文章编号：1005-8451（2019）4-0036-05

关于Ceph优化企业云后端存储方案的研究

解辰辉，刘承亮，曲左阳
（中国铁道科学研究院集团有限公司　电子计算技术研究所，北京　100081）

摘　要：以Ceph为基础为企业构建数据私密程度较高的企业云分布式存储系统，同时支持对象存储，块存储和文件存储。研究一种在实际环境中可以使用的混合存储方案。通过尝试社区版本的Ceph与基于OpenStack的企业云进行融合，发现其在应用过程中存在诸多问题。因此，根据企业云自身的业务特点，优化Ceph 对于企业云后端存储的支持，在Ceph原有的框架基础上，增加部分扩展，使得原生Ceph更加适合企业的生产环境。

关键词：企业云；后端存储；OpenStack；Ceph
中图分类号：U29：TP39　**文献标识码：**A

Optimizing enterprise cloud backend storage scheme based on Ceph

XIE Chenhui, LIU Chengliang, QU Zuoyang
(Institute of Computing Technologies, China Academy of Railway Sciences Corporation Limited, Beijing　100081, China)

Abstract: Based on Ceph, a distributed enterprise cloud storage system with high degree of data privacy was built for enterprises. The system also supported object storage, block storage and file storage. This article studied a hybrid storage scheme that could be used in real environment. There were many problems in the application process by trying to integrate the community version of Ceph with the enterprise cloud based on OpenStack. Therefore, according to the business characteristics of enterprise cloud, the article optimized the Ceph support for enterprise cloud backend storage, and added some expansion based on the original framework of Ceph, made native Ceph more suitable for enterprise production environment.

Keywords: enterprise cloud; backend storage; OpenStack; Ceph

OpenStack 是目前流行的开源云平台技术，是企业实现私有云平台提供 IaaS 形式服务的重要解决方案[1]。OpenStack 所包含的组件相对较多，各个组件间存在依赖关系，如每个组件都会依赖 Keystone，Nova 还依赖于 Glance，Neutron 和 Cinder；Swift，Glance 和 Cinder 需要后端存储的支持。

原生的 OpenStack 并不支持统一存储，云主机服务 Nova、镜像服务 Glance、云硬盘服务 Cinder 的后端存储各不相同。后果是内耗严重，单纯从创建虚拟机这一操作来看，通常需要 1~3 min。这样的设计缺乏合理的横向扩展性，当系统压力增大时，必然会出现各种问题。在构建云平台时，须对存储进行重新设计。早先业界不少学者或企业在为 OpenStack 优化 Swift 上做了大量的工作。本文尝试将云平台所有数据存储在Ceph资源池里，包括创建虚拟机，迁移，扩容，缩容等所有操作都可以避免不必要的数据传输[2]。提出一套更加适用于企业生产环境的优化方案：通过调整元数据备份策略，网络传输方式以及根据企业环境调整配置策略，提升整体性能。

收稿日期：2018-05-30
基金项目：中国铁道科学研究院集团公司科研开发基金项目（1751DZ0305）
作者简介：解辰辉，工程师；刘承亮，高级工程师。

1　Ceph分布式存储关键技术

Ceph 是一种高性能的统一分布式存储系统，具有高可靠性和可扩展性。Ceph 可以通过一套存储系统同时提供对象存储、块存储和文件存储系统 3 种功能，以便在满足不同应用需求的前提下简化部署和运维。其中，对象存储，既可以通过使用 Ceph 的库，利用 C、C++、Java、Python、PHP 代码访问，也可以通过 Restful 网关以对象的形式访问或存储数据，兼容亚马逊的 S3 和 OpenStack 的 Swift。块存储，作为块设备，可像硬盘一样直接挂载。文件系统如同网络文件系统一样挂载，兼容 POSIX 接口。在 Ceph 系统中分布式意味着真正的去中心化结构以及没有理论上限的系统规模可扩展性[2]。Ceph 的系统层次，如图 1 所示。

在 Ceph 的架构中，对象存储由 LIBRADOS 和

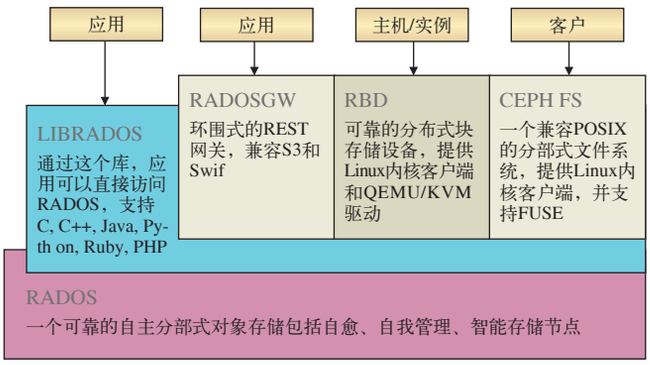

图1　Ceph的系统层次

RADOSGW 提供，块存储由 RBD 提供，文件系统由 CEPH FS 提供，而 RADOSGW, RBD, CEPH FS 均需要调用 LIBRADOS 的接口，而最终都是以对象的形式存储于 RADOS 里[3]。Ceph 集群的节点有 3 种角色：

（1）Monitor，监控集群的健康状况，向客户端发送最新的 Crush map（含有当前网络的拓扑结构）。

（2）OSD，维护节点上的对象，响应客户端请求，与其他 OSD 节点同步。

（3）MDS，提供文件的 Metadata，提供高层应用接口 CephFS[4]。

Ceph 是分布式的存储，它将文件分割后均匀随机地分散在各个节点上，Ceph 采用了 CRUSH 算法来确定对象的存储位置，只要有当前集群的拓扑结构，Ceph 客户端就能直接计算出文件的存储位置，直接跟 OSD 节点通信获取文件而不需要询问中心节点获得文件位置，这样就避免了单点风险。Ceph 已经是一套比较成熟的存储系统了，是 OpenStack 比较理想的存储后端，也可以作为 Hadoop 的存储后端。

Ceph 和 Gluster 都是灵活存储系统，在云环境中表现非常出色。在速率上两者通过各自不同的方式，结果不相伯仲[5]。这里之所以选用 Ceph 除了因为其更容易与 Linux 做集成且对于 Windows 较为友好，更是看重 Ceph 访问存储的不同方法有可能使其成为更流行的技术。Ceph 已经是主线 Linux 内核(2.6.34)的一部分，由于其具有高性能，高可靠和高扩展性，Ceph 作为一个优秀的开源项目得到更多的关注。

2　企业云后端存储集成Ceph

使用 Ceph 作为企业云（基于 OpenStack）的后端存储主要有 2 种思路：(1) 由于私有云本身包含 Swift 组件作为对象存储，Ceph 用 C++ 编写而 Swift 用 Python 编写，性能上应当是 Ceph 占优。但是与 Ceph 不同，Swift 专注于对象存储，作为 OpenStack 组件之一经过大量生产实践的验证，与 OpenStack 结合很好，目前，不少人使用 Ceph 为 OpenStack 提供块存储，但仍旧使用 Swift 提供对象存储[6]。(2) 将 Ceph 统一作为 Nova/Glance/Cinder 的存储后端[7]，如此一来，Nova, Glance, Cinder 之间没有数据传输，快速创建虚拟机，只需要管理一个统一存储[8]。本文采取第 2 种实现方式进行更深一步的探讨。

2.1　实验环境简介

整合项目实验资源（包含设备利旧），搭建实验集群[9]。总体集群部署分布，如图 2 所示。

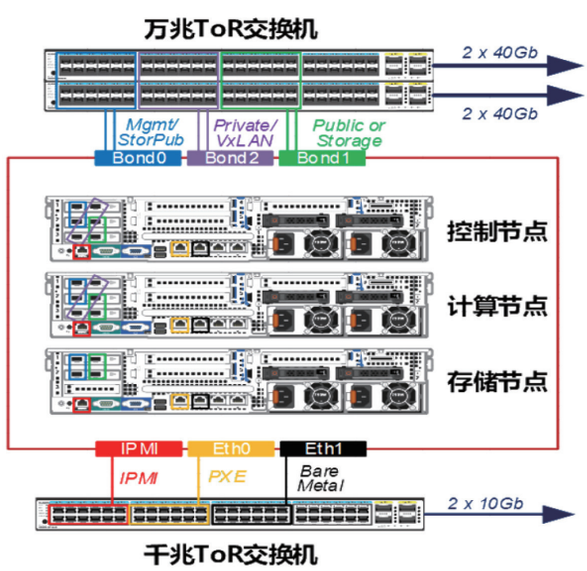

图2　实验集群部署功能示意图

出于未来业务发展需要，集群中部分物理存储节点选配了 3.2 TB NVME SSD 卡。所有存储节点标配 2 块 2.5 寸 600 GB 10K SAS 热插拔硬盘，2 块 480 GB SSD（intel S3520 系列）热插拔硬盘，8 块 6T 7.2K SATA 热插拔硬盘。

宿主机集群通过部署节点进行 PXE 网络部署，操作系统发行版为 CentOS Linux release 7.3.1611 (Core)，部署基于 M 版 OpenStack 深度开发的企业云，部署的网络架构，如图 3 所示。

控制节点：存储 public 网 - 双万兆，L3 public 网 - 双万兆，业务 Private 网 - 双万兆，部署和物理机监控 IPMI- 单千兆，管理网 - 双万兆（管理网和存

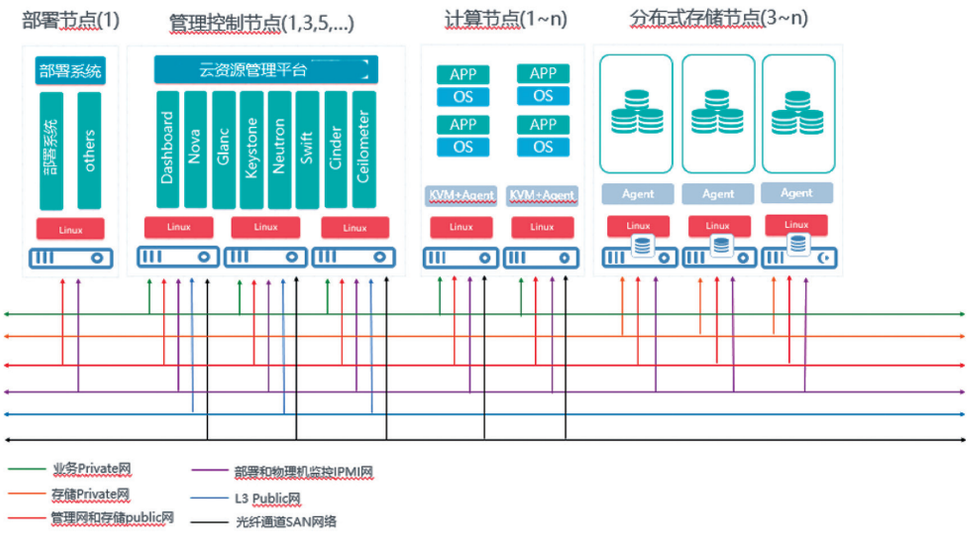

图3　集群部署网络架构图

储 Public 网合并）。

计算节点：存储 Public 网 - 双万兆，业务 Private 网 - 双万兆，部署和物理机监控 IPMI- 单千兆，管理网 - 双万兆（管理网和存储 Public 网合并）。

存储节点：存储 Public 网 - 双万兆，存储 Private 网 - 双万兆（可用 IB），部署和物理机监控 IPMI- 单千兆，管理网 - 双万兆（管理网和存储 Public 网合并）。

2.2　社区Ceph支持企业云后端存储

将 Ceph FS 作为 Nova 节点的本地文件系统。作为 OpenStack 中的共享实例存储，可以在 OpenStack 中使用 Ceph 块设备镜像，Ceph 块设备镜像被当作集群对象。还可以使用 OpenStack Glance 将镜像存储在 Ceph 块设备中。这样 OpenStack 的 Nova Glance 和 Cinder 之间没有数据传输。高可用集群只需管理一个统一存储 [10]。本实验采用社区版本的 Ceph version 10.2.5。

在本文的实验场景中，Ceph 资源池总容量约为 479 TB，设置为 3 副本，可用容量约为 159 TB。所有节点上 2 块 ssd 作为日志盘，ssd 分为 4 个区，每个分区大小为 40 G，每个 ssd 分区对应一个 osd。ceph 中分为 image、volumes、backups 3 个池，每个池设置为 3 副本。由于宿主节点有 2 种不同的存储介质，为了发挥硬件资源的最大效力将高速 SSD 存储和普通机械 SATA 存储划归不同的资源池进行测试 [11]。按照官方推荐架构做融合，修改 OpentStack 控制节点中 /etc 目录下的配置文件。融合架构，如图 4 所示。

在企业云平台上，实例化 1 台 Ubuntu14.04 LTS amd_64 的云主机，之后分别在 2 个不同存储介质的 Ceph 池中实例化 2 块 80 G 的云硬盘。启动云主机，分别挂载 2 块云硬盘进行 fio 测试；将 ssd 挂载到主机后 fdisk -l 显示硬盘路径为 /dev/vdb；采用读写混合模式 fio 测试如下：

fio-filename=/dev/vdb -direct=1 -iodepth 1 -thread -rw=randrw -rwmixread=70 -ioengine=psync -bs=16k -size=80G -numjobs=30 -runtime=100 -group_reporting -name=ssd

将 ssd 卸载，挂载 sata 硬盘重复上述步骤并进行读写混合模式 fio 测试，测试如下：

fio -filename=/dev/vdb -direct=1 -iodepth 1 -thread -rw=randrw -rwmixread=70 -ioengine=psync -bs=16k -size=80G -numjobs=30 -runtime=100 -group_reporting -name=sata

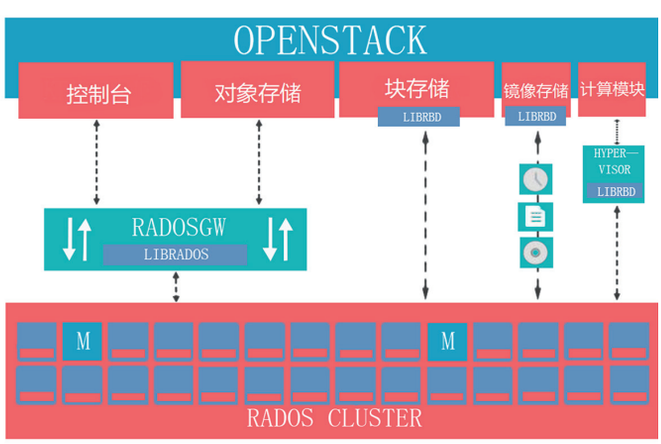

图4　融合架构

随后，卸载云硬盘，注销实例后重新申请，重复上述步骤 6 次。测试结果如表 1、表 2 所示。

所测试的主要指标包括 IO：总的输入输出量；bw：磁盘吞吐量；iops：磁盘每秒 IO 次数；depths：队列深度，为 1；blocksize：块大小，默认 16 K。

从上述测试结果中可以看出，使用社区版 Ceph

池化资源之后，SSD 资源池的读写速率只有 SATA 资源池的 2 倍左右。

表1　社区版Ceph-FIO-SSD池测试结果

关键指标		Ceph-SSD池					
		1	2	3	4	5	6
读	bw(KB/s)	330 136	365 295	332 364	335 429	329 473	302 375
	iops	20 633	20 453	21 023	20 868	20 089	22 186
写	bw(KB/s)	141 713	146 601	145 744	139 912	144 320	139 847
	Iops	8 857	9 012	8 463	8 356	8 217	8 541
硬盘使用率		100%	100%	100%	100%	100%	100%

表2　社区版Ceph-FIO-SATA池测试结果

关键指标		Ceph-SATA池					
		1	2	3	4	5	6
读	bw(KB/s)	161 171	159 791	166 523	162 351	160 716	163 224
	iops	10 073	10 000	10 524	10 740	10 221	10 092
写	bw(KB/s)	68 983	69 300	68 172	70 778	69 862	68 051
	Iops	4 311	4 326	4 278	4 390	4 328	4 291
硬盘使用率		100%	100%	100%	100%	100%	100%

2.3　优化Ceph对于企业云后端存储的支持

如上文所述结果，社区版 Ceph 存在一定的性能损耗。根据行业统计，社区版 Ceph 如果没有好的运维开发团队，存储节点数很难超过 20 个节点以上。在网络通信，线程调度，内存管理等方面都有很大的提升空间，可以做如下改进：

（1）Ceph 在读写数据的过程中使用 FileStore，在写数据块时先写日志再写数据，由于双写导致性能大打折扣。因此，可将元数据和 data 分离，只将元数据写入日志，以此来提升效率。

（2）增加热点预读冷池休眠功能，通过将访问频次较高的热数据存储在高速缓存中，提升读写性能。将热度下降的数据逐步在机械硬盘里落盘，同时控制冷存储池硬件，减少能耗，增长磁盘寿命。如此一来提高整体的效率。

（3）针对网络通信进行优化，主要通过聚合 TCP 链路来实现。同时增加对数据接口的支持，包括 FC，ISCSI 在内的冗余链路，保证业务链路的安全。并添加压缩和灾备策略，包括 1~6 的数据副本，纠删码等不同的策略。改进架构，如图 5 所示。

在具体的使用过程中，可根据集群状况对一些具体配置进行优化。本文中的配置优化简述如下：

（1）由于 Ceph-OSD 进程和 Ceph-MSD 进程都

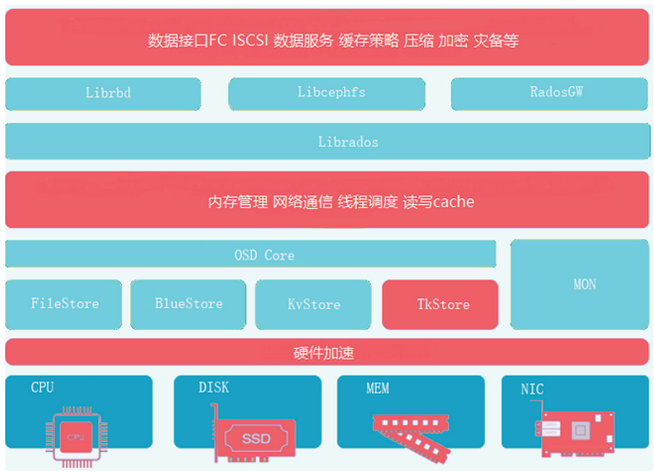

图5　改进优化的Ceph

会消耗大量的 CPU 资源，因此，通过在节点主机 BIOS 设置 CPU 等硬件为最佳性能模式（默认为均衡模式），提高效率，是一个较为简单却行之有效的方案。

（2）关闭 NUMA，可以在 BIOS 中关闭 NUMA 或者修改 Ceph 节点的配置，在 /etc/grub.conf 文件添加 numa=off 来关闭 NUMA。

（3）调整 PG 数量。在本文的实验环境中共有 118 个 osd；使用 3 副本；划分 6 个 pool，因此，Total PGs=5 900，每个 pool 的 PG 计算结果为 983，按照以 2 为底的指数函数取近似，设为 1 024。

TotalPGs=(Total_number_of_OSD*100)/max_replication_count

（4）设置预读，通过数据预读并且记载随机访问内存方式，提高磁盘读操作。

echo "8192" > /sys/block/sda/queue/read_ahead_kb

总之，高性能硬件带来效率提升的同时也会增加成本，需要企业根据自身情况和应用场景进行权衡。在架构和配置调优的过程中，不能仅仅关注性能，系统的稳定性同样重要，上述调优方式较为保守但行之有效。按照 2.2 节所述测试步骤，进行测试。优化 Ceph-FIO-SSD 池测试结果，如表 3 所示。

通过上述测试结果可知，优化后，Ceph 的效率较之社区版 Ceph 效率有较为明显的提升，为原有效率的 130%。

3　结束语

Ceph 是面向大型存储的应用，用以解决企业各

表3　优化Ceph-FIO-SSD池测试结果

关键指标		优化Ceph-SSD池					
		1	2	3	4	5	6
读	bw(KB/s)	436 279	436 943	439 025	435 619	439 051	434 190
	iops	27 267	27 666	29 125	28 223	28 416	27 191
写	bw(KB/s)	212 246	214 946	212 037	211 598	214 549	211 031
	Iops	13 266	13 434	13 252	13 224	13 409	13 189
硬盘使用率		100%	100%	100%	100%	100%	100%

种存储业务上的复杂问题。虽然 Ceph 的设计初衷定位为 PB 级的分布式文件系统，但在落地生产环境的过程中尚有很大的提升空间。本文探讨了一种可用于实际生产环境的 Ceph 实施优化方案，较之原生 Ceph 性能得到了显著提升，但在生产环境中也需要专业的技术服务团队作技术支撑。

参考文献：

[1] 熊振华．基于 OPENSTACK 云存储技术的研究 [D]. 长春：吉林大学，2014.

[2] 唐国纯，罗自强．云计算体系结构中的多层次研究 [J]. 铁路计算机应用，2012，21（11）：4-7.

[3] Weil SA, Brandt SA, Miller EL, et al. CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data[C]// SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing , Tampa, FL, USA, 2006:31.

[4] Weil SA, Leung AW, Brandt SA, et al. RADOS:a scalable, reliable storage service for petabyte-scale storage clusters[C]// International Petascale Data Storage Workshop. DBLP, 2007:35-44.

[5] Weil S A, Brandt S A, Miller E L, et al. Ceph: a scalable, high-performance distributed file system[C]//Symposium on Operating Systems Design and Implementation. USENIX Association, Seattle, WA, USA, 2006:307-320.

[6] Johanes J, Johari MF, Khalid M, et al. Comparison of Various Virtual Machine Disk Images Performance on GlusterFS and Ceph Rados Block Devices[C]//International Conference on Informatics Applications, 2014.

[7] Azagury A, Dreizin V, Factor M, et al. Towards an Object Store[C]//MASS Storage Systems and Technologies. IEEE, 2003:165-176.

[8] Wei Kong, et al. Multi-level image software assembly technology based on OpenStack and Ceph[A]. IEEE Beijing Section、Global Union Academy of Science and Technology, Chongqing Global Union Academy of Science and Technology. Proceedings of 2016 IEEE Information Technology,Networking,Electronic and Automation Control Conference(ITNEC 2016) [C]//IEEE Beijing Section, Global Union Academy of Science and Technology, Chongqing Global Union Academy of Science and Technology, 2016:4.

[9] Gudu D, Hardt M, Streit A, et al. Evaluating the performance and scalability of the Ceph distributed storage system[C]// Big Data (Big Data), 2014 IEEE International Conference, Washington, DC, USA ,2014:177-182.

[10] 赵铁柱．分布式文件系统性能建模及应用研究 [D]. 广州：华南理工大学，2011.

[11] 李　翔．Ceph 分布式文件系统的研究及性能测试 [D]. 西安：西安电子科技大学，2014.

[12] 王　斌．Ceph 存储在铁路站车 Wi-Fi 运营服务系统中的应用 [C]// 中国智能交通协会．第十一届中国智能交通年会优秀论文集．北京：电子工业出版社，2016.

责任编辑　徐侃春