存储架构解决方案

目录

[存储架构解决方案 1](#_Toc502332584)

[一、系统架构 3](#_Toc502332585)

[1. 大容量数据存储现状 3](#_Toc502332586)

[2 解决方案 3](#_Toc502332587)

[1.2.1 基于CEPH的统一存储方案 3](#_Toc502332588)

[1.2.2 系统架构图（对象存储架构图） 4](#_Toc502332589)

[二、实施方案 4](#_Toc502332590)

[1. 硬件需求理论详述 4](#_Toc502332591)

[CPU 4](#_Toc502332592)

[RAM内存 4](#_Toc502332593)

[数据存储 5](#_Toc502332594)

[硬盘驱动器 5](#_Toc502332595)

[固态硬盘 5](#_Toc502332596)

[衡量SSD性能的指标 5](#_Toc502332597)

[硬盘控制器 6](#_Toc502332598)

[其他注意事项 6](#_Toc502332599)

[2. 最低硬件配置推荐 6](#_Toc502332600)

[生产集群实例 7](#_Toc502332601)

[网络 7](#_Toc502332602)

[现有Ubuntu系统的服务器处理器配置： 8](#_Toc502332603)

[3. 设备配置标准清单 9](#_Toc502332604)

# 一、系统架构

## 1. 大容量数据存储现状

详述目前遇到的问题。

## 2 解决方案

### 1.2.1 基于CEPH的统一存储方案

#### CEPH介绍

#### CEPH优势

#### CEPH架构

### 1.2.2 系统架构图（对象存储架构图）

# 二、实施方案

本部分从搭建ceph集群的硬件需求方面，先从理论详述，再结合实际需求，给出最低硬件配置推荐。

## 1. 硬件需求理论详述

**Ceph OSDs：**守护进程的功能是存储数据，处理数据的复制、恢复、回填、再均衡，并通过检查其他OSD守护进程的心跳来向Ceph Monitors提供一些监控信息。当Ceph存储集群设定有2个副本时，至少需要2个OSD守护进程，集群才能达到active+clean状态。

**Monitors：**Ceph Monitor（类似Zookeeper）维护着集群状态的各种图标。

**MDSs：**元数据服务器为Ceph存储元数据。

Ceph把客户端数据保存为存储池内的对象。通过CRUSH算法，Ceph可以计算出那个PG（归置组）应该持有指定的对象，进一步计算出那个OSD守护进程持有该归置组

**Ceph硬件需求：**Ceph为普通硬件设计。规划集群硬件时，需要均衡几方面的因素，包括区域失效和潜在的性能问题。硬件规划要包含把使用Ceph集群的Ceph守护进程和其他进程恰当分布。**通常，我们推荐在一台机器上只运行一种类型的守护进程。**推荐把使用数据集群的进程安装在别的机器上。

### CPU

Ceph元数据服务器（MDSs）对CPU敏感，它会动态的重分布它们的负载，所以元数据服务器应该有足够的处理能力（**如4核或以上的CPU**）。Ceph的**OSD**运行了很多服务（RADOS、CRUSH计算数据存放位置等）所以OSD需要一定的处理能力（如双核CPU）。Ceph的监视器（Monitors）只简单地维护着集群运行图的副本，因此对CPU不敏感，但是得确保给Ceph保留足够的处理能力。

### RAM内存

元数据服务器（MDSs）和监视器（Monitors）必须可以尽快的提供它们的数据，所以他们应该有足够的内存，至少每个进程1G。OSD的日常运行不需要那么多内存（每个进程500M），但是在恢复期占用的内存会比较大（通常每个进程处理每个TB的数据需要大约1G内存）。通常提供的内存越多越好。

### 数据存储

来自操作系统的并行操作和到单个硬盘的多个守护进程并发读、写请求操作会极大的降低性能。文件系统局限性也要考虑：btrfs尚未稳定到可以用于生产环境的程度，但它可以同时记日志并写入数据，而xfs和ext4却不能。

**重要：**因为Ceph在发送ACK前必须把所有数据写入日志（至少对xfs和ext4来说是），因此均衡日志和OSD性能相当重要。

### 硬盘驱动器

推荐使用独立的驱动器用于安装操作系统和软件，**每个OSD守护进程占用一个驱动器。**Ceph允许你在每块硬盘驱动器上运行多个OSD，但这会导致资源竞争并降低总体吞吐量；Ceph也允许把日志和对象数据存储在相同的驱动器上但这会增加记录写日志并回应客户端的延迟。**Ceph最佳实践指示，应该分别在单独的硬盘运行操作系统、OSD数据和OSD日志。**

### 固态硬盘

一种提升性能的方法是使用固态硬盘（SSD）来降低随机访问时间和读延时，同事增加吞吐量。SSD和硬盘相比每GB成本通常要高10倍以上，但访问时间至少比硬盘快**100**倍。

**重要：**建议发掘SSD的用法来提升性能，建议核实SSD的性能指标，并在测试环境下衡量性能。

### 衡量SSD性能的指标

1、写密集语义：记日志涉及写密集语义，所以得确保选用的SSD写入性能和硬盘相当或好于硬盘；

2、顺序写入：在一个SSD上为多个OSD存储多个日志时也必须烤炉SSD的顺序写入极限；

3、分区对齐：采用了SSD 的常见问题是喜欢分区，如果忽略分区对齐，会导致SSD的数据传输速率变慢。

SSD用于对象存储太昂贵了，可以把OSD的日志存到SSD，把对象数据存储到独立的硬盘可以明显提升性能。osd日志的默认选项是在/var上，可以把它挂载到一个ssd或者ssd分区上，这样就可以把日志和对象数据分开存储。

提升CephFS文件系统性能的一种方法是从CephFS文件内容里分离出元数据。Ceph提供了默认的metadata存储池来存储CephFS元数据，所以你不需要给CephFS元数据创建池，但是可以给它创建一个仅指向某主机SSD的CRUSH运行图

### 硬盘控制器

硬盘控制器对写吞吐量也有显著影响，要谨慎选择，以免产生性能瓶颈。

### 其他注意事项

可以在同一主机上运行多个OSD，但要确保总吞吐量不超过为客户端提供读写服务所需要的网络带宽；还要考虑集群在每台主机上所存储的数据占总体的百分比，如果一台主句所占百分比太大而挂了，可能会导致full ratio的问题，这会导致Ceph挂了而数据丢失。如果每台主机运行多个OSD，也得保证内核是新的。

**OSD数量较多的主句会派生出大量线程，尤其是在恢复和重均衡期间。很多linux内核默认的最大线程数较小，可以把kernel.pid\_max值调高一些。理论最大值是4194303。可以把这个配置项加入到/etc/sysctl.conf文件中。**

## 2. 最低硬件配置推荐

****

### 生产集群实例

PB级生产集群也可以使用普通硬件，但应该配置更多内存、CPU和数据存储空间来解决流量压力。

如下是DELL实例中配置：



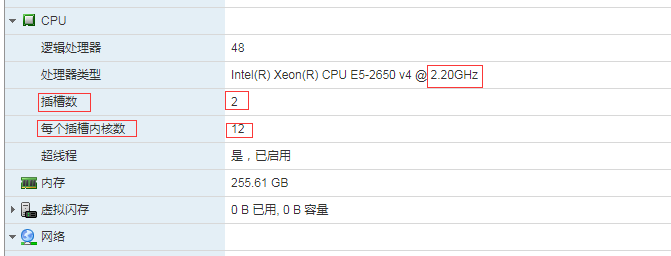
### 网络

建议每台机器最少两个千兆网卡，网卡应该可以处理所有OSD硬盘的总吞吐量，分别用于公网和集群网络。集群网络用于处理由数据复制产生的额外负载（考虑万兆网卡）。在一个PB级集群中，OSD磁盘失败是常态，而非异常

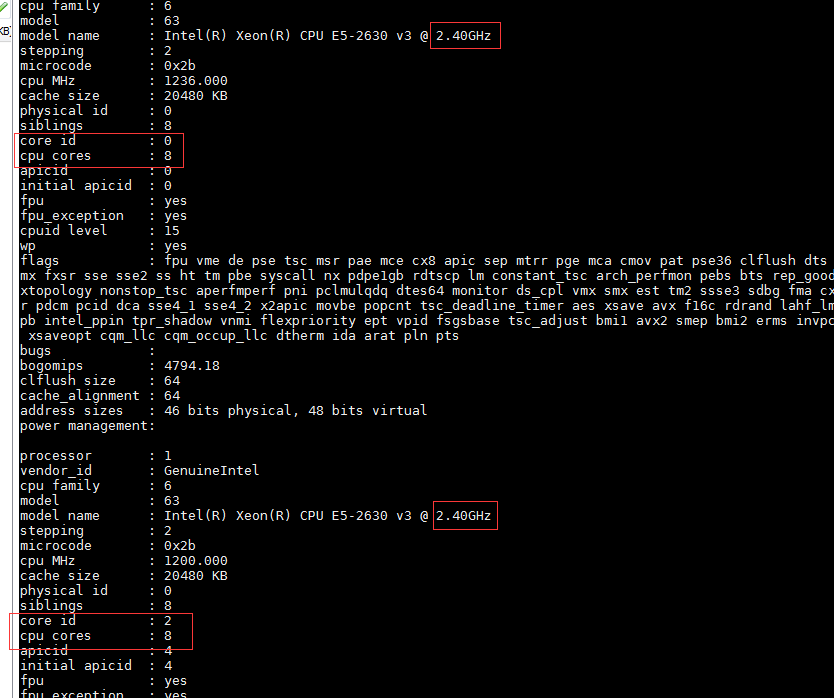
每个osd守护进程至少有一个cpu核

计算公式如下：（cpu个数\*核数\*cpu在GHZ的速率）/osd的个数需要大于等于1

例如：一台服务器拥有一个双插座，12核，2.20Ghz的cpu，就足以支持50个osd

（2\*12\*2.2）/ 50 = 1.056 

### 现有Ubuntu系统的服务器处理器配置：



插槽：2个，每个核数：8核，cpu在Ghz中的速率为：2.4Ghz

内存：ceph mds很大程度上取决于数据缓存，需要大量的内存（1TB需要1GB以上），内存越高，cephfs性能越好；Osd会要求数量可观的内存，一般每个OSD守护进程需要1G内存，从性能上将每个守护进程2G是更好的选择

网络：少量节点ceph集群，千兆网络速率是可以满足正常运行的，如果是一个中型或大型的网络（几十个节点），应该使用万兆网络

硬盘：当一个osd接受请求写一个object时，会先把object写到osd对应的日志盘，然后发送一个写确认给客户端，很快日志数据会同步到数据盘，使用ssd做日志盘，可以减少访问时间，降低写延迟，提升吞吐量

日志盘：在ssd上获取高性能，ssd和osd的比例应该为1:4,也就是说4个OSD数据硬盘可以共享一个ssd

现有Ubuntu有8个硬盘插槽，现有市场机械硬盘有3T、4T（比较贵），需要保存的数据量大致为：

2U服务器可插8个磁盘 每个磁盘3T，可插24T

3U服务器可插16个磁盘，每个磁盘3T，可插48T

450T 743路

1211T（1.18PB） 2000路



## 3. 设备配置标准清单

（例如按2000路存1年计算，可能需40-50台服务器），……