中国移动
China Mobile

中移苏研

存储产品规划和实践经验分享

云计算产品部 刘军卫

中国移动苏州研发中心

2017年4月

公司及部门简介

中国移动 China Mobile

中移（苏州）软件技术有限公司（中国移动苏州研发中心）

- 中国移动全资子公司，注册资本7亿，园区占地480亩，建筑面积36万平方米
- 现有正式员工750余人，硕士以上占75%，研发人员86%，远期规模4500人
- 云计算、大数据、IT支撑系统以及部分应用产品等开发和应用
- 中国移动IT能力内化和业务创新发展的中坚力量
- 已负责总部一级私有云、公有云、性能管理、云化OA等一级平台的建设工作，助力集团加快IT系统1+N两级架构的集中化进程，促进研发运营一体化。

云计算产品部

目前共有210人左右，研发人员超过80%，面向内外部客户提供以下产品和服务：

- 提供云计算相关的标准化和定制化产品及解决方案；
- 提供构建云计算资源池的软硬件集成服务和技术支撑服务；
- 提供云计算咨询、应用云化迁移和容器化迁移服务；

产品和服务已经在集团公有云、集团一级私有云及省/专业公司项目中商用，

累计部署规模超过20000台服务器。

行业地位

以市场化机制为手段推动能力内化，推动技术实力积累与能力提升，

目前在开源、标准和行业组织等多方面均取得了长足进步。

开源	推进开源合作，解决社区Bug并回馈社区：2016主流开源项目贡献超过250个补丁。Openstack社区贡献**国内前10**，黄金会员、中国首个 **OpenStack SuperUser**；Ceph社区贡献**社区第4**；Linux基金会银牌会员，**国内贡献前5**。
标准	牵头标准制定，降低技术风险：牵头制定**工信部云服务能力评估**等**4项**云计算国家标准，参与研制《信息技术 云计算参考架构》等**8项**云计算行业标准。 工信部《云服务能力标准》增强级 **开源云计算产品联盟**（OSCAR）和**中国开源云联盟**（COSCL）**副理事长**单位
内外部奖励	**双奖**：“科技进步与业务服务创新奖励评选”情况：苏州研发中心获得科技进步类**一等奖1项**（2016），**二等奖1项**（2015），**三等奖2项**（2016）。其中一等奖项目“基于开源社区的定制化Linux操作系统应用与推广”得到李正茂副总裁和与会专家领导的高度评价。 **外部奖励**：中国通信学会科学技术奖（2014）**二等奖1项**；中国电子学会科学技术奖（2016）**三等奖1项**；

存储产品线 - 产品规划

中国移动 China Mobile

统一存储管理平台

- 自动化部署
- 用户管理
- 监控告警
- 资源统计
- 性能分析
- 存储池管理
- 块存储管理
- 对象存储管理
- 文件存储管理

网盘 | **存储网关** | **和目视频** | **无纸化办公** | **归档备份**

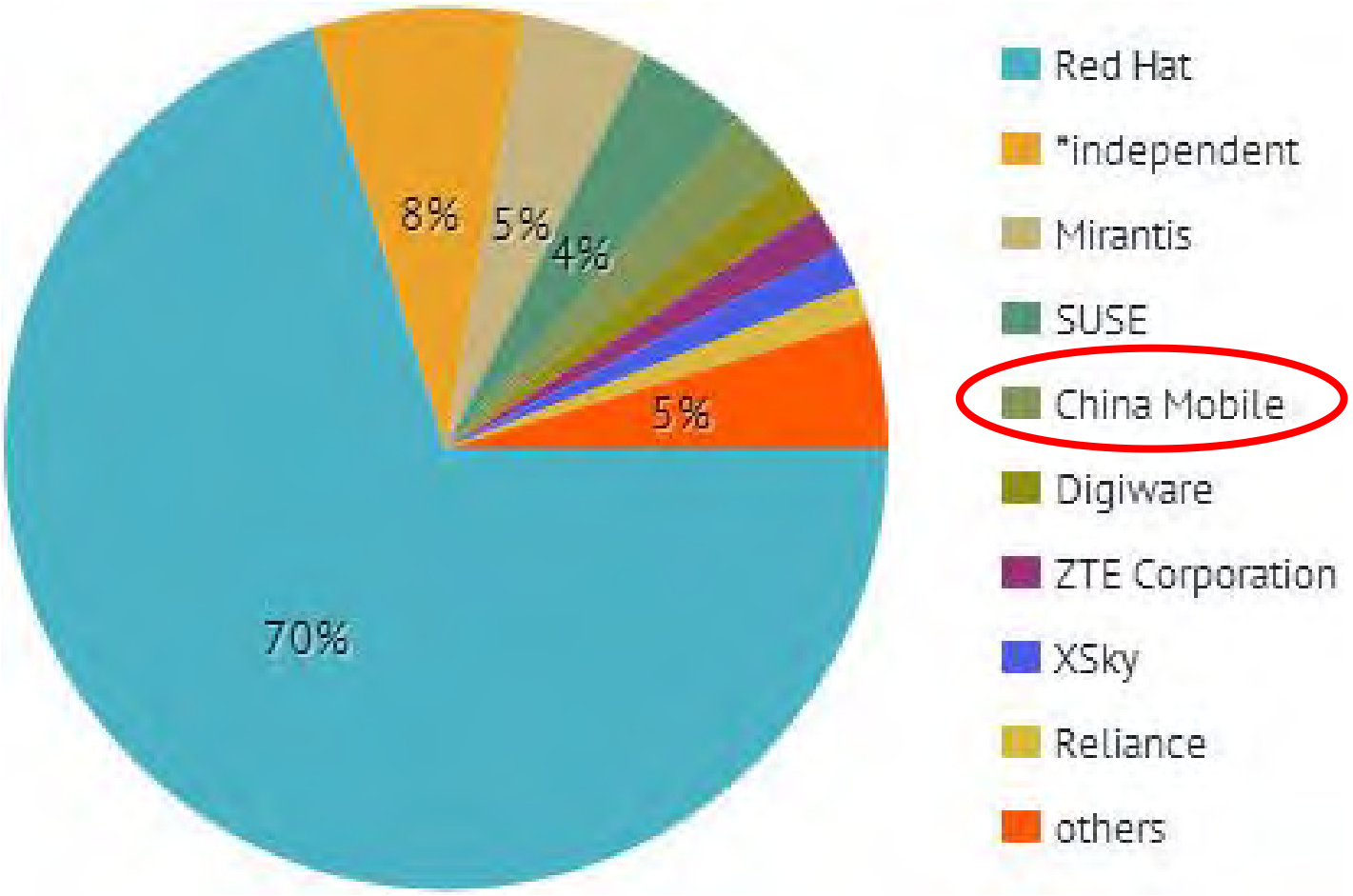
云存储 | **云硬盘** | **云化NAS** | **云化CDN**

S3 | Swift | Qemu | iSCSI | NFS | CIFS

对象存储（BC-oNest） | **块存储**（BC-EBS） | **文件存储**（BC-EFS） | **超融合存储**（BC-Cube）

一体机、定制化服务器

存储产品线 – 社区贡献



#	Company	Commits
1	Red Hat	1254
	*independent	138
2	Mirantis	81
3	SUSE	74
4	China Mobile	47
5	Digiware	31
6	ZTE Corporation	28
7	XSky	27
8	Reliance	24
9	Intel	23
10	EISOO	8
11	UMCloud	8
12	OVH	8
13	Mellanox	6
14	Cloudwatt	5
15	Virtuozzo	4
16	UnitedStack	3
17	Kylin Cloud	3
18	IBM	2
19	CERN	2
20	EasyStack	2

存储产品线 – 对象存储

中国移动 China Mobile

- 基于CEPH研发的以**对象形式存储**和管理**海量非结构化数据**的存储系统
- 支持**跨地域、跨数据中心和跨机房**的数据容灾保护，适配X86服务器
- 支持**AWS S3**和**Openstack Swift API**，以及简单易用的工具和SDK
- 在数据备份、公有云服务、文件归档、视频存储等领域应用广泛，目前已经商用超过**30PB**

高性能

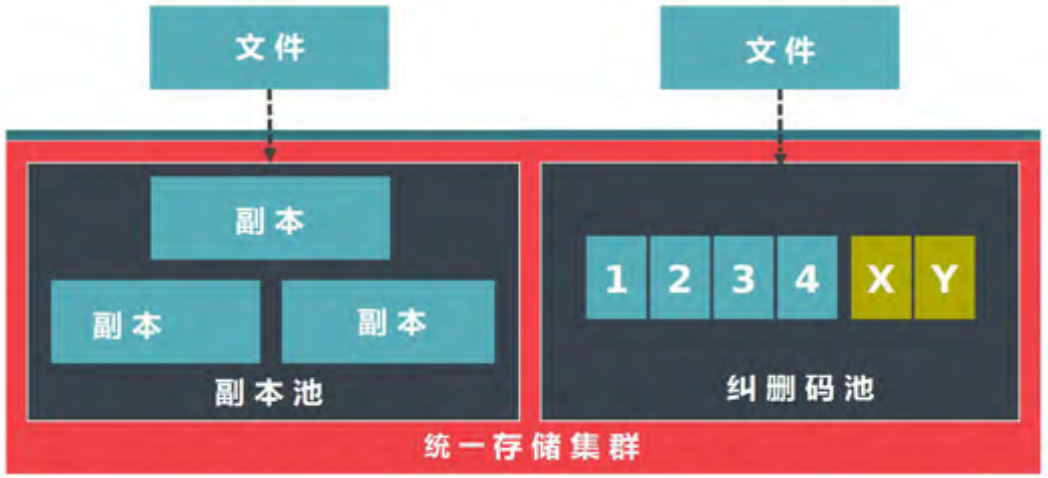
- 无需查表，算算就好
- 支持SSD缓存加速
- 自动负载均衡
- 最小化网络传输

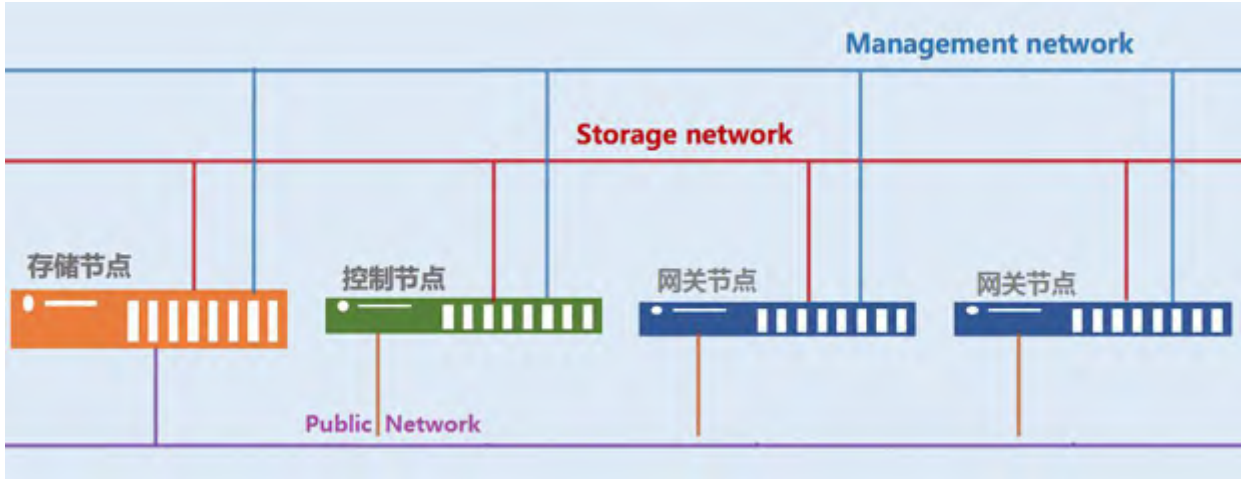
弹性扩展

- 分布式架构，支持水平扩展
- 多数据中心，统一命名空间
- 故障域隔离，存储池隔离
- 集群式管理，无单点瓶颈

自动化

- 自动化部署
- 可视化管理
- 实时监控告警
- 不停服升级和扩容





存储产品线 – 块存储

中国移动 China Mobile

- 基于CEPH深度定制的分布式块存储系统
- 支持KVM、VMWARE、Xen、Baremetal、K8S等常用场景
- 在私有云、研发云、公有云、视频存储等领域应用广泛，目前已经商用超过**35PB**

接口层：QEMU　ISCSI　NBD

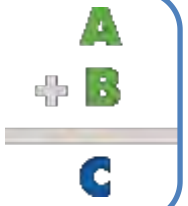
服务层：快照　克隆　增量备份　热迁移　VAAI　实时镜像　跨存储迁移　QoS

引擎层：分布式哈希　强一致副本　全冗余设计　并行重建　自动均衡　智能修复　瘦存储　线性扩展　硬件感知　故障域隔离　存储池隔离　热点缓存

硬件层：X86服务器　SAS/SATA/SSD　磁盘控制器管理　1GE/10GE　磁盘错误/慢盘检测　SSD寿命监控

存储管理：资源监控　性能监控　滚动升级　在线扩容　告警管理　日志管理　磁盘定位　数据盘漫游　部件更换　可视化

简单　初始投入减少 **85%**



高效　运营成本减少 **55%**

节省　总体拥有成本减少 **65%**

存储产品线 – 文件存储

中国移动 China Mobile

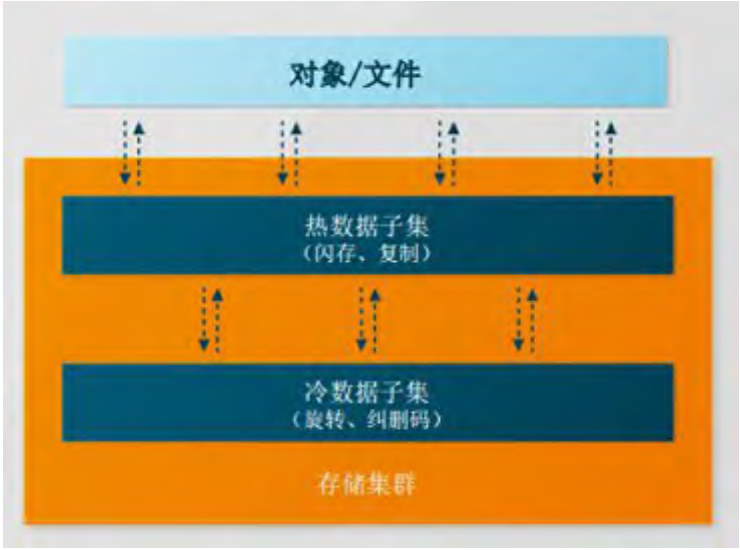
- 纯软件定义分布式文件存储，单机群最大1028节点/192PB，支持软硬一体机解决方案，商用部署超过5PB
- 支持容器Kubernetes平台(持久化存储)，大数据Hadoop平台（HDFS），Openstack Manila（弹性文件服务）
- 支持主流的CIFS/NFS/FTP文件访问协议，高性能IB RDMA网络
- 支持跨数据中心备份容灾，多数据中心

纠删码

- 更高的空间利用率
- 校验码计算硬件加速
- 更小的网络开销
- 摆脱硬件RAID

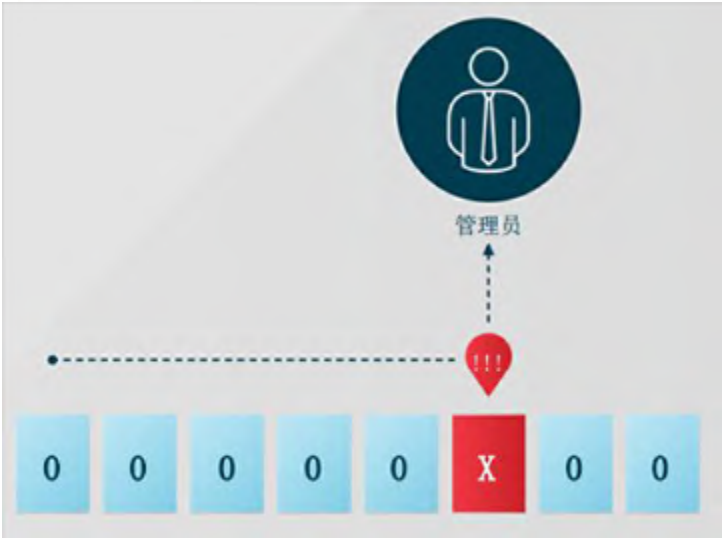
分层存储

- 冷热数据基于访问自动分层
- 高性价比混合存储加速方案
- 热数据：固态硬盘、复制
- 冷数据：旋转硬盘、纠删码



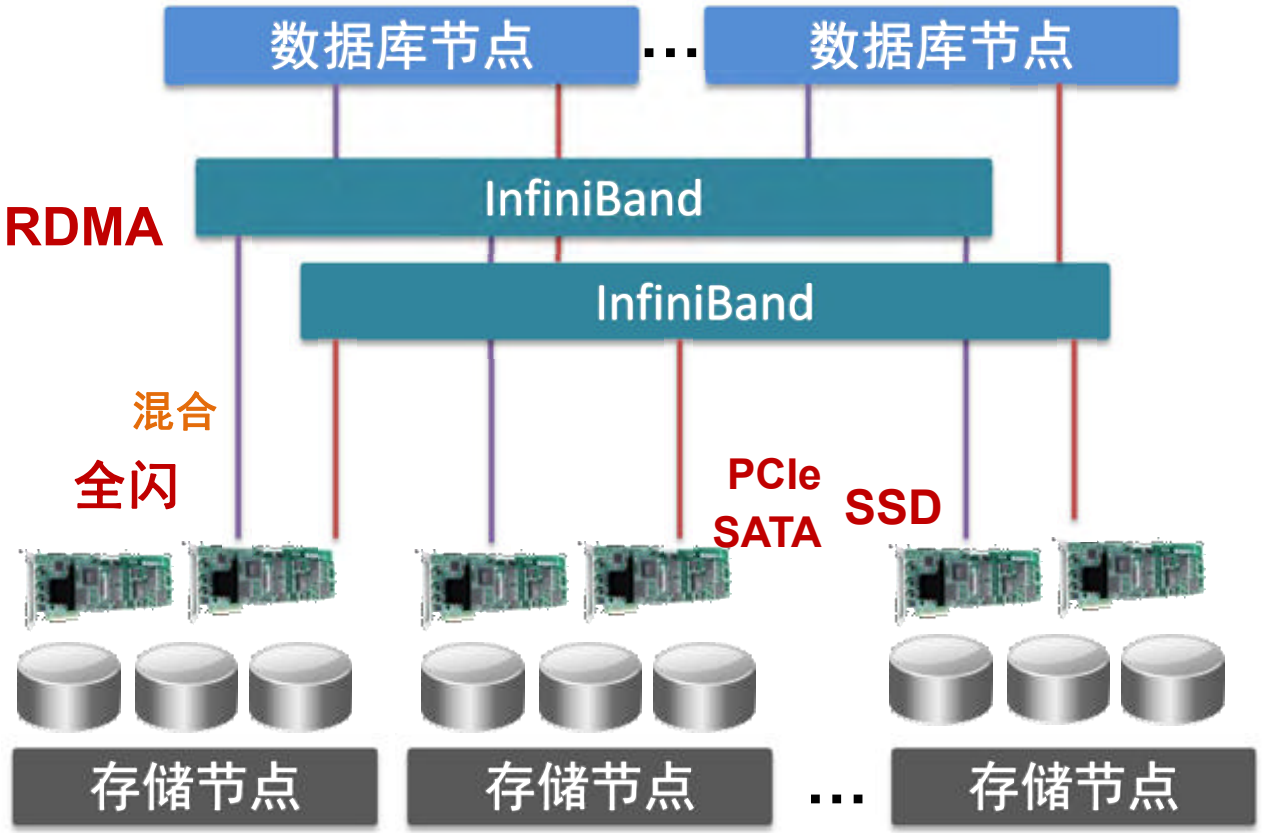
静默数据损坏检测

- 自动扫描检测
- SHA256算法硬件加速
- 自动恢复损坏数据到其他



存储产品线 – 存储一体机

- 基于X86服务器、PCIe/SATA SSD、IB构建高性能、低延迟的块存储一体机
- 满足Oracle、DB2、Mysql、Sybase等数据库及高性能块存储场景
- 在集团/省公司数据集市、IAP、无纸化等多类系统中广泛使用



优势

- 计算、存储、交换机全冗余，IB链路双活
- SSD、HDD热插拔
- SSD全生命周期管理
- 数据 多副本
- 一体化快速交付、统一运维
- 超融合、分离部署任选
- 标准2+3配置400W TPM+

专业

- 专门针对数据库场景深度优化
- OLTP、OLAP
- 支持共享式/非共享式集群

开放

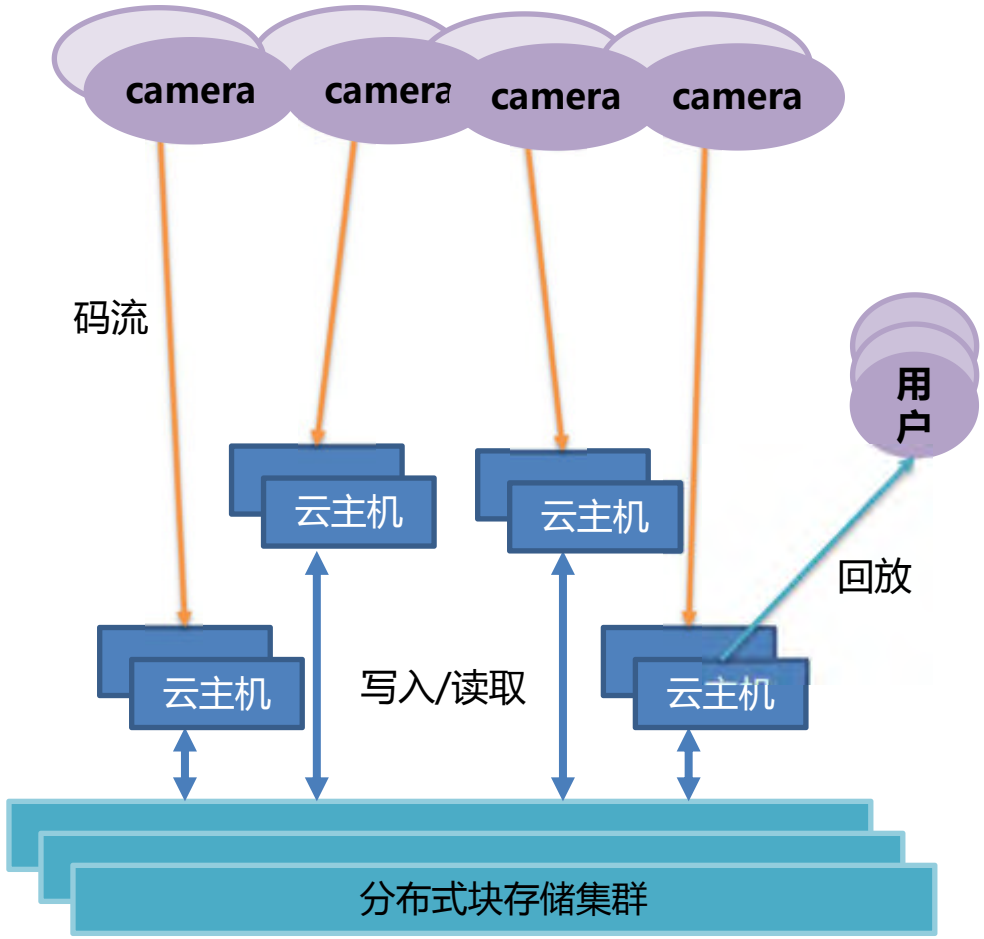
- 兼容国内外所有主流及开源数据库
- 采用x86架构通用硬件组件
- 无固件锁定，开放第三方备品备件

灵活

- 多系列产品从容应对不同场景
- 节点级在线扩展
- 优化策略在线调整适应场景变化

存储产品线 – 和目视频案例

- 某省公司和目视频实现使用分布式块存储集群，并完全云化
- 目前承载高清码流视频10000路+/10PB，预计近几年扩容至60000路。



云化 优势

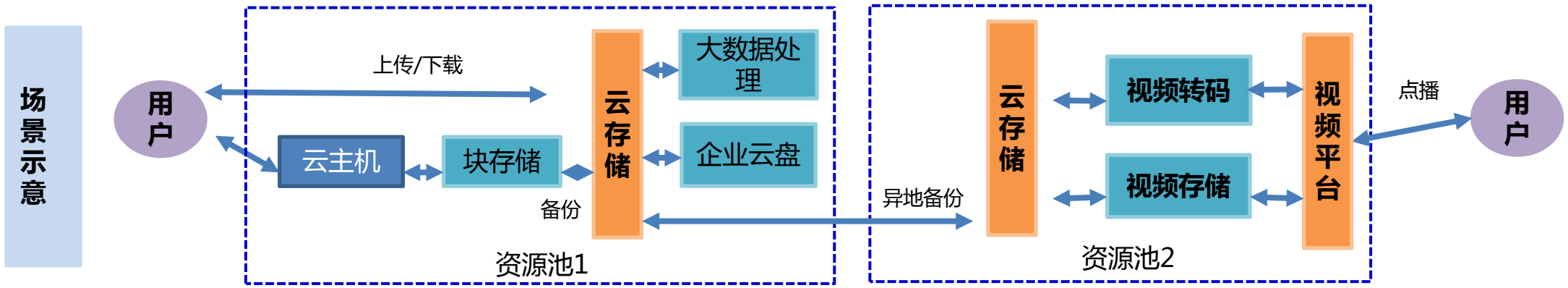
- 根据客户需求，快速开通业务
- 减少初始投资成本，快速扩容集群
- 自动化运维，极大减少成本
- 跨节点冗余，数据永不离线

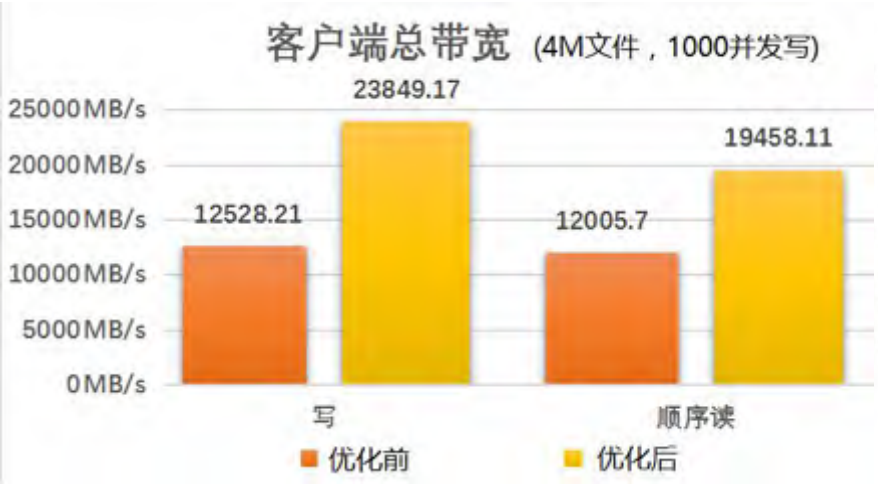
运行数据

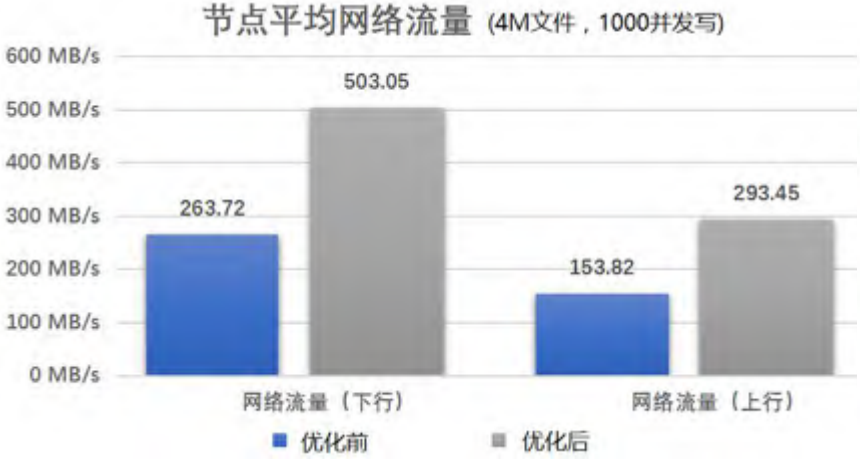
- 总写带宽10000MB/s，单集群写2000MB/s+
- 集群扩容至上线运行业务，以**周**为单位

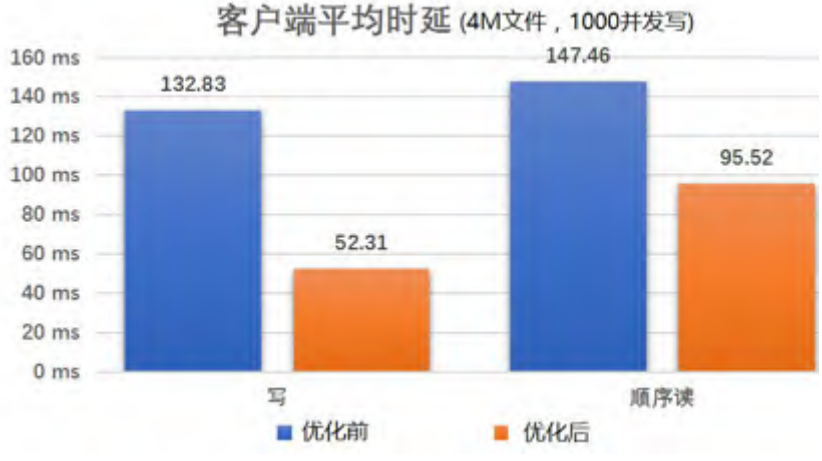
存储产品线 – 公有云案例

- 移动公有云自2014年正式运营，目前总体规模超过3000台物理机，提供弹性计算、云存储、数据库等多个产品
- 移动公有云对象存储分布在北京和广州两个数据中心，总规模达到30PB，今年预计继续扩容新数据中心

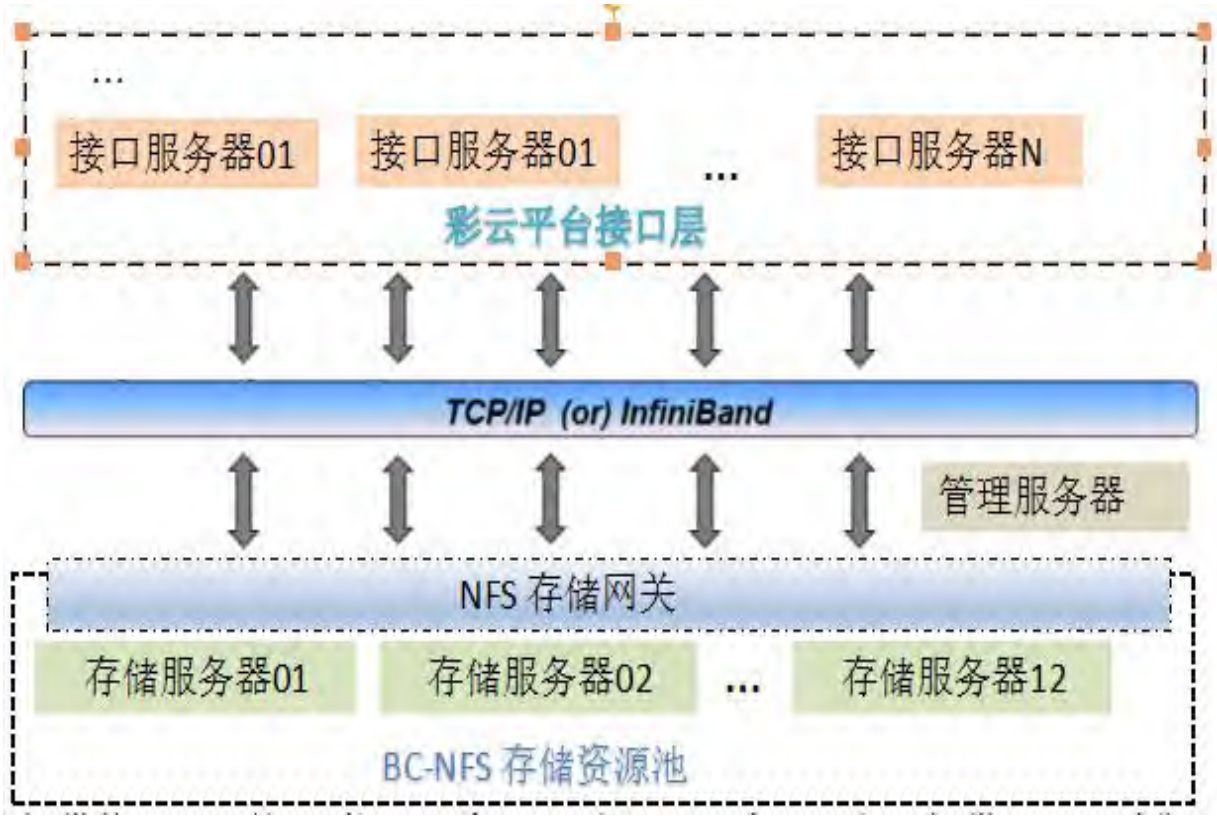




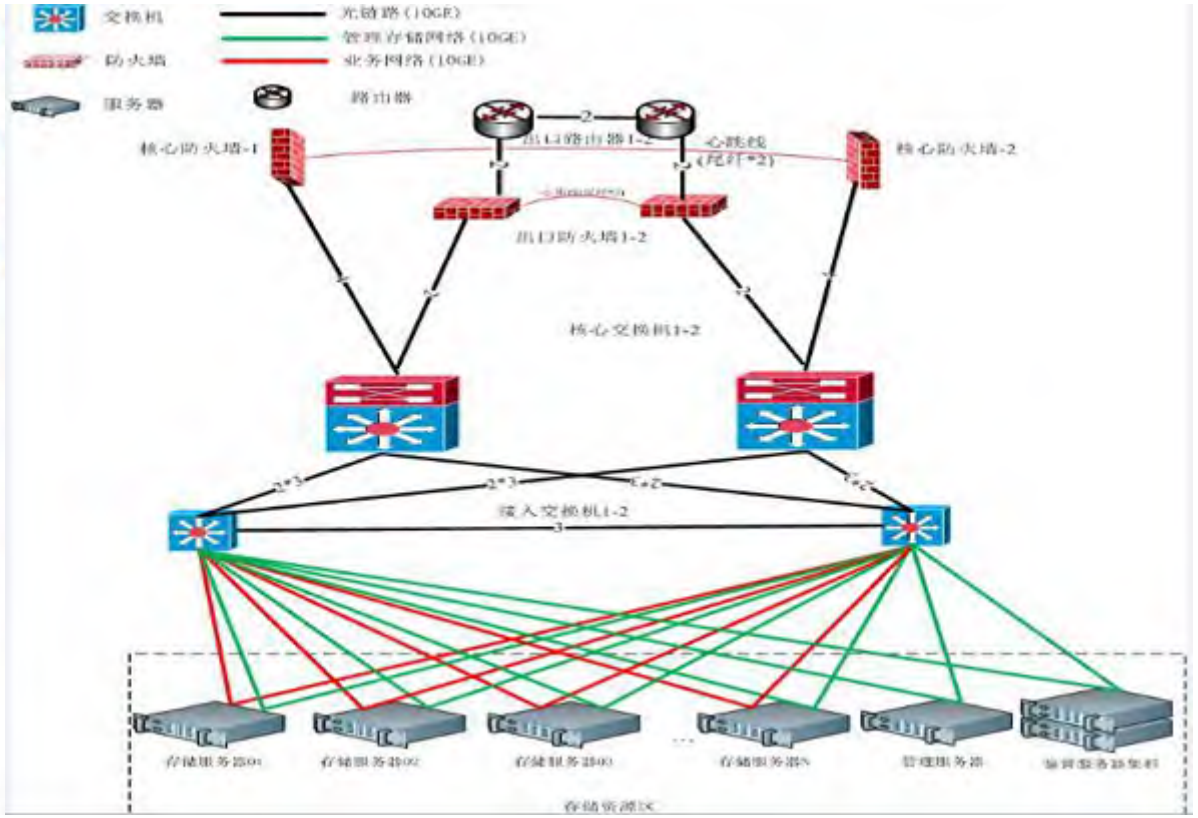




存储产品线 – 和彩云案例



业务逻辑架构



组网架构

2016年3月23日，苏研分布式文件系统正式割接上线，第一阶段部署规模为80台存储服务器集群，共计**2.9PB**裸容量，承载和彩云平台100%存储业务量。截止2017年5月15日，累计存储文件大小超过**1.1PB**、文件数量超过**10亿**，集群运行稳定，业务高峰期系统IOPS未出现饱和，CPU和内存平均利用率不超过15%。彻底解决了互联网公司和彩云业务底层商业存储故障频发的问题，同时帮助用户有效降低了存储设备建设成本。

中国移动 China Mobile

首页 存储池 块存储 对象存储 硬件管理 性能分析 告警管理 admin

10.9 TB
可用容量
已使用 784.0 MB
总容量 10.9 TB

11
存储池数量

0
块设备数量

3
节点数

固态硬盘 0个
机械硬盘 6个

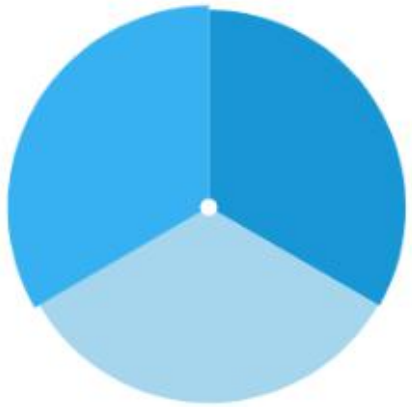
低
风险度

节点总数 3
正常节点数 3
故障节点数 0

系统IOPS
read write


系统CPU使用率
4%
of 154.2 GHz

系统内存使用率
1%
of 377.1 GB

节点容量使用占比


告警事件

系统MBPS(B/s)
read write


TOP3节点CPU使用率

10.128.2.76	5%
10.128.2.79	5%
10.128.2.78	4%

系统时延(ms)
read write

TOP3节点IOPS

10.128.2.76	12 (IO/s)

TOP3节点MBPS

10.128.2.76	57.5 (KB/s)

中国移动 China Mobile

首页 | 存储池 | 块存储 | 对象存储 | 硬件管理 | 性能分析 | 告警管理 | admin

10.9 TB
可用容量

已使用 784.0 MB
总容量 10.9 TB

0%

系统IOPS

系统MBPS(B/s)

系统时延(ms)

存储池

+创建存储池 1/1 | 1 跳转

	存储池名称	存储策略	IOPS	吞吐量	状态
◉	rbd	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	default.rgw.control	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	default.rgw.data.root	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	default.rgw.gc	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	default.rgw.log	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	qw	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	qq	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	122	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	default.rgw.users.uid	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	default.rgw.users.keys	3副本	0 (IO/s)	0 (Bytes/s)	正常
○	default.rgw.meta	3副本	0 (IO/s)	0 (Bytes/s)	正常

中国移动 China Mobile

首页 存储池 块存储 对象存储 硬件管理 性能分析 告警管理 admin

用户管理 桶管理 服务管理 负载均衡管理

1/1 | 1 跳转 + 创建对象用户 收集用户信息

名称	ID	邮箱	是否启用	最大对象个数	最大容量	操作
admin	21232f297a57a5a743894a0e4...		是	无限制	无限制	

子用户

1/0 | 1 跳转

子用户名称	权限	s3_ak	s3_sk	swift_sk	操作

中国移动 China Mobile

首页 存储池 块存储 对象存储 硬件管理 性能分析 告警管理

CPU总数： 6

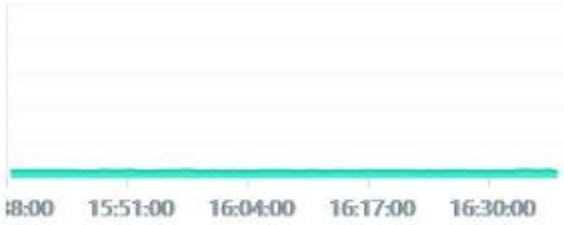
总内存： 377.1 GB

磁盘总数： 6个

数据盘总容量： 10.9 TB

缓存盘总容量： 8.2 TB

系统CPU利用率(百分比) cpu



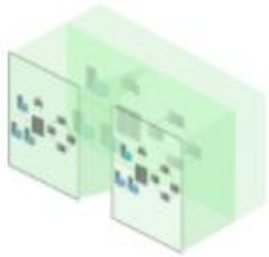
系统内存利用率(百分比) mem

系统磁盘使用率(百分比) disk

集群概况 • 节点信息

物理服务器集群

该视图可以查看物理机集群的整体概况



扩容服务器节点 修改管理网VIP地址

node0001 CPU 磁盘
10.128.2.76 内存 网卡
磁盘使用率：0%

node0002 CPU 磁盘
10.128.2.78 内存 网卡
磁盘使用率：0%

node0003 CPU 磁盘
10.128.2.79 内存 网卡
磁盘使用率：0%

性能项添加区

- 系统IOPS
- 系统MBPS(B/s)
- 系统时延(ms)

+ 性能观察项

性能分析区

近一小时 近一天 近一周 近一个月

告警 事件 日志收集

手动修复 1/3 1 跳转

	级别	状态	告警ID	告警名称	对象索引	产生时间	告警描述
	紧急	已修复	A1030	服务器无法连接	节点 10.128.2.76	2017-05-12 16:09:55	服务器无法连接
	紧急	已修复	A1060	存储服务异常	节点 10.128.2.76:mon.a	2017-05-11 17:57:46	集群仲裁服务异常
	紧急	已修复	A1060	存储服务异常	节点 10.128.2.76:osd.0	2017-05-11 17:57:46	硬盘数据管理服务异常
	重要	已修复	A1041	数据冗余度降级	存储池	2017-05-15 14:33:23	数据冗余度降级
	重要	已修复	A1021	网口故障	节点 10.128.2.76:bond4.1002:22	2017-05-12 16:54:28	网口链路断开
	重要	已修复	A1041	数据冗余度降级	存储池	2017-05-12 16:11:50	数据冗余度降级
	重要	已修复	A1021	网口故障	节点 10.128.2.76:em1	2017-05-12 15:07:23	网口链路断开
	重要	已修复	A1021	网口故障	节点 10.128.2.76:em2	2017-05-12 15:07:23	网口链路断开
	重要	已修复	A1021	网口故障	节点 10.128.2.76:bond4.1002:22	2017-05-11 17:58:06	网口链路断开
	重要	已修复	A1021	网口故障	节点 10.128.2.76:bond4.1002	2017-05-11 17:57:50	网口链路断开

谢谢！