Ceph基础知识和基础架构认识

1 Ceph基础介绍

Ceph是一个可靠地、自动重均衡、自动恢复的分布式存储系统，根据场景划分可以将C
文件系统服务。在虚拟化领域里，比较常用到的是Ceph的块设备存储，比如在OpenSt
OpenStack的cinder后端存储、Glance的镜像存储和虚拟机的数据存储，比较直观的是
虚拟机实例的硬盘。

Ceph相比其它存储的优势点在于它不单单是存储，同时还充分利用了存储节点上的计算
数据存储的位置，尽量将数据分布均衡，同时由于Ceph的良好设计，采用了CRUSH算法
障的问题，且随着规模的扩大性能并不会受到影响。

2 Ceph的核心组件

Ceph的核心组件包括Ceph OSD、Ceph Monitor和Ceph MDS。

Ceph OSD：OSD的英文全称是Object Storage Device，它的主要功能是存储数据、
间进行心跳检查等，并将一些变化情况上报给Ceph Monitor。一般情况下一块硬盘对应
一个分区也可以成为一个OSD。

Ceph OSD的架构实现由物理磁盘驱动器、Linux文件系统和Ceph OSD服务组成，对于
的支持了其拓展性，一般Linux文件系统有好几种，比如有BTRFS、XFS、Ext4等，BT
境所需的稳定性，一般比较推荐使用XFS。

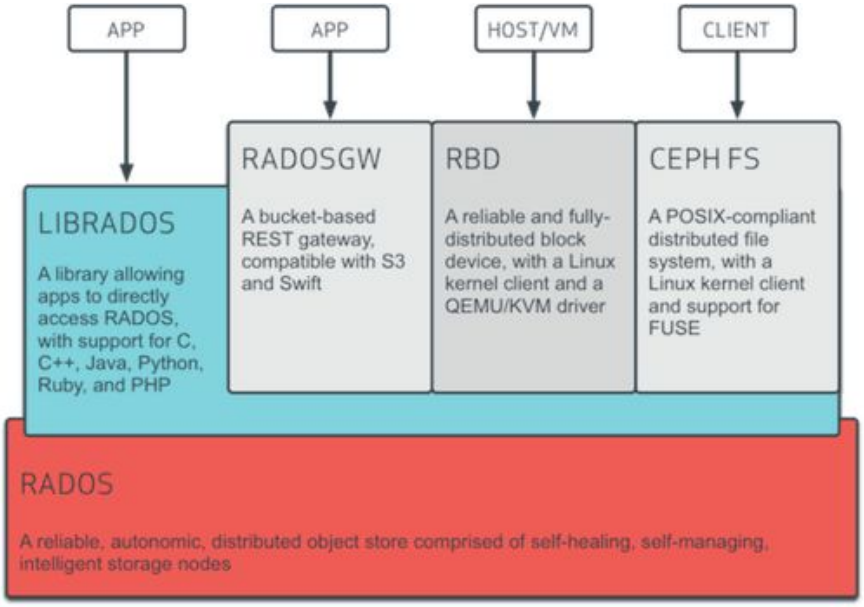
伴随OSD的还有一个概念叫做Journal盘，一般写数据到Ceph集群时，都是先将数据写
将Journal盘中的数据刷新到文件系统中。一般为了使读写时延更小，Journal盘都是采
好，Ceph中引入Journal盘的概念是因为Journal允许Ceph OSD功能很快做小的写操作
journal，然后刷新到文件系统，这给了文件系统足够的时间来合并写入磁盘，一般情况
负载。

Ceph Monitor：由该英文名字我们可以知道它是一个监视器，负责监视Ceph集群，维
的各种Map图，比如OSD Map、Monitor Map、PG Map和CRUSH Map，这些Map统
键数据结构，管理集群中的所有成员、关系、属性等信息以及数据的分发，比如当用户需
Monitor获取最新的Map图，然后根据Map图和object id等计算出数据最终存储的位置

Ceph MDS：全称是Ceph MetaData Server，主要保存的文件系统服务的元数据，但

查看各种Map的信息可以通过如下命令：ceph osd(mon、pg) dump

3 Ceph基础架构组件



从架构图中可以看到最底层的是RADOS，RADOS自身是一个完整的分布式对象存储系
高可靠、高可拓展、高性能、高自动化都是由这一层来提供的，用户数据的存储最终也
是Ceph的核心。

RADOS系统主要由两部分组成，分别是OSD和Monitor。

基于RADOS层的上一层是LIBRADOS，LIBRADOS是一个库，它允许应用程序通过访
语言，比如C、C++、Python等。

基于LIBRADOS层开发的又可以看到有三层，分别是RADOSGW、RBD和CEPH FS。

RADOSGW：RADOSGW是一套基于当前流行的RESTFUL协议的网关，并且兼容S3和

RBD：RBD通过Linux内核客户端和QEMU/KVM驱动来提供一个分布式的块设备。

CEPH FS：CEPH FS通过Linux内核客户端和FUSE来提供一个兼容POSIX的文件系统。

4 Ceph数据分布算法

在分布式存储系统中比较关注的一点是如何使得数据能够分布得更加均衡，常见的数据
Crush是一种伪随机的控制数据分布、复制的算法，Ceph是为大规模分布式存储而设计

数据依然能够快速的准确的计算存放位置，同时能够在硬件故障或扩展硬件设备时做到尽
心为这些特性设计的，可以说CRUSH算法也是Ceph的核心之一。

在说明CRUSH算法的基本原理之前，先介绍几个概念和它们之间的关系。

存储数据与object的关系：当用户要将数据存储到Ceph集群时，存储数据都会被分割成
个object的大小是可以设置的，默认是4MB，object可以看成是Ceph存储的最小存储单

object与pg的关系：由于object的数量很多，所以Ceph引入了pg的概念用于管理obje
某个pg中，一个pg可以包含多个object。

pg与osd的关系：pg也需要通过CRUSH计算映射到osd中去存储，如果是二副本的，则
[osd.1,osd.2]，那么osd.1是存放该pg的主副本，osd.2是存放该pg的从副本，保证了

pg和pgp的关系：pg是用来存放object的，pgp相当于是pg存放osd的一种排列组合，
osd.3，副本数是2，如果pgp的数目为1，那么pg存放的osd组合就只有一种，可能是[
到osd.1和osd.2，如果pgp设为2，那么其osd组合可以两种，可能是[osd.1,osd.2]和
排列组合，pgp就是代表这个意思。一般来说应该将pg和pgp的数量设置为相等。这样设

先创建一个名为testpool包含6个PG和6个PGP的存储池
ceph osd pool create testpool 6 6
通过写数据后我们查看下pg的分布情况，使用以下命令：

```
ceph pg dump pgs | grep ^1 | awk '{print $1,$2,$15}'
dumped pgs in format plain
1.1 75 [3,6,0]
1.0 83 [7,0,6]
1.3 144 [4,1,2]
1.2 146 [7,4,1]
1.5 86 [4,6,3]
1.4 80 [3,0,4]
```

第1列为pg的id，第2列为该pg所存储的对象数目，第3列为该pg所在的osd

我们扩大PG再看看
ceph osd pool set testpool pg_num 12
再次用上面的命令查询分布情况：
1.1 37 [3,6,0]

1.9 38 [3,6,0]
1.0 41 [7,0,6]
1.8 42 [7,0,6]
1.3 48 [4,1,2]
1.b 48 [4,1,2]
1.7 48 [4,1,2]
1.2 48 [7,4,1]
1.6 49 [7,4,1]
1.a 49 [7,4,1]
1.5 86 [4,6,3]
1.4 80 [3,0,4]
我们可以看到pg的数量增加到12个了，pg1.1的对象数量本来是75的，现在是37个，而
38，加起来是75，而且可以看到pg1.1和pg1.9的osd盘是一样的。
而且可以看到osd盘的组合还是那6种。

我们增加pgp的数量来看下，使用命令：
ceph osd pool set testpool pgp_num 12
再看下
1.a 49 [1,2,6]
1.b 48 [1,6,2]
1.1 37 [3,6,0]
1.0 41 [7,0,6]
1.3 48 [4,1,2]
1.2 48 [7,4,1]
1.5 86 [4,6,3]
1.4 80 [3,0,4]
1.7 48 [1,6,0]
1.6 49 [3,6,7]
1.9 38 [1,4,2]
1.8 42 [1,2,3]
再看pg1.1和pg1.9，可以看到pg1.9不在[3,6,0]上，而在[1,4,2]上了，该组合是新加
的组合。

通过实验总结：

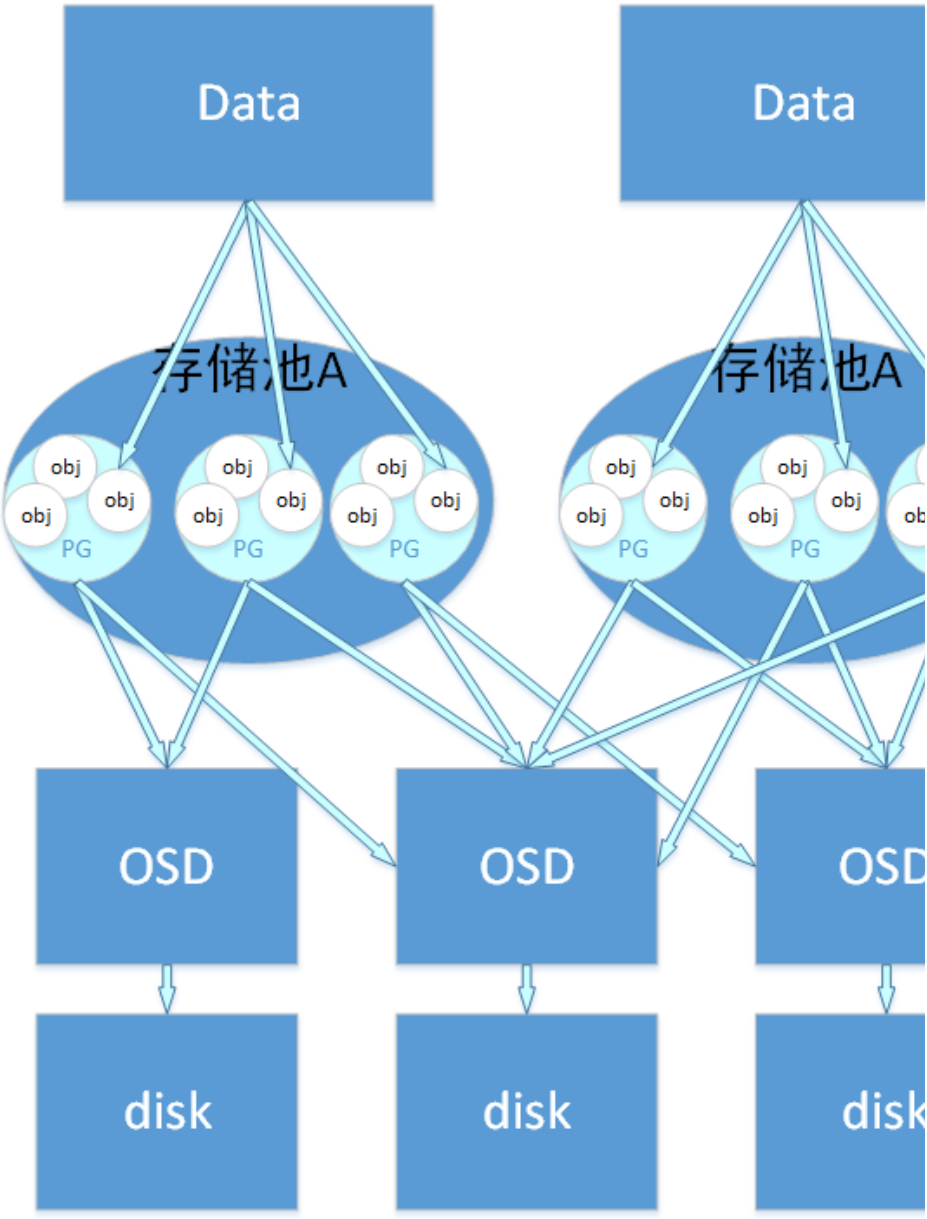
（1）PG是指定存储池存储对象的目录有多少个，PGP是存储池PG的OSD分布组合个数

（2）PG的增加会引起PG内的数据进行分裂，分裂相同的OSD上新生成的PG当中

（3）PGP的增加会引起部分PG的分布进行变化，但是不会引起PG内对象的变动

pg和pool的关系：pool也是一个逻辑存储概念，我们创建存储池pool的时候，都需要指储池的，就有点像object是属于某个pg的。

以下这个图表明了存储数据，object、pg、pool、osd、存储磁盘的关系



本质上CRUSH算法是根据存储设备的权重来计算数据对象的分布的，权重的设计可以根
量大小可以将1T的硬盘设备权重设为1，2T的就设为2，在计算过程中，CRUSH是根据
定数组最终的存储位置的。

Cluster Map里的内容信息包括存储集群中可用的存储资源及其相互之间的空间层次关系个服务器，每个服务器有多少块磁盘用以OSD等。

数据分布策略是指可以通过Ceph管理者通过配置信息指定数据分布的一些特点，比如管Host起不来时，数据能够不丢失，CRUSH可以通过将每个pg的主从副本分别存放在不同Host，还可以指定机架等故障域，除了故障域，还有选择数据冗余的方式，比如副本数

下面这个式子简单的表明CRUSH的计算表达式：

CRUSH(X) -> (osd.1,osd.2.....osd.n)

式子中的X就是一个随机数。

下面通过一个计算PG ID的示例来看CRUSH的一个计算过程：

（1）Client输入Pool ID和对象ID；

（2）CRUSH获得对象ID并对其进行Hash运算；

（3）CRUSH计算OSD的个数，Hash取模获得PG的ID，比如0x48；

（4）CRUSH取得该Pool的ID，比如是1；

（5）CRUSH预先考虑到Pool ID相同的PG ID，比如1.48。

分类： Ceph

标签： Ceph

« 上一篇：Glance组件解析

» 下一篇：Cinder组件解析

posted @ 2017-12-24 01:19 luohaixian 阅读(29676) 评论(2) 编辑 收藏

评论列表

#1楼 2018-08-24 17:22 March On

赞一个。
感觉LZ非常适合写论文搞研究。弱问下在哪个学校？读研呢？

#2楼[楼主] 2018-09-22 09:42 luohaixian

@ March On
过奖了，我也只是初学者，本人本科毕业，已经出来工作了，互相学习。

相关博文：

- OpenStack H版与 Ceph 整合的现状
- ceph应用情况分析
- OpenStack 使用Ceph 配置指导
- "CEPH浅析"系列之六——CEPH与OPENSTACK
- Ceph与Gluster之开源存储的对比

最新新闻：

- 百度杀毒软件正式谢幕：不再提供下载
- 南极企鹅也没能幸免！体内发现抗生素耐药性基因
- 当部分开源公司考虑封闭的方法
- 希捷四十载：如何做好一家非常规存储公司？
- 酷派起诉锤子欠钱不还：罗永浩回应正在解决

» 更多新闻...