ceph

Ceph中国社区 | IT大咖说 知识共享平台

CEPHALOCON APAC 2018

THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

CEPH在中国电信集约天翼高清项目中的应用与实践

中国电信云公司CDN运营中心

苏帅

2018/03/23

ceph

Ceph中国社区

IT大咖说 知识共享平台

目 录

ceph

Ceph中国社区

IT大咖说 知识共享平台

背 景

业务介绍：

集约天翼高清是一个OTT项目，主要包括内容中心和CDN两部分。

内容中心

中国电信CDN

ceph

业务介绍：

内容中心（北京，南京，广州）	CDN
直播点播内容引入	总节点数：287
直播点播内容处理	出口总带宽：19.9T
直播点播内容存储（点播内容需要用到分布式存储）	注册用户数：340万+

背　景

ceph

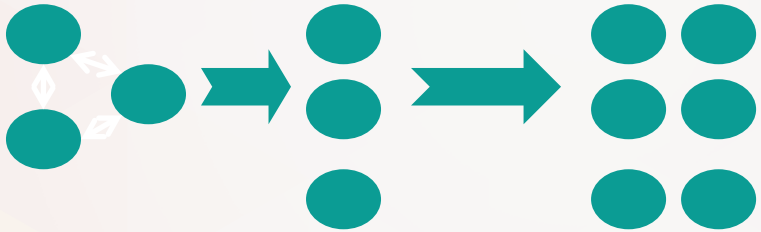
业务总体架构层次：

调度

DNS　HTTP 30x　IP地址库　自动化调度

内容分发

内容中心　父层　边缘



内容管理

内容引入　内容存储　预加载　刷新

运营管理

自动化运维　监控告警　日志分析　客户管理　计费

总体架构分为四部分：

- 内容管理
- 内容分发
- 调度
- 运营管理

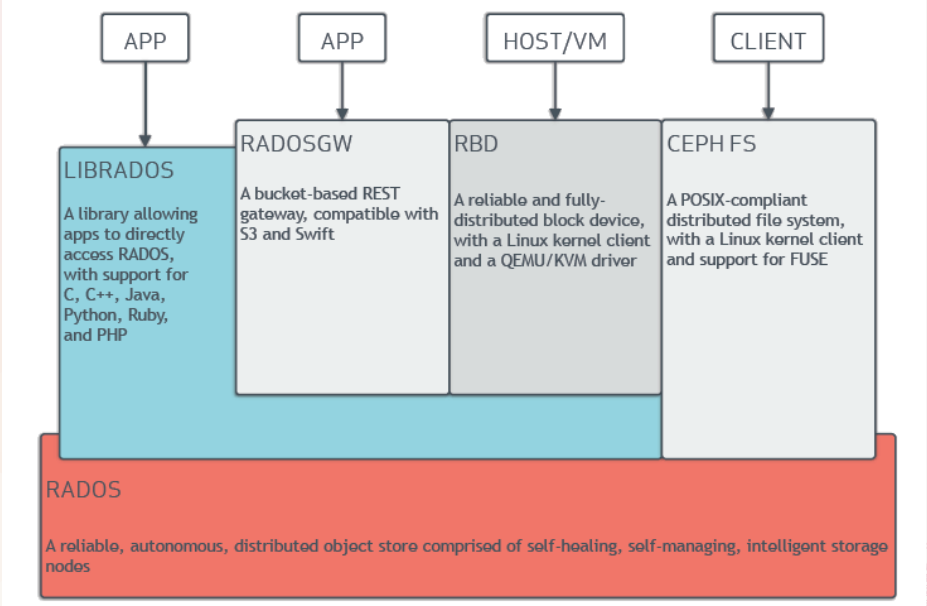
背 景

ceph

点播业务存储需求:

- 大容量，可弹性扩容或缩容
- 高性能，读写并发性能可线性扩展
- 高可靠性，保证数据完整性，安全性并可持续提供服务
- 易接入，提供完整的数据读写接入方案

CEPH体系架构



功能特性：

- 块存储
- 对象存储
- 文件存储

背 景

ceph

ceph分布式文件系统，具有以下优点：

高可靠性

- 没有单点故障
- 多副本
- 自动重新均衡
- 故障自动修复
- 数据安全性

高可扩展性

- 使用普通的x86服务器，支持TB到PB级的扩展

高性能

- 数据分布均衡，并行度高

接口统一

- 支持块存储，对象存储文件存储，基本是包含了市面上所有的流行存储类型

ceph

Ceph中国社区

IT大咖说 知识共享平台

应用与实践

ceph

规模：

北京	南京	广州
裸容量：1.5PB	裸容量：1.5PB	裸容量：1.8PB
实际使用量：753TB	实际使用量：753TB	实际使用量：753TB
OSD: 450	OSD:450	OSD:510
MON：5	MON:5	MON:5
副本数：2	副本数：2	副本数：2
版本号：10.2.9	版本号：10.2.9	版本号：10.2.9

ceph

服务器配置：

CPU:E5-2630v3　　CPU核数：32

内存：64G

SATA盘:40T （10块4T）　　SSD:960G (2块480G)

网卡：4000M　　20000M(千兆对外提供服务，万兆用于内部同步)

部署架构：

每台服务器：10个osd, 每块盘为一个OSD，两块ssd做日志盘,划分10个分区

ceph

演进路线：

2015

搭建北京ceph

容量：450T

2016

扩容北京ceph至880T

建设南京中心ceph,容量：1.5PB

2017

扩容北京ceph至1.5PB

建设广州中心ceph,容量：1.8PB

2018

按需求计划扩容北京，广州，南京中心ceph容量达3PB

ceph

存储

点播视频文件的存储，转码服务对视频转码完成后，通过librados写入CEPH

备份

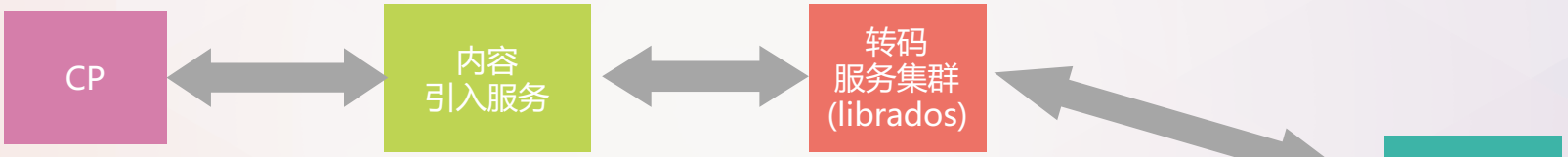
视频文件备份,多套CEPH互为备份，内容由单独的服务同步

源站

做为视频源站提供CDN回源服务，采用原生librados封装的NGINX模块

ceph

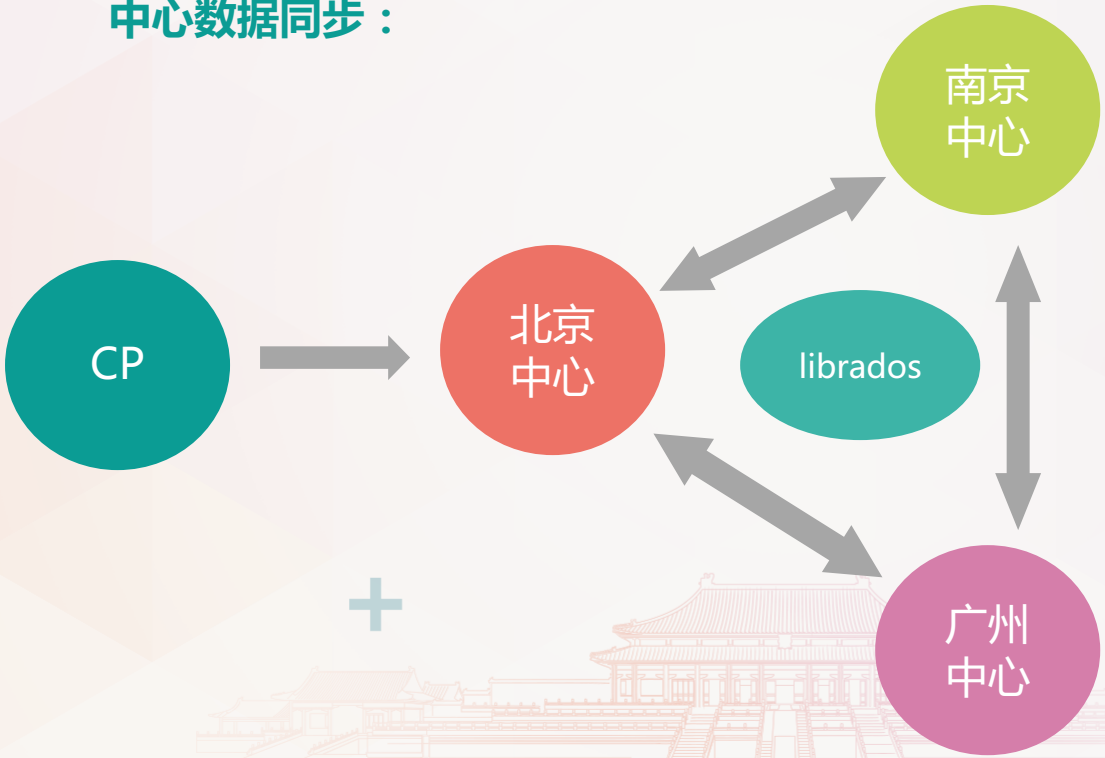
内容引入：



中心回源：

ceph

中心数据同步：

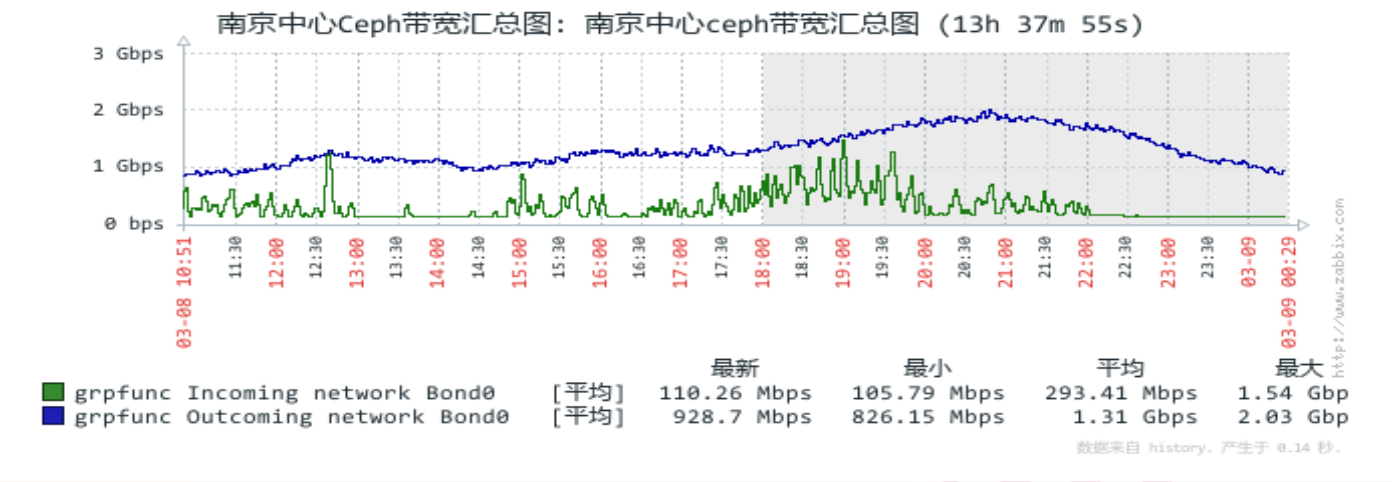


各个数据中心之间存储数据互为备份

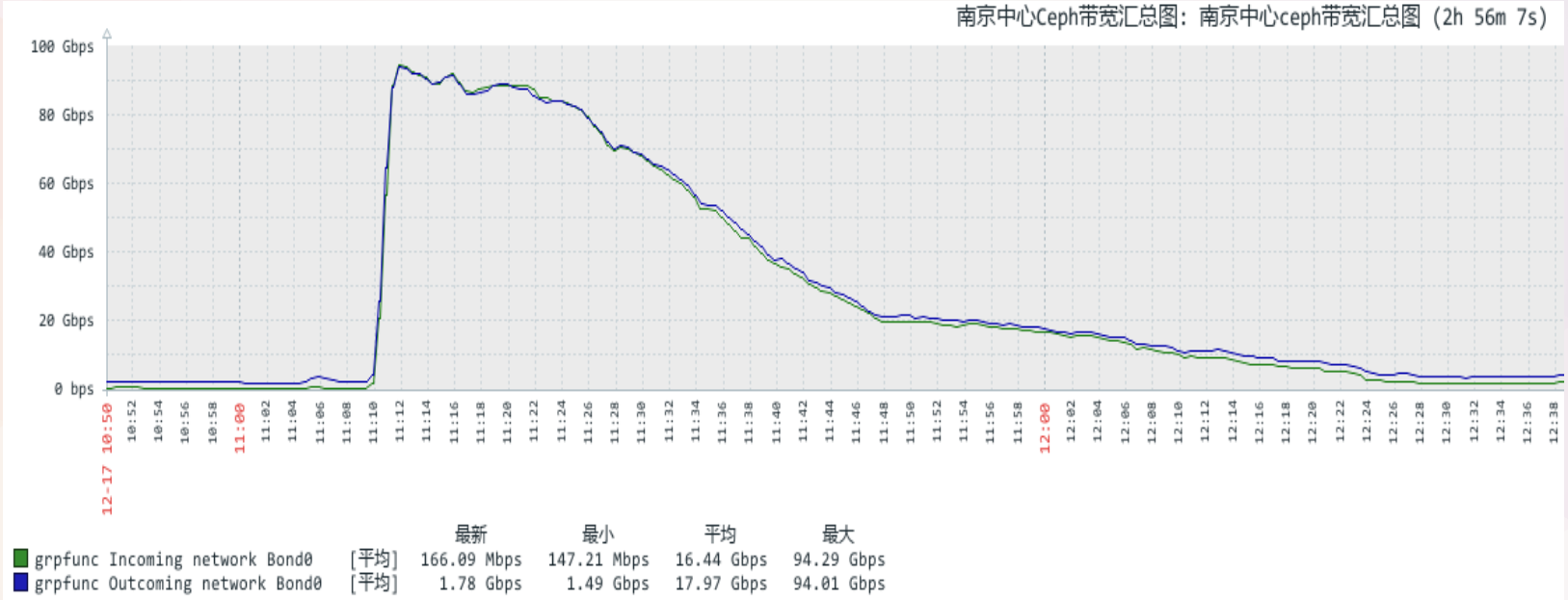
ceph

日常数据统计：

日常读	50-300 op/s
日常写	0-200 op/s



日常数据统计：


南京中心Ceph带宽汇总图：南京中心ceph带宽汇总图 (2h 56m 7s)
100 Gbps
80 Gbps
60 Gbps
40 Gbps
20 Gbps
0 bps
12-17 10:50
11:00
12:00
最新 最小 平均 最大
grpfunc Incoming network Bond0 [平均] 166.09 Mbps 147.21 Mbps 16.44 Gbps 94.29 Gbps
grpfunc Outcoming network Bond0 [平均] 1.78 Gbps 1.49 Gbps 17.97 Gbps 94.01 Gbps

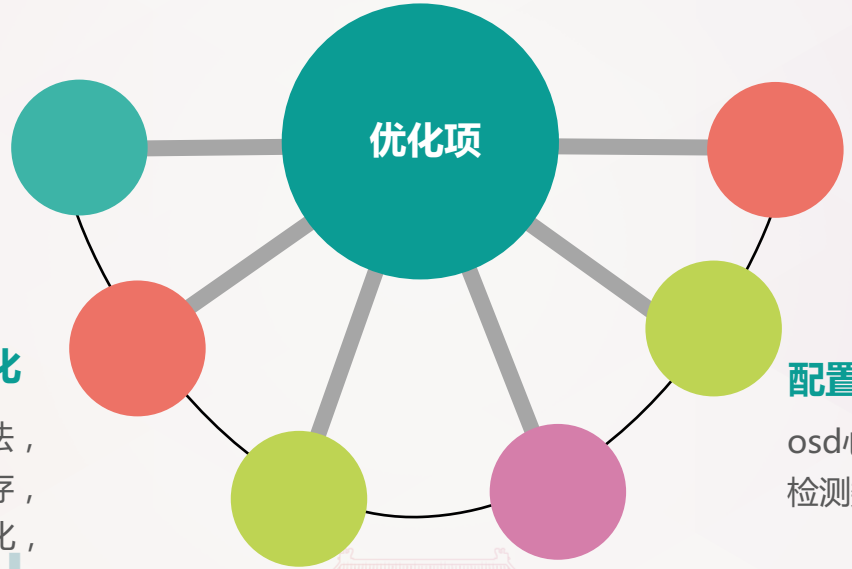
ceph

Ceph中国社区

IT大咖说 知识共享平台

问题及优化

问题及优化



资源池规划

每个资源池固定大小，达到一定量新建池

故障域的划分

减少故障影响范围

系统优化

设置sata和ssd盘调度算法，设置sata和ssd盘的raid缓存，万兆网卡MTU优化,,操作系统pid优化，CPU主频优化

配置文件参数

osd心跳时间，osd线程数，scrub检测数据一致性的时间优化

网络分离

集群网络与客户端网络分离

文件系统

ceph osd的xfs 文件系统参数优化

问题及优化

pg状态（down+peering处理）	查看该pg状态详情，找出具体哪个osd的问题导致修复不能恢复。标记改osd为lost（有可能会丢数据）并重启osd状态恢复
pg状态（incomplete处理）	先使用ceph提供的工具备份好pg，然后把改pg标记为complete状态即可恢复
pg状态（inconsistent处理）	磁盘出现静默错误就会出现，如果主副本数据完成直接可以使用ceph repair命令进行修复，如果是主副本有问题，应把副本拥有的pg拷贝到主副本。

OSD所在磁盘文件系统损坏	导致OSD启动失败，使用**xfs**文件系统提供的**repair**命令进行修复，修复完成后启动对应**osd**。
Journal损坏	更换journal 分区即可。
Mon节点问题	删除此mon节点，待修复后再添加进去。mon节点出问题后，会导致所有连接此mon的服务异常，应用服务层也应删除此节点配置。

后续计划

后续计划

•建设融合CDN,一张网络，一个平台，统一规划，集约运营，对内支撑集约天翼高清，对外支撑商业用户

•根据业务需求，北京，南京，广州中心单集群扩容至3PB

• rgw元数据和真实数据分离，元数据使用ssd磁盘，真实数据使用EC方式存到sata盘

• 多数据中心融灾方案使用multisite，rgw之间的数据同步网络和业务网络分离

• 做为边缘集群的内容存储

ceph

Ceph中国社区

IT大咖说 知识共享平台

THANK YOU 谢谢！