

第4章 分布式存储Ceph

云计算导论和应用实践

存储领域的"Linux" - 统一分布式存储解决方案

本章目录

- 4.1 Ceph 简介
- 4.2 Ceph 的特点
- 4.3 Ceph 的架构
- 4.4 Ceph I/O 算法流程
- 4.5 Ceph 读写数据流程
- 4.6 Ceph 操作实践

4.1 Ceph 简介

Ceph 可以称之为存储领域的"Linux"

- **开源的软件定义存储系统**: 提供统一的分布式存储解决方案
- **三种存储类型**: 同时提供块存储、文件存储和对象存储
- **核心特性**: 高性能、高可靠性、高扩展性
- **云计算集成**: 与OpenStack云计算平台紧密结合
- **应用领域**: 在云计算和大数据领域占据领导地位

发展历程

- **2003年**: 由加州大学圣克鲁兹分校Sage Weil开发
- **2006年**: 遵循LGPL协议开源
- **2012年**: 成立Inktank公司提供专业服务
- **2014年**: 被RedHat收购

4.2 Ceph 的特点(1)

- **高性能**
 - 摒弃传统集中式存储元数据寻址方案
 - 采用**CRUSH算法**，数据分布均衡，并行度高
 - 支持上千个存储节点，PB到ZB数据规模
- **高可靠性**
 - 通过副本保证数据可靠性
 - 支持**纠删码**节省物理存储空间
- **高度自动化**
 - 数据自动复制、自动均衡
 - 自动故障检测和自动故障恢复

4.2 Ceph 的特点(2)

- **高可扩展性**

- 理论上存储节点可无限扩展
- 性能随存储节点数线性增长

- **特性丰富**

- 统一平台同时支持三种存储类型
- 支持多种语言驱动

核心优势：分布式架构带来的高性能、高可靠性和高扩展性，使Ceph成为理想的企业级存储解决方案

4.3 Ceph 的架构

Ceph在统一系统中提供对象、块和文件存储，核心是**RADOS**（可靠自修复分布式对象存储）

核心组件

- **RADOS**: 负责存储对象，确保数据一致性和可靠性
- **LIBRADOS**: 简化访问RADOS的库，支持C、C++、Java、Python等
- **RADOSGW**: Ceph对象网关，兼容Amazon S3和OpenStack Swift API
- **Ceph RBD**: Ceph块设备，提供块存储服务
- **CephFS**: Ceph文件系统，提供POSIX兼容的分布式文件系统
- **OSD**: 对象存储设备，存储实际用户数据
- **Monitor**: 监视器，负责系统状态检测和维护

核心机制: **"无须查表，算算就好"**

4.4 Ceph I/O 算法流程

对象存储基础

- 每个对象作为一个文件存储（无分层无目录）
- 每个对象包含：**ID**、**Binary Data**、**Metadata**

三次映射过程

- **第一次映射：file → object**
 - $\text{ino (文件唯一ID)} + \text{ono (对象序号)} = \text{oid (对象ID)}$
 - 默认以4MB切分文件块
- **第二次映射：object → PG**
 - $\text{hash(oid)} \& \text{mask} \rightarrow \text{pgid}$
 - $\text{mask} = \text{PG总数}m-1$, m 为2的整数幂
- **第三次映射：PG → OSD**
 - 采用CRUSH算法： $\text{CRUSH(pgid)} \rightarrow (\text{osd1}, \text{osd2}, \text{osd3})$
- 6 • 生产环境至少3副本存储

4.5 Ceph 读写数据流程

数据存储流程

- **步骤1**: 客户端访问Ceph monitor获取cluster map副本
- **步骤2**: 数据转化为一个或多个对象
- **步骤3**: 对象以PG数为基数做散列运算
- **步骤4**: 通过CRUSH查询确定主→次→再次OSD位置
- **步骤5**: 客户端直接与OSD通信存储数据

读写性能特点

读操作

- 只需两步完成
- 性能较优
- 直接从主OSD读取

写操作

- 需要主OSD完成写入
- 次OSD和再次OSD完成副本拷贝
- 所有副本写入完成才算成功

核心优势: 所有操作在客户端完成, 实现"无须查表, 算算就好", 不影响集群服务器端性能

4.6 Ceph 操作实践 - 硬件要求(1)

硬件基础

- **服务器**：行业标准服务器，运行标准Linux发行版
- **磁盘配置**：
 - 生产环境主要使用SATA磁盘
 - SSD用于加速（存储元数据）
 - 支持分层存储：SSD + SAS + SATA

网络要求

- **公共网络**：前端对外提供存储服务
- **集群内部网络**：处理OSD心跳、对象复制和恢复流量
- **带宽计算**：10个OSD × 250Mbit/s × 8 = 20Gbit/s
- **推荐配置**：每节点至少2个10Gbit/s端口

4.6 Ceph 操作实践 - 硬件要求(2)

CPU和内存

- **CPU**: 可用核数 > OSD数, 单个OSD绑定一个CPU核
- **内存**: 每个OSD预留4-8GB作为缓存

硬件配置建议

- 生产环境建议使用企业级硬件
- 根据业务需求合理规划资源
- 预留足够的扩展空间

其他注意事项

- **电源**: 建议配置冗余电源
- **散热**: 确保良好的散热条件
- **机柜空间**: 预留足够的机柜空间用于扩展

实验环境配置(1)

虚拟机配置

主机名	公共网络	集群网络
node01	192.168.1.101	192.168.2.101
node02	192.168.1.102	192.168.2.102
node03	192.168.1.103	192.168.2.103

准备工作

- 各节点网卡IP地址配置
- 主机名命名和解析
- 时区设置和时钟同步
- 各节点间SSH免密认证
- 系统升级和优化（关闭防火墙、SELinux等）
- 添加软件仓库（可选）

实验环境配置(2)

基本术语

- **fsid**: 集群唯一标识符
- **cluster name**: 集群名称 (默认ceph)
- **monitor name**: 监视器名称
- **keyring**: 密钥环, 用于认证和授权

注意事项

- 确保所有节点时间同步
- 网络配置需要严格按照规划执行
- 建议使用独立的集群网络
- 妥善保管密钥信息

Monitor 节点部署(1)

安装Ceph软件

```
# yum update -y  
# yum install ceph -y  
# ceph -v
```

生成配置文件

```
# uuidgen  
# vi /etc/ceph/ceph.conf  
[global]  
fsid=79e959e7-a534-46ef-94b7-eaee96e4c4ee  
mon initial members=node01  
mon host=192.168.1.101  
public network=192.168.1.0/24  
cluster network=192.168.2.0/24  
auth cluster required=cephx
```

注意事项

10 确保配置文件权限正确

Monitor 节点部署(2)

创建密钥环

```
# ceph-authtool --create-keyring /tmp/ceph.mon.keyring --gen-key -n mon. --cap mon 'allow *'
```

```
# ceph-authtool --create-keyring /etc/ceph/ceph.client.admin.keyring --gen-key -n client.admin --cap mon 'allow *' --cap osd 'allow *' --cap mds 'allow *'
```

密钥环说明

- mon.keyring: Monitor服务密钥
- client.admin.keyring: 管理员密钥

Monitor 节点部署(3)

创建引导密钥

```
# ceph-authtool --create-keyring /var/lib/ceph/bootstrap-osd/ceph.keyring --gen-key -n client.bootstrap-osd --cap mon 'profile bootstrap-osd'
```

合并密钥环

```
# ceph-authtool /tmp/ceph.mon.keyring --import-keyring /etc/ceph/ceph.client.admin.keyring  
# ceph-authtool /tmp/ceph.mon.keyring --import-keyring /var/lib/ceph/bootstrap-osd/ceph.keyring
```

密钥环说明

- bootstrap-osd.keyring: OSD引导密钥,用于自动部署OSD
- 合并后的mon.keyring包含所有必要的认证信息

OSD 节点部署(1)

OSD部署步骤

- **准备磁盘**: 为每个OSD准备独立磁盘
- **创建OSD**: 使用ceph-volume工具创建OSD
- **激活OSD**: 激活并启动OSD服务
- **验证状态**: 检查OSD状态和集群健康

常用命令

```
# ceph-volume lvm create --data /dev/sdb
# systemctl start ceph-osd@0
# systemctl enable ceph-osd@0
# ceph osd tree
# ceph -s
```


OSD 节点部署(2)

存储池管理

- 创建存储池: `ceph osd pool create`
- 设置副本数: `ceph osd pool set`
- 查看存储池: `ceph osd lspools`
- 删除存储池: `ceph osd pool delete`

注意事项

- 创建存储池时需要考虑PG数量
- 删除存储池前确保数据已备份
- 及时监控存储池使用情况

对象存储网关 (RGW) - 基本特性

RGW特性

- **API兼容**: 兼容Amazon S3和OpenStack Swift API
- **多租户支持**: 支持多租户和Keystone身份验证
- **RESTful接口**: 提供标准的RESTful API

核心优势

- 标准API支持, 便于集成
- 企业级多租户管理
- 灵活的访问控制

应用场景

- 云存储服务
- 数据备份归档
- 媒体资源存储
- 大数据存储

对象存储网关 (RGW) - 部署配置(1)

部署配置

```
# 创建RGW存储池
# ceph osd pool create .rgw.root 16 16
# ceph osd pool create default.rgw.control 16 16

# 启动RGW服务
# systemctl start ceph-radosgw@rgw.node01
# systemctl enable ceph-radosgw@rgw.node01
```

部署步骤说明

- 首先创建必要的存储池
- 启动并设置RGW服务自启动
- 确认服务状态正常运行

对象存储网关 (RGW) - 部署配置(2)

用户管理

```
# 创建管理员用户
# radosgw-admin user create --uid="admin" \
  --display-name="admin user" --system

# 查看用户信息
# radosgw-admin user list
# radosgw-admin user info --uid=admin
```

配置注意事项

- 确保存储池副本数配置合理
- 妥善保管用户认证信息
- 定期备份用户数据

S3 客户端操作

s3cmd工具

```
# 安装s3cmd
# pip install s3cmd
# 配置文件 ~/.s3cfg
[default]
access_key = *****
secret_key = *****
host_base = 192.168.1.101:7480
use_https = False
```

常用操作命令

- s3cmd ls - 列举buckets
- s3cmd mb s3://bucket - 创建bucket
- s3cmd rb s3://bucket - 删除bucket
- s3cmd put file s3://bucket - 上传文件
- s3cmd get s3://bucket/file - 下载文件
- s3cmd rm s3://bucket/file - 删除文件
- s3cmd sync /path/ s3://bucket - 同步目录
- s3cmd info s3://bucket - 查看bucket信息

本章小结

主要内容回顾

- **Ceph简介**: 存储领域的"Linux", 统一分布式存储解决方案
- **核心特点**: 高性能、高可靠性、高扩展性、高度自动化
- **系统架构**: RADOS核心, 支持块、文件、对象三种存储
- **算法机制**: 三次映射, "无须查表, 算算就好"
- **实践操作**: Monitor、OSD、RGW部署和管理

技术要点

- CRUSH算法
- PG机制
- 副本和纠删码
- 自动故障恢复
- S3 API兼容
- 多租户支持
- 分层存储
- 性能优化

学习建议: 继续深入学习官方文档, 实践更多高级特性如缓存分层存储、性能调优等