



华南理工大学
South China University of Technology

工程硕士学位论文

开源分布式存储系统 Ceph 测试及在桌面虚拟化
平台中的应用

作者姓名	洪亮
工程领域	软件工程
校内指导教师	袁华 副教授
校外指导教师	李惊生 高级工程师
所在学院	软件学院
论文提交日期	2016 年 5 月

Testing of Distributed Storage System Ceph and Application of Desktop Virtualization Platform

A Dissertation Submitted for the Degree of Master

Candidate: Hong Liang

Supervisor: Prof. Yuan Hua

Senior Engineer Li Jingsheng

South China University of Technology

Guangzhou, China

分类号: TP3

学校代号: 10561

学 号: 201220208898

华南理工大学硕士学位论文

开源分布式存储系统 Ceph 测试及在桌面虚拟化平台中的应用

作者姓名: 洪亮

指导教师姓名、职称: 袁华 副教授

申请学位级别: 工程硕士

工程领域名称: 软件工程

论文形式: ☐ 产品研发 ☐ 工程设计 ☒ 应用研究 ☐ 工程/项目管理 ☐ 调研报告

研究方向: 软件工程技术

论文提交日期: 2016 年 4 月 30 日

论文答辩日期: 2016 年 6 月 5 日

学位授予单位: 华南理工大学

学位授予日期: 年 月 日

答辩委员会成员:

主席: 韩国强

委员: 高英、郑东曦、陈泽琳、陈虎

华南理工大学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：洪亮

日期：2016年6月1日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华南理工大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅（除在保密期内的保密论文外）；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。本人电子文档的内容和纸质论文的内容相一致。

本学位论文属于：

☐ 保密，（校保密委员会审定为涉密学位时间：____年____月____日）于____年____月____日解密后适用本授权书。

☒ 不保密，同意在校园网上发布，供校内师生和与学校有共享协议的单位浏览；同意将本人学位论文提交中国学术期刊（光盘版）电子杂志社全文出版和编入 CNKI《中国知识资源总库》，传播学位论文的全部或部分内容。

（请在以上相应方框内打“√”）

作者签名：

洪亮

日期：2016.6.1

指导教师签名：

袁华

日期：2016.6.2

摘 要

随着虚拟化技术的全面发展，桌面虚拟化和服务器虚拟化对计算能力的依赖，逐渐地转为对存储设备的性能和安全性需求。分布式存储系统的发展有效提高了整个系统的安全性和并发能力，保障了存储服务的高可用性，其易扩展的系统结构和低廉的硬件设备成本，在生产应用中也更具备竞争力。然而分布式存储系统原理较复杂，存储硬件与软件、网络多种要素相结合，并且技术本身正处在应用发展的上升阶段，更迭较快，实际应用中需要比较大的学习开发成本。分布式开源存储系统 Ceph 作为一款新兴的通用型存储底层系统，提供多种存储类型，常见于对象存储类型的应用。但是其块存储的应用，特别是在桌面虚拟化应用场景中，缺少指导调优的资料和经验，其复杂分布式的层次结构也增大了研究测试难度，使很多企业的 IT 部门无法很好地应用。面对分布式存储系统与传统存储截然不同的运行模式，在服务对象也有较大差异的桌面虚拟化应用背景下，对分布式存储 Ceph 的测试与应用有很大的实用研究意义。

本文从桌面虚拟化应用场景入手，介绍了国内外分布式存储的研究现状以及 Ceph 开源分布式存储系统的基本原理及优缺点。仔细研究传统存储在桌面虚拟化应用中所遇到的性能难点，如存储启动风暴而导致用户体验差等阻碍桌面虚拟化发展的关键问题，并结合桌面虚拟化应用场景的特点分析出对分布式存储的性能需求，设计出一套在该应用场景下具体可行的分布式存储性能测试方案。该测试方案针对桌面虚拟化对存储的特殊需求，从本地磁盘基准、分布式块存储和虚拟机内虚拟磁盘，三个层次协同完成对分布式存储系统的性能参照关联测试，找出三者之间的相互影响，根据测试结果调整参数获得最佳性能。最后在桌面虚拟化平台实际应用中，使用分布式存储提供的虚拟机磁盘支撑能力进行并发测试，评估测试结果。根据测试结果说明分布式存储 Ceph 经过测试后应用，可以满足 80 台桌面虚拟化并发性能需求，验证了测试方案有效可行，为生产方案的硬件选型及性能测试提供了有力的支撑。

关键词： 分布式存储 ； 桌面虚拟化 ； 测试

Abstract

With the development of virtualization technology, desktop virtualization and server virtualization in the past reliance on computing performance, gradually turned to meet the demand of I/O device performance and safe. The development of a distributed storage system has effectively improve the safety redundancy of the entire storage system and the hyper converged infrastructure, which ensure high availability storage service, it is easy to expand the system architecture and low hardware cost, in actual production more competitive. However, the lack of a distributed storage system itself is too complex, involving two levels of storage hardware and software, and is in the early stages of the rise of open source development, which rapidly change, resulting in the production of its application requires large learning costs. As a universal design underlying storage systems, distributed storage is widely used, so it lacks official tuning method for a specific application guidance, its complex system configuration increases the size and difficulty of research and testing, many corporate IT departments stop research of the difficulties. Faced with the traditional storage distinct mode of operation, especially in the case of traditional storage service objects are large differences desktop virtualization application background is very characteristic test research significance.

From the desktop virtualization platform, this paper introduces the basic principles, advantages and disadvantages and domestic and foreign research status from distributed storage to Ceph open-source distributed storage system. Combined with the characteristics of desktop virtualization scenarios, sorting out a feasible distributed storage performance and stability testing program. Special programs for desktop virtualization storage needs, from the local disk, distributed block storage and VM's disk, these three levels to complete the distributed storage system performance comparison test. Desktop virtualization platform in practical application, the use of virtual machine disk storage provides the ability to support distributed testing, test results are complete finishing 80 desktop virtualization traffic load level select the best distributed storage as its backend Select the stored program summary, and where insufficient research to make improvements.

Key words: Distributed Storage; Desktop Virtualization; Testing

目 录

摘 要.....	I
Abstract	II
目 录.....	III
第一章 绪论.....	1
1.1 研究背景与意义.....	1
1.2 国内外研究现状与分析.....	4
1.3 研究目标与主要内容.....	7
1.4 论文结构.....	8
第二章 分布式存储的技术说明.....	9
2.1 分布式存储系统.....	9
2.1.1 分布式存储基本协议与种类.....	9
2.2 Ceph 分布式存储介绍.....	10
2.2.1 Ceph 架构.....	11
2.2.2 Ceph 的逻辑层次.....	15
2.3 对比主流分布式存储技术.....	16
2.4 本章小结.....	18
第三章 分布式存储测试需求分析.....	19
3.1 应用场景.....	19
3.2 分布式存储测试需求分析.....	22
3.2.1 桌面虚拟化对分布式存储服务的需求分析.....	22
3.2.2 硬件需求分析.....	24
3.3 测试方案与测试用例.....	26
3.3.1 测试流程.....	26
3.3.2 测试方法.....	28
3.3.4 测试环境规划.....	29
3.3.5 测试工具的设计和选择.....	31
3.3.6 Ceph 性能测试用例设计方法.....	31
3.3.7 测试代码的设计方案.....	32

3.3.8 Ceph 性能测试用例设计	33
3.3.9 性能测试与系统调优	36
3.4 本章小结	37
第四章 Ceph 分布式存储系统测试及结果分析	38
4.1 测试部署	38
4.1.1 存储服务器以及测试平台配置	38
4.1.2 Ceph 测试集群网络配置	39
4.1.4 网络环境验证	39
4.2 性能测试	40
4.2.1 本地磁盘存储测试	40
4.2.2 桌面虚拟化对存储的具体性能测试	42
4.2.3 Ceph 块存储测试	43
4.2.4 测试 Ceph 提供的虚拟机磁盘性能	46
4.3 Ceph 扩容功能测试	49
4.4 异常测试	50
4.5 测试结果分析与调优	51
4.5.1 测试结果分析	51
4.5.2 性能调优	53
4.6 本章小结	54
第五章 分布式存储 Ceph 在桌面虚拟化中的应用	56
5.1 Ceph 分布式存储部署	56
5.2 桌面虚拟化后端连接 Ceph	58
5.3 验证 Ceph 在桌面虚拟化中的应用效果	59
5.3.1 解决启动风暴问题	59
5.3.2 满足虚拟机的应用需求	61
5.4 本章小结	62
结束语	63
参考文献	64
致 谢	66

第一章 绪论

在计算性能和网络性能高速发展的今天，以虚拟化为基础的云计算将这些过剩的资源重新汇聚到一起再按需分配，而作为基础设施中的存储却无论是从性能到安全性还存在难以突破的瓶颈。面对私有云平台主要实现的桌面虚拟化场景中，传统存储系统提供虚拟机镜像和高负载时都不可避免的出现了大规模“启动风暴”导致用户体验差的问题，本章节主要分析桌面虚拟化后端存储的主流选择及遇到的问题，为后续的分布式存储测试提供应用场景需求的研究方向，为基于桌面虚拟化应用的开源分布式存储测试测试方案设计实施打下基础。

1.1 研究背景与意义

企业在技术改革转型中越来越青睐尝试新技术。桌面虚拟化技术通过持续强有力的推进发展，在远程桌面传输协议上取得长足进步，优化了用户体验，不断的适应更多的常规应用场景需求，但是作为虚拟机背后最重要的大容量虚拟机镜像及数据存储，始终被存储性能与安全性约束^[1]。经过厂商的大力推广和技术的快速发展，分布式存储因为功能与性能全面提高代替传统共享存储成为必然趋势，也受到了共享存储的物理磁盘性能极限与其功能结构及性能问题一直有难以突破的物理性有着直接的影响。

桌面虚拟化中的分布式存储技术也逐渐成为了用户关心的重点，产品方案中是否有分布式存储，性能足以支撑多少虚拟机，是否解决了以前共享存储或本地存储情况下的启动风暴，是否有扩展方案和高可用，这些都成为了产品市场竞争力的关键。市面上常见的桌面虚拟化厂商，制定方案时为了性能和质量保证，一般都选择存储区域网络设备和网络共享存储设备等企业级商业存储设备，因为客户端与存储的连接依赖于相互之间的网络连接通信，所以带宽性能和扩展性受到局限，在集群规模变大的时候，性能和扩展性就会因为自身或者客户端硬件接口和带宽不足或者中间交换设备的增加受到限制，只是简单地增加网络设备规模，也容易出现单点故障，不利于企业大规模的数据安全要求。

虚拟化作为云计算中最基本的并广泛应用在数据中心基础设施建设的技术，在如今成为厂商争夺的热点，而虚拟化作为一种新的计算存储网络提供形式，与传统的物理计算机系统在对硬件的需求上转为了对虚拟资源的调用，跨越的硬件软件以及数据层次增加，逻辑变得更为复杂，新型的多集群、去中心化的网络结构，但是也由此提供了多种

多样的资源服务的形式。由于上述存储硬件设备的局限，跟随软件定义硬件的潮流，适时出现了软件定义存储，这个概念最早来源于威睿（VMware）公司提出的软件定义数据中心^[2]。软件定义的数据中心是指数据中心的服务器中的计算能力、存储能力、网络、安全设备等物理资源可以通过软件化方法或者硬件虚拟化来重新进行定义和构造，并且能够弹性有效快速自动地管理和分配这些资源^[3]。所以分布式存储这种新的软件系统在存储方面得到了更多的底层调用的机会，接管了集群内的物理硬件，抽象出设备里的软件和固件层，与操作系统或上一层次的管理软件服务衔接，结合分布式软件处理故障的方式，充分实现集群服务的 HA（High Available，高可用），从而完整了现有的分布式存储解决方案，将各节点的本地存储能力上升为分布式存储的多服务与多类型设备的整体提供，从性能到扩展性、数据安全性都到了极大的提升^[4]。但是在桌面虚拟化的应用场景下，用户对存储的直观感受无可避免地还是体现在“启动风暴”的问题上，这个问题如何解决成为了分布式存储在这一领域使用的重要突破。

保持良好的 I/O（Input/Output）吞吐性能还需要在极短时间内处理大量的 I/O 高峰并发访问，称之为“I/O 风暴”^[5]。而“启动风暴”则是因为虚拟桌面是被用户自身所驱动的，用户通常在早上上班时开启并登录到他们的虚拟桌面，在准备下班时保存工作注销关机。大多数启动登录操作很可能发生在工作时间前后半小时之内，随着大批量虚拟桌面启动读取虚拟机磁盘镜像，会在进入操作系统引导至登录桌面完成启动程序和服务运行时产生大量的读请求形成风暴。与此对应地在完成一天的工作时，用户在关机时保存工作日志，保存自己的工作并关闭桌面，用程序的保存和操作系统关闭了大量的随机读写操作系统。随着程序的保存和操作系统关闭产生大量的随机读写 I/O。这两种短时间内发生的 I/O 峰值只能通过虚拟化平台中的存储系统来协调，其中任何响应 I/O 请求的性能延迟直接影响用户的体验，甚至导致整个平台故障的发生，所以灵活使用廉价缓存的分布式存储是解决此类问题的绝佳方法^[6]。

桌面虚拟化厂商为了增强自身产品的市场竞争力，在使用了分布式存储的虚拟化项目上，制定解决方案只会有虚拟化桌面服务数量上的说明，只能依据服务器配置和用户规模、体验等方法推测出其分布式存储的应用方案。而且至今有的分布式存储方案还停留在研发测试试验，并未形成最终商业化产品，所以在业界对于此的测试方案和测试数据也非常的稀缺，厂商自身从测试分析得到的优化方向也作为核心竞争力基本不对外公开，如果企业内部的信息化部门想搭建分布式存储只能在开源社区的爱好者和专业研究人士的经验教训中去总结学习。而根据互联网上已有的测试方案和结果来看，大多数测

试忽略了具体的应用场景特点分析，本地存储性能、分布式存储块设备性能及虚拟机内部测试性能，三者没有相关联系起来，过于单独测其中的某一个指标，描述的局限性非常大，而且带来的结果也没有具体实际的参考意义，三者关系复杂又相互影响，很难得知性能瓶颈具体原因，给性能优化带来了很大的困扰，但是碍于桌面虚拟化对存储的访问复杂，也使得测试难以设计。虚拟机对后端存储访问结构如下图 1-1：

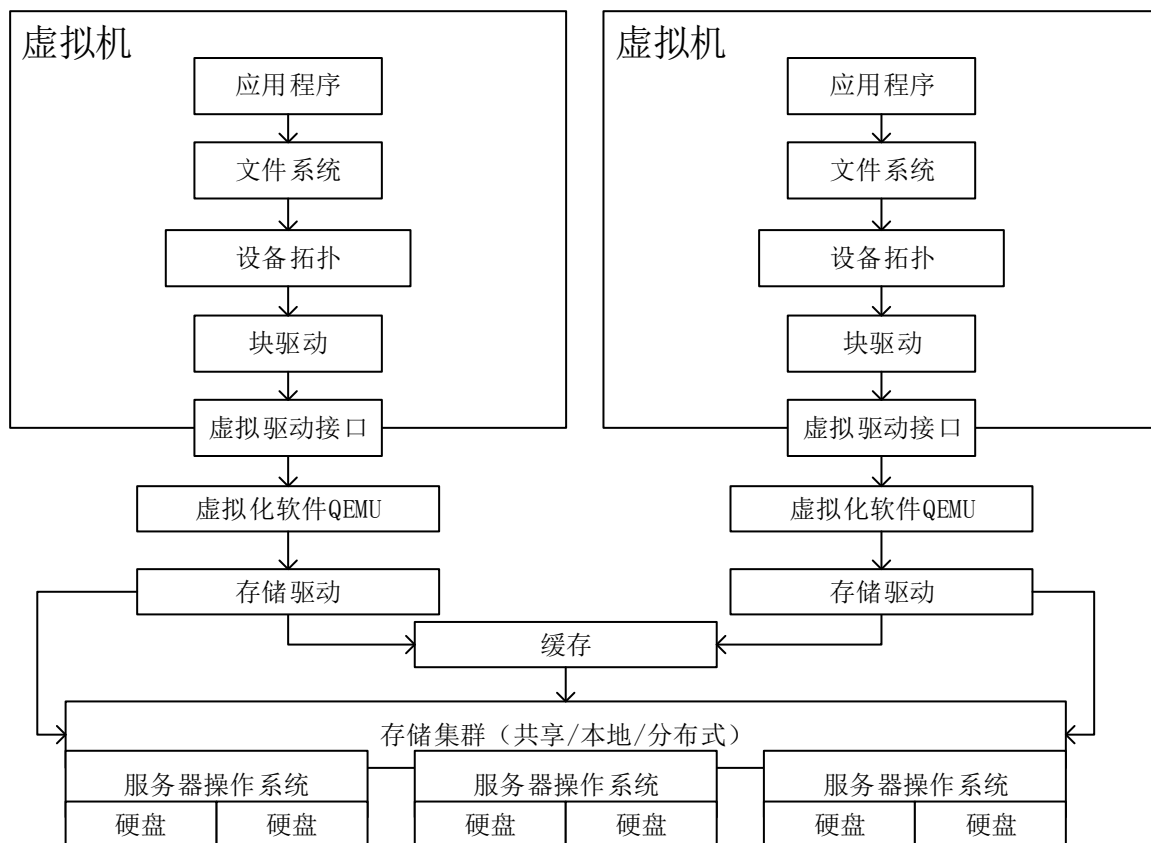


图 1-1 虚拟机应访问存储结构图

从上图中可以得知桌面虚拟化上的应用访问此存储时需要经过多个层级，到达存储集群，所以存储后端必须有足够的性能和安全性，以保证整个业务的正常运行。在具体研发测试试验中，使用传统存储集群由于性能局限，导致大规模虚拟机启动风暴问题也没有实质性的解决，所以本文以此为切入点，对特定应用场景进行分布式存储测试，形成解决在桌面虚拟化应用上开源分布式存储的测试方案。

桌面虚拟化中的启动风暴导致的用户体验等问题，经过分析发现是因为传统存储的 IOPS 性能不足。面对一般客户在小规模生产环境中在无法承担使用昂贵的高性能商业存储的情况下，根据研究决定尝试使用开源分布式存储进行应用解决。所以需要对分布式存储方案的应用进行预研、测试其与既有桌面虚拟化平台的兼容性和支撑力探索，也需要根据其中的测试得到的数据变化动态，给虚拟化管理平台提供实时的负载均衡所需

要的数据分析和监控指标。本文的研究对虚拟化管理平台资源的稳定性、监控完整性以及用户体验这些桌面虚拟化平台核心竞争力起着至关重要的作用，并且最终能以较为详尽的数据事实支撑产品介绍和工程实施方案的设计。通过分布式存储解决桌面虚拟化的启动风暴，降低开机时间，并且确定满足大批量虚拟机并发开机性能需求，并且通过实际应用进行验证分布式存储 Ceph 有能力支撑大量虚拟机的并发访问。

1.2 国内外研究现状与分析

在谷歌、亚马逊、阿里巴巴等技术领先的大型跨国互联网公司，还有 IBM、红帽、微软这类软件解决方案厂商极力推广的云计算和大数据背后，都不难发现他们看中分布式存储系统的发展前景。这其中，谷歌最初将 GFS (Google File System) 分布式存储文件系统应用于搜索的后端存储，也根据 YouTube 和 Gmail 等应用的需要升级到了第二代，亚马逊著名云服务 AWS 后端提供存储的 S3 和 Amazon EC2 后端的块层次的存储卷 EBS (Elastic Block Store)，阿里云背后飞天平台使用名为“盘古”系统的分布式存储，IBM 的 XIV 分布式的网格级磁盘存储系统，红帽收购的两大分布式存储系统 Ceph 以及 GlusterFS，微软全面支持及集成 Apache Hadoop 的 HDFS (Hadoop Distribute File System) [7]。而 NetAPP、易安信(EMC)等存储公司也提供了专门为虚拟化平台服务的分布式存储系统，富士通公司也开始推广直接基于开源分布式存储 Ceph 的存储设备 ETERNUS CD10000 等产品，而国内的华为也开始宣传使用了分布式存储技术 FusionStorage 作为存储底层的 OceanStor V3 融合存储[8]。为了能共享技术发展成果和缩短开发周期，使用开源技术并参与开源项目为一个共同的项目提交自己的代码，也极大的推动了平台技术的发展，所以开源平台以及项目成为了厂商各取所需和共同进步的重要技术来源。分布式存储作为新的研究方向成为了各个厂商的技术发展关键，加上开源社区的技术生态建设也使得更多的厂商、技术公司和专家爱好者们投入到研发和使用中，给分布式存储技术带来快速发展。国内由于起步较晚，在虚拟化底层核心和存储技术基础比较薄弱，所以普遍采用直接集成开源分布式存储或者修改其中部分组件和算法来满足商用需要。但是大多数技术平台直接集成相关存储技术，却又缺乏明确的应用场景案例和资料，缺少合理的测试理论依据，难以给分布式存储的应用设计合理的实施方案并进行商业化生产的尝试。

目前存储投入及分布式系统需求的快速增长，在新的趋势下结合桌面虚拟化应用场景由于性能瓶颈和安全考虑，逐步开始选取开源分布式存储作为后端，有针对虚拟化存

储特性优化及实现虚拟化管理系统所需接口，对桌面虚拟化性能需求和多种存储提供方式做出分析选择。现今全世界最大也是最受瞩目的云计算管理平台项目 OpenStack 是 NASA（National Aeronautics and Space Administration 美国宇航局）和 Rackspace 协作开发并推进其开源的，其在 2014 年 OpenStack 社区中受访者对块存储使用分布的调查调查的结果如下图

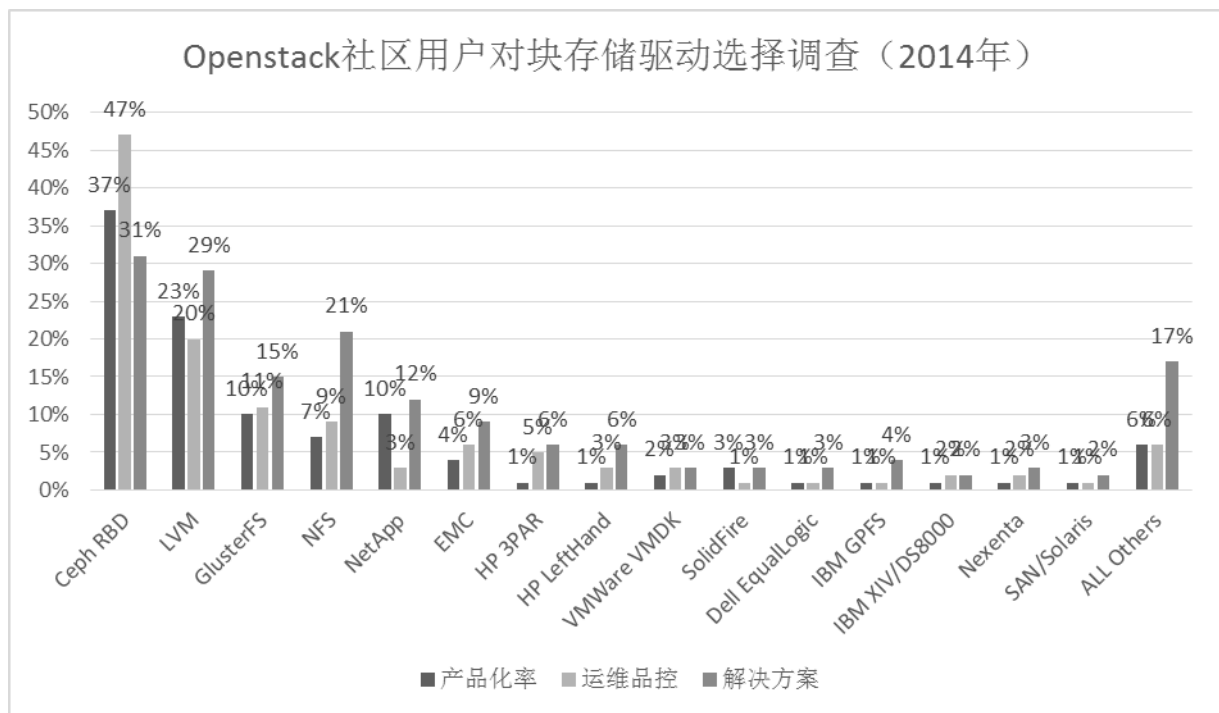


图 1-2 OpenStack 社区用户对块存储驱动选择调查（2014 年）^[9]

根据 OpenStack 社区的调查后，排名第一的 Ceph 分布式存储系统作为云计算虚拟化平台块存储后端是用户们关注的焦点并且遥遥领先于其他产品，由此可说明分布式存储较之其他选择更有强大的生命力和持续支持发展的动力。开始针对开源分布式存储 Ceph 进行研究，在笔者对 Ceph 中国社区的开发者进行统计，大部分 Ceph 存储的开发者和产品都集中在公有云存储服务和企业内部存储中，合计占了 90%，针对桌面虚拟化的使用比较少只有 4%，仅有三家具有较强硬件能力的厂商将 Ceph 集成使用在桌面虚拟化方案中完成产品化，如下图 1-3：

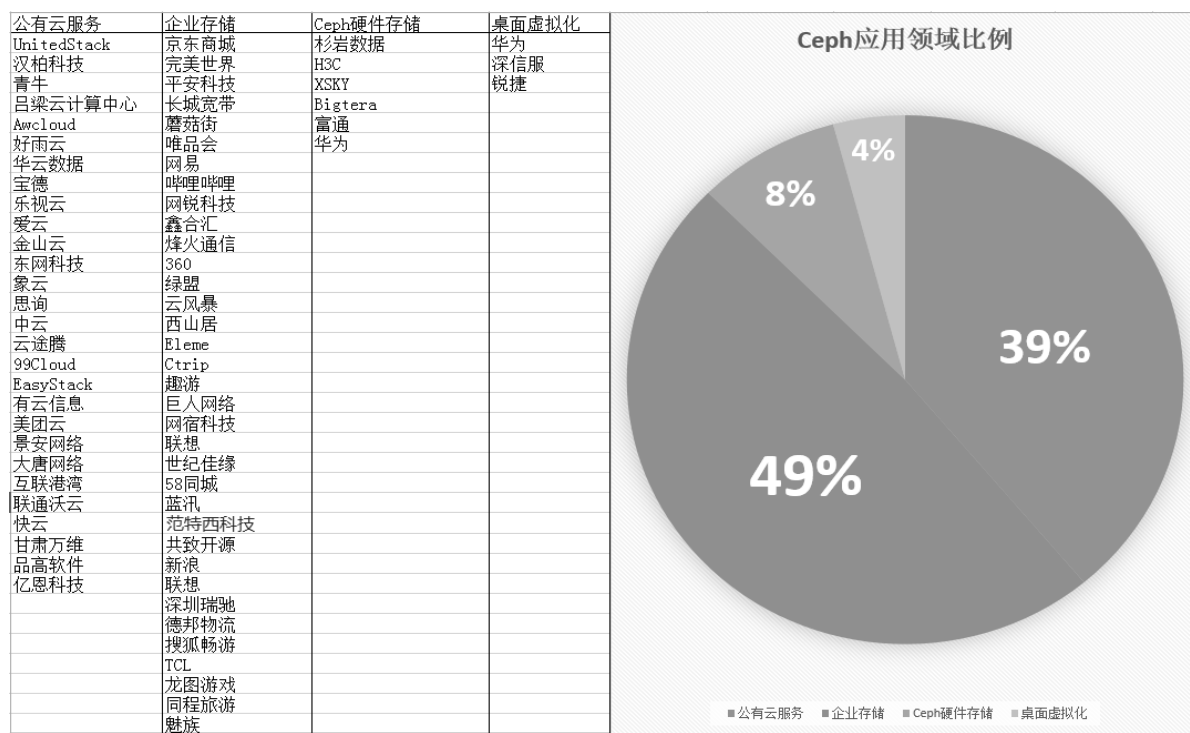


图 1-3 Ceph 中国社区开发者统计

集成和测试这些分布式存储也成为了许多云计算平台厂商的工作重点，具体的使用方法和调优测试数据等也作为技术优势而没有公开，所以为了验证一项技术是否能支撑业务系统，只有通过测试来验证是否符合用户需求并且进行调优后才能够正常地应用到生产环境中。由于开源软件社区的特点，分布式存储软件本身由于丰富的组织形式和广泛的应用场景，官方并没有给出业务级的测试方案，仅仅对分布式存储本身提供性能和功能的数据收集，而且使用的软硬件资源也趋于理想化，在实际生产环境并没有很强的指导意义。

区别于传统存储成熟的测试技术和简单的方法，现在缺乏一种有效全面的测试方案判断开源分布式存储是否符合何种规模的桌面虚拟化，而且虚拟化应用不同于常见应用有固定的性能表现，桌面虚拟化对存储的使用也越过多个层级，评判存储功能是否达到要求，无法在方案制定和介绍产品时给出量化的结果和可信的数据，所以需要制定一个合适的测试方案，以便能在研发和生产环境中对分布式存储的使用上得到合理可靠完整的测试结果或者可复制的测试方案。面对如此多种的分布式存储选择，用户亟需通过全面的评估手段和详实有效的测试数据，来对不同的应用场景下选择合适的存储系统作为虚拟化平台的可靠保障。目前，虽然存储系统已经存在多种性能测试工具，主要有 IOZone、FIO、IOmeter、Vdbench、Postmark 等免费开源测试软件，但是只能得到一些简单的独立数据，需要通过多种数据的交叉对比组合，需要人工确定数据与参数之间的

关联性，形成系统完整的评估体系^[10]。还有一些第三方测试组织及服务商的标准化评估基准测试方法如 SPEC 组织的 SFS 和 SPC 组织的 SPC-1、SPC-2、SPT，这类专用测试工具的通过是模拟用户使用场景下的负载，真实反映在实际使用条件下的性能表现^[11]。但是一般企业由于实力和资金有限无法提交至大型测试组织做出整体测试。为了能提供研发人员自行多维度评估分布式存储在虚拟化平台中的使用情况，一套精简有效的测试方法或方案能在短时间将多种测试工具得到的数据系统地组织起来，为分布式存储性能分析和方案及性能指导优化非常必要。在没有桌面虚拟化应用场景的测试的具体数据与参考确切资料情况下，由于桌面虚拟化方案常常需要为客户的服务器利旧做准备，没有固定标准的硬件支撑，难以评估 Ceph 存储支撑能力以及和硬件性能之间的关系，所以 Ceph 存储是否足以支撑适用于桌面虚拟化模式下的工作负载，在大规模并发访问中性能的稳定性，软硬件平衡的性能系数与大规模按需扩展的能力都是文本的研究意义所在之一。

1.3 研究目标与主要内容

本文的研究目标首先是了解分布式存储和 Ceph 的基本概念，并且与其他主流的分布式技术进行对比，学习分布式存储的特点和 Ceph 的工作原理有利于系统的分析可能存在的性能瓶颈和应用调优。为了解决启动风暴问题需要获得桌面虚拟化对于存储的具体性能要求，设计一个简单可行的分布式存储测试方案，能根据分布式软件的特殊性，在一致的硬件环境中测试 Ceph 本身的性能，并且对比服务器本地存储环境性能，分析之间联系，再分别测试出 Ceph 分布式存储提供的块设备性能和提供给虚拟机作为磁盘的性能，在虚拟桌面特定性能指标下，服务器能提供正常性能保障的最大虚拟机数量，将三者测试数据联合起来进行结果分析，研究出分布式存储优化的方案，最终在桌面虚拟化特定的应用中，能验证优化所带来的提高分布式存储性能和可用性，减少启动风暴等不良因素的发生。通过测试，得出虚拟机的存储性能需求，并与分布式存储 Ceph 提供的虚拟化存储设备集成后提供的虚拟机集群性能结合，探索 Ceph 分布式存储对虚拟机的性能支撑能力，最后确定测试方法正确可行并发现分布式存储中存在的性能瓶颈，并对性能瓶颈进行修改验证系统的生产环境运行参数设置是否合理，或确定该参数，完成对 Ceph 分布式存储的应用进行调优，以获得不同备选方案的性能表现，为方案选择提供性能数据支持。

测试评估分布式存储系统与以往常见的传统存储设备测试不同，需要综合考量更多

的影响因素，并且需要对分布式存储下的各节点需要先进行传统物理存储能力和文件系统的测试以保证测试数据的准确有效，测试结果才能更有实际意义和说服力，下面先分别结合文件系统和传统存储的测试方法介绍。

本文主要进行了如下工作：

1. 阐明分布式存储 Ceph 基本原理，对分布式存储使用中遇到实际问题进行解析。
2. 根据桌面虚拟化对存储性能需求的特点进行分析。
3. 制定合适的测试方案并进行测试，测试从本地存储到 Ceph 分布式存储提供的块设备性能和在特定性能指标下服务器能提供正常性能保障的最大虚拟机数量。
对测试数据结果进行分析和推敲。
4. 最终将测试调优后的分布式存储 Ceph 应用在桌面虚拟化中做出测试验证。

1.4 论文结构

本论文分为五个章节，内容如下：

第一章为绪论，主要介绍了以桌面虚拟化为特定的应用场景，分布式存储代替传统存储的重要意义和趋势，对分布式存储与其测试现状的分析，表明了针对特殊应用场景的分布式存储测试的必要性和研究意义，并概述论文的主要目标工作与文章结构。

第二章主要介绍了分布式存储结构及原理，介绍了开源分布式 Ceph 的设计理念和运行方式，并且与主流的分布式存储进行对比。

第三章在桌面虚拟化的特殊场景中对分布式存储的性能与功能需求，分析分布式存储测试中的包括桌面虚拟机所需要的吞吐量和 IOPS(Input/Output Operations Per Second)，即每秒钟存储进行读写（I/O）操作的总次数等关键指标，其次根据分布式存储特性制定有效测试方案，介绍测试过程，设计测试用例。

第四章配置分布式存储的测试环境，并进行测试统计并呈现测试数据，根据数据做出分析，对测试方案做出验证。在性能测试中对本地磁盘、桌面虚拟化存储、Ceph 块存储以及 Ceph 作为桌面虚拟化存储后端进行测试，分析结果在实际应用场景中的价值，并对测试方案进行调优和改进。

第五章描述分布式存储 Ceph 在桌面虚拟化中的应用，首先部署分布式存储 Ceph，桌面虚拟化如何连接和调用 Ceph，最后通过测试验证 Ceph 在桌面虚拟化中的实际性能。

第六章为结语，对本文及测试工作及测试过程进行梳理总结，然后对以后测试研究方向的展望最后是参考文献与致谢。

第二章 分布式存储的技术说明

在前一个章节中，主要介绍桌面虚拟化应用场景下，分布式存储的契合的特点、业务要求和发展现况。这一章将介绍主流的分布式存储原理及问题，Ceph 分布式存储介绍，分析分布式存储测试中的关键指标，在桌面虚拟化的特殊场景中对分布式存储的性能与功能需求，其次分析分布式存储测试性能的基准，根据分布式存储特点制定测试方案。

2.1 分布式存储系统

分布式存储系统的根本是计算机通过网络相互通信交换数据，将自身存储能力成为一个抽象的存储服务，利用多链路带宽和多点优势，达到超过相比传统存储的性能和安全性突破的效果^[12]。

2.1.1 分布式存储基本协议与种类

分布式存储系统的基本在于分布式概念存储安全性评判的体现，一般分布式系统中的核心是由两个重要的协议实现：Paxos 算法一致选举协议和两阶段协议^[13]。Paxos 算法一致性选举协议用于多个成员节点间，根据 Paxos 算法的选举协议规则仲裁达成一致后，选举某个合乎仲裁结果的成员作为总控制节点，以此保证分布式系统中的强一致性。确保事务在多个节点之间操作的原子性，这些操作只产生全部成功或者全部失败两种结果，不会产生其他模糊或者无法明确的结果，避免操作纠纷和失误，这就需要两阶段提交协议的实现了^[14]。保证主控制器服务的服务完整和集群操作的原子性，这既是分布式系统核心理念，也为整个系统实现高可用的解决方案提供了支持。

现阶段分布式存储系统在存储方面的研发重点主要在于数据存储冷热分层、高速缓存多层级、状态信息的持久化，而且集群中需要实现存储数据的跨节点冷热迁移，多份数据间容错纠错，大规模同时读写的数据一致性这些商业存储的基本功能。现今分布式存储领域主要分为以下几种：块存储（Block Storage）、对象存储(Object Storage)、文件系统(Filesystem)和表存储(Table Storage)这四种分布式存储技术，在桌面虚拟化虚拟机的使用暂时只采用了分布式存储的中的块存储作为使用对象^[15]。

分布式块存储：类似于物理磁盘，操作系统对磁盘的调用直接找到块设备上事先标示的地址进行操作。因为是直接从存储底层访问数据并操作，相比传统存储存取方式来

说,访问使用块存储的读写性能最高,所以常见的大规模数据库服务都直接部署在块存储设备上。分布式块存储以市面上常见的计算机硬盘作为基本存储单元,分布在不同计算机上的分布式存储节点通过网络点对点拓扑(peer-to-peer,简称 P2P)技术联通,由此形成分布式网格存储。存储服务管理内部事务和数据时采用分布式算法保证安全性来管理存储资源,这样多节点之间利用以太网带宽和总线速度优势,可以通过软件在硬件结构上模拟第二层磁盘阵列冗余功能,将阵列卡功能扩展到一个集群规模,简化单机的 RAID 技术,采取集群多节点多盘冗余的方式进行存储备份来保证数据安全,可以将原有的 Raid 卡成本转为容量和盘位,从而降低了集群存储成本,又保障了性能和安全之间的平衡^[16]。

根据以上介绍分析出分布式存储系统具有以下特点:

- 1) 安全性能: 分布式系统的核心就是安全,通过多节点保障了不发生单点失效的可能,将故障的几率分摊下去,系统自动容错也使得安全性能大幅增加。
- 2) 高扩展性: 轻松扩展节点数量,满足集群规模的不断发展,同时规模越大系统的性能和安全性也随之提高,并且自动负载均衡保障性能不出现瓶颈。
- 3) 成本低廉: 对硬件依赖极低,可以部署在标准的个人计算机甚至嵌入式计算机上。因为扩展非常方便平滑,通过工具可实现自动化的部署扩展接入分布式存储系统。
- 4) 性能优异: 利用了集群内多节点,节点多盘位的特点,充分使用计算机内部总线带块和节点间网络通信带宽性能,为整个集群的存储使用提供可靠的高性能的数据访问。
- 5) 简单易用: 接口标准方便外部调用,按需提供存储资源,底层透明,可以方便地实现监控、运维手段,易于其他系统集成。

2.2 Ceph 分布式存储介绍

Ceph 是 Sage Weil 在 2007 年 12 月在开源分布式文件系统的毕业论文中创造的存储系统。随着云计算的发展和分布式存储的热潮,开源社区的交流和快速推进,在 2010 年 3 月,经过 Linux 内核小组审核并验证,决定在 Linux 2.6.34 内核的正式版本中集成了这个非常有发展潜力的软件,2014 年 4 月,红帽公司以 1.75 亿美元收购 inktank 公司,而 inktank 的创始人正是 Ceph 的创始人 Sage Weil,收购该公司后其主要目的是为了推进软件定义存储 Ceph 的发展,并着力将 Ceph 开发成为能够提供商业化解解决方案的技术之一^[17]。所以在有了大厂商推动的背后,作为虚拟化、云计算关注的重点,Ceph 也成

为了 OpenStack 这个最大的云计算管理平台社区的生态系统中呼声最高的开源存储技术方向。

在 Ceph 官网的定义：为极快的读写速度与稳定可靠的服务、简单有效的扩展性而设计的分布式存储系统。Ceph 分布式存储系统提供了块存储、文件系统存储和对象存储，这三种基本的存储类型可以实现不同场景下的完整实现和运营维护^[18]。而分布式在 Ceph 存储系统中则体现了无中心的结构设计和集群系统的大规模可扩展性。下面来详细介绍一下 Ceph 这种分布式存储系统。

2.2.1 Ceph 架构

Ceph 分布式存储系统可以按角色分为四个部分（见下图 2-1）：

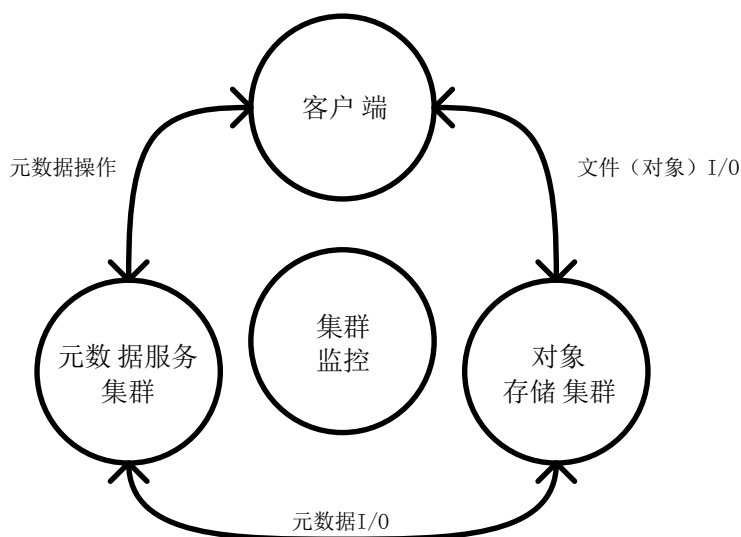


图 2-1 Ceph 分布式存储系统的角色示意图

如上图所示，四个部分分别为：1.客户端 Clients（存储服务访问），2. 对象存储集群 Object storage cluster（存储数据和元数据作为对象存储），3. 元数据服务集群 Metadata server cluster（同步分布式存储中的元数据并提供缓存），4.集群监控 Cluster monitors（履行监控协同功能，维护集群状态）。

Ceph 分布式存储通过元数据服务集群，对元数据进行读取来查找存储地址^[19]。元数据服务集群用于存放管理数据地址，并且调度管理新数据分配存储地址。元数据存储分布在分布式集群中与对象存储集群进行元数据的 I/O 操作。实际的 I/O 操作产生在客户端访问和对象存储集群两者之间。存储中的 POSIX 功能和所产生的常用操作就交由元数据服务器进行托管，较为简单的读写操作通过对象存储集群来执行，这样提高了整个存储

数据管理的速度，避免管理决断和执行集中在一点^[20]。

以下分别对上述角色进行分析介绍：

1. 元数据服务集群

元数据集群用于管理元数据的访问请求，元数据统一存储在对象存储设备(Object Storage Device,简称 OSD)上，元数据服务集群只需要处理已经缓存和同步的元数据请求^[21]；元数据集群上使用动态子树分割来管理缓存的元数据信息（如下图 2-2）：

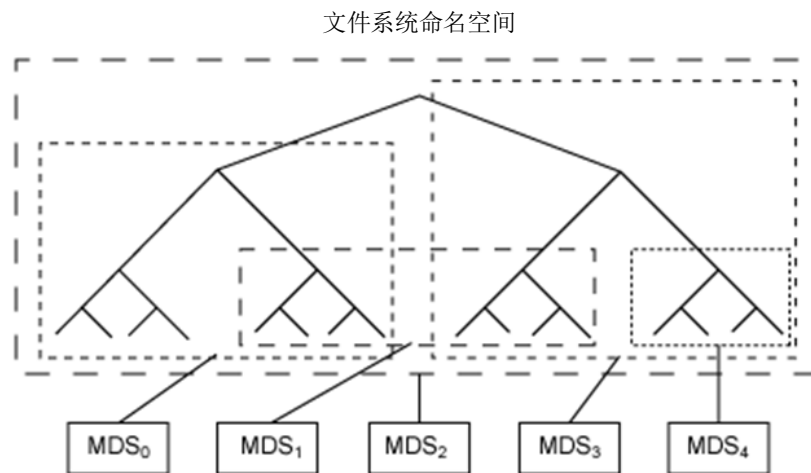


图 2-2 Ceph 动态子树分割示意图

子树分割实时地将元数据信息自动分布到一组节点上，不同的元数据服务器统计其分支下的权重，给每一个元数据服务生成对应的具有描述负载分布的权值树。定期查询元数据服务的分布负载，必要时自动迁移其中大小合适的子树，以此来维持整个元数据系统的负载均衡^[22]。这其中共享部分的存储与命名空间的状态锁，维持系统内存中迁移数据的一致性，降低对系统安全的影响，使得 Ceph 的架构和机制得到保证，确保在数据或状态发生错误前后以及错误过程中，性能基本不会受到影响。

Ceph 在元数据服务上有两大特点：一个是文件索引元数据的结构简单，可以把索引结构存储在与关联的目录项中，此外，Ceph 分布式存储将所有的目录信息（目录和索引）统一存储在对象存储集群中，所以客户端通过访问上一层目录就能预加载索引信息，极大加快读取速度；另一个是 Ceph 存储中所有元数据服务端上都有一个更新操作的记录日志，元数据服务能够快速有效地寻找有效的更新操作，减少随机读写的次数，提高性能。

2. Ceph 对象存储集群

与常见的对象存储相同，Ceph 服务中的节点包括基本存储功能和一些自动的扩展功能。一般的存储只能被动地接受控制器的指令，Ceph 的对象存储设备不仅能响应控制器指令，还可以自主判断控制并与其他对象存储设备连接通讯^[23]。常见于对象存储设备

中对象到块的映射操作流程如下（图 2-3）

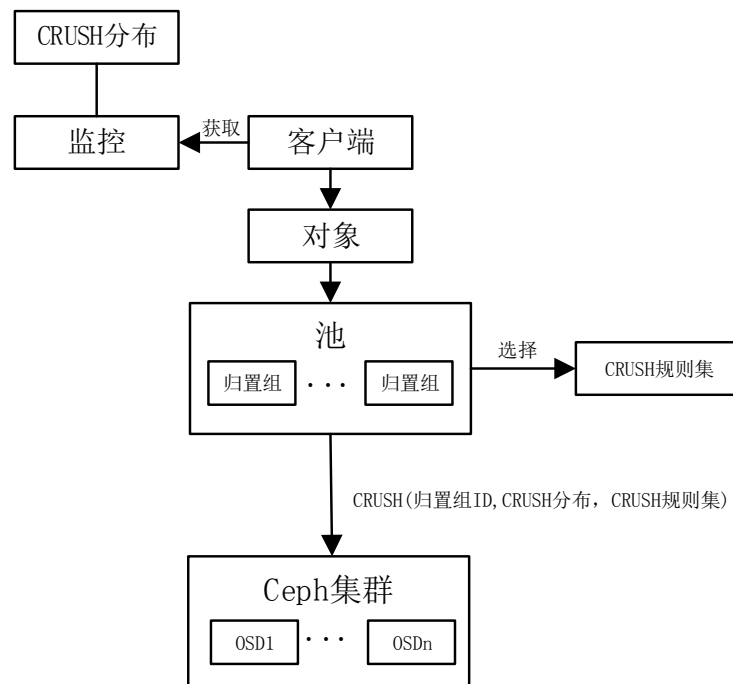


图 2-3 对象到块的映射过程流程图

这一过程中，服务自身做出决定通过何种最优渠道存储对象。因为 Ceph 使用 CRUSH 伪随机算法将数据进行分布，存储系统资源池就能透明地管理对象到块的映射。当监控得到设备故障信息，进行其他存储节点复制数据进行灾备，同时开始进行服务检查与故障恢复^[24]。对象数据存储有以下两种：一是数据迁移，为扩大容量增加磁盘进入系统，集群节点被随机的抽取部分存储数据到新增加的磁盘上，平滑扩展存储空间和保持性能的负载平衡；二是数据分发：通过 hash 算法结合对象计算得到对象归置的组别，然后再通过 CRUSH 算法进行对象存储设备的对应。如下图 2-4 所示。

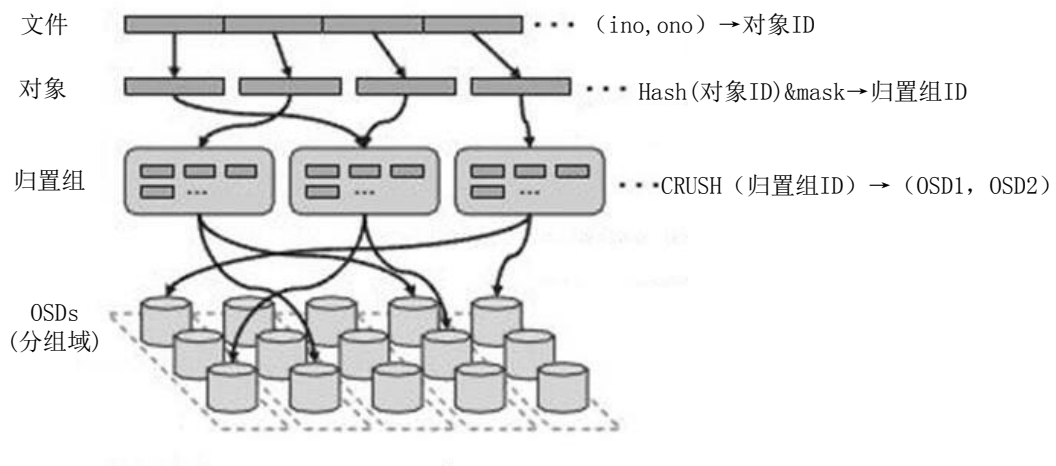


图 2-4 数据分发示意图

- 1) 对象到归置组(Placement Group, 简称 PG)的对应。同一归置组的对象会被放置到相同的对象存储设备集合中去。归置组是对象按照一定规则结合 Hash 算法产生的逻辑集合^[25]。
- 2) 对象存储 RADOS (A Reliable, Autonomous, Distributed Object Storage, 可靠自动的分布式对象存储) 系统根据集群映射表将归置组分配到相应的对象存储集合中。对象存储设备地址与归置组中的对象数据的存储地址相同^[26]。通过 CRUSH 算法计算出对象存储设备地址, 但这不一定是最终的数据存储目标, 需要经过初步的过滤筛选, 分布式系统集群规模的特点, 各种错误异常状态会使得部分节点可能失效, 如果筛选后无法满足使用则需要锁死避免误操作。集群映射中关于设备状态的区分如下表 2-1 所示。

表 2-1 设备状态

	-	进(in)	出(out)
-	-	分配归置组	不分配归置组
启用	在线 & 连接	活动状态	在线 & 空闲
禁用	无法连接	无法连接 & 无法重新映射	失败

对象存储的映射分发, 集群映射是通过资源抢夺的方式进行与对象存储设备之间的更新。在获取存储进行连接之前都需要交换对象映射的信息, 保证集群映射的一致。集群映射在客户端及服务端两者间独立的更新, 有效降低了整体集群映射分布压力。所有的对象存储设备都会同步预存最新的集群映射, 而且到所有增量映射信息也会进行缓存, 对象存储设备的所有消息都会嵌入增量映射, 并与其通信的相邻对等方的集群映射的版本保持同步监控^[27]。如果对等方得到的监控版本过期失效, 对象存储设备会重新共享对等方来说的增量映射, 使两者版本一致, 当对等方收到同步监控消息中校验位自身版本过期, 就会从其监控消息的增量映射中查询本地的增量映射然后修改, 保持一致。

对象存储集群中的自动管理中, RADOS 使用 3 种不同的副本分配办法, 主拷贝: 读写操作主对象存储设备上进行, 并同时更新其他副本; 链式: 链式读写, 读写分离; 伸缩型: 主拷贝和链式的结合: 并行更新副本和读写分离。

3.Ceph 集群监控

Ceph 分布式存储即使对自身对象存储映射的管理, 但是有些错误操作是在对象存储过程自己产生的。所以需要监控服务在对象存储设备出现错误和存储设备改变时, 能实时检测并保证完整有效一致的集群映射, 避免引发数据安全事故。

首先是集群节点失效进行监控，一般在线的元数据服务固定频率地向监控发送信息进行状态确认，当出现了元数据服务在规定时间内不予集群监控交互时，就可以怀疑该元数据服务已经下线，集群监控立即改变状态，同时将状态消息下发至集群中的各存储节点^[28]。如果元数据服务节点自身在于监控的通信中无法收到状态回馈，自动更改状态为失效，避免无效操作保证集群数据一致。当元数据服务失效时，集群的数据读取服务维持正常，因为存储数据已经落在磁盘等长效存储设备上。当集群中的存储节点发生失效，集群上的剩下节点就启动了故障恢复操作，重新将失效节点上的数据迁移到正常节点上。

以上就是 Ceph 的整个架构和每一个角色可以执行的动作，显示了 Ceph 分布式存储系统完整的安全性能和解决办法，提供了安全可靠的多节点管理。

2.2.2 Ceph 的逻辑层次

下面从 Ceph 逻辑层次来看业务的分级，如下图 2-6 所示

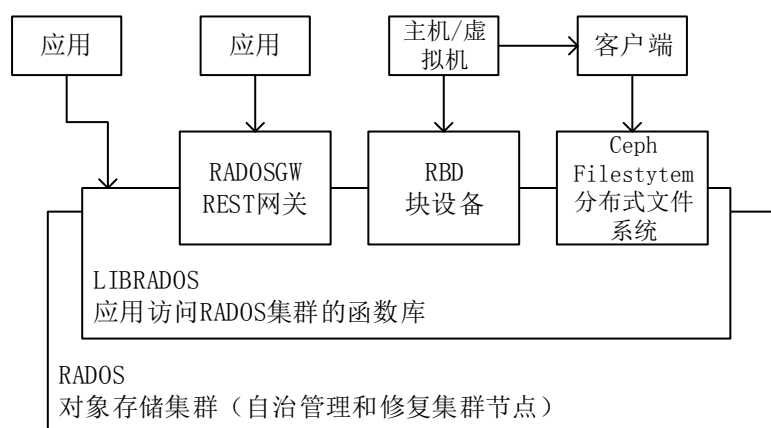


图 2-6 Ceph 逻辑层次结构图

自下而上 Ceph 分布式存储整个逻辑分为以下 4 层：

1. 对象存储系统 RADOS，所有数据都是由它来完成最终存储写入存储介质。而 Ceph 的特性基本都是由它展现，所以也成为了 Ceph 存储的基础与关键。RADOS 分布式对象存储集群由前面介绍的对象存储设备和集群监控组成。两者相互通信，检查存储节点状态，协同评估整个存储的状态，构成整个分布式存储系统的集群映射。客户端程序从 RADOS 与对象存储设备和集群监控的通信中确认集群映射，再进行利用算法计算数据的存储地址，然后直接访问对应地址的对象存储设备，完成任务。如果集群映射数据变化很小，客户端访问可以不通过元数据服务器进行，也不需要进行查询存储表，就能读

取数据，极大的节约处理时间。当对象存储设备出现故障或 RADOS 节点增加这两种情况才会导致集群映射的变化，而正常情况下，集群映射变化的频率也远低于客户端访问的频率^[29]。

2. RADOS 基础库-librados

librados 对底层组织和封装，并对外提供接口，所以可以利用 librados 对下层进行应用开发，librados 提供的接口调用限制在对象存储功能的范围。程序可从外部调用分布式存储系统上的 librados 接口，再通过套接字（socket）同 RADOS 集群中的存储节点传递消息并实行存储操作，极大地方便很多第三方使用者集成和二次开发^[30]。

3. 高级应用接口

应用接口包括了：RADOS 网关、可靠的块设备（Reliable Block Device，简称 RBD）和 Ceph 文件系统（Ceph File System），其作用是在 librados 库的基础上提供更抽象、更方便应用或客户端使用的上层接口。RADOS 网关设备可以与亚马逊 S3 和 OpenStack 中 Swift 组件兼容的 RESTful 接口进行通信。块设备拥有通用的块存储设备接口，云计算平台中使用它为虚拟机创建存储卷，在 KVM(Kernel-based Virtual Machine，基因内核的虚拟机)/QEMU(开源虚拟化仿真软件)中已经集成了块设备的驱动^[12]。Ceph 文件系统还并不成熟，所以暂时还未有用户作为产品使用^[31]。

4. 应用层

这一层主要用于适配外部调用，可扩展及二次开发 Ceph 内部组件及接口应用于新的功能。

以上就是 Ceph 分布式存储的整体介绍，下面将针对分布式存储都会面对的问题进行分析，为分布式存储测试方案的规划做好准备。

2.3 对比主流分布式存储技术

一个完整的虚拟化管理平台都需要对象存储，文件系统，块设备存储这三大存储服务，它们针对对应特有的应用场景。三种存储如果分别用三个软件来实现比较符合 Linux 操作系统的理念，不过如果在一套平台内同时维护 3 套存储服务，会造成比较大的资源争夺浪费，而且研发成本也比较高。根据上文对各种分布式存储技术的介绍，还有桌面虚拟化存储的要求，我们可以看到，主流的分布式开源系统中 Ceph 是唯一能同时提供三种存储方式服务的，Gluster 是比较成熟的分布式存储系统，但是只提供文件存储和对象存储服务，而其他分布式存储都是只提供其中一种服务。所以在虚拟化场景中没有

自主研发存储的情况下，开源分布式存储解决方案只能在占有率最大的 Ceph 和 Gluster 之间做选择，并且现在两种存储方案都被红帽公司所收购，所以根据红帽官方的推荐和社区中的推广，虚拟化厂商更愿意选择较有发展前景和丰富的社区资源的 Ceph 分布式存储，然而根据在大规模中的应用比较，无论在功能还是性能，Ceph 都有绝对的优势。基本信息对比如下表 2-2。

表 2-2 Ceph 同 GlusterFS 的差别

名称	Ceph	GlusterFS
架构	基于 RADOS 提供块、文件系统、对象等调用接口。系统由副本和数据分布组成。使用 hash 和 Crush 算法算地址，有元数据服务器	GlusterFS 提供文件系统 / 对象。用 hash 算法定位数据，hash 算法作用于存储池内的所有存储服务器。无元数据服务器。
数据分布	条带化数据到不同的节点。	无
默认块大小	64KB，可以调整至 1MB。	无
扩展性能	线性扩展	线性扩展
缓存/分层存储能力	Ceph 的文件日志可以写到 SSD 固态硬盘中，支持冷热数据分层和代理缓存。	正在开发中。
坏盘下系统重建能力	数据分布到多节点，多盘可以从完整的副本中接收数据，重建时间快，不增加单盘负载。	无
安装和管理	综合部署管理方法	cli 集群管理，可平滑升级

GlusterFS 分布式系统虽然因为组件中无元数据服务器，而避免了单点故障和性能出现瓶颈，提高了海量小文件应用性能，但是造成了数据一致性问题复杂，遍历性能弱，无法把控。使客户端负载过重，过度消耗节点本身的计算能力和内存。而且使用用户空间效率很低，数据需要多次与内核空间交换，而且地址转换树通常深达 10 层以上效率低。接口与数据易于使用，方便管理和节点间迁移。但是因为数据以原生方式存储的，可以直接拷贝读取，这在安全性上是很大的漏洞，而且这依赖于操作系统对文件系统的局限。GlusterFS 采用拷贝技术提供高可用性，其空间利用率为复制数的倒数，所以造成存储利用率低下。

Ceph 在很多情况下都超过了 GlusterFS，更加符合了桌面虚拟化的性能和功能需求，如上文介绍，Ceph 对块存储的支持，利于大文件的使用然而在桌面虚拟化中，分布式存储在虚拟化平台中的主要作用就是存储操作系统镜像文件，一个虚拟机镜像往往有 5-15G 大小，并且还需要增加逻辑卷作为数据补充，所以大文件的读写性能比小文件读

写更加重要，而桌面虚拟化对 IOPS 的敏感和弹性要求，Ceph 分多副本多条带提高带宽和性能，而且块设备直接与物理设备相同，能给虚拟机提供高利用率的性能。而且 Ceph 拥有元数据服务器，在一致性上性能较高。一个计算机存储系统，缓存层的设计直接影响性能，Ceph 操作日志能够存放在高效率的固态介质存储设备中。而且可以对数据进行缓存或冷热数据分层，Ceph 在恢复损坏的磁盘时也更有优势，与 GlusterFS 不同的是数据分散在多个节点里做法，Ceph 可以有更多的磁盘能够同时输入副本数据。极大减少数据重建的时间，避免单盘负载过大。在集群规模越大时，优势越明显。而且从开源代码可维护性和红帽着力推荐 Ceph 来看，Ceph 能获得更大的商业及社区支持，有利于后续开发。

2.4 本章小结

本章节描述了分布式文件存储系统的原理，然后就分布式文件系统的核心概念和技术实现进行分析。对 Ceph 分布式文件系统的做出分析并说明分布式存储存在的问题和异常，以及最终为何选择了 Ceph 作为桌面虚拟化后端存储提供者。

第三章 分布式存储测试需求分析

3.1 应用场景

桌面虚拟化的应用一般常见于教育行业中的电子教室与多媒体教学系统，将传统个人主机更换成可统一管控瘦客户端，降低维护成本，网络维护简单，防病毒安全性高，可批量创建的操作系统与应用程序模板更换灵活快速，无需本地磁盘安装多系统与还原卡，数据统一置于数据中心，教师与学生的个人虚拟桌面可在课室宿舍办公室中安全登录使用，与排课、学生信息等业务系统衔接，极大的方便了教学、管理、备课、上机、实验，云平台弹性地提供各种资源，充分地提高了硬件设备的使用率。IT 企业和其他对桌面应用服务相互切换要求较为频繁和多操作系统工作的场景，为了能集中管理数据或者移动办公需要，也同样地需要桌面虚拟化来提供安全保障，而这些应用领域可以看出，桌面虚拟化应用场景剧有以下几个关键点：

- 1.并发访问大，因为应用场景普遍有规律的工作时间和业务安排，会有大批量的虚拟机会集中地在某一时刻同时启动和运行；
- 2.复杂业务环境多变，每台虚拟机运行不同的应用软件，但是计算能力和存储能力会一致的向下宿主机申请资源，造成性能资源弹性供应提供利用率，但是就会对存储的随机读写能力要求高；
- 3.桌面虚拟化的安全，虚拟机或者宿主机损坏或者错误时，存储可以多副本安全恢复，甚至迅速迁移至别的宿主，不会造成业务中断，并且存储数据不会留存在客户连接终端，统一存储在后端，为企业资源安全 and 数据管理提供保障，这一点要求使分布式存储成为桌面虚拟化的最优解决方案。

而桌面虚拟化对存储的特殊要求体现在整个虚拟机的生命周期中，所以必须对每一个阶段进行详细的分解。而根据桌面虚拟化对存储的需求的特点，可以有针对性的测试出分布式存储的瓶颈。在以上工作场景中分布式存储的具体应用也发生了改变并进行详细描述：

- 1.第一个特点是不同于传统存储的访问方式。一台物理服务器上通过虚拟化技术虚拟出大量的虚拟机，通过虚拟机管理平台 Nova 对 Cinder 管理存储资源池分配创建并指定存储路径，分布式存储后端通过操作系统连接访问到物理服务器的真实存储系统上，返回的块存储数据与镜像结合成为虚拟机使用的存储服务，相比传统直接由计算机操作

系统访问硬件存储，多了一层分布式存储系统，并且需要管理和调度才能进行使用，而且在传统计算机系统只有一套系统去调用单一存储，而现在是多个虚拟化的操作系统虚拟机对统一存储进行访问，加大了并发量，所以第一个需求就是对分布式存储的并发访问的性能需求非常大。虚拟机访问 Ceph 存储的逻辑如下图 3-1。

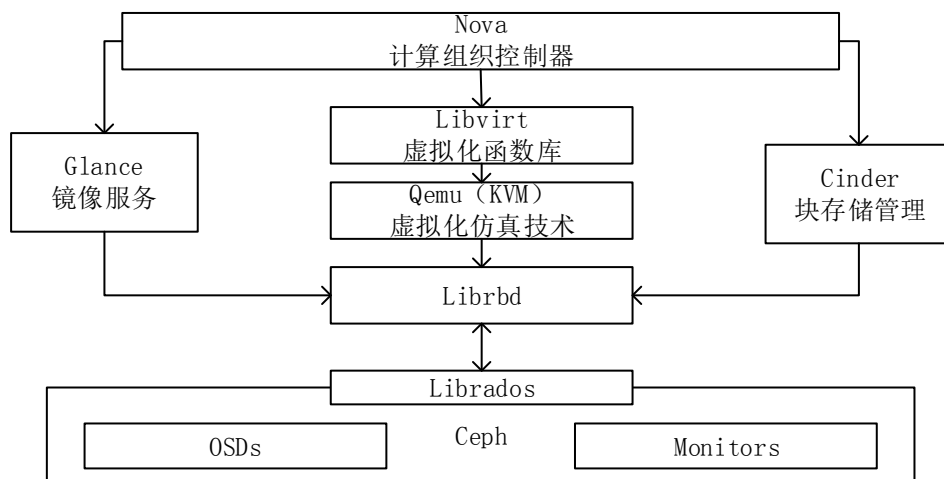


图 3-1 虚拟机访问 Ceph 存储逻辑结构图

2.第二个特点是这些虚拟机根据需要启用不同软件和环境配置的基准镜像，但是会有大量的虚拟机基本都采用相同的操作系统，这就造成许多用户的虚拟机启动时会读取许多重复数据，譬如启动相同的操作系统时，相同的系统文件与配置、统一部署的软件系统，如果不加识别，虚拟机与存储会频繁的访问传输相同的数据块，由于操作系统本身容量庞大，大量数据读写频繁必将急剧地增加存储 I/O 压力，当达到 IOPS 性能瓶颈时就会产生虚拟机启动风暴，拖长了启动登录时间，甚至是进入锁死状态，直接带来不良的客户体验。常见的虚拟桌面如微软的 Windows7 以及之前的桌面操作系统在设计时，并没有考虑现有云计算虚拟化的等新的计算机提供方式的支持，在桌面虚拟化中每个独立的 Windows 操作系统虚拟机都会争抢消耗大量的磁盘 IO 和吞吐量。由于桌面虚拟化，虚拟机相互之间做出不可感知的资源隔离，因此虚拟化平台需要虚拟出较多数量的 Windows 桌面操作系统时，所有操作系统使用传统存储（SAN 或 NAS）设备时，由于相互隔离无法感知对方，误认为是物理计算机并相互争夺磁盘 IO 资源，以满足尽可能最大限度地满足自身对硬件资源需求。而在统一平台中，多个虚拟机在同一宿主主机中，后端使用同一存储，所以要对 IOPS 的峰值和平均值进行测量，避免极大争夺，保证每台虚拟机的 IO 资源均等或者按级分配。如果 Windows 7 桌面操作系统的因此而造成的 IO 性能受到限制，用户的桌面体验的性能就会随之降低，用户将发现操作系统启动和登录时间越来越长，所需的应用程序也运行卡顿，由于桌面虚拟化性能低下，原本

正常的工作业务节奏变慢而被用户排斥。所以一般个人计算机现在采用大 IOPS 的 SSD (Solid State Drives) 固态硬盘来做操作系统启动盘, 加快系统和应用程序的启动速度。所以解决相同数据块的缓存, 解决 IOPS 暴增瓶颈成为第二个需求。

3.第三个特点是桌面虚拟机 I/O 变化特点, 在虚拟机启动和登录时, 由于集群中一个节点存在很多虚拟机, 所以相当于将多个操作系统的 IOPS 需求聚集在了一起, 由于操作系统本身容量庞大, 大量数据读写频繁急剧地增加存储 I/O 压力, 当达到 IOPS 性能瓶颈时就会产生虚拟机启动风暴, 拖长了启动登录时间, 甚至是进入锁死状态。需要对虚拟机启动时的 IOPS 进行测量, 并且统计启动登录时间。登录后, 用户会根据工作需要使用不同的办公软件或者 ERP(Enterprise Resource Planning,企业资源计划系统)与 OA(Office Automation,办公自动化)辅助系统, 有的客户还需要专业型的生产力工具及开发编程软件, 需要满足多种业务负载模式, 在实际使用中用户负载在稳态时对引荐资源的需求中表中所描述的用户负载分为 4 种:

轻度用户: 运行通用的办公软件, 无多媒体和其他应用软件需求

正常用户: 除包含轻度用户需求外, 使用 ERP 和 OA 系统协助工作, 使用浏览器。

全能用户: 除包含正常用户需求外, 需要多媒体播放, 运行较多应用程序

重度用户: 除包含全能用户需求外, 还需要使用大型工程类软件辅助工作。

在打开应用软件时进行对软件文件本身顺序读写,这时考验存储的吞吐量, 在使用中产生碎片型的随机读写, 在这期间磁盘带宽会闲置下来, 降低对存储带宽需求, 变化为读写性能需要, 转化为运行中的虚拟机对存储 IOPS 性能的要求。所以须测试出虚拟机 I/O 特点即整个生命周期的 I/O 数据作为整个测试和需求的基准, 然后才能分析出分布式存储上提供的瞬时性能是否能满足虚拟机的应用需求。

4.第四个就是符合云计算的特点: 按照用户需要弹性提供资源和服务。因为用户的规模和数据量是动态变化的, 在前期方案制定后, 实际上线之后, 用户会根据需要提供更大的存储容量, 兼容更多的存储设备, 增加存储性能, 或者是加大存储的备份数来获得更高的安全保障, 所以存储系统可以动态的扩展, 而不必中断业务。但是云计算中暂时虚拟机还未能达到跨节点使用资源, 所以虚拟机对计算节点和存储节点, 在单台物理服务器上其实还是有相对的绑定关系, 所以在建设之初或者扩展规模时, 计算资源以及存储资源相互制约, 决定了解物理服务器的虚拟机实际承载能力, 所以为了便于桌面虚拟化建设方案的规划, 以及分布式存储在桌面虚拟化应用中的实际意义, 探索一定硬件规模下满足使用需求的虚拟机数量是至关重要的。

3.2 分布式存储测试需求分析

分布式存储可以为许多应用提供存储服务，但是其不同场景下所需要的功能和性能的不同。所以下文将会根据在桌面虚拟化这一应用场景中，有针对性的对分布式存储服务能力进行需求分析以及测试。

3.2.1 桌面虚拟化对分布式存储服务的需求分析

所以在桌面虚拟化应用场景中测试分布式存储需要根据上述桌面虚拟化对存储业务场景的需求特点和工作方式，进行相应细致的性能需求分析。下文均以 OpenStack 虚拟化管理平台和 Qemu(KVM)技术虚拟出微软 Windows7 操作系统的虚拟机为例，分析虚拟机对存储的所需性能及功能。

3.2.1.1 提取需求分析要点

桌面虚拟化既要在加载基准镜像操作系统时的大数据量顺序读写，又要满足运行时的小数据块随机读写，所以根据上述桌面虚拟化的存储特点，总结出的测试需求要点有以下几个：

- 1) 在本地存储的情况下确定虚拟机的开机 IOPS 性能需求，其中虚拟机启动时间按照常规机械硬盘的不大于 1 分钟，为用户可接受的较快启动时间，正常启动一般在 2 分钟左右，所需要对这两种启动时间进行 IOPS 模拟控制，控制启动时间测试所需 IOPS 大小。

表 3-1 虚拟机启动时间需求

Windows 7	快速启动	正常启动
开机时间	<1 分钟	<2 分钟

- 2) 对不同用户负载下的 IOPS 和存储带宽进行性能需求如下表

表 3-2 虚拟机不同应用负载需求

用户负载	操作系统	使用软件	vCPU 用量	虚拟内存用量
轻度	Windows 7	Office	1	1GB
正常	Windows 7	ERP、chrome	2	2GB
全能	Windows 7	Media	4	4GB
重度	Windows 7	AutoCAD	8	8GB

- 3) 对比本地存储与使用 Ceph 分布式存储提供的块存储性能。
- 4) 根据 1、2 项测试结果，保障虚拟机开机（IOPS）性能情况下，Ceph 分布式存储可提供的最大虚拟机数量，并且获得最优性能的临界数量。

- 5) 在一定的硬件规模条件下,保障虚拟机和应用负载时高并发的数据吞吐性能情况下,Ceph 分布式存储可提供的最大虚拟机数量,并且获得最优性能的临界数量。
- 6) Ceph 存储系统的灾难恢复能力的高可用性,在多副本时其中一个 OSD 或主机节点下线时,验证存储文件完整。
- 7) Ceph 分布式系统的弹性扩展能力。增加删除 OSD 节点,实现动态扩容。

3.2.1.2 根据需求点建立测试业务模型

得到性能测试需求要点后,根据桌面虚拟化系统的实际使用情况,我们可以进行测试模型的建立,也就是建模。我们需要首先测试本地磁盘和分布式存储提供的块存储两个种存储的基本性能作为底层存储基准,然后在本地存储的情况下测试单一台虚拟机对存储的性能需求基准,并在多个应用负载下的性能差异,由此得到虚拟机的存储性能需求基准,最后通过这些性能基准,从体验和高并发两方面对整个使用分布式存储的桌面虚拟化集群进行统一测试,最终确定虚拟机存储的性能和数量的临界值。如果用流程图表示,则可表示为图 3-2。

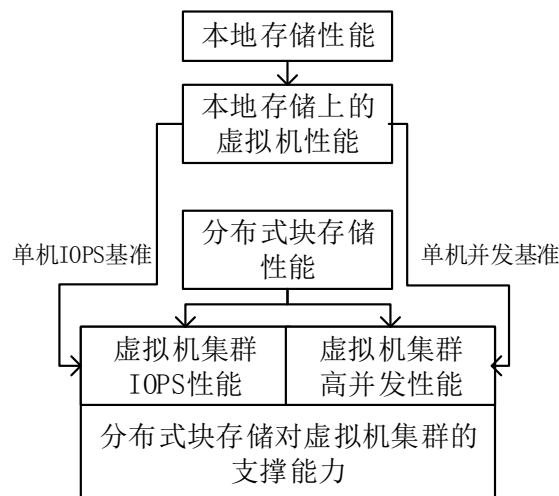


图 3-2 测试流程图

虚拟机在整个生命周期中,由于操作系统在启动关闭时无法再虚拟机内部使用软件检测,所以采集性能数据需要内外两部分一起收集,具体流程如下图 3-3:

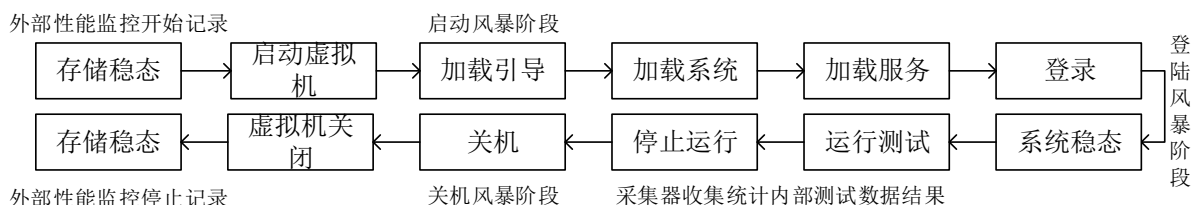


图 3-3 虚拟机生命周期流程图

3.2.2 硬件需求分析

Ceph 分布式存储系统可以在多种类型的通用硬件上运行，通过各种方式的网络连接，完成一个高性能和安全保障的存储系统集群。在实施部署 Ceph 分布式存储之前，需要综合考虑整个方案的硬件配置，因为如果只有两个节点以下或者网络达不到千兆的情况下，并不能体现分布式存储安全性和性能的优势，而且服务器的硬件也会对上层的软件甚至操作系统产生限制和冲突，所以需要明确硬件配置，从 Ceph 的每个组件出发，估计出节点服务器的计算，存储、网络等硬件资源的需求，探知满足系统运行的要求的最低硬件要求，或者根据测试给出推荐的硬件配置。其中最低需求来自于 Ceph 的官方指导文件中的理论值，本文需要确定在特定的桌面虚拟化场景中的平衡数据。

1. 计算能力需求分析

分布式存储工作时，工作负载至散布各个节点，属于计算密集型节点，所以需要比较强的计算能力，按照 Ceph 官方要求采用双核或以上性能的中央处理器^[16]。而对象存储设备节点构成的集群上面寻找存储数据的地址还要同步一致性的版本，需要 CPU 性能。监控节点只有简单的维护集群映射的主副本，所以不需要太强的计算能力。除此之外还应该考虑到主机节点上桌面虚拟化所需要的 QEMU、Libvirt（虚拟化工具函数库）等进程是计算密集型的进程。如果 Ceph 分布式存储集群的节点同时既做客户端又做对象存储设备，并且上面运行了拟机，所以相当消耗计算资源。由于客观情况约束，客户开始选择高密度的超融合的机器，将存储和虚拟化平台全部融合在一起，比如四节点机架式高密度服务器，既节省数据中心的机位空间，又降低成本。

2. 数据存储需求分析

底层存储设备受物理性能约束，直接局限了存储系统能力，但是依靠分布式存储集群中因为根据算法进行分布，越多的存储设备越能够体现分布式存储的性能和安全性优势，降低物理性能约束，但是要综合考虑存储磁盘盘位也就是存储控制器数量和存储性能的相互影响。而且还有诸多正常情况下的存储网络等设备的损耗。所以分布式存储系统的构建时候会根据这些性能损失点对冷热数据增加缓存层交换，或者增加存储缓存代理预载入。

固态介质的存储设备在由于其强劲的性能得到青睐，存储厂商也开始主推各种固态存储产品，传统的高性能 RAID 卡厂商 LSI 都为此研发了专门的融合固态硬盘做缓存的优化软件。固态介质的存储设备特点是随机读取性能高延迟小，达到万级的 IOPS 性能

让传统硬盘望尘莫及，但是伴随而来的容量偏小和价格昂贵两个问题也一直制约着大规模使用，但是在分布式存储系统方案制定的时候，通过固态介质的存储设备来做缓存或者数据交换量大的节点以满足性能要求高的关键业务。下表 3-3 对比了不同存储介质的性能价格等差异。

表 3-3 不同存储设备对比表

类别	每秒读写次数 (IOPS)	每 GB 价 格 (元)	顺序读取	顺序写入
SSD 盘	70000	10	400MB/s	160MB/s
SAS 磁盘	180/300	5	200MB/s	190MB/s
SATA 磁盘	90	0.5	170MB/s	150MB/s

如表中所示，固态硬盘单位容量存储成本比 SAS 或者 SATA 磁盘都要高出很多，但是固态介质存储设备的 IOPS 性能强劲，性价比高，适合对性能要求高的关键业务场景。

为避免对象存储设备和操作系统对硬盘性能资源的争夺，因此为操作系统、Ceph 日志和数据盘分开在三个盘上保证性能和测试的准确性。

4. 网络部署需求

网络的部署与规划直接影响到分布式存储集群的性能表现。每个存储节点的网络设备性能需要达到处理本地对象存储设备节点产生的数据能力。对象存储设备集群利用网络进行读写操作，同步副本。新节点上线或者故障节点下线，进行数据迁移以存储系统分布平衡。节点间的通信操作会形成相应的带宽性能影响，加大通信链路网络的业务需求，而且在虚拟化桌面应用中，大量的压缩的桌面图像视频流量也占用了大量的带宽。所以需要分网，内部一条网络给 OSD 与客户端通讯，一条分为各存储节点间 OSD 集群间通信，一条为虚拟机对外服务通信接口。由于一般服务器只有两个或以上千兆端口，所以只有采取网卡虚拟化，将多个网口做绑定后，负载均衡动态分布流量给三条链路，并使用 Qos 技术，限制最小带宽和根据重要性调整策略，尽量减少瓶颈。

5. 集群配置需求

分布式系统涉及的软硬件种类很广，Ceph 分布式存储系统中可配置调整的参数也非常多，因此如何进行配置才能使系统性能达到最优化成为了测试的关键。而且相互之间性能关联紧密，Ceph 分布式存储将存储数据以对象的结构存放到对象存储设备资源中，归置组中的 object 分布到一组对象存储设备中，所以对象存储集群和归置组的配置都可能会直接影响性能。

3.3 测试方案与测试用例

本节根据前几节对 Ceph 分布式存储特点、桌面虚拟化对存储的性能需求分析为测试方案包含：测试目的、测试的方法、测试环境的规划、测试工具的设计和选择、测试用例的设计方法、测试代码的设计方案。

3.3.1 测试流程

在使用分布式存储的时候，为了更加符合桌面虚拟化要求，需要充分了解分布式存储性能特点和桌面虚拟化性能需求,这样才能提出符合要求的测试方案和测试模型,测试过程中需要用严谨的业务流程方案来规范，也是测试能达到预期效果的保证，该方案来自于工作中的长期研发，并参考其他先进范例综合而成。

Ceph 作为一个全新的分布式存储系统，其系统的性能、安全性和扩展性都需要一个区别于传统物理存储设备的测试方法来验证，通过测试分析可以不断的优化 Ceph 本身的不足，并能为以后判断监控系统运行状态和运用在其他领域做出技术铺垫。测试分为存储测试需求分析、计划、实施、数据分析与系统优化。具体流程如下图：

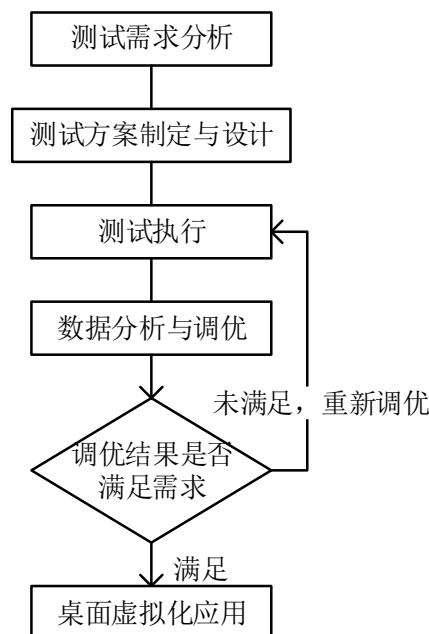


图 3-4 测试流程图

1. 测试需求分析

确定测试的对象和目的。分析 Ceph 分布式存储的逻辑构造、功能性能和使用环境场景,提取需求点，分析性能测试的要求以及理论的系统性能极限得到最终测试目标，从

而制定出详细的测试策略。

2. 测试方案制定与设计

分析测试中使用的方法。依照前面的得出的结论，指定有效可行的测试流程计划，进行工程化过程形成方案，以标准的软件工程测试方法对整个测试做出细化指导。测试方案是测试工作的根本。依照制定的方案中的桌面虚拟化应用场景设计具体的测试用例，使用与之对应的测试软件程序采集测试数据等。

3. 测试执行

根据测试用例使用测试软件进行分布式存储的测试。部署安装测试环境、编写测试脚本代码或调整软件、在桌面虚拟化应用下进行测试采集数据。测试执行阶段流程图如下：

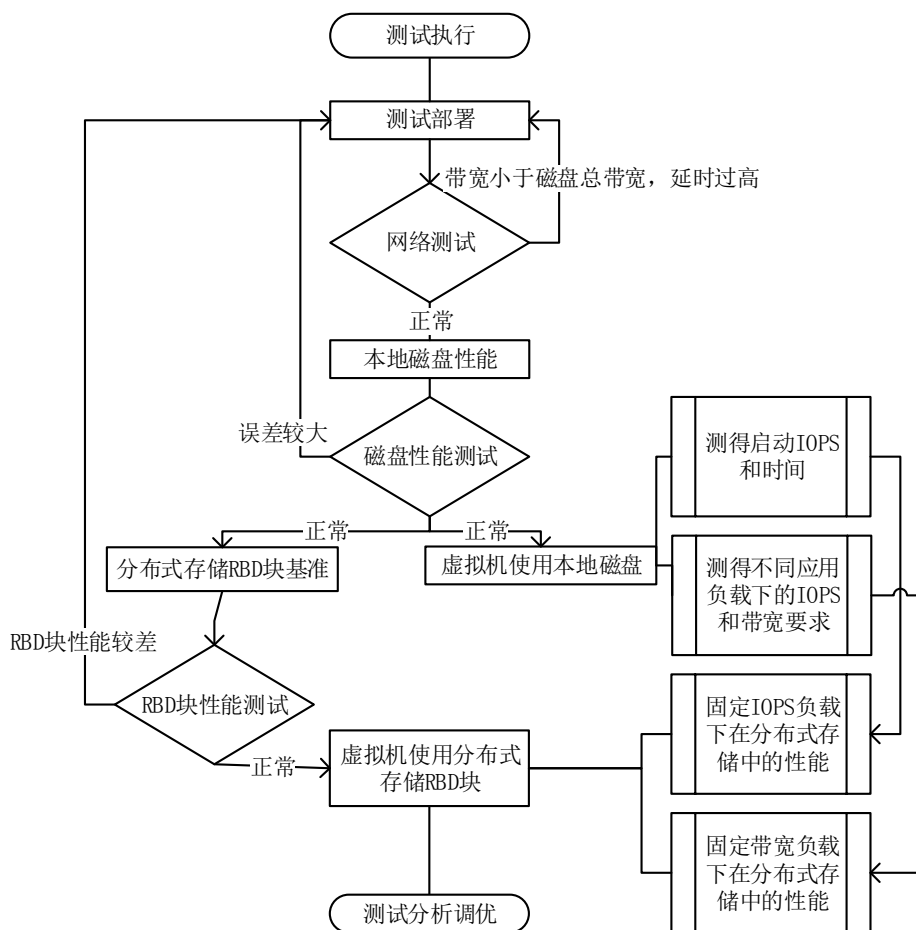


图 3-5 测试执行阶段流程

4. 测试数据分析与系统调整优化

在测试完成后对得到的数据完成归纳、校验测试记录，归集生成有效的测试结果，再按照测试需求和目标完成最终的测试分析报告。在分析测试结果时，根据其中的不足

进行调整修改。最终确定分布式存储最佳性能调节参数，寻找存储不同参数对结果影响的变化规律，找出性能的制约因素，最终提高存储性能和应用场景的适应性。

3.3.2 测试方法

分布式存储测试在结合 SPC 和 SPEC 等测试组织的基准测试方法，根据开源测试软件的特点和用法，总结出以下主要的测试内容，根据以下工具的组合可以得到较为完整的测试方案。由于应用场景固定桌面虚拟化，而且针对的是制约桌面虚拟化体验的核心问题：启动风暴问题，所以测试方向主要集中在非功能性测试。

1. 性能测试

性能测试评估一个分布式存储系统的最关键的指标，同时也是对测试技术和方案提出新要求的领域，用户一般都直接根据性能测试结果直接决定是否采用分布式存储，但是根据分布式存储系统在不同场景下的性能表现又各不相同，只能通过分布式存储系统对比传统存储测试基准来评估是否适合特定的应用场景，也可以为分布式存储系统性能调优提供依据，本文的核心思想也在于此。分布式存储系统在桌面虚拟化应用场景中性能测试部分包括 IOPS（Input Output Per Second, 每秒并发操作数）和吞吐量，分别是对存储系统的频繁小文件读写、并发数据吞吐的处理能力进行记录。这两项的测试由自动化测试软件进行，并且要从存储系统和虚拟机内部镜像两个方面展开主要包括小数据频繁读写、并发大数据存取、虚拟桌面镜像拷贝快照，虚拟桌面启动登陆风暴等场景下的 IOPS、吞吐量。其中包括：

- 1) 本地磁盘性能测试，检查本地磁盘性能是否正常。
- 2) 桌面虚拟化在本地存储的性能，以此对照采用分布式存储的虚拟机性能差别，并同时测得在不同应用负载下存储需求的差别，便于后期使用测试软件进行稳定的模拟。
- 3) 测试分布式存储提供的块设备 RBD 的基准性能，检查是否正常，根据结果调整最优参数。

以上测试完成后，即可对整个系统进行压力测试，检测支撑能力，并找到瓶颈以及应用调优的关键参数。

2. 压力测试

在分布式存储在高工作负载下，测试分布式存储节点的连接性能，IOPS 性能，吞吐量压力性能，监控分布式存储与虚拟机运行情况，记录存储系统与网络资源消耗情况，

从而为真实生产环境运营维护提供极限的实验依据。这项测试自动化方式进行，使用测试软件对存储系统和虚拟机内部进行负荷模拟，同时使用功能测试的方法来验证功能的完整性，并监视 Linux 性能，记录计算和存储、网络等相关的统计信息。对于桌面虚拟化场景中影响用户体验的关键就在于因存储性能影响的启动风暴问题，通过对分布式存储性能的测试分析，找出分布式存储或者虚拟化技术对该问题的影响因子，不断调整可控参数，最终总结分布式存储调优方法，完成针对性的性能优化的目标。其中定量测试，所以控制了两个参数的具体步骤：

- 1) IOPS，随机写 IO 情况下：固定为本地存储虚拟机性能测得的 IOPS 最高值，测得最大能启动虚拟机的数量。
- 2) 吞吐量，顺序写 IO：固定 MBPS（吞吐率）为本地存储虚拟机性能测得的吞吐率均值，测得最大能启动虚拟机的数量。

经过上述压力测试，可以均衡得到避免启动风暴的最大支撑能力，以及在虚拟机持续读写情况下，分布式存储的性能。

3. 功能测试

扩展能力作为分布式存储主要特点，也是存储系统测试的一部分。

验证弹性扩展能力。对正在工作的存储节点，进行增加硬盘操作，实现动态扩容，分布式存储系统通过后台复制协议将数据同步到新增存储节点。

- 1) 在每个节点中插入一块新硬盘，在原有 Ceph 正常工作的同时系统挂载后，更新 Ceph 配置使其成为新的 OSD 节点，加入到存储集群中。
- 2) 检查 Ceph 总体容量是否相应上升，进行在线扩容功能的验证

4. 异常测试

无单点故障的高可用性作为分布式存储突出特点，当某个副本所在的存储节点发生故障时，分布式存储系统能够将存储服务自动的切换到其他的副本从而保证数据的可用性。分布式存储系统通过复制协议将数据同步到多个存储节点，并确保多个副本之间的数据一致性。验证分布式存储的异常测试,以下为具体步骤：

- 1) 通过将由多对象组成的块存储映射到某一挂载卷上，格式化为可用文件系统后进行读写。
- 2) 下线副本映射表中对应的一个 OSD 后，校验文件是否与下线前一致。

3.3.4 测试环境规划

因为日常实验和研发环境系统中的资源（CPU、内存、磁盘、网络）是有限的，与大规模部署测试不同，前期只能通过小规模的性能测试分析根据规律找出可能出现的资源瓶颈做出优化，虽然性能测试分析的结果不一定精确的，但通过软件模拟和估算的结果与实际值相比较，确定最终的性能规律可以符合科学对误差的允许范围内，为大规模测试和产品上线做出性能预计。

本次测试物理硬件规划上根据需求和实际生产环境，采用市面上的超融合高密度机架式服务器，是分布式存储与计算节点整合的高性价比形式，与生产环境吻合。单个 2U 机架高度的服务器中含有 4 个独立的双路服务计算节点，多于 3 个盘位用以满足分布式存储安全性的要求，网卡同时含有万兆和千兆两种接口，将存储网络和通信连接网络可相互隔离以保证通信链路不产生干扰，足够的带宽也保证了不会成为性能瓶颈，计算能力等也充分保证了和体现了真实生产环境，配套的网络设备也提供相应连接能力即可保证。硬件规划如下图：

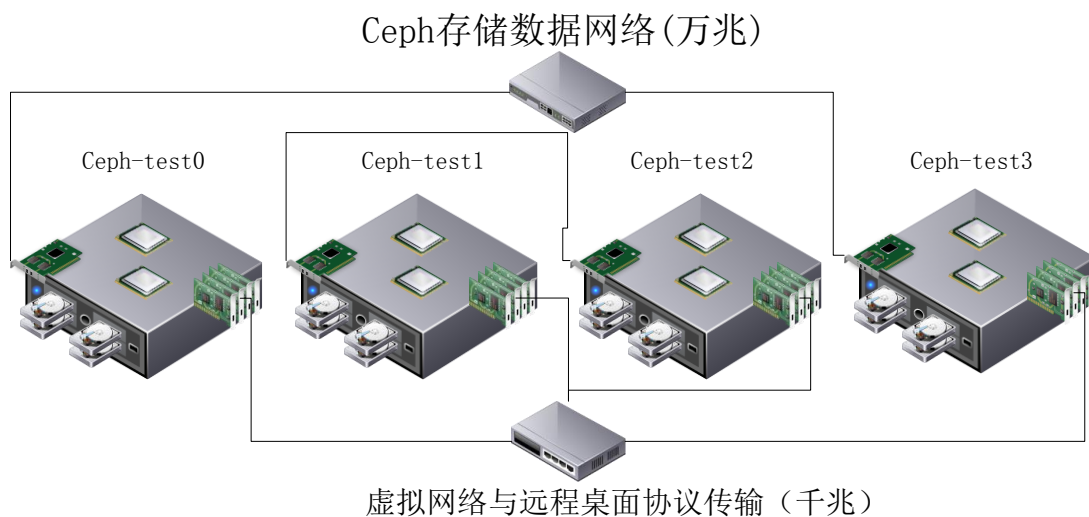


图 3-6 硬件测试环境规划示意图

软件环境规划根据需求和实际生产环境，底层采用红帽的 CentOS 6.5 64Bit 系统，分布式存储 Ceph，虚拟化部件 libvirt, qemu，虚拟化机操作系统 Windows 7，虚拟化设备驱动 Virtio，虚拟化管理平台 OpenStack。其结构图如下：

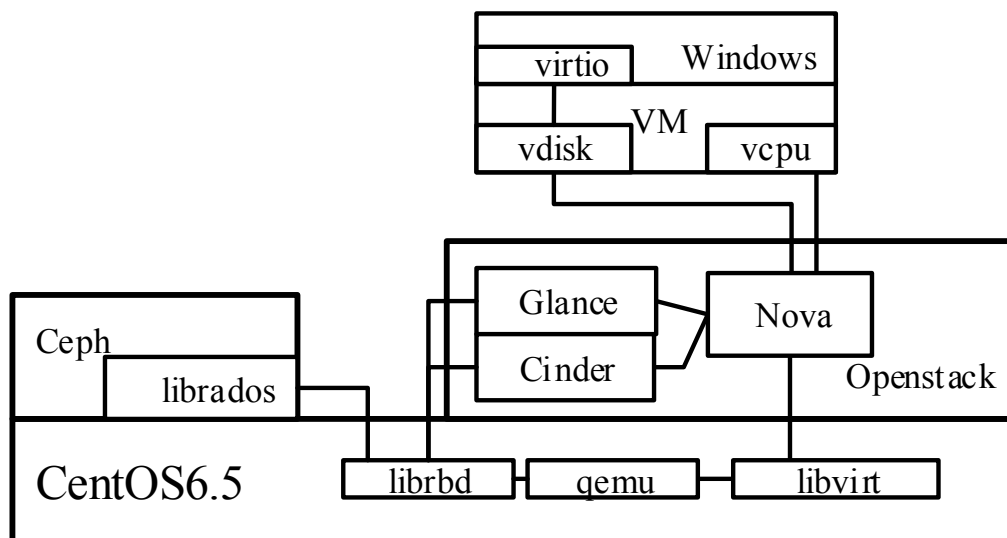


图 3-7 软件环境架构图

3.3.5 测试工具的设计和选择

测试工具由于整个层次跨越了本地磁盘，服务器操作系统，分布式存储软件，虚拟机四层。所以需要每一层都进行相应的监控，以发现性能衰减或者异常的位置。首先使用 IPerf 先对网络进行带宽等性能测试验证，检查是否存在瓶颈。然后在本地磁和服务器操作系统上使用常用 fio 进行读写操作记录 IOPS 性能，同时采用分布式存储 monitor 检测各个 OSD 的性能和集群的存储状态，在虚拟机内部采用 iometer 进行存储负载的模拟和记录。最后通过数据汇总进行可视化，做出判断和测试结果统计。

3.3.6 Ceph 性能测试用例设计方法

Ceph 底层为客户端开放了一组服务 API，由于可以为虚拟机提供虚拟磁盘的块存储，Ceph 的 RBD 块存储现在是 Ceph 整个分布式存储里面用户使用和关注度最高的组件，而桌面虚拟级作为客户端直接使用 RBD 块存储，所以基于桌面虚拟化应用场景，本次研究只针对块存储接口进行用例分析。基于块的存储接口将一个块转化为一个字节序列，这是最常见的存储数据的方式，所以块设备接口使得虚拟块设备成为了与大型分布式存储系统交互的理想方式。

需要测试硬件在操作系统下本身的性能，来确定系统和硬件本身没有问题，确定基准的存储性能，然后测试 Ceph RBD 块设备在物理机器上的性能，客户端向 RBD 块设备接口做文件的读取与写入操作，记录对 RBD 裸块设备以及用 Linux 常见的文件系

统格式和红帽操作系统中的文件系统的性能。然后接入桌面虚拟机，在虚拟机内和操作系统测试 IOPS 值和吞吐量，观察是否存在性能瓶颈，修改块设备参数，找到最优化的解决方案。最后将虚拟机的虚拟硬盘换为 RBD 块设备在虚拟机内进行测试，观察性能瓶颈与裸设备的特点是否相似，并根据特点调优参数，得到解决方案。

3.3.7 测试代码的设计方案

1.测试网络

通过 IPerf 工具进行网络信息收集与测试^[18]。对 TCP/UDP 的带宽进行检查，探寻通信延迟和链路中的丢包概率，同时对其他网络设备性能进行检测，四个节点相互进行测试验证，并记录带宽。

```
iperf -c Ceph-test0 -t30
```

```
iperf -c Ceph-test1 -t30
```

```
iperf -c Ceph-test2 -t30
```

```
iperf -c Ceph-test3 -t30
```

2.测试本地磁盘

而服务器一般使用的基于 Linux 的操作系统，会自动开启 4K 大小的缓存页。在操作系统执行存取操作时，会预先缓存数据逻辑，调用文件系统同步命令将数据同步回去，将页面缓存写入长期性存储设备中，加快读写性能。所以测试时需要关闭页面缓存以免影响物理设备的真实性能使用参数 `oflag=direct.sync`，将读写数据直接在存储设备上操作。网存储设备上写数据，按照测试用例测试块大小为 4M。

```
dd if=/dev/zero of=/dev/sdc bs=4M count=1000 oflag=direct.sync
```

接下来使用 FIO 对硬盘进行读写 IOPS 测试，在 fio 使用中会有几个参数，`ioengine` 后的参数代表使用 `libaio` 异步 IO 库函数发起 IO 引擎的请求；`bs` 后面的参数为每次读写的块大小；`direct` 代表直写如存储设备，不落在操作系统 Cache，所以设置为 1 开启；`rw` 代表读写模式，测试中使用随机读；`size` 代表寻址空间范围大小，一般设置为硬盘的大小表示整个磁盘空间都能被用于寻址；`filename` 后面为数据路径；`iodepth` 代表 IO 的请求队列深度；`runtime` 测试时长为 60 秒

1) 随机读:

```
fio -ioengine=libaio -bs=4k -direct=1 -thread -rw=randread -size=100G
=filename=/dev/vdb -ame="EBS 4KB randread test" -iodeph=9 -runtime=60
```


2) 随机写:

```
fio -ioengine=libaio -bs=512k -direct=1 -thread -rw=randwrite -size=100G
=filename=/dev/vdb -ame="EBS 512KB randwrite test" -iodeph=64 -runtime=60
```

3) 顺序写:

```
fio -ioengine=libaio -bs=512k -direct=1 -thread -rw=write -size=100G
=filename=/dev/vdb -ame="EBS 512KB seqwrite test" -iodeph=64 -runtime=60
```

3.进行 ceph 块设备测试

对 ceph 块设备进行测试,使用不同的文件系统,进行对比测试后选用最佳性能的文件系统性能作为基准,并且客户端使用不同数量的 RBD 块来测试性能验证最佳的配置。

1) 随机读:

```
fio =filename=/mnt/rbd0/tt -bs=4k -direct=1 -rw=randread -size=30G
-name="Ceph's rbd 4K randread test" -numjobs=64 -runtime=120
```

2) 随机写:

```
fio =filename=/mnt/rbd0/tt -bs=4k -direct=1 -rw=randwrite -size=30G -name="Ceph's
rbd 4K randwrite test" -numjobs=64 -runtime=120
```

3) 随机读写:

```
fio =filename=/mnt/rbd0/tt -bs=4k -direct=1 -rw=randrw -size=30G -name="Ceph's rbd 4K
randread&write test" -numjobs=64 -runtime=120
```

3.3.8 Ceph 性能测试用例设计

Ceph 的整个测试用例分三个部分,首先使用本地磁盘做正常底层操作系统级别的存储性能测试,测出数据成为存储性能的基准。然后测 Ceph RBD 分布式存储自身的块设备性能,确定裸块存储设备性能,最后测试用 Ceph 块设备做桌面虚拟机虚拟磁盘的性能。

1.测试本地磁盘性能用例设计

确定测试环境后,在已有的操作系统上进行存储性能测试,主要关注随机读写下的 IOPS 和顺序读写下的吞吐量。使用 dd 和 fio 工具对存储进行顺读写和随机读写的测试,记录 IOPS 和吞吐量,测试用例如表 3-4

表 3-4 本地磁盘性能测试用例

磁盘类型	IO 大小	吞吐量	完成测试内容
SATA 硬盘	4M	最大	顺序读
SATA 硬盘	4M	最大	顺序写
	IO 大小	IOPS	
SATA 硬盘	4K	null	随机读
SATA 硬盘	4K	null	随机写

2. 桌面虚拟化在本地存储的性能用例设计

桌面操作系统对传统存储的要求一般是从 IOPS 和容量的两个方面，从本地存储直接提供给虚拟机使用，确定所需其他资源不成为虚拟机瓶颈，同时以作为与使用分布式存储的虚拟机作为对比参考。

首先测定虚拟机开机所需资源，以防止其他计算资源成为瓶颈。记录所需资源表测试用例如下：

表 3-5 Windows 桌面操作系统的虚拟机开机所需资源

Windows 7	快速启动	正常启动
开机时间	<1 分钟	<2 分钟
IOPS		

通过自动化脚本测试上述软件，并同时采用 iometer 工具记录相关 IOPS 统计得出各种负载所需 IOPS 性能和带宽峰值，如测试用例表 3-6：

表 3-6 不同用户负载下对硬件资源的需求

用户负载	操作系统	vCPU 用量	内存用量	平均 IOPS	带宽峰值
轻度	Windows 7	1	1GB		
正常	Windows 7	2	2GB		
全能	Windows 7	4	4GB		
重度	Windows 7	8	8GB		

完成计算瓶颈排除后，对单机单虚拟机进行全生命周期的 IOPS 记录，获取 IOPS 峰值，为 IOPS 需求的上限，也可作为解决启动风暴的关键参数，对应用进行模拟后得出应用的带宽峰值。

3.测试 RBD 裸块存储设备性能用例设计

对块设备测试文件的常见存储操作，测试其性能。这其中要注意测试的文件系统类型和客户端 RBD 块设备数量，由于块设备与系统中直接识别的未初始化物理存储设备一致，是无法直接使用的，所以分别使用 Linux 常见的文件系统格式对块存储设备进行

格式化，并测试多个文件系统格式的性能差别。使用不同数量的客户端向 块设备进行并发读写文件，测试不同变量下的性能，测试用例如表 3-7

表 3-7 RBD 块设备性能测试用例

客户端数量	单个客户端 RBD 设备量	RBD 文件系统类型	测试文件大小	单次 IO 大小	测试线程数量	完成测试内容
1	1/2/3/4	RBD /XFS/EXT4	30G	4K	64	随机读
1	1/2/3/4	RBD /XFS/EXT4	30G	4K	64	随机写
1	1/2/3/4	RBD /XFS/EXT4	30G	4K	64	随机读写
2	1	RBD /XFS/EXT4	30G	4K	64	随机读
2	1	RBD /XFS/EXT4	30G	4K	64	随机写
2	1	RBD /XFS/EXT4	30G	4K	64	随机读写

4.测试 Ceph 提供的虚拟机磁盘性能用例设计

使用 QEMU 搭建虚拟机平台，通过服务接口连接 RBD 快存储提供虚拟磁盘，在虚拟机操作系统内使用与本地磁盘测试相同的工具进行 IOPS 和吞吐量的测试。由于是测试服务器提供保证虚拟机服务能力的最大性能，需要定量测试，所以控制了两个参数，一个是 IOPS，一个是吞吐量，测试在这两种控制情况下能最大能启动满足上诉指标虚拟机的数量。随机 IO：固定为 100IOPS，延时小于 20ms,顺序写 IO：固定 MBPS（吞吐率）为 60MB/s，测试用例如表 3-8

表 3-8 虚拟机磁盘性能测试用例

控制类型	队列深度	IO 块大小	完成测试内容	虚拟机数量
固定 IOPS	2	4KB	随机读、写	20 台以上
固定吞吐量	2	4KB	随机读、写	20 台以上
固定 IOPS	2	4KB	顺序读、写	20 台以上
固定吞吐量	2	4KB	顺序读、写	20 台以上

5.测试 Ceph 扩容功能用例设计

Ceph 扩容功能设计用例如下表：

表 3-9 Ceph 扩容功能测试用例

用例名称	Ceph 扩容功能
预置条件	1. Ceph 存储系统正常运行中 2. 副本数与节点数一致
测试步骤	1. 在每个节点插入一块 X 容量大小新硬盘 2. 进行系统挂载 3. 更新 Ceph 配置，使新加入硬盘成为 OSD 节点
预期结果	Ceph 总体存储容量增大 X 容量大小

6.异常测试用例设计

存储异常测试用例设计如下：

表 3-10 Ceph 异常测试用例

用例名称	Ceph 异常测试
预置条件	1. Ceph 存储系统正常运行中
测试步骤	1. 创建一个 1G 的 RBD 块映射到系统挂载目录中 2. 进行目录挂载并格式化为 xfs 格式 3. 对格式化后文件放入大小为 900M 的测试文件 test 4. 校验文件 md5 值 5. 根据 rbd map 查出对应副本所在的 OSD 6. 卸载该 OSD 硬盘 7. 再次校验 md5 值
预期结果	两次校验值相同
测试结果	

3.3.9 性能测试与系统调优

由于 Ceph 分布式存储功能相对完整和稳定，所以现在限制着 Ceph 分布式存储在桌面虚拟化的推广的最大限制就是启动风暴和后期的规模支持能力，而启动风暴的根本就是存储系统 IOPS 性能的体现，特别是将以前单台计算机及操作系统的 IOPS 集中到了一台服务器上，容易出现瓶颈，所以需要确定分布式存储系统在不同桌面虚拟化负载下性能的表现判断容量和性能的虚拟机容纳量。

分布式存储系统测试需要采集使用中的多种数据，跟踪数据及增长规律，找出对存储系统有性能影响的关键点，从而调整对应参数来改善存储系统性能及适应能力，最终挤压系统的性能极限，或者为新型分布式存储设计做出设计。分布式存储系统测试有以下几个方面组成：

1. 测试系统能力

在模拟的仿真负载场景下测试，确定存储系统和虚拟机体验是否可以达到用户的使用需求。

2. 识别性能上限和漏洞

不断调节虚拟机对存储性能的负载，逼近理论极限，查找制约性能增长的限制因素，确定相关的影响因子权重，优化存在问题的部件，填补存储系统的漏洞。

3. 验证功能性能

执行不同业务负载，探索存储性能参数范围，验证服务的完整性、通信响应时间、

无故障率、容灾恢复、读写疲劳等，根据数据结果分析，掌握存储软硬件性能边界，为虚拟化管理平台动态调整负载提供数据支撑。

4. 系统调优

根据上述优化参数调整存储系统，对性能瓶颈进行突破，并填补系统缺陷，并重复针对性测试，最终优化分布式存储系统性能。

3.4 本章小结

本章的首先介绍了分布式存储的测试的流程，针对桌面虚拟化下的分布式存储测试进行需求分析,为本地存储能力测试以及 Ceph RBD 块存储接口的功能及基准性能测试，对桌面虚拟化所使用的后台存储介绍了测试方案，设计了测试用例。

第四章 Ceph 分布式存储系统测试及结果分析

首先介绍本文 Ceph 分布式存储的测试环境，介绍测试过程和方法，并统计并呈现测试数据，根据测试数据做出分析，验证测试方案。最后优化并对整个测试方案做出总结，对其先进性、实用性、可靠性和局限性做出评判。

4.1 测试部署

部署测试分为存储服务器集群的硬件及操作系统，网络以及虚拟化平台以及安装 Ceph 分布式软件。

4.1.1 存储服务器以及测试平台配置

整个软件平台以及 Ceph 此次部署在四个存储节点服务器上，拓扑如图 4-1 所示。

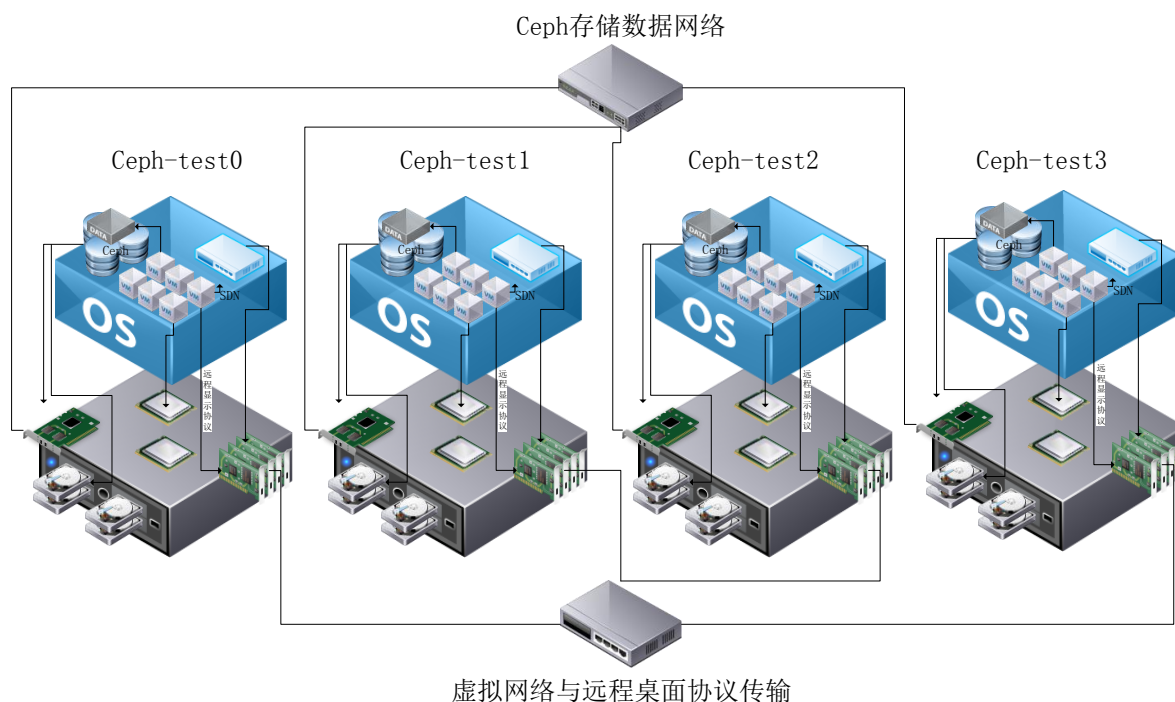


图 4-1 测试集群拓扑

其中服务器具体的硬件型号及版本信息如表 4-1 所示，内含四个服务器节点，资源平均分配。使用融合式架构的情况下，服务器操作系统与虚拟化管理平台软件（包括 Ceph 管理端），安装在第一块磁盘；第二块 SSD 磁盘为 Ceph 日志缓存盘；后面三块磁盘为 Ceph 的 OSD。虚拟桌面的磁盘 virtio 驱动通过 qemu 与 Ceph 的 RBD 驱动连接获得存储能力。桌面虚拟化的计算部分通过虚拟化管理平台软件生成，将计算资源与 Ceph 提供的虚拟机盘和虚拟交换机提供的网络端口结合，形成完整的虚拟机。此次用

于测试集群的软硬件配置如下表 4-1 所示：

表 4-1 测试集群的软硬件配置

CPU	英特尔第二代至强处理器 24 核 48 线程
内存	256G
磁盘	12 块 2.5 寸 SATA 硬盘 4 块 2.5 寸 100G SSD
网卡	4 个万兆端口 4 个四端口千兆网卡
操作系统及软件要求	
OS:RedHat CentOS 6.5 64Bit	
分布式存储版本： Ceph 0.8.2	
虚拟化部件： libvirt 1.2.2, qemu 2.2.0	
虚拟化机操作系统： Windows 7 Ulitmate 64bit	
虚拟化设备驱动： Virtio-0.100	
虚拟化管理平台： openstack	

4.1.2 Ceph 测试集群网络配置

由于虚拟机需要对外提供服务，而对象存储设备相互之间需要访问，因为保证虚拟化桌面远程传输的网络通道与分布式存储后台数据通信通道互不干扰，所以在网络上进行隔离，测试环境中虚拟化桌面网络使用 4 千兆口进行链路聚合，两个万兆口链路聚合用作存储通信私有网络，网络信息如表 4-2 所示。

表 4-2 网络配置

IP（虚拟化桌面/分布式存储）	主机名	组件名
10.12.18.1/10.12.19.1	Ceph-test0	OSD ,monitor
10.12.18.5/10.12.19.5	Ceph-test1	OSD
10.12.18.10/10.12.19.10	Ceph-test2	OSD
10.12.18.15/10.12.19.15	Ceph-test3	,OSD

4.1.4 网络环境验证

网络作为整个 Ceph 分布式存储通信的基础，网络的带宽与稳定性是影响分布式存储正常工作的最关键因数，虚拟机桌面传输协议所使用的 10.12.18.0/24 网段千兆网链路聚合后达到 4GB 带宽，存储集群的内部 10.12.19.0/24 网段万兆网链路聚合后达到 20GB 带宽。

在服务器端端与连接的客户端。由 20Geb 网络连接的节点之间的网络带宽为

19.41Gbits/sec,理论速度为 2500MB/s, 操作过程如下

```
root@Ceph-test0:~# iperf -c Ceph-test1 -t 30
-----
Client connecting to compute1, TCP port 5001
TCP window size: 23.5 KByte (default)
-----
[ 3] local 10.12.18.1 port 54202 connected with 10.12.18.5 port 5001
[ ID] Interval Transfer Bandwidth
[ 3] 0.0-30.0 sec 32.8 GBytes 19.41Gbits/sec
```

图 4-2 网络万兆端口带宽测试记录图

由 4Geb 网络连接的节点之间的网络带宽为 3939Mbits/sec,性能为 500MB/s,操作过程如下。

```
root@Ceph-test0:~# iperf -c Ceph-test1 -t 30
-----
Client connecting to compute1, TCP port 5001
TCP window size: 23.5 KByte (default)
-----
[ 3] local 10.12.19.1 port 36032 connected with 10.12.19.5 port 5001
[ ID] Interval Transfer Bandwidth
[ 3] 0.0-30.0 sec 32.8 GBytes 3939Mbits/sec
```

图 4-3 网络千兆端口带宽记录图

各节点相互之间全部经过验证,证明网络环境正常。避免了网络瓶颈对存储造成影响。部署的测试环境与生产环境相符,各组件运行正常,无软硬件故障,硬件配置也未有明显瓶颈。接下来是性能测试。

4.2 性能测试

测试从本地既有存储性能、分布式存储裸块设备性能及虚拟机内部测试性能等三个方面根据测试用例要求进行性能测试,得出并测试结果,提出优化的解决方案,验证测试方案与测试方法是否有效。

4.2.1 本地磁盘存储测试

将读写数据直接在存储设备上操作。网存储设备上写数据,按照测试用例测试块大

小为 4M。

1. 本地存储设备的顺序读写性能操作如下

```
dd if=/dev/zero of=/dev/sdd bs=4M (4K) count=1000 oflag=direct,sync
1000+0 records in
1000+0 records out
4194304000 bytes (4.2 GB) copied, 23.1389 s, 181 MB/s
```

图 4-4 本地存储顺序读结果记录图

2. 测试顺序写语句如下图

```
fio -ioengine=libaio -bs=512k -direct=1 -thread -rw=write -size=100G -filename=/dev/vdb -name="EBS
512KB seqwrite test" -iodepth=64 -runtime=60
```

图 4-5 本地存储顺序写测试代码记录图

吞吐量 MBPS 为 184226KB/s, 约为 180MB/s。iops 达到 340。以上几项测试结果具备一般代表性, 与 dd 结果的数值相近。

2. 接下来使用 FIO 对硬盘进行读写 IOPS 测试。首先对所有硬盘的总 IOPS 进行测试 4K 随机读性能, 测试代码如下图

```
#fio -ioengine=libaio -bs=4k -direct=1 -thread -rw=randread -size=100G -filename=/dev/vdb -name="EBS
4KB randread test" -iodepth=8 -runtime=60

fio 2.0.7

Starting 1 threads

Jobs: 1 (f=1): [100% done] [22977K/0K /s] [5744/0/00 iops] [eta 00m:00s]

EBS 4KB randread test: (groupid=0, jobs=1) :err=0:pid=7203:Sun JAN 20 18:50:44 2015

    read: io=1853.4MB, bw=24012KB/s, iops=6030, runt=60058msec

    slat(usec): min=2, max=179, avg=6.67, stdev=4.65

    clat(usec): min=2, max=221116, avg=2656.19, stdev=31653.46

    lat (usec): min=60, max=221117, avg=2528.08, stdev=31653.46
```

图 4-6 本地存储随机读性能测试结果记录图

这其中的读了 1853.4MB 大小的数据, 速率为 24012KB/s, iops 等于 6030, 第二条下划线处显示 IO 请求的平均响应时间为 2.65ms, 第三条下划线处说明 94% 的 IO 请求的响应时间是小于等于 1.7 ms 的。第四条下划线处表示该硬盘的利用率已经达到了 99.96%。

3. 随机写

fio 测试随机写语句如下图

```
$fio -ioengine=libaio -bs=4k -direct=1 -thread -rw=randwrite -size=1000G -filename=/dev/vdb \
-name="EBS 4K randwrite test" -iodepth=64 -runtime=60
```

图 4-7 本地存储随机写测试代码记录图

结果 IOPS 是 5700。IO 在 5.42ms 内响应, 96% 的 IO 请求的响应时间在 6.24 ms 之

内。

本地磁盘存储性能汇总：

表 4-3 本地磁盘性能测试结果

磁盘类型	IO 大小	吞吐量	完成测试内容
SATA 硬盘	4M	200 MB/s	顺序读
SATA 硬盘	4M	181 MB/s	顺序写
	IO 大小	总 IOPS	
SATA 硬盘	4K	6030	随机读
SATA 硬盘	4K	5700	随机写

4.2.2 桌面虚拟化对存储的具体性能测试

测试一般情况下 Windows7 桌面操作系统的虚拟机所需要的存储性能。使用本地存储的虚拟机，通过测试软件得到的开机 IOPS 性能与开机时间的关系如下图 4-8 所示

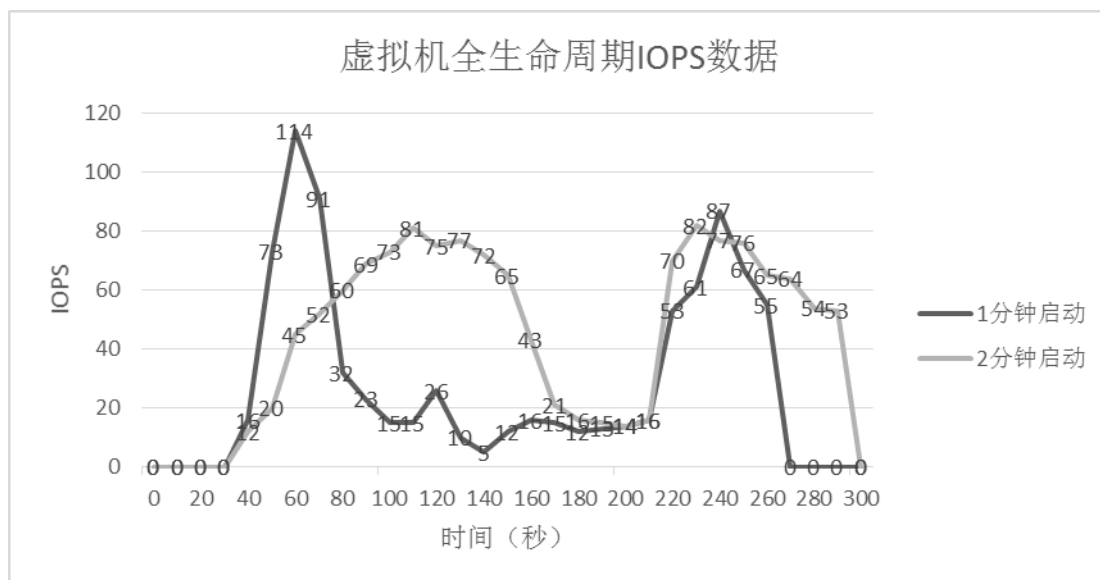


图 4-8 虚拟机 IOPS 外部监控曲线图

1 分钟内看到 IOPS 在第一次启动时最高的时候为 114，启动完毕后迅速下降达到 15 左右，登陆操作后又小幅上涨到 26，在无操作情况下稳定在 10-15 左右，当触发关机操作时会突长到 87 并且回落，观察到虚拟机操作系统在完成关机以后 IOPS 恢复至 0。最后统计开机 IOPS 所需资源为 114

表 4-4 Windows 桌面操作系统的虚拟机所需资源

Windows 7	快速启动	正常启动
开机时间	<1 分钟	<2 分钟
IOPS	114	81

在已启用的虚拟机中通过自动化测试，同时启动并使用按键精灵脚本模拟各种不同

负载的软件操作，并导出 iometer 工具记录相关 IOPS 引荐资源的得到参数如下表 4-5：

表 4-5 不同用户负载下对硬件资源的需求

用户负载	操作系统	vCPU 用量	虚拟内存用量	平均 IOPS (稳态)	带宽峰值 mbps
轻度	Windows 7	1	1GB	15	161
正常	Windows 7	2	2GB	26	219
全能	Windows 7	4	4GB	46	376
重度	Windows 7	8	8GB	63	450

峰值带宽最高应用达到 450mbps 即 56.25MB/s，所以为满足最终负荷的带宽峰值分布式存储应该提供高于 60MB/s 的带宽。表中 IOPS 数据仅显示无操作状态下的系统稳态时数据，启动时和关闭时的 IOPS 会达到 110 左右，如下图所示：

4.2.3 Ceph 块存储测试

首先对 Ceph 分布式存储中原始的块设备进行测试，

本次对 Ceph 块设备进行基准测试，测试存储常见的操作时存储系统的性能，块设备接口基准测试命令如下：

```
fio -filename=/mnt/rbd0/tt -direct=1 -rw=randrw -bs=4k -size=30G
-numjobs=64 -runtime=120 -group_reporting --name=test
```

图 4-9 块存储测试命令记录图

3.测试结果

Ceph 块设备和使用 ext4 和 xfs 文件系统格式化的 RBD 块设备的在随机读能相差在 700-900IOPS 左右，是符合误差范围的性，如图 4-10 所示。

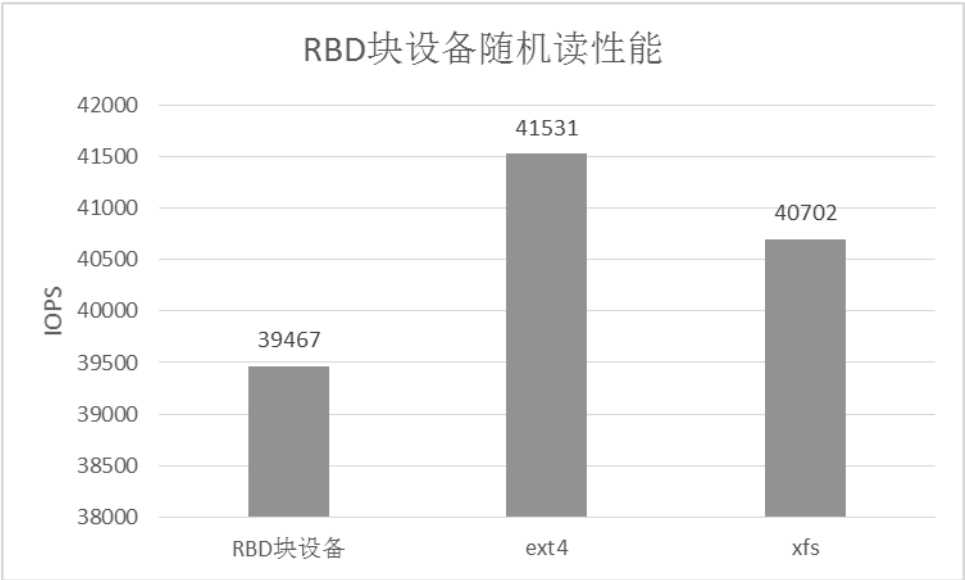


图 4-10 三种类型 RBD 设备的随机读性能柱状图

基于 ext4 文件系统的 RBD 和裸 RBD 的性能在随机读这项上面一样，相差 100IOPS 左右，但是基于 xfs 的 RBD 的随机写与随机读写性能比较高，如图 4-11 所示。

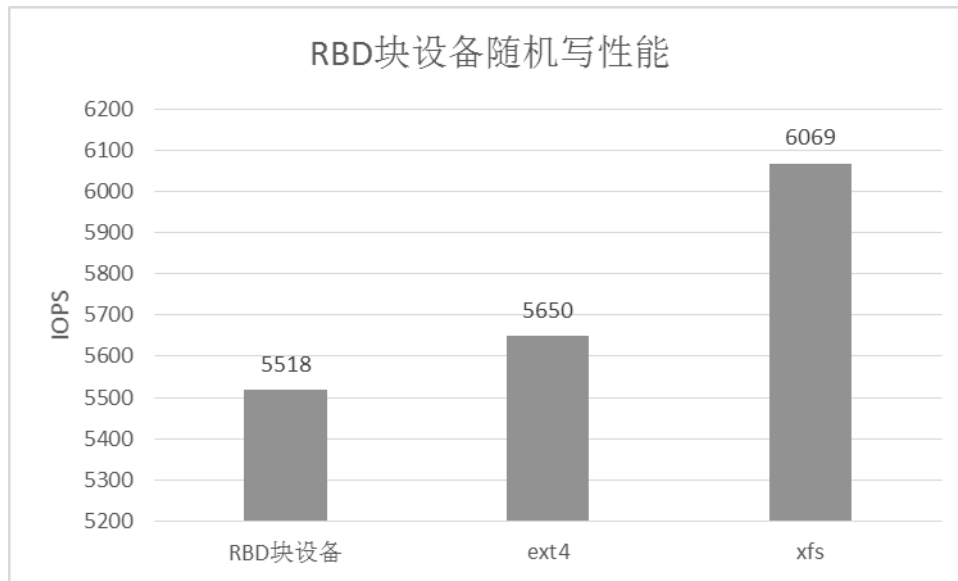


图 4-11 三种类型 RBD 设备的随机写性能柱状图

基于 ext4 文件系统的 RBD 的随机读写性能和基于 ext4 文件系统的 RBD 的性能与随机读写一样，而裸 RBD 的随机读写性能比较高，图 4-12 所示。

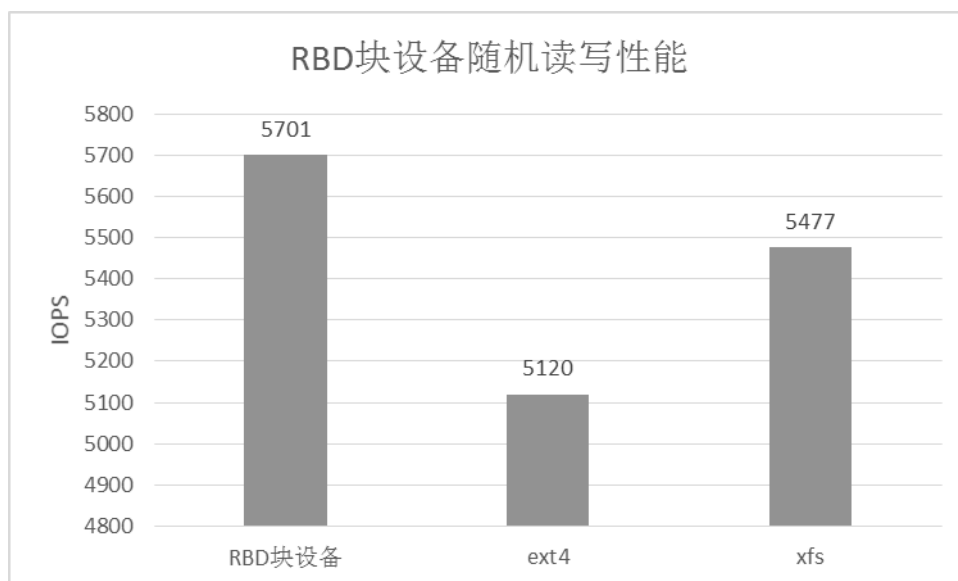


图 4-12 三种类型 RBD 设备随机读写性能柱状图

2. RBD 块设备接口性能调优测试

分别在一个客户端中创建 1 至 4 个 RBD 块设备，对每个 RBD 块设备同时并发产生读写。测试 RBD 块存储数量的系统读写性能。测试结果得出当一个客户端构建两个 RBD 块设备时，随机读性能为理想状态，继续增加后衰减。如图 4-13。

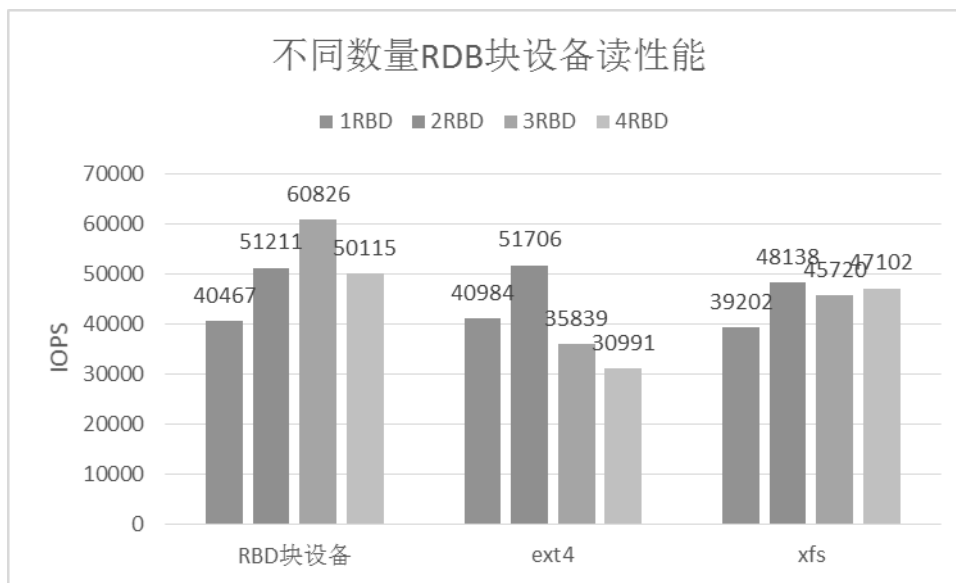


图 4-13 随机读性能柱状图

存储系统的随机写达最理想状态发生在块设备为三的情况下，如图 4-14。

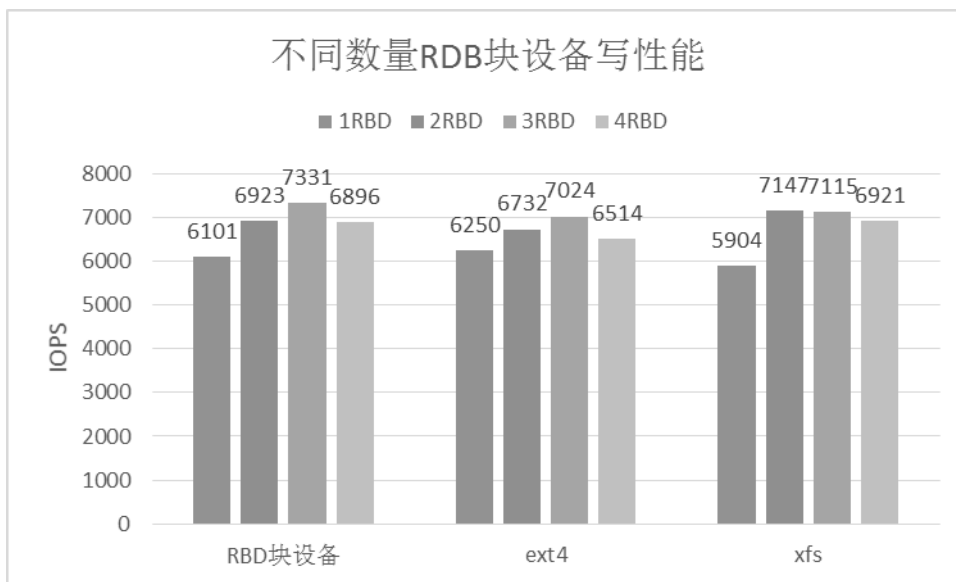


图 4-14 随机写性能柱状图

随机读写的性能与随机写情况基本一致，但是只有 1 个 RBD 块设备时候无法发挥分布式存储特点，性能低下，如图 4-15

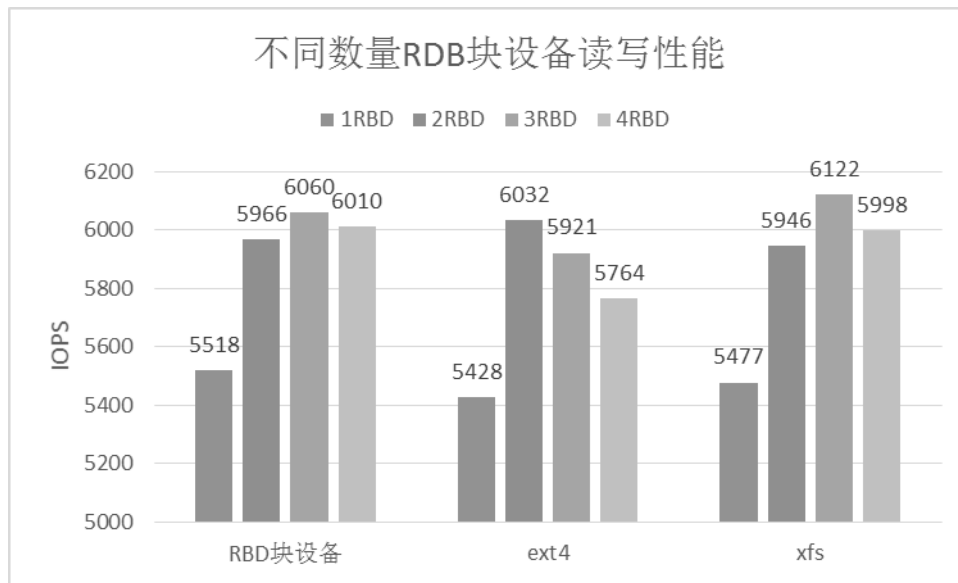


图 4-15 不同 RBD 设备数量情况下接口的随机读写性能柱状图

4.2.4 测试 Ceph 提供的虚拟机磁盘性能

然后在虚拟机内使用存储测试工具，按照测试用例对以下情况测试：1.在虚拟机最大 IOPS 100 下，测试 4k 随机读写情况 2.在虚拟机最大带宽 60MB/s 下，测试 64k 顺序读写情况 3.测试结果分析

1. 随机 IO 性能

测试工具的使用与前面测试本地磁盘相同。图 4-16 表示每个虚拟机最大 IOPS 为 100，队列深度 2，4KB 随机读操作测试结果。

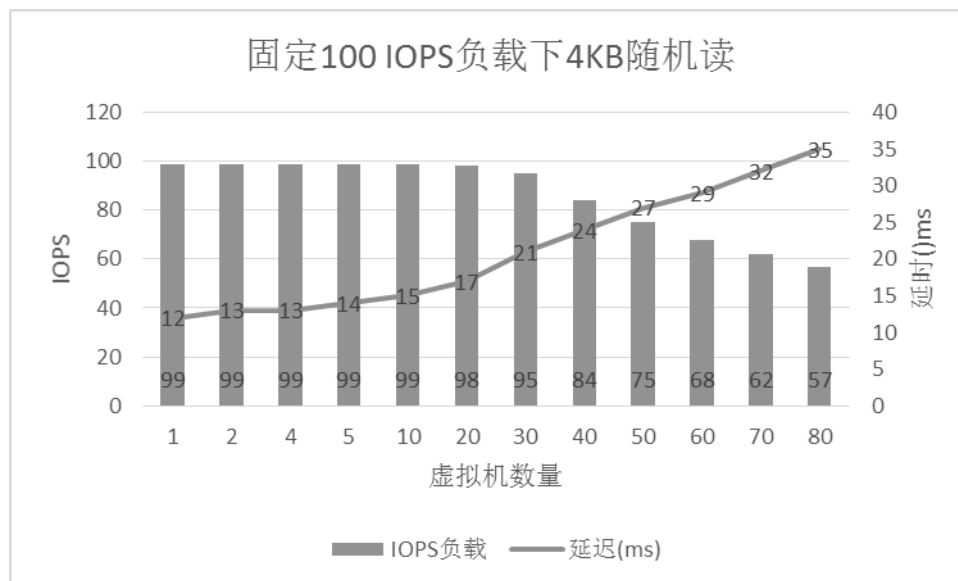


图 4-16 IOPS100 时 4KB 随机读曲线图

从图可得，每个虚拟机的操作延时随着虚拟机数量增多而增大，同时虚拟机的 IOPS

随之下降。当 30 台虚拟机时，可达到每台虚拟机 95 IOPS，延时 21ms，整个集群 IOPS 总量为 2858。所以如果虚拟机的应用负载会经常造成在 4KB 随机读这样类似的操作下，环境最多支撑 30 台虚拟机。

图 4-17 表示每个虚拟机最大 IOPS 为 100，队列深度 2，4KB 随机写操作测试结果。

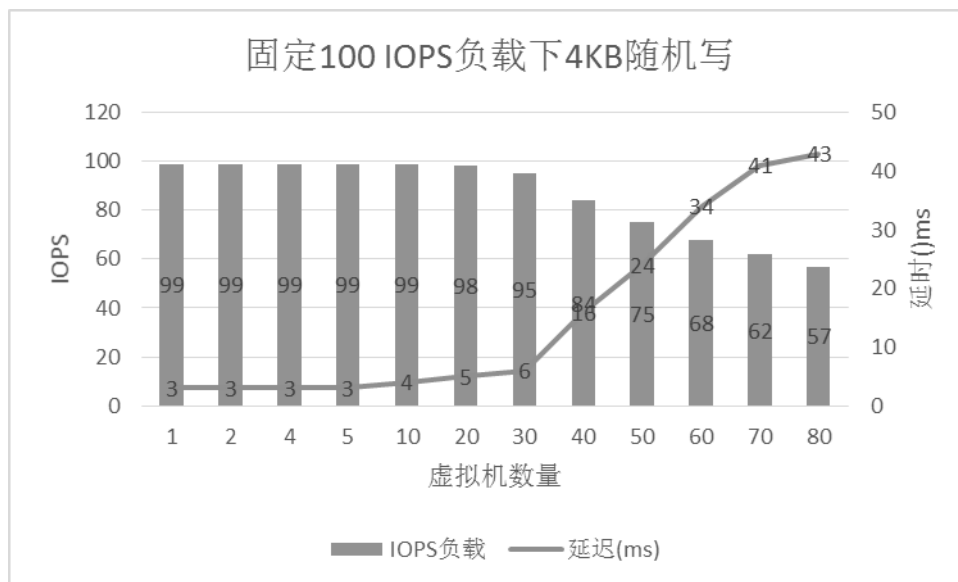


图 4-17 IOPS100 时 4KB 随机写曲线图

当少于 30 台虚拟机时，4KB 随机写操作延时不大于 3ms，虚拟机内操作流畅。而虚拟机台数大于 30，延时猛增到 24ms，IOPS 下降到了 82。这是因为存储服务器使用固态硬盘作为日志以及硬盘文件存储内存页，当数据负载超过固态硬盘及缓存能力，需要把数据写入硬盘，延时会变大。所以如果虚拟机的应用负载会经常造成 4KB 随机写操作下，环境最多支撑 30 台虚拟机，总 IOPS 的需要为 3000。

经上文测试，每个 SATA 硬盘盘读 IOPS 为 80IOPS，写 IOPS 为 200（由于写缓存），11ms 延时。

表 4-6 Ceph 4KB 随机读写性能

模式	最大吞吐量	吞吐量 (QoS)	理论吞吐量	效率
4KB 随机读	1582	2858	3600	79%
4KB 随机写	3357	3000	4000	75%

表 4-6 中第二列表示集群随机读写 IOPS 的峰值，当 80 台虚拟机时，集群随机读总 IOPS 为 4582，随机写总 IOPS 为 3357。理论上，集群总 IOPS 随着虚拟机的个数增加而增大。第三列值表示当满足性能指标前提下，集群 IOPS 总值。第四列表示 40 块 SATA 硬盘的 IOPS 理论值。由第三列及第四列的值，可算出整个集群的 4KB 随机读操作性能达到理论值的 79%，随机写操作达到 75%。

2.顺序 IO 性能

图 4-18 表示虚拟机最高吞吐率 60MB/s,队列深度 64, 64KB 顺序读的测试结果。

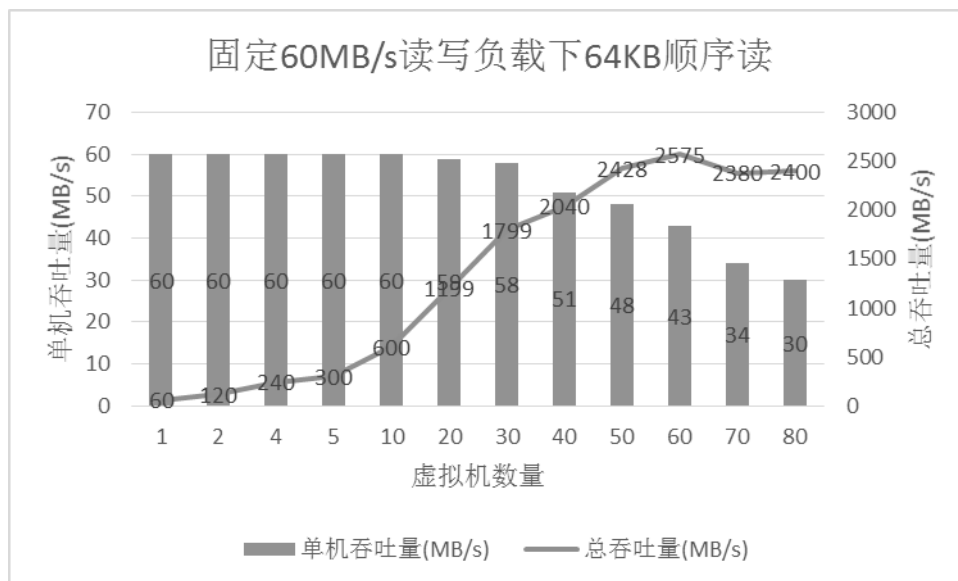


图 4-18 最大吞吐为 60MB/s 时 64KB 顺序读曲线图

随着虚拟机个数增加，虚拟机内磁盘的吞吐率逐渐下降。集群顺序读吞吐率峰值出现在 60 台虚拟机，总值为 2575MB/s。所以如果虚拟机的应用负载在经常造成 60MB/s 的读操作时，测试环境最多支撑 30 台虚拟机，总吞吐量为 1799MB/s。

图 4-19 表示虚拟机最高吞吐率 60MB/s,队列深度 64, 64KB 顺序写的测试结果。

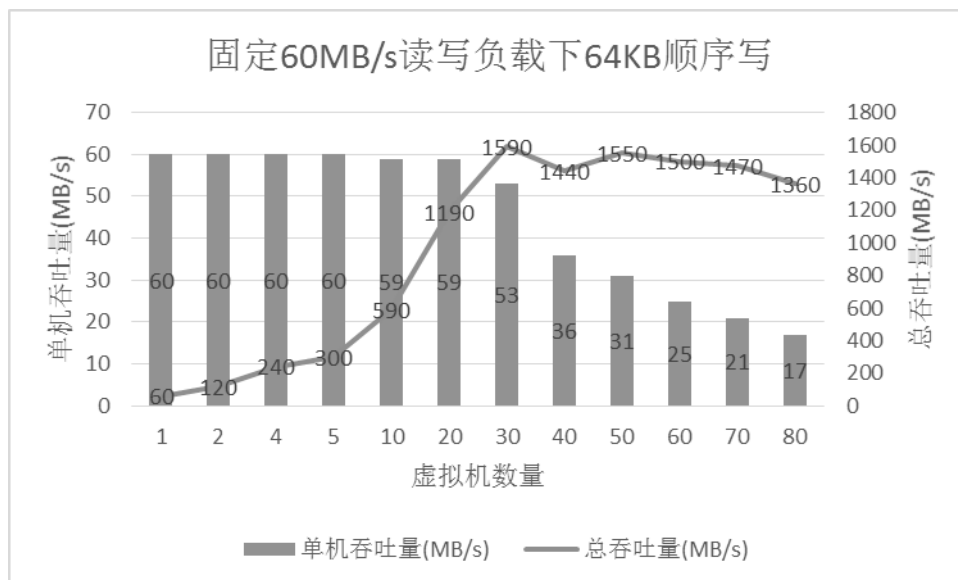


图 4-19 大吞吐为 60MB/s 时 64KB 顺序写曲线图

从图可知，虚拟机内吞吐率随着虚拟机个数增加而下降。集群顺序写的吞吐率峰值出现在 50 台虚拟机，总值 1487MB/s。所以如果虚拟机的应用负载在经常造成 60MB/s 的写操作时，测试环境最多支撑 20 台虚拟机，此时总吞吐率为 1197MB/s。

经测试，SATA 盘顺序读写的吞吐率均为 160MB/s，10Gb 网卡带宽约 900MB/s。因此，整个环境 40 个 SATA，理论吞吐率为读 6400MB/s，写 3200MB/s，4 张网卡总带宽为 3600MB/s，这样网卡带宽也不会对此形成瓶颈。因此可得表 4-7，存储 64KB 顺序读性能只达到理论值的 57%，顺序写性能达到 37%。

表 4-7 Ceph 顺序读写性能

模式	最大吞吐量	吞吐量 (QoS)	理论吞吐量	效率
64KB 随机读	2759	2263	6400	59%
64KB 随机写	1487	1197	3200	37%

当存储随机 IO 达到理论 75%的性能，顺序写操作性能却只剩下 37%。因为 Ceph 存储会将每个文件分成多个固定大小的对象，而这些对象根据 CRUSH 算法散列到不同的对象存储服务器的硬盘上，不同文件的对象会映射到相同的硬盘上。因此虚拟机上的顺序 IO 映射到硬盘上变成随机 IO，这会造成延时变大以及总的吞吐率减少，所以在虚拟机需要顺序读写时，Ceph 的 IO 效率低下，根据分布式存储特点进行顺序读写的优化。

4.3 Ceph 扩容功能测试

在扩容测试中，在副本数设置为 4 与节点数一致，每个节点已有 3 个 OSD，存储总容量为 3TB，并且正在使用 2213MB 存储容量，剩余 797MB，如下图所示：

```
[root@osd1 ~]# ceph -s
cluster f649b128-963c-4802-ae17-5a76f36c4c76
health HEALTH_OK
monmap e1: 4 mons at {mon1=10.12.18.1/0,mon2=110.12.18.4/0,mon3=10.12.18.10/0,mon4=10.12.18.15/0}
osdmap e13: 12 osds: 12 up, 12 in
pgmap v24: 64 pgs, 1 pools, 0 bytes data, 0 objects
          2416 MB used, 797 MB / 3213 MB avail
```

图 4-20 Ceph 扩容前状态截图

现再插入一块 1TB 的硬盘，进行系统挂载后，更新 Ceph 配置

```
[osd.4]
host = osd4
devices = /dev/sda5
```

图 4-21 Ceph 配置更新记录图

启用 osd4，扩容后再次确认如下图所示：

```
[root@osd1 ~]# ceph -s
cluster f649b128-963c-4802-ae17-5a76f36c4c76
health HEALTH_OK
monmap e1: 4 mons at {mon1=10.12.18.1/0,mon2=110.12.18.4/0,mon3=10.12.18.10/0,mon4=10.12.18.15/0}
osdmap e13: 16 osds: 16 up, 16 in
pgmap v24: 64 pgs, 1 pools, 0 bytes data, 0 objects
          2520 MB used, 1593 MB / 4113 MB avail
```

图 4-22 Ceph 扩容后状态截图

完成扩容功能测试用例如下表：

表 4-8 Ceph 扩容功能测试用例

用例名称	Ceph 扩容功能
预置条件	3. Ceph 存储系统正常运行中 4. 副本数与节点数一致
测试步骤	4. 在每个节点插入一块 X 容量大小新硬盘 5. 进行系统挂载 6. 更新 Ceph 配置，使新加入硬盘成为 OSD 节点
预期结果	Ceph 总体存储容量增大 X 容量大小
测试结果	与预期结果一致，扩容功能达到要求

4.4 异常测试

对 Ceph 分布式户进行异常测试，进行硬盘下线操作，检查在节点 1 中进行以下操作：

```

rdm create testdevice --size 1024 [-m {mon-10.12.18.1}] [-k /path/to/ceph.client.admin.keyring]
#创建一个 1G 大小的 RBD 块
sudo rbd map foo --pool rbd --name client.admin [-m {mon-10.12.18.1}] [-k
/path/to/ceph.client.admin.keyring]
#在节点上,将镜像映射到块设备
sudo mkfs.xfs -m0 /dev/rbd/rbd/testdevice
#在 cephclient 节点上格式化块设备
sudo mount -t xfs /dev/rbd/rbd/testdevice /mnt/testdevice
#挂载该设备.
cp test /mnt/testdevice
#拷贝 test 文件到挂载目录下
md5sum /mnt/testdevice/test
#对目录中的文件进行 md5 值计算
13df384c47dd2638fd923f60c40224c6
    
```

图 4-23 创建 RBD 拷贝文件记录图

得出 test 文件的 MD5 值为 13df384c47dd2638fd923f60c40224c6，之后进行 OSD 查找和卸载，执行如下操作：

```

rdm info testdevice
#查看 rbd 块设备信息
  rbd image 'testdevice':
    size 1024 MB in 256 objects
    order 22 (4096 kB objects)
    block_name_prefix: rb.0.fbe4e.2ae8944a
    format: 1

ceph osd map pool rb.0.fbe4e.2ae8944a.000000000000
#块设备首地址映射 osd 地址
osdmap e50 pool 'pool' (15) object 'rb.0.fbe4e.2ae8944a.00000000' ->
pg 15.5cf0bd7d (15.3d) -> up ([1,0,2,3], p1) acting ([1,0,2,3], p1)
#查到数据存放目录为 15.3d._head primary osd=1 replicate osd=0,2,3
    
```

图 4-24 根据 RBD 进行 OSD 查找记录图

将 OSD1,0,2,3 对应的硬盘一一进行断电拔除。

```
ceph osd tree
# id      weight      type  name      up/down reweight
-1        4          root default
-2        1          host  osd1
0         1          host  osd.0    down    1          #osd.0下线
-3        1          host  osd2
1         1          host  osd.1    up      1
-4        1          host  osd3
2         1          host  osd.2    up      1
-5        1          host  osd4
3         1          host  osd.3    up      1

md5sum /mnt/testdevice/test                                #重新对目录中的文件进行md5值计算
13df384c47dd2638fd923f60c40224c6
```

图 4-25 移除 osd.0 后 MD5 计算值记录图

重新计算 test 文件的 md5 值也与之之前所得一致。所以证明在对应硬盘发生故障异常时，分布式存储也能保证各副本的一致性，从而保证上层业务业务连续性。

存储异常测试完成用例如下：

表 4-9 Ceph 异常测试用例

用例名称	Ceph 异常测试
预置条件	2. Ceph 存储系统正常运行中
测试步骤	8. 创建一个 1G 的 RBD 块映射到系统挂载目录中 9. 进行目录挂载并格式化为 xfs 格式 10. 对格式化后文件放入大小为 900M 的测试文件 test 11. 校验文件 md5 值 12. 根据 rbd map 查出对应副本所在的 OSD 13. 卸载该 OSD 硬盘 14. 再次校验 md5 值
预期结果	两次校验值相同
测试结果	两次校验值相同

4.5 测试结果分析与调优

4.5.1 测试结果分析

根据前文测试中获得的数据及整理结果，我们可以做出如下分析：

1.在分布式存储设备测试过程中首先需要本地磁盘与文件系统间有一定的性能影响，虽然本次磁盘的存储性能与吞吐量与 Ceph 块裸盘的性能和虚拟机磁盘性能在吞吐量上不会形成瓶颈，但是 IOPS 会有明显局限。

2.桌面虚拟化对存储的性能需求，当虚拟桌面数量较多时吞吐量的约束远小于 IOPS 产生的约束。为避免多虚拟机并发启动、登录而造成的启动风暴和登陆风暴的问题，可

以适当配置高 IOPS 存储设备或这增加存储数量提供更大带宽和增加 IOPS。但是存储数量越多，控制器越多，在计算能力上也耗费更多资源，需要平衡已有设备和计算资源的争抢。

3. Ceph 参数和块设备数量

三个块设备的时候随机写和随机读写性能为理想状态，而且在分布式存储测试中块使用 XFS 文件系统的支持明显好于 EXT4 文件系统，所以在给虚拟机提供磁盘的时候可以根据需要提供两个虚拟机磁盘，以增加磁盘读写性能，使用三个块设备达到平衡读写，并且可以使用 XFS 作为文件系统使用。

4. 保证用户体验的最大支撑数量

在满足虚拟机最低 IOPS100 的性能要求时，虚拟机在对 4KB 随机读写操作需求下，环境中每节点最多支撑 30 台虚拟机，能保证每台虚拟机的用户体验，超过临界值，每一台虚拟机平均减少 10 个 IOPS，当超过 70 台时，延迟达到 30ms 以上造成严重的卡顿。特别在随机写时，超过 30 台以后延迟会突然出现 10ms 的跃升，根据虚拟机在后端大小判断应该是持续测试时间超过 20 分钟以后，突然出现了后端容量增大的问题，导致写数据增加，延迟增长，如下表 4-10 及图 4-26：

表 4-10 虚拟机在 ceph 后端随时间变化写大小变化

分钟	4MB	16MB
2	356	800
20	506	1053
720	4936	5614

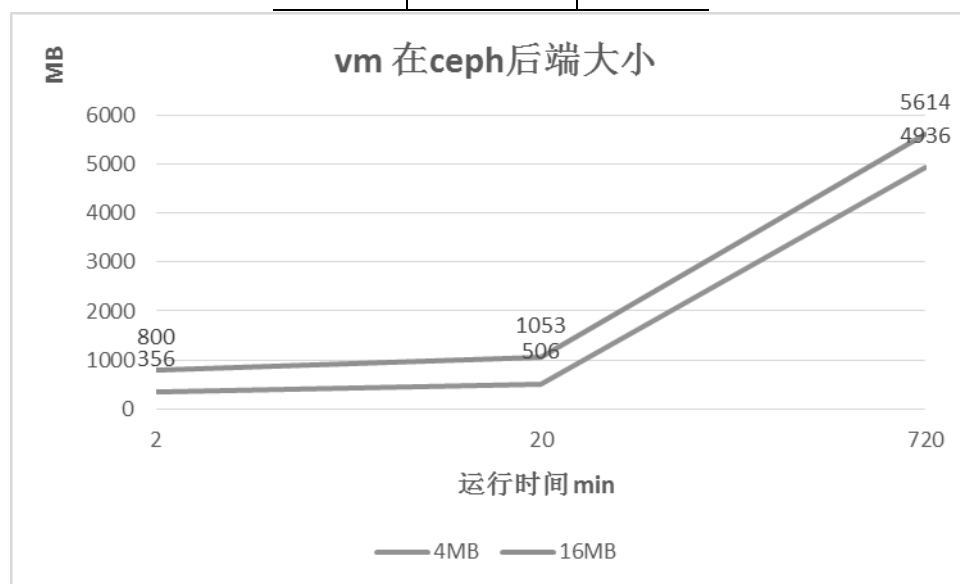


图 4-26 虚拟机在 ceph 后端大小曲线图

5. 保证虚拟机带宽的最大支撑数量

当保证虚拟机带宽恒定为 60MB/s 时，虚拟机在 64KB 顺序读写下，最佳情况能支撑 30 台虚拟机，超过 30 台后，开始衰减，而且同样地在写方面出现了比较大幅度的性能下降，从 30 台时的 58MB/s 降到了 40 台时的 36MB/s，接近 37%，分析与 4KB 随机写的情况相似，容量增大速率增加，导致的写性能下降。

6. 综合第 4 点和第 5 点，可以得出，在计算能力充足的情况下，单 2 块机械 SAS 硬盘作为 OSD，1 块 SSD 作为缓存日志盘的情况下，可以支撑的 30 台左右的虚拟机，以保证操作流畅和读写性能。如果继续增加虚拟机的数量，需要额外天机 OSD，以满足性能需求。

在测试分布式存储性能时，关键点在于，需要多层次相互验证性能瓶颈，从本地磁盘性能，分布式存储块性能，和虚拟机内部的虚拟磁盘读写，三个层次跨越两个操作系统进行对照参考分析。

4.5.2 性能调优

所以根据以上得到的测试结果可以通过以下方法来优化性能：

1. 增大对象容量或者增加高 IOPS 设备

根据本地存储和虚拟机开机及负载测试中得出，单一虚拟机的 IOPS 性能需求已经超过了传统的单块硬盘能力，而且，在固定 IOPS100 负载的测试中可以得知，集群有能力开 40 台以上虚拟机，却因存储能力下降延迟增大，IOPS 也迅速开始下滑，无法维持，所以需要提高 IOPS 性能。而且在每个节点使用固态硬盘作为缓存盘后，在加入固态硬盘盘时需要进行分区对齐，避免分区不当带来的性能下降。O_DIRECT 和 O_DSYNC 写模式及写缓存 OSD 在写日志文件时，使用的 flags 是：flags |= O_DIRECT | O_DSYNC，O_DIRECT 表示不使用 Linux 内核 Page Cache；O_DSYNC 表示数据在写入到磁盘后才返回，由于磁盘控制器也同样存在缓存，而 Linux 操作系统无法调用管理设备缓存，O_DSYNC 在到达磁盘控制器缓存之后会立即返回给客户端，并无法保证数据真正写入到磁盘中，Ceph 致力于数据的安全性，对用来作为日志盘的设备，应禁用其写缓存。同一个 Ceph 集群里存在传统机械盘组成的存储池，以及 SSD 组成的快速存储池，可把对读写性能要求高的数据存放在 SSD 池，而把备份数据存放在普通存储池。

2. 重新配置 CRUSH 数据读写规则

根据第 2、3 条测试分析，多个块同时作为使用以平衡读写性能，使主备数据中的主

数据落在较高性能的 OSD 上，而高性能的 OSD 节点不用来组成独立的存储池，而是配置 CURSH 读取规则，让所有数据的主备份落在高性能 OSD 上。Ceph 集群内部的数据备份从固态硬盘的主 OSD 往机械硬盘的副 OSD 写数据。这样，所有的 Ceph 客户端直接读写的都是固态硬盘 OSD 节点，既提高了性能又节约对 OSD 容量的冗余要求。配置重点是 CRUSH 读写规则的设置如下：

```
rule ssd-primary {
    ruleset 1
    step take ssd
    step chooseleaf firstn 1 type host #从 SSD 根节点下取 1 个 OSD 存主数据
    step emit
    step take sas
    step chooseleaf firstn -1 type host #从 SATA 根节点下取其它 OSD 节点存副本数据
    step emit
}
```

图 4-27 CRUSH 调优参数记录图

3. 改变 Ceph 缓存分层

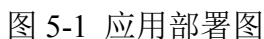
在保证用户体验的最大支撑数量时，分析得出的虚拟机数量临界值，根据虚拟机磁盘突然增大，认为是缓存已经用完，而需要向后端开始申请空间，解决这个问题可以考冷热数据分离技术，用固态硬盘组成一个存储池来作为缓存层，后端用相对慢速/廉价的设备来组建冷数据存储池。构建一个代理程序处理缓存层和存储层的数据的自动迁移，对客户端透明操作透明。第一种是回写模式，Ceph 客户端直接往缓存层写数据，写完立即返回，代理程序再及时把数据迁移到冷数据池。当客户端取不在缓存层的冷数据时，代理负责把冷数据迁移到缓存层。也就是说，Ceph 客户端直接在缓存层上进行 IO 读写操作，不会与相对慢速的冷数据池进行数据交换。这种模式适用于可变数据的操作，如照片、视频编辑等等。第二种是只读模式，Ceph 客户端在写操作时往后端冷数据池直接写，读数据时，Ceph 把数据从后端读取到缓存层。这种模式适用于不可变数据，如页面应用，数据库应用等。CRUSH 算法是 Ceph 集群的核心，在深刻理解 CRUSH 算法的基础上，利用固态硬盘的高性能，可利用较少的成本增加，满足企业关键应用对存储性能的高要求。

4.6 本章小结

本章首先对搭建 Ceph 分布式存储测试的分析过程做了简要介绍，进行对本地磁盘

基准测试、RBD 性能、RBD 提供虚拟磁盘的能力进行试验，总结了实验结果，提出了对基于桌面虚拟化的分布式存储性能优化的方案,而且在测试和使用中也发现了一些有待回答的问题。

根据测试结果显示，Ceph 分布式在调优之后，存储性能和安全冗余上都满足了桌面虚拟化在使用体验和大规模分布式系统的后端存储要求，与传统直连存储和共享存储不同，分布式存储 Ceph 的 IOPS 性能充分保障用户体验，其本身的分布式特点也保障了用户数据在资源池中的存储安全。下面就分布式存储 Ceph 在桌面虚拟化环境中进行部署应用。应用具体部署图 5-1 如下：



对 Ceph 分布式存储进行安装配置，其操作步骤如下所示：

在各个服务器节点分别获取更新系统并安装 Ceph 分布式存储系统。执行命令如下：

56

2. 配置 Ceph

安装完后需要按测试需要修改配置，其中包括集群标识号、验证方法设置、集群的成员、服务器主机名、服务器主机 IP、数据存储路径、日志存储路径以及其它一些运行时选项。测试集群配置文件如下图所示。

```
[global] ↵
fsid = bb262a63-deac-4d78-9653-eb8017b27c3 #集群 ID↵
mon initial members = mon↵
mon host = 10.12.18.1                #monitor 地址↵
auth cluster required = none↵
auth service required = none↵
auth client required = none↵
osd journal size = 1024                #对象存储设备的日志大小↵
filestore xattr use omap = true↵
osd pool default size = 3                #对象存储设备默认大小↵
osd pool default min size = 1↵
osd pool default pg num = 333            #归置组数量↵
osd pool default pgp num = 333↵
osd crush chooseleaf type = 1↵
osd max backfills = 1↵
osd max recovery threads = 1            #对象存储设备的最大恢复线程↵
osd recovery op priority = 1↵
osd client op priority = 63↵
osd recovery max active = 1            #最大可活动的恢复数量↵
public network = 10.12.18.0/24          #公网 IP 段↵
cluster network = 10.12.19.0/24        #存储集群网络 IP 段↵

[osd] ↵
osd mount options xfs = rw,noatime,inode64,logbsize=256k,logbufs=8,delaylog↵
```

图 5-2 Ceph 配置记录图

3. 创建数据存储路径

在主机上使用为每个对象存储设备建立日志的存储的目录，并使用操作系统的文件系统格式格式化对应磁盘后挂载，Ceph-test0 存储节点为例，命令如下图 5-3 所示。

```
#cd /var/lib/ceph/osd    #进入对象存储设备目录↵
#mkdir ceph-0            #建立对应的目录↵

#mkdir ceph-1↵
#mkdir ceph-2↵
#mkfs.ext4 /dev/sdc1 -f   #磁盘使用 ext4 文件系统格式↵
#mkfs.ext4 /dev/sde1 -f ↵
#mkfs.ext4 /dev/sdf1 -f ↵
#mount /dev/sdc1 /var/lib/ceph/osd/ceph-0 #挂载↵
#mount /dev/sde1 /var/lib/ceph/osd/ceph-1 #挂载↵
#mount /dev/sdf1 /var/lib/ceph/osd/ceph-2 #挂载↵
```

图 5-3 格式化块存储记录图

4. 运行 Ceph 分布式存储服务

按上述操作配置完成后，即可使用 Ceph 分布式存储系统。进行状态检查如图 5-4

```
$cd /etc/ceph↵
$mkcephfs -a -c /etc/ceph/ceph.conf -k ceph.keyring #编译系统↵
$Sudo service ceph -a start #启动系统↵
$Ceph health #查看系统状态↵
HEALTH OK↵
```

图 5-4 Ceph 状态检查记录图

5.创建块存储测试环境。

操作过程如下 5-5 所示：

```
#ceph↵
>>rbd create rbd-image0 -size 32768 #创建大小为 32M 的 rbd 镜像 0 ↵
>>rbd list #列出 rbd ↵
>>rbd map rbd/rbd-image0 #映射 rbd 关系↵
>>rbd showmapped #显示 rbd 映射↵
#mount /dev/rbd0 /mnt/rbd0 #挂载 rbd0 到操作系统中↵
```

图 5-5 创建块存储记录图

5.2 桌面虚拟化后端连接 Ceph

1. Ceph 块存储地址创建 RBD 卷

将虚拟机所需要的制作好的 Windows 镜像文件 RAW 导入到 rbd 里面，然后生成母镜像快照的地址 parent 后创建卷。查询卷信息代码如下

```
$ rbd info volume21021/volume-438d531a-a8d4-48ae-aa59-93c3b400b227
rbd image 'volume-438d531a-a8d4-48ae-aa59-93c3b400b227':
size 10240 MB in 640 objects
order 24 (16384 kB objects)
block_name_prefix: rbd_data.3292bb736c2
format: 2
features: layering
parent: rbd/37078763-c9df-47c5-ac10-8a3711c78a27@snap
overlap: 10240 MB
```

图 5-6 查询卷信息记录图

2.在虚拟机配置中使用 RBD 卷

创建卷后，在 qemu 的虚拟机 xml 配置文件中修改虚拟磁盘地址，这一部分可以交

由 Openstack 的 nova 和 Cinder 卷管理模块批量生成卷和修改虚拟机配置文件。修改虚拟机 xml 文件如下：

```
<disk type='network' device='disk'>
  <driver name='qemu' type='raw' cache='none'/>
  <source protocol='rbd' name='volume21021/volume-438d531a-a8d4-48ae-aa59-93c3b400b227'>
    <host name='10.0.0.1' port='6789'/>
  </source>
  <target dev='vda' bus='virtio'/>
  <alias name='virtio-disk0'/></disk>
```

图 5-7 xml 文件修改记录图

然后就可以通过 QEMU 启动以 RBD 存储提供虚拟磁盘的 Windows 桌面的虚拟机。

5.3 验证 Ceph 在桌面虚拟化中的应用效果

根据之前的测试结果与调优，通过计算资源的相互隔离，保证预留存储资源冗余，对缓存进行冷热分级后，从本文研究分析出的桌面虚拟化两个关键问题进行应用效果的检验。应用效果验证分布式存储是否能够满足启动风暴的 IOPS 需求，并在实际虚拟机应用负载下的支撑能力。

5.3.1 解决启动风暴问题

4 个计算节点按照测试方法重新测试结果，同时启动 80 台虚拟机，80 台虚拟机同时启动通过虚拟机内的 IOPS 性能测试软件记录进行汇总后监控情况如下图：

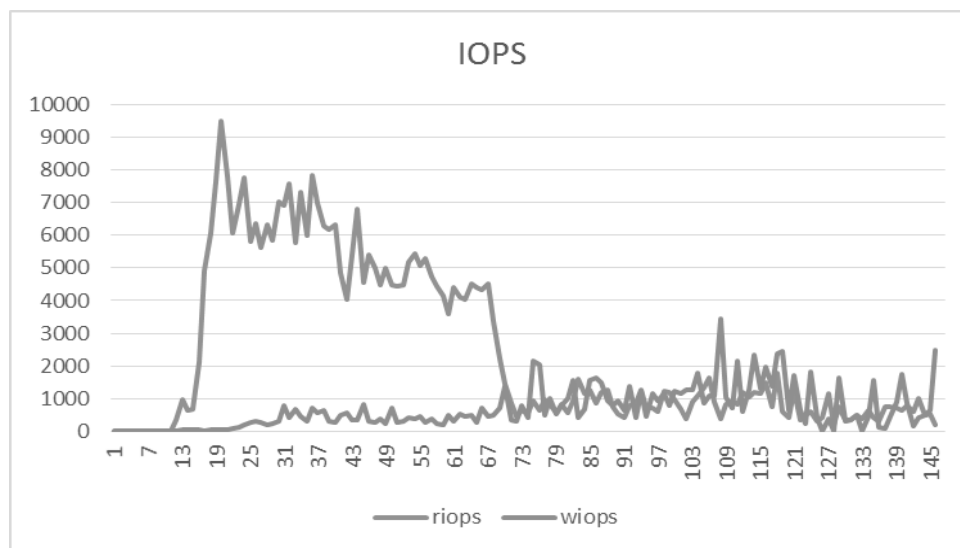


图 5-8 IOPS 监控数据曲线图

80 台虚拟机的总 IOPS 支撑能力超过 9000，平均每台虚拟机的 IOPS 可以达到 110

与原测试数据相符，并且通过调优后，整个分布式存储系统能够完全满足集群正常虚拟化能力的启动风暴的 IOPS 性能要求，虚拟机流畅运行，成功满足了虚拟机启动所需要的 IOPS，避免了 IO 锁死状态。

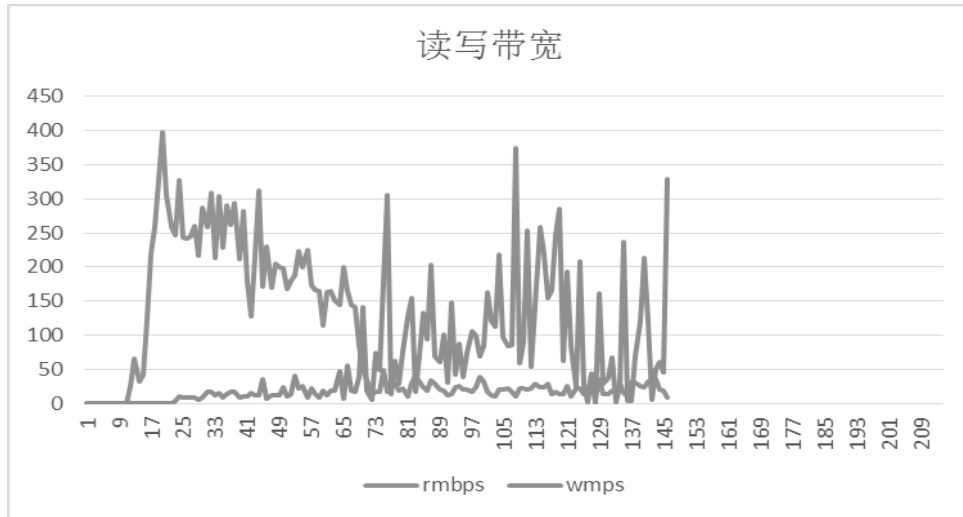


图 5-9 节点存储带宽监控数据曲线图

Ceph 分布式存储达到了大规模虚拟机并发 IO 访问的需求，并且在开机阶段的存储访问延迟也非常小，在开机后虚拟机内操作体验流畅，点击进入应用程序后才产生正常的读写延迟，读写延迟数据如图 5-10：

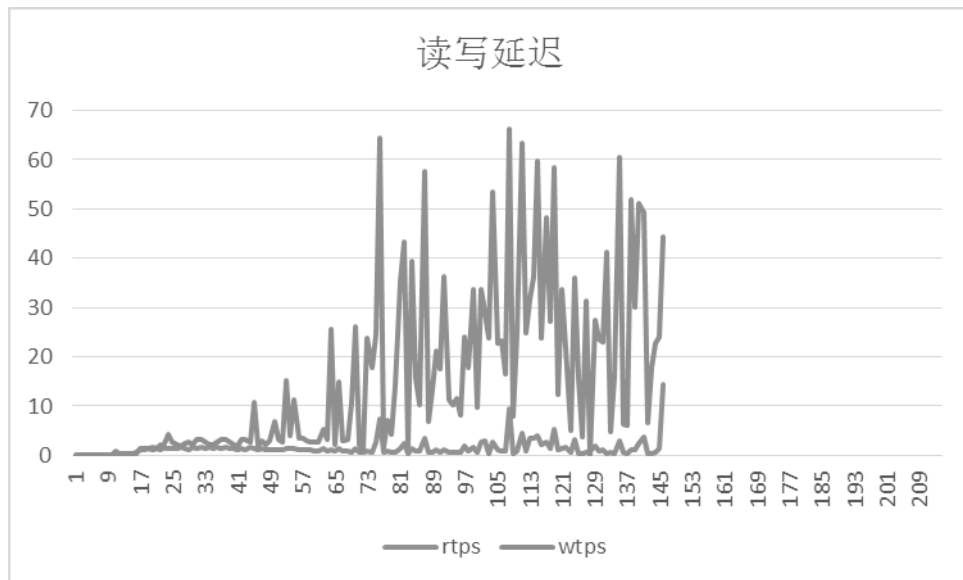


图 5-10 节点存储读写延迟监控数据曲线图

通过测试结果可以看到，测试结果和过程符合测试用例的设计与预计结果，而且所得到的优化结果也满足了测试的需要，验证了测试调优和方案可行性，改进后的存储解决了传统存储的启动风暴问题。

5.3.2 满足虚拟机的应用需求

在测试实际桌面虚拟化应用中，集群生成 80 台 win7 虚拟机同时启动并使用按键精灵脚本模拟办公，通过按键精灵脚本不断重复打开并浏览自动输入绘制 word, ppt, pdf, 进行正常的模拟办公软件操作。其测试数据结果如下：

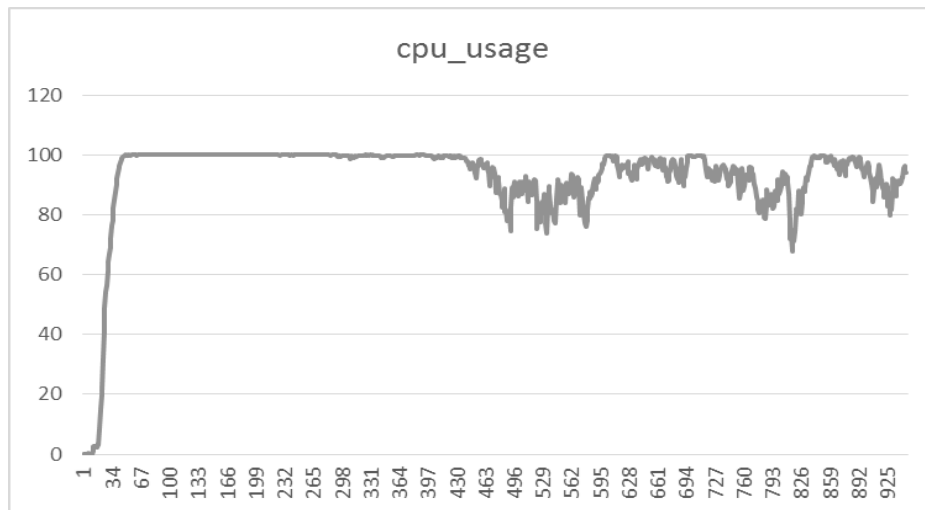


图 5-11 节点 CPU 使用率曲线图

在测试过程 CPU 使用率在这一指标已经达到满负载，应用使用率达到顶峰，对存储的使用也达到了最高需求。

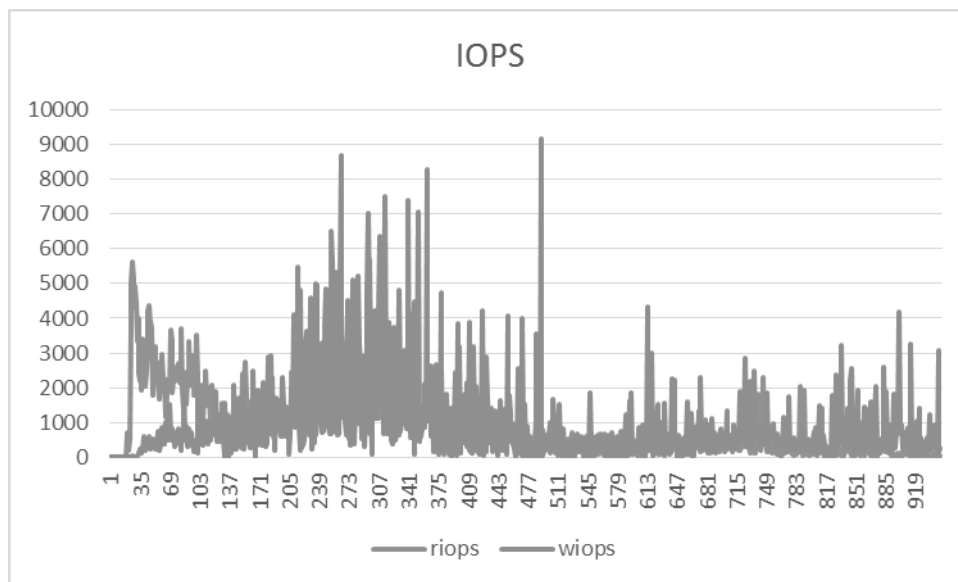


图 5-12 节点 IOPS 统计曲线图

分布式存储在 IOPS 性能方面，与启动风暴一致，分布式存储也能提供了中间出现的 9000IOPS 的应用程序读写频率的峰值，保证了应用程序的 IO 资源性能。

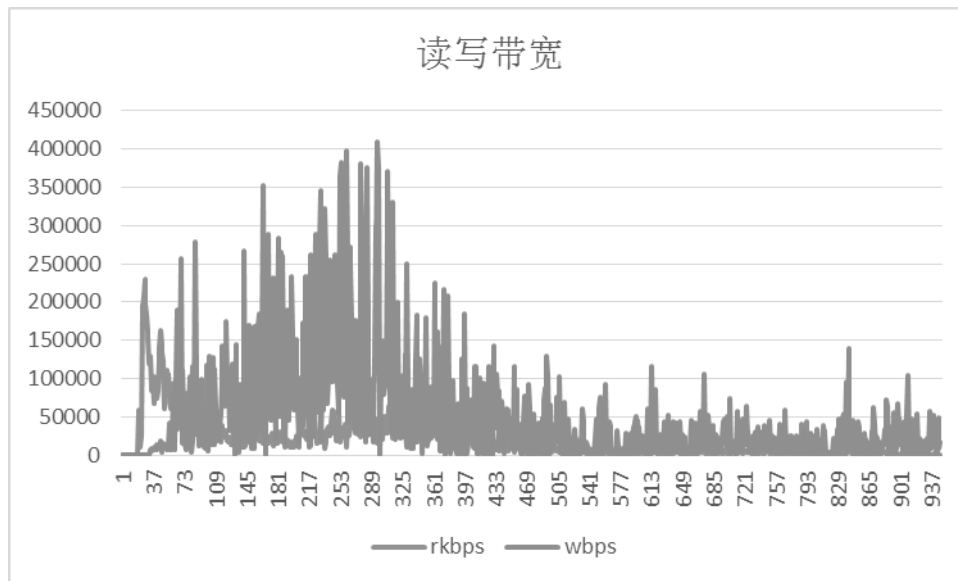


图 5-13 分布式存储读写带宽曲线图

读带宽达到 400Mbps 的性能峰值，即 50MB/s 的带宽性能，写性能在，250Mbps 即 31.25MB/s 左右，以与之前的测试基准对比，满足了虚拟机应用对后端存储的带宽实际需求，为业务提供了较高的读写性能。所以根据测试结果可以得出，分布式存储 Ceph 在经过调优和测试之后可以满足对大规模应用并发访问的存储性能需求。

5.4 本章小结

通过对分布式存储 Ceph 在桌面虚拟化中的应用，验证了 Ceph 的性能可以避免虚拟机启动风暴，并且在大并发访问下保持足够的存储带宽，为用户体验和应用提供了有效的支撑，也为以后工程实施方案的制定获得有效数据支持，同时也验证了分布式存储测试方案的可行性，完成了本文的研究目标。

结束语

随着云计算大数据的研究越来越深入，虚拟化技术带来的资源整合按需分配的方式受到更多人的认可，而桌面虚拟化的发展，从对存储系统的容量、吞吐量以及安全等方面有更高的要求。然而，分布式存储系统在性能分析研究、性能测试、性能优化等方面还有很多空白，也提供了更大的发展潜力。根据本文对分布式存储 Ceph 的测试和在桌面虚拟化中的应用，现总结本文所做的工作如下：

1. 对国内外桌面虚拟化中应用分布式存储的现状做出调查分析，并确立了本文的研究目标。
2. 根据参考文献与其他资料的学习，简单介绍了分布式存 Ceph 的工作原理和与常见的 GlusterFS 存储系统做出对比。
3. 针对桌面虚拟化的应用场景，对存储的使用方法和特点进行了分析，对桌面虚拟化的应用场景和存储性能需求进行了分析，并针对 Ceph 分布式系统测试进行方案设计。
4. 根据测试方案进行测试，得到本地硬盘，桌面虚拟化性能需求，分布式块存储的性能，最后对测试结果进行分析，并针对可以优化的地方重新作出测试调整，使之能满足更大规模的应用。
5. 最后在实际部署分布式存储 Ceph，并将其与桌面虚拟化后端连接，实际测试分布式存储 Ceph 的性能，根据结果判断最终满足了桌面虚拟化的需求。

以上就是本文的工作，通过这次论文的应用研究，学习到了很多关于分布式系统和虚拟化系统的知识，并且也发现了测试中还有很多可以更细致的测量点，相关的测试结果和调优方法也给对分布式存储 Ceph 在桌面虚拟化中的应用起到了推动作用。今后将在以下几方面进行更深入的学习研究：进一步学习分布式存储系统的特性，并尝试借鉴其他分布式存储的优点和设计思维，优化 Ceph 的存储性能，扩大 Ceph 的应用规模。根据这次的研究，也对今后开发云平台存储相关的监控起到很大帮助。也希望能在桌面虚拟化和 Ceph 的使用上探索更自动化的测试方法，并且根据应用场景特点修改数据分布算法，提高适应性能，增强其安全能力的情况下降低性能影响。

参考文献

- [1]. Pawlowski B, Juszczak C, Staubach P, et al. NFS version 3: Design and implementation[C] Summer Usenix Conference. 1994:137--151.
- [2]. 姜跃. 面向桌面虚拟化的分布式镜像存储研究[D]. 华中科技大学, 2012.
- [3]. 武杰. 虚拟化技术在大规模桌面网格中的研究与应用[D]. 中国科学院大学, 2013.
- [4]. O'Connor M A. Method of enabling heterogeneous platforms to utilize a universal file system in a storage area network: US, US 6564228 B1[P]. 2003.
- [5]. 张磊. 面向多学科虚拟实验平台的高可用分布式存储系统[D]. 华中科技大学, 2012.
- [6]. Soltis S R, Ruwart T M, O'keefe M T. The global file system[C]. NASA CONFERENCE PUBLICATION, 1996: 319-342.
- [7]. Schmuck F B, Haskin R L. GPFS: A Shared-Disk File System for Large Computing Clusters[C] Usenix Conference on File & Storage Technologies Monterey. USENIX Association, 2002:231--244..
- [8]. 刘明亮. 高性能云计算平台存储系统配置关键技术研究[D]. 清华大学, 2014.
- [9]. Menon J, Pease D A, Rees R, et al. IBM Storage Tank—a heterogeneous scalable SAN file system[J]. IBM Systems Journal, 2003, 42(2): 250-267.
- [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]. Ghemawat S, Gobioff H, Leung S T. The Google file system[C]. ACM SIGOPS Operating Systems Review, 2003, 37(5): 29-43.
- [13]. 刘碧薇. 基于虚拟化技术的管理服务部署及优化方法[D]. 北京邮电大学, 2015.
- [14]. Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data[J]. ACM Transactions on Computer Systems (TOCS), 2008, 26(2): 4.
- [15]. Weil S A. Ceph: Reliable, scalable, and high-performance distributed storage[D]. Berkeley: University of California, 2007.
- [16]. 程靓坤. 基于 Ceph 的云存储系统设计与实现[D]. 中山大学, 2014.
- [17]. Weil S A, Brandt S A, Miller E L, et al. Ceph: A scalable, High-Performance-Distributed File System[C]. USENIX Association, 2006: 307-320.
- [18]. 马思翔. 基于对象存储的混合块存储系统的研究[D]. 上海交通大学, 2015.

- [19].Josephson W K, Bongo L A, Li K, et al. DFS: A file system for virtualized flash storage[J]. ACM Transactions on Storage (TOS), 2010, 6(3): 14.
- [20].Eric Siebert. How-to-avoid-VDI-boot-storm-problems-using-SSD[R]. LAX: Intel, 2013
- [21].傅彦铭. 分布式软件性能测试工具设计及其关键模块的实现[D]. 西南交通大学, 2008
- [22].Weil S A, Leung A W, Brandt S A, et al. RADOS: a scalable, reliable storage service for petabyte-scale storage clusters[C]. International Workshop on Petascale Data Storage: Held in Conjunction with Supercomputing. ACM, 2007:35-44.
- [23].Marc Staimer. Choosing the right SSD application for addressing VDI boot storms[R]. Shenzhen: Intel Developer Forum, 2014
- [24].王东兴. 分布式块级别存储系统的设计与实现[D]. 哈尔滨工业大学, 2013.
- [25].Ray Sun. Ceph performance optimization summary. <http://xiaoquqi.github.io/blog/2015/06/28/ceph-performance-optimization-summary/>, 2015-06-28
- [26].王霄飞. 基于 OpenStack 构建私有云计算平台[D]. 广州: 华南理工大学, 2012
- [27].刘小飞. 大型分布式系统性能测试研究与实践[J]. 现代计算机(专业版), 2011 年 17 期:17-20
- [28].张毕涛, 辛阳. 基于 Ceph 的海量小文件存储的优化方法[R]. 中国辽宁沈阳: 中国通信学会, 2014
- [29].沈良好, 吴庆波, 杨沙洲. 基于 Ceph 的分布式存储节能技术研究[J]. 计算机工程, 2015, 41(8):13-17.
- [30].王迪. 基于 oVirt 的虚拟机池化平台压力测试系统的研究与实现[D]. 电子科技大学, 2015.
- [31].刘力. 基于被测系统虚拟化的云交叉测试方法的研究[D]. 东南大学, 2014

致 谢

硕士研究生学习生活就要结束了，本文将为我的研究生生涯画上一个句号，两年半的时间里，很多老师和同学给予了我无私的帮助和指导，我才能顺利地完成我的学业。在论文写作期间，我有幸得到了袁华老师的指导，衷心感谢袁华老师的教导和培育。老师有渊博的计算机学术知识，严谨的治学态度，这令我受益匪浅，对我今后的工作有很大的帮助。另外，在毕业论文的写作过程中，从最初选题，到关键点研究，再到最后的写作，老师给了我很多指导，严格要求提出很多有益的建议。

在研究生这个阶段，我有幸认识一起就读研究生课程的同学，与他们不多的相处却让我感受到了同学的深厚友谊！并有幸认识了李惊生总经理并担任我的校外导师，和工作中多位同事，衷心的感谢他们在我工作和学习期间对我无私的指导，让我在在项目中磨练自己，不断成长。

最后感谢我的父母，在我的研究生学习阶段给予的理解与支持。

IV - 2 答辩委员会对论文的评定意见

洪亮同学的硕士论文“开源分布式存储系统 Ceph 测试及在桌面虚拟化平台中的应用”基于实际问题，所研究的内容具有一定的实用和推广价值。

论文作者不但对分布式开源存储系统 Ceph 在桌面虚拟化应用场景中应用潜力进行研究，还深入研究了其调优问题，设计出一套在该应用场景下具体可行的测试方案，并通过实验验证了方案的可行性。

论文对国内外相关研究有一定的了解，研究内容明确具体，研究思路清晰，实验详尽，论文工作难度及工作量适中，反映作者具有较强的理论基础、较好的科研和工程能力，达到硕士学位论文水平。

学位申请人答辩过程讲述清楚，对评阅意见中提出的问题或质疑已作出明确的回复，答辩委员判定学位申请人的回复已达到评阅专家的要求。答辩过程讲述清楚，回答问题基本正确。经答辩委员会无记名投票，同意该同学通过硕士论文答辩，同意授予硕士学位。

论文答辩日期：2016 年 6 月 5 日

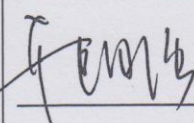
答辩委员会委员共 5 人，到会委员 5 人

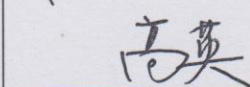
表决票数：优秀（0）票；良好（4）票；及格（1）票；不及格（0）票

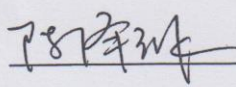
表决结果（打“√”）：优秀（ ）；良好（√）；及格（ ）；不及格（ ）

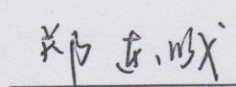
决议：同意授予硕士学位（√） 不同意授予硕士学位（ ）

答辩
委员
会成
员签
名

 (主席)

 高英



 郑东

