



对象存储系统的自动化管理技术

胡豆豆¹, 罗飞², 张园¹

(1. 中国电信股份有限公司上海研究院, 上海 200122;

2. 华东理工大学, 上海 200237)

摘要: 物联网视频监控领域中的对象存储系统不仅轻松突破了传统存储系统 SAN 的性能瓶颈, 而且可以实现性能与容量的线性扩展; 然而, 对象存储系统资源的远程管理成为困扰管理者的一个问题。针对该问题, 研究能够实现对象存储集群的自动化管理的方法和相关技术, 提高存储集群的使用效率。该自动化管理系统功能主要包括: 集群组件的远程启动、停止、重启; 运行状态监测; 集群扩张和日志查看及存储等功能。基于 Ceph 对象存储集群, 实现了面向 Web 应用服务平台的存储集群自动化管理系统, 最终有效地对对象存储系统进行监视和管理, 并能够自动化扩展集群的节点和存储容量。

关键词: 对象存储; 自动化; 云计算

中图分类号: TP315

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017306

Automatic management technology of object storage systems

HU Doudou¹, LUO Fei², ZHANG Yuan¹

1. Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

2. East China University of Science and Technology, Shanghai 200237, China

Abstract: The object storage system not only breaks the bottleneck of the traditional SAN storage system, but also achieves the expansion of performance and capacity linearly in video monitoring field. However, the remote management for the resources of the object storage system has become a problem for managers. Aiming at this problem, the method and related technology of automatic management of object storage cluster was studied, so as to improve the efficiency of storage cluster. The functions of the automatic management system included remote launching, stopping and resetting of cluster components, monitoring of running status, cluster expansion and log viewing and storage. Based on Ceph object storage cluster, the automated management system for Web application service platform was achieved, monitoring and management was realized effectively ultimately, and the nodes of the cluster and the storage capacity were expanded automatically.

Key words: object storage, automation, cloud computing

1 引言

随着互联网、Web 2.0 的快速发展, Web 应用创建出数百亿的小文件, 上传海量的照片、视频、音乐, Facebook 每天新增数十亿条内容, 每天发送数千亿封邮件。据 IDC 统计, 未来 10 年间数据将增长 44 倍, 2020 年全球数据将增长到 35 ZB, 其中, 80% 为非结构化的数据, 且大部分是非活跃数据^[1-3]。

在物联网视频监控领域中, 随着高清化、智能化的推进以及平安城市等城市公共安防的发展, 面对如此庞大的数据量, 仅具备 PB 级扩展能力的块存储和文件存储显得无能为力; 人们需要一种全新的架构存储系统, 这样的存储系统需要具备极高的可扩展性, 能够满足人们存储视频数据时对存储容量 TB 到 EB 规模的扩展需求^[4,5]。

对象存储是一种基于对象的存储设备, 具备智能、自我管理的能力, 通过 Web 服务协议(如 REST、SOAP)实现对象的读写和对对象资源的访问。对象存储以对象 ID 为基础, 扁平化管理所有存储和桶, 根据对象 ID 便可直接访问数据, 解决 NAS 复杂的目录树结构在海量数据情况下数据查找耗时长的的问题, 这使得对象存储具备极强的扩展性, 能够轻松实现单一名字空间内支持数百万级文件的存储。

对象存储系统的出现满足了数据归档和云服务两大需求。通过面向对象的概念构建的存储系统, 能够有效地利用集群各个节点的性能, 并且有效地将多个客户端请求均匀地分布到集群中的各个节点, 充分地利用了对象存储集群的特性。集群管理员可以通过集群管理系统实时地监视集群各个部分的状态, 也可以根据需要进行调整单个节点上组件的运行状态。当存储集群需要扩张时, 通过自动化的管理系统也能够很快地对集群进行调整以及在各个存储节点之间进行动态的数据迁移, 从而保持服务的持续可用^[6]。

但随着存储资源的堆积和存储集群节点数量的不断上升, 如何通过集群自动化管理系统进行高效的管理和控制成为了一个困扰企业的问题^[7]。本文针对该问题, 研究基于对象存储文件系统 Ceph 的自动化管理系统。在 Ceph 存储集群上, 通过构建一个 Web 项目, 实现对 Ceph 集群的远程监视和控制(包括集群的常用操作、存储节点的增加和删除等); 以此来模拟如何对 Ceph 存储集群进行自动化管理, 从而推进它的商用价值^[8-10]。

2 系统方案

自动化管理的目的是帮助用户解决存储集群在运行时的远程监视、管理和控制的问题, 即管理员通过浏览器随时随地以访问 Web 应用程序的形式来访问对象存储系统, 实现对远程对象存储集群各组件的监视、控制、节点的横向扩展及收缩, 进而改善集群资源的使用性能、提高利用率和降低运营成本。

对象存储系统自动化管理的需求主要可归纳为如下几点。

(1) 监视功能

包括集群当前运行状态监视、存储节点分布和状态监视、单个节点上运行的集群组件监视、集群中已添加节点的分布和运行状态监视。

(2) 组件的常规操作

包括 MON 的重启、MDS 的停止/启动、OSD 的停止/启动等。

(3) 存储节点动态调整

包括 OSD 的添加/删除。

(4) 群日志的查看

包括管理员操作记录查看、集群状态日志查看、存储节点监视日志查看、单节点组件运行状态日志查看、集群节点分布和运行状态的监视日志查看。

系统采用基于 MVC 框架开发的 B/S(浏览器/服务器)模式的 Web 应用程序。用户可通过访



问 Web 浏览器的方式对存储集群进行简单的控制和管理。本文系统的模型拥有较多的处理任务,将集群的监视、操作调用与实现和与数据库中日志信息的交互都放在该层来实现。只要网络连接正常,互联网用户就可以通过打开浏览器随时随地访问应用系统,从而实现对局域网内集群的所有机器的访问并进行远程管理和控制。模块划分为:集群监视模块、存储节点添加/删除模块、组件控制模块、日志存储/调用模块 4 个主要模块,系统功能结构如图 1 所示。

为了提高管理系统的用户体验,采用名为 Bootstrap 的 Web 前端开发框架。Bootstrap 是 Twitter 推出的一个开源的、用于前端开发的工具包。当管理员需要监视、操作或者调整集群时,需要在 Web 页面填写相应的表格,并提交到 Web 服务器进行相关操作。但是和集群相关的操作并不能在很短的时间内完成,例如:集群中各个节点是否都已启动、添加存储节点 OSD 等。因此当提交此类操作时,即使忽视数据在广域网中的传输和建立 SSH 连接的时间,都很难将操作的反馈瞬间呈现在页面上。所以,鉴于上述原因,本文使用了 AJAX (asynchronous JavaScript XML) 技术,实现了页面和 Web 服务器之间的异步通信。

当用户在 Web 页面点击提交事件处理的按钮

时,视图会将该请求的相关信息传递到控制器,控制器对用户请求进行查看并作出判断,执行相应的模型操作,即对应的功能模块会创建一个 SSH 协议,打开 Java 项目与 SSH 服务器的会话,通过该会话可以远程执行命令和 Shell 访问。本文采用了一个名为 Ganymed SSH-2 for Java 的包,专门用于实现 SSH 协议,建立 SSH 服务器连接,并支持 SSH 会话。Ganymed SSH2 for Java 是一个以纯 Java 的形式实现 SSH-2 协议的库(支持 J2SE1.4.2 和 5.0 以上的版本)。它使得在 Java 项目中与 SSH 服务器建立连接,同时支持 SSH 的会话(远程命令执行和 Shell 访问),本地和远程端口交互,本地流转发、X11 转发和 SCP 功能等。

自动化管理系统的后台除了关于 MySQL 数据库的日志操作外,其余的集群控制都是远程调用集群上的脚本实现的。这些脚本均存放在存储集群的主管理节点上,所以当 Ganymed SSH2 for Java 与集群主管理节点建立 SSH 会话后,调用对应的 Shell 脚本即可执行所需要的管理操作。

3 系统实现

系统中 Web 控制页面显示的数据都是通过执行后台脚本,截取到所需的特征数据,并经过整理后呈现出来的。

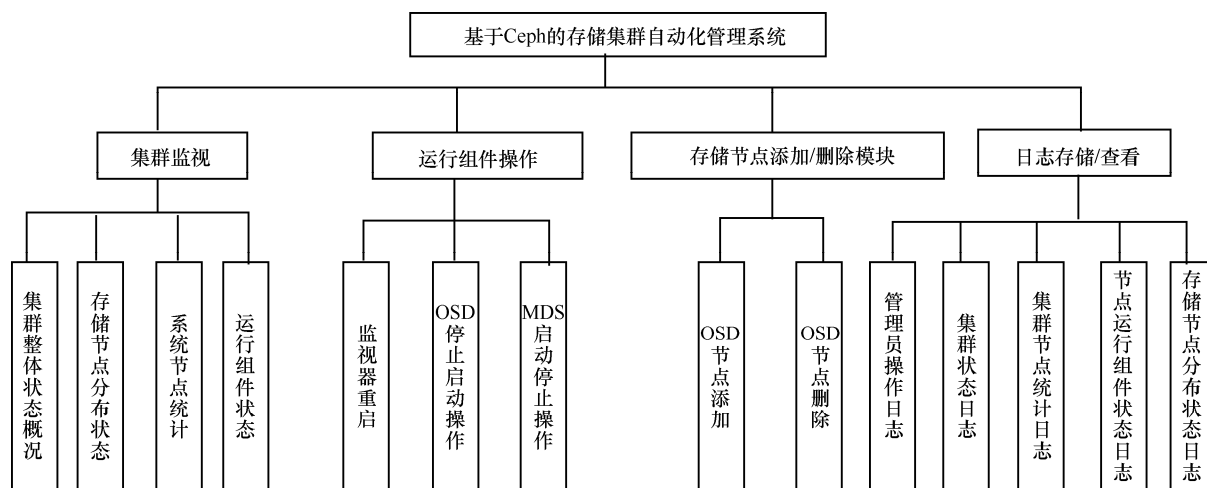


图 1 系统功能结构

3.1 全局状态监控

全局状态监测在终端输入 `ceph -s` 指令时可以查看集群整体的当前状态，快速地查看需要的关键信息，见表 1。首先从指令 `ceph -s` 输出集群当前的通用状态信息，进而通过 `grep` 和 `sed` 的筛选和替换功能实现对集群特征数据的采集。

表 1 集群整体状态

信息类别	总体信息	当前信息	状态信息
MON (监控节点)	MON 个数	主 MON 节点的 IP 地址	MON 状态列表
MDS (元数据服务器)	MDS 个数	MDS 所在节点 IP 地址	MDS 状态列表
OSD(对象存储节点)	OSD 个数	可访问 OSD 个数	OSD 状态列表
存储容量/GB	存储容量总和	可使用容量	已使用容量

3.2 存储节点分布状态监测

Ceph 是一个高可用性和高可靠性的管理硬件和软件。它不具备单点故障，并且可以为数据提供还原服务。假如系统出现故障，需要到放置组和相关的守护进程下查找问题。当 Ceph 集群遇到一些故障时，还可能阻止用户访问特定的对象，但可能访问其他对象。当运行时遇到故障，只需要按照步骤监测守护进程和放置组就能知道故障原因，然后排除。

由图 2 可知，在集群中一个 OSD 的状态或者在这个集群中 (in) 或者在这个集群外 (out) 的；并且，它或者是运行中 (up) 的，或者是不运行 (down) 的。如果一个 OSD 的状态是 up，它或者在集群 (可以读取和写入数据) 或者在 out 状态。如果它曾是 in 状态但最近被移出集群了，Ceph 将会迁移配置至其他 OSD 中。如果 OSD 不在集群中，基于 Ceph 的 CRUSH 算法，该 OSD 将不会配置在一个放置组中。如果 OSD 的状态是 down，它也会是 out 状态。

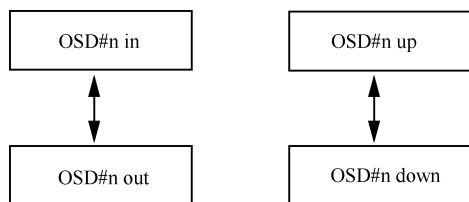


图 2 OSD 状态转换

执行 `Ceph Osd tree` 指令能够确定 `ceph-osd` 后台是否正常运行，根据控制台的输出，可以很容易地从中获取所需要的特征数据，并且有效地组织在一起，作为日志存储到日志并刷新页面中的对应部分视图。

3.3 集群节点分布监测

由于集群中的节点会逐渐增多，所以需要在 Web 页面上直观表示节点名称和 IP 地址，形式见表 2。

表 2 集群节点配置

IP 地址	节点 Hostname
192.168.232.158	server01
192.168.232.160	server02
192.168.232.170	server04

为了获取集群中加入的节点 IP 地址和相关的 Hostname，前提必须是将 Hostname 和 IP 地址写入系统文件 `/etc/hosts` 中。

同时，管理系统能够自动化管理存储节点 OSD 的添加/删除操作。集群上放置的每一个对象 (object) 都可以根据 Ceph 的 CRUSH 算法计算出具体的存放位置。所以，当 OSD 发生变化时，集群也需要做相应的调整。CRUSH 算法是 Ceph 对象存储的核心，决定了每一个对象的存放位置。当需要存放一个对象时，只需要知道对象的 ID、集群架构图和存放策略，就可以计算出应该存放在哪个 OSD 上。因此 Ceph 的元数据不需要包含位置信息，每个客户端都可以自己计算出对应的 OSD，大大减少了对元数据的请求。

RADOS 是使用 CRUSH 算法的分布式对象存储，由 OSD 和 MON 组成，OSD 承担数据的存储、备份以及恢复工作，MON 保证集群架构图与实际情况相符。RADOS 的特点之一就是数据的备份、同步以及恢复等都由 OSD 自发地完成。

RADOS 中 OSD 独自管理数据的能力是建立在 CRUSH 算法上的，CRUSH 算法又是建立在 CRUSH map (包含 OSD 的数量、组织结构以及存储策略)



上的。因此拥有一份最新的 CRUSH map，对 OSD 是至关重要的，否则 OSD 无法确定隶属同一个 PG 的 OSD，也就无法完成数据的备份、同步等操作。

RADOS 使用一个由 MON 组成的小集群生成最新的 CRUSH map。当增/减 OSD、修改 OSD 组织结构、修改存放策略或者发现 OSD 失效时，更新请求被发送到 MON 集群，MON 生成一份新的 CRUSH map。同时 MON 集群也对外提供 CRUSH map 的访问服务，其他节点可以从 MON 集群中获取一份最新的 CRUSH map。MON 集群中采用 Lease 机制确保每个 MON 对外提供的 CRUSH map 的一致性，采用 Paxos 协议选举出一个 leader，由 leader 负责生成新的 CRUSH map。

另外，OSD 节点之间互相通信时会比较各自拥有的 CRUSH map 的 epoch，用最新的 CRUSH map 替换旧的 CRUSH map。当客户端要从 OSD 中读取对象时，如果客户端中还没有 CRUSH map：客户端首先从 MON 集群中读取一份最新的 CRUSH map，然后计算出存放该对象的 PG 对应的 OSD，然后客户端与目标 OSD 比较各自的 CRUSH map 的 epoch，同步到最新的 CRUSH map。如果客户端中已经有 CRUSH map，在和 OSD 通信时，如果 OSD 的 CRUSH map 版本更新，那么更新客户端的 CRUSH map。如果客户端和 OSD 的 CRUSH map 版本一致了，但都不是最新的，使用旧的 CRUSH map 计算出的 OSD 上可能没有存放要访问的对象，导致操作失败，这时客户端从 mon 集群读取最新的 CRUSH map。

3.3.1 OSD 添加

新增一个 OSD 需要分为两个部分：准备（prepare）和激活（activate）。准备步骤的指令为：`ceph-deploy osd prepare ceph1: /dev/vdb`，步骤如下所示：

```
push 配置文件到 ceph1:/etc/ceph/ceph.conf
push ceph.bootstrap-osd.keyring 到 ceph1:/var/
lib/ceph/bootstrap-osd/ceph.keyring
```

```
udevadm trigger --subsystem-match=block
--action=add
```

```
ceph-disk-prepare --fs-type xfs --cluster ceph --
/dev/vdb //创建分区、格式化等操作
```

激活步骤的指令为：`ceph-deploy osd active ceph1:/dev/vdb`，步骤如下所示：

```
ceph-disk-activate --mark-init upstart --mount
/dev/vdb //挂载 Ceph 分区
```

```
initctl emit --no-wait -- ceph-osd cluster=ceph
id=0 //添加 ID 为 0 的 OSD 守护进程，并使其处于
非等待状态，即添加完该 osd.0 后直接启动。
```

3.3.2 OSD 删除

删除 OSD 守护进程时需要根据该 OSD 的不同状态执行一些列的指令，从而达到有效的集群 CRUSH 调整，并且完成该 OSD 上存储对象的完整迁移，最后完全将该 OSD 守护进程从集群中移除。步骤如下所示：

```
service ceph stop osd.0 //停止该设备
```

```
ceph osd crush remove osd.0 //从 CRUSH 中移
除与该 OSD 相关的数据，此时运行 ceph osd tree
指令可以看到该存储进程 osd.0 处于 down 状态
```

```
ceph auth del osd.0 //从认证中删除
```

```
ceph osd rm 0 //从集群中完全移除该 OSD 的
所有数据
```

4 系统测试

为了完成系统测试，本文在 Ubuntu12.04 的虚拟机上搭建基于 Ceph 的初始集群。该集群初始化规模为 3 个节点，并通过 `ceph-deploy` 完成 Ceph 的安装。首先将 Ceph 资料库添加到 `ceph-deploy` 的管理节点，然后安装 `ceph-deploy`。注意：以不同的用户登录上节点安装 `ceph-deploy` 时不需要在命令前面添加 `sudo` 或者以 `root` 身份执行，因为远程的主机将不会执行 `sudo` 命令。测试过程包括监视功能测试、存储节点添加/删除功能测试等。

4.1 监视功能测试

监视功能所显示当前集群的状态如图 3 所示, 包括: mon 节点的个数、所在节点、运行状态; mds 节点的个数、所在节点、运行状态; OSD 节点的个数、运行状态、存储容量等信息; 其中, OSD 存储容量已用部分 13 837 MB, 剩余可用部分为 4 210 MB, 总共的存储容量为 19 021 MB。

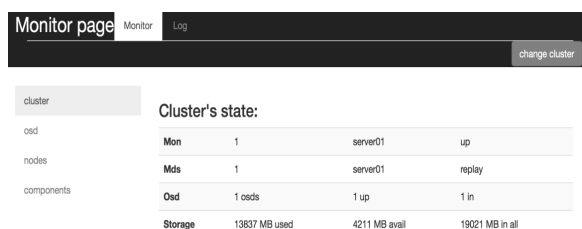


图 3 集群状态监视

集群中各存储节点状态如图 4 所示, 显示了集群中存储节点 OSD 的分布情况: ID 为 0 的 OSD 节点的权重为 0.019 99 (1 TB 大小的容量权重为 1, 所以 0.019 9 权重说明 ID 为 0 的 OSD 存储容量大约为 20 GB 左右) 以及状态为 up。



图 4 集群存储节点分布

集群内节点状态统计如图 5 所示, 这里显示的节点是已经被添加到集群的, 包括节点 IP 地址、节点名称以及该节点是否在运行中。从图 5 可以看到, 目前只有节点 IP 地址为 192.168.232.158、节点名为 server01 的节点 power 栏显示的是 up, 即在运行中。

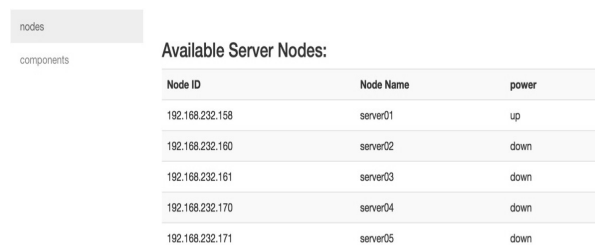


图 5 集群内节点状态统计

4.2 存储节点添加/删除功能测试

在 server02 上添加一个 OSD 守护进程, 操作和结果分别如图 6 和图 7 所示。

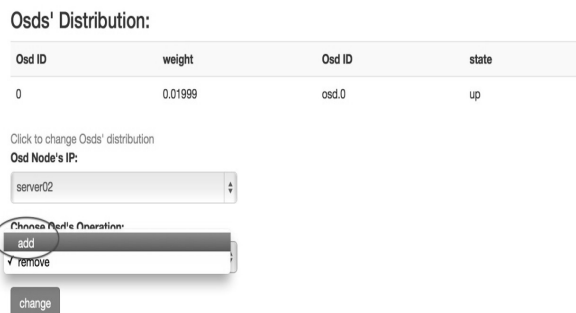


图 6 添加 OSD.1 操作



图 7 OSD.1 添加结果

将主管理节点 server01 上面的 OSD 守护进程 osd.0 删除, 操作和结果分别如图 8 和图 9 所示。

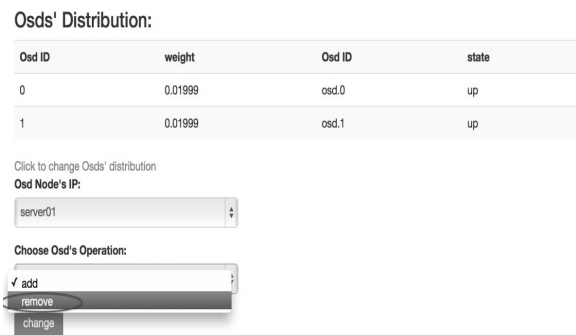


图 8 删除 OSD.0 操作示意

4.3 日志查看功能测试

查看某个时间段的存储节点分布情况, 结果如图 10 所示。



Cluster's state:

Mon	1	server01	up
Mds	1	server01	replay
Osd	1 osds	1 up	1 in
Storage	12511 MB used	5537 MB avail	19021 MB in all

Osds' Distribution:

Osd ID	weight	Osd ID	state
1	0.01999	osd.1	up

图9 删除 OSD.0 的结果

Osd Tree Distribution:

Osd ID	weight	Osd Name	Osd State	Time
0	0.01999	osd.0	up	2014-05-28 07:35:25
0	0.01999	osd.0	up	2014-05-28 07:35:43
0	0.01999	osd.0	up	2014-05-28 07:36:02
1	0	osd.1	down	2014-05-28 07:36:02
0	0.01999	osd.0	up	2014-05-28 07:36:21
1	0.01999	osd.1	up	2014-05-28 07:36:21
1	0.01999	osd.1	up	2014-05-28 07:36:33
0	0.01999	osd.0	down	2014-05-28 07:36:33
1	0.01999	osd.1	up	2014-05-28 07:36:51

Time From: 2014-05-28 07:35 To: 2014-05-28 23:55 Get Logs!

Cluster Nodes:

Node ID	Name	power
---------	------	-------

图10 某时间日志查看结果

5 结束语

基于对象存储系统 Ceph, 研究了一个对象存储集群的自动化管理系统。管理系统的功能主要集中在 3 个部分——集群状态监视、OSD 存储节点添加与删除和集群状态日志查看。以平台化方式设计, 采用基于 MVC 框架进行设计和实现。用户可通过访问 Web 浏览器的方式对存储集群进行简单的控制和管理。

作为底层存储集群的管理系统, 这些功能都是必备的, 但还是远远不够的。要真正实现自动化的存储集群管理系统, 故障排除与恢复和负载均衡优化是不可缺失的主要功能。而本系统中仅有的一些功能做得也不是非常完善, 这些工作都有待在未来进行优化。

参考文献:

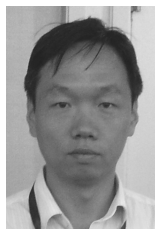
- [1] 郭立, 金海, 韩宗芬. 集群系统的单一系统映像研究[D]. 武汉: 华中科技大学, 2003.
GUO L, JIN H, HAN Z F. Research on single system image for cluster[D]. Wuhan: Huazhong University of Science and Technology, 2003.
- [2] 孙乔, 邓卜侨, 王志强, 等. 一种基于分布式服务器集群的可扩展负载均衡策略技术[J]. 电信科学, 2017, 33(9): 190-196.
SUN Q, DENG B Q, WANG Z Q, et al. A scalable load balancing strategy based on distributed server cluster[J]. Telecommunications Science, 2017, 33(9): 190-196.
- [3] 郭延辉, 陈征, 龚松春. 一个集群文件系统的虚拟化支持平台研究与实现[J]. 电信科学, 2012, 28(9): 34-37.
WU Y H, CHEN Z, GONG S C. Research and implementation of virtual platform on cluster file system[J]. Telecommunications Science, 2012, 28(9): 34-37.
- [4] 兰超, 张勇, 邢春晓. 面向复杂数据的对象存储系统[J]. 计算机科学与探索, 2014, 8(2): 128.
LAN C, ZHANG Y, XING C X. Elastic object store system for complex data[J]. Journal of Frontiers of Computer Science & Technology, 2014, 8(2): 128.
- [5] 蒋多元, 陈海雄. 云数据中心集群间网络性能优化的探讨[J]. 电信科学, 2015, 31(5): 138-142.
JIANG D Y, CHEN H X. Discussion of inter-cluster network performance optimization in cloud data center[J]. Telecommunications Science, 2015, 31(5): 138-142.
- [6] WEIL S A, POLLACK K T, BRANDT S A, et al. Dynamic metadata management for petabyte-scale file systems[C]//2004 ACM/IEEE Conference on Supercomputing (SC'04), November 6-12, 2004, Tucson, AZ, USA. New York: ACM Press, 2004.
- [7] XIN Q, MILLER E L, SCHWARZ T J E. Evaluation of distributed recovery in large-scale storage systems[C]//IEEE International Symposium on High Performance Distributed Computing(HPDC), June 4-6, 2004, Honolulu, Hawaii, USA. New Jersey: IEEE Press, 2004.
- [8] 藤本和久, 细谷睦, 岛田健太郎, 等. 集群型存储系统及其管理方法: CN200410045333.2[P]. 2004-05-20.
FUJIMOTO W, SAGO M, SHIMADA K, et al. Cluster storage system and its management method: CN200410045333.2[P]. 2004-05-20.
- [9] 胡学庆. 集群环境下海量数据存储管理技术的研究[D]. 大庆: 东北石油大学, 2010.
HU X Q. Research on massive data storage management technology in cluster environment[D]. Daqing: Northeast Petroleum University, 2010.
- [10] 黄昌华, 康慕宁, 韩朴杰. Hyper_v 虚拟机集群管理系统的设计与实现[J]. 计算机与现代化, 2013(4): 166-170.

HUANG C H, KANG M N, HAN P J. Design and implementation of integrated management system for cluster of Hyper_v virtualMachine[J]. Computer and Modernization, 2013(4): 166-170.

[作者简介]



胡豆豆（1985-），男，现就职于中国电信股份有限公司上海研究院，主要从事视频监控系统智能、云存储方向产品开发及研究工作。参与多项工业和信息化部物联网专项、“863”计划等科研项目工作，累计发表专利4项，发布 ITU-T 视频监控类国际标准2项，2013年获得上海市科学技术进步奖二等奖。



罗飞（1978-），男，博士，华东理工大学副教授、硕士生导师，主要从事嵌入式系统、云计算、大数据等方面的开发与研究工作。曾获得2012年上海市科学技术进步奖一等奖、2013年上海产学研合作优秀项目奖。



张园（1986-），女，现就职于中国电信股份有限公司上海研究院，主要从事视频监控系统智能、云存储方向产品开发及研究工作。参与多项工业与信息化部物联网专项、“863”计划等科研项目工作，累计发表专利3项，发布 ITU-T 视频监控类国际标准2项。