基于Ceph对象存储集群的高可用设计与实现

杨 飞[1],朱志祥[2],梁小江[2]
(1 西安邮电大学,陕西 西安 710061;2 陕西省信息化工程研究院,陕西 西安 710061)

摘 要:为了实现一种基于ceph对象存储集群的高可用设计方案,先搭建ceph集群生态系统,然后设计和部署ceph对象存储集群,实现多对象网关发布对象存储服务.通过keystone统一认证中心保证多区域的用户访问的安全性,最后整合haproxy和keepalived,设计和实现基于ceph对象存储集群的高可用设计方案.在相同实验环境下,首先测试ceph存储集群健康状况,保证整个ceph存储集群正常运行,最后通过大量的网络压力测试和分析,证明本系统能实现ceph对象存储集群的高可用性.

关键词:ceph;高可用性;keystone;对象存储;haproxy
中图分类号:TP302.1 **文献标识码**:A **文章编号**:1000-7180(2016)01-0060-05

Design and Implementation of a High Availability Cluster Based on Ceph Object Storage

YANG Fei[1], ZHU Zhi-xiang[2], LIANG Xiao-jiang[2]
(1 Xi'an University of Posts and Telecommunications, Xi'an 710061, China;
2 Institute of Communication Technology, Xi'an 710061, China)

Abstract: In order to design and achieve a high availability cluster ,which is based on ceph object storage . At the first, set up a healthy ceph storage cluster,then, design and deploy the ceph object storage cluster that is based on ceph storage cluster, which can achieve the service of multi-object storage . ensure security of cloud storage capabilities from multiple areas of users by keystone Unified Certification Center, finally, design and implement the high available cluster of ceph object storage whih integrating Haproxy and Keepalived. Under the same experimental environment, firstly, we must ensure the entire ceph storage cluster is healthy,so we test the ceph storage cluster . Then, ceph object storage cluster must include the function of manipule and manage data. Last but not the least , through analysising a mount of network pressure tests,whick can prove this system can achieve a high availability scheme,which is based on ceph object storage cluster.

Key words: ceph;high availability;keystone;object storage;haproxy

1 引言

近年来,随着大数据应用的爆发性增长和网络应用的快速普及,网络数据呈海量的增长方式.直接推动了存储、网络以及计算技术的发展.大数据分析应用需求正在影响着数据存储基础设施的发展.

随着结构化数据和非结构化数据量的持续增长,以及分析数据来源的多样化,对存储系统的可靠性、容量、可扩展性、IO性能等方面提出更高的要求[1],当前存储系统的设计已经无法满足大数据应用的需要.

本文以ceph分布式系统为研究对象,通过设计ceph对象存储集群,实现多个对象网关对应用请求数据的存储和管理,提高应用的请求、响应和吞吐能力.通过统一的keyston认证体系,实现ceph对象存储集群的数据安全性和隔离性.整合Haproxy和Keepalive实现基于ceph对象存储集群的高可用性.

收稿日期:2015-04-01;**修回日期**:2015-05-14

2　整体设计框架

为保证基于 ceph 对象存储集群的的高可用性，本系统主要包括：搭建 ceph 集群生态系统、设计 ceph 对象存储集群、Keystone 统一认证系统和 ceph 对象存储集群的高可用设计与实现.

通过实验测试与分析，本系统能实现 ceph 对象存储集群的安全性和隔离性，多对象网关发布对象存储服务、有效的解决 ceph 对象存储集群在大量并行网络应用请求下的高可用性. 图 1 所示为系统整体设计框架.

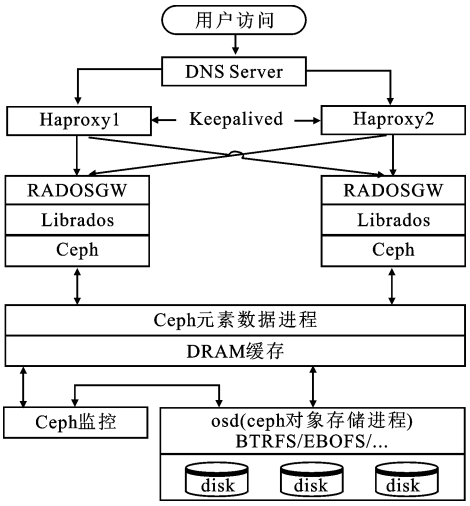


图 1　系统整体设计架构

Ceph 集群生态系统是一个完整的对象存储系统，实现 ceph 对象存储集群中用户数据的存储和管理[2].

Librados 对 RADOS 进行抽象和封装，向上层提供 API，满足基于 RADOS 的应用开发需求.

RADOSGW 在 librados 库的基础上提供抽象层次更高、更便于应用或客户端使用的上层接口. RADOSGW 是一个提供与 Amazon S3 和 Swift 兼容的 RESTful API 的 gateway，通过设计 RADOSGW 来提供 ceph 对象存储的应用开发. 其中包括用户的认证体系、用户的管理、配额管理、对象的操作等[3].

Haproxy1 和 Haproxy2 是整个 ceph 存储集群中设置 RADOSGW 的节点，实现多对象网关发布对象存储服务. 通过 haproxy 和 keepalived 实现基于 ceph 对象存储集群的高可用性.

2.1　集群框架

Ceph 的集群框架包括客户端、元数据服务集群、对象存储集群、集群监视器[4]. 图 2 所示为 ceph 集群框架.

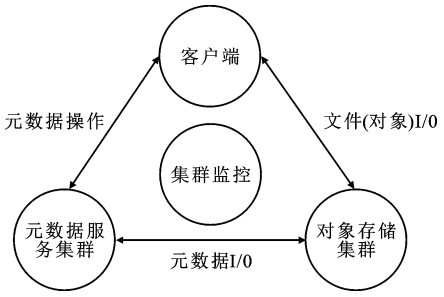


图 2　ceph 集群框架

客户端：数据用户，在大多数文件系统中，所有的控制和智能在内核的文件系统源本身中执行.

集群监视器：执行监视功能，当对象存储设备发生故障或者新设备添加时，监视器就检测和维护一个有效的集群映射.

对象存储设备：ceph 存储节点不仅包括存储，还包括智能. 对象存储设备能作为目标和启动者，支持与其他对象存储设备的通信与合作.

元数据服务器：缓存和同步分布式元数据，元数据服务器将文件名转变为索引节点、文件大小，和 ceph 客户端用于文件 I/O 的分段数据布局.

2.2　对象存储集群设计

在 ceph 对象存储集群健康的状态下，用户通过 DNS Server 节点访问 ceph 对象存储服务，在整个 ceph 对象存储集群中，根据实际生产需求设计 ceph 对象存储集群的 RADOSGW，形成庞大的多区域访问节点，实现用户访问的多区域访问模式，满足实际生产环境下不同用户的交互和操作. Keystone 统一认证实现用户访问的安全性和对存储数据的隔离性. 通过统一认证的用户可以操作和管理 ceph 对象存储集群，包括数据存储、数据管理、数据的整合等服务. 图 3 所示为 ceph 对象存储集群设计框架.

2.3　高可用设计方案

在 ceph 对象存储集群的基础上，设计和实现基于 ceph 对象存储集群的高可用方案. 图 4 所示为高可用性设计框架.

Haproxy1 和 Haproxy2 提供 ceph 对象存储集群中的对象存储服务. 通过设置 keepalived，在 haproxy1 和 haproxy2 节点上不断进行故障检测，保证 ceph 对象存储集群提供正常的云存储服务[5].

在 DNS Server 节点安装和配置 Haproxy，用户可以通过访问 DNS Server 实现对 ceph 对象存储的轮询访问[6]，当网络访问量急剧增长时，能够有效的解决网络的 IO 瓶颈，减轻服务器节点的压力.

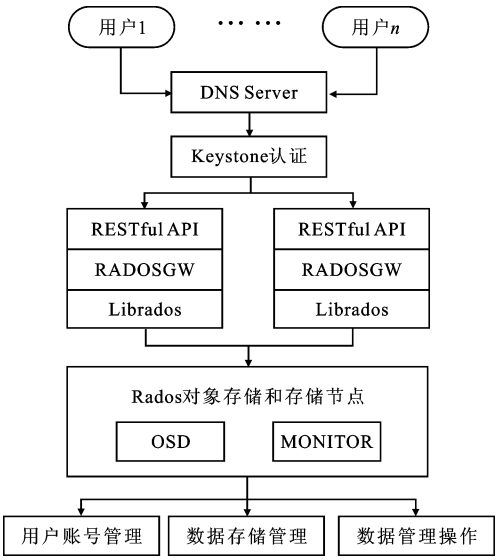


图 3 ceph 对象存储集群设计框架

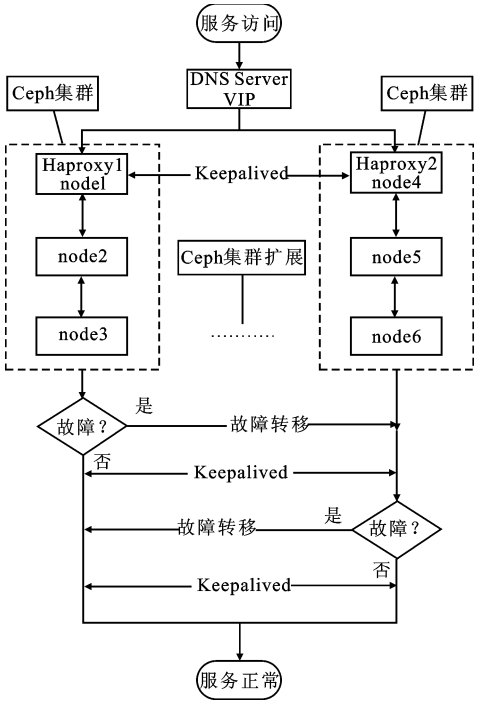


图 4 高可用性设计框架

Keepalived 不断检测 ceph 对象存储集群的运行状态，若 Haproxy1 服务器死机或工作出现故障，则将有故障的服务器重新启动，当 ceph 对象存储集群工作正常后，keepalived 自动将 Haproxy1 服务器加入到 ceph 对象存储集群中，不需要人工干涉. 保证了整个系统的高可用性[7].

3 系统实现和测试结果

3.1 系统实现

ceph 集群框架：通过 ceph-deploy 部署和规划本实验所需的 ceph 集群，检测其健康状态.

ceph 对象存储集群设计：在 ceph 存储集群中，分别选取 Haproxy1 和 Haproxy2 两个节点，作为 ceph 对象存储集群的接口. 设计和部署 ceph 对象存储的 RADOSGW，RADOSGW 是一个提供与 Amazon S3 和 Swift 的 RESTful API 的 gateway，以供相应的对象存储应用开发使用[8].

通过 keystone 统一认证体系，合法用户能够实现对 ceph 对象存储的数据操作和数据管理等功能.

ceph 对象存储集群的高可用设计方案：在 ceph 对象存储集群的基础上，设计和部署多区域数据访问节点，通过 keystone 的统一认证，合法用户才能通过经 haproxy 和 keepalived 设置后的 DNS Server 节点实现对基于 ceph 对象存储集群的数据存储、数据管理和数据整合[9].

3.2 测试过程

3.2.1 云存储测试

根据生产需求设计和部署 ceph 对象存储集群，在 ceph 对象存储集群中分别选取 Haproxy1 和 Haporxy2 作为对象网关节点，对外提供和发布对象存储服务.

在 DNS Server 节点的均衡负载配置文件中添加两个不同区域的监控节点信息. 用户请求服务时，通过数据监控界面可以查看当前的数据操作记录.

ceph 对象存储的云存储服务，主要包括用户账户管理、数据存储管理和数据管理操作.

3.2.2 高可用测试

安装配置 keepalived，不断地检测 Haproxy1 和 Haproxy2 的健康状况，保证了 Haproxy1 和 Haproxy2 提供稳定可靠的网络服务.

当 keepalived 正常运行时，在 Haproxy1 查看 IP，发现 DNS Server 节点 IP. 在 Haproxy2 查看 IP，发现没有 Dns Server 节点 IP，说明 DNS Server 访问了 Haproxy1 节点. 同理，在 Haproxy2 查看 IP，发现 DNS Server 节点 IP. 在 Haproxy1 查看 IP，发现没有 Dns Server 节点 IP，说明 DNS Server 访问了 Haproxy2 节点.

当任意一个 Haproxy 节点失效后，若 Haproxy1 节点坏掉，DNS Server 自动跳到 Haproxy2 节点. 当 Haproxy2 节点坏掉后，DNS Server 自动跳到 Haproxy1.

在 DNS Server 节点上面安装配置 haproxy，为了保证服务器信息的保密性，设置用户名和密码. 监控和检测 Haproxy1 和 Haproxy2 节点，通过 DNS

Server 节点对对象网关节点进行轮询访问，减轻了服务器的访问压力，通过访问 DNS Server 节点的 IP 和端口，实现对 Haproxy1 和 Haproxy2 实时检测和监控. 图 5 所示为负载均衡监控检测界面.

DNS Server	Queue			Session rate			Sessions					Bytes	
	Cur	Max	Limit	Cur	Max	Limit	Cur	Max	Limit	Total	LbTot	In	Out
Haproxy1	6	0	—	0	10		0	0	—	7	7	276	278
Haproxy2	5	0	—	0	8		0	0	—	8	8	150	266

图 5　负载均衡监控检测界面

通过编写和执行健康检测脚本，保证 keepalived 和 haproxy 服务的持续稳定运行. 当 DNS Server 服务停止后，自动化检测脚本在设定的时间后，重新启动 haproxy 服务，对系统进行监控和故障检测，提高整个 ceph 对象存储集群的高可用性.

3.3　测试结果

在相同的实验环境下，使用 webbench 软件对整个 ceph 对象存储集群进行网络压力测试，并行请求数呈 2 的幂次方增长，测试并行数据处理速率和成功请求概率. 表 1 为优化前测试，表 2 为优化后测试.

表 1　优化前测试

测试	并行数	成功数	失败数	速率/(M/s)
优化前测试	256	517 202	0	0.61
	512	512 572	40	0.60
	1 024	503 510	467	0.60
	2 048	496 707	1 016	0.59
	4 096	495 997	2 239	0.59
	8 192	508 968	7 230	0.61

表 2　优化后测试

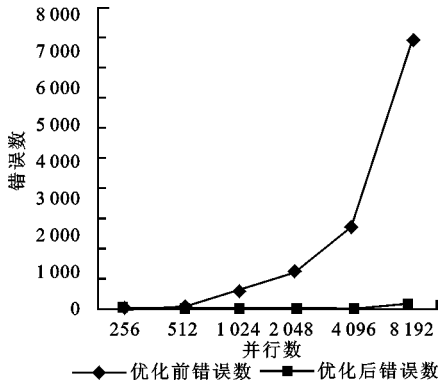
测试	并行数	成功数	失败数	速率/(M/s)
优化后测试	256	1 917 680	0	0.81
	512	1 919 326	0	0.79
	1 024	1 528 596	0	3.32
	2 048	2 046 238	0	4.70
	4 096	2 157 721	65	5.74
	8 192	3 607 091	168	4.58

通过大量的实验测试和分析，并行服务请求失败数和并行服务访问速率成为整个 ceph 对象存储集群的性能测试主要的指标. 表 3 为综合对比测试.

由表 3 对比测试结果显示，在相同实验环境下，优化后用户并行请求数呈 2 的幂次方增长时，错误请求数明显降低，数据访问的速率明显提高了. 如图 6 所示为整体性能对比图.

表 3　综合对比测试

测试对比	优化前		优化后	
并行数	失败数	速率/(M/s)	失败率	速率/(M/s)
256	0	0.61	0	0.81
512	40	0.60	0	0.79
1 024	467	0.60	0	3.32
2 048	1 016	0.59	0	4.70
4 096	2 239	0.59	65	5.74
8 192	7 230	0.61	168	4.58



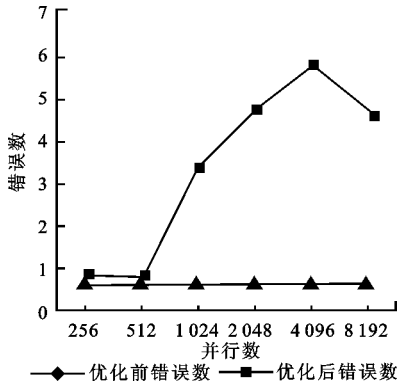


图 6　整体性能对比图

4　结束语

基于 ceph 对象存储集群的云存储应用，用户可以进行多区域对象服务请求，可以实现对用户的数据存储、数据管理、数据整合. 统一的 keystone 认证系统保证用户数据的安全性和隔离性，通过设置软

负载均衡减轻服务器压力，有效提升服务请求的吞吐能力. 自动化的健康检测设计，实现服务器的高可靠性和稳定性.

当并行服务请求急剧增长的情况下，本系统能够有效降低数据并行访问的失败请求数，并行请求错误率平均降低了 38%左右，并行服务平均速率提升近 4 倍. 本系统能够实现整个 ceph 对象存储集群的高可用性.

参考文献：

[1] 李翔，李青山，魏彬. Ceph 分布式文件系统的究及性能测试[J]. 西安电子科技大学，2014，29(5)：1-15.

[2] 符永康. 云存储中数据安全关键技术研究及系统实现[D]. 北京：北京邮电大学，2013.

[3] 蔡官明. 开放式云存储服务平台设计及移动云盘应用开发[D]. 广州：华南理工大学，2013

[4] Weil S A, Brandt S A, Miller E L, et al. Ceph: A scalable, high-performance distributed file system [C]// Proceedings of the 7th Symposium on Operating Systems Design and Implementation, OSDI. USA, Jeattle, 2006：307-320.

[5] Hsiao H C, Chung H Y, Shen H, et al. Load rebalancing for distributed file systems in clouds[J]. Parallel and Distributed Systems, IEEE Transactions on, 2013, 24(5)：951-962.

[6] 冷学健. 基于分开式存储负载均衡的设计与实现[D]. 哈尔滨：哈尔滨工程大学，2012.

[7] 胡利军. Web 集群服务器的负载均衡和性能优化[D]. 北京：北京邮电大学，2010

[8] Wu T, Lee W, Lin Y, et al. Dynamic load balancing mechanism based on cloud storage[C] // Computing, Communications and Applications Conference (ComComAp). HongKong, IEEE, 2012：102-106.

[9] 王芳，陈亮. 对象存储系统中基于负载均衡的设备选择算法[J]. 华中科技大学学报：自然科学版，2007，35(10)：46-49.

作者简介：

杨 飞 男，(1989-)，硕士研究生. 研究方向为云计算与大数据、计算机系统结构. E-mail：yangfeigogo@sina.com.

朱志祥 男，(1959-)，博士，教授. 研究方向为信息安全研究.

梁小江 男，(1983-)，软件工程师. 研究方向为云计算与大数据处理.

(上接第 59 页)

[5] Krishnamurthy V, Hershberger K, Eplett B, et al. SiGe power amplifier ICs for 4G (WIMAX and LTE) mobile and nomadic applications[C]// 2010 IEEE Radio Frequency Integrated Circuits Symposium (RFIC). USA, California, 2010：569 -572.

[6] Kang J, Yoon J, Min K, et al. A highly linear and efficient differential CMOS power amplifier with harmonic control[J]. IEEE J. Solid-State Circuits, 2006, 41(6)：1314-1322.

[7] Grebennikov A. RF and microwave power amplifier design[M]. USA：McGraw-Hill professional engineering, 2005.

[8] 阮颖，陈磊，田亮，等. 基于 0.18μm SiGe BiCMOS 工艺的高线性射频功率放大器[J]. 微电子学，2010，40(4)：469-472.

[9] 胡锦，陶可欣，郝明丽，等. 基于 SiGe 工艺的高增益射频功率放大器[J]. 微电子学与计算机，2012，29(2)：18-21.

[10] François B, Reynaert P. A fully integrated watt-level linear 900-MHz CMOS RF power amplifier for LTE-applications[J]. Microwave Theory and Techniques, IEEE Transactions on, 2012, 60(6)：1878-1885..

[11] Wu R, Li Y, Lopez J, et al. A monolithic 1.85GHz 2-stage power amplifier with envelope tracking for improved linear power and efficiency[C]// 2012 International Symposium on VLSI Design, Automation, and Test(VLSI-DAT). Taiwan, Hualian, 2012：1 -4.

作者简介：

王 巍 男，(1967-)，博士，教授. 研究方向为集成电路设计、半导体器件.

蔡文琪(通讯作者) 女，(1991-)，硕士. 研究方向为微波与射频集成电路. E-mail：747169109@qq.com.

莫 啸 男，(1990-)，硕士. 研究方向为模拟预设 IC 设计.

胡 凤 女，(1991-)，硕士. 研究方向为射频集成电路.