



# 打造58**千亿级**分布式存储平台

江军平



# 关于我

## 58集团基础架构部 技术总监

- 分布式存储平台
- 分布式缓存平台
- 微服务框架及支撑平台
- IM即时通信平台
- 核心基础服务

## 腾讯高级架构师

- 微博&微视&视频后端技术负责人

## 微软STC研发工程师



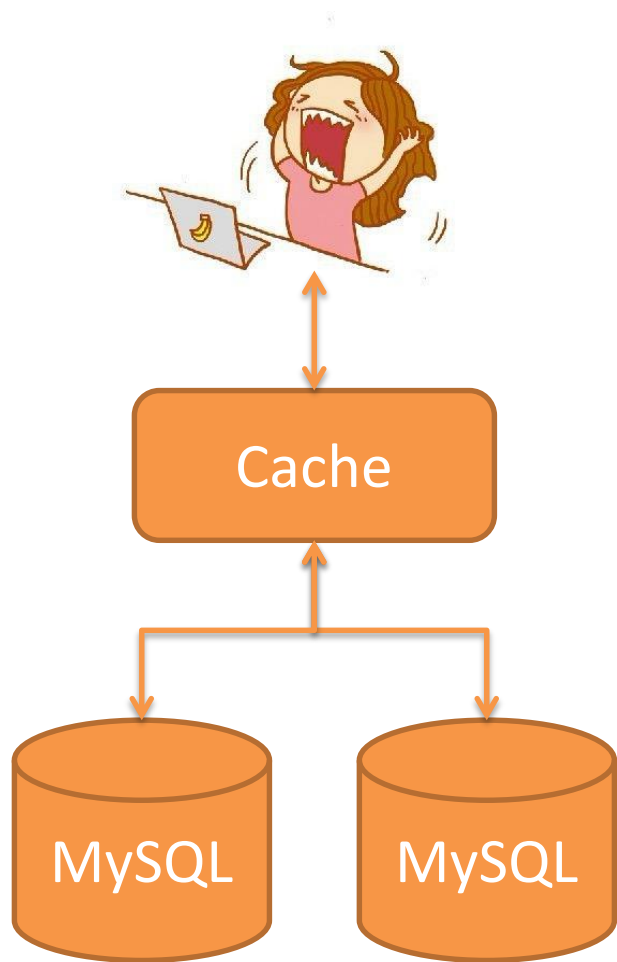
@stevejiang



# Agenda



# 存储架构的痛点和挑战



扩容

性能

扩展性

可用性

规模

- 爆炸式增长

团队

- 小团队 → 超大团队

业务

房产、招聘、二手、.....

Listing Service

列表、详情、发布、登录、用户中心

Core Service

帖子、用户、类目、搜索、推荐、商业、IM

Cache

WCache

WRedis

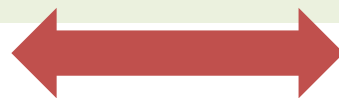
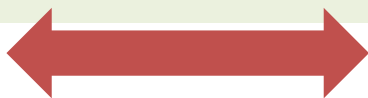
Data

MySQL

WTable

WList

微服务支撑平台



# WTable分布式存储平台

58自研



分布式



NoSQL

Key-Value

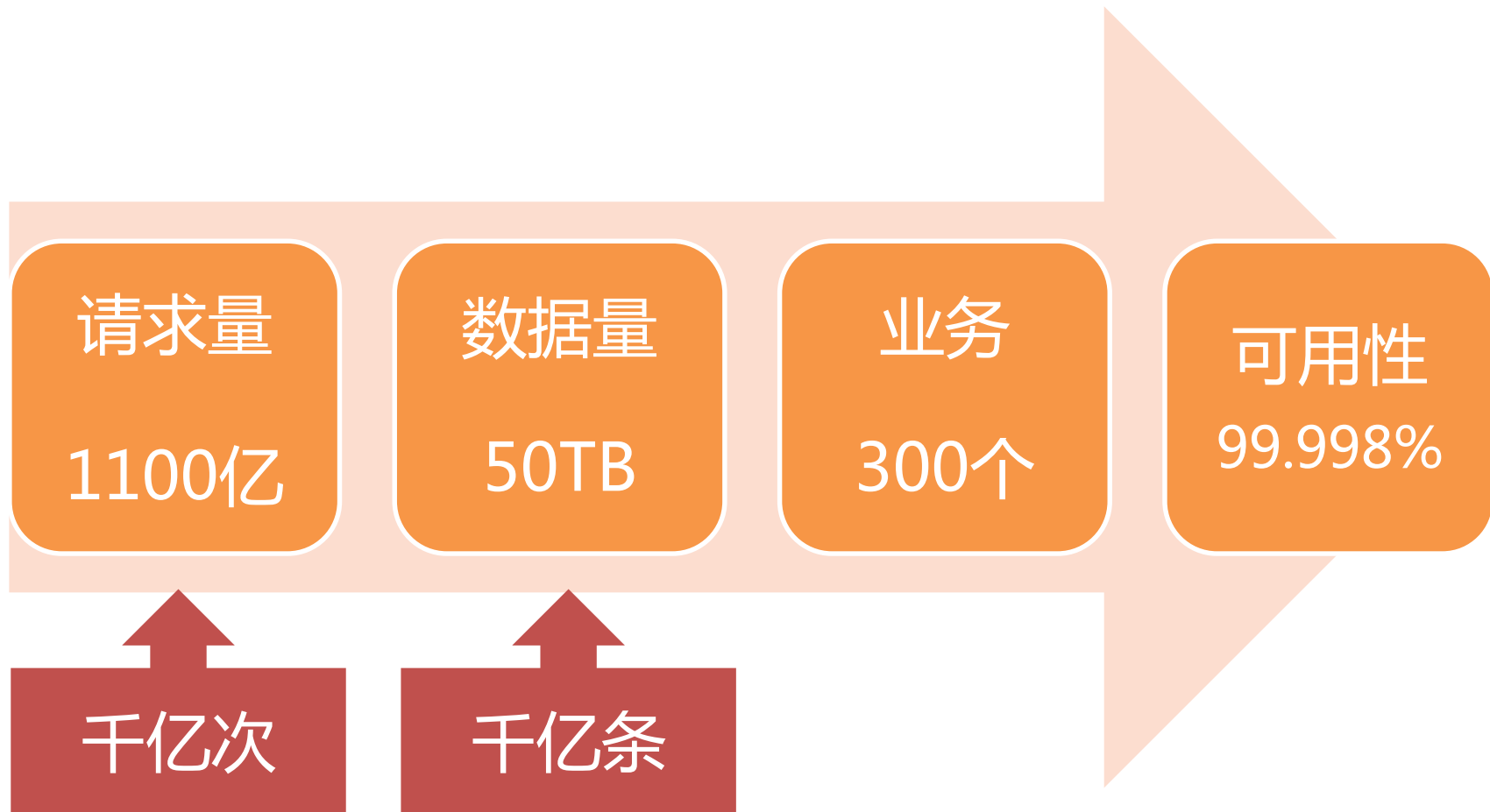
•  $\{\text{rowKey}, \text{colKey}\} \rightarrow \{\text{value}, \text{score}\}$

Key-List

• 按colKey排序，按score排序

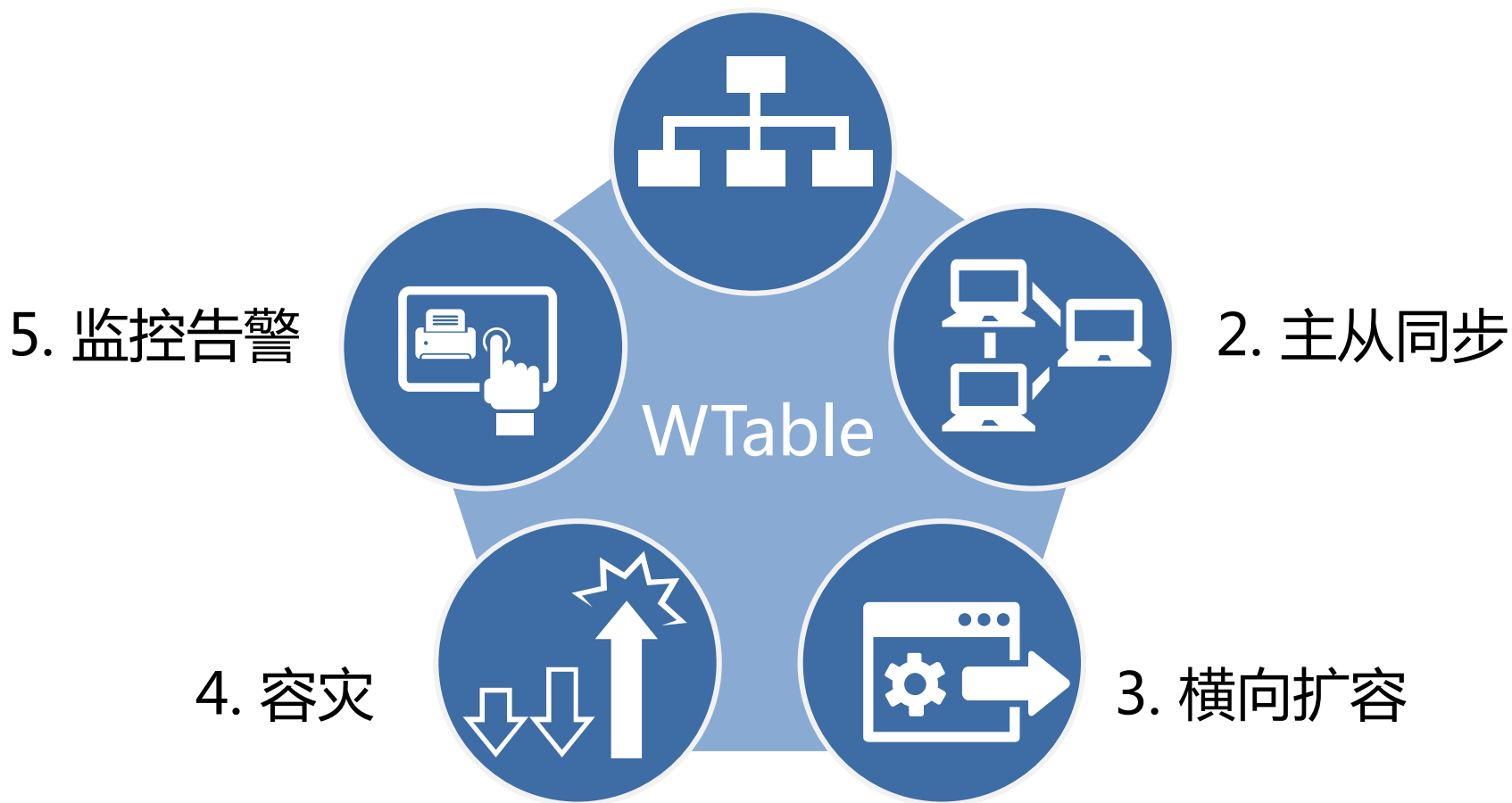
|         | colKey1      | colKey2      | colKey3      | ... |
|---------|--------------|--------------|--------------|-----|
| rowKey1 | value, score | value, score | value, score | ... |
| rowKey2 | value, score | value, score | value, score | ... |
| ...     | ...          | ...          | ...          | ... |

# WTable的规模

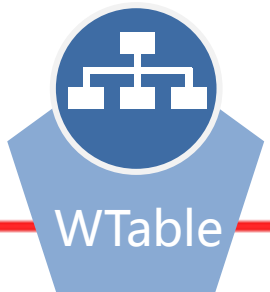


# WTable关键技术

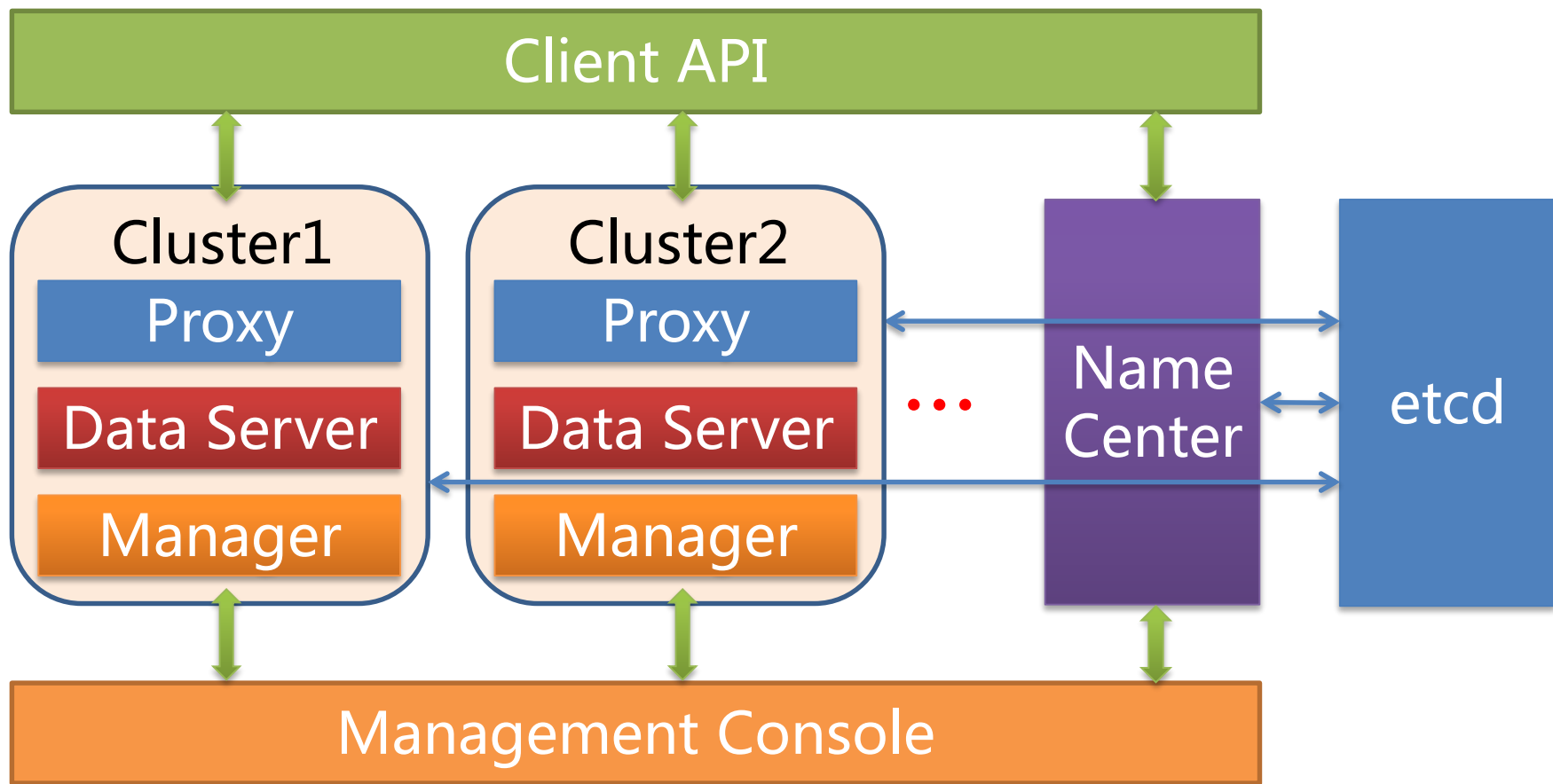
## 1. 整体架构







# 1. 整体架构



1

隔离

+

共享

大集群？

小集群？

混合模式

2

Proxy

+

多一层

客户端变简单

更好容灾

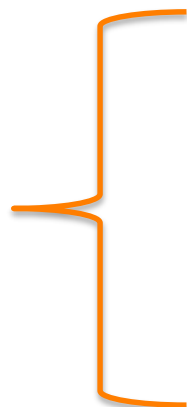
并行处理

3

引擎

+

性能



BerkeleyDB ?

LevelDB ?

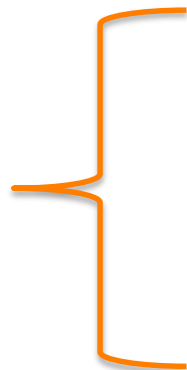
RocksDB

4

逻辑层

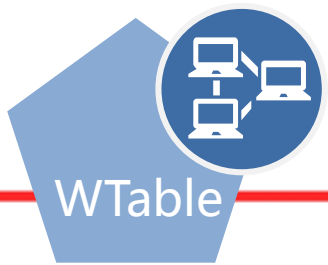
+

效率



C/C++

Golang



## 2. 主从同步

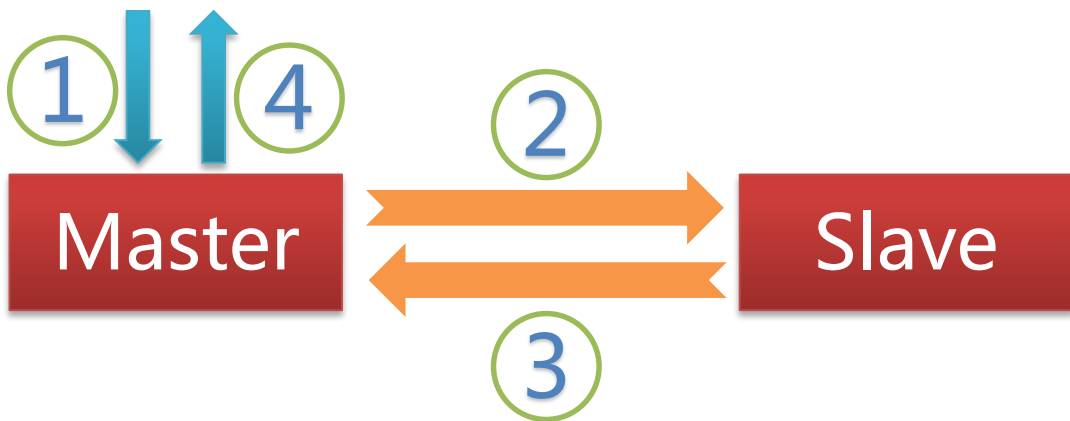
### 异步实时推送

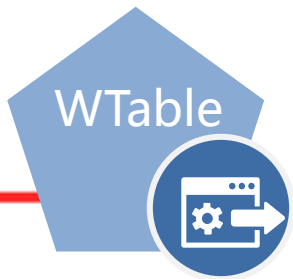
- 不一致风险
- 毫秒延迟
- 高性能+高可用



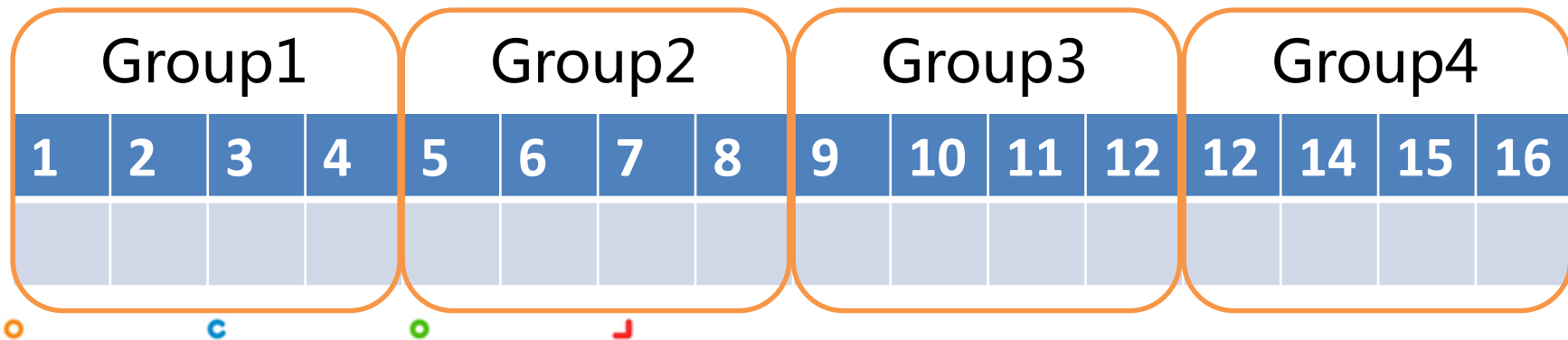
### 半同步复制

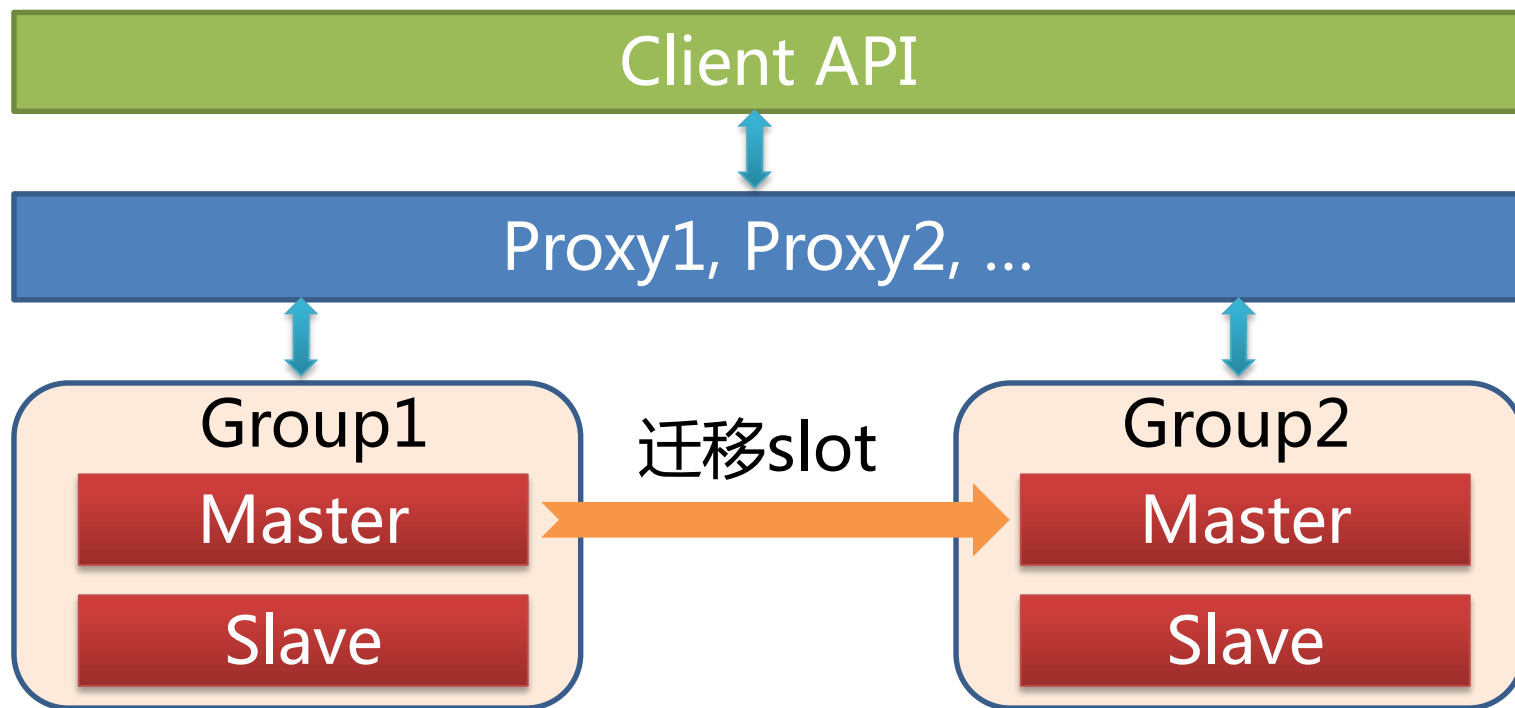
- 一致性强
- 写性能下降
- 可用性下降



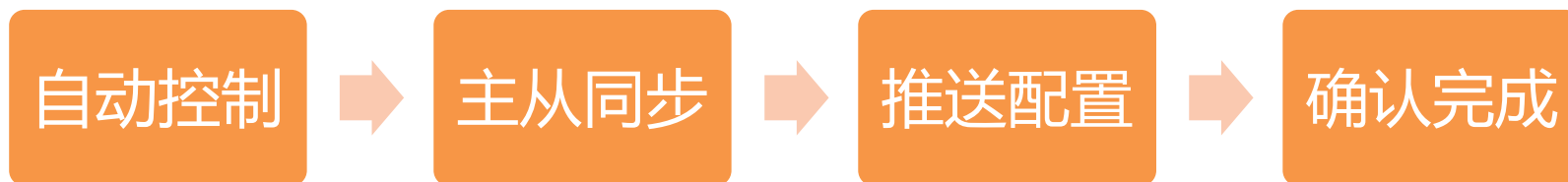


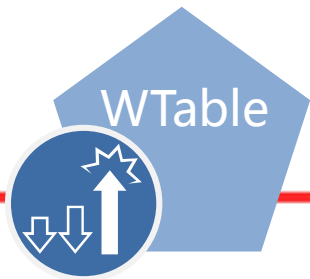
# 3. 横向扩容



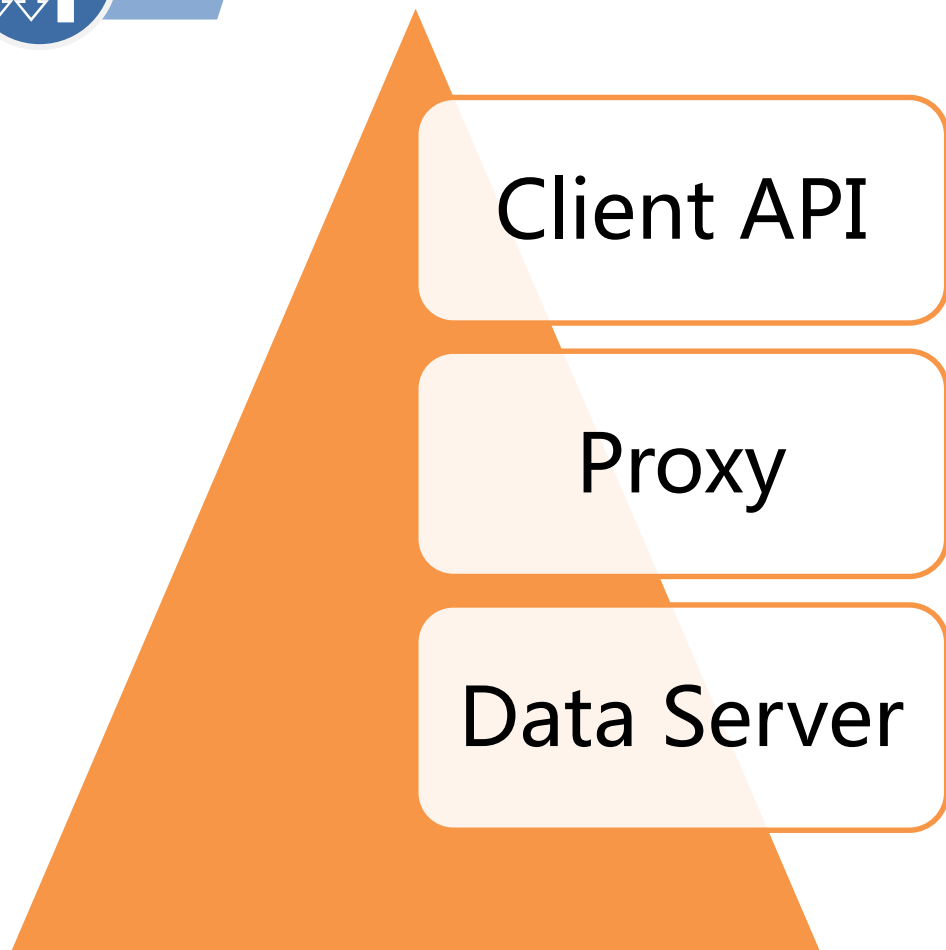


## 一键扩容

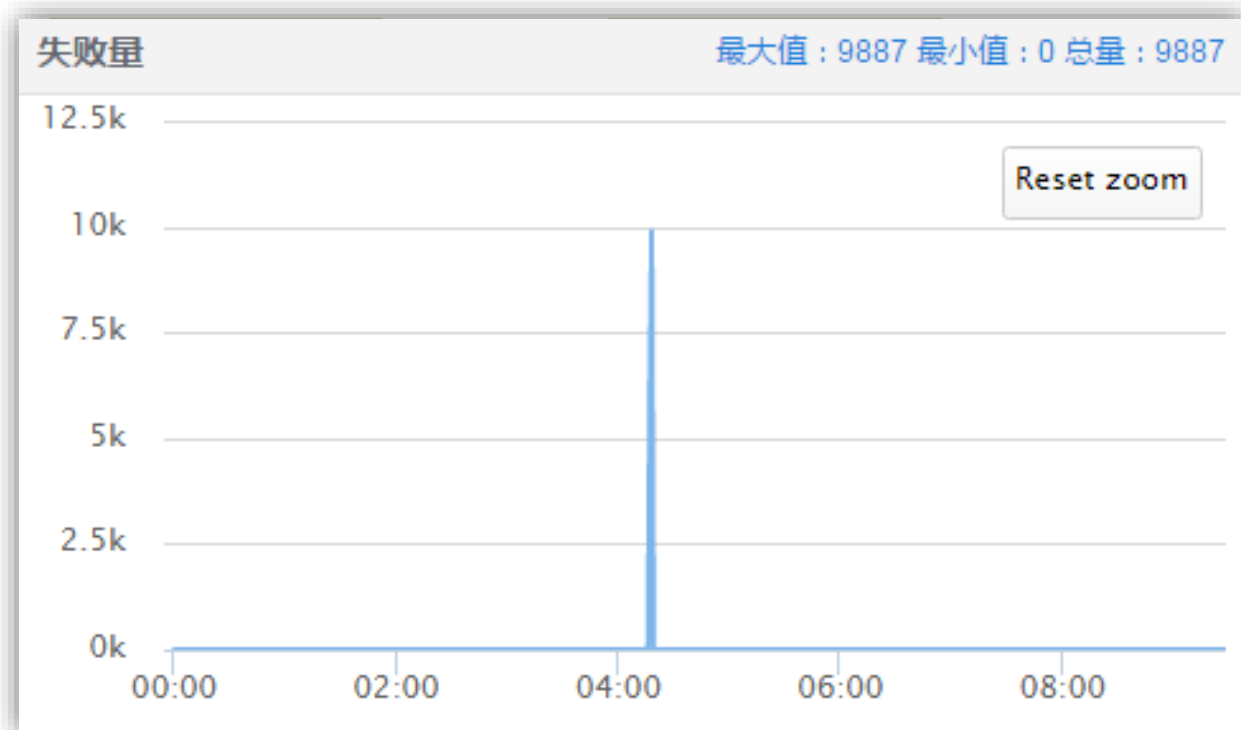




## 4. 容灾



- 电源故障造成某集群1/4的机器掉电（单个机柜）
- 秒级自动恢复，业务几乎无影响



异常检测

自动摘除

跨机柜部署



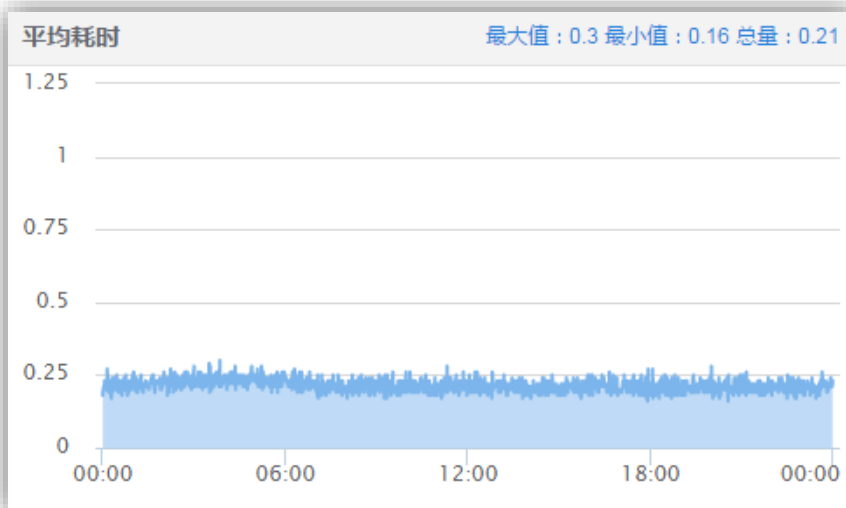
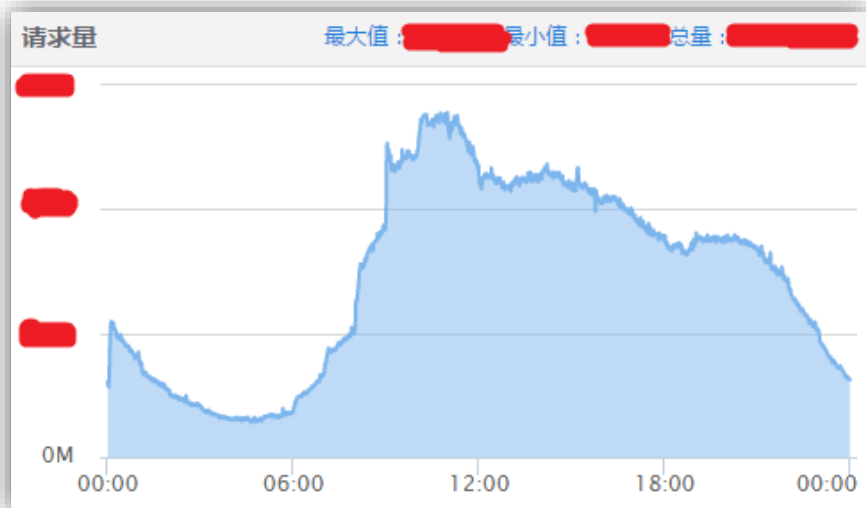




WTable

## 5. 监控告警

让生活更简单 



基于日志



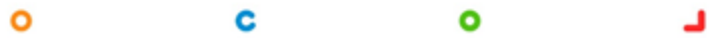
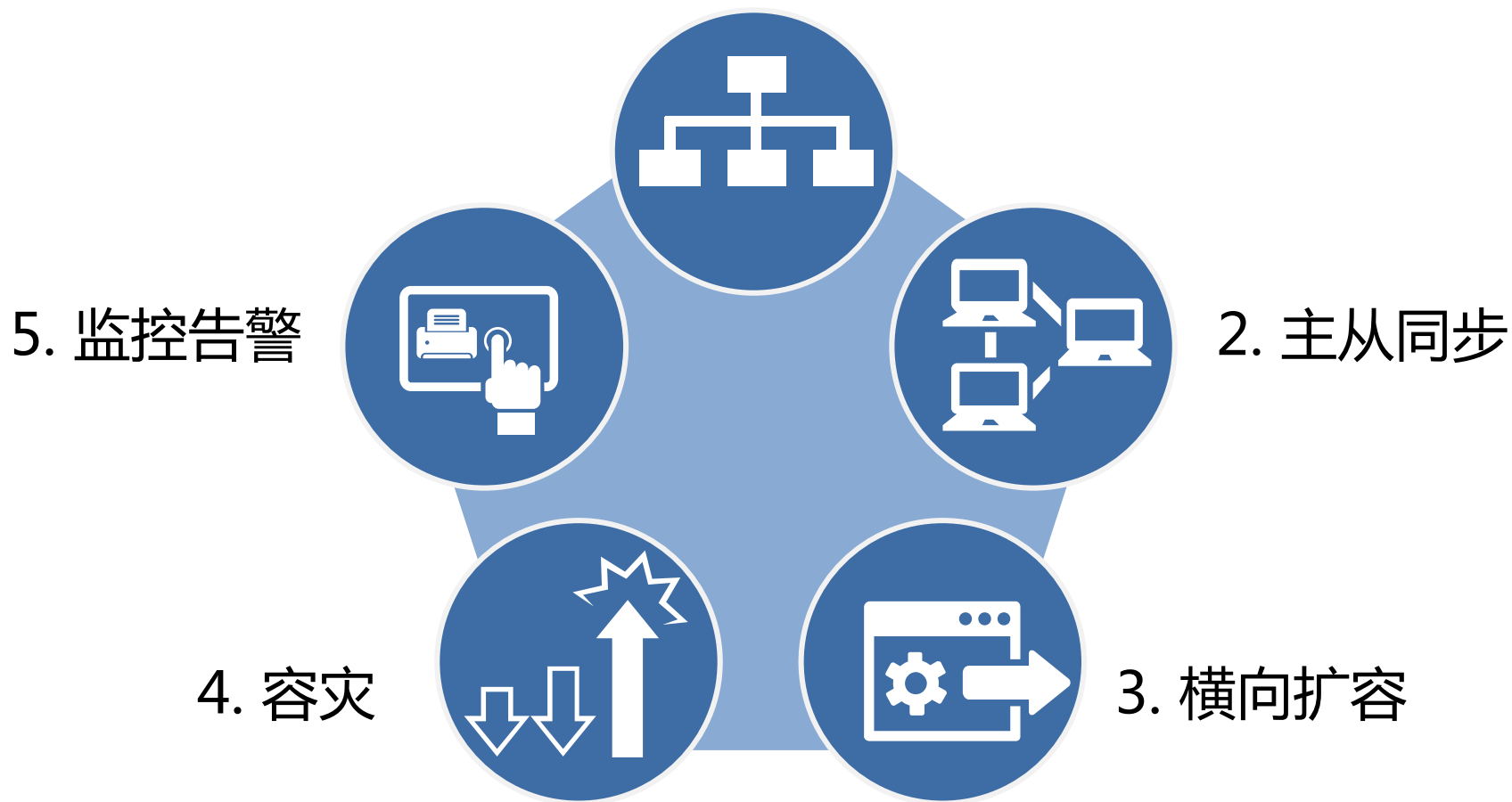
UDP+Agent

开销极小



# WTable关键技术

## 1. 整体架构



# 典型案例：核心帖子服务

让生活更简单



- 58核心帖子服务，**百亿**个帖子，每天**百亿**次调用
- id → {帖子信息}

扩容

可用性

爆炸式增长



诚意出售 不限购 精装修 龙湖唐宁one 南向开 2年

五道口 - 唐宁ONE二手房

近地铁 不限购 土巴免礼包

经纪人：卫峰 今天

350 万 67829元/m<sup>2</sup>

1室1厅1卫 (51.6m<sup>2</sup>)



地质大学 东西通透2居 无税随时看房 1年

五道口 - 地质大院二手房

有钥即看 土巴免礼包

经纪人：丁春玲 我爱我家 今天

524 万 80776元/m<sup>2</sup>

2室1厅1卫 (64.87m<sup>2</sup>)

# 典型案例：核心帖子服务

让生活更简单



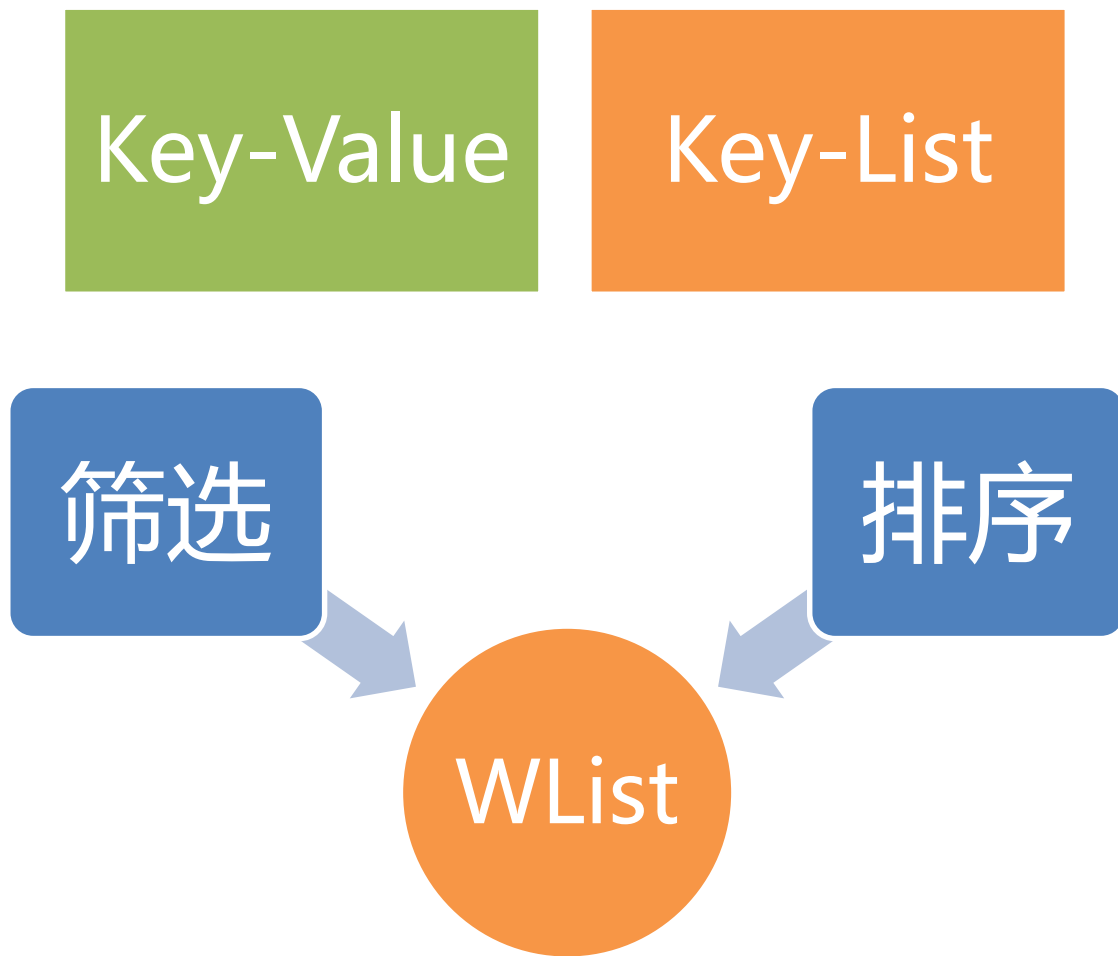
## ➤ WTable存储设计

|           | colKey=1                  | colKey=2 |
|-----------|---------------------------|----------|
| rowKey=id | {userId, 时间, 状态, 类别, ...} | {内容}     |

## ➤ 迁移效果



# 还有哪些痛点



# WList列表存储平台

分布式



Key-List



筛选



排序

## ➤ 类SQL语法

- where category=8&date=14567616\_14594400
- orderby date desc

## ➤ 应用场景举例

- 用户发帖&收藏列表
- 简历投递&下载记录
- 好友列表



- 基于WTable架构设计
- Schema
  - 字段类型int, float, string
  - 字段可随意扩展
- Schema如何设计才能做到效率高？
  - etcd存储schema配置
- 如何在list很长的情况下快速响应？
  - 预排序
  - 缓存
  - 截断



# 总结

WTable

WList

千亿级

架构

关键技术

案例





# Thanks!

让生活更简单



58集团技术专场