存储框架模型在地震资料大数据中的应用①

金 弟, 庄锡进, 王启迪, 曹晓初, 王宗仁

(中国石油杭州地质研究院 计算机应用研究所, 杭州 310023)

摘 要: 油气勘探领域的地震资料是一种典型大数据. 运用超大规模并行处理应用软件对地震资料数据进行叠前偏移成像处理, I/O 存储子系统成为整个系统的主要性能瓶颈. 针对该问题, 本文以存储主机集群、动态存储多路径与并行文件系统为关键技术进行有机结合, 提出了一种存储框架模型, 分析优越性, 并对其进行部署. 在典型应用场景下对系统的功能、性能进行测试, 结果验证了系统的优点.

关键词: 地震资料大数据; 存储主机集群; 并行文件系统; 存储多路径

Application of Storage Framework Model in Seismic Big Data

JIN Di, ZHUANG Xi-Jin, WANG Qi-Di, CAO Xiao-Chu, WANGZ Zong-Ren

(Department of Computer Application, Petrochina Hangzhou Research Institute of Geology, Hangzhou 310023, China)

Abstract: Seismic data in the field of oil and gas exploration is a typical kind of big data. The I/O storage system has become a major performance bottleneck of the whole system when pre-stack migration imaging is processed for seismic data using large scale parallel processing application software. Adopting the key technologies of the storage host cluster, dynamic storage multipath and parallel file system, this paper puts forward a storage framework model, analyzes the superiority, and does its deployment. The function and performance of this system is tested in typical application scenarios and the results verify the advantages of the system.

Key words: seismic data; storage host cluster; parallel file system; storage multipath

1 引言

在油气勘探过程中, 利用人工地震波激发采集的地震资料是一种海量数据[1]. 国内外对地震资料大数据分析处理主要集中在叠前深度偏移、逆时偏移、介质偏移、反演偏移等偏移成像技术[2,3]与排序、整理[4], 对其存储方面的管理模型、数据访问基础架构[5]、地震数据管理系统建设[6]国内有些研究, 而在存储框架模型角度相对比较少.

利用高性能计算集群系统与超大规模并行处理应用软件对地震资料数据进行偏移成像处理分析过程中, 负责地震资料大数据读写操作存储子系统成为性能瓶颈的核心环节, 存在以下问题:

①多客户端大规模 I/O 并行请求形成高聚合带宽, 使存储主机节点端出现拥堵与单点故障.

②多客户端并行读写同一个或多个叠前地震大数据文件, 在 Linux 操作系统环境下的文件系统层出现大量文件读写的 I/O 排队等待.

③存储主机的大规模 I/O 并行服务, 形成的高吞吐量, 在存储主机与存储控制器之间的数据链路通道出现拥堵与单点故障.

有鉴于此, 本文采用基于分布式与并行化的 I/O 读写思路对存储框架模型进行设计, 提升地震资料大数据 I/O 读写过程中各核心环节上的可靠性与性能. 为针对基于地震资料大数据环境下负责读写操作的存储子系统构建提供了一种途径.

2 存储框架模型设计

2.1 存储主机集群

集群(Cluster)就是一组计算机, 利用并行计算处理[7]的原理, 它们作为一个整体向用户提供一组资源,

① 收稿时间:2015-05-08;收到修改稿时间:2015-06-15

具有低成本、高效率[7]. 本文把集群技术应用到存储主机, 如图 1 所示, 客户端的大规模读写请求形成的高聚合访问带宽通过万兆以太网络或 IB (InfiniBand)等高速网络(简称前端网络)分布式加载至 n 个存储主机搭建的集群执行并行 I/O 响应. 实现并行 I/O 与单点故障切换,提升 I/O 执行性能、增强可靠性与可扩展性.

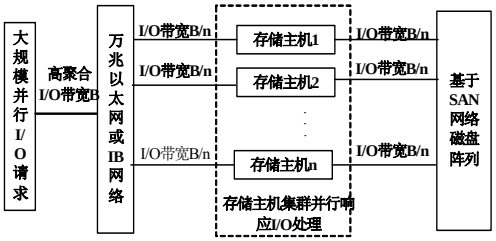

图 1　存储主机集群示意图

2.2 动态存储多路径

存储多路径是指在存储区域网络体系架构下, 多个 I/O 节点(也称存储主机)并发访问存储资源, 每个存储主机到存储设备的 LUN(Logical Unit Number)之间有多条 I/O 数据流, 即每个存储主机读写每个 LUN 有多条数据通路[8]. 如图 2, 通过 SAN 网络(简称后端网络)第 i 个存储主机至具有 h_m 个主机端口的存储控制器有 P_{im} 条路径可路由. 使用存储多路径负载均衡算法[8]在每个存储主机可用的存储路径之间动态智能的分配 LUN 读写流量, 分布式 LUN 数据访问, 提升 LUN 级的读写性能. 实现提高在存储主机与存储控制器之间的数据链路流量带宽, 避免出现拥堵与单点故障.

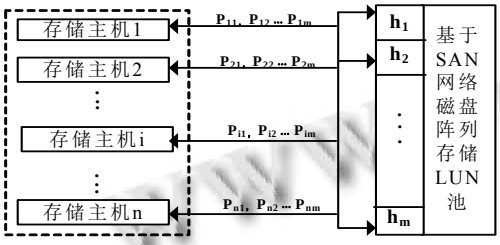

图 2　动态存储多路径示意图

2.3 并行文件系统

与常规网络文件系统(如 NFS, CIFS)相比, 并行文件系统在文件读写密集型大数据应用中具有高吞吐量、高文件读写带宽、可扩展的特点[9]. 为了有效利用存储主机群相当可观的 I/O 带宽资源, 采用并行文件系统方法对应用程序的大规模文件级读写请求服务使用读写代理负载均衡算法[9]动态加载到存储主机集群中的多个 I/O 代理执行节点上去, 实现多个文件并行读写操作与单个文件并发读写, 充分发挥存储主机集群的优势, 大幅度减少文件读写的I/O排队等待. 同时提供更大的文件存储容量和聚集的 I/O 带宽, 并随文件读写规模扩大而动态扩展文件系统容量. 并行文件系统拓扑结构如图 3, n 个存储主机并行执行文件级数据块读写操作, 元数据服务器负责 LUN 级分布式组织至文件系统级映射、文件属性等元数据信息管理.

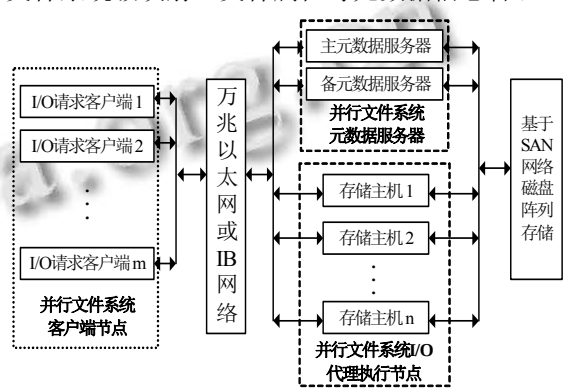

图 3　并行文件系统拓扑

2.4 模型构建与分析

存储框架模型构建采用集群技术应用于存储主机端、采用并行文件系统方法在文件系统层上实现并行读写以及存储后端的 LUN 层数据通道采用存储多路径. 设计应用于地震资料大数据存储读写的存储框架模型如图 4 所示, 主要实现文件级分布式并行 I/O 与 LUN 级分布式并行 I/O.

2.4.1 文件级分布式并行 I/O

① 文件级 I/O 服务执行者的智能选择. i 个分布式 LAN 客户端(Distributed LAN Client,DLC)作为文件级 I/O 请求的发起者. 通过元数据专用网络从元数据服务器(MetaData Server,MDS)获取文件访问位置、数据块分配等与文件系统相关元数据信息. 利用数据专有网络把并行大数据地震文件读写请求, 动态负载均衡分配给 n 个分布式 LAN 服务端(Distributed LAN Servert,DLS), 并行处理文件级 I/O.

② 文件级的 DLC 并行请求映射至 LUN 级. 利用并行文件系统客户端(Parallel File System Client,PFSC)通过数据专用网与元数据专用网获取元数据服务器的元信息, 把并行文件级的读写转化到多个存储主机并行对多个 LUN 读写.

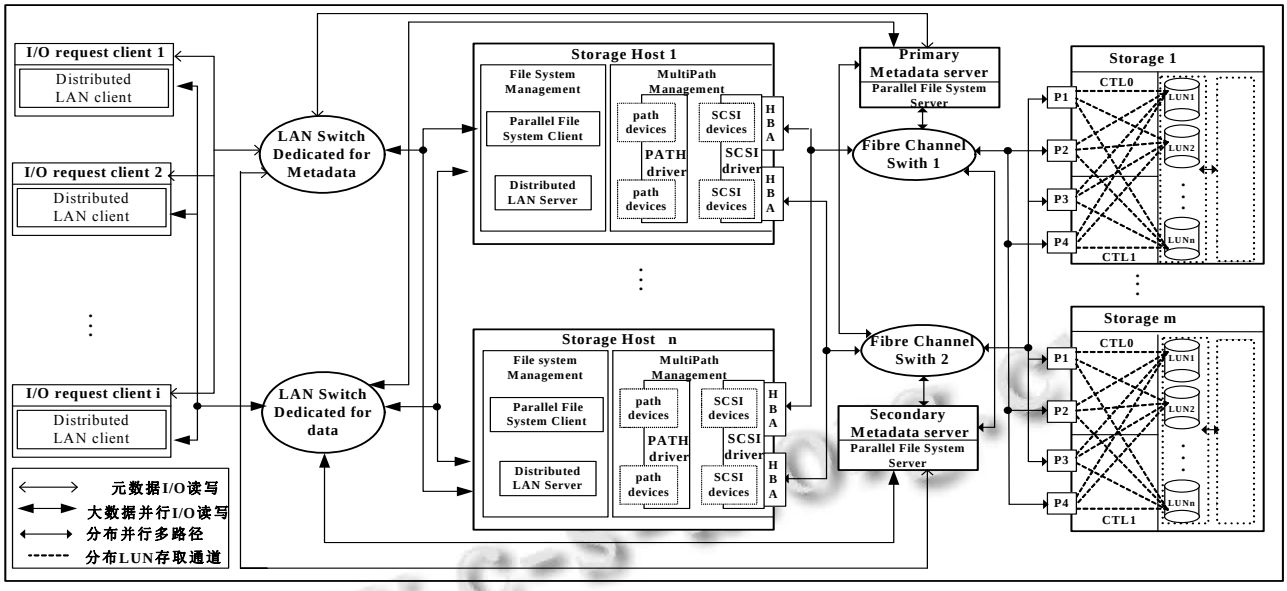

图 4 地震资料大数据环境下的存储框架模型

2.4.2 LUN 级分布式并行 I/O

① 分布式 LUN 集的生成. 对挂载不同磁盘通道的阵列中抽取磁盘驱动器构成RAID组, 再对RAID组划分成 LUN, 最后把 LUN 分布式映射到存储控制器与存储主机群的不同端口, 实现存储主机对一个逻辑层 LUN 读写横跨至多个磁盘驱动器在不同磁盘通道的物理层操作.

② SCSI驱动层的LUN至路径驱动层LUN映射. 在多路径环境下, SCSI 驱动(SCSI driver)识别的一个 SCSI 设备(SCSI devic)以 LUN 的形式呈现给存储主机时存在重复多个 LUN, 即一条路径对应一个 LUN. 路径驱动(PATH driver)对重复的 LUN 进行合并, 并关联多条可分配的路径. 存储主机在执行多个 LUN 读写时, 通过数据通道负载均衡算法智能分配所有 LUN 的 I/O 请求分散到可用的路径资源, 优化多个LUN 在多条路径同时读写操作. 实现 LUN 级 I/O 动态分配存储路径.

与联机事务处理应用关注存储系统 IOPS 性能指标不同, 地震资料大数据处理应用更关注的是大文件读写、并行大 I/O 块的吞吐量性能指标. 图 4 中的文件级分布式并行 I/O 与 LUN 级分布式并行 I/O 机制从文件读写请求级至数据通道的 LUN 级等不同层面有效提升吞吐量.

3 模型部署与实现

针对高效的部署与实现存储框架模型, 提出数据存储组织方式与数据访问机制结合的方法. 在数据存储组织方式上, 从物理层磁盘驱动器至逻辑层 LUN 的数据元素布局, 结合并行文件系统的优势与参数配置, 利用分布与并行 I/O 的先进机制优化数据存储组织. 在数据访问机制方面, 利用并行文件系统与多路径软件融合在存储主机集群的优化部署, DLC 与 DLS 间的多数据通道并行传输使用高速低延迟的 IB 网络, 利用冗余 SAN 网络, 在存储主机集群与存储控制器的多通道传输.

3.1 部署流程

基于上述存储框架模型的存储子系统部署流程如图 5 所示, 分为以下三个步骤.

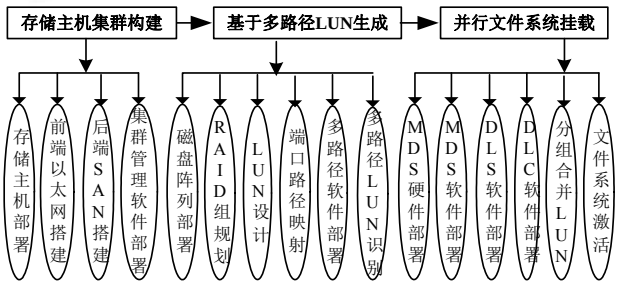

图 5 存储系统部署流程图

① 存储主机集群构建. 存储主机集群节点部署采用 12 台 IBM System3650, 操作系统为 Red Hat Enterprise Linux Server release 5.5. 1 台 Mellanox Grid Director4200 网络设备进行前端 IB 网络搭建. 后端

SAN 网络搭建采用 2 台 Brocade 5100 8GB FC 交换机. 存储主机集群管理软件为 xCAT2.3.

② 基于多路径 LUN 生成. 2 台 HDS AMS2500[10] 为存储控制器、基于 SAS 硬盘搭建磁盘阵列, 作为后端海量地震数据分布式 LUN 集. 地震资料数据盘采用采用 RAID5 与元数据盘采用 RAID10 进行 RAID 组规划设计. 每个 RAID5 划分 2 个 LUN. LUN 端口路径映射依据每个 LUN 对应 4 个物理路径. 采用 HDLM 6.5[11]部署于存储主机集群中的每个节点, 实现动态多路径软件部署与多路径 LUN 识别, 完成基于多路径的动态 LUN 级访问的数据存储组织方式与存储访问机制的构建.

③ 并行文件系统挂载. 以 Quantum SNFS4.1 [12,13] 为基于元数据服务器架构的并行文件系统. SNFS Server 作为元数据服务器 MDS, 硬件部署采用 2 台 IBM System3650 进行双机 HA. SNFS I/O Client 部署在存储主机群作为 PFSC 与 DLS. SNFS Client 部署在 I/O 并行请求发起端作为 DLC.

3.2 实施关键技术

3.2.1 存储主机集群安装配置

① 存储主机 IB 网络包驱动安装. tar xzvf MLN X_OFED_LINUX-1.5.2-2 .1.0-RH EL5.5.tar.gz; mlnxofed install –all.

② 修改 IB 网络驱动模块参数. 编辑文件/etc/modprobe.conf 内容如下: options mlx4_vnic net_admin=0.

③ 修改 IB 网卡配置文件与加载启动网卡. 编辑文件 /etc/sysconfg/network- scripts/ifcf g-eth1 内容如下: VNICIBPORT=mlx4_0:1; 编辑文件 /etc/infiniband /openib.conf 内容如下: M LX4_VNIC=yes; 重新启动 IB 网络服务: /etc/init.d/mlx4_vnic_confd restart.

④ xCAT 安装. 设置 xCat 相关环境变量: source /etc/profile.d/xcat.sh; 用 yum 安装: yum install xCAT.

⑤ xCAT 相关表与参数. 配置 network 表, 设置 NFS、DNS、DHCP、TFTP、HTTP 服务;编辑文件/ /etc/hosts 加入所有存储主机节点, /etc/resolv.conf 指向管理服务器.

3.2.2 并行文件系统 StorNext 安装配置

①软件包安装. MDS 端安装: ./snfs_full_RedHat50 AS_26x86_64.bin ; install.stornext; DLS 端安装: rpm -ivh snfs-redhat-x86_64.rpm; DLC 端安装: rpm –ivh snfs-client-4.1.1_x86_64.rpm.

②MDS 配置. SNFS 生成的每个并行文件系统对应一个配置文件, 由对文件系统功能、性能管理的全局参数与仅对条带组(StripeGroup,SG)管控的条带参数二部分构成. 通过这个核心配置文件把基于多路径生成的 LUN 集依据设置参数策略构建一个并行文件系统, 其关键参数设置与描述如表 1.

表 1　MDS 端配置参数

	名称	描述
全局	FsBlockSize = 64K	块分配粒度
	BufferCacheSize=256M	元数据信息缓存
	AllocationStrategy= balance	选择 Stripe Group 分配算法
条带	Type=Data	条带组类型
	StripeBreadth=64K	条带宽度, 需匹配
	mulipathMethod=rotate	选择 LUN 分配算法

③DLS 端配置. 编辑文件/usr/cvfs/config/fsnameservers 加入 MDS 主机 IP; 执行命令/usr/cvfs/bin/sndpscfg –e 配置 DLS 与 DLC 数据传输主要参数如表 2. 编辑文件/etc/fstab 在 DLS 端配置加载并行文件系统, 一行对应一个并行文件系统, 字段格式如表 3.

表 2　DLS 端配置参数

名称	描述
interface=eth1	与 DLC 传输数据网络接口
transfer_buffer_size_kb=256	socket 传输的缓冲区大小
tcp_window_size=64	与 DLC 连接时 TCP window 大小
Daemon_threads=8	DLS 启用的核心线程数

表 3　DLS 端并行文件系统加载字段格式

	名称	描述
	并行文件系统名称	与 MDS 的核心配置文件名称一致
	挂载点	空闲 Linux 目录
	类型	并行文件系统类型 cvfs
挂载参数	diskproxy=proxyserver	指定挂载的文件系统的主机为 DLS
	auto_concwrite=yes	允许多线程并行写文件
	cachebufsize=128K	每个 Cache buffer 大小
	buffercache_iods=16	执行 cache buffer I/O 后台进程数量
	mnt_retry=8	挂载文件系统失败重试次数

④DLC 端配置. 编辑文件/usr/cvfs/config/fsnameservers 加入 MDS 主机 IP; 编辑文件/etc/fstab, 加载 DLC 端文件系统, 一行对应一个并行文件系统, 字段格式如表 4.

表 4 DLC 端并行文件系统加载字段格式

名称		描述
并行文件系统名称		与 MDS 的核心配置文件名称一致
挂载点		空闲 Linux 目录
类型		并行文件系统类型 cvfs
挂载参数	diskproxy=proxyc lient	指定挂载的文件系统的主机为 DLC
	proxypath= balance	DLC 向多个 DLS 并行 I/O 请求时, 负载均衡 I/O 采用的算法.
	proxyclient_wto=15	DLC 向 DLS 写请求等待时间超时
	proxyclient_rto=15	DLC 向 DLS 读请求等待时间超时

3.2.3 动态多路径 HDLM 安装配置

① 设置环境变量. set path=($PATH : /opt/Dynamic LinkManager/bin).

② 安装 HDLM driver. 确认 HDLM 安装版本, ./installhdlm –v; 执行 HDLM 安装脚本, ./ installhdlm 安装; 确认安装成功, rpm –qi hdlm; 启动 HDLM 服务, ./dlmstart.

③ 设置动态多路径负载均衡算法. 在所有可用的多个数据通道, 分配多路径 I/O 采用的算法. 设置扩展 least I/O 负载均衡算法: dlnkmgr set –lb on –lbtype exlio.

④ 设置路径健康检查. 检查在线路径的状态, 监测到路径错误, 把该路径置于 offline 状态, 进行数据流的重新路由. 设置检查周期 10 秒: dlnkmgr set –pchk on –intvl 10.

⑤ 设置故障自动恢复检查. 检查故障路径的修复状态, 自动恢复到 online 状态. 设置检查周期 10 秒: dlnkmgr set –afb on –intvl 10.

⑥ 设置错误监控制. 设置监控周期 20 秒, 错误级别为 2: dlnkmgr set –iem on –intvl 20 –iemnum 2.

⑦ 确认检查当前配置. dlnkmgr view –sys –sfunc.

4 系统测试与分析

通过上述方法, 对所部署存储子系统在典型地震资料大数据偏移成像分析处理的应用环境下进行功能与性能测试. 功能测试通过故障注入法进行. 性能测试采用地震资料并行处理应用软件对采集的地震数据体进行偏移成像时, 监控采集 I/O 动态实时变化数据.

4.1 功能测试

主要目的是测试系统的可靠性、可用性、冗错性. 操作系统全部采用 RedHat Linux 5.5, 在地震资料数据进行逆时偏移应用环境下, 主要测试结果如表 5.

表 5 功能测试

测试方案	测试结果
模拟存储主机群节点端软硬件故障, 对任一存储主机正常关机	数据流切换到其它正常存储主机, 存储访问连续进行. 解决 DLC 请求时存储主机端单点故障
模拟数据链路通道故障, 对任一存储主机端口拔掉光纤线	数据流路径切换到其它正常多路径中去, 存储访问连续进行, 解决数据链路通道单点故障
模拟数据链路通道故障, 对任一存储控制器端的主机端口拔掉光纤线	数据流路径切换到其它正常多路径中去, 存储访问连续进行, 解决数据链路通道单点故障
模拟 MDC 服务器软硬件故障, 对主 MDC 服务器正常关机	并行文件系统服务从主 MDC 切换到从 MDC,存储访问连续进行, 解决 MDC 单点故障

在测试完成后, 对逆时偏移后的地震资料数据进行了一致性检测. 结果表明故障发生前后未发生数据不一致现象, 能正确完成地震数据体 I/O, 同时故障前后对用户使用不受影响.

4.2 性能测试

对国内某盆地的三维地震资料数据, 主要采集参数信息如表 6, 使用 Omega2012[14]大型地震资料并行处理软件选择部分地震数据体 2TB 进行叠前时间偏移成像时进行测试, 并行度为 64 个客户端 I/O 请求, I/O 读写地震数据体横跨三个并行文件系统的 12 个 LUN.

表 6 某盆地三维地震资料采集参数信息

覆盖面积	280km2	采集面元	25*25m
覆盖次数	120 次	接收道数	1760 道
接收线距	300m	接收道距	50m

对地震资料大数据进行叠前时间偏移分析处理主要分为: 存储子系统把地震数据分发加载到客户端(DLC)本地 scratch 盘队列、客户端对队列地震数据体进行偏移成像分析处理、客户端分析处理的地震数据结果返回给存储子系统三个循环阶段. 其中第一与第三阶段是地震大数据密集 I/O, 第二阶段是地震大数据密集计算. 存储子系统性能测试关注第一与第三阶段.

首先在地震资料大数据偏移成像分析处理第一阶段期间, 对 DLC 端并行请求时, 负责并行 I/O 响应处理的存储主机群前端网络端口的性能进行测试, 用以分析存储主机集群 I/O 动态负载均衡与并行处理文件

级读写的性能效率. 由图6的测试结果可以看出, 12个存储主机的I/O网络带宽都在2Gb/s至4Gb/s, 形成的聚合带宽平均3.5GB/s. 实现了DLS对DLC读写请求文件级的分布式并行 I/O. 消除存储主机端 I/O 堵塞, 减少文件读写I/O排队等待.

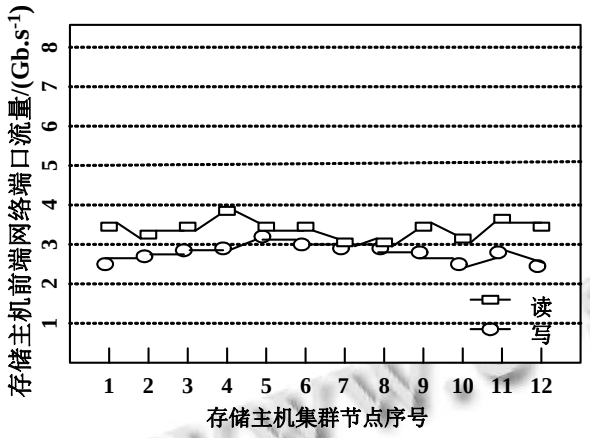

图 6　存储主机集群前端网络端口带宽

接着对地震资料大数据偏移成像分析处理第三阶段期间的某一时刻, 随机选取存储主机集群中的 5 号存储主机, 随机监控一个LUN对应的4条多路径(多路径序号为 1 至 4)的数据读写流量. 测试结果如图 7 所示, 我们可以看出, 每条路径的读最高速度和最低速度分别为 71MB/s 与 62MB/s. 每条路径的写最高速度和最低速度分别为 61MB/s 与 52MB/s. 测试数据表明, 对任何一个 LUN 的 I/O 请求, 都能负载均衡至多条数据路径, 实现 LUN 级的分布式并行 I/O. 提升存储主机与存储控制器间数据链路通道总带宽.

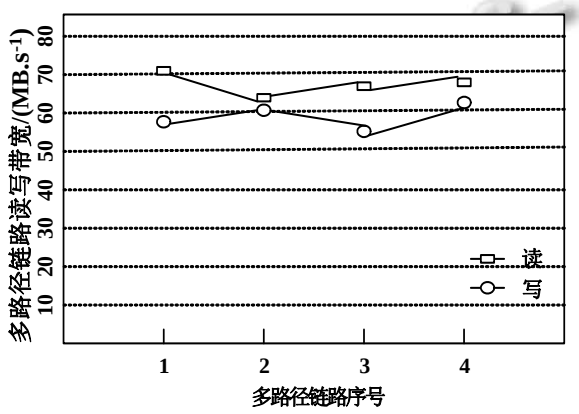

图 7　存储多路径 I/O 带宽

最后针对地震资料大数据偏移成像分析处理第一阶段的数据分发高峰期, 选取三个文件系统涉及的 12 个 LUN 的读写速度进行测试. 由图 8 可以看出, 每个 LUN 的读和写分别在 220MB/s 与 200MB/s 上下波动. 这说明面对客户端的大规模 I/O 请求访问, 通过文件级的分布式并行读写, 实现了最后映射到对多LUN级的分布式并行读写. 测试结果表明存储框架模型的设计是非常有效.

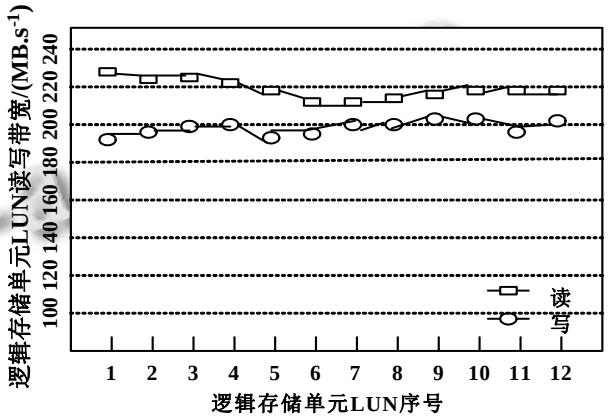

图 8　多个 LUN 读写速度

5　结语

本文结合存储主机集群、动态存储多路径、并行文件系统技术, 提出了一种存储框架模型. 在油气勘探领域的地震资料大数据环境下进行部署, 目前在国内某些盆地矿权区块地震资料处理项目上得到广泛应用, 取得了良好效果. 基于该存储框架模型, 在存储子系统的运行过程中, 根据地震资料处理业务需求, 如何改进设计优化相关配置参数, 进一步提升系统性能, 是笔者下一步的研究方向.

参考文献

1 赵满.地震数据并行访问策略的研究[硕士学位论文].大庆:东北石油大学,2013.
2 李振春.地震偏移成像技术研究现状与发展趋势.石油地球物理勘探,2014,49(1):1–22.
3 刘守伟,王华忠,陈生昌.三维逆时偏移 GPU/CPU 机群实现方案研究.地球物理学报, ,2013,56(10) :3487–3496.
4 程学旗,靳小龙,王元卓,郭嘉丰,张铁赢,李国杰.大数据系统和分析技术综述.软件学报,2014,25(9):1889–1908.
5 涂新莉,刘波,林伟伟.大数据研究综述.计算机应用研究,2014,31(6):1612–1617.
6 高辉,李建军,曹瑜,李涛,杨俊丰.塔里木油田 Expeditor 地震

数据管理系统建设.石油地球物理勘探,2008,43(增刊1):182–185.
7 李敏,张宜生,李德群.用于并行计算的 PC 集群系统构建.计算机应用研究,2009,26(3):1042–1044.
8 金弟,庄锡进,曹晓初,王启迪,王宗仁.基于多路径地震资料处理集群存储系统.计算机研究与发展,2012,49(增刊):42–46.
9 霍严梅,杨可新,胡亮,鞠九滨.并行文件系统研究综述.小型微型计算机系统,2008,29(9):1631–1636.
10 HDS. Hitachi Storage Navigator Modular User's Guide. Santa Clara USA: Hitachi Data Systems Press, 2010: 23–52.
11 HDS.Hitachi Dynamic Link Manager Software User's Guide for Linux. Santa Clara USA: Hitachi Data Systems Press, 2010: 17–64.
12 Quantum. Stor Next File System Installation Guide. Seattle USA: Quantum Press, 2010: 27–48.
13 Quantum. Stor Next File System User's Guide. Seattle USA: Quantum Press, 2010: 41–150 .
14 Schlumberger. Omega 2012.1 Install Guide. Houston USA: Schlumberger, 2012: 17–24 .