参考资料：

https://juejin.cn/post/6844903544961630221

对于分布式存储系统，数据分布在各个节点中，需要一种有效的数据分布机制，不但能够被快速寻址，并且数据的分布能够考虑到容量均衡和负载均衡。目前现有分布式存储系统的数据分布实现中，很多采用中心节点分配数据分布地址并查表索引，随着数据量的增大，寻址开销也越来越大。

其他分布式存储系统，比如CEPH的数据分布采用哈希计算寻址，可以解决上面的问题，但也存在几个不足：

1）哈希计算分配数据地址，存在一定的随机性，随着系统的使用，各个

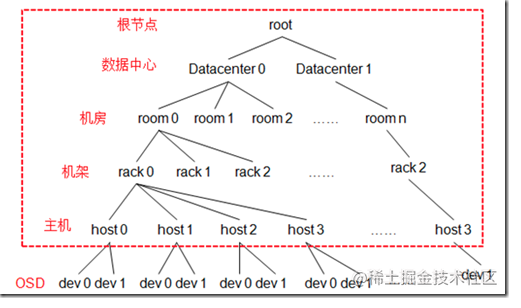
存储节点的容量使用很不均衡，在一些不利配置的条件下，最大和最小容量之间最高相差50%，即使对数据进行复杂的重新分布均衡优化，也仅能达到90%左右的容量使用率。造成容量的浪费；

2）扩容或硬盘故障会导致数据分布发生变化，导致几乎整个系统产生数据迁移。比如在Ceph中，数据最小地址单位（比如CEPH的归置组Placement group）是动态变化的，当存储池中出现故障或扩容时，会导致大量数据发生迁移，由于哈希分布的随机性，当输入发生变化后，输出是整系统联动的，会导致整个存储池都参与数据的重新分布，过分影响用户的正常使用；

1、数据分布原理

1.1、osd map

osd map是个树形结构，其中HDD,SSD等存储设备作为叶子节点OSD。其他节点根据物理结构进行抽象，比如机房抽象，主机抽象，主机由于存在唯一的IP等信息可以被软件实时检测并自动抽象，其他节点可以人工配置抽象。其中数据中心，机房，机架，主机等中间层级都是可选的，可以根据实际需要配置。它可以作为crush算法的输入属性，用来生成对应归置组Placemen group的osd列表，如下图所示



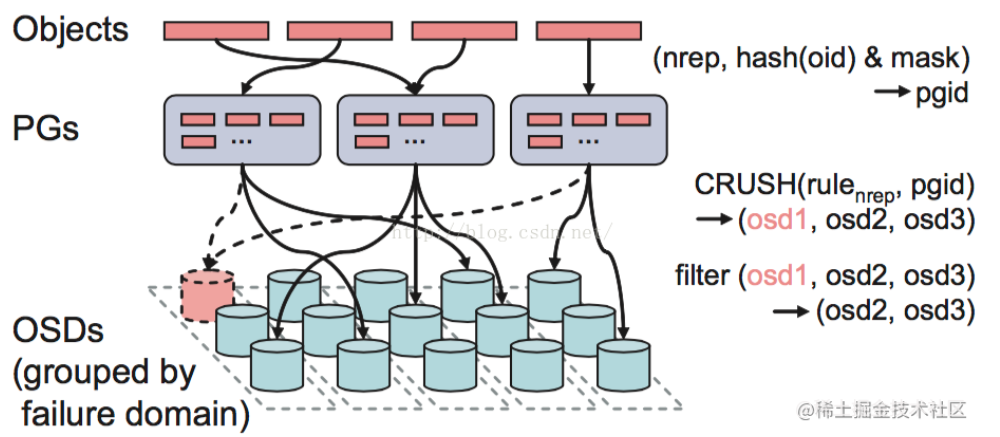
1.2、crush算法

crush算法主要是用来控制数据分布，它的输入包括osd map，x（pgid），定位规则rule，输出是一组osd Dev配对列表(osd0, osd1, osd2 … osdn)。其中pgid是数据最小地址单位Placement Group，pgid根据数据冗余规则包含了多个osd配对列表，比如pgid1=（1,2），pgid2=（2,3），数据通过哈希算法映射到不同的pgid，并且再进一步分布到不同的存储osd Dev中。

crush算法利用强大的多重整数hash函数根据osd map、定位规则rule、以及pgid计算出独立的完全确定可靠的映射关系 ：

*CRUSH(x=pgid,osd map, rule) = (osd0, osd1, osd2 … osdn)*

pg内的存储设备配对应该尽量满足故障域分离，比如分布到不同的host或者不同的机架，以便达到分布式容灾的效果。大致过程如下，crush算法遍历osd map的叶子节点，通过伪随机算法分配一个osd，接着根据crush中定义的take type类型，寻找不同（disk，host，rack等）下的其他叶子节点osd，依据同样规则分配足够crush中定义的osd size个数。最后会得到一个该pgid下的osd列表，其中osd是不同的，并且分别属于不同故障域（比如在不同的主机host中）。如下图所示。



1.3、数据寻址流程

如上图所示，一个PG会被映射到n个OSD上，而每个OSD上都会承载大量的PG，即PG和OSD之间是“多对多”映射关系。在实践当中，n至少为2，如果用于生产环境，则至少为3。一个OSD上的PG则可达到数百个。事实上，PG数量的设置牵扯到数据分布的均匀性问题。

数据写入时，数据object先映射到PG，首先要根据object名称计算数据object的Hash值并将结果和PG数目取余，以得到数据object对应的PG编号。然后，通过CRUSH算法将PG映射到一组OSD中。最后把数据object存放到PG对应的OSD中。这个过程中包含了两次映射，第一次是数据object到PG的映射，第二次是PG到osd的映射。PG是抽象的存储节点，它不会随着物理节点的加入或则离开而增加或减少，因此数据到PG的映射是稳定的。

2、原因分析

通过上面的原理分析，我们继续分析，为何会产生前言提到的两个缺点。

1）容量不均衡。在数据分布的两次映射中，对象到PG的映射是通过哈希值取模，对象名是不可控的，因此只有在数据量大的情况，可以达到一个大致的均衡分布。再就是第二次映射，ceph会生成一系列的pg数目，包含了一些osd列表，它是一种伪随机算法，pg的数目也会影响数据均衡的效果，也会产生一些不稳定因素。

2）扩容或故障导致大量数据重建。在crush算法中，PG是相对稳定的，但是osd map包含了osd的设备状态，当osd map中增加新的设备，或者某些设备发生了故障，会导致PG内的osd列表发生变化以便可以进行数据重建或均衡。比如某个host0的osd0发生了故障，因为osd和pg之间的关系是多对多的关系，所以包含osd0的pg都要进行数据重建，基本上这些pg会分布到几乎所有主机host上，并在所有主机上产生数据迁移，并影响整体系统性能。而且这种性能衰减并非线性的，硬盘的特点是要么写要么读，随机读写的混合任何产生伴随着巨大的性能衰退。