存储虚拟化在石油物探的应用①

金 弟[1]，庄锡进[1]，杨 俊[2]

[1](中国石油杭州地质研究院 计算机应用研究所，杭州 310023)
[2](北京君杰时代科技发展有限公司，北京 100011)

摘 要：随着石油物探向高密度单点、高精度、高分辨率地震技术方向发展，地震资料处理和解释领域正面临着数据存储挑战。要满足存储资源的统一集中管理、异构平台的地震数据共享、提升存储系统的利用率、I/O 动态负载均衡等需求，提出了存储设备级与存储网络级结合的存储虚拟化思想，设计了一种实现方案，对地震资料处理与解释中的存储系统进行了逻辑卷级和文件系统级虚拟化，并对其优越性进行了分析与验证。

关键词：地震资料处理解释；存储虚拟化；逻辑卷级虚拟化；文件系统级虚拟化

Application of Storage Virtualization to Petroleum Exploration

JIN Di[1], ZHUANG Xi-Jin[1], YANG Jun[2]

[1](Department of Computer Application, Research Institute of Hangzhou Geology, Hangzhou 310023, China)
[2](Beijing Junjie Times Technology Co.. Ltd, Beijing 100011, China)

Abstract: With the petroleum exploration development toward high-density, high accuracy, high resolution seismic technology, seismic data processing and interpretation are facing data storage challenges. Storage virtualization is used to satisfy the demand of storage resource centralized management, sharing of seismic data between heterogeneous platforms, enhanceing the utilization of storage systems, I / O dynamic load balancing. This paper gives a design of storage virtualization by combination of storage device level and storage network level. Storage virtualization of logical level volume and file system level are applied in seismic data processing and interpretation storage system. The advantage is analyzed and verified.

Key words: seismic data processing Interpretaion; storage virtualization; logical level virtualization; file system level virtualization

1 引言

石油物探的深水海洋地震资料处理解释是大规模企业级高端存储系统重要应用领域之一，作者所在的海洋地震资料处理解释中心(简称处理解释中心)随着不断膨胀的地震数据和不断增加的存储设备，整个数据储存系统在系统管理、用户使用方面面临着诸多问题，本文针对这些存在的问题通过存储虚拟化技术进行了探讨、分析和解决。

2 存储虚拟化概念

存储虚拟化是存储领域的研究热点与核心技术，在基础的层面，存储虚拟化定义为在物理存储设备和低级逻辑存储设备之上，能够提供简化的逻辑存储资源视图的提取层[1]。在系统层面上看，存储虚拟化将通过多个存储阵列对提取层进行扩展，不但能够隐藏单个物理驱动器的复杂性，还能够隐藏整个物理存储子系统的复杂性[1]。也就是说存储虚拟化是存储的抽象，使用户能够在一个“更高的抽象层”来显示储存资源的视图，将所有的存储资源置于一个统一的、可用的大存储池中，为用户提供一个统一的逻辑视图。存储虚拟化技术就是在如何提高存储设备的管理效率、利用效率、优化性能，如何整合不同类型的存储资源

① 收稿时间:2011-05-13;收到修改稿时间:2011-06-12

便于共享，如何向用户提供统一的访问接口等前提下提出的。存储虚拟化在技术上分两种结构模式即对称结构和非对称结构[2]，实现方式分为三个层次：基于主机的虚拟化、基于存储设备的虚拟化和基于存储网络的虚拟化[3]。

3　存储虚拟化方案设计

3.1 存储系统现状

作者所在的处理解释中心原有存储系统拓扑结构如图 1 所示，由地震资料处理 SAN 存储系统（简称处理 SAN 存储系统）和地震资料解释 SAN 存储系统（简称解释 SAN 存储系统）组成,物理上互相独立,分别采用 IBM DS4800 和 SUN STK6540 存储设备构成，I/O 节点（也称存储主机）由 Linux 和 Solaris 异构操作系统（简称 OS）组成。存储资源分散管理，文件系统共享性差，两个 SAN 存储系统之间利用率和 I/O 负载等无法均衡，特别要新增存储系统时这些问题表现的尤其突出。本文利用新引进的 128 节点（1024 CPU 核）集群系统中的存储设备，运用逻辑卷级虚拟化和文件系统级虚拟化结合的虚拟化技术，来整合现有存储资源，构建深水海洋地震资料处理解释存储系统一体化的虚拟存储环境。

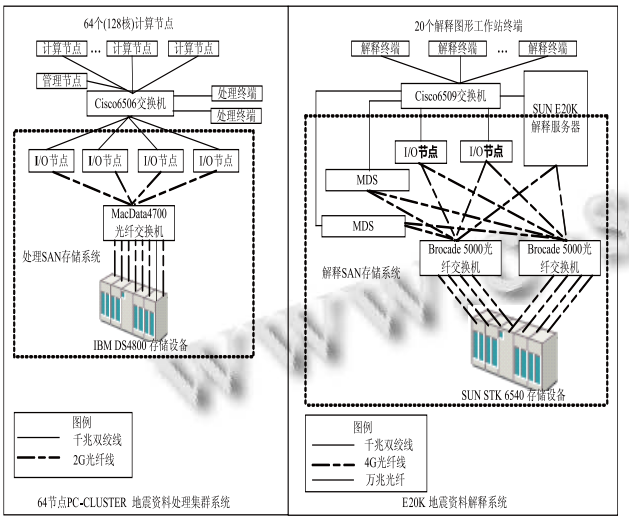

图 1　虚拟化前存储系统拓扑结构

3.2 逻辑卷级虚拟化

逻辑卷级虚拟化（简称卷虚拟化）是存储系统虚拟化的基础，主要目的是屏蔽存储设备的差异构建统一逻辑卷池。本文采用性能上最优的基于存储设备级虚拟化的方式。通过存储系统 HDS USP VM（简称本存储系统）[4]，将原有分散的处理 SAN 存储系统和解释 SAN 存储系统（简称外存储系统）的异构存储设备的逻辑卷映射成虚拟的存储视图，和本存储系统的逻辑卷一起构建统一逻辑卷池，并将该逻辑卷池的访问接口提供给所有 I/O 节点。所有 I/O 节点与分配给它的逻辑卷交互。该虚拟化方法的核心技术是卷映射，即外存储系统的异构存储设备的逻辑卷到统一逻辑卷池的映射。

卷映射的主要思想：使用分配外存储系统的每个逻辑卷（简称外卷）到本存储系统的每个虚拟卷的一个管理号进行关联，通过这个管理号标识在本存储系统建立的虚拟卷，实现在本存储系统层面上操纵外卷，进行了本存储系统的逻辑卷(简称内卷)和外卷的统一。管理号由外卷组号和顺序号构成，确保其唯一性。在分配一个管理号的同时，也相应的在外卷和虚拟卷之间建立一条或多条由本存储系统 External Port 端口和外存储系统 Local Port 端口物理相连接的映射路径，每条映射路径可以由一对或多对外卷和虚拟卷共享，用于连接本存储系统 Target Port 端口的所有 I/O 节点对这个由内卷和虚拟卷组成的统一逻辑卷池进行相同方式访问，屏蔽了外卷、虚拟卷及内卷的物理和逻辑的差异性，实现了逻辑卷级的虚拟化，如图 2 所示。表 1 为深水海洋地震资料处理解释存储系统进行逻辑卷级虚拟化的主要设计数据。

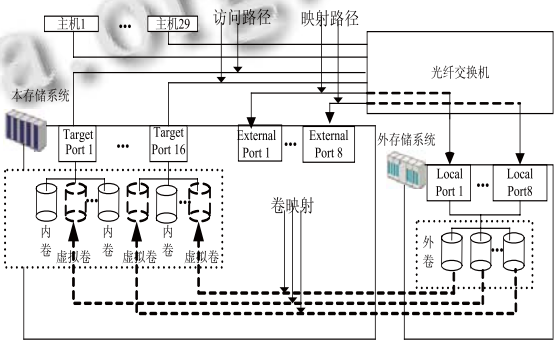

图 2　逻辑卷级虚拟化卷映射机制

3.3 文件系统级虚拟化

通过逻辑卷级虚拟化，实现向主机提供统一存储视图，但逻辑卷是 OS 管理访问的逻辑存储单元，用户和应用对存储资源直接访问形式是基于 OS 之上的文件系统作为存储单元。故要实现真正存储虚拟化，

只有实现了用户和应用级层面上文件系统虚拟化，才真正达到虚拟化目的，完全体现出存储虚拟化的优越性。

表 1 逻辑级虚拟化主要设计数据

存储设备	管理号	端口数	映射路径/访问路径
IBM DS4800（8T，12 个外卷）	01:01 至 01:12	Local Port 8 个	奇数管理号负载均衡共享 4 条映射路径，偶数管理号负载均衡共享 4 条映射路径，外存储系统无 I/O 节点访问路径。
SUN STK6540 (40T，58 个外卷）	02:01 至 01:58	Local Port 8 个	奇数管理号负载均衡共享 4 条映射路径，偶数管理号负载均衡共享 4 条映射路径，外存储系统无 I/O 节点访问路径。
HDS USP VM（100T，63 个内卷）	03:01 至 03:63	External Port 8 个，Target Port 16 个	本存储系统无映射路径，29 个 I/O 节点动态负载均衡共享 16 条 I/O 节点访问路径。

本文采用基于存储网络虚拟化的机制，使用非对称结构方式，采用昆腾 SNFS[5]实现统一逻辑卷池的管理，它能提供一个基于逻辑卷之上的的中间层逻辑块设备，形成统一文件系统池，从而实现文件系统级的虚拟化，最终实现两级虚拟化。两级虚拟化结构图如图 3。

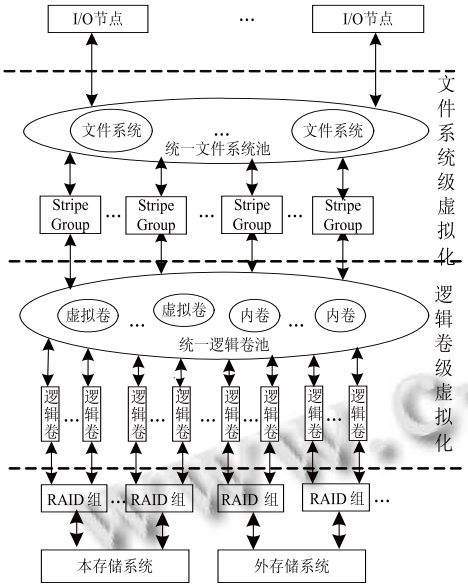

图 3 两级虚拟化结构

文件系统级虚拟化实现的思路：通过逻辑卷级虚拟化生成了由内卷和虚拟卷组成的统一逻辑卷池，根据用户对存储资源的需求，对内卷和虚拟卷再虚拟化成跨平台的统一文件系统池，用户和应用通过任何 I/O 节点使用这个统一文件系统池资源。在设计结构上，借鉴采用非对称方式的存储虚拟化结构的思想，传统 IP 网络作为存储控制通道，高速光纤网络作为存储数据通道。所有 I/O 节点的 I/O 请求首先通过控制通道传送到元数据服务器，获取逻辑卷池元数据控制信息后通过数据通道获取所需数据。存储数据在专用的数据通道上传输，减少了网络延迟，增加了带宽的可用性，从而提高了系统性能，同时还避免了系统的单点故障和瓶颈。系统结构如图 4，实线为数据 I/O 流，虚线为控制 I/O 流。通过元数据服务器负责协调访问、I/O 节点负载均衡，为 I/O 节点提供文件访问位置、数据块分配等信息，并确保多个 I/O 节点端对统一文件系统池并行 I/O 读写时保持数据的一致性和完整性。元数据服务器的体系结构由基于操作系统层次的底层存储逻辑卷管理和基于操作系统之上的应用层文件系统管理两部分组成，如图 5。

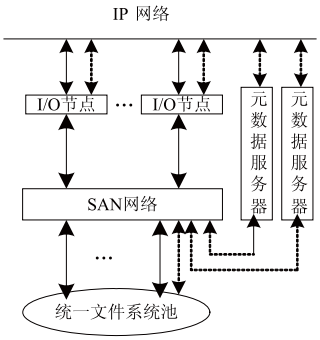

图 4 存储虚拟化非对称结构

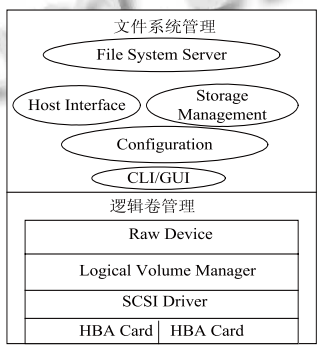

图 5 元数据服务器体系结构

3.4 虚拟存储系统分析

采用两级虚拟化技术构建的深水海洋地震资料处理解释存储系统虚拟存储环境结构如图 6，64 节点集群系统、128 节点集群系统及 E20K 解释系统三个系统的存储子系统虚拟化成一个线框所表示的存储系统，三个系统的任何 I/O 节点对这个虚拟化的存储系统的

资源访问和管理对用户和应用来说，就像一个物理存储系统。两级虚拟化后 I/O 请求对不同存储物理设备读写时实际的数据流向如图 7 和图 8。对虚拟化后的存储系统已经在国内外某些深水海洋区域的二维、三维地震资料处理解释项目上得到了广泛应用，从实际运行情况来看，在以下几个方面体现出了明显的优势。

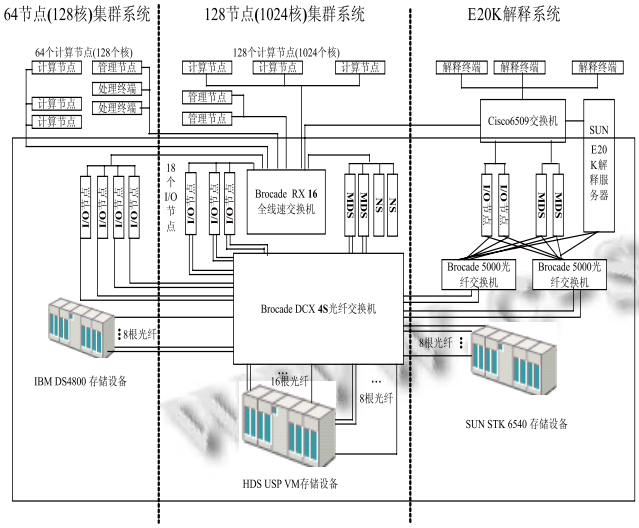

图 6　虚拟化后存储系统拓扑结构

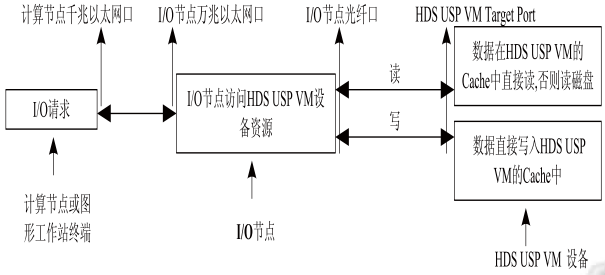

图 7　I/O 请求本存储设备数据读写流向图

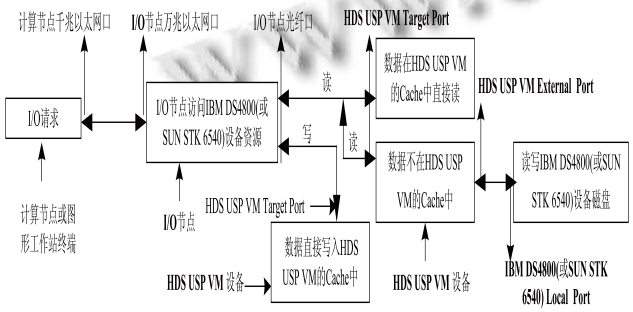

图 8　I/O 请求外存储设备数据读写流向图

(1) 集中统一管理。将原有独立 2 个 SAN 存储系统和新增的 SAN 存储系统建立统一的 SAN 网络，三个异构存储设备资源连接到这个 SAN 网络，由 HDS USP VM 控制器接管所有存储资源，形成了一个大的卷存储池，对处理和解释的存储资源实现集中统一全局管理。

(2) 数据共享与提高利用率。三个系统的所有 I/O 节点都可以访问这个虚拟化存储池资源。应用软件和用户通过这些 I/O 节点间接的按需动态使用，避免传统情况下出现处理（或解释）系统的物理存储设备资源紧缺，而解释（或处理）系统的物理存储设备资源利用率低下的情况。

(3) 性能优化与负载均衡。29 个 I/O 节点都能执行应用软件和用户的 I/O 请求，通过多路径软件 HDLM[6]，I/O 节点数量的增加能实现最大规模的多路径并行 I/O 读写和数据通道的冗余，尤其对深水海洋地震资料处理中的大规模、大容量的三维高密度单点、高精度、高分辨率的地震采集数据加载，性能提高特别明显。从表 2 实测数据看 (以 IBM DS4800 存储设备为例，使用 Linux dd 命令测试 10 次取平均值，测试文件大小为存储设备内存的 2 倍即 16GB)，虚拟化后使原有存储系统能充分利用 HDS USP VM 的高速 Cache，提高了对原有存储系统读写的 I/O 吞吐量，特别对于写操作有很大程度优化提升，而且虚拟化后写操作比读操作的性能还好。

表 2　虚拟化前后实测数据对比

类型	环境	文件大小	时间(s)	速度(MB/s)
读操作	虚拟化前	16GB	86	190
	虚拟化后	16GB	85	192
写操作	虚拟化前	16GB	91	181
	虚拟化后	16GB	63	262

(4) 可用性、可扩展性及可靠性。从应用软件和用户的角度，提供简单统一的存储资源，使用更方便。新增存储设备资源能方便在线的加入统一存储池后，就可以共享存储虚拟化的优点。发生 I/O 节点故障的 I/O 请求通过元数据服务器会自动的切换到其他 I/O 节点上，不影响应用软件和用户对存储资源的使用。

4　结语

存储虚拟化是一个支撑存储容量和存储服务的基础平台。本文通过逻辑卷级虚拟化和文件系统级虚拟

（下转第 76 页）

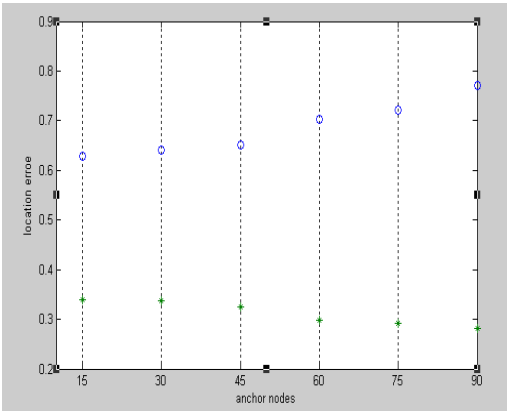

图 3　节点数量与平均定位误差的关系

如上图所示在锚节点数从 15,30，…,90 的变化过程中，可以看到这两种算法的定位误差与锚节点数的关系图。蓝色圆圈代表的是 DV-HOP 定位算法的误差图，绿色星形代表的是改进后的 DV-HOP 定位误差图。从图中，不然可以发现，改进后的算法的定位误差明显低于 DV-Hop 算法,可达到 28%,并且会随着锚节点的增加误差值会逐渐缩小。

5　结论

通过对 DV-Hop 算法分析,总结 DV-Hop 算法的特点,在此基础上提出了改进算法。改进算法是基于一种几何斜率的定位方法，来选取合适的锚节点，从而提高定位精度。经过仿真实验表明在锚节点数量一定的前提下,其定位误差明显好于 DV-Hop 算法,增强了系统的鲁棒性。因此,改进后的算法是有效和可行的。

参考文献

1 任丰原,黄海宁,林闯.无线传感器网络.软件学报,2003,14(7):1282−1291.
2 邱岩.无线传感器网络节点定位技术研究.计算机科学,2008,35(5):47−50.
3 段渭军,王建刚,王福豹.无线传感器网络节点定位系统与算法的研究和发展.信息与控制,2006,35(2):239−245.
4 王福豹,史龙,任丰原.无线传感器网络中的自身定位系统和算法.软件学报,2005,16(5):857−868.
5 He T, Huang CD, Blum BM, Stankovic JA, Abdelzaher T. Range-free localization schemes in large scale sensor networks. Proc. of the 9th Annual Int'l Conf. on Mobile computing and Networking. San Diego: ACM Press, 2003.81− 95.
6 B WHL, H CJ. Global positioning system. 1997.
7 Bahl P, Padmanabhan VN. RADAR:Aninbuilding RF-based user location and tracking system. Proc. of IEEE Infocom 2000, Tel-Aviv, Israel. 2000.2:775−784.
8 李善亮,黄刘生,吴俊敏.基于连通性的传感器节点定位算法研究.计算机工程,2008,34(7):115−117.
9 孙利民,等.无线传感器网络.北京:清华大学出版社,2005.
10 杨旸.传感器网络节点定位技术研究.杭州:浙江大学,2006.

（上接第 16 页）

化对虚拟化存储系统进行了扩展和丰富，并在石油物探的深水海洋地震资料处理解释系统的存储子系统中进行了应用，取得了较好的效果。

参考文献

1 杨光年,郭荣亮,张国政.存储技术的发展及整合与虚拟化应用.计算机与数字工程,2008,36(3):140−144.
2 谢长生,金伟.SAN 网络存储虚拟化实现方式研究与设计.计算机应用研究,2004,4:191−193.
3 谭生龙.存储虚拟化技术的研究.微计算机应用,2010,21(1):33−38.
4 Hitachi Universal Volume Manager User's Guide.Hitachi Data Systems.2010.
5 StorNext File System.ADIC Educational Services.2006.
6 Hitachi Dynamic Link Manager Software User's Guide for Linux. Hitachi Data Systems.2010.