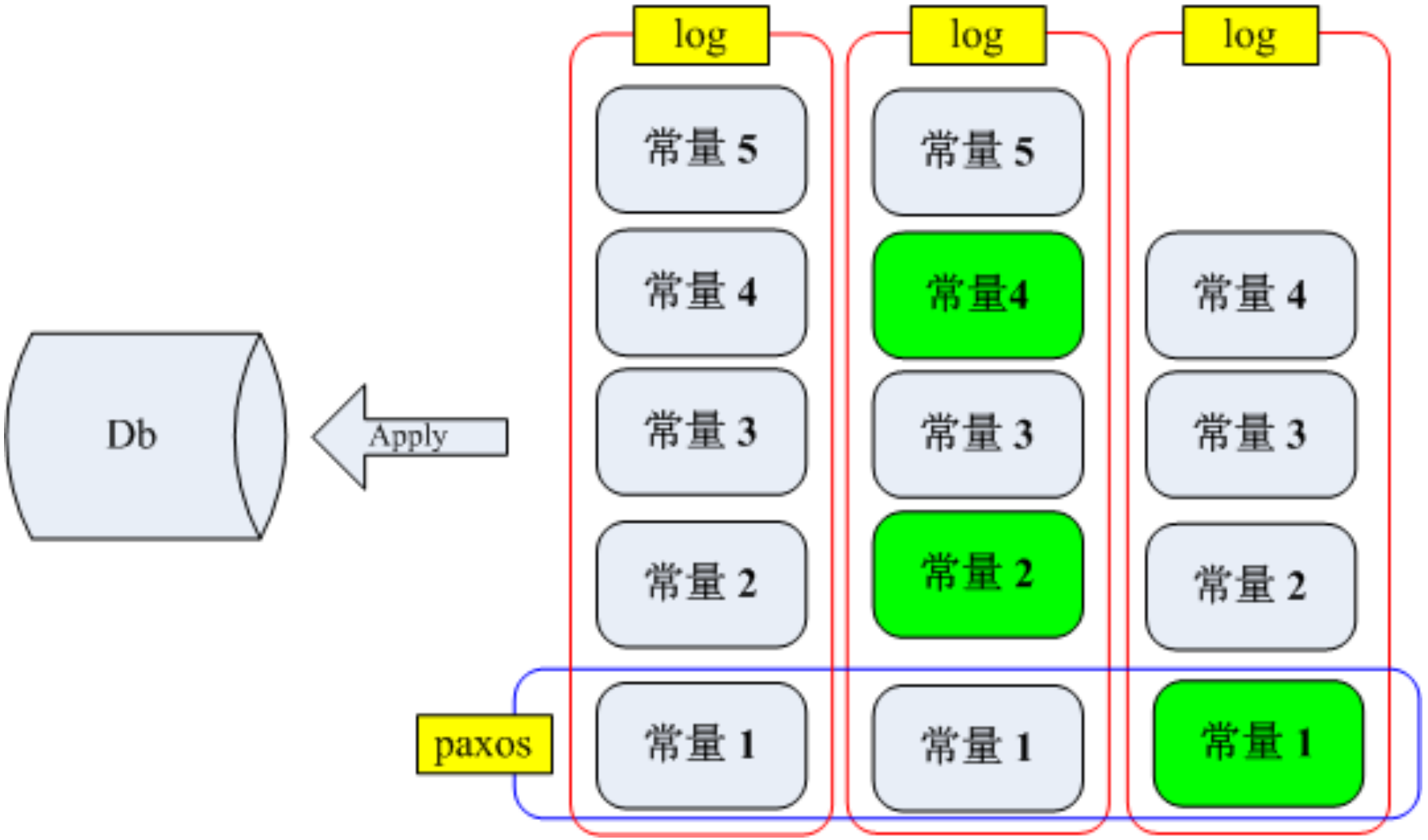
应用简单快速

数据容灾

服务可用性

……

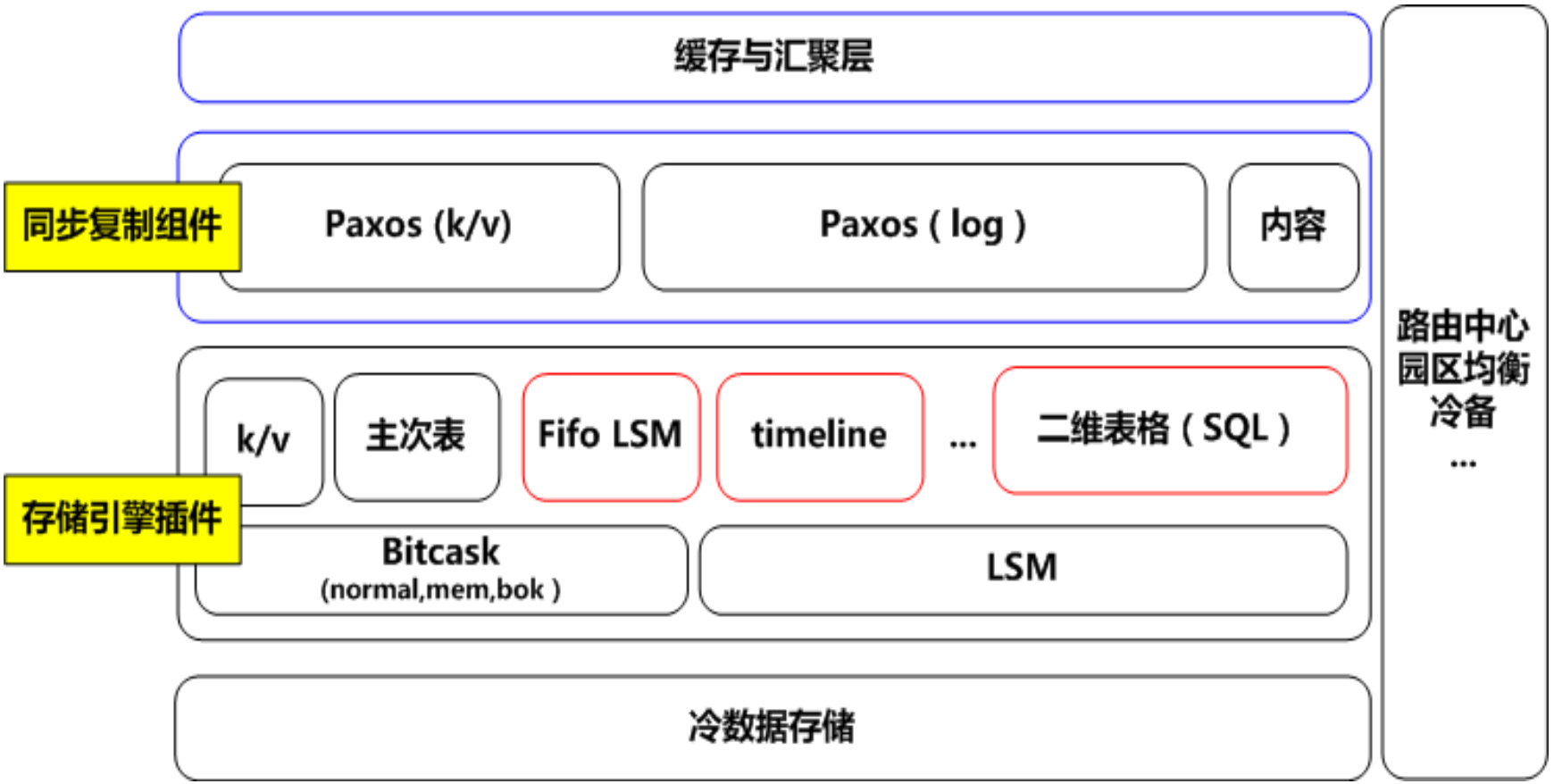
Raft ?


log
log
log
常量 5
常量 5
常量 4
常量4
常量 4
常量 3
常量 3
常量 3
常量 2
常量 2
常量 2
Db
Apply
paxos
常量 1
常量 1
常量 1

微信PaxosStore

- Paxos工程实现在海量服务上的权衡与创新
  - 多主多写（去租约无leader模型）
    - 过载保护
    - 无切换不可用时间
  - 隔离性：细粒度，单机kw+的Paxos(log)实例
  - 高性能：协议交互优化

微信PaxosStore / 系统架构

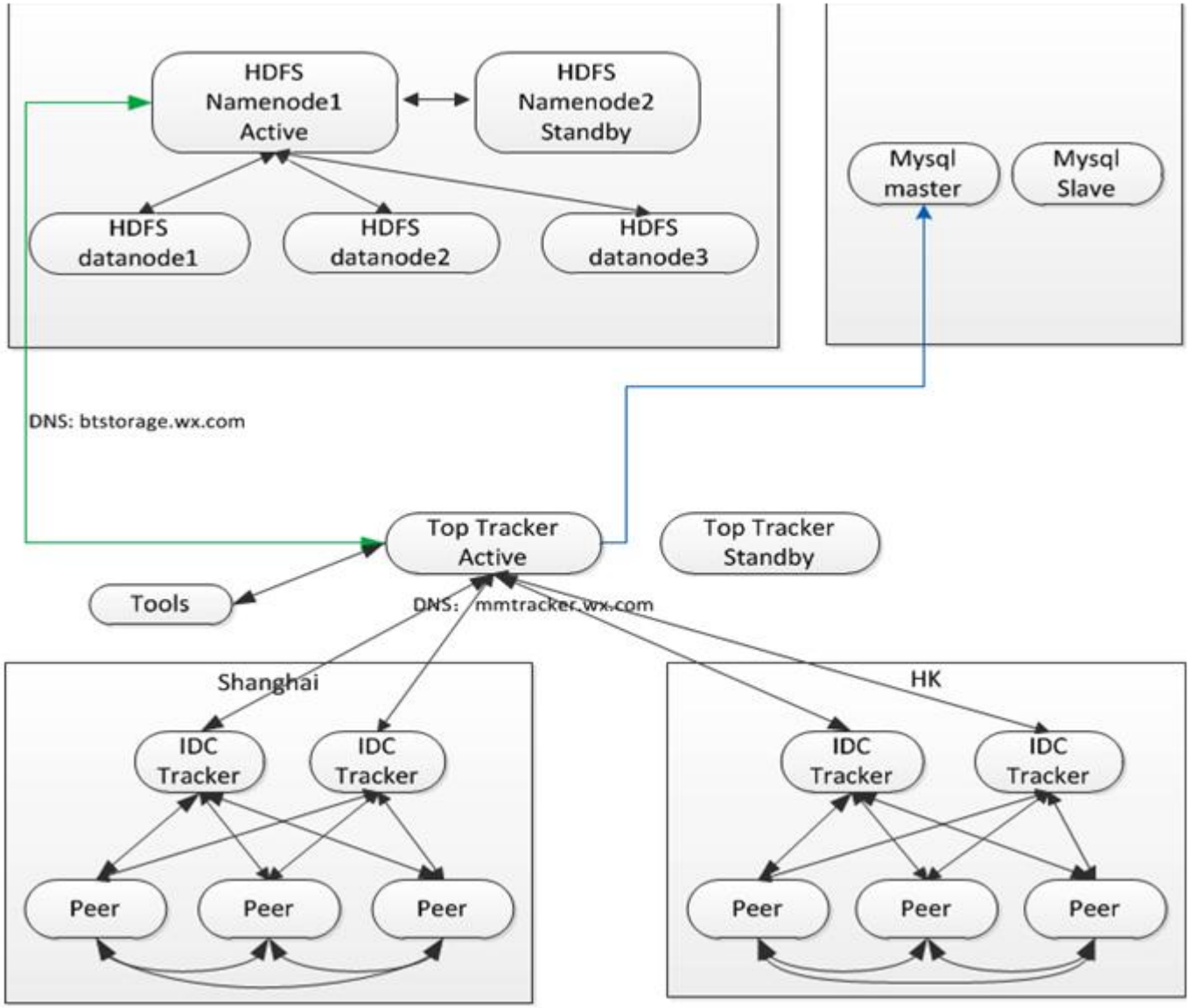


基础架构

发布系统

- 三军未动，粮草先行
- 基于**BitTorrent**的传输方案
- Facebook/Twitter/EBay

发布系统

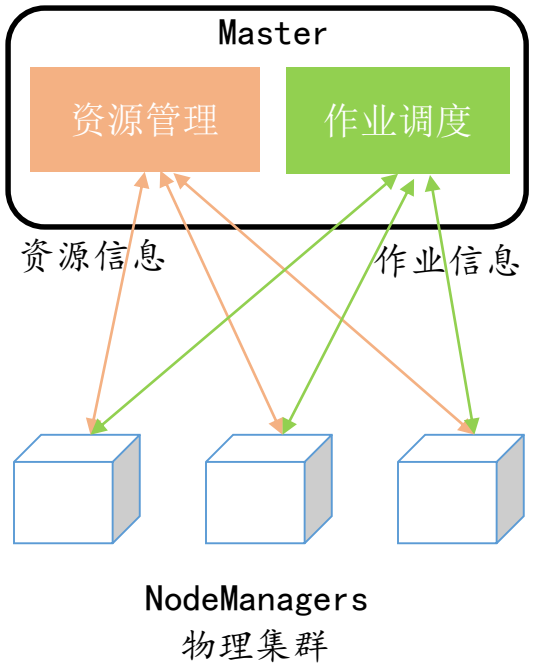

HDFS
Namenode1
Active
HDFS
Namenode2
Standby
HDFS
datanode1
HDFS
datanode2
HDFS
datanode3
Mysql
master
Mysql
Slave
DNS: btstorage.wx.com
Top Tracker
Active
Top Tracker
Standby
Tools
DNS: mmtracker.wx.com
Shanghai
HK
IDC
Tracker
IDC
Tracker
IDC
Tracker
IDC
Tracker
Peer
Peer
Peer
Peer
Peer
Peer

基础架构

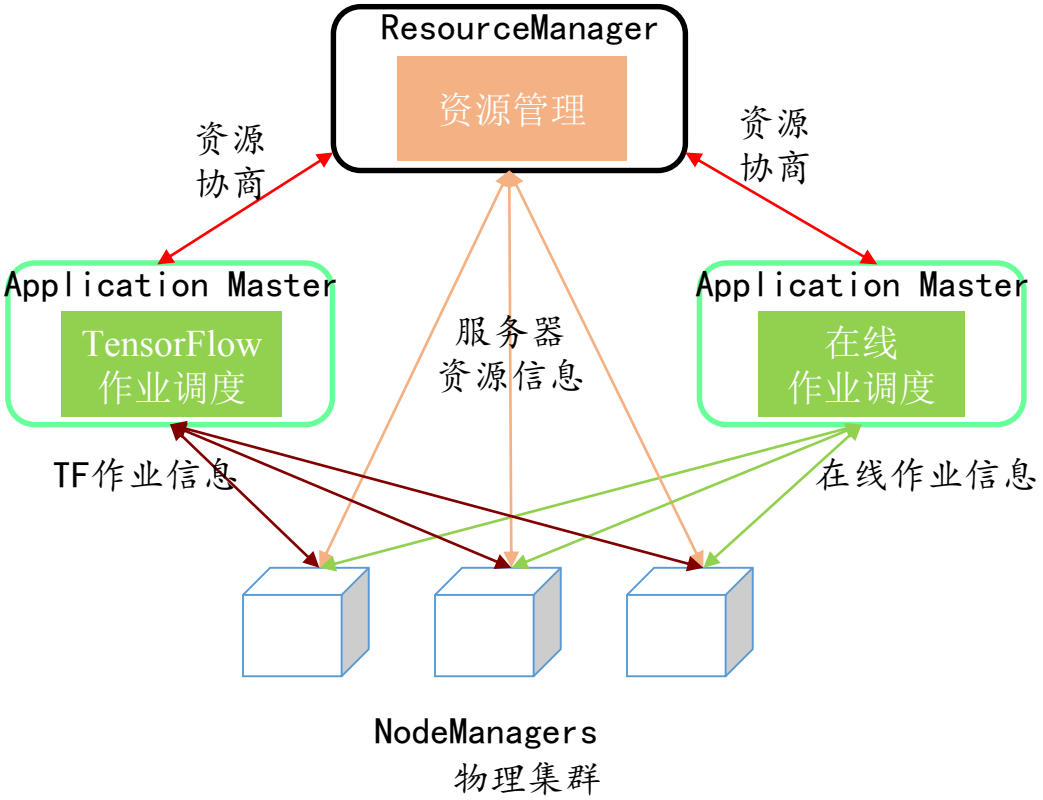
容器与集群管理

- Cgroup
- Docker
- 微信Yard

微信Yard



中央调度器
（JobTracker，Torca）



双层调度器
（Mesos，Yarn）

基础架构

需求

高可靠、高可用：完善的容灾能力，能管理微信在线服务

可扩展、伸缩性好：能方便的接入各种不同的离线计算框架

独立性：各个作业资源相互独立，不争用

大规模：能管理数万级别服务器的集群和作业

选择

双层调度器模型

监控与测试

性能监控与故障分析

- 分钟级的调用关系
- 可用性指标
- 手动介入原则

测试方法

- 分布式系统测试代价高
- Mock
- 单机集成

跨业务服务关系

代码管理

- 依赖管理
- 编译系统

服务解耦

- Gateway
- broker

THANKS

ArchSummit
全 球 架 构 师 峰 会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ