一种采用带外虚拟化技术的网络存储系统

孟晓烜[1,2],那文武[1,2],朱旭东[1,2],柯 剑[1,2],许 鲁[1]

[1](中国科学院 计算技术研究所,北京 100080)
[2](中国科学院 研究生院,北京 100039)
E-mail: mengxx@ict.ac.cn

摘 要: 介绍一种基于带外虚拟化技术的网络存储系统,简称 BW-VSDS,它具有以下特点: (1)采用两级带外虚拟化数据管理模型以充分发挥单个存储节点的 I/O能力并释放存储网络的承载能力; (2)采用分布式数据存储管理协议以协同多个存储节点有效实现高级数据存储语义; (3)支持多种数据传输协议以适用于不同的应用环境. 目前该系统已应用于视频监控、信息处理和企业办公等多个领域.

关 键 词: 网络存储;存储区域网;存储虚拟化;带外数据管理
中图分类号: TP333 **文献标识码**: A **文 章 编 号**: 1000-1220(2009)11-2123-05

Network Storage System Using Out-of-band Virtualization Approach

MENG Xiao-xuan[1,2], MENG Wen-wu[1,2], ZHU Xu-dong[1,2], KE Jian[1,2], XU Lu[2]
[1] (*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*)
[2] (*Graduate School of the Chinese Academy of Sciences, Beijing* 100039, *China*)

Abstract: This paper presents a network storage system using out-of-band virtualization approach, called BW-VSDS. It has following characteristics: (1) adopt a two-level out-of-band virtualization data management model which can fully utilize the I/O capacity of individual storage devices as well as maximize the utilization of underlying storage network; (2) use distributed data storage management protocols which coordinate multiple independent storage devices to realize advanced data storage semantics; (3) support multiple network data transport standards in case of different applications. Presently, BW-VSDS has applications in fields of video monitor, information processing and enterprise computing, etc.

Key words: network storage; storage area network; storage virtualization; out-of-band data management

1 引 言

随着信息存储量的爆炸式增长和数据重要性的日益凸显,网络存储正在逐步取代传统直连式存储 (DAS, direct attached storage),它具有以下优点:

(1) 提高存储资源利用率,研究表明[1]直连存储中的存储设备利用率仅为 40%,而网络存储中则高达 90%;

(2) 降低存储系统的总体拥有成本 (TCO, total cost of owner),存储管理成本中最主要的是人力成本,网络存储可以有效降低大规模存储中的人力成本的开销.

据统计[1],直连存储的 TCO约为 0.84 $/MB,而网络存储则仅为 0.35～0.38 $/MB.

存储虚拟化[2]是网络存储系统中普遍采用的一种数据管理技术,它通过一定手段实现对存储资源的集中式管理,屏蔽了组成物理存储介质的异构性并为使用者提供大容量、高性能和多功能的存储系统. 网络存储虚拟化可以分为带内和带外两类[3,4],如图 1所示,在带内虚拟化网络存储系统中,所有数据通路都经由唯一 I/O导向器,它不仅负责存储资源的虚拟化管理同时在数据通路上实现各种数据存储管理语义;在带外虚拟化网络存储系统中,应用服务器和存储设备直接进行数据传输,存储资源的管理由独立于数据传输通路的存储虚拟化服务器提供. 相比之下,一方面,带内虚拟化方式不仅存在由 I/O导向器带来的性能瓶颈和单点故障等问题,同时它不能充分发挥交换式存储网络的承载能力和单个存储设备的 I/O能力,此外 I/O导向器的 I/O转发操作也会相应增加数据传输通路的 I/O延迟;另一方面,通过对数据传输通路的有效控制,带内虚拟化方式易于实现各种高级数据存储管理语义,如在线数据迁移、数据复制和数据版本 (快照)控制等.

目前网络存储从系统结构上主要分为 SAN (storage area network)和 NAS (network attached storage)两类[1],其中前者由专用的网络和设备构建,提供块级数据访问接口,而后者基于服务器直连盘阵架构,在传统数据网上提供文件级数据访问接口. 本文主要介绍由国家高性能计算机工程技术中心自主研制的网络存储系统 --BW-VSDS (Blue Whale - virtual storage device system),它是一种基于带外虚拟化技术的 SAN系统,

收稿日期: 2008-06-17 基金项目:国家"九七三"重点基础研究发展规划基金项目 (2004CB318205)资助. 作者简介:孟晓烜,男,1980年生,博士研究生,研究方向为网络存储,缓存管理;那文武,男,1977年生,博士研究生,研究方向为网络存储,RAID技术;朱旭东,男,1979年生,博士研究生,研究方向为网络存储,数据访问模式挖掘.

目前已实际应用于视频监控、科学研究和企业办公等领域. 相比于其它 SAN系统,BW-VSDS具有以下特点:

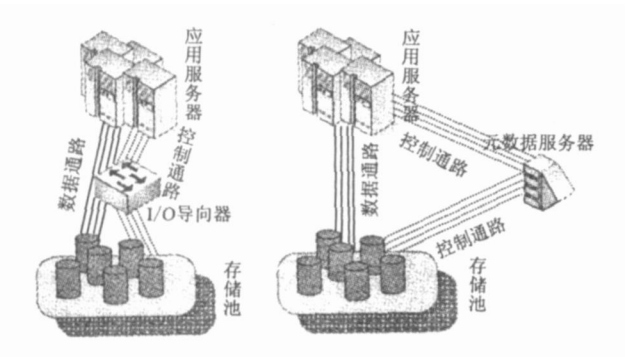

图 1 网络存储虚拟化原理示意图

Fig 1 Illustration of network storage virtualization

(1)采用两级带外虚拟化数据管理模型以充分发挥单个存储节点的 I/O能力并释放存储网络的承载能力;

(2)采用分布式数据存储管理协议以实现高级数据存储管理语义;

(3)支持多种块级数据传输协议以适用于不同的应用环境.

2 基本原理

BW-VSDS系统由元数据服务器、存储节点和应用服务器三种功能实体组成,如图 2所示,元数据服务器采用带外虚拟化方式将分布在多个存储节点中的离散存储资源聚合为一个

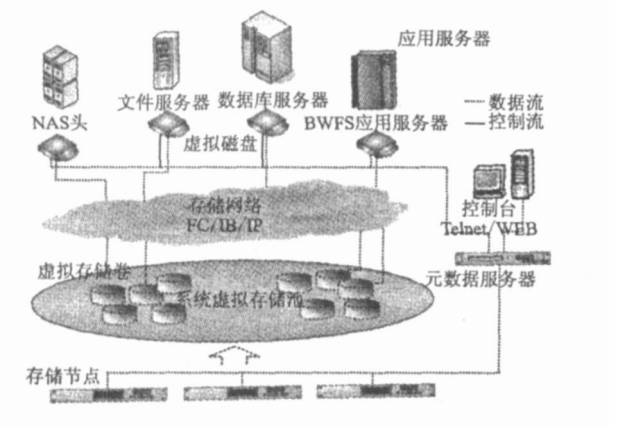

图 2 BW-VSDS系统原理示意图

Fig 2 Illustration of BW-VSDS architecture

统一的虚拟存储池,它根据应用的具体需求划分出具有不同属性的虚拟存储卷并授权给相应的应用服务器,后者则通过存储代理以虚拟磁盘的方式直接访问位于存储节点中的存储资源. 本小节我们将阐述为实现上述功能 BW-VSDS所采用的两项关键技术.

2.1 两级带外虚拟化数据管理模型

BW-VSDS系统中将存储虚拟化功能划分为存储节点内和存储节点间两级 (如图 3所示):

本地虚拟化:存储节点通过 RAID或基于磁盘属性的分组池化技术集中管理与其直连的物理存储资源 (磁盘或磁盘阵列),在此基础之上,它以逻辑存储卷为单位对存储资源进行划分,并在节点内提供各种高级数据存储管理功能,如数据读写缓存、数据复制、数据版本控制和数据加密等,其中逻辑存储卷是 BW-VSDS系统中最小的可管理单元.

全局虚拟化:元数据服务器集中管理系统中的存储资源,它首先将存储节点中离散的存储资源聚合为一个统一的虚拟存储池,接着根据应用服务器的具体需求为其分配具有相应属性 (存储空间的大小、数据的可靠性和读写性能等)的虚拟存储卷,其中后者是由分布于不同存储节点中的逻辑存储卷在按照一定的地址映射规则聚合而成,位于应用服务器中的虚拟磁盘驱动 (见下页图 5)根据相关元数据 (对应于图 3中的静态路由转发表)完成数据读写的转发.

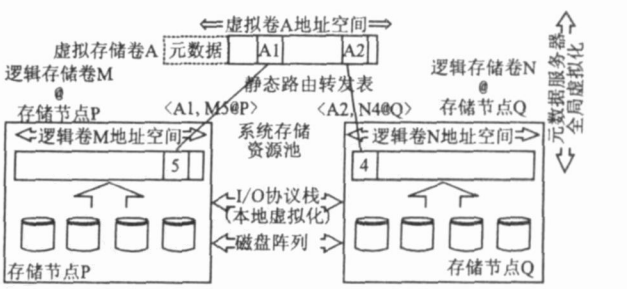

图 3 两级虚拟化管理模型

Fig 3 Two layer virtualization management model

相比于 MagicStore系统[4]中采用的带外虚拟化技术,BW-VSDS系统的特点在于将数据存储的管理和存储资源的管理进一步分离,其中元数据服务器只专注于系统存储资源的管理,而各种高级数据存储语义由每个存储节点在本地提供,这样彻底将元数据服务器从数据传输通路中释放出来,这不仅减轻了前者的负载,提高了系统的可扩展性,同时避免了 MagicStore系统在数据读写过程中访问动态元数据而引入的额外网络延迟开销,从而进一步降低了端到端的数据传输延迟. 这些相互独立的存储节点之间则通过相应的分布式数据存储管理协议以协作 (见 2.2节)实现对虚拟存储卷的高级数据存储功能.

2.2 分布式数据存储管理

如前所述,带内虚拟存储系统中的 I/O导向器通过重定向应用服务器的读写请求在数据传输通路中实现各种高级数据存储管理语义;而带外虚拟存储系统的特点在于元数据服务器并不直接参与数据的读写,为了实现多个存储节点之间的协同工作需要某种分布式数据管理机制的支持,MagicStore系统采用一种存储空间的动态影射技术,其基本原理为:为了完成一次读写请求,应用服务器需首先向元数据服务器查询完成该读写操作所需的地址影射信息,因此元数据服务器虽然不支直接参与读写转发但实际控制着读写操作的进程. 在这类带外虚拟化存储系统中,数据存储管理功能对存储节点是透明的,其优点在于实现简单而代价是增加了每次读写操

作的延迟.相反,在 BW-VSDS系统中,智能化的存储节点在本地实现对单个组成逻辑存储卷的数据存储管理,而多个独立存储节点在元数据服务器的统一协调下协同工作,从而不仅实现了对复合虚拟存储卷的数据存储管理,同时保证了存储数据的一致性.目前 BW-VSDS系统已实现了对虚拟存储卷的在线扩容、数据版本控制和在线数据复制等 3种分布式数据存储管理协议.下面我们主要介绍数据版本控制协议的基本工作原理 (如图 4所示):

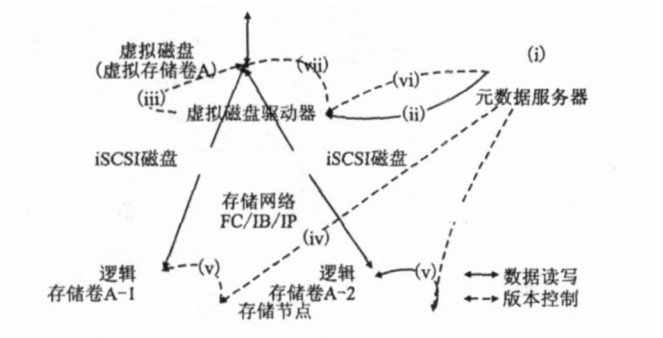

图 4　分布式数据版本控制协议原理示意图

Fig 4　Illustration of distributed data versioning control protocol

数据版本控制用以解决虚拟存储卷快照创建时多个组成逻辑存储卷的数据版本一致性问题.

(1)系统管理员为满足应用或系统在线数据备份等需求,分别从应用服务器或元数据服务器发起虚拟存储卷的快照创建请求;

(2)元数据服务器在接受该请求后向当前在线共享该虚拟卷的所有应用服务器发送快照创建命令;

(3)后者依次执行对前台读写请求的阻塞和本地缓存的同步;

(4)当所有应用服务器都完成 (3)步操作后,元数据服务器向相关存储节点发起对组成逻辑存储卷的本地快照创建命令; (5-7)元数服务器在所有存储节点完成快照创建操作后,通知应用服务器快照操作完成,后者随即释放正被阻塞的读写请求.由上可知一次虚拟存储卷的快照创建操作的总用时为 $T_{VirtSnapOpt}=\max\{2.5RTT_{net}\}+\max\{T_{LogicalSnapOpt}\}$,其中在高速局域网环境下 RTT_{net}为毫秒级,而 $T_{LogicalSnapOpt}$也为毫秒级,因此该操作不仅在实现上对前台应用是透明的同时也几乎不影响应用的读写性能.

3　系统实现

如图 5所示,BW-VSDS系统软件从结构上可划分为全局虚拟化、本地虚拟化和存储代理三个子系统,他们分别运行于系统中的三种功能实体:元数据服务器、存储节点和应用服务器.本节将依次介绍这三种子系统各自不同的软件组成 (参见表 1)和系统功能,以及三者之间的通讯机制.

3.1　全局虚拟化子系统

该子系统运行于元数据服务器是整个系统的核心枢纽,它一方面将位于后端存储节点中离散存储资源聚合统一管理;另一方面以虚拟存储卷为单位为前端应用服务器分配所需的存储资源.它由全局虚拟化引擎、系统管理接口,Web服务器三个模块构成,其中全局虚拟化引擎实现了系统存储资源管理机制及多种资源管理策略 (具体参见文献 [5]);系统管理接口将用户管理员命令转换为对全局虚拟化引擎的相应操作;Web服务器对外提供功能丰富的系统管理界面以方便管理员在控制台实施远程访问.

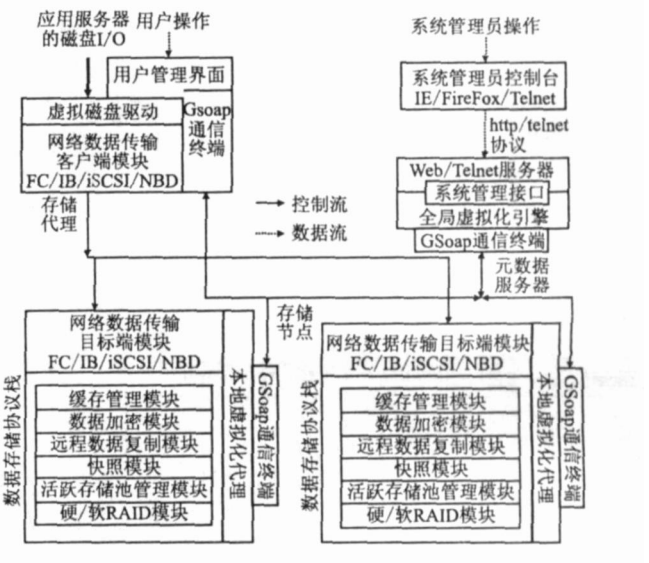

图 5　BW-VSDS系统软件结构示意图

Fig 5　Software structure of BW-VSDS

3.2　本地虚拟化子系统

该子系统运行于存储节点,它主要实现以下功能:

(1)代理全局虚拟化子系统完成以逻辑存储卷为单位的本地存储资源管理;

(2)协助全局虚拟化子系统实现对逻辑存储卷的多种数据存储管理语义;

(3)为前端应用服务器提供网络块级 (或文件级)数据访问接口.

如图 5所示,该系统包括: I/O协议栈、本地虚拟化代理和网络数据传输目标端等 3个模块,其中 I/O协议栈在 Linux平台下实现为一组层叠式虚拟块设备驱动,它完成对本地存储资源的池化管理并提供各种高级的数据存储语义,每个逻辑存储卷可根据应用的需求相应配置其 I/O协议栈 (图 5中给出了目前已实现的各模块);本地虚拟化代理通过配置 I/O协议栈和网络数据传输目标端以代理完成从元数据服务器接收到的各种存储管理命令;网络数据传输目标端模块支持多种业界标准的网络数据传输协议,包括 FC、IB、iSCSI、NBD和 NFS等五种,其中 NBD是针对小型局域网设计的一种轻量级数据传输协议,而 NFS作为一种文件级数据访问协议,适用于单存储节点的 NAS应用模式.

3.3　存储代理

该子系统运行于应用服务器,它根据从元数据服务器获取的关于虚拟存储卷的元数据信息,在应用服务器本地主机

系统内以虚拟磁盘的方式提供对后端存储节点中存储资源的访问接口.如图 5所示,存储代理由虚拟磁盘驱动、网络数据传输客户端和用户管理界面三个模块组成,其中虚拟磁盘驱动在功能上相当于逻辑卷管理器,其基本原理为:根据虚拟存储卷中元数据信息,在数据传输通路中完成从虚拟存储卷地址空间到逻辑存储卷地址空间的线性影射,其中元数据信息组织为一组静态路由转发表 (见图 3),系统目前已支持线性叠加和条带化两种地址影射方式;网络数据传输客户端同时支持 3.2小节中所述的各种数据传输协议,它将从虚拟磁盘驱动接收到的本地读写请求打包封装为一组数据传输协议帧,通过存储网络按序发送至存储节点,并由后者完成数据的物理存取;用户管理界面响应来自用户的各种管理操作,如虚拟存储卷的上线/离线、快照创建等,此外它还响应从元数据服务器接收到的各种异步管理命令或消息 (见 2.2节),如在线扩容通知、快照创建命令等.

表 1 软件子系统运行平台及模块实现语言

Table 1 Platform and Implementation language of software subsystems

软件子系统	系统平台	组成模块	实现语言
全局虚拟化子系统	Linux	系统虚拟化引擎	Perl
		系统管理接口	Perl
		Web服务器	Java/Perl
		GSoap通信终端	Perl
本地虚拟化子系统	Linux	本地虚拟化代理	Perl/C
		I/O存储协议栈	C
		网络传输目标段	C
		GSoap通信终端	Perl
存储代理子系统	Linux/Window	用户管理界面	C/VC++
		虚拟磁盘驱动	C
		网络传输客户端	C
		GSoap通信终端	C/C++

3.4 子系统间的通讯机制

如图 5中所标志的带外控制流,在 BW-VSDS系统中我们统一采用 GSoap协议实现软件子系统间的带外存储管理通讯,包括以下两方面:

元数据服务器与存储节点:实现对逻辑存储卷的各种管理操作以及对存储节点的状态查询等操作交互;

元数据服务器与应用服务器:实现对虚拟存储卷的各种管理操作以及对系统的状态查询等操作交互,其中前者包括:虚拟存储卷的上线注册、离线注销、元数据获取、在线扩容、快照创建等操作;后者包括:授权虚拟卷的列表查询等操作.

4 数据通路分析

本节首先介绍 BW-VSDS系统中端到端的数据传输通路,随后给出针对该数据传输通路的 I/O性能评测结果.

4.1 数据传输通路

图 6给出了 BW-VSDS系统中完整的数据传输通路,下面我们以读操作为例简述系统 I/O处理的主要流程 (假定网络数据传输协议为 NBD):

(1)应用服务器本地文件系统接收到应用程序的读请求,若本地文件系统缓存命中则直接返回,否则将请求转发至虚拟磁盘驱动;

(2)虚拟磁盘驱动首先根据请求的地址偏移在静态路由转发表中查询对应的地址映射项,后者描述了请求数据所在的网络地址 <存储节点, 逻辑存储卷, 地址偏移 >;接着根据查询得到的网络地址将请求转发至对应的 NBD客户端;

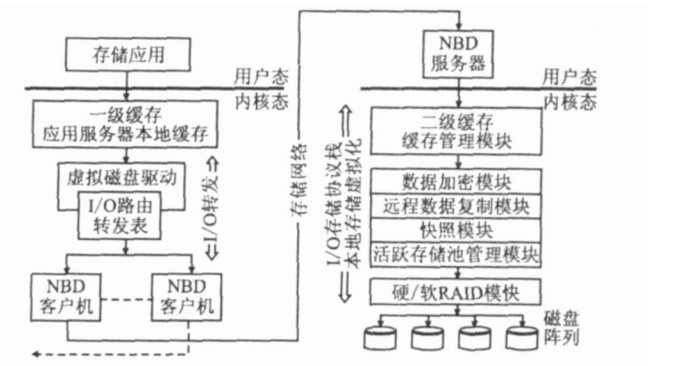

图 6 BW-VSDS中数据传输通路剖析

Fig 6 Analysis of I/O path in BW-VSDS

(3)NBD客户端将收到的请求打包封装为 NBD数据帧,并发送至后端 NBD服务器 (注:NBD数据传输链接在虚拟磁盘创建时已预先建好);

(4)NBD服务器将收到数据帧解析还原为初始读请求,并通过系统 I/O接口发送给 I/O协议栈中的缓存管理模块;

(5)若请求数据在缓存中命中,则请求处理完毕,否则该请求被转发至 I/O协议栈中的下一层模块;

(6)在依次经过 I/O协议栈中各层处理后,请求最终到达实际存储的物理磁盘;

(7)磁盘控制器在结束读操作处理后,读取的数据沿原路依次返回至请求发起的应用程序;

4.2 性能评测

本小节对上述数据传输通路进行 I/O性能评测,实验环境配置参见表 2, xdd是美国明尼苏达大学开发的磁盘性能测试工具.我们对比测试应用服务器中虚拟磁盘在单/双存储节点两种模式下的读/写性能,其中在双节点模式中地址映射采

表 2 实验环境配置

Table 2 Experiment settings

存储节点	Gentoo-2.6.12-SMP Kernel
	Intel(R) Xeon(TM) 2.4GHz CPU
	2 * 512MB DDR2 Memory
	3ware-9500RAID Card
	RAID5: 7(d5 + p1 + s1) ×SATA 7200RPM
	Intel(R) 82573E Gigabit Ethernet Controller
应用服务器	Gentoo-2.6.12-SMP Kernel
	Intel(R) Xeon(TM) 2.4GHz CPU
	2 * 512MB DDR2 Memory
	Intel(R) 82573E Gigabit Ethernet Controller
测试工具	xdd6.5

用粒度大小为 64KB的条带化方式以最大化存储节点间读写操作的并发度,此外数据传输协议使用 NBD.

图 7给出了在不同读写粒度下虚拟磁盘的各种顺序读/写性能指标:

(1)双节点模式下虚拟磁盘的读/写吞吐率较单节点最大可提升 97.02%/71.53% (见图 7(a)(b)),相应的,前者平均读/写响应时间最小仅为后者的 51.52%/66.67% (见图 7(c));

(2)在单节点模式下,应用服务器的读/写网络带宽利用率最低仅为 35.40%/51.81%,而双节点在最低时也高达 65.01%/88.87%,近乎是前者的两倍 (见图 7(d)).

实验数据充分验证了 BW-VSDS系统所采用的带外虚拟化技术能够 (1)充分发挥单个存储节点的 I/O能力; (2)提高网络带宽利用率; (3)在多存储节点模式下可以有效提升应用服务器的 I/O性能.

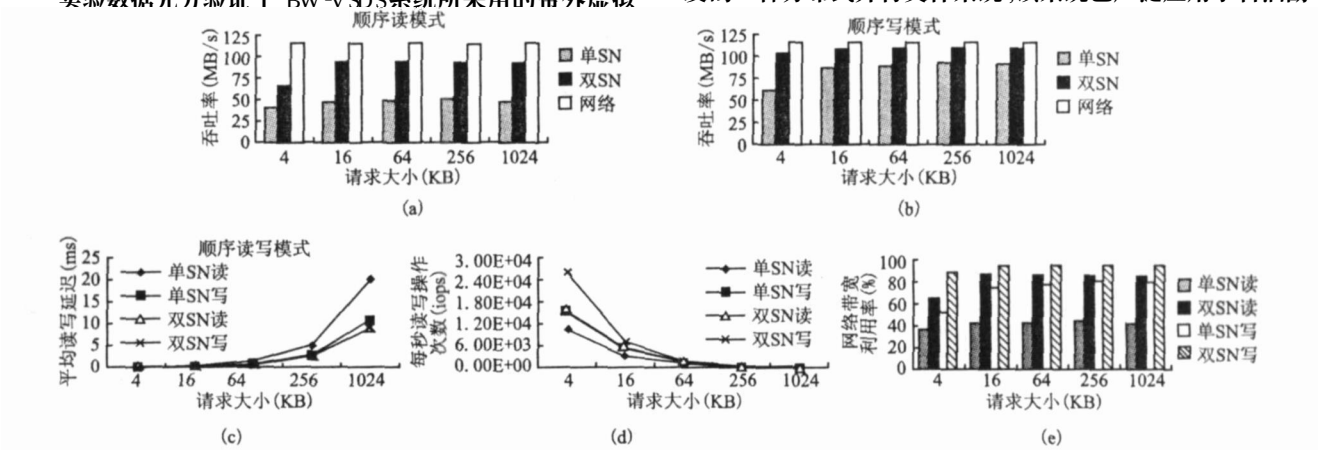

图 7　顺序读写模式下单/双 SN的性能对比

Fig 7　Comparison of sequential I/O performance of single/double SN

5　应用案例

目前 BW-VSDS系统能够提供 TB级的存储容量,已实际应用于奥运场馆的视频监控、研究院所的多媒体信息处理以及中小型企业的日常办公系统[7]等领域.此外最新的 BWFS系统[6]集成了 BW-VSDS系统,整合后的 BWFS只提供文件结构的管理和数据存储的组织而逻辑存储空间则由 BW-VSDS提供,BWFS也是国家高性能计算机工程技术中心自主研发的一种分布式并行文件系统,该系统已广泛应用于石油勘探、大规模科学计算、非线性编辑、企业办公和 WEB服务等多个领域.

6　总　结

本文详细介绍了一种基于带外虚拟化技术的 SAN系统,其特点在于采用两级带外虚拟化数据管理模型以充分发挥单个存储节点的 I/O能力并释放存储网络的承载能力;采用分布式数据存储管理协议以协同多个存储节点有效实现各种高级数据存储语义;支持多种网络数据传输协议以适用于不同应用环境.该系统目前已经应用于视频监控、信息处理和企业办公等领域.

References:

[1] Lofgren A. Decisions about storage area networks and network attached storage should not be based purely on capacity[EB/OL]. http://www.forrester.com/research, Sep. 2002.

[2] Bilas A. Block-level visualization: how far can we go[C]. Sardinia, Italy: Proc of 2th IEEECS International Symposium on Global Data Interoperability-Challenges and Technologies, 2005, 99-102.

[3] Shu J W, Li B G, Zheng W M. Design and implementation of a SAN system based on the fiber channel protocol[J]. IEEE Transactions on Computers, 2005, 54(4): 439-448.

[4] Zhang G Y, Shu J W, Xue W, et al. Design and implementation of an Out-of-Band virtualization system for large SANs[J]. IEEE Transactions on Computers, 2007, 56(12): 1654-1665.

[5] Wang Min, Li Jing, Fan Zhong-lei, et al. A service model for virtual resource management and its implementation[J]. Chinese Journal of Computers, 2005, 28(5): 856-863.

[6] Yang De-zhi, Huang Hua, Zhang Jiang-gang, et al. A distributed file system with large capacity, high throughput and high scalability[J]. Journal of Computer Research and Development, 2005, 42(6): 1028-1033.

[7] Liu Hai-yang, Meng Xiao-xuan, Ku Yi-nong, et al. Analysis and optimization of I/O performance for embedded NHD system[J]. Journal of Chinese Computer Systems, 2007, 28(7): 1334-1338.

附中文参考文献:

[5] 王　敏,李　静,范中磊,等. 一种虚拟化资源管理服务模型及其实现[J]. 计算机学报, 2005, 28(5): 856-863.

[6] 杨德志,黄　华,张建刚,等. 大容量、高性能、高扩展能力的蓝鲸分布式文件系统[J]. 计算机研究与发展, 2005, 42(6): 1028-1033.

[7] 刘海洋,孟晓烜,库依南,等. 嵌入式 NHD系统 I/O性能分析及优化[J]. 小型微型计算机系统, 2007, 28(7): 1334-1338.