分布式存储运维之道

腾讯社交网络NoSQL集群

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现

IT168.com ChinaUnix ITPUB

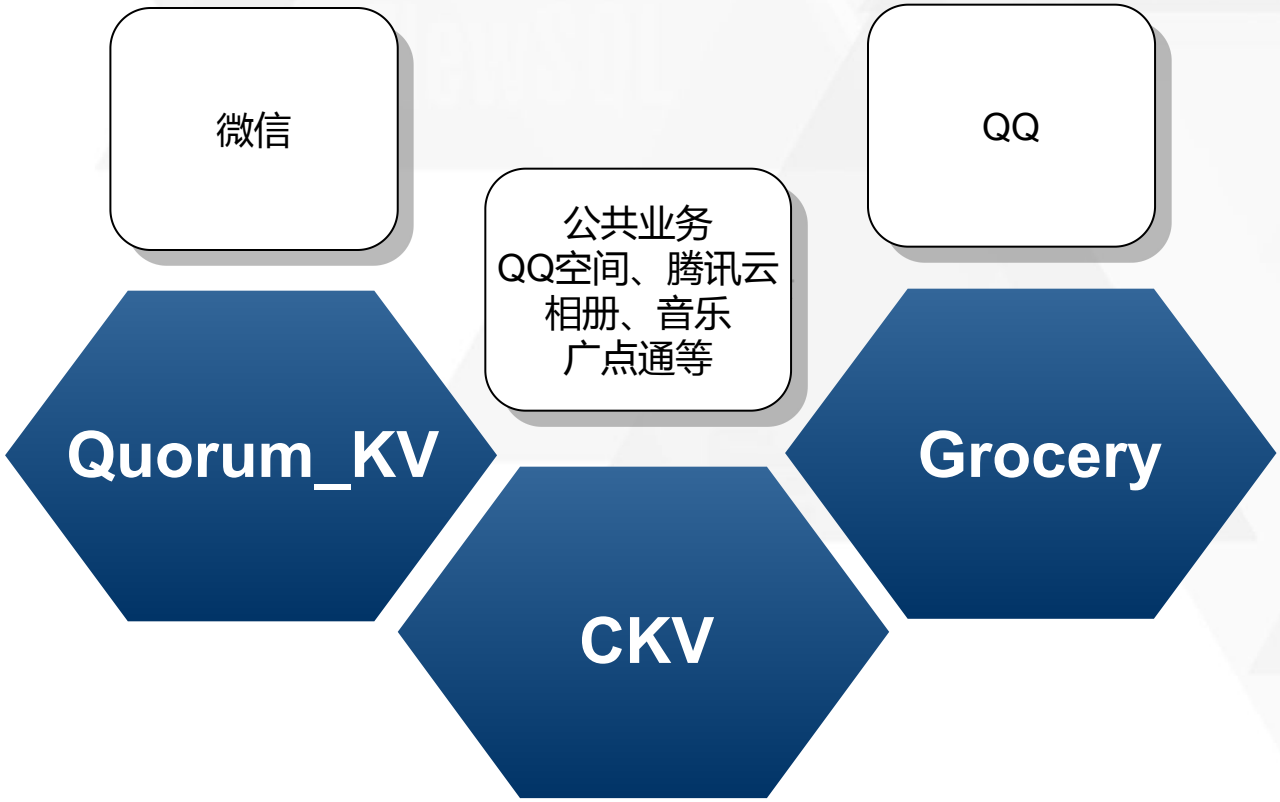
关于我

- millerzhou（周小军）
- 高级运维工程师
- @腾讯社交网络平台技术运营中心
- 曾负责天涯整体运维
- 擅长网站架构、云计算、分布式存储、关系数据库
- 马拉松新手（成绩440）

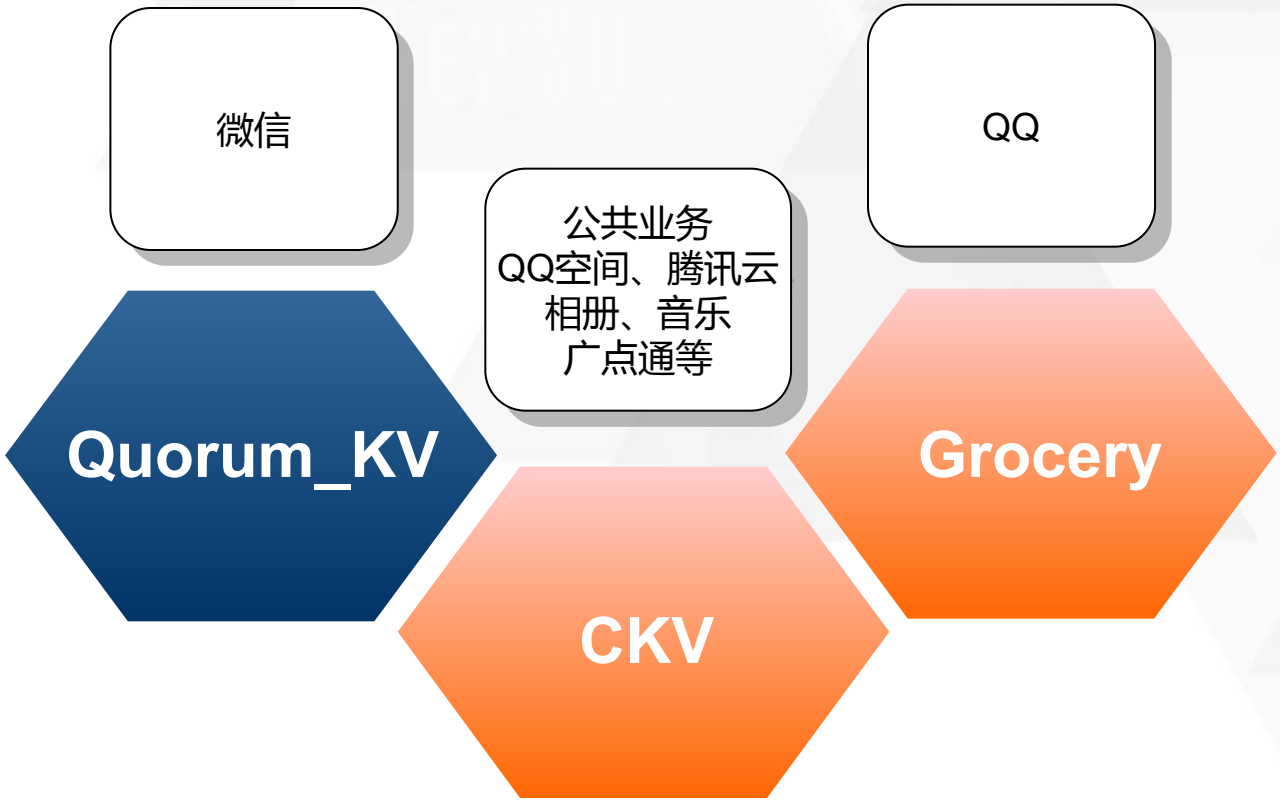
提纲

- 腾讯三大存储系统
- 存储运维体系
- 挑战和运维实践
  - -> 成本
  - -> 自动伸缩
  - -> 动态平衡
  - -> 网络优化
  - -> 跨城容灾

三大KV存储系统



三大KV存储系统



社交网络KV存储现状

10000+存储服务器

部署四地

70+存储仓库

150+TB内存容量

1亿-QPS

QQ

QQ音乐

QQ空间

广点通 腾讯效果推广

QQ会员 VIP.QQ.COM

腾讯微云

CKV典型业务

- QQ空间：Feeds、计数
- 广点通：用户画像、广告发布、计费
- 手机QQ：红点，游戏活动
- 世界杯彩票、春节红包
- 秒杀：抢F码
- 腾讯云：嘀嘀打车

CKV存储结构

KEY % 10000

桶 桶 桶 桶 桶 桶 桶

固定1万个桶

路由映射

内存存储单元（1GB） 内存存储单元（1GB）

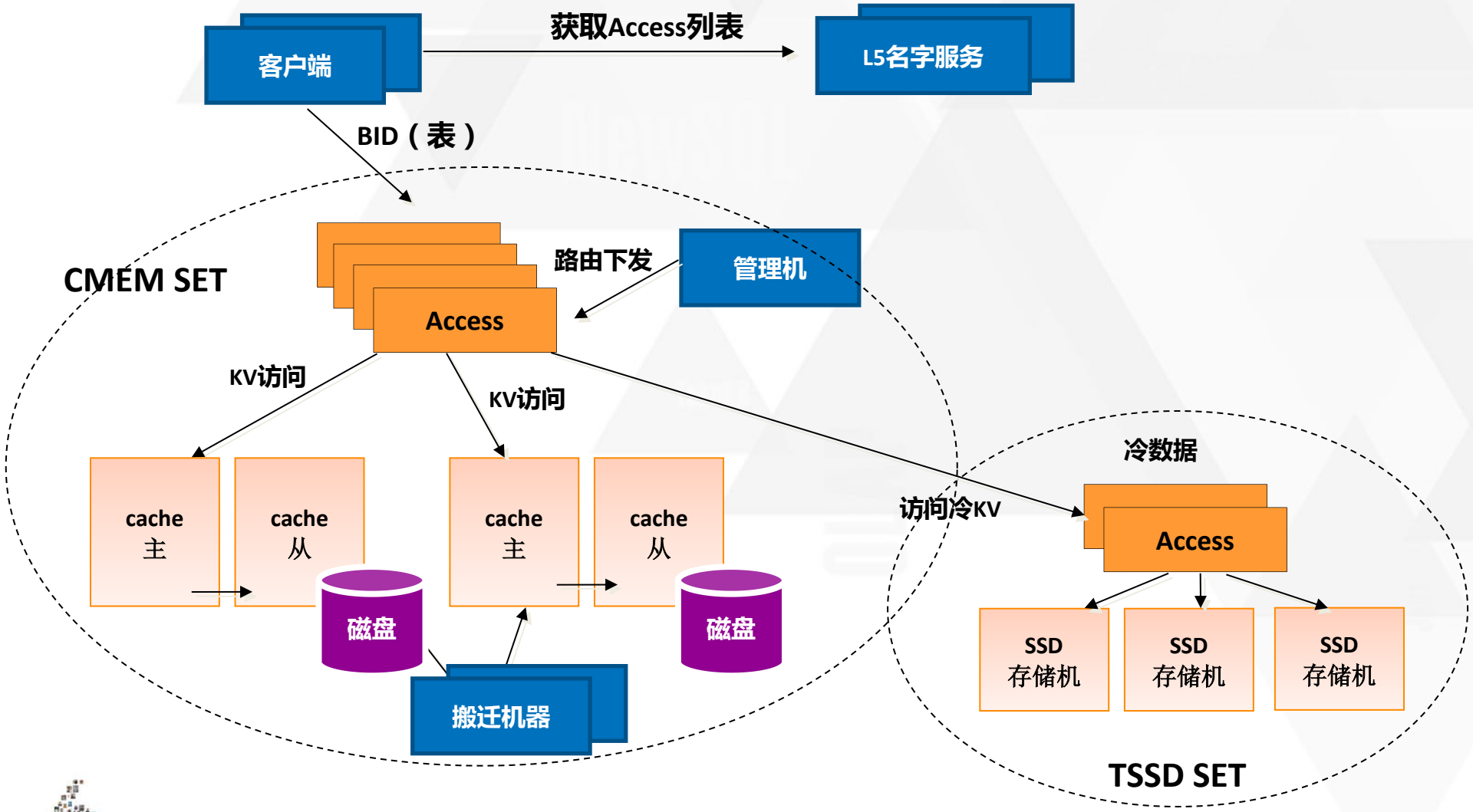
内存存储单元（1GB） 内存存储单元（1GB）

... 56*1GB内存存储

1个内存存储单元（CU）
最小存储1个桶，
最大存储1万个桶。

主存储 备存储

CKV存储部署模式


客户端
获取Access列表
L5名字服务
BID（表）
CMEM SET
Access
路由下发
管理机
KV访问
KV访问
cache
主
cache
从
cache
主
cache
从
磁盘
磁盘
搬迁机器
冷数据
访问冷KV
Access
SSD
存储机
SSD
存储机
SSD
存储机
TSSD SET

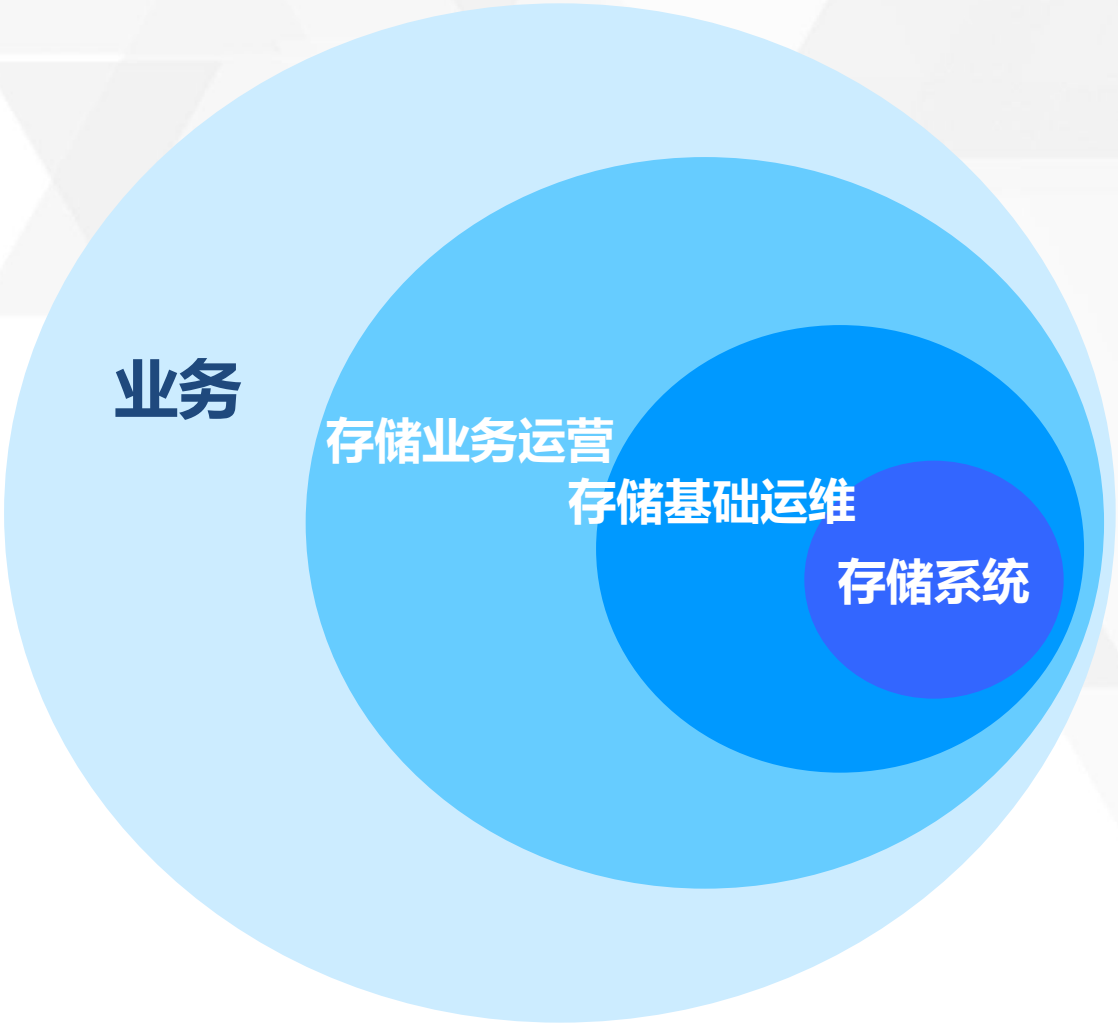
存储在社交架构中的位置

接入层 — **Qzhttp、Nginx**

逻辑层 — **SPP、SF2**

存储层 —
- **KV存储：CKV、Grocery**
- 关系数据库：**CDB**
- 文件存储：**TFS**

存储运维体系



运维的角色

存储运维

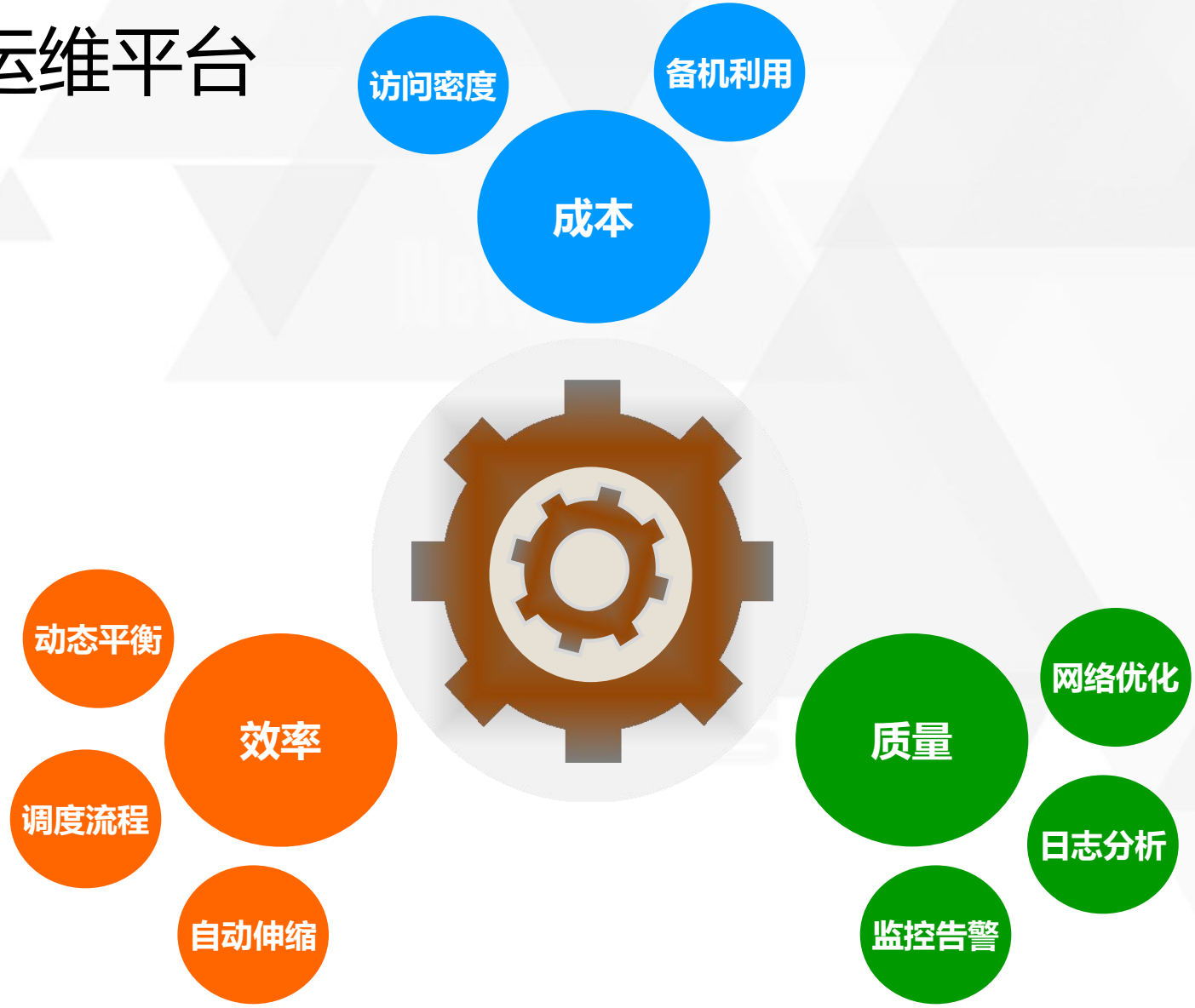
业务规划

打造工具

存储运维平台

面向业务的存储运维系统

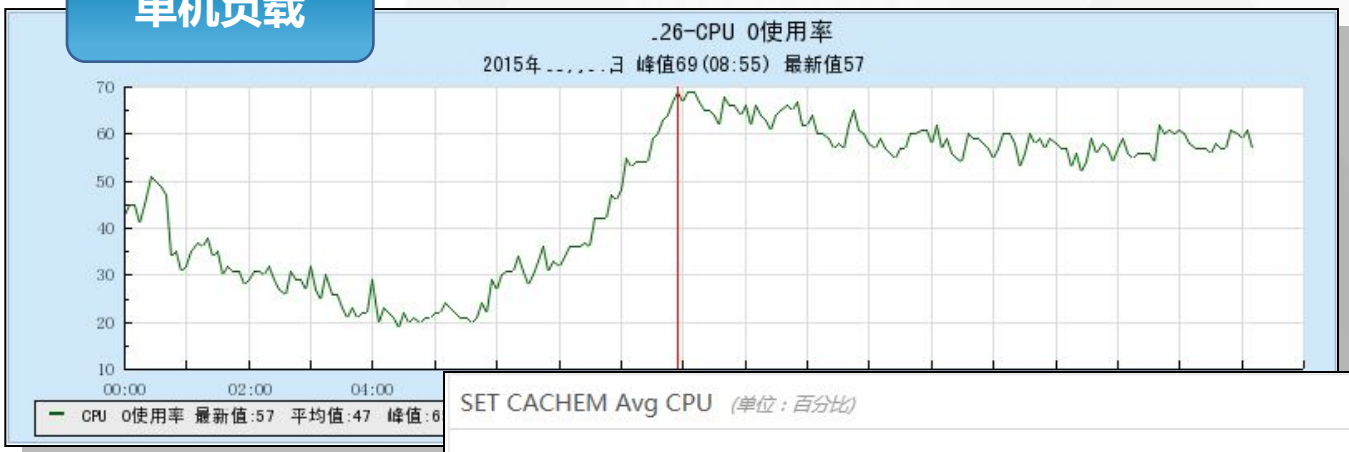
存储运维平台



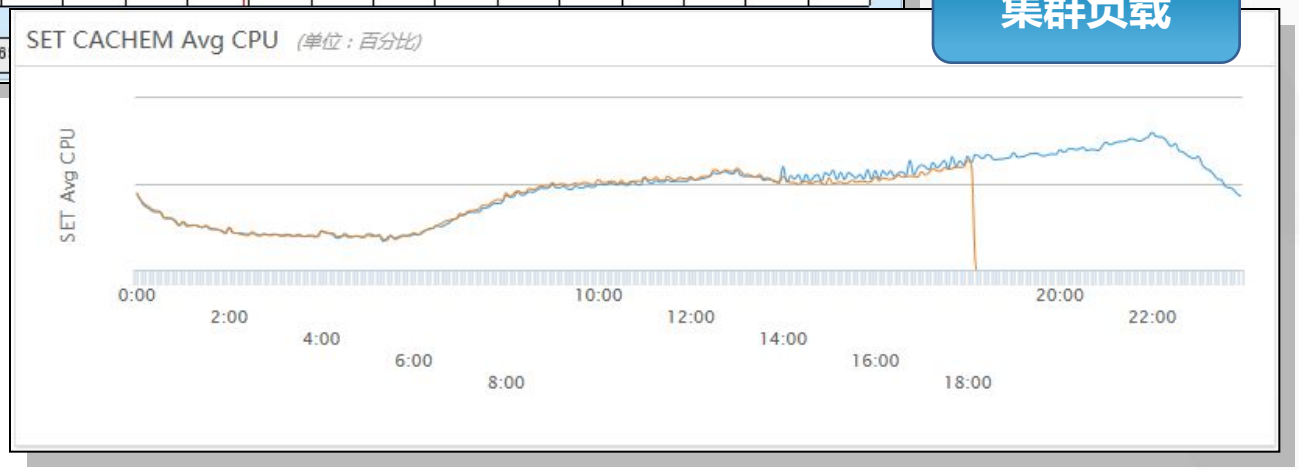
成本

存储集群中，怎么定义业务的负载容量？

单机负载



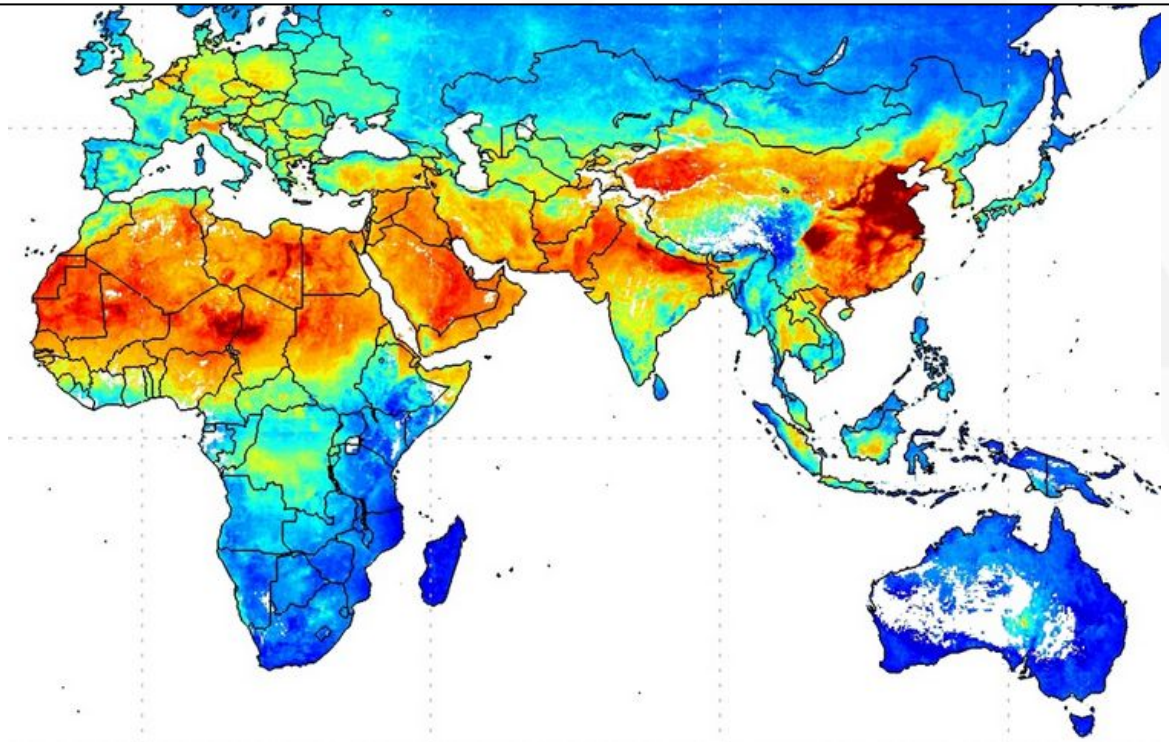
集群负载



可度量的成本指标

访问密度模型

-> 按存储容量和QPS来计算业务负载
-> 业务负载容量的核心指标
-> 公式：QPS / 存储容量

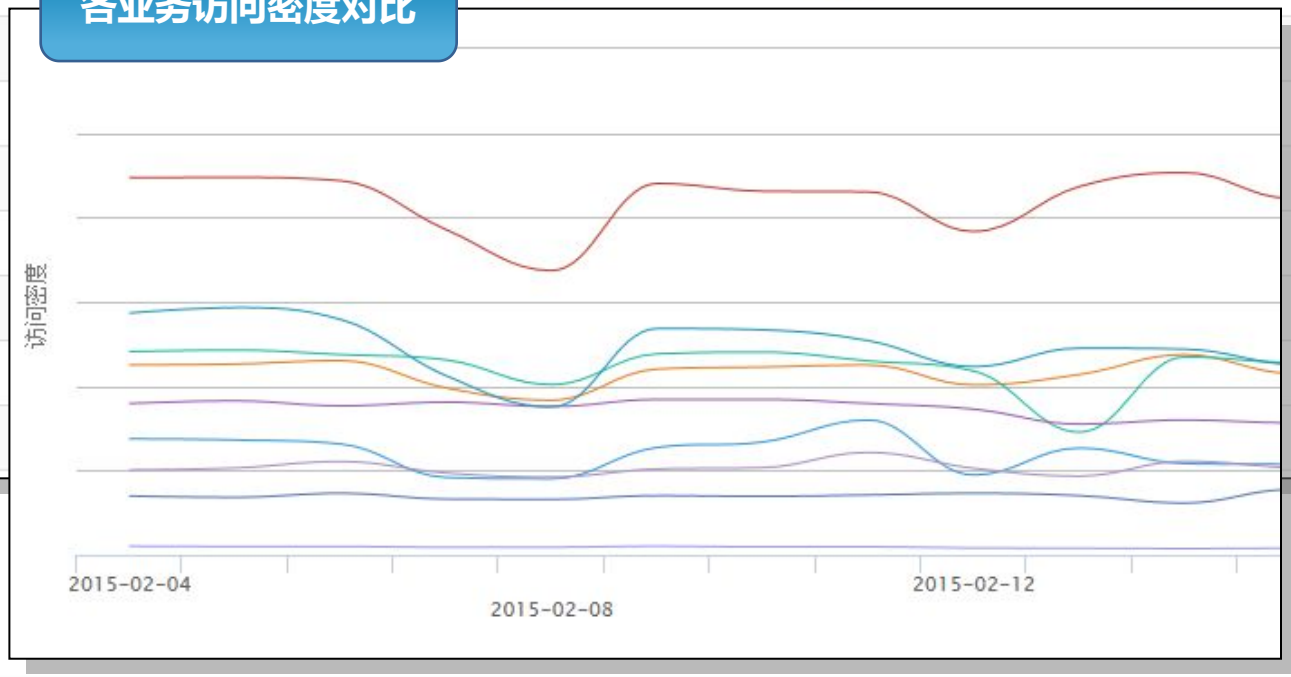


可度量的成本指标

业务的访问密度表

业务名称	申请空间 (GB)	已用空间 (GB)	存储容量 (%)	请求量	访问密度
	10721.53	7440.01	69.39	3,941,507	367.63
	7678.16	5055.54	65.84	3,701,672	482.1
	3628.20	2272.67	62.64	1,361,789	375.33
		1279.42	61.64	1,305,233	628.83
					318.78
					140.61
					473.46
					427.86
					77.36
					0
					5

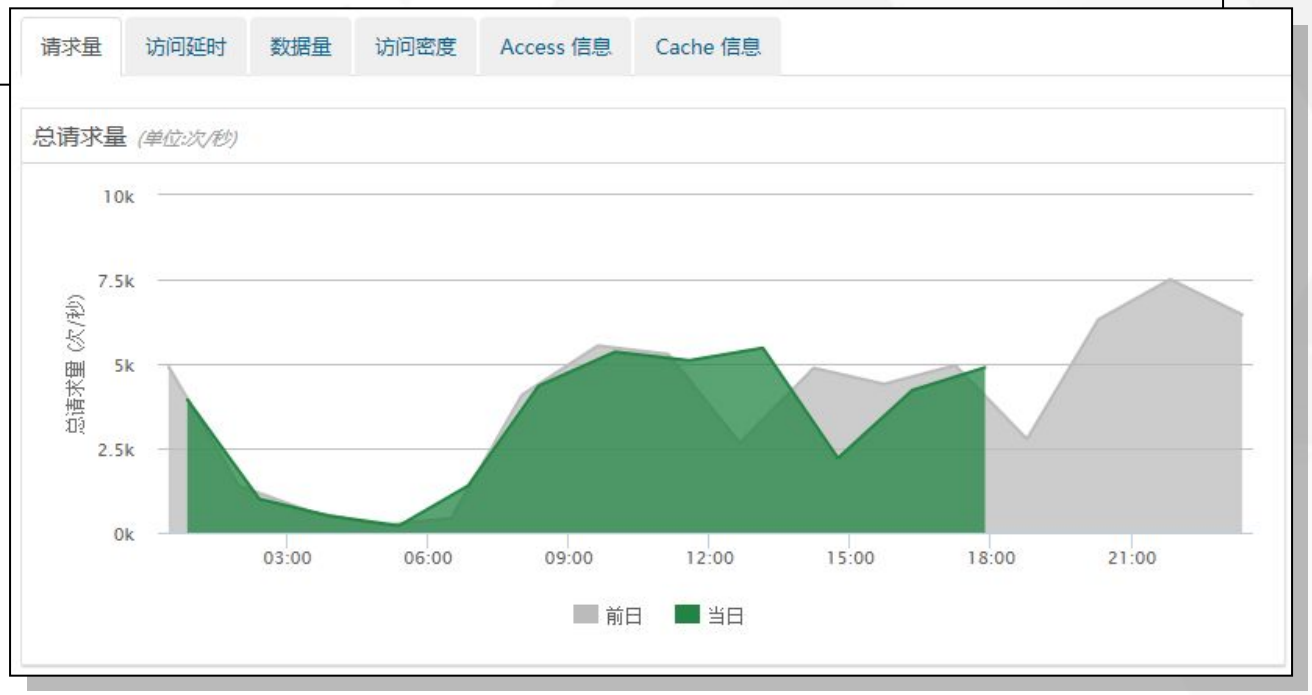
各业务访问密度对比



存储成本优化

提升访问密度的方法

-> 业务逻辑优化
-> 数据压缩
-> 淘汰过期数据
-> 冷热分层存储
-> 存储碎片整理

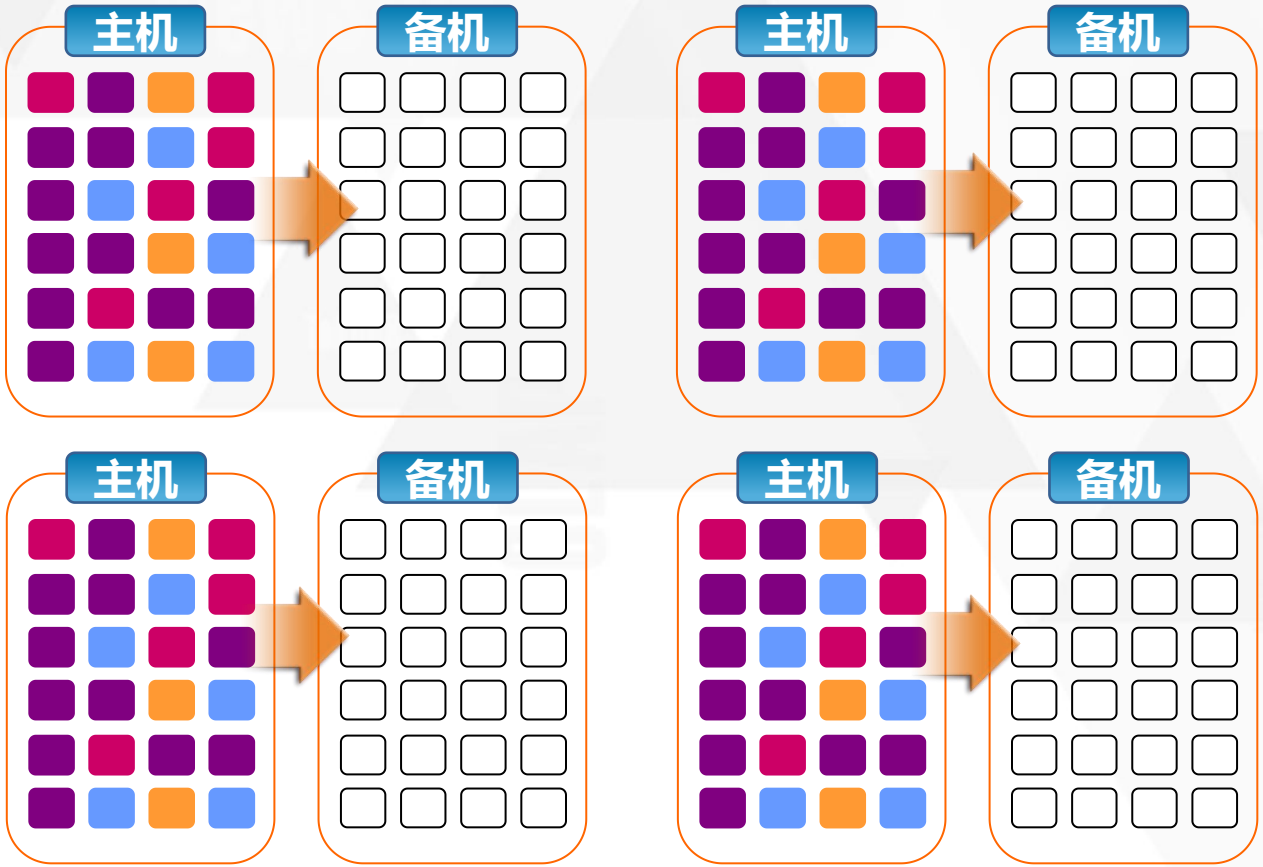


存储成本优化

访问密度趋势

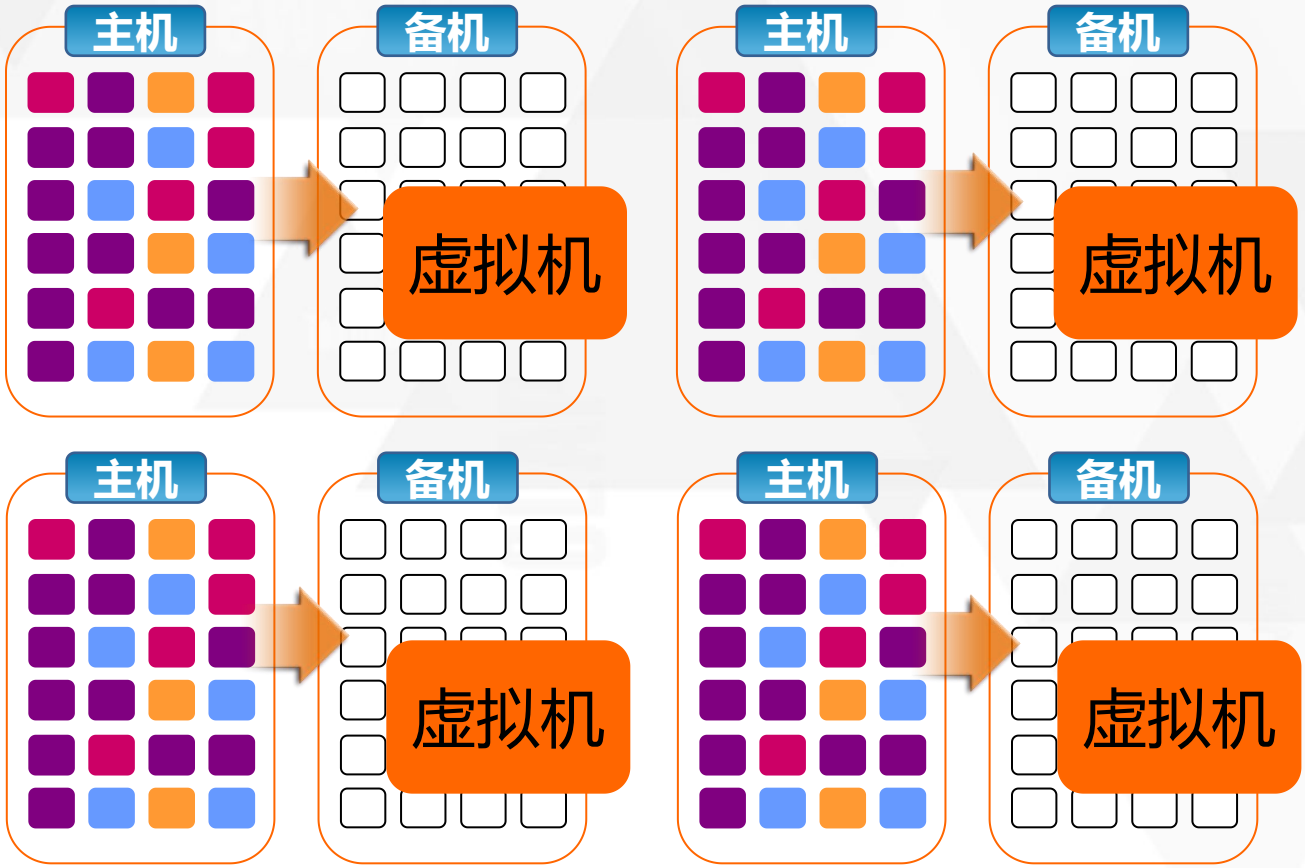
冗余备机利用

主读写，备低负载造成的资源空闲



冗余备机利用

寄居蟹虚拟机调度系统



适用于高计算、低磁盘和内存的长尾业务

冗余备机利用

虚拟机被自动销毁

寄居蟹虚拟机调度系统

故障

主机

主机

备机

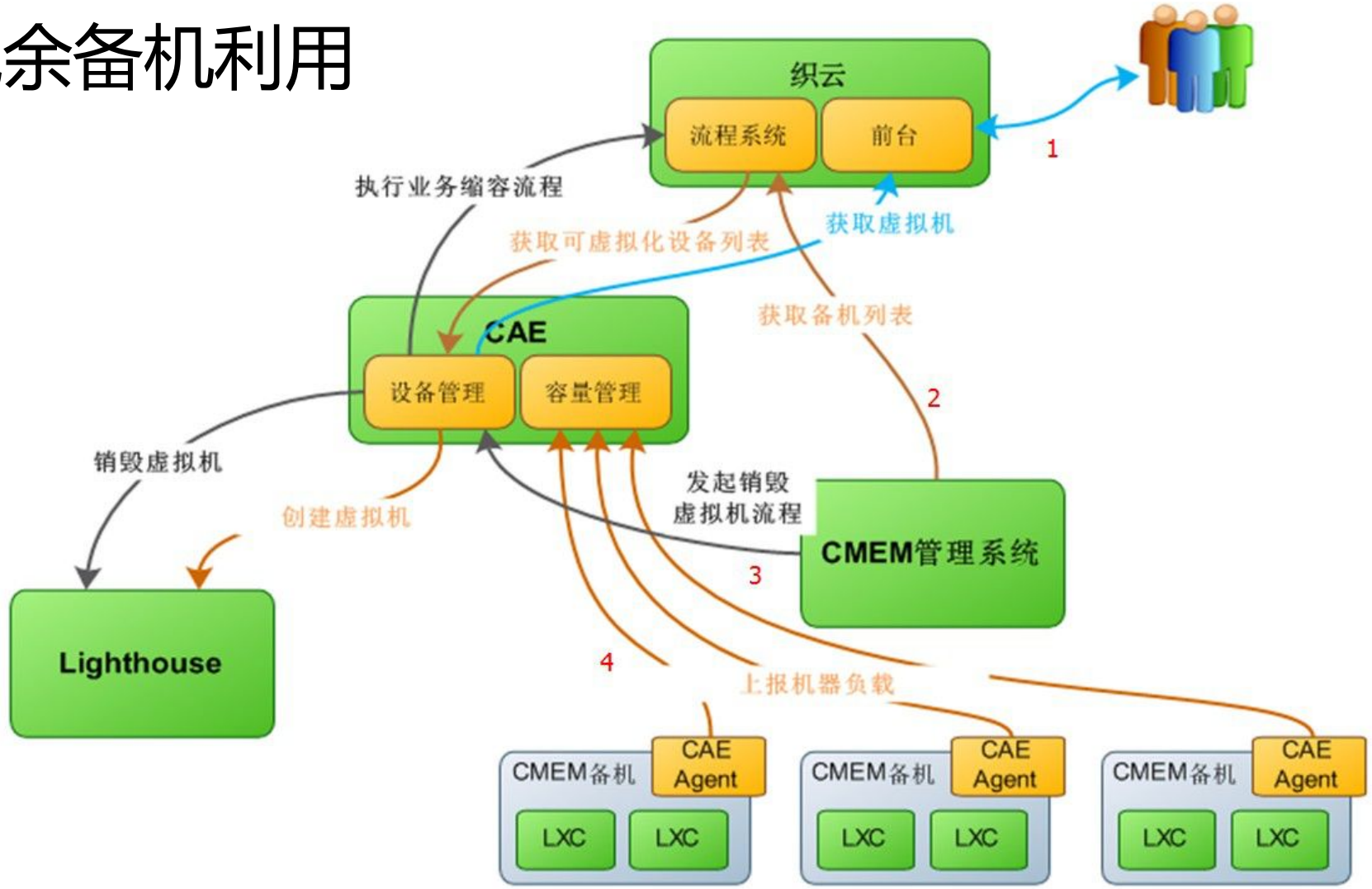
虚拟机

生产环境上千个虚拟机在备机上运行

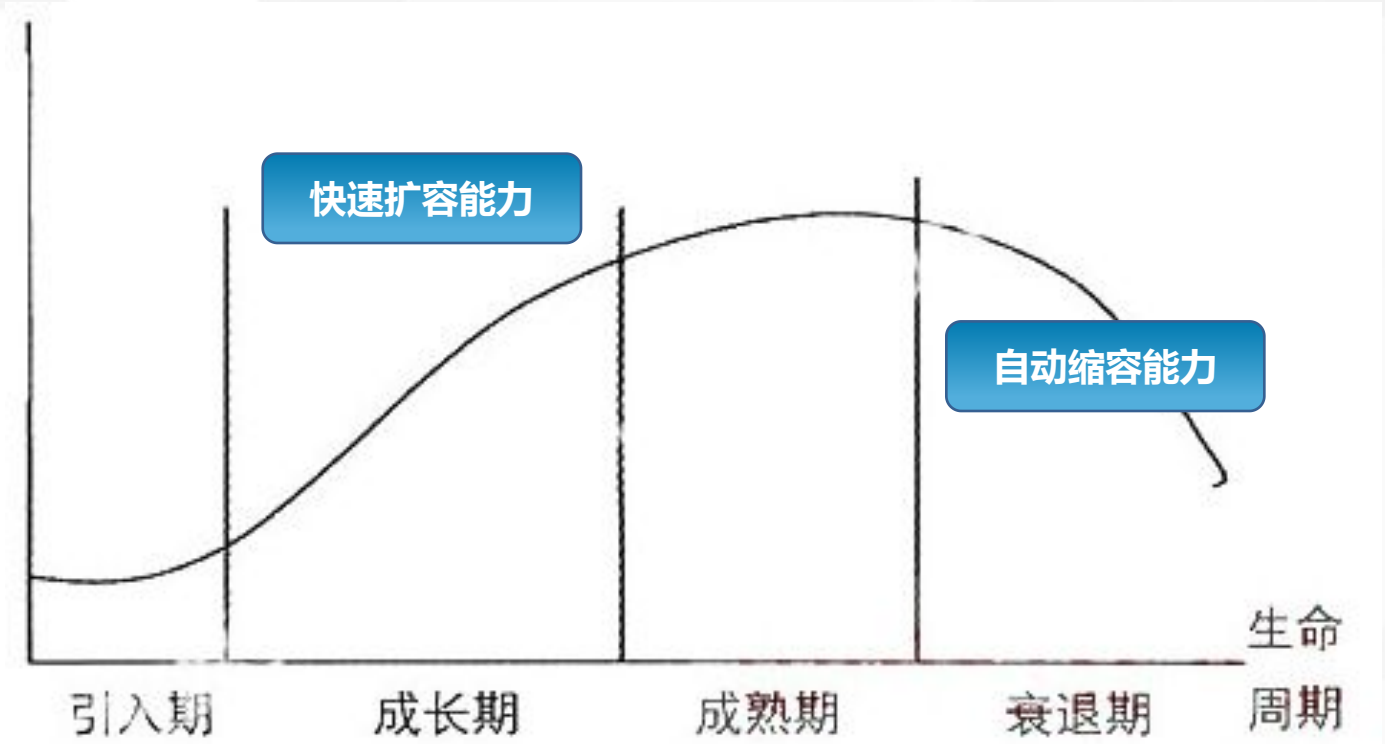
主备切换流程

1、管理机检测到主死机
2、主自动切换到备
3、管理机通知调度系统
4、调度把虚拟机销毁
5、管理机启动数据搬迁
6、调度重上线虚拟机

冗余备机利用


织云
流程系统
前台
1
执行业务缩容流程
获取可虚拟化设备列表
获取虚拟机
获取备机列表
CAE
设备管理
容量管理
2
销毁虚拟机
创建虚拟机
发起销毁
虚拟机流程
CMEM管理系统
3
Lighthouse
4
上报机器负载
CMEM备机
CAE Agent
LXC
LXC
CMEM备机
CAE Agent
LXC
LXC
CMEM备机
CAE Agent
LXC
LXC

业务自动伸缩



业务自动伸缩

- 通过L5实现接入机扩/缩容
- 通过增/减存储块实现扩/缩容
- 扩容对业务基本无损
- 调度定期扫描业务容量，自动伸缩

业务自动伸缩

自动缩容

机房	SET	Bid	缩容cu个数	迁移后使用率	缩容次数
深圳机房	仓库	101020086	12	71%	2
深圳机房	ayC	20500026	3	63%	1
深圳机房	ayC	20500025	1	0%	1
深圳机房	ayC	20500024	10	77%	2
深圳机房	ayC	20500021	16	79%	3
深圳机房	ayC	20500020	3	9%	3
深圳机房	ayC	20500019	5	70%	3
深圳机房	ayC	20500009	5	0%	5
深圳机房	ayC	20500008	5	0%	5
深圳机房	ayC	20500012	11	0%	8
深圳机房	ayC	20500013	2	76%	1
深圳机房	ayC	20500011			
深圳机房	ayC	20500014			
深圳机房	ayC	20500015			
深圳机房	营销	101020377			

自动扩容

机房	完成个数	失败个数	涉及cu个数	成功率
机房	32	2	93	94%
机房	11	2	13	84%
自营	9	0	53	100%
机房	18	8	18	69%
5D机房	11	1	68	91%
总计	81	13	245	86%

每周二百多起实例自动扩缩容

业务自动伸缩

鹦鹉螺系统邮件 2015年1月29日

CMEM缩容报告：

hi han
bid 四级模块

.qq.com][CMEM]

在 2015-01-28 发生了缩容。
缩容可能由于满足如下策略:

1、容量首次发现低于70%以下，持续3天
2、缩容到容量占用80%以上或者cu 个数为1为止
3、内部业务bid上线30天后，才会触发自动缩容

请关注该业务bid是否申请容量过大或者业务尚未上线而造成
需要恢复容量，烦请联系

邮件由系统自动发送，请勿直接回复！有疑问请联系:

鹦鹉螺由 SNG - 社交网络运营部 - 数据运维组 开发维护

动态平衡

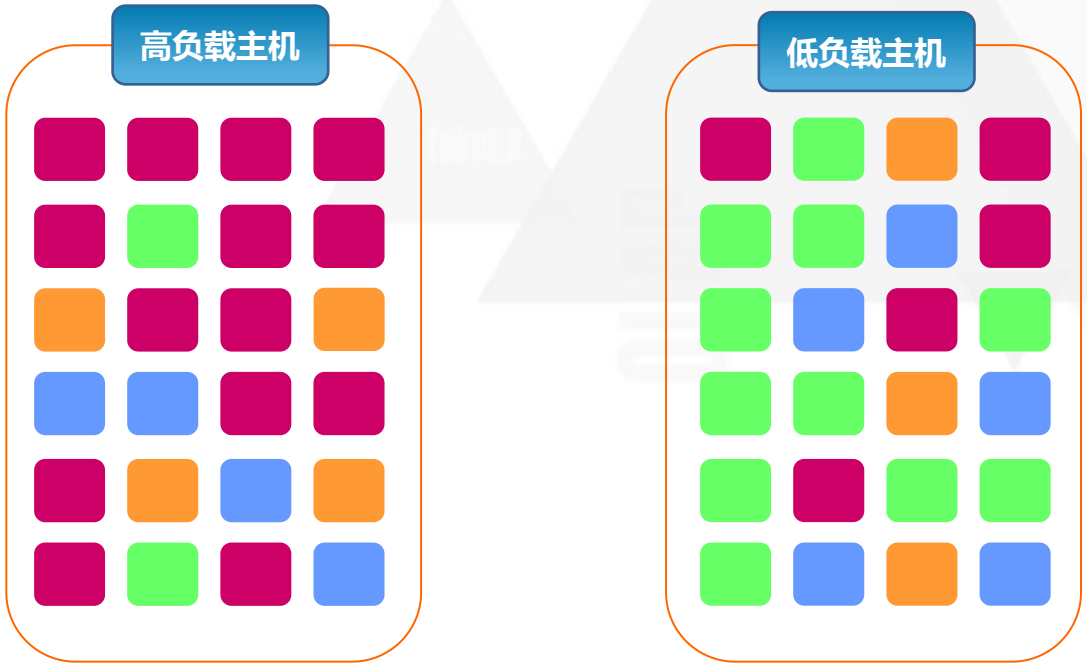
胖子和长尾业务混布造成负载不均

高负载块

中负载块

低负载块

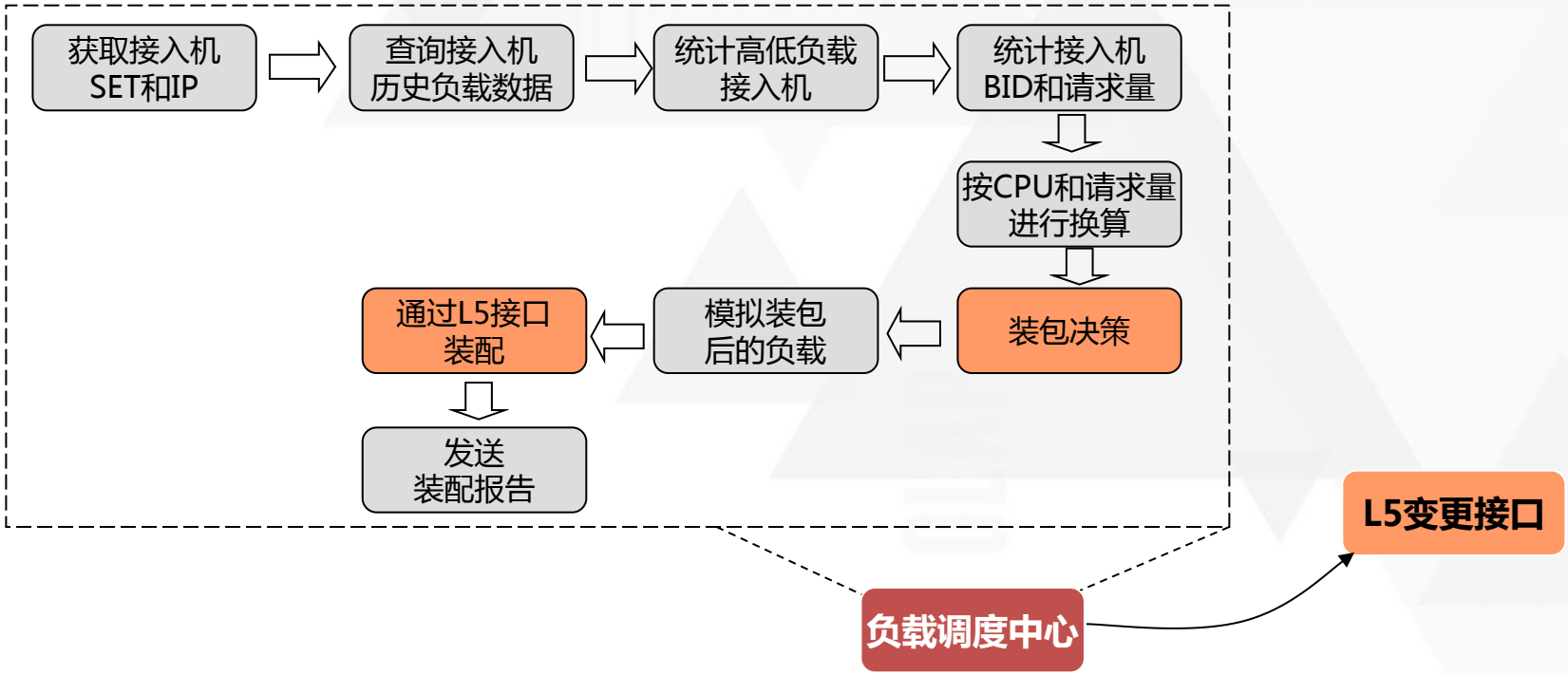
极低负载块



动态平衡

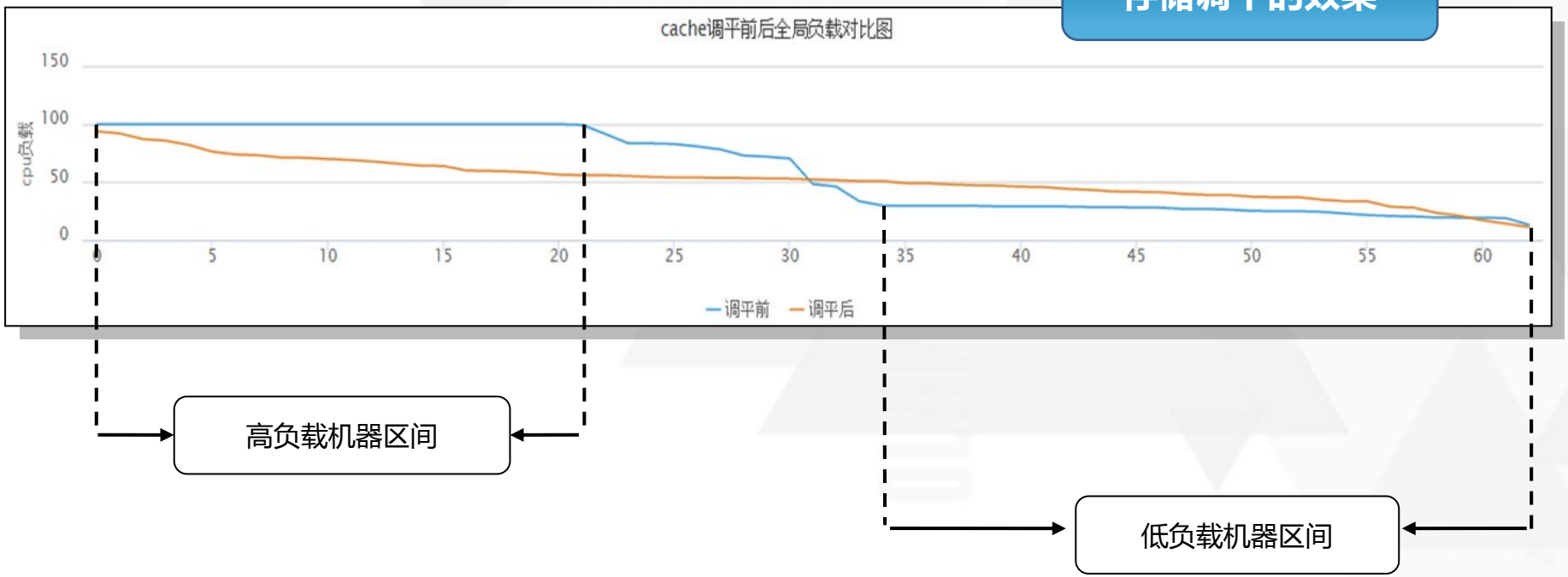
动态平衡

调平逻辑流程


获取接入机
SET和IP
查询接入机
历史负载数据
统计高低负载
接入机
统计接入机
BID和请求量
按CPU和请求量
进行换算
装包决策
模拟装包
后的负载
通过L5接口
装配
发送
装配报告
负载调度中心
L5变更接口

动态平衡

存储调平的效果



动态平衡

【CMEM超高负载日报 2015[illegible]】

没有满足条件的设备

【接入机/存储机调平日报】[20

接入机调平结果

当天无调平记录

各set负载分布

set名	master_ip	高负载数目
[illegible]_SET1	[illegible]32	0
[illegible]仓库	[illegible]29	0
[illegible]_SET1	[illegible]220	0
[illegible]_SET1	[illegible]15	0
[illegible]_SET2	[illegible]9	0
[illegible]_SET1	[illegible]59	0
[illegible]_SET2	[illegible]15	0
[illegible]_SET1	[illegible]217	0
[illegible]ET1	[illegible]238	0

存储机调平结果

各set调平结果分析

set名	master_ip	任务总数
[illegible]_SET2	[illegible]15	6
[illegible]_SET1	[illegible]59	11
[illegible]_SET1	[illegible]20	7
[illegible]_SET1	[illegible]215	19
[illegible]ET1	[illegible]142	9
[illegible]_SET2	[illegible]29	16
[illegible]_SET3	[illegible]78	9
[illegible]_SET1	[illegible]17	17
[illegible]SET1	[illegible]88	5
合计		99

每天有几十起
动态平衡作业
自动运行

质量优化

- 主备跨机架部署
- 大仓库内划分逻辑仓库
- 日志上报和分析
- 以业务为维度进行延迟定位
- 跨IDC或跨城容灾

质量优化

日志挖掘定位网络问题

Bid 101021048 起始时间 2014-11-14 13:00:00 结束时间 2014-11-14 14:00:00 查询

单点源ip异常

access	返回码	最大延迟	异常数目	大于50ms请求数
.41	-11001	3935.394	4	4730
.142	-11001	3994.334	4	4466
.163	-11001	2999.609	4	4272
.39	-11001	2998.377	4	4236

单点目标ip异常

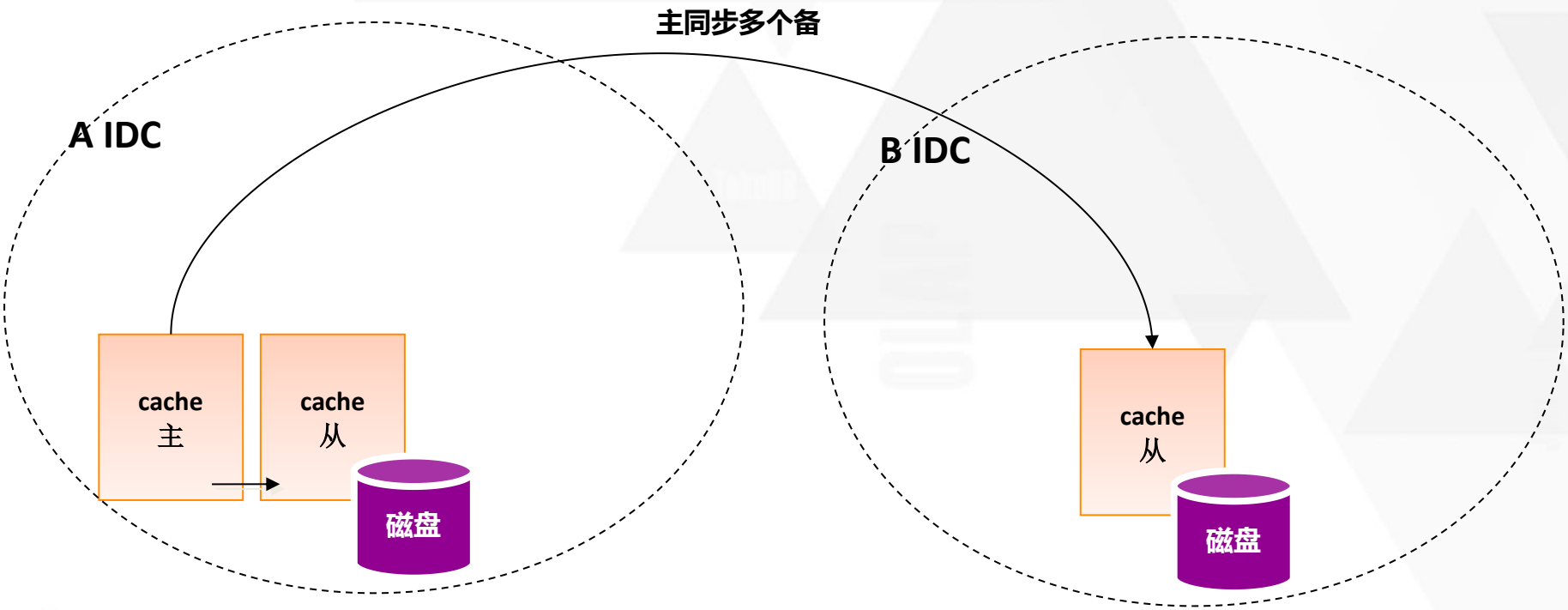
access/cache	返回码	最大延迟	异常数目	大于50ms请求数
.139	-11001	3994.334	16	17704

源&目标ip联合异常

access	access/cache	返回码	最大延迟	异常数目	大于50ms请求数
.41	10.185.21.139	-11001	3935.394	4	4730
142	10.185.21.139	-11001	3994.334	4	4466
.163	10.185.21.139	-11001	2999.609	4	4272
.39	10.185.21.139	-11001	2998.377	4	4236

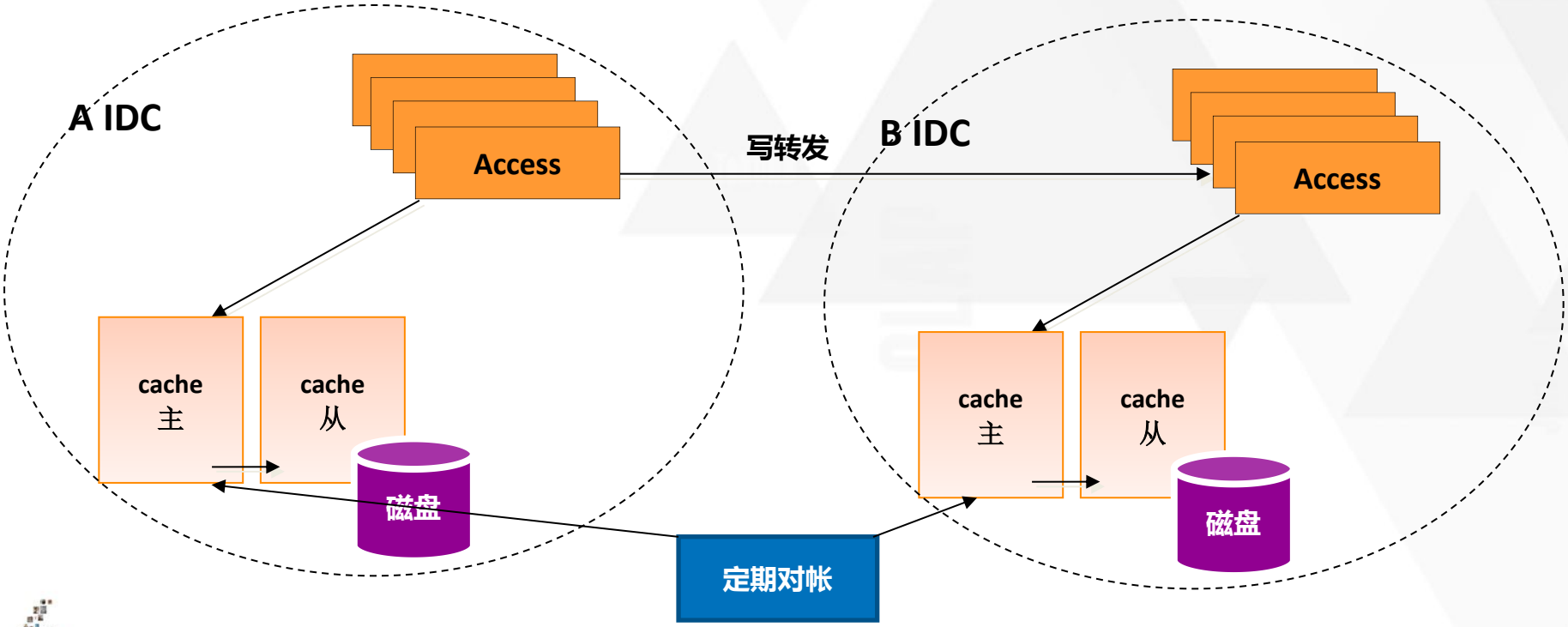
日志异步复制

1、一主多备
2、A主B备同一个IDC，C备D备部署于不同IDC
3、A主日志异步复制到B、C、D



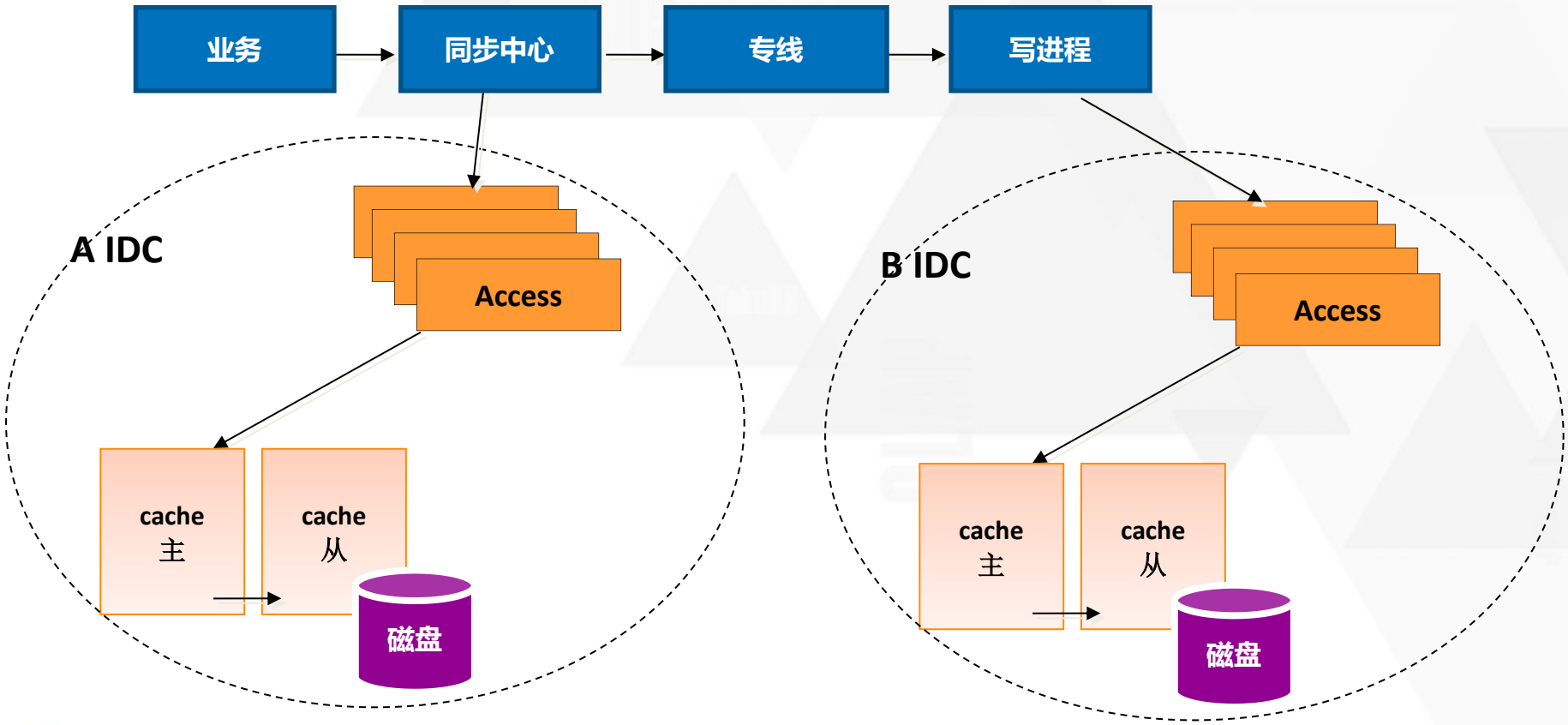
仓库复制

1、源接入机实时同步到复制仓库的目标接入机
2、同步故障时落盘流水，从流水恢复
3、对帐中心定期对帐



同步中心

1、业务写同步中心
2、同步中心写多地的仓库
3、错误时记流水，从流水恢复



未来挑战：云化

- 快速搬迁能力
- 业务数据透明
- 存储隔离
- 运维大数据分析能力
- 性能监控和自愈能力
- 更智能的调度平台

回顾

- 存储系统介绍：亚洲最大的存储集群
- 存储运维体系：为业务服务的运维平台
- 成本：访问密度模型、备机资源池
- 自动伸缩：每周二百多个实例自动扩缩容
- 动态平衡：每天几十个作业
- 质量优化：跨交换机部署、日志上报及定位
- 跨城容灾：日志异步复制、仓库复制、同步中心

ChinaUnix ITPUB

THANKS