GOPS2018
Shenzhen

GOPS

全球运维大会 2018

2018.4.13-4.14 中国·广东·深圳·南山区 圣淘沙大酒店（翡翠店）

海量社交业务多活及调度实战

李剑锋　腾讯QQ业务运维负责人

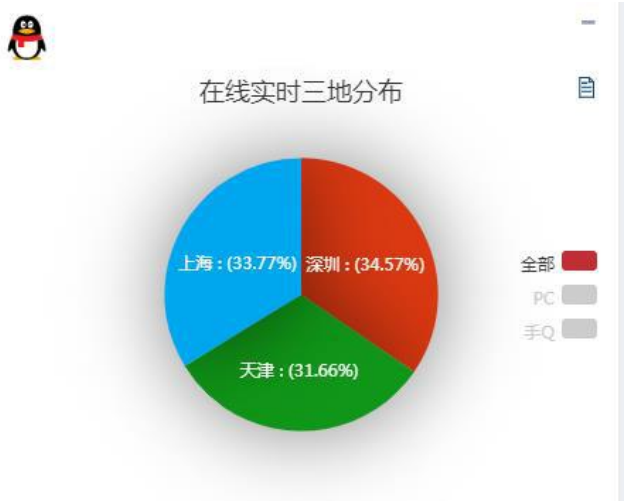
目录

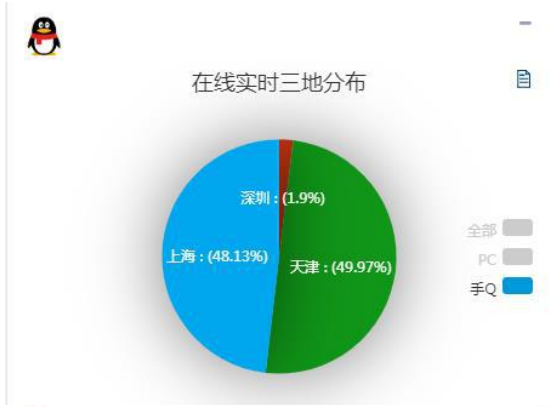
为什么要做异地部署？

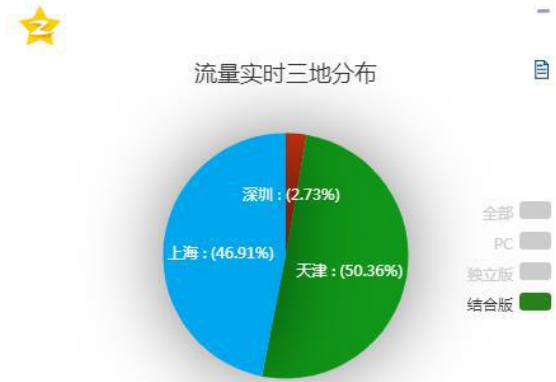
- 被动
  - 灾难性事件应对
- 主动
  - 大型变更地区级灰度
  - 优化接入质量

承载7500万QQ用户的天津IDC随时可能出问题

多地部署现状





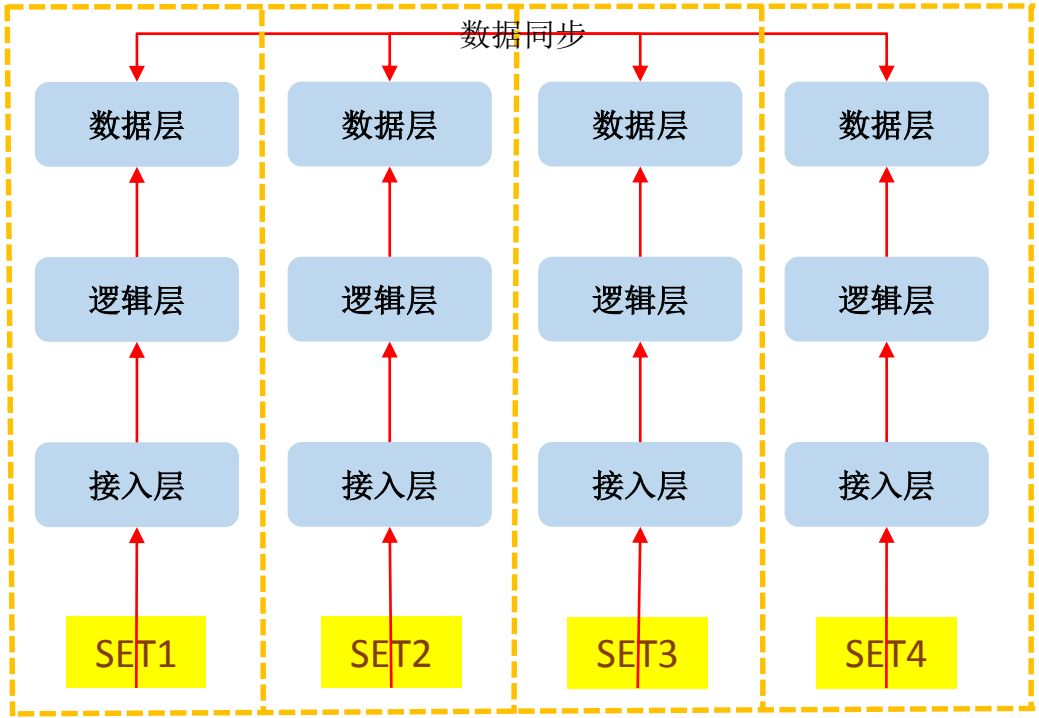


容量评估不可被完全信任
需要周期性的进行压测演习

目录

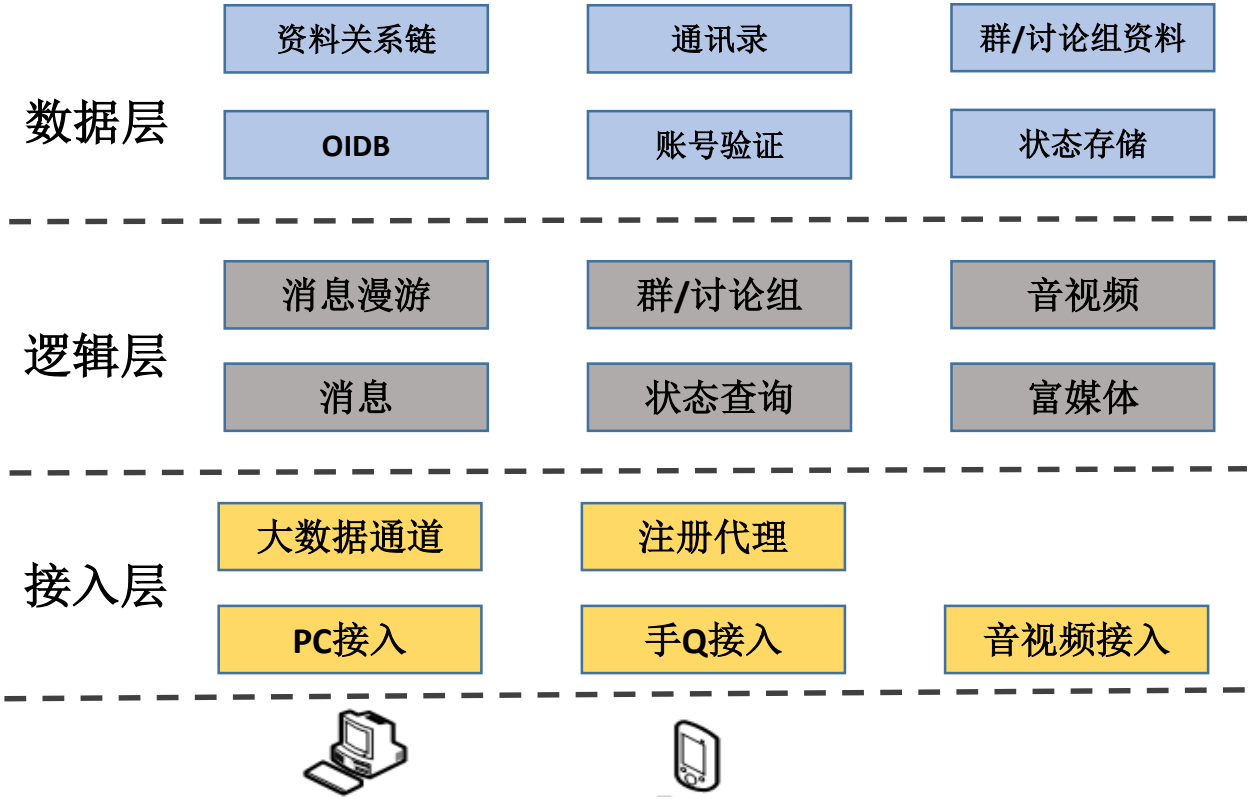
Set化部署

- 数据与逻辑分离
- 数据的多地同步
- 平行扩缩能力
- 就近调度能力



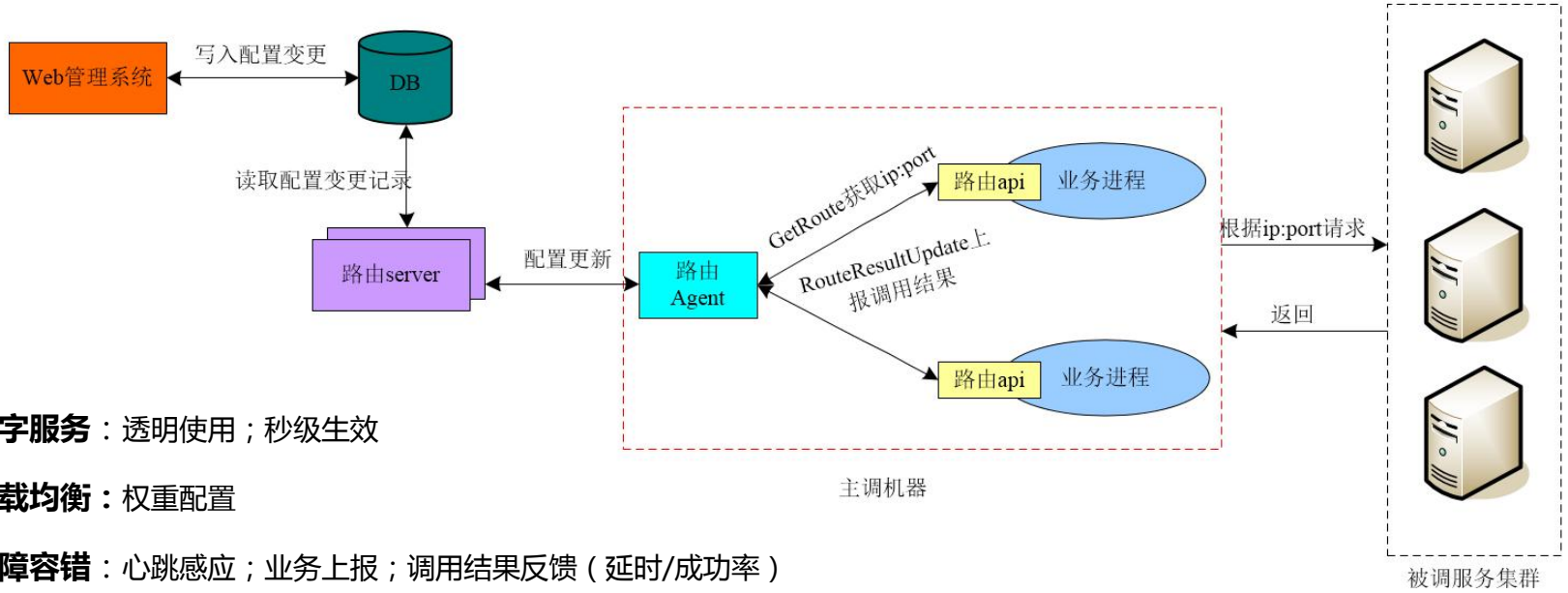
SET： 具备完整逻辑和数据，可以支持用户独立服务的小容量系统

QQ核心架构框架



- 资料写数据同步
- 状态写数据同步
- 异地用户消息穿越

无状态统一调度服务-织云路由



- **名字服务**：透明使用；秒级生效
- **负载均衡**：权重配置
- **故障容错**：心跳感应；业务上报；调用结果反馈（延时/成功率）
- **就近访问**：自动异地容灾切换

目录

QQ核心数据分类

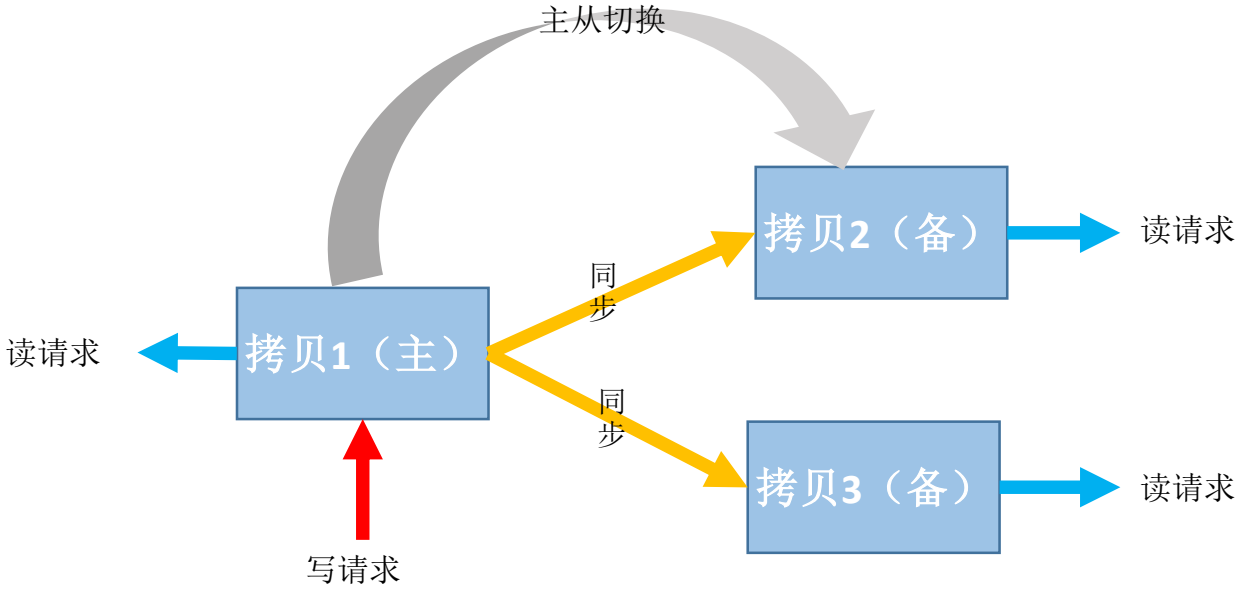
◆资料关系链数据

- ●高可靠性保证
- ●读多写少
- ●同步流量小

◆状态数据

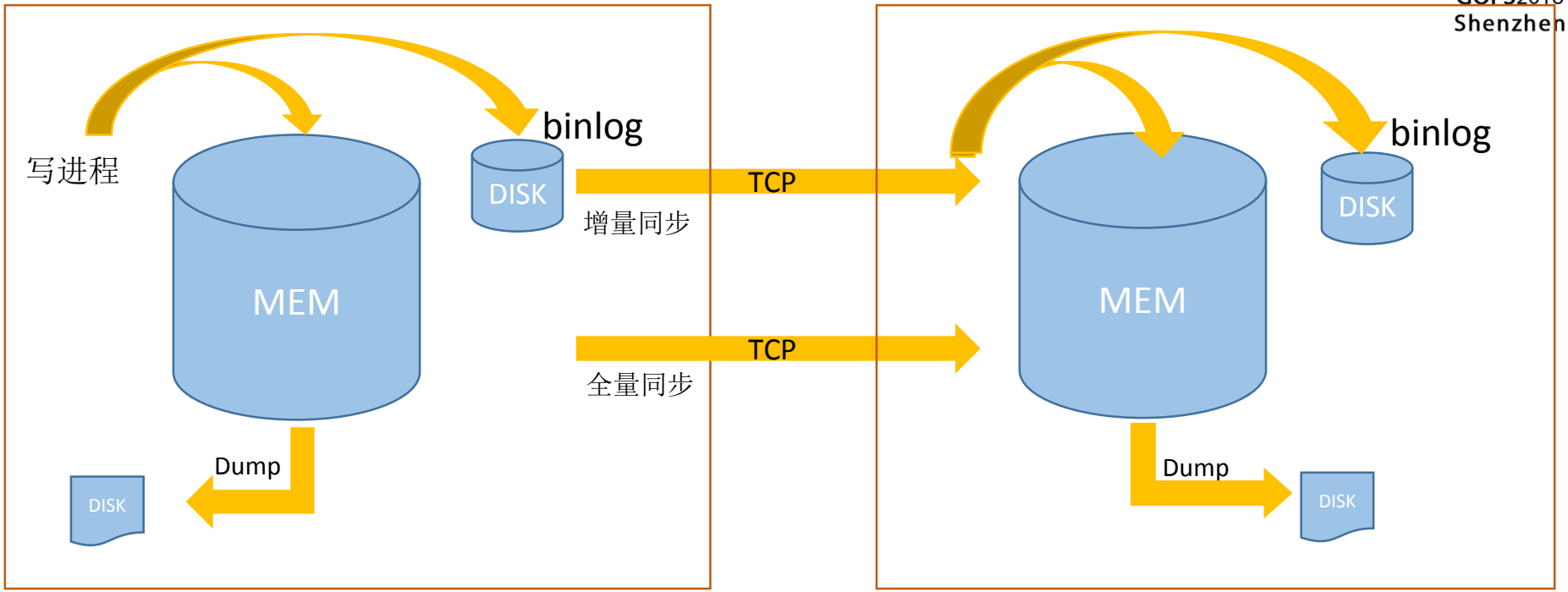
- ●可再生数据
- ●读多写多，地区间同步流量大
- ●数据下沉到应用，本地同步流量大

资料关系链数据同步



- 防止主key冲突
- 故障自动切换
- Seq保证一致性
- 读写无法均衡

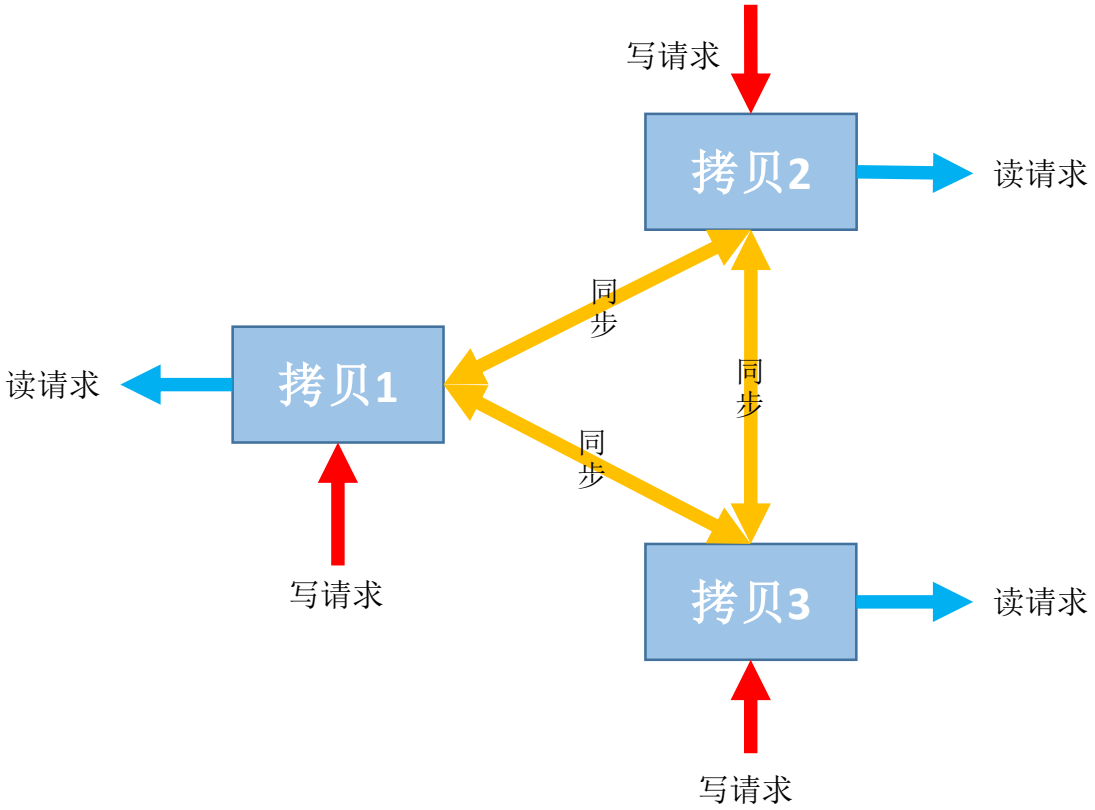
资料关系链数据同步



主拷贝（Seq=N）

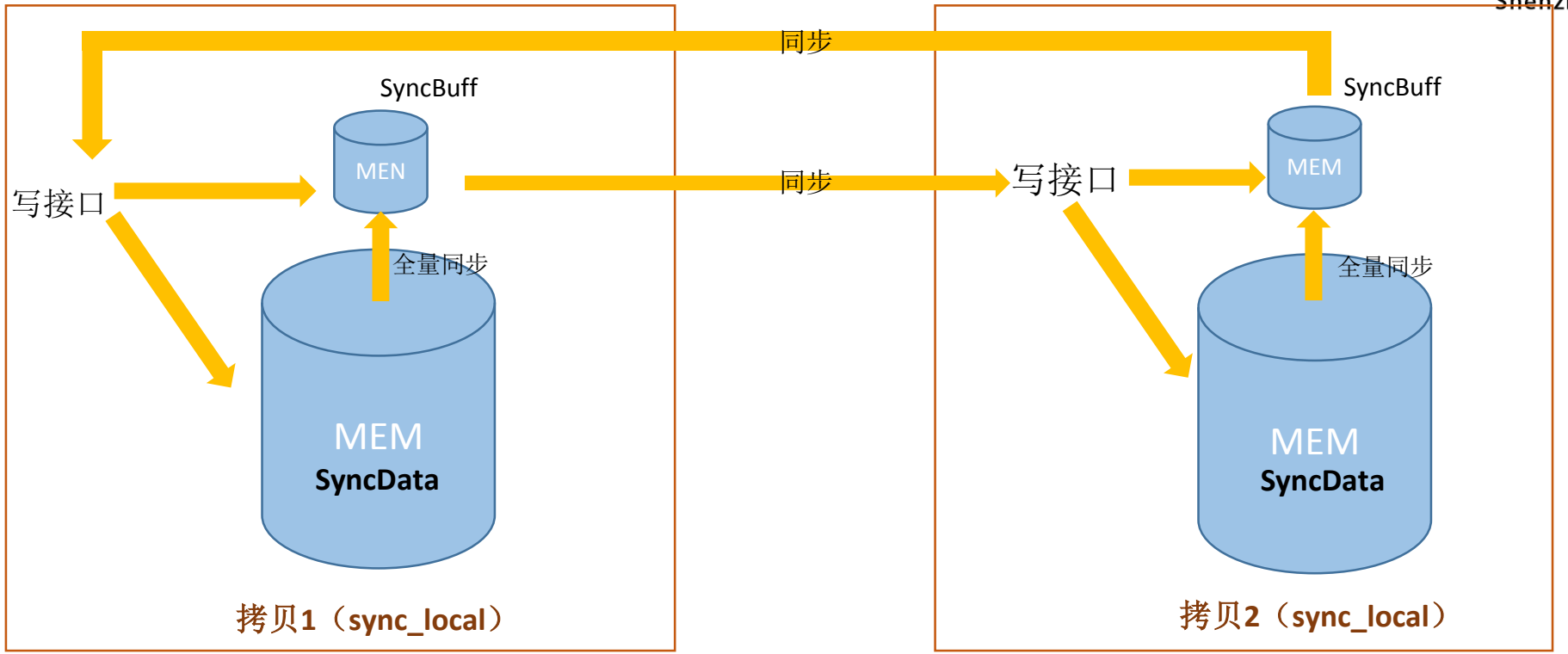
备拷贝（Seq=M）

状态数据同步



- 业务保证key不冲突
- 全内存，读写效率高
- 数据不一致容错
- 增量同步和全量同步结合
- 读写可以均衡

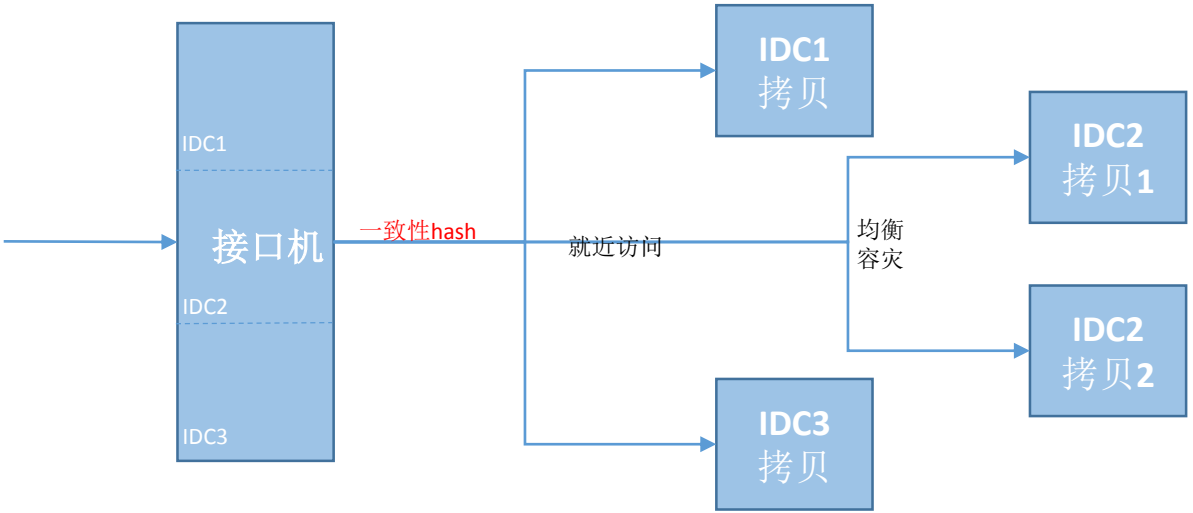
状态数据同步（地区间）

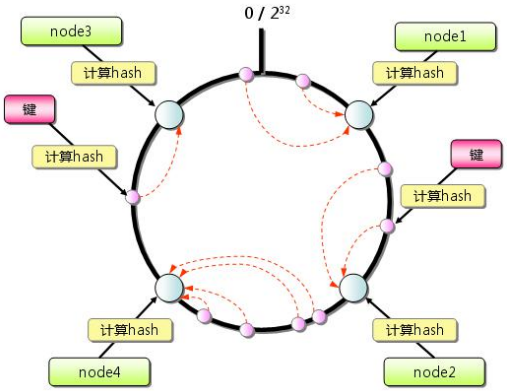

同步
SyncBuff
MEN
写接口
同步
全量同步
MEM
SyncData
拷贝1（sync_local）
SyncBuff
写接口
MEM
全量同步
MEM
SyncData
拷贝2（sync_local）

数据切分与寻址

- 互联网系统的分布式特点，单节点能力有限
- 业务大请求与大容量需求
- set容量模型依赖数据水平伸缩能力
- 数据切分的关键点：主KEY
- 切分方法
  - 一致性hash：key通用；算法通用；回查困难
  - （key段）号段切分：数字key；人为划分；回查直观

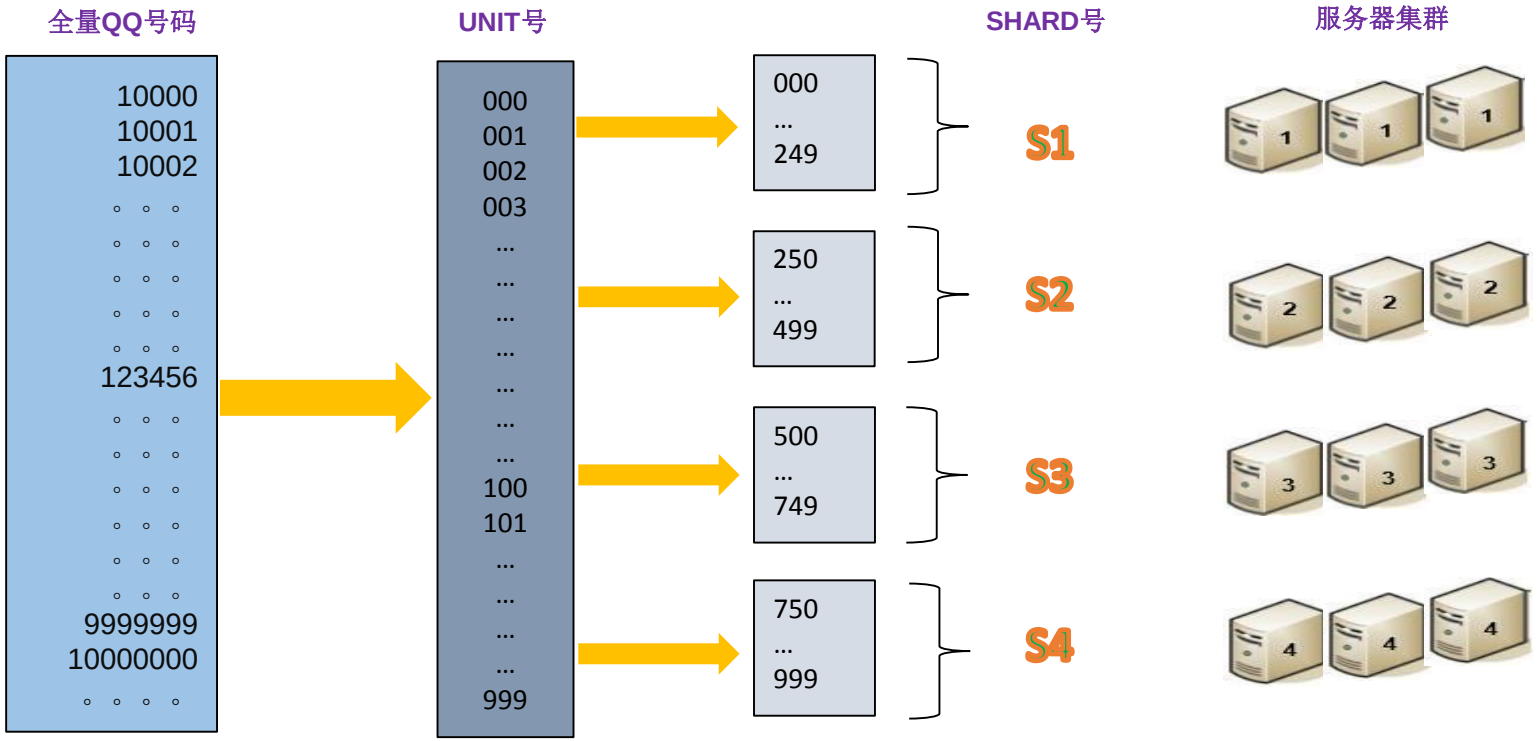
一致性hash数据寻址



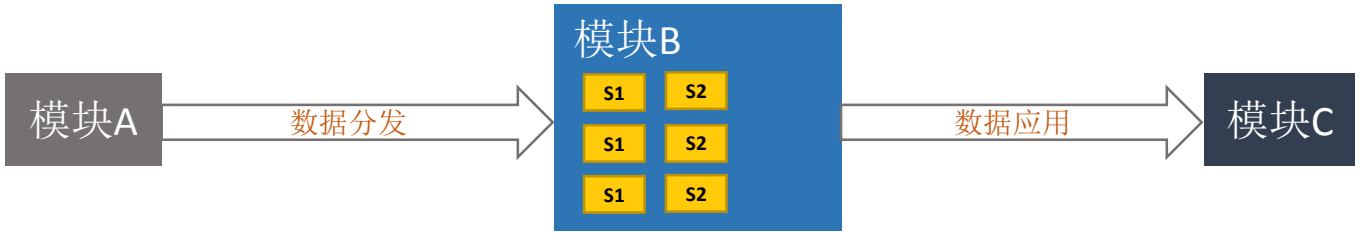


- 可以拥有多台接口机，多个拷贝；每个拷贝可以拥有多台机器，相互备份
- 逻辑层发送请求包到接口机是不用寻址的，任意接口机都可以处理对任意Key的请求
- 接口机转发请求包到拷贝cache采用一致性哈希算法，cache处理后发送回应包给来源接口机，接口机再转发回应包给来源逻辑层
- 接口机有超时机制，超时未收到cache的回应包就会生成超时回应包给来源逻辑层
- 当增加机器（扩容）时，数据只会从老cache迁移到新cache，而不会在老cache之间做无用的移动

号段切分寻址方案


全量QQ号码
UNIT号
SHARD号
服务器集群
10000
10001
10002
123456
9999999
10000000
000
001
002
003
100
101
999
000 ... 249
250 ... 499
500 ... 749
750 ... 999
S1
S2
S3
S4

号段切分热扩容方案1



模块B
读取使用

并集

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

模块C
读取使用

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

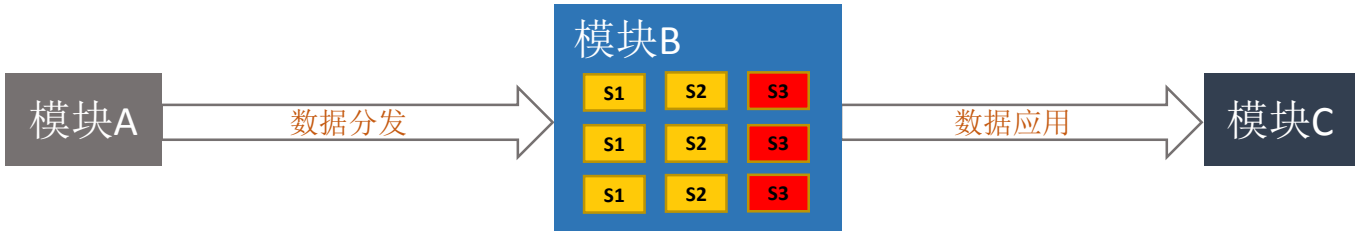
模型假设：

- 模块B从模块A拉取数据，并供模块C调用
- 模块B分成2个SHARD，3个拷贝

模块B配置：

- 对于模块B，总共有三行SHARD切分配置
- 第一行和第二行取并集后确定自身所负责从模块A拉取的数据
- 第三行配置供模块C确定需要模块B的哪个SHARD拉取数据
- 日常运营三行配置一致

号段切分热扩容方案2



第一次变更

模块B
读取使用

并集

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	**3**	2	2	**3**

模块C
读取使用

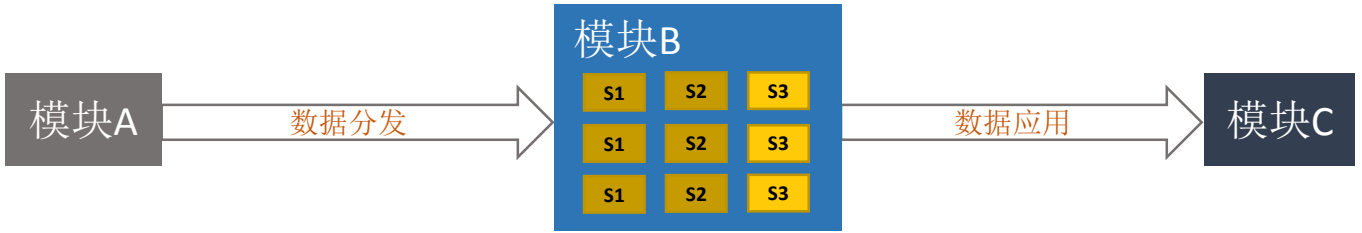
UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

当前状态：

- 对模块B自身来说， SHARD 1、2负责的UINT列表无变化，拉取数据无变化
- 对模块C所有号段依然从SHARD 1、2拉取数据
- 模块B新扩容SHARD3设备配置负责UINT3、6，并开始拉取数据
- SHARD3设备处于只从模块A拉取数据的不健康状态，直到数据拉取完整

号段切分热扩容方案3



第二次变更

模块B
读取使用

并集

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	3	2	2	3

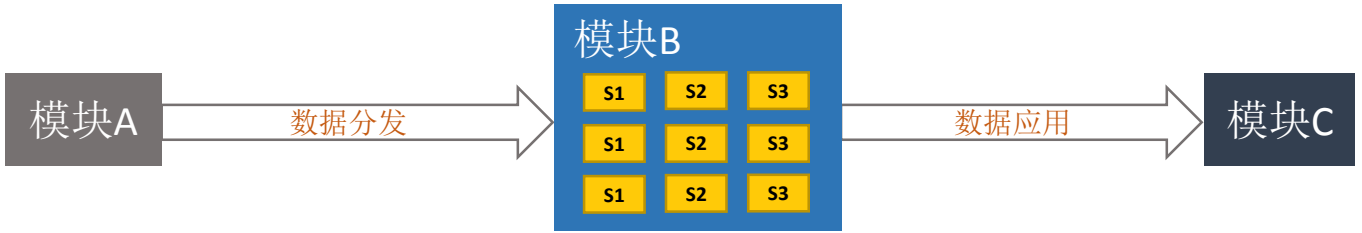
UNIT	1	2	3	4	5	6
SHARD	1	1	3	2	2	3

模块C
读取使用

当前状态：

- 模块C的UNIT3、6开始从模块B新扩容SHARD3服务器上调用
- 模块B原有SHARD1、2服务器接收冗余数据
- 直到确认新扩容SHARD3服务器正常提供服务

号段切分热扩容方案4



第三次变更

模块B
读取使用

并集

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	3	2	2	3

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	3	2	2	3

模块C
读取使用

UNIT	1	2	3	4	5	6
SHARD	1	1	1	2	2	2

UNIT	1	2	3	4	5	6
SHARD	1	1	3	2	2	3

当前状态：

- 去掉模块B的SHARD1、2服务器冗余数据拉取
- 扩容完毕

目录

调度基本原理

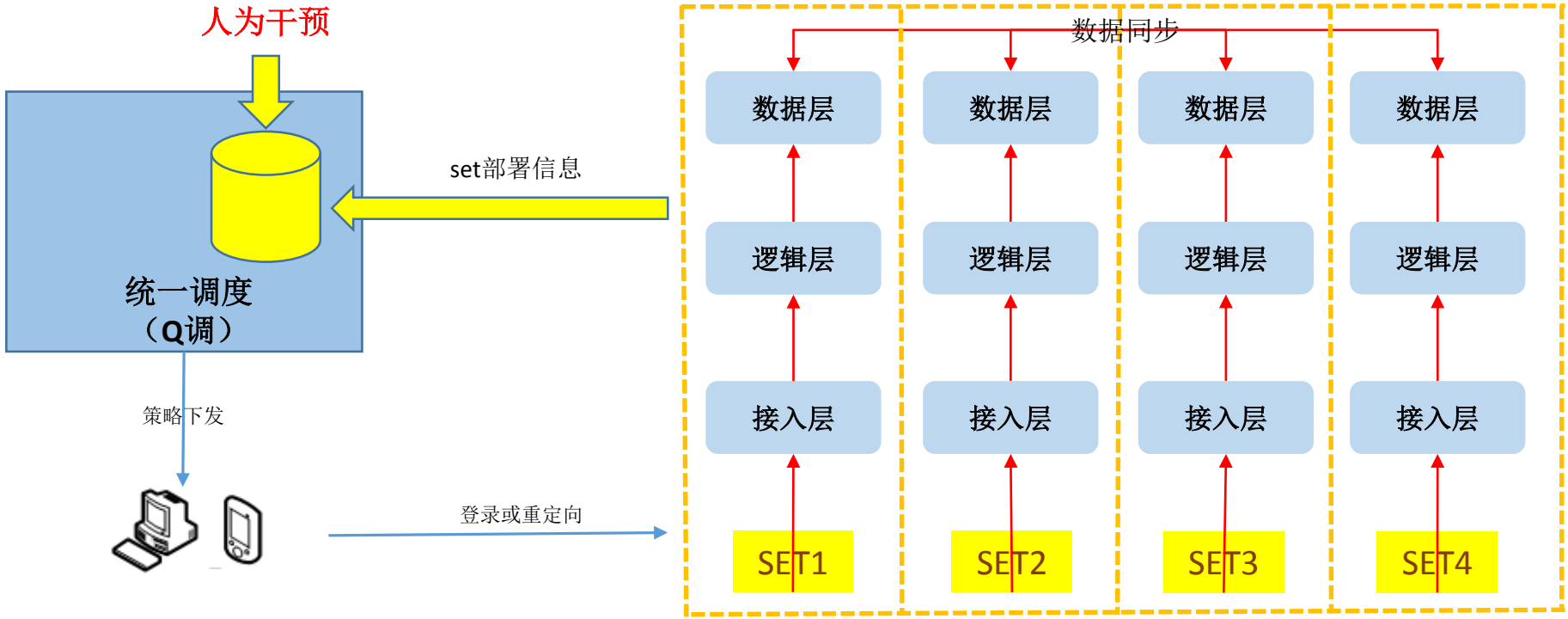
◆PCQQ

- ●重定向：客户端与服务器的专有协议
- ●优先尝试服务器给的IP

◆手机QQ

- ●下发更新用户登录服务器列表
- ●用户按照列表以此登陆服务器

调度框架


人为干预
统一调度
（Q调）
set部署信息
策略下发
登录或重定向
数据同步
数据层
逻辑层
接入层
SET1
SET2
SET3
SET4

调度分类

分类	逻辑	应用场景
屏蔽调度	● 更新上线用户的登录IP列表，屏蔽掉目标IDC服务器。 ● 用户下一次登录将不会登录到目标IDC。 ● 迁移速度依赖用户上下线频繁度。	容量压测性演习调度
弱禁调度	■ 更新上线用户的登录IP列表，屏蔽掉目标IDC服务器。 ■ 对登录到目标IDC的用户下发重新登录命令，用户将重新登录到非目标IDC服务器。 ■ 迁移速度依赖用户上下线频繁度。 ■ 会引起登录命令请求波动。	故障主动性防御调度
强禁调度	◆ 更新在线用户登录IP列表，屏蔽掉目标IDC服务器，并将再线用户按指定比率（用户量）踢下线强制重登录。 ◆ 迁移速度依赖于人为配置。 ◆ 会引起登录命令较大程度的波动。	灾难性事件应对

GOPS2018
Shenzhen

腾讯织云
CLOUD OPERATIONS CONSOLE

高效运维社区

开放运维联盟

Thanks

腾讯运维体系专场

荣誉出品

想第一时间看到高效运维社区的新动态吗？