World of Tech WOT

51CTO传媒

2014全球软件技术峰会

Software Technology Summit

深圳站

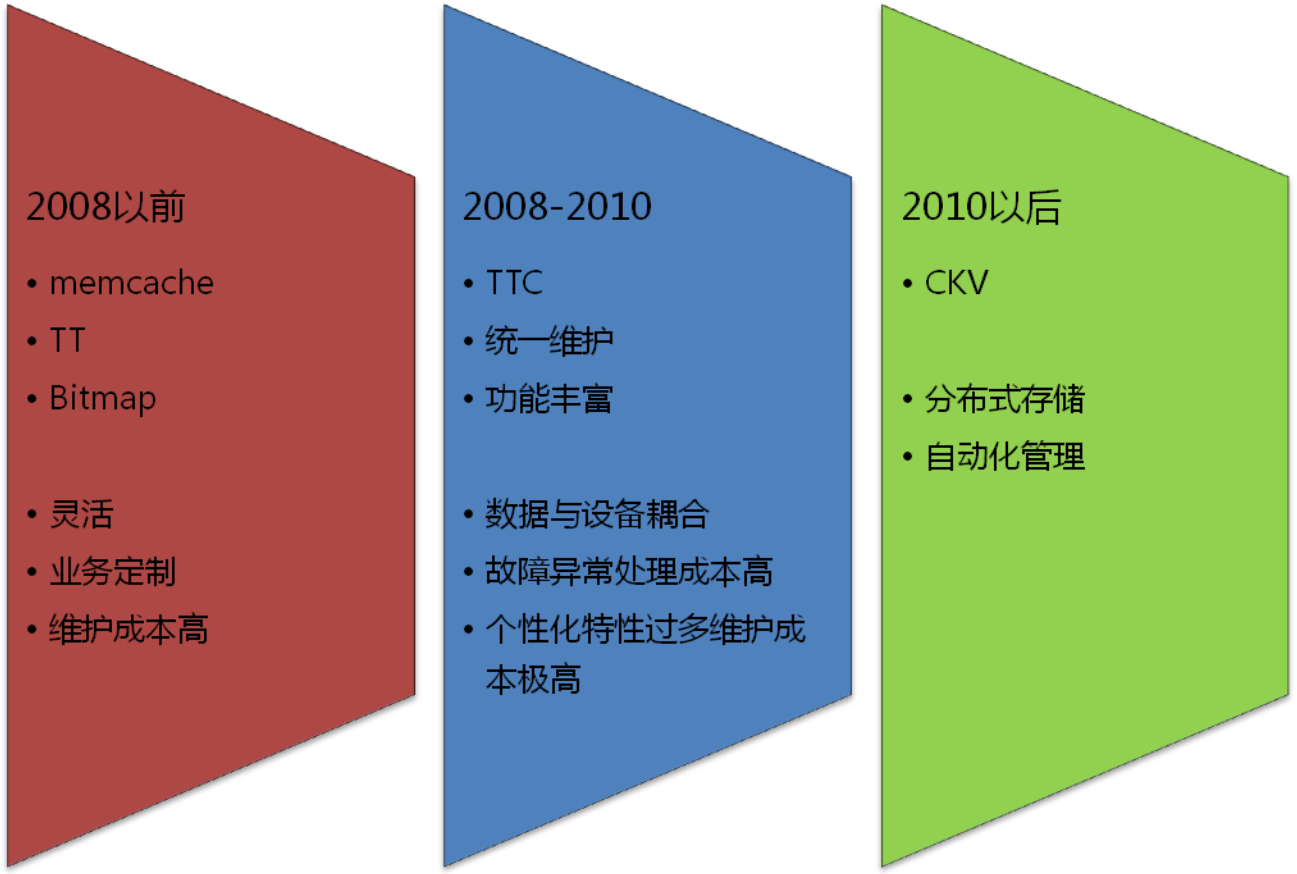
CKV分布式存储揭秘

邹润谋
腾讯社交网络运营部高级DBA

2014年11月

目录

腾讯社交网络内存发展演变



目录

概述

什么是CKV

- 分布式的内存/SSD存储系统
- Key-Value模型数据
- 双机热备+流水磁盘备份
- 微秒级响应速度
- 支持多协议接入
- 存储无理论上限

使用CKV的优点

容灾能力

- 提供双机热备容灾+定期冷备+附加流水，数据安全性高，永久存储，数据定点恢复

扩缩容能力

- 根据实际业务数据量，平滑的增加或减少存储量，扩容或缩容，前端无需变动，接入能力扩容通过名字服务变更，与业务解耦合

自动化能力

- 自动部署、自动扩缩容、死机自动切换搬迁等

数据冷热自动调度

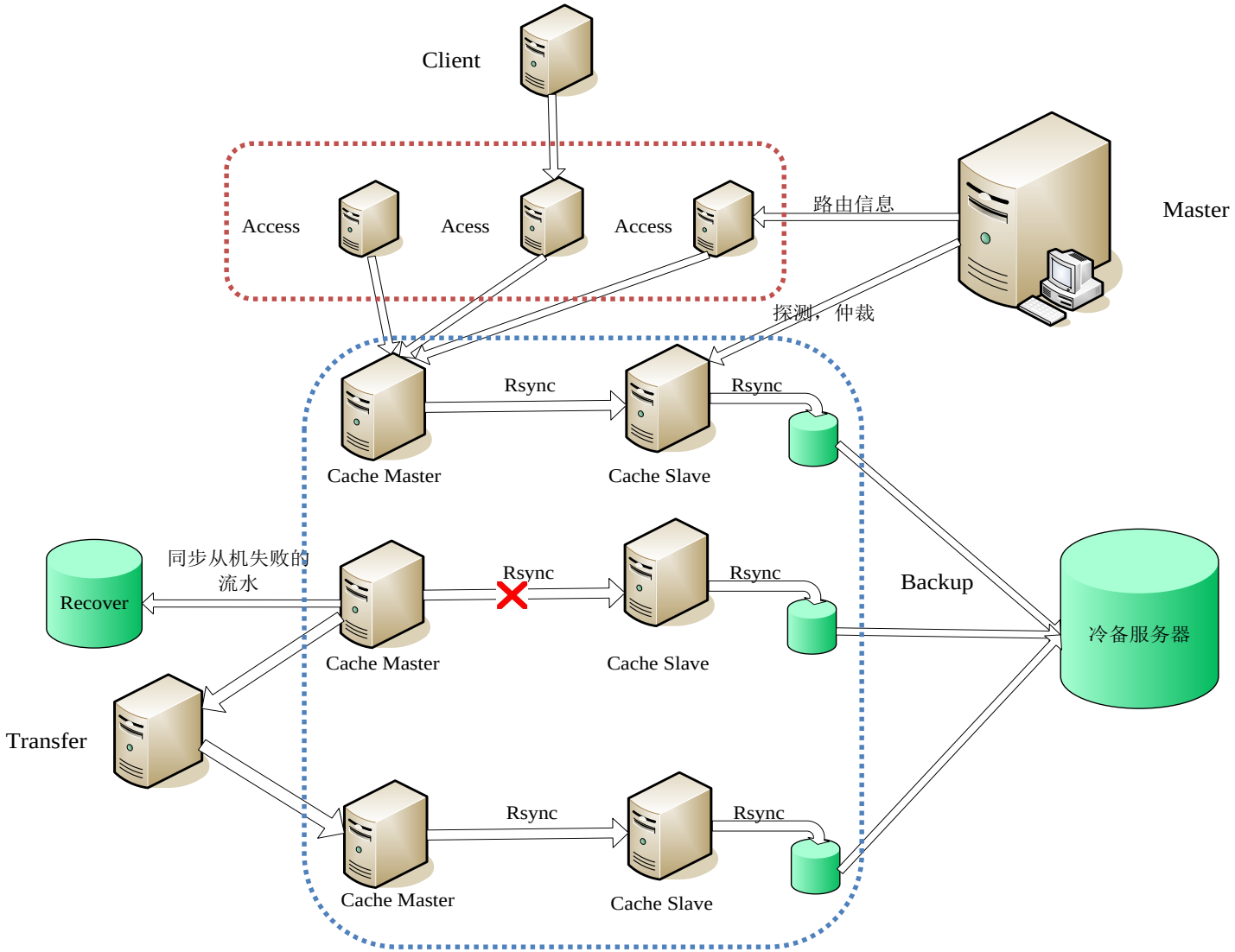
- 解决性能与成本矛盾，满足业务全生命周期的数据存取需求

CKV在腾讯社交网络运营现状

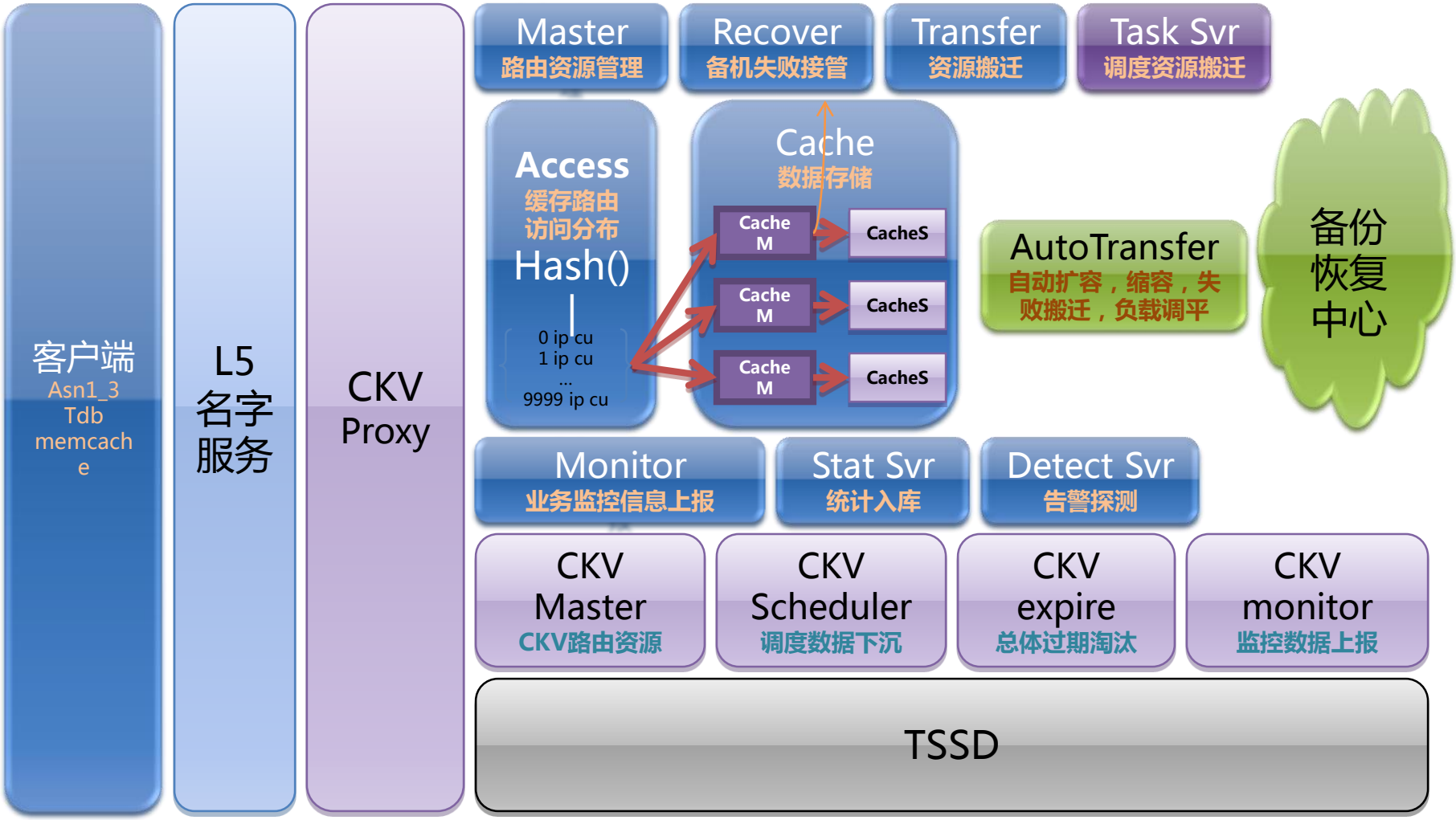
- 大范围业务覆盖，空间/广点通/相册/QQ/开放平台等
- 超过3000个子业务模块接入
- 设备规模超万台，TB级内存存储+PB级SSD存储
- 日峰值访问量超过4千万/秒

目录

CKV最早模型

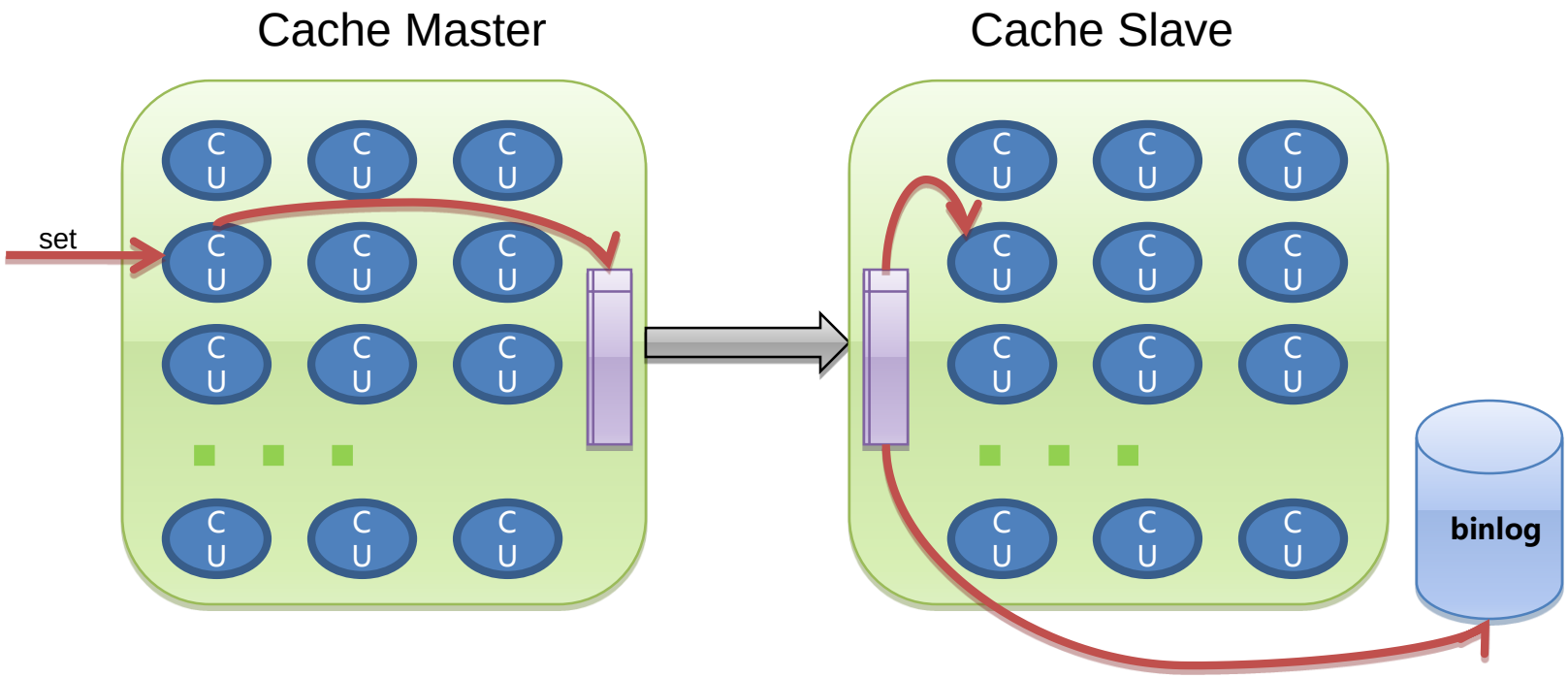

Client
Access
Acess
Access
路由信息
Master
探测，仲裁
Rsync
Rsync
Cache Master
Cache Slave
同步从机失败的
流水
Recover
Rsync
Rsync
Backup
冷备服务器
Cache Master
Cache Slave
Transfer
Rsync
Rsync
Cache Master
Cache Slave

CKV组件模型


客户端
Asn1_3
Tdb
memcach
e
L5
名字
服务
CKV
Proxy
Master
路由资源管理
Recover
备机失败接管
Transfer
资源搬迁
Task Svr
调度资源搬迁
Access
缓存路由
访问分布
Hash()
0 ip cu
1 ip cu
...
9999 ip cu
Cache
数据存储
Cache
M
CacheS
Cache
M
CacheS
Cache
M
CacheS
AutoTransfer
自动扩容，缩容，失
败搬迁，负载调平
备份
恢复
中心
Monitor
业务监控信息上报
Stat Svr
统计入库
Detect Svr
告警探测
CKV
Master
CKV路由资源
CKV
Scheduler
调度数据下沉
CKV
expire
总体过期淘汰
CKV
monitor
监控数据上报
TSSD

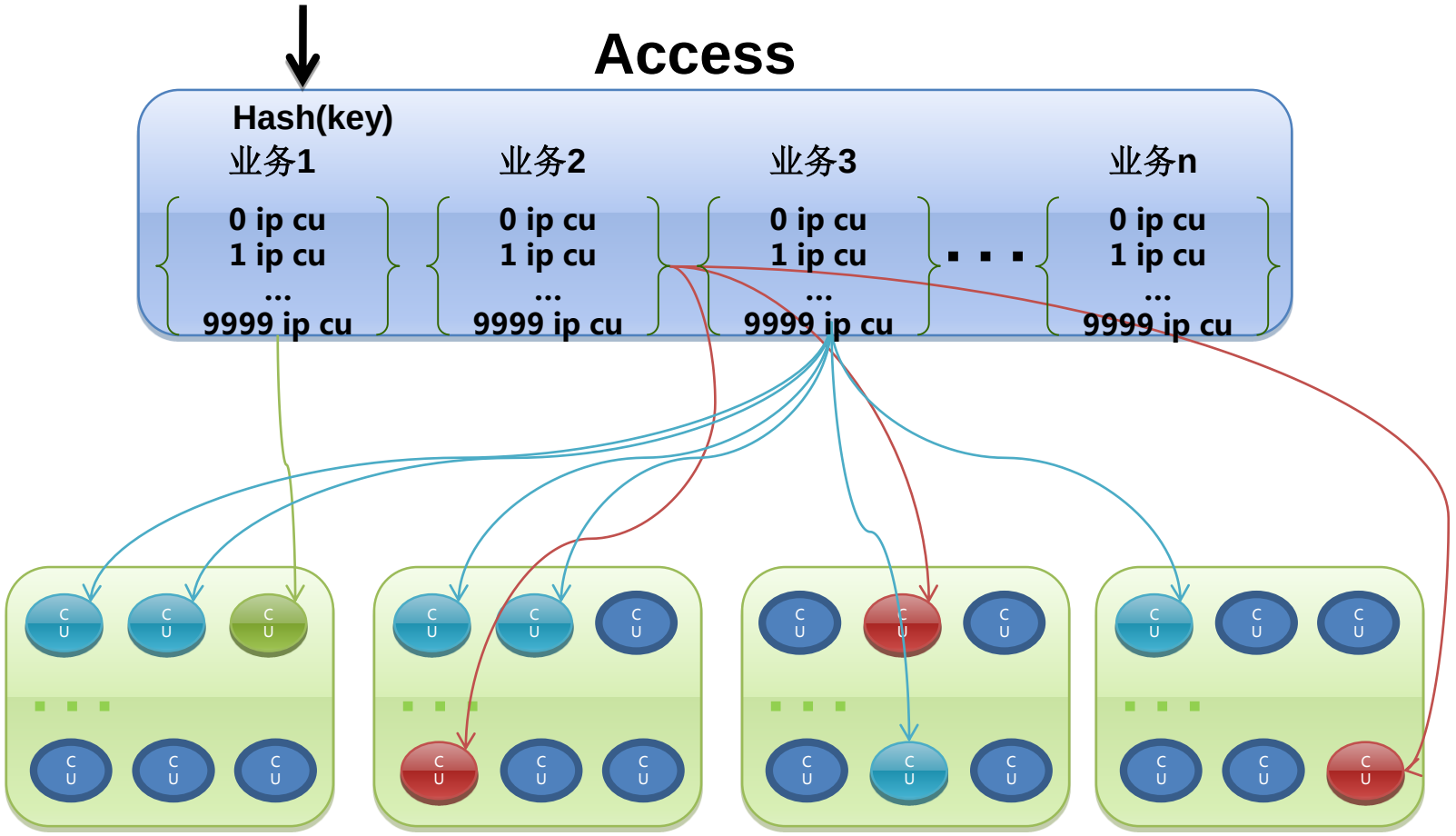
目录

存储组织



CU:CacheUnit，CKV中最小存储单元，默认为1GB，一对机器提供56个CU
里面包括K/V存储，索引，元数据

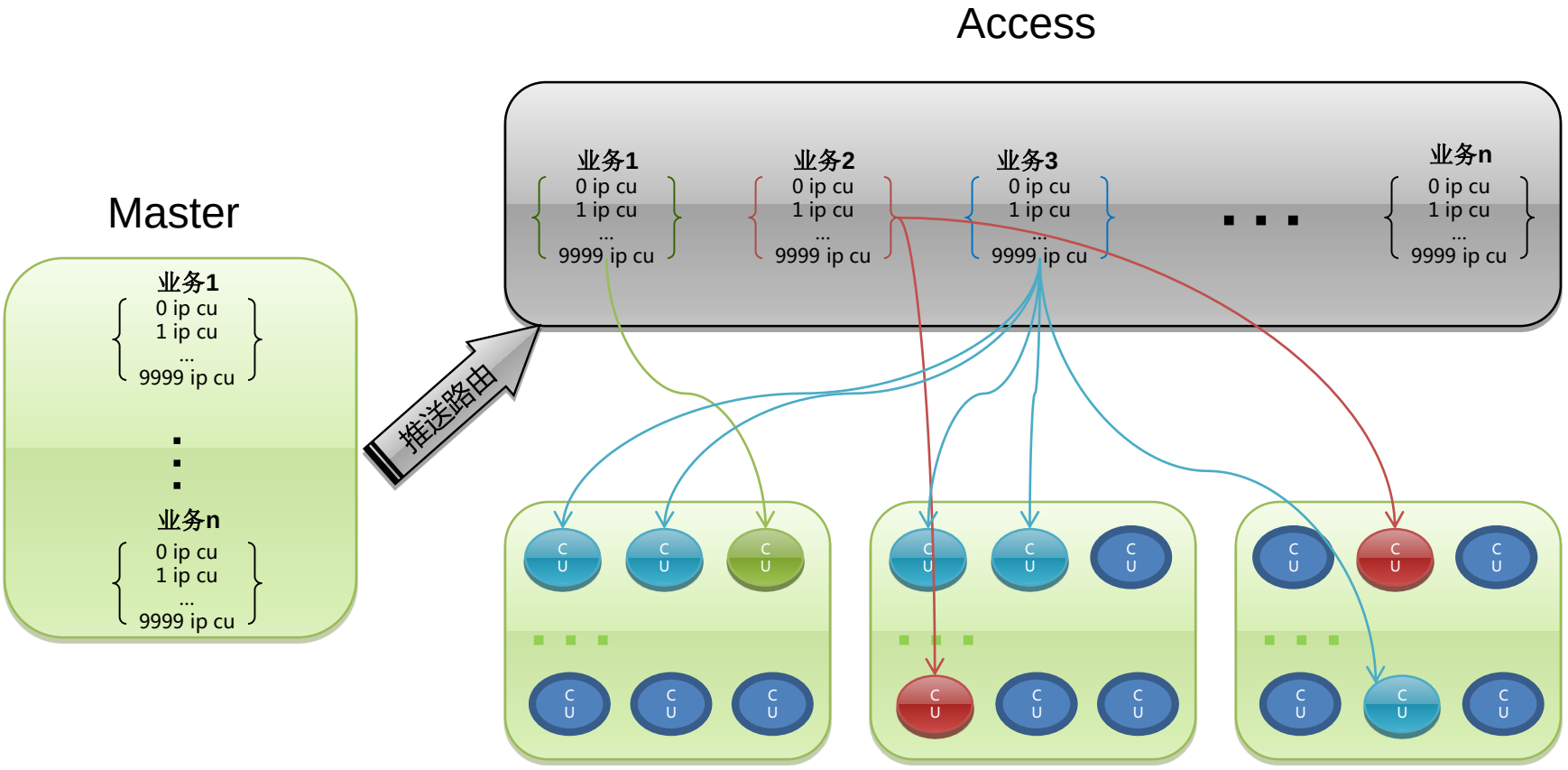
数据路由



路由表由1万个格子（桶）组成，每个格子定义后端指向的CU

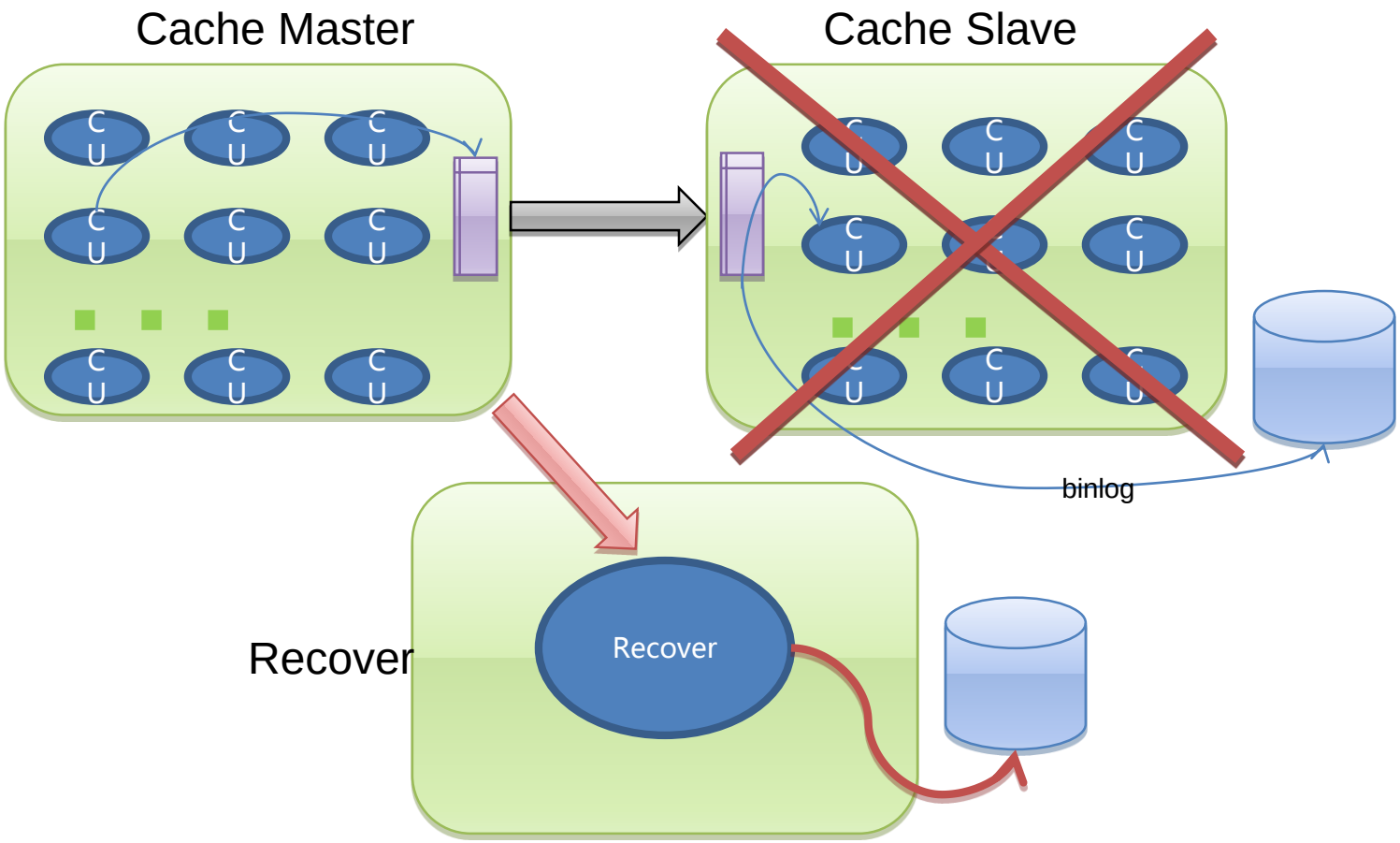
一个格子最多对应一个CU，一个业务最大容量是:CU*10000

最小模型



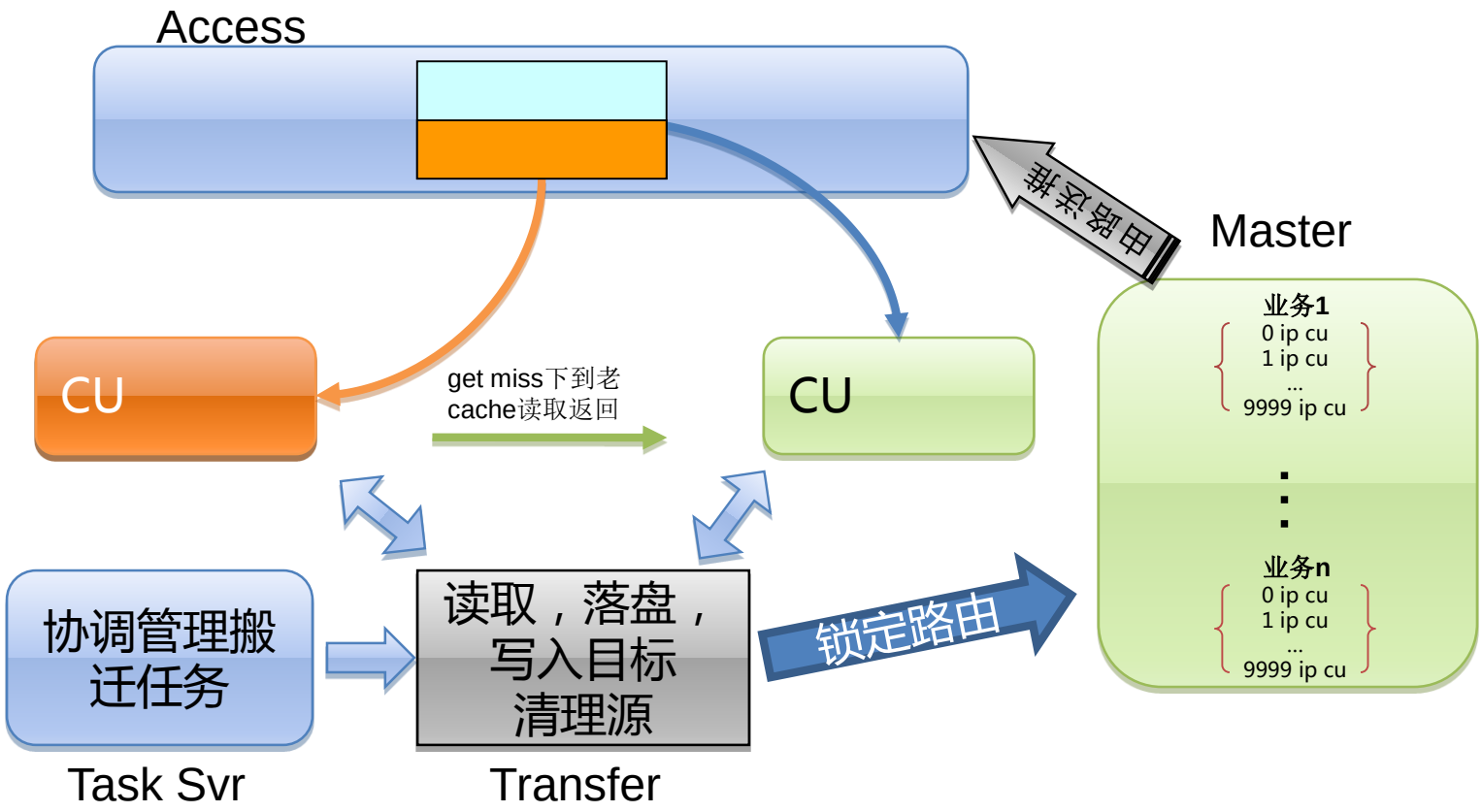
管理存储资源，资源路由关系，变更时推送

死机接管



备机恢复后自动到Recover拉取变更列表恢复

数据搬迁



多任务并发搬迁，异常自动重搬，优先级判断

所有扩缩容/死机切换搬迁等都是有基本的搬迁流程组成

目录

CKV自动化管理

自动部署

- 百台设备的部署时间<1小时

自动扩缩容

- 每周超过200次自动扩缩容动作，运维无需参与

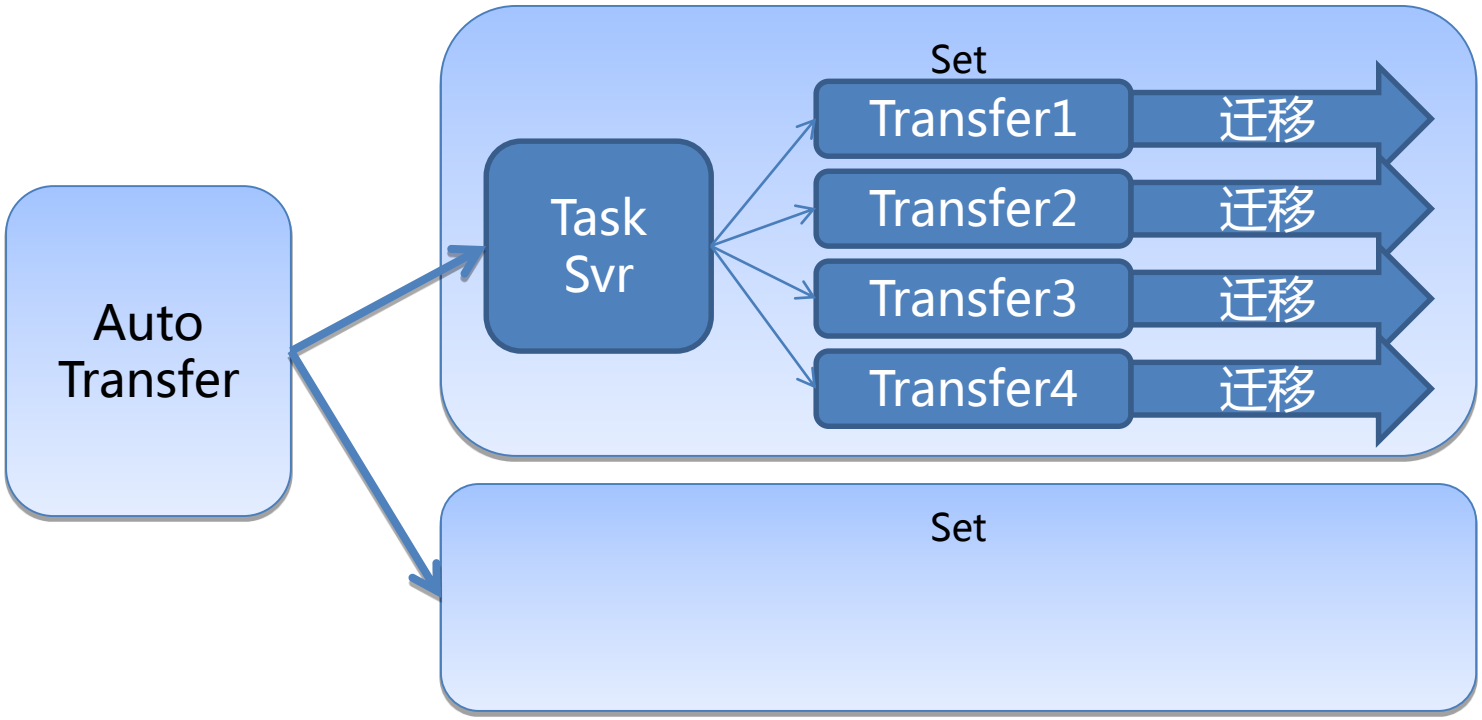
自动切换恢复

- 平均每天超过1台机器死机，系统自动切换搬迁恢复

自助业务接入

- 每周超过20个新业务接入自动化流程

自动搬迁



定期扫描死机、扩容、缩容、调平、生成任务给Task Svr

实例弹性

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	1	21%	1
深圳机房	ayC	20500014			
深圳机房	ayC	20500015			
深圳机房	营仓	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

业务自助接入

CKV业务上线

业务信息（以负载做设备需求考核,请务必如实填写）

实例名称	Qzone 英文	业务预计上线日期	2014-11-2 上线时间
实例描述	Qzone空间上线 中文	存放地点	深圳
模块信息	---请选择--- ---请选择--- ---请选择--- ---请选择---	是否开启Expire	是 Expire详细信息
		数据是否容忍丢失	是
申请人	runmouzou;miller	是否需要存储操作流水	是 流水产生机器差异
申请方运营负责人	为业务接口人，最好能为一主一备，使用半角分号(;)分隔	容量推算及其他	

上线记录数	50000000 条	网卡出流量峰值	Mbps
预估未来3个月增加记录数	1000000 条	网卡入流量峰值	Mbps
最终记录数	100000000 条	3个月后所需存储空间	G
平均单条记录大小	84 字节	当前所需存储空间	G
该业务每秒读峰值	150000 次/秒	最终所需存储空间	G
该业务每秒写峰值	35000 次/秒		
Key长度	4 字节		
协议	asn13		

提交

目录

CKV精细化运维

趋势分析预测

- 实时分析业务增长趋势以及未来一段时间的增长预测

设备负载调平

- 将分布式环境中的资源与机器进行拆装操作，消除高低负载

业务模拟拨测

- 模拟业务请求并将延迟成功率入库分析，提前发现异常

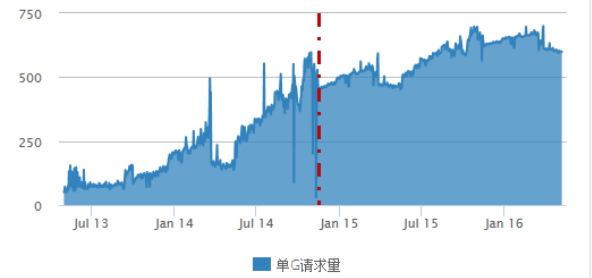
业务Profile

- 分析每个业务的访问模型，冷热配比，访问密度，将业务部署在最适合的介质和设备上

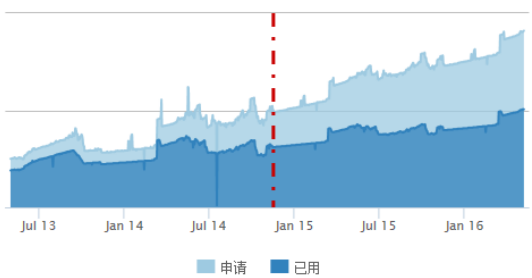
趋势分析预测

基准数据 12个月 预测区间 18个月 预测

访问密度 (次每秒)



存储量 (TB)

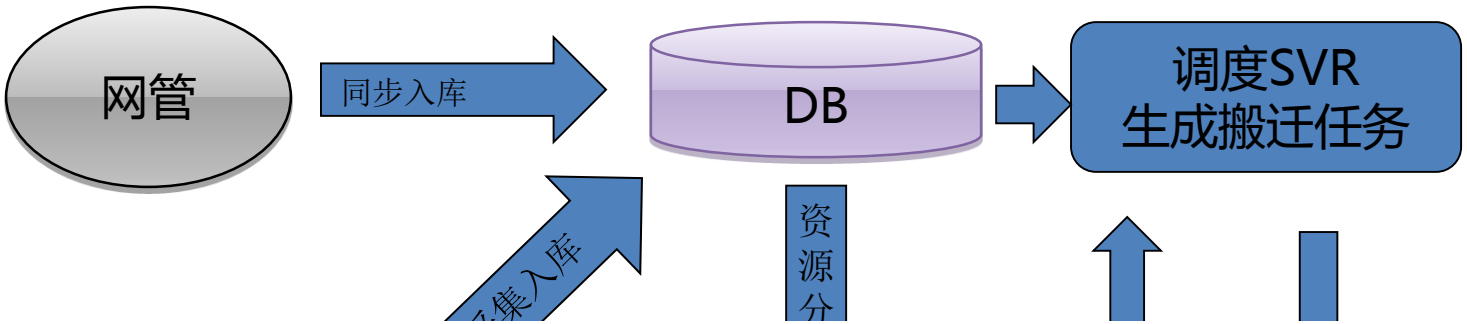


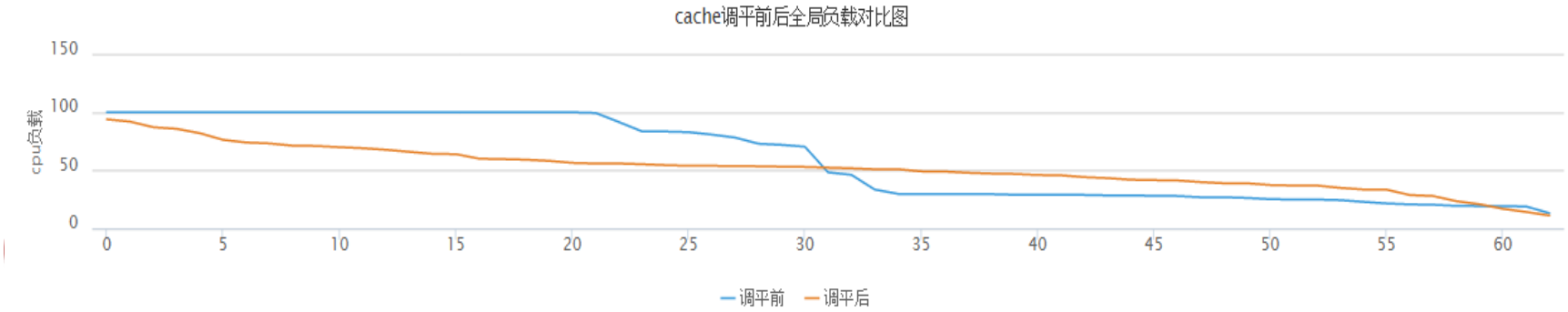
请求量 (单位:万次每秒)

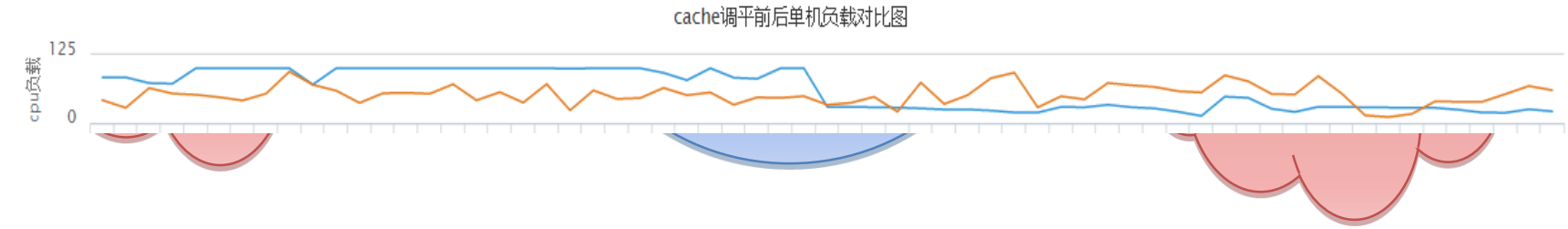
筛选：

业务名称	申请空间 (GB) 今天 | 20160510 | ↑↓	已用空间 (GB) 今天 | 20160510 | ↑↓	存储容量 (%) 今天 | 20160510 | ↑↓	请求量 今天 | 20160510 | ↑↓	访问密度 今天 | 20160510 | ↑↓
	| 52.57%↑	| 28.57%↑	| 15.74%↓	| 113.98%↑	| 40.26%↑
	| 154.92%↑	| 102.71%↑	| 20.49%↓	| 275%↑	| 47.1%↑
	7 | 7.6%↓		100%↓	0.37%↓	| 7.83%↑
	938.85%↑	789.43%↑	| 14.5%↓	62.34%↑	| 36.22%↓
	| 97.35%↑	| 94.19%↑	| 1.61%↓	81.36%↑	2 | 8.1%↓
	462.87%↑	| 635.5%↑	| 30.69%↑	256.11%↑	| 36.73%↓

设备负载调平


网管
同步入库
DB
调度SVR
生成搬迁任务
集入库
资源分


cache调平前后全局负载对比图
cpu负载
150
100
50
0
0
5
10
15
20
25
30
35
40
45
50
55
60
调平前
调平后


cache调平前后单机负载对比图
cpu负载
125
0

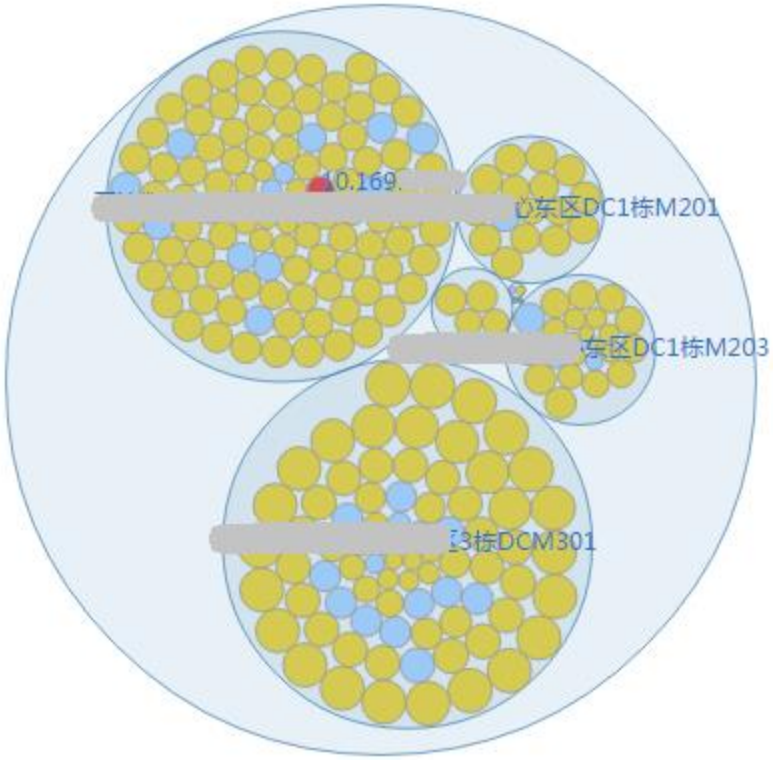
业务模拟拨测

Bid 101021048 起

Bid 108010021 →查询 查询

10.169 [SET_G21Q][逻辑SPP]

调用总数71808调用成功:65448 调用失败:63 成功率:99.90%



单点源ip异常

access	请求数
.41	
.142	
.163	
.39	

单点目标ip异常

access/cache	请求数
139	

源&目标ip联合异常

access					于50ms请求数
.41					30
142					56
.163					72
.39	139	-11001	2998.377	4	4236

谢谢大家

邮箱: runmouzou@tencent.com

微信: runmou

欢迎加入腾讯社交网络数据运维团队