中国科学技术大学

工程硕士学位论文



**基于ceph的云存储运维系统的设计和实现**

作者姓名： 孙君伟

学科专业： 软件系统设计

导师姓名： 周英华

完成时间： 二○一六年八月三日

University of Science and Technology of China

**Limited**

A dissertation for master’s degree

of engineering



**Design and implementation of cloud storage operation and maintenace system based on ceph**

Author’s Name： Junwei Sun

Speciality： The design of software system

Supervisor： Ms. Yinghua Zhou

Finished time: August 3nd, 2016

书脊

|  |
| --- |
| **基于C**  **E**  **P**  **H的云存储运维系统的设计和实现**  **孙君伟**  **中国科学技术 大学** |

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文,是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：\_\_\_\_\_\_\_\_\_\_\_ 签字日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

□公开 □保密（\_\_\_\_年）

作者签名：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 签字日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

摘要

此次主要讨论基于ceph的云存储运维系统的设计和实现。基于ceph的云存储运维系统，是为了实现将网络中相互独立，型号各异，价格低廉的存储设备统一组织起来，并对外提供统一接口。从而实现具有备份、容灾、易于维护特性的存储系统。基于ceph的云存储系统充分利用每一个磁盘，能提供高稳定性和高并发的特性。同时具备易于扩展的特点。

基于ceph的云存储运维系统是基于ceph基础之上对磁盘进行监控的系统，在ceph层可以实现的功能如下：

1. 容灾备份：

当出现突发情况时，造成局部磁盘数据损坏而造成的数据丢失，可以通过ceph的容备特性，等排出故障磁盘后通过存储在其他地方的数据来恢复原先坏掉磁盘的数据。对于一些重要的信息，以实现将原来的两地三中心，变成三地三中心，值要不同时破坏掉对应数据的所有存储场所，对应的数据都是安全可靠的

1. 高效存取：

现在存储数据PB级数据已经很常见了，EB级数据也时有出现，这就面临着一个问题，如何更好的存储这些数据。对于大数据，如何有效存储至关重要，这不仅仅关系到数据存储的安全，还包括读取数据的效率。基于ceph的云储存运维系统，提供了很好的存储方案。其设计充分发挥了每台对应服务器的计算能力，摆脱了传统单一中心节点的瓶颈，实现了真正无中心的特性。其中有八个字概括这一特性：无需查表，算算就好。这充分体现了整过数据存取的过程中不会因单点故障，而造成系统故障，同时保证了存取的高效

1. 负载均衡

在数据存储的过程中，难免会出现磁盘损坏，磁盘扩容，以及不同磁盘集聚的磁盘使用率的失衡，这样将会影响到数据存取效率。如果能做到将同一个大文件合理的存储在不同磁盘上，将会使数据存储效率有明显的提升，充分利用系统提供的高并发特性，实现数据的并发存取。避免单个磁盘因频繁被操作而缩短使用寿命。

在用户界面层将实现和用户交互的监控界面，隐藏底层操作细节，通过用户的点击实现对应的操作

关键字：容灾备份，均衡，并发ABSTRACT

The design and implementation of cloud storage operation and maintenance system based on CEPH is mainly discussed. In order to achieve the network in each other independent, different models, inexpensive storage devices unified organization, and to provide a unified interface to the outside world, in order to achieve an easy storage, backup and disaster tolerance system maintenance characteristics, the cloud storage system based on CEPH was appearance. The cloud storage system based on CEPH makes full use of every disk, which can provide high stability and high concurrency. At the same time, the cloud storage system has the characteristics of easy to expand.

CEPH, which based cloud storage operation and maintenance system, is the monitor system, based on the disk. In the CEPH layer, we achieved the function as follows:

1. back and up:

It is possible which in the emergency situation, the disk in someplace was damage and resulting in the disk data was loss. Using the CEPH which has the characteristic of content preparation, the disk data can be recover when replacing the broken disk by other place disk data. For some important information, it has three places storage data and it have three center point of primary, rather than two places for storage data and three center point of primary. Using the system, if not make the same data was destroy at the same time, the corresponding data are safe and reliable

1. efficient access:

Now the data which stored disk has TB level is very common. And the data which has PB level was also appearance. That will make a problem, how to efficient store these data. For big data, effectively store is essential. It is not only related to the safety of data storage, but also includes the efficiency of reading data. The cloud storage system based on CEPH provides a good method for storing data. The design gives full play to the computing power of each corresponding server, gets rid of the bottleneck of the traditional single center node, and realizes the real non center characteristic. There are eight words summed up the characteristics: no look-up table , it is of for calculating. This fully reflects that the whole process of data access will not cause the system failure as a single point breakdown, while ensuring access to efficient

1. load balance

In the process of data storage, it is inevitable that there will be a disk damage, disk expansion, as well as the disk of different disk use rate of imbalance, which will affect the efficiency of data access. If you can store a big file in a good method, the data storage efficiency will be significantly improved. These will full use of the system to realize the character of the high concurrency, the realization of data access. To avoid a single disk has shorten the service life due to frequent operation.

In the user interface layer, I will be implemented the function that used monitor the state of disk, hide the underlying operating details, through the user's click to achieve the corresponding operation.

Key Words: Backup, balanced, concurrent

1. 绪论
   1. 选题背景和意义

（1）个人需求

当今社会，互联网已经普及到每个人，数据信息比如照片、电影、工作文件自然而然的出现，这就需要存储设备去保存这些信息，且随着社会快速的发展，人们需求不断扩大，对他们的要求也不断提高，这些数据信息的体积也越来越大。U盘、光盘、硬盘等凭借大容量，相对体积小，价格适中，存储稳定等特性轻易的战胜了传统的软盘，成为当代青年日常生活必需品。而随着网络磁盘的到来，很多人又将目光转移到了网盘。它相对于当代实体存储介质，有不可比拟的有点。当采用网盘存储数据时，无论在哪，只要能联网就随时可用联网终端设备存取自己的资料。也为多人协同工作带来了极大的便利。很多网盘，比如360云盘、百度网盘等都提供免费的，较大容量的网盘，对于普通用户，完全可以满足日常需求。这是云存储给个人带来的便利。使人们找到了一种和物理介质存储设备一样具有存储性能的方式，也减少了因为为物理设备的损坏或丢失而造成的不可恢复的灾难。

（2）大型应用

随着云计算和物联网的兴起，智能应用成为当今一大主题，智能应用主要涉及到信息的收集，整理和应用。因涉及范围广，需求各异，会形成海量复杂的信息数据，这些信息如果还是存储在由政府，企业提供的孤立的存储介质上，那这种代价将会非常高[1]。随着后期数据的不断增加，管理起来会越来越复杂。还要购买昂贵的储存介质，配备有专门的人员去维护。同时，存储介质的升级换代也是个很大的问题。信息的可靠性也无法得到保证。综合各种因素，采用普通的，孤立的存储介质将会花费高昂的代价，效果也不理想。这些弊端通过云存储系统就可以很好的解决，云存储系统可以按需分配磁盘空间，后期可以轻松扩容，在数据安全上也会有很好的保证。在管理上有专门的人员负责。这样就不会存在因一次性购买大量存储介质而造成的浪费，也不用为扩容时去买新的存储介质。花费代价上远小于自己独立管理。给政府，企业带来极大便利。下面通过生活中一个例子去感受下。

当代数字化迅速发展的今天，安全防护监控在日常生活中的需求也在不断升级。安全监控将会产生大量数据，随着社会的发展，与安全监控匹配的存储也经历好几代的发展，从初始的VCR模拟存储到DVR数字数据存储，又从DVR数字数据存储到现在的网络

集中存储。因分辨率的提高，现在产生的数据是PB级数据，如果采用基于ceph的云存储运维系统，将会很好的满足这种需求。

* 1. 国内外研究发展状况及发展趋势

1.2.1 ceph概述

此次毕业论文设计是基于ceph[2]分布式系统的，接下来让我们看下ceph分布式系统在国内外现状和发展。ceph是一种具有高性能，高扩展性和高可用性的分布式存储系统。可以提供对象存储，块存储和文件系统存储三种功能[3]，从而使其满足不同应用需求下的部署和运维。ceph可以部署在上千台服务器上。ceph项目由Sage Weil在加州大学Santa Cruz分校攻读博士期间研究的课题。该项目是从2004年开始，在2006年的OSDI学术会议上，Sage Weil发表了介绍ceph的论文，并提供了ceph项目的下载链接。从此，ceph开始广为人知。ceph分布式系统是由Sage Weil在2007年提出，作为其博士论文设计的新一代分布式系统。在2010年三月，应用到Linux内核。随着ceph的不断发展，Sage Weil于2011年创建了Inktank公司，主导ceph的开发和社区维护。随后，ceph分布式系统的发布周期为三个月。Ceph分布式系统虽然非常优秀，但随着ceph的不断发展，还是有些问题需要改进或优化。比如：

<1> 数据双倍写入的性能：因为ceph分布式系统为了支持事务，引入了日志机制。也就是说，一份数据需要写两遍，即日志和本地系统文件，使得实际的磁盘输出的吞吐量只有物理性能的一半。

<2> IO路径太长：这种情况在客户端和服务器端都有存在，比如对于osd，一个IO的操作需要经过message、OSD、FileJournal、FileStore多个部分才可以实现整个过程，而每个部分都会涉及到队列和线程的切换，有些模块在对IO进行处理时还会涉及内存拷贝，因此造成整体性能不高。

<3> 兼容性：ceph分布式系统刚开始是基于HDD设计的，所以没有全面考虑到SSD。所以没有能充分发挥新型存储介质的性能。特别是在延迟和IOPS方面，问题较为突出。

ceph分布式操作系统对外提供丰富的操作接口，给管理网络磁盘提供了便利。使得基于ceph的云存储运维系统自然而然的出现。

1.2.2 基于ceph的云存储运维系统的现状

基于ceph的云存储运维系统，是在ceph分布式系统的基础上，通过calamari做了进一步封装，实现了对磁盘信息的监控和部分操作。不同厂商会提供不同用途的存储设备，根据不同用途，会有不同的监控系统，其主要重点实现在云存储系统上。下面来了解下基于ceph的云存储系统的一些现状。它在ceph分布式系统和云计算的基础上发展而来的，通过集群的应用，将网络上不同的存储介质统一组织起来，并对外提供存储功能的一个新兴存储方式。相对于传统的孤立存储，基于ceph的云存储系统能提供更加便携，安全，可靠的性能[4]。下面将通过基于ceph的云存储系统的优势和劣势来分析下基于ceph的云存储系统的现状：

优势：基于ceph的云存储运维系统可以按客户需求，分配对应存储空间，且随着后期的需求的变化，可以及时方便的进行扩容。且在对设备进行升级或扩容时，给客户提供极大便利，基本不会影响客户正常业务，节省了客户维护方面的开销[5]。同时也节省升级服务器的费用。在数据安全方面，云存储会自动检测故障磁盘，当出现故障磁盘时，系统会根据预先设计的规则，将故障磁盘上的数据自动迁移到正常运行设备上[6]。从而保证了数据的安全，并且不影响客户的正常操作。随着供应商的日益增多，使用费用将会不断降低。

劣势：基于ceph的云存储运维系统在存储介质稳定性方面有很好的保证，但在信息安全（主要指被非法用户获取）方面存在明显不足。因为数据存储是通过网络进行存储，所以有可能被中途拦截，同时对网络磁盘进行管理维护的工作人员可以轻易获得存储在云磁盘上的数据。虽然通过加密可以降低信息泄露的风险，但在数据处理方面将变得比较复杂，使得数据处理能力下降[7]。能耗大也是云存储面临的另一重大挑战。一方面，随着基础规模的扩大，能量消耗将会急剧增加。另一方面，随着数据量的增加，启动设备组件增加，所需能源越来越大。有统计，在2010年，其对电能的消耗为2x1011千瓦时，约占全球总电量的1.3%，且其比例还在不断增加[8]。存储系统会在能源消耗中占很大比例，仅次于计算资源耗电量，约占30%[9]。因此，现在云存储的能耗也是云存储的一个重要的关注点。基于ceph的云存储运维系统会涉及到负载均衡，在小规模集群的环境下，负载均衡体现的效果不是很好[10]。

1.2.3 基于ceph的云存储运维系统发展趋势

安全性：基于ceph的云存储运维系统在其存储方面，安全性一直以来都是基于ceph的云存储系统的最大焦点之一[11]。如何保证用户的数据不被盗用，保证用户的数据在传输过程中不被篡改，将会成为云存储发展的一个重要目标之一。只有很好的解决了这些基本的安全问题，才能让更多的用户放心选择使用基于ceph的云存储系统之上的各种应用。

多样性：由于客户需求的多样性，基于ceph的云存储运维系统将显示各种各样的应用，同时会使得数据存储更加复杂，这就要求基于ceph的云存储运维系统在应对将来更多需求，能提供更加丰富和完善的机制。备份、容错、恢复和快照等方面做的更加完善[12]。

易用性：物联网在不断发展，将会出现用特定设备访问特定数据。这将会使得基于ceph的云存储运维系统在和各个终端设备上有更好的兼容，完善用户体验。同时，采用更先进的技术提高数据读写速度。

* 1. 云存储系统设计的先进性和意义

对于个人：当多人协作，共同完成一件事时，就会涉及到信息的一致性。采用网络数据共享可以很好的做到信息同步，并能将自己的工作及时告知他人，防止出现重复劳作，同时保证了其他人可以及时获得自己工作的内容以便他们后续工作。

对于企业：通过基础ceph的云存储系统，可以节省一大笔开销，比如在更新设备，升级服务，存储管理等方面，基于ceph的云存储系统显然比企业独立存储系统占优势[13]。同时通过基于ceph的云存储系统，在磁盘故障方面有很好的保证，基于ceph的云存储系统会自动迁移故障磁盘数据，故障恢复快，使企业不必担心因磁盘故障造成的业务中断[14]。在磁盘扩容，服务升级方面有着孤立磁盘无法比拟的优势[15]。

* 1. 本人主要工作

当前负责运维系统的前端界面开发和用户前端与服务器交互的后端开发。还做磁盘性能测试。

1.4 论文组织

1. 理论概述

2.1 云存储

2.1.1 云存储出现场景

随着互联网的发展，以及云计算和物联网的出现，人类产生的数据信息在以极快的速度产生。具体有多快，我们通过一个名人提出的定律感受下，图灵获奖者Jim Gray曾提出的据增长经验定律——网络环境下每18个月产生的数据量等于有史以来数据量之和。通过他提出的定律，我们可以更好的感受下当前社会中数据增长的速度。随着而来的问题就是如何能安全、有效的存储这些数据，同时保证在人们需要这些数据时，有能安全、高效的呈现给用户。此时，云储存就自然而然的应运而生。

2.1.2 云存储的关键技术

随着不同种类数据的急剧增加和人们对信息安全需求的不断提高，拥有安全，性价比高的存储机制成为企业，政府等机构的关注的焦点。而云存储具备：

1. 易扩充

云储存具有按需定制所需空间。随着社会的发展，数据几乎成爆炸式的方式出现，作为公司，企业，政府他们将来预测不到自己需要多少存储。如果还是采取传统方式，自购存储设备，必然会出现前期存储闲置，后期存储设备紧张。这两种情况都会造成极大的浪费。而如果才用当前所提供的云储存技术，按需申请存储空间，后期出现大量数据时，可以申请扩容，方便便捷。避免了闲置浪费和后期因存储紧张造成的不便。

1. 易管理

随着信息量的急剧增加，存储介质也在发生翻天覆地的变化，存储之间的差异在所难免，云存储会将这些存储设备协调起来，通过云存储上的软件优化存储性能。同时云存储提供专业化管理策略，对资源进行了整合。从而极大的提高了存取效率。

1. 高性能

云存储采用的是分布式存储方式，可以根据数据访问，使用频率，磁盘使用率的不同进行自动均衡，充分利用每个服务节点，极大的发挥系统的并发性能。让客户在为大文件的存储而感到烦恼。

1. 安全

云存储所提供的服务对象范围广泛，因此出现了对服务需求的多样化，有些数据追求可靠，有些追求私密，有的追求完整，有的追求高效率。这就要求云存储能做到满足不同用户不同需求，只有这样，云存储才能被大众所接受。为此，云存储在提供物理隔离的同时还提供了权限控制，除此之外，云储存还能够对对应的数据进行加密保护。为了提高存取效率，云存储还进行了切片处理，这不仅提高了存取速度，在一定程度上还有保密效果，才用这样存储，防止非法用户通过单个存储节点获取所有对应信息。

2.2 ceph机制

2.2.1 集群框架机制

在ceph的框架有以下几个部分组成：

1. 客户端

通过客户端，用户可以实现根据自己的需求，进项相应的操作。

1. 元数据服务器

用户处理来自客户端传输的一些数据，将这些数据进行编序、切割，形成便于搜索和存储的数据单元

1. 对象存储设备

对象存储设备，不仅仅具备存储功能，还具备通信能力，可以实现和其他设备间的通讯

1. 集群监控器

集群监控器主要功能是监控，当系统中出现有故障的设备，或者当系统主插入了先的存储设备，集群监控器能感知这些变化，并作出相应信息的更新，保证信息表的准确性

其结构图2.1如下：



图2.1 ceph集群框架

2.2.2 存储机制

基于ceph的存储有其独特的存取方式，也正是这种存取方式，是的基于ceph的存储系统具有高并发的特性，同时这种存储方式也消除单点故障瓶颈。如图2.2是ceph存储机制：



图2.2 ceph存储机制

通过图5我们可以看到，存取的实现主要有以下几个部分：

1、file -> object

实现方式是给每个文件一个唯一id，然后再将较大的文件拆分成事先规定好的，易于在网络中传输的各个块，每个数据块都有自己的编号，用文件id+块编号就可以在整个磁盘系统中为数据块指定唯一的一个编号，从而将文件拆分成易于存储的objects，方便后续操作。

2、objects -> PGs

将文件分割成诸多objects后就需要将这些文件映射到对应的PGs上去，映射规则比较简单，可通过公式：hash(oid)&mask -> pgid实现对应的映射。其中里面有些限制。之所以采用这种方式，是为了让一个文件所分割的各个部分均匀的放在网络磁盘上，为了实现这种情况，要保证PG总数为2的整数次幂。

3、PGs -> OSD

这个过程是将object的逻辑组织单元的PG映射到实际存储单元osd上，采用CRUSH算法，将pgid带入其中，可以得到一组中有n个的osd，其中n既是为了实现可靠性而设置的一个值，代表副本数。在生产环境下通常为3。

到此，基于ceph分布式系统下的低层，完成了从file -> object -> PG -> OSD整个过程的映射，从整个过程看，我们发现没有一个地方会涉及到查询全局性的表来获取数据，从而实现高并发特性。这也是ceph之所以被很多企业所接受的最重要的一个原因。到此整个ceph的存储过程已经完成。

此外，在ceph上面还有cluster map，在维护磁盘负载均衡方面发挥重要特性，也是为磁盘高效率做出了很大的贡献。随着不断向网络磁盘存储数据，网络磁盘数据就会出现分布不均衡的现象。这就需要通过均衡机制去实现负载均衡。下面讲述cluster map的特性。

在cluster map中含有版本号、各个OSD的网络地址、各个OSD的状态、以及CRUSH算法配置参数。其中需要说明的是OSD的状态，OSD的状态包括：

up&in：该OSD运行正常，且至少承载了一个PG。

up&out：该OSD运行正常，但并未承载任何PG。

down&in：该OSD发生异常，但仍然承载至少一个PG。

down&out：该OSD发生异常，且已经不再承载任何PG

其中cluster map可以收集到这些信息，并以增量形式进行扩散，其中版本号是递增的形式，当出现版本不一致时，以较新的为准。且cluster map的信息是以异步且lazy的形式扩散的，从而避免了广播风暴。最终会将cluster map扩散到全体OSD和client端。OSD会根据cluster map进行数据维护，而client则是根据OSD进行数据的寻址。从而使得基于ceph的云储存具有高效的存取性能。

2.2.3 高可用性机制（需重新作图）

2.3 总结

第三章 需求分析

3.1 系统概述

3.2 系统总体架构

3.3 云存储性能需求分析

3.4 云存储功能需求分析

3.4.1 管理系统需求

3.4.2 应用系统需求

3.5 总结

第四章 系统设计

4.1 系统架构设计

4.2 云存储底层设计

4.3 接口设计

4.4 云存储管理系统的设计

4.4.1 集群管理

4.4.2 系统管理

4.5 总结

第五章 系统开发和实现

5.1 虚拟化层的开发与实现

5.1.1 集群环境搭建前的准备

5.1.2 ceph基础环境配置

5.2 接口层的开发与实现

5.3 存储管理系统的开发与实现

5.4 总结

第六章 系统的部署和应用

6.1 存储底层ceph基础环境安装和部署

6.2 存储应用云盘系统的部署与安装

6.3 系统测试用例及功能表现

6.3.1 云存储管理系统

6.3.2 云盘系统

6.4 总结

第七章 总结

参考文献

附录

致谢