1.

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Ceph分为三大块，分别是对象存储、块设备存储和文件系统服务。在虚拟化领域里，比较常用到的是Ceph的块设备存储，比如在OpenStack项目里，Ceph的块设备存储可以对接OpenStack的cinder后端存储、Glance的镜像存储和虚拟机的数据存储。

2.

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、复制数据、平衡数据、恢复数据等，与其它OSD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OSD，由OSD来对硬盘存储进行管理，当然一个分区也可以成为一个OSD.

伴随OSD的还有一个概念叫做Journal盘，一般写数据到Ceph集群时，都是先将数据写入到Journal盘中，然后每隔一段时间比如5秒再将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盘都是采用SSD，一般分配10G以上，当然分配多点那是更好，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；一个随机写入首先写入在上一个连续类型的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，一般情况下使用SSD作为OSD的journal可以有效缓冲突发负载。

**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，维护Ceph集群的健康状态，同时维护着Ceph集群中的各种Map图，比如OSD Map、Monitor Map、PG Map和CRUSH Map，这些Map统称为Cluster Map，Cluster Map是RADOS的关键数据结构，管理集群中的所有成员、关系、属性等信息以及数据的分发，比如当用户需要存储数据到Ceph集群时，OSD需要先通过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，但对象存储和块存储设备是不需要使用该服务的。

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

3.

**Ceph数据分布算法 （CRUSH）：**

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，常见的数据分布算法有一致性Hash和Ceph的Crush算法。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的，数据分布算法必须能够满足在大规模的集群下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可能小的数据迁移，Ceph的CRUSH算法就是精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多个object，每个object都有一个object id，每个object的大小是可以设置的，默认是4MB，object可以看成是Ceph存储的最小存储单元。

**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object，每个object最后都会通过CRUSH计算映射到某个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，则每个pg都会映射到二个osd，比如 [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，保证了数据的冗余。

**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，我举个例子，比如有3个osd，osd.1、osd.2和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[osd.1, osd.2], 也可能是[osd.2, osd.3]，也可能是[osd.1, osd.3] 但只能其中选一个，如果选择的是[osd.1, osd.2]那么所有的pg主从副本分别存放到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[osd.1,osd.3]，是不是很像我们高中数学学过的排列组合，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。

**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，都需要指定pg和pgp的数量，逻辑上来说pg是属于某个存储池的，就有点像object是属于某个pg的。

