# 面向云计算的集群文件系统性能对比测试①

刘建平[1], 张 辉[2], 于铁忠[3], 吴 恒[4]

[1](解放军第 401 医院信息科, 青岛 266071)
[2](山东乾云启创信息科技有限公司, 济南 250101)
[3](石家庄学院 计算科学和工程系, 石家庄 050035)
[4](中国科学院 软件研究所, 北京 100190)

**摘 要**: 虚拟机使用集群文件系统是云计算主要架构模式之一, 主要分为共享式集群文件系统和分布式集群文件系统两大类. 其中, 后者是近年来学术界关注的热点, 在产业界也逐渐得到深入应用. 采用测试法, 选取 fio 作为测试工具, 以IOPS作为度量指标, 对比两类集群文件系统的性能优劣. 实验结果显示, 在单物理机节点配置下进行单台虚拟机 IOPS 写性能测试, 共享式文件系统要好于分布式集群文件系统大约 40%, 但读性能要差 1.5 倍; 对于多台虚拟机并发测试, 则当支持虚拟机运行的共享式/分布式集群文件系统规模小于 16 台时, 前者和后者性能基本一致. 否则, 后者会明显好于前者.

**关键词**: 集群文件系统; IOPS; 性能

## Performance Comparison Between Two Kinds of Clustered File System in Cloud Computing

LIU Jian-Ping[1], ZHANG Hui[2], YU Tie-Zhong[3], WU Heng[4]

[1](The 401 hosptial of PLA, Qingdao 266071, China)
[2](Shandong Massclouds CO., LTD, Jinan 250101, China)
[3](Computer Science Engineering, Shijiazhuang University, Shijiazhuang 050035, China)
[4](Institue of Software, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: Clustered file system is a key component of cloud computing. It can be divided into two major groups, which are shared-disk file system and distributed file system, in which the latter is the hot spot of academic research in recent years. This paper mainly focuses on IOPS, which is an important performance indicator for VM(virtual machine). We employ fio as the testing tool and experimental results show that, (1) when a VM runs on sharing-disk file system and distributed cluster file system both with 3 storage nodes, the write IOPS of the former is higher than the latter about 40%, but the read IOPS of latter is 1.5 times as much as the former; (2) when multiple VMs run simultaneously on sharing-disk file system and distributed cluster file system both with less than 16 storage nodes, the write IOPS of VMs for both clusters are almost the same; otherwise, the write IOPS performance of the latter is better than the former.

**Key words**: clustered file system; IOPS; performance

虚拟化是构建云计算的主要基础设施之一, 它通过对物理资源的抽象和封装, 实现物理资源的分时复用, 具有同时运行多台虚拟机, 提高物理资源平均利用率, 降低了IT部门成本的优势[1]. 如图1所示, 虚拟机作为信息化系统和物理资源的中间层, 通常采用文件(raw、qcow2 等) 进行封装 (即虚拟机镜像), 运行时虚拟机产生的持久化数据, 也存放在虚拟机镜像中[1]. 其中, 集群文件系统是当前主流的虚拟机镜像存储方案, 其实现机制会影响虚拟机硬盘性能, 从而间接影响信息化系统(如 ebay 等)性能, 比如硬盘读写会直接影响数据库的访问延迟. 为表述方便, 本文采用 IOPS (Input/Output Operations Per Second) 作为虚拟机硬盘性能的度量指标.

① 收稿时间:2015-05-26;收到修改稿时间:2015-07-02

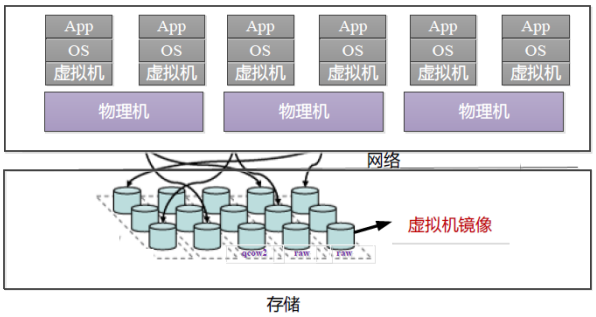


图 1 虚拟机管理架构

当前, 集群文件系统可根据实现机制的不同, 细分为共享式集群文件系统和分布式集群文件系统两大类. 其中, 共享式集群文件系统主要包括 VMWare VMFS[3], IBM SmartCloud GPFS[4], Oracle VM OCFS2[5] 等, 其对外表现为具体文件格式, 类似于 FAT32, NTFS、ext4. 它通常需要和磁盘阵列协同工作, 是当前产业界用于虚拟机镜像存储的常用解决方案. 分布式集群文件系统主要包括 Ceph[6]、Glusterfs[7]、Moosefs 等, 安装在多台普通服务器上, 协同工作对外提供满足 FUSE 协议的存储服务, 它是当前学术界关注的重点, 也是产业界发展的重要趋势之一[2]. 比如, 我国著名的公有云 Ustack 使用 Ceph 作为虚拟机存储方案.

已有研究工作[8,9]面向小文件读写场景, 对共享式和分布式集群文件系统的可靠性和性能进行了对比测试. 但这些工作并未考虑到虚拟机场景具有大文件(一个虚拟机镜像通常为 10GB 左右), 并发大(比如阿里云有 40 万台虚拟机[10])的特点. 因此, 共享式和分布式集群文件系统分别适用于哪种规模的虚拟机场景, 尚属需探索的问题.

本文选取 fio 作为测试工具, 面向虚拟机场景, 对比共享式和分布式集群文件系统的 IOPS, 其性能结果可作为云计算实施的参考. 具体而言, 论文选取典型共享式集群文件系统 OCFS2 和典型分布式集群文件系统 Ceph 作为虚拟机镜像存储和运行支撑载体, 其原因是: (1)OCFS2 和 Ceph 是开源可免费获取的; (2)OCFS2 是商用虚拟化厂商 Oracle VM 和 SUSE 推荐的存储方案之一; 而 Ceph 是商用虚拟化厂商 Ustack 的存储方案, 具有代表性. 在单物理机节点配置下进行单台虚拟机 IOPS 写性能测试, 共享式文件系统的写性能要好于分布式集群文件系统大近 40%, 但读性能要差 1.5 倍; 而对于多台虚拟机并发测试, 则当支撑虚拟机运行的物理机规模小于 16 台时, 前者和后者性能基本一致. 否则, 随着物理机规模和虚拟机个数的增加, 后者会明显好于后者.

## 1 分布式文件系统简介

### 1.1 OCFS2

如图 2 所示, OCFS2 是一种典型的共享式集群文件系统, 可支持 IP 和光纤交互两种协议, 采用分布式锁机制实现多台安装 Xen/KVM 的物理服务器对同一硬盘资源读写访问的全局视图, 防止了写冲突, 因此又称为“Shared disk file systems”. 从性能视角来说, OCFS2 采用文件访问的数据流与控制流分离策略, 具有高吞吐数据访问能力. 从可靠性视角来说, OCFS2 主要依靠磁盘阵列的 RAID 机制实现虚拟机文件的冗余保护, 确保虚拟机镜像文件的完整性.

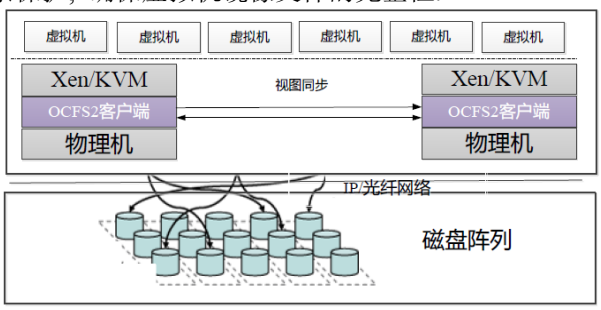


图 2 OCFS2 存储方案

### 1.2 Ceph

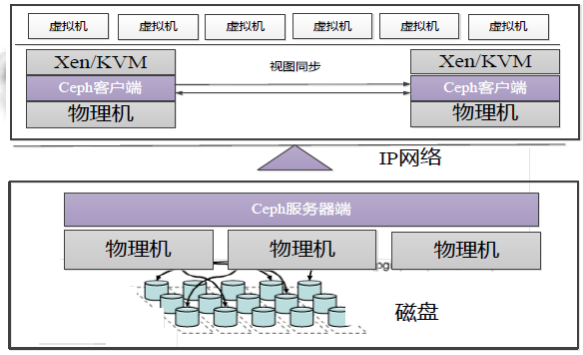


图 3 Ceph 存储方案

如图 3 所示, Ceph 只支持 IP 交换协议, 需要安装在多台普通服务器上, 使其协同工作对外提供存储服务, 它是磁盘阵列的一种可替代方案[6]. Weil[6]认为 Ceph 本质是基于对象存储(Object Storage Device, OSD)的分布式集群文件系统, 通过建立文件到对象, 块到

对象的映射为虚拟机提供存储和运行支撑服务. 从性能视角来说, Ceph 采用元数据管理数据的管理方式, 通过高效 Hash 实现大大减少了元数据产生量, 降低了数据定位的开销, 从而达到提高硬盘读写性能的目标. 从可靠性视角来说, Ceph 本身具有灵活副本管理机制, 可以通过参数配置实现虚拟机文件的冗余, 提高可靠性.

## 2 集群文件系统性能对比

### 2.1 实验目标

对比不同规模下, 虚拟机使用共享式和分布式集群文件系统的IOPS性能, 回答两种集群文件系统分别适用于哪种规模的虚拟机场景. 实验主要包括以下 2 个内容: (1)单台虚拟机使用集群文件系统的硬盘 IOPS 性能及 CPU 资源开销; (2)不同虚拟机规模下, 集群文件系统峰值 IOPS 能力.

### 2.2 实验环境

如图 4 所示, 整个实验包括虚拟化服务器, Ceph 服务器, 万兆交换机和磁盘阵列四个部分组成:

(1)磁盘阵列映射出 1 块 512GB 硬盘, 作为 OCFS2 存储;

(2)Ceph 服务器 3 台, 每台配置 512GB 大小的 SSD 硬盘, 型号与磁盘阵列中一致, 配置三副本;

(3)虚拟机硬盘使用 rbd 协议创建.

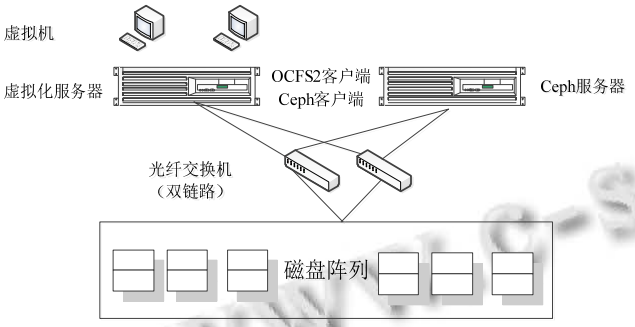


图 4 实验架构

其中, 具体软硬件配置如表 1 所示.

表 1 软硬件配置

| 名称 | 配置 |
|---|---|
| 磁盘阵列 | 磁盘阵列模拟 iSCSI, 512GB 硬盘 1 块, SSD |
| 虚拟化服务器 | 24 台, CPU 2.6GHz, 24 cores, 内存 32GB |
| Ceph 服务器 | 3 台, 每台配置 512GB 硬盘 1 块, SSD, CPU 2.6GHz, 8 cores, 内存 8GB |
| Ceph 软件 | Ceph 0.87 |
| 虚拟化软件 | Xen 4.4.1 |
| 虚拟机配置 | 1 vCPU, 1G 内存, OS 为 CentOS 6.5 |

### 2.3 测试工具和脚本

fio 是一个文件系统的 benchmark 工具, 它可以实例化多线程或进程模拟各种 io 操作, 可以测试不同的操作系统中文件系统的读写性能. 其可配置参数:

① ioengine: 负载引擎, 设置成 libaio, 发起异步 IO 请求.

②bs: IO 块大小, 本文设置为 4k, 4k 是内存 page 页的默认大小

③direct: 设置为 1, 表示绕过操作系统 Cache;

④ type: 包括顺序写 write、顺序读 read、随机写 randwrite、随机读 randrea 等;

⑤ size: 寻址空间, IO 会落在 [0, size)这个区间的硬盘空间上, 本文设置成 100GB, 避免虚拟机 OS 缓存对测试结果的影响;

⑥ filename: 测试对对象;

⑦ iodepth: 队列深度, 设置成 32;

⑧runtime: 测试时长, 设置成 600 秒;

本文选取 read, write, randrea, randwritewrite 四种模式作为对比分析, 可分别模拟视频上传和播放(read, write 模式, 虚拟机文件内容填充适用于该场景)和 NoSQL 读写场景(re-read, re-write, 虚拟机文件内容修改适用于该场景), 具有代表性.

## 3 结果分析

### 3.1 单台虚拟机运行场景

如图 4 所示, 在虚拟化服务器上部署一台虚拟机, 分别运行在OCFS2上Ceph上, 测试配置如小节2.3所示.

如图 5-图 8 所示, 虚拟机运行在 OCFS2 上, 其硬盘顺序写、随机写性能会快于虚拟机运行在 Ceph 文件系统上的能的 40%左右, 但顺序读、随机读性能要慢大约 1.5 倍. 这是因为 Ceph 存储接口(FileStore)为了支持事务, 引入了日志(Journal)机制. 所有写操作都需要先写入日志, 然后再写入文件系统, 因此实际磁盘输出约为其物理性能的一半. 而对于读操作, 得益于 Ceph 的三副本设置, 可以并发读入, 因此理论上可快大约 2 倍.

如图 9-图 12 所示, 物理机安装 OCFS2 客户端的 CPU 开销相对于物理机安装 Ceph 客户端, 会明显降低 15%左右, 这是因为 OCFS2 的客户端基本运行在内核态, 而 Ceph 客户端基本运行在用户态. 另外, Ceph 服务器端还需要消耗大约 20%的 CPU 开销.

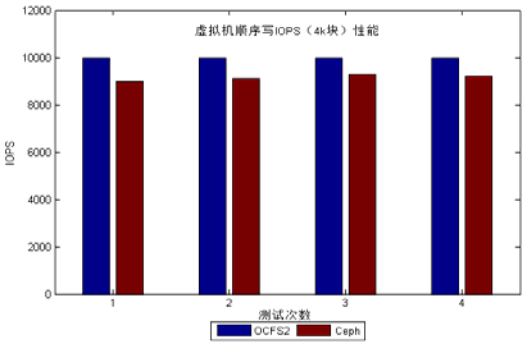


图 5 单台虚拟机 4k 块 IOPS 顺序写性能对比

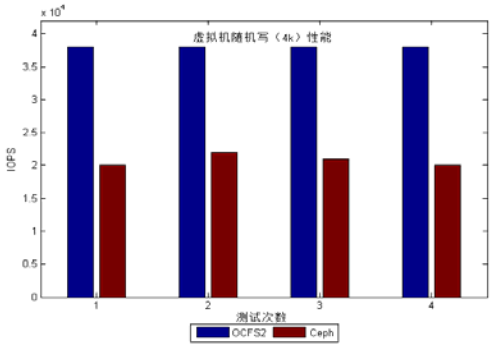


图 6 单台虚拟机 4k 块 IOPS 随机写性能对比

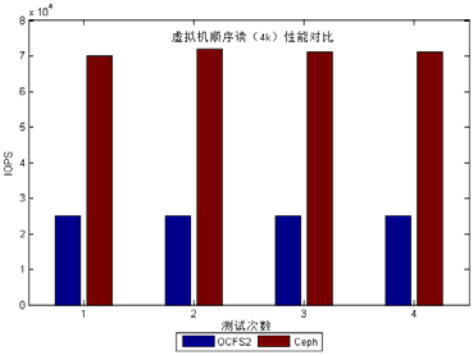


图 7 单台虚拟机 4k 块 IOPS 顺序读性能对比

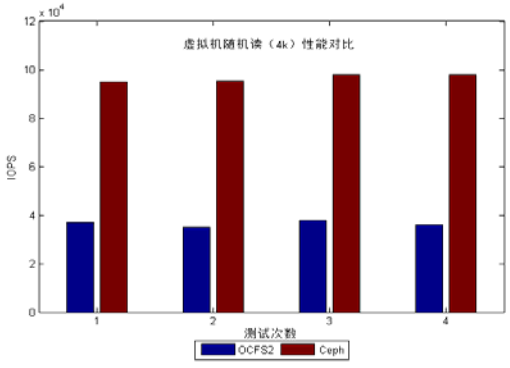


图 8 单台虚拟机 4k 块 IOPS 随机读性能对比

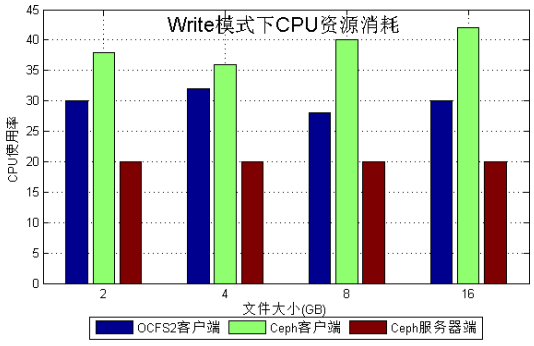


图 9 顺序写模式下 CPU 开销对比

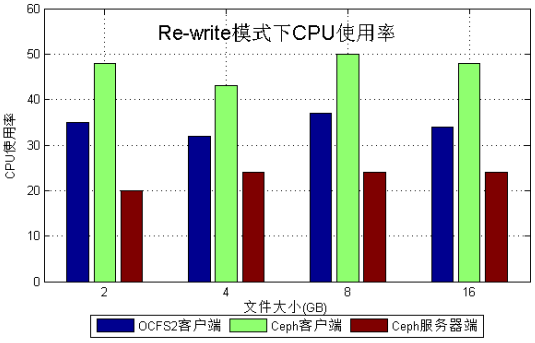


图 10 随机写模式下 CPU 开销对比

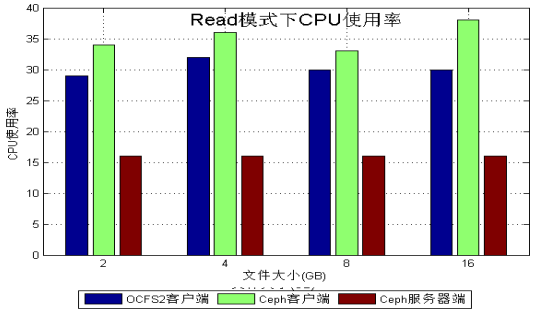


图 11 顺序读模式下 CPU 开销对比

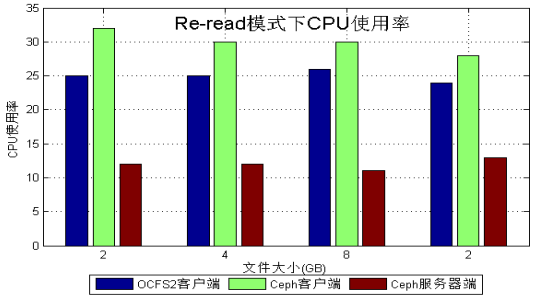


图 12 随机读模式下 CPU 开销对比

### 3.2 多台虚拟机运行场景

如图 4 所示, 在虚拟化服务器上同时部署 16 台虚拟机, 每台物理机上部署一台虚拟机, 分别运行在 OCFS2 上 Ceph 上, 运行小节 2.3 的测试配置. 假设每台虚拟机写 IOPS 限速为 300IOPS, 读为 500IOPS, 则 16 台虚拟机并发的总 IOPS, 会小于 SSD 供给能力, 分别在虚拟化服务器规模是 8 和 24 两种场景下, 进行测试总体 IOPS 读写输出结果.

如图 13-图 16 所示, 当集群规模小于 16, 即物理服务器规模为 8 时, 随着虚拟机个数增加(16 到 32 台), Ceph 和 OCFS2 的读写数据基本一致, 但当集群规模大于 16, 为 24 节点时候. Ceph 比 OCFS2 性能要好, 这是因为 OCFS2 设计的物理集群规模为 16, 超过 16 个节点时, 其内部通信开销增加, 导致性能下降.

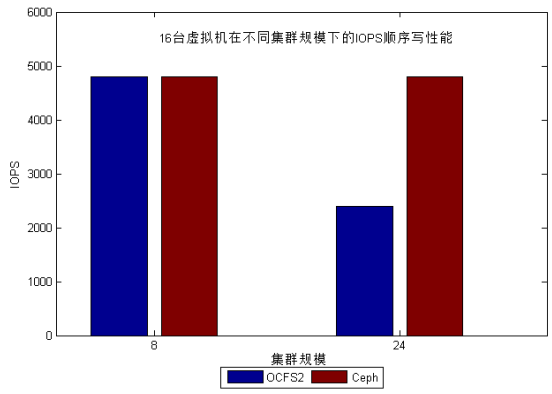


图 13　多台虚拟机顺序写性能对比

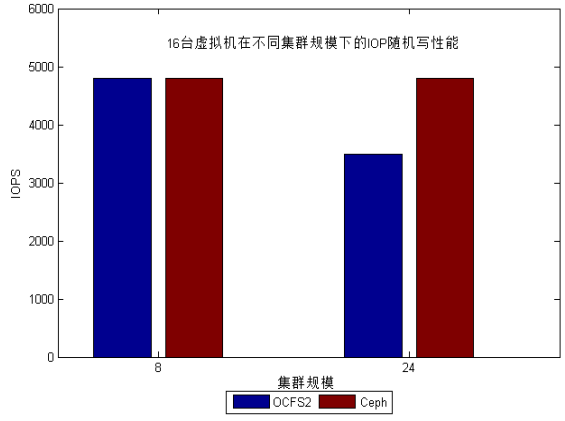


图 14　多台虚拟机随机写性能对比

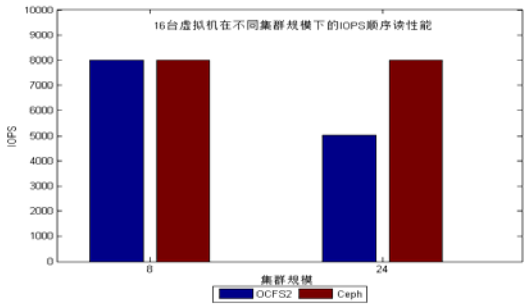


图 15　多台虚拟机顺序读性能对比

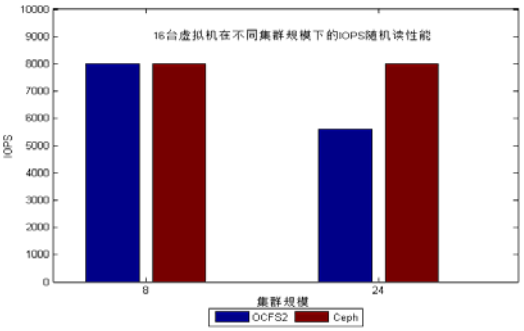


图 16　多台虚拟机顺序写性能对比

## 4　结语

本文主要关注使用者视角, 对两种典型的文件系统实现进行了性能分析. 并不是从分布式文件系统的角度加以阐述的, 因此没有进行分布式文件系统可靠性, 可扩展性测试. 该方面测试也可作为未来工作的计划.

在相同硬件配置下的, 单台虚拟机运行在共享式文件系统的写性能要好于运行在分布式集群文件系统大近 40%, 但读性能要差 1.5 倍; 而对于多台虚拟机并发测试, 则当支撑虚拟机运行的物理机规模小于 16 台时, 前者和后者性能基本一致. 否则, 随着物理机规模和虚拟机个数的增加, 后者会明显好于后者.

这是因为 Ceph 等用户态分布式文件系统是一种可完全替代磁盘阵列的解决方案, 多用于百度等互联网企业, 通常使用普通物理服务器进行, 由于普通服务器的可靠性和可用性相对较低故此类文件系统通常考虑并实现了副本机制, 性能相对 OCFS2 等内核态文件系统较低. 而 OCFS2 等内核态文件系统则使用与磁盘阵列场景.

## 参考文献


1 Thereska E, Ballani H, O'Shea G, et al. Ioflow: A software-defined storage architecture. Proc. of the Twenty-Fourth ACM Symposium on Operating Systems Principles. ACM. 2013. 182–196.

2 Guan JC. Research and backup implementation based on VMware VMFS. Electric Power Information Technology, 2011: 7–17

3 Gupta K, Jain R, Koltsidas I, et al. GPFS-SNC: An enterprise storage framework for virtual-machine clouds. IBM Journal of Research and Development, 2011, 55(6): 1–2.

4 Fasheh M. OCFS2: The Oracle clustered file system, version 2. Proc. of the 2006 Linux Symposium. 2006. 289–302.

5 Weil SA, Brandt SA, Miller EL, et al. Ceph: A scalable, high-performance distributed file system. Proc. of the 7th Symposium on Operating Systems Design and Implementation. USENIX Association. 2006. 307–320.

6 Beloglazov A, Piraghaj SF, Alrokayan M, et al. Deploying OpenStack on CentOS using the KVM Hypervisor and GlusterFS distributed file system[Technical Report] CLOUD-TR-2012-3 (2012): 1–49.

7 Yang D, Wang Y, Liu P. Fault-tolerant mechanism combined with replication and error correcting code for cloud file systems. Journal of Tsinghua University, 2014, 54(1): 137–144.

8 Shirinbab S, Lundberg L, Erman D. Performance evaluation of distributed storage systems for cloud computing. International Journal of Computers and Their Applications, 2013: 195–207.