

非对称集群存储系统的块虚拟化方法

范中磊

(长安大学 信息工程学院, 陕西 西安 710064)

摘 要: 非对称集群存储系统以其在高 I/O 性能、高可扩展和易管理等方面的优势而成为研究的热点. 存储虚拟化是提高其可扩展性和易管理性的关键技术之一. 从构建智能存储路由器的思路出发, 提出了一种分层映射的块虚拟化方法, 可实现资源的集中式管理和分布式映射. 该方法支持物理存储资源的动态添加, 从而可使整个存储系统具有良好的可扩展性和可管理性.

关键词: 集群存储; 存储虚拟化; 块虚拟化; 可扩展; 可管理

中图分类号: TP333

文献标识码: A

文章编号: 1000-7180(2011)06-0005-03

Block Virtualization Method for Asymmetric Cluster Storage System

FAN Zhong-lei

(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: Asymmetric cluster storage systems are becoming a focus due to the high I/O performance, high scalability, easy management, etc. Storage virtualization is one of the key technologies to improve the scalability and manageability of cluster storage systems. From the perspective of building an intelligent storage router, this paper presented a block virtualization method with layered mapping, by which central management and distributed mapping for resource were implemented. This approach supports physical storage resources to be joined in storage system dynamically, and enables good scalability and manageability for the whole storage system.

Key words: cluster storage; storage virtualization; block virtualization; scalability; manageability

1 引言

集群存储系统通过将集群技术应用到存储系统, 可以有效地满足数据的爆炸性增长对存储系统在管理和可扩展性等方面的需求. 从系统结构的角度出发, 集群存储系统可以划分为对称和非对称结构两种形式^[1-2].

近年来, 非对称结构的集群存储系统以其在高 I/O 性能、高可扩展能力和易管理等方面的优势而受到广泛重视和应用. 非对称集群存储系统所使用的物理存储设备主要包括块存储设备、对象存储设备和文件存储设备三种形式^[3-5].

存储虚拟化技术通过添加虚拟化层, 将实际的物理存储实体与存储的逻辑表示分离开来, 从而解除了应用服务器和物理存储设备的绑定关系, 它是

提高集群存储系统可扩展性、可管理性, 降低成本的核心技术之一^[6-8].

本文针对采用块存储设备的非对称集群存储系统, 从构建智能存储路由器的思路出发, 提出了一种分层映射的块虚拟化方法, 该方法可实现资源的集中式管理和分布式映射, 支持物理存储资源的动态添加, 从而使整个系统具有良好的可扩展性和可管理性.

2 分层映射块虚拟化方法

图 1 给出了采用分层映射块虚拟化方法的非对称集群存储系统的逻辑结构. 其中, 应用服务器组成应用服务器集群, 对外提供服务, 系统支持应用服务器的动态添加; 元数据服务器构成元数据服务器集群, 提供元数据管理服务. 元数据服务器集群可以由两台元数据服务器构成的高可用集群, 也可以是

由多元数据服务器构成的具有负载均衡能力的服务集群;存储服务器构成存储服务器集群,它为应用服务器提供存储空间,数据访问是在它和应用服务器之间直接完成的,无需经过元数据服务器。这样,元数据服务器只承担少量文件系统元数据相关的工作负载,而大量的数据 I/O 负载都由存储服务器承担。据 Panasas 统计^[3],采用非对称结构的集群存储系统中,大约有超过 90% 的负载由存储系统承担。

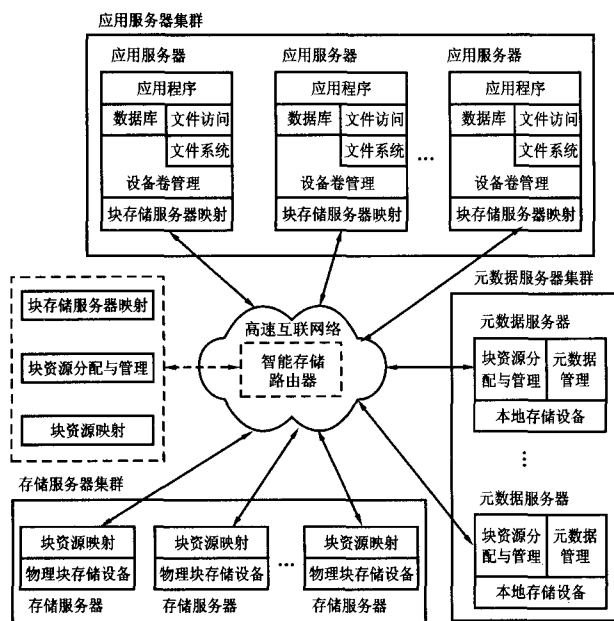


图1 分层映射块虚拟化方法

基于智能存储路由器的思路,本文提出了分层映射的块虚拟化方法。智能存储路由器就是假想的位于网络中的存储专用的路由和交换设备,它可以是由多个智能存储路由器组成的集群。智能存储路由器主要实现逻辑和物理块资源的分配和管理、应用服务器上逻辑块到存储服务器的映射,以及存储服务器上逻辑块到物理块的映射;智能存储路由器提供交换功能,实现数据在应用服务器和存储服务器之间的传递。分层映射就是将逻辑块到物理块的映射分成逻辑块到存储服务器的映射以及逻辑块到该存储服务器上物理块的映射两个步骤。

智能存储路由器目前的实现是将块存储服务映射、块资源分配与管理以及块资源映射三个模块分别放在应用服务器、元数据服务器和存储服务器。也就是说,逻辑块到存储服务器的映射在应用服务器完成,逻辑和物理块资源的分配与管理由元数据服务器完成,逻辑块到物理块的映射由存储服务器完成。将来的实现可以将这三个模块集成在一个真实的智能存储路由器设备中,实现基于网络设备的存储虚拟化。

3 资源分配与映射

资源的分配是由某个元数据服务器承担的。当有多个元数据服务器时,则通过选举产生负责资源分配的元数据服务器。资源的分配是基于资源当时的状况和预定义的分配策略,由资源分配和管理模块进行决策来完成的。资源分配的基本流程如下:

(1) 当一个存储服务器加入系统后,会向资源分配和管理模块注册,注册内容包括物理存储资源的信息(如空间大小、空间属性)、访问协议、访问端口等;

(2) 资源分配和管理模块接收到存储服务器的注册请求后,将注册信息添加到全局物理资源表,该表用于记录有关物理资源的管理、分配和使用情况;

(3) 当一个应用服务器加入系统后,它也会向资源分配和管理模块注册,注册内容包括节点信息、用户信息和资源需求信息等;

(4) 资源分配和管理模块接收到应用服务器的注册请求后,根据当时的物理资源状况和用户预定义的策略进行决策,为该应用服务器分配全局逻辑空间以及所对应的物理存储服务器,并将相关信息记录到全局逻辑映射表;之后它会通知相应的存储服务器该逻辑空间的物理存储空间由其提供;

(5) 存储服务器在收到来自资源分配和管理模块的资源分配信息之后,建立逻辑块资源到物理块资源的映射,管理并完成将逻辑数据块写入相应的物理块设备。

这里有两点需要说明。一是资源分配和管理模块既支持基于策略的自动分配,也提供接口供管理员手动进行分配。二是为了提高物理设备的共享能力,存储服务器上逻辑块资源与物理块资源的映射是在有数据写入,真正需要物理存储空间时建立的,而不是接收到资源分配和管理模块所分配的逻辑块资源时立刻就建立的。即映射是动态的、按需的,而不是静态的。

逻辑资源到物理资源的映射是分层的和按需的。用户对存储资源进行访问时,主要利用的是应用服务器上的块存储服务模块所维护的局部逻辑映射表,这个映射表是资源分配和管理模块所维护的全局逻辑映射表的一个缓存,只含有应用服务器常用的映射关系记录。资源映射采用类似 Linux 内核中内存管理的寻页缺失(page fault)机制,其基本过程如下:

(1) 用户对逻辑块资源进行访问,块存储服务模块首先根据用户请求地址检索局部逻辑映射表,如检索命中,得到存储服务器的访问信息,直接访问存储服务器。如未命中,则执行如下(2)~(4);

- (2) 请求资源分配和管理模块更新局部逻辑映射表;
- (3) 资源分配和管理模块更新局部逻辑映射表;
- (4) 块存储服务器模块检索局部逻辑映射表命中,访问相应的存储服务器;
- (5) 存储服务器的块资源映射模块接收逻辑块访问请求,查找映射表,找到对应的物理块号,完成数据读写访问.如果是写操作,且没有建立映射,则分配物理块资源,建立映射关系,完成数据写操作.

4 系统实现与测试

基于上述方法,在 Linux 下实现了一个具有基本功能的原型系统.该系统包括三个子系统:系统管理子系统、虚拟化子系统和系统软件子系统.其中:系统管理子系统包括存储系统服务、用户系统服务和管理台服务;虚拟化子系统包括虚拟空间服务(包括虚拟空间的配置管理服务和虚拟空间的监视服务)、虚拟化软件、虚拟设备驱动和物理设备服务(包括物理设备管理服务和物理设备监视服务);系统软件子系统包括 I/O 处理、配置代理、管理代理、监视代理、事件处理及操作系统核心.

表 1 给出了系统测试的硬件环境.系统配置了 2 台存储服务器,存储服务器上的百兆网口用于元数据管理,千兆网口用于数据传输,用两台主要是为了测试系统的可扩展性,即动态添加存储空间的能力;由于系统采用的是数据和元数据分离的非对称结构,1 台元数据服务器足以支撑大量的应用服务器请求,暂不考虑元数据服务器失效和大量应用服务器时可能成为瓶颈的问题;应用服务器 2 台用千兆口,2 台用百兆口,主要用来模拟不同应用服务器的配置和测试系统的网络性能.

表 1 系统测试的硬件配置

设备名称	数量	配置
存储服务器	2	2.4 GHz Xeon CPU;2 GB 内存;40 GB 系统盘,480 GB 数据盘;1 个百兆口,1 个千兆口
元数据服务器	1	1.5 GHz P4 CPU;512 MB 内存;40 GB 系统盘;1 个百兆口
应用服务器 A	2	1.5 GHz P4 CPU;512 MB 内存;40 GB 系统盘;1 个千兆口
应用服务器 B	2	1.5 GHz P4 CPU;512 MB 内存;40 GB 系统盘;1 个百兆口
网络交换机	1	24 口,百兆/千兆自适应

该原型系统包括三个子系统,子系统又包括若

干个子模块表.测试内容主要包括各子模块的单元测试、各子模块构成的子系统的局部集成测试、各子系统构成的整个系统的集成测试等.测试内容涵盖了虚拟和物理空间的自动与手动分配、静态和动态的资源分配与映射、虚拟和物理空间的动态扩展能力、元数据服务器的管理与请求处理能力(用于验证元数据服务器不会成为系统瓶颈)、大数据量的 I/O 能力、网络带宽性能等.测试结果证明了该方法的正确性和可行性.

5 结束语

尽管本文所述方法在原型系统得以实现和验证,但是还需要进一步的研究和完善.例如,如何与其他已有的非对称集群存储系统进行集成,如何实现元数据服务器的扩展和高可用,如何构建集成的智能存储路由器设备,系统可扩展能力的测试与评估等.

参考文献:

[1] Wikipedia, the free encyclopedia. List of file systems [EB/OL]. [2010-08-07]. http://en.wikipedia.org/wiki/List_of_file_systems.

[2] 朱晓辉,陈苏蓉. 基于流媒体的小型分布式视频点播系统研究[J]. 微电子学与计算机, 2010, 27(10): 69-73,78.

[3] PANASAS, INC. Panasas white paper: object storage architecture[EB/OL]. (2007-04-01)[2010-10-25]. http://www.panasas.com/docs/Object_Storage_Architecture_WP.pdf.

[4] Microsystems Sun, Inc Lustre file system: high-performance storage architecture and scalable cluster file system [EB/OL]. (2008-10-01)[2010-10-25]. http://wiki.lustre.org/index.php/Lustre_Publications.

[5] PVFS2 Development Team. Parallel virtual file system, version 2 [EB/OL]. (2003-09-01)[2010-10-25]. <http://www.pvfs.org/cvs/pvfs-2-8-branch-docs/doc/pvfs2-guide.pdf>.

[6] 曾龙海,张博锋,张丽华,等. 基于云计算平台的虚拟集群构建技术研究[J]. 微电子学与计算机, 2010, 27(8): 31-35,40.

[7] Nahum Nelson. 高性能存储虚拟化架构[J]. 软件世界, 2007(20):65-67.

[8] 谭生龙. 存储虚拟化技术的研究[J]. 微计算机应用, 2010, 31(1):33-38.

作者简介:

范中磊 男,(1970-),博士,教授.研究方向为分布式计算与存储.