基于 SNFS 地震资料处理并行共享存储系统①

金 弟, 庄锡进, 王启迪, 王宗仁, 曹晓初

(中国石油杭州地质研究院 计算机应用研究所, 杭州 310023)

摘 要: 针对传统并行共享存储系统在深水海洋地震资料处理应用中存在的缺点, 设计部署了一种新的基于 SNFS 并行共享存储系统. 通过对其进行功能与性能测试分析, 结果表明该系统具有优越性, 能满足实际石油物探应用需求. 在此基础上, 进一步提出了对该存储系统优化的建议与方法.

关键词: 并行共享存储系统; 地震资料处理; SNFS; 聚合吞吐量

Seismic Data Processing Parallel Shared Storage System Based on SNFS

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

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

Abstract: In view of the traditional parallel storage system in deep water offshore seismic data processing applications shortcomings, this paper designs and deploys a new parallel shared storage system based on SNFS. On the analysis of function and performance test, the results show that the system has the superiority, and it can meet the actual application demand of petroleum geophysical exploration. Finally, it puts forward the optimization suggestions and methods of the storage system.

Key words: parallel shared storage system; seismic data processing; snfs; aggregate throughput

油气勘探领域的深水海洋地震资料处理是高端企业级大规模并行共享存储系统应用的重要领域之一. 本文是在作者所在单位建设地震资料处理平台的存储子系统下开展研究, 其主要工作是根据地震资料处理对存储系统的大数据大规模并行聚合读写工程应用需求,提出一种基于 SNFS(StorNext File System)并行共享存储系统解决方案.

1 SNFS文件系统

SNFS[1,2]是由昆腾公司研发的高性能共享并行文件系统, 运行模式如图 1. SNFS 对 LUN 进一步数据布局构建, 对不同应用需求提供众多优化参数, 其存储数据组织如图 2. 它主要构成如下: (1) DLC(Distribute LAN Client). 基于 CVFS 协议接口, 其数据流量负载均衡分布到多个 DLS(Distribute LAN Server), 把数据高速存储扩展到基于 LAN 的应用; (2) DLS. 处理由 DLC 通过 LAN 网络对存储资源并行 I/O; (3) SNFS Client. 担任存储主机角色, 通过 SAN 网络发起对存储设备并行读写, 一般与 DLS 属于同一个物理主机; (4) MDC(Meta Data Controller) 也称 SNFS Server. 处理元数据请求的服务器, 管理访问共享卷和元数据. 确保并行读写共享数据一致性和最新元数据统一视图; (5) FJR(Fast Journal Recover). 对元数据和存储卷的修改, 在文件系统异常时, 确保数据完整性和一致性, 一般 MDC 和 FJR 属同一个物理主机.

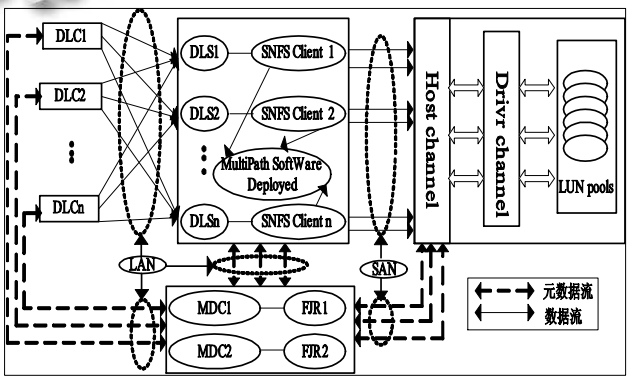

图 1 SNFS 运行机制模式

① 收稿时间:2012-06-18;收到修改稿时间:2012-08-13

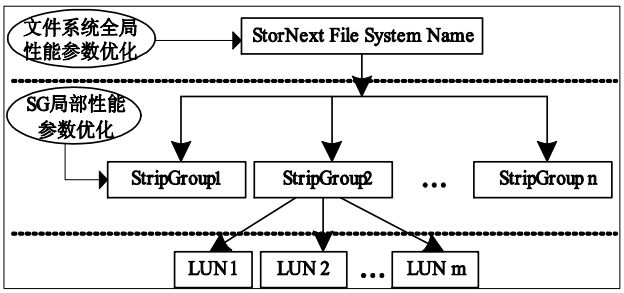

图 2　SNFS 存储数据组织

2　并行共享存储系统设计

2.1 总体框架

根据地震资料处理对存储需求, 通过分析研究部署的一种基于 SNFS 地震资料处理并行共享存储系统结构如图 3, 由并行应用客户端、存储前端网络、存储主机、存储后端网络、多存储控制器及磁盘阵列 6 部分组成. 其特点: 1) 具有 256 个计算节点, 3072 个 CPU 核的大规模 DLC 并行访问客户端. 2) InfiniBand QDR 40Gb/s 高速网, DLC 与 DLS 间的多数据通道并行传输. 3) I/O 请求动态负载均衡到多个存储主机, 具备流量控制、故障自动切换、多路径管理及宽带聚合. 4) 冗余 SAN 网络, 存储主机与存储控制器的多通道传输. 5) 主机通道与驱动器通道接口构成多存储控制器. 其优点: 1) 充分发挥多网络、多存储主机、多 I/O 通道、多存储控制器、盘阵、并行文件系统等各个关键存储部件融合优势. 2) 满足对海量海洋三维地震数据进行大文件大规模并行共享读写、具有高聚合吞吐量、高可靠可用性及易扩展性.

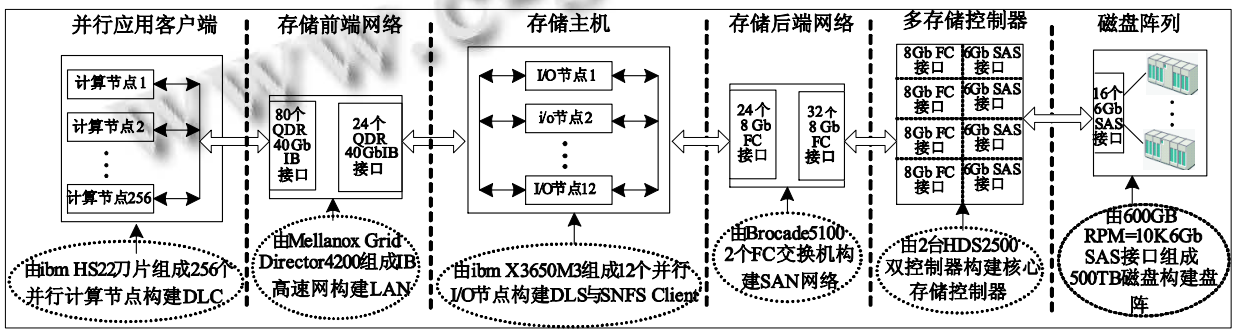

图 3　并行共享存储系统部署图

2.2 数据存储组织方式

合理的数据存储组织方式是高效的数据访问机制的前提, 高效的存储访问机制将并行共享存储的优越性得到最好发挥. 因此对于设计部署一个并行共享存储系统, 存储组织方式与存储访问机制是其核心, 对存储系统的功能和性能产生重要影响, 本节与 2.3 节分别对这二个方面进行详细阐述.

采用数据存储组织方式模型如图 5, 实现从物理层磁盘驱动器到逻辑层 LUN 的数据元素布局. 每一层设计原则: 充分发挥硬件性能, 利用分布与并行 I/O 的先进机制优化数据存储组织, 其设计方法如下: 数据存储组织模型描述见表 1, 存储组织方式设计采用矩阵表示, 行数表示为托盘位置,列数表示为磁盘驱动器在托盘中的排列位置.

表 1　数据存储组织模型描述

名称	数学描述	含义
磁盘驱动器	$D_{(m,n)}, m=1,...,56$ $n=1,...,15$	共 56 个托盘, 每个托盘 15 块磁盘驱动器构成 840 块磁盘驱动器.
托盘	$T_m, m=1,...,56$	每个托盘由 2 个输入, 2 个输出 SAS 通道，共享 15 块磁盘驱动器, 共 56 个托盘.
RAID 组	$R_k, k=1,...,104$	采用 8D+1P 构建用户数据 RAID 组, 共 102 个，2D+2P 构建元数据 RAID 组,共 2 个.
热备盘	$H_u, u=1,...,16$	16 个磁盘驱动器作为全局热备盘
池	$P_i, i=1,...,34$	3 个用户数据 RAID 组构建 1 个池, 共 34 个池.
逻辑单元号	$L_j, j=1,...,76$	1 个池均匀划分 2 个用户数据 LUN, 共 68 个. 2 个元数据 RAID10 均匀划分 8 个元数据 LUN.

行列唯一标识磁盘驱动器. (由于整个存储布局设计是完全对称的分 2 组, 故下面设计只给出 1 组, 存储访问机制设计也相同).

$$M_1=\begin{bmatrix} D^1_{(1...4,1...6)} & D^2_{(1...4,7...12)} \\ \vdots & \vdots \\ D^{13}_{(25...28,1...6)} & D^{14}_{(25...28,7...12)} \end{bmatrix}$$

$$M_2=\begin{bmatrix} D^{15-1}_{(1...4,13...15)} & D^{15-2}_{(5...6,15)} \\ D^{16}_{(5...16,13...14)} & D^{18}_{(17...24,15)} \\ D^{17}_{(17...28,13,14)} & D^{19}_{(25...28,15)} \end{bmatrix}$$

其中,

$[R_{3i-2},R_{3i-1},R_{3i}]=D^i, R_{103}=D^{19}$

$P_i=D^i, i=1,2,...,17$, $[L_{2i-1},L_{2i}]=P_i$

$H_r=D^{18}_{(m,15)}, r=m-16, r=1,...,8\ m=17,...,24$

$[L_{35},L_{36},L_{37},L_{38}]=R_{103}$, R_{103} 为元数据 RAID10.

从上述存储组织方式设计可知, 对用户数据与元数据分别采用 RAID5 与 RAID10 满足不同的实际性价比需要, 根据托盘规模预留相应的全局热备盘应对硬盘物理破损. 对用户 RAID 组按池的方式进行构建与在池层面进行 LUN 的划分, 为实现存储自动精简配置(Dynamic Provisioning)提供数据组织基础.

2.3 存储访问机制

本文给出的存储访问机制是基于并行作业共享的多存储主机、多端口、多路径模式, 图 4 给出了存储访问机制的分层设计模式. 由图 4 可知 4 层设计实现了冗错、故障切换与恢复、数据共享, 在性能上满足了地震数据流并行、负载均衡、高带宽聚合的需求.

采用集合分类及归属方法进行存储访问机制设计, 根据表 1 和表 2 定义的存储元素, 具体设计阐述如下:

(1) 驱动器通道端口 $S_{r,v}$ 划分 4 个子集合, 每个子集合包含 2 个端口: $S^v=\{S_{1,v},S_{2,v}\}$. 托盘 T_m 划分 4 个子集合, 每个子集合包含 7 个托盘:

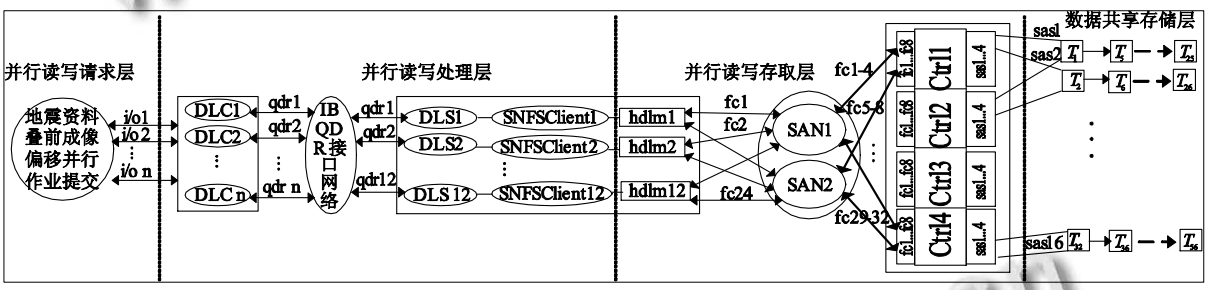

图 4 并行共享 I/O 存储访问机制图

表 2 存储访问机制设计

名称	数学描述	含义
控制器	$C_r, r=1,...,4$	4 个控制器.
驱动器通道端口	$S_{r,v}, v=1,...,4$	每个控制器的驱动器通道有 4 个路径端口, 共 16 个.
主机通道端口	$O_{r,p,t}, p=1,...,8\ t=1,2$	每个控制器的主机通道有 8 个路径端口, 每 4 个端口路径归属 1 个 SAN 网, 共 32 个.
条带组集	$G_s, s=1,...,24$	对 76 个 LUN, 分组为 16 个数据条带组集, 8 个元数据条带组.
文件系统	$F_w, w=1,...,8$	合并 LUN 生成 8 个并行文件系统

$T^v=\{T_{4*x+v}\mid x=0,...,6\}$. 子集合之间归属: $T^v\to S^v$.

(2) 对主机通道端口 $O_{r,p,t}$ 划分 4 个子集合.

$$O^1=\{O_{1,1,1},O_{1,2,1},O_{2,5,2},O_{2,6,2}\}$$

$$O^2=\{O_{1,3,1},O_{1,4,1},O_{2,7,2},O_{2,8,2}\}$$

$$O^3=\{O_{1,5,2},O_{1,6,2},O_{2,1,1},O_{2,2,1}\}$$

$$O^4=\{O_{1,7,2},O_{1,8,2},O_{2,3,1},O_{2,3,2}\}$$

(3) 对多主机识别的 LUN 归属到上述子集合.

$L_{2i-1}\to O^1, L_{35}\to O^1, i=1,...,9$

$L_{2i}\to O^2, L_{36}\to O^2, i=1,...,9$

$L_{2i-1}\to O^3, L_{37}\to O^3, i=10,11,...,17$

$L_{2i}\to O^4, L_{38}\to O^4, i=10,11,...,17$

对归属后的 LUN, 通过 Storage Navigator[3]与 HDLM [4]部署与配置方法实现每个 LUN 在 4 个主机通道端口, 即对每个 LUN 的访问数据在 4 条并行 I/O 路径上实现智能动态流量管理.

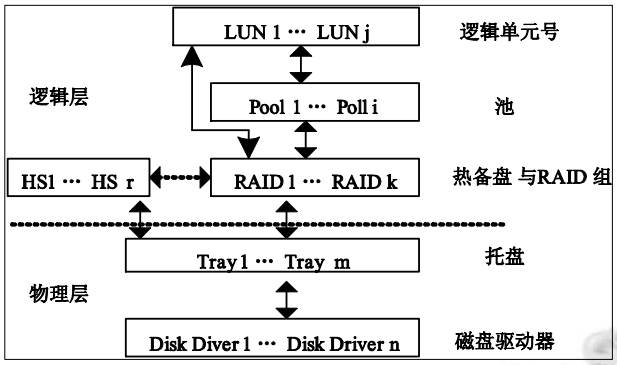

图 5　存储组织方式模型

(4) 根据图 2 存储数据组织, 对条带组 G_s 集合关系表征如下.

$$G_1=\{L\mid L_{2i-1}\},G_2=\{L\mid L_{2i}\}\ i=1,...,4$$
$$G_3=\{L\mid L_{2i-1}\},G_4=\{L\mid L_{2i}\}\ i=5,...,8$$
$$G_5=\{L\mid L_{2i-1}\},G_6=\{L\mid L_{2i}\}\ i=9,...,13$$
$$G_7=\{L\mid L_{2i-1}\},G_8=\{L\mid L_{2i}\}\ i=14,...,17$$
$$G_9=\{L\mid L_{35}\},G_{10}=\{L\mid L_{36}\}$$
$$G_{11}=\{L\mid L_{37}\},G_{12}=\{L\mid L_{38}\}$$

(5) 根据划分条带组 G_s 集合, 使用 SNFS 部署与配置方法[1,2]分布式生成 4 个并行共享文件系统.

$$F_i=\{G_i,G_{i+2},G_{34+i}\}\ i=1,2$$
$$F_{i-2}=\{G_i,G_{i+2},G_{32+i}\}\ i=5,6$$

通过上述的设计方法描述, 驱动器通道接口、主机通道接口、LUN 映射主机、SNFS 条带组划分及并行共享文件系统生成是存储访问机制的主要关键设计环节, 每个环节设计充分利用现有存储元素, 以提高并行性、可靠性、共享性为目标.

3　系统测试

与联机事务处理应用关注存储系统 IOPS 性能指标不同, 地震资料处理应用更关注的是并行大 I/O、大文件读写共享存储系统的吞吐量性能指标, 其影响常规处理和偏移处理. 使用 iozone[5] Benchmark 测试工具与实际地震资料数据测试并行读写存储聚合吞吐量、聚集 I/O 性能效率.

3.1 Benchmark 性能测试

在存储性能测试中, 选择的参数与测试的结果有很大关系.

1) Stripe Group 中的 Stripe Breadth 大小.

Stripe Breadth(简称 SB)是图 2 中 StripeGroup(简称 SG)的重要参数[6], 描述在 SG 内一次读写每个 LUN 的大小, 是 SNFS 并行读写的基本单位. 其大小直接影响并行文件系统读写性能. 图 6 是在 32 节点并行访问规模, 共享粒度为 8 个文件系统, 记录块大小为 4MB 读写环境下, 不同 SB 大小的聚合吞吐量变化. 图 6 结果数据表明在 SB 为 1MB 时, 吞吐量读写达到峰值, SB 为 1/8KB 时吞吐量读写达到谷值, 对 SB 参数性能优化有参考价值.

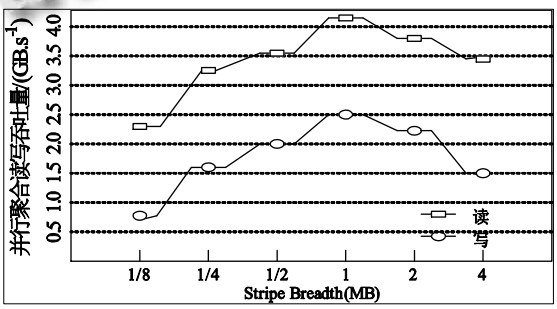

图 6　不同 Stripe Breadth 读写聚合吞吐量

2) 存储访问并行规模

增大存储访问并行规模能充分挖掘并行存储的优势, 提高 I/O 聚合带宽和数据吞吐量, 但有一个达到最佳性能的上限值制约. 图 7 为记录块大小为 4MB, SB 为 1MB, 共享粒度为 8 个文件系统环境下, 不同并行规模的吞吐量变化, 图 7 结果数据表明在 64 个节点并行访问时达到读写峰值, 分别为 4.3GB/s 与 2.6GB/s, 在 4 个节点时, 达到读写谷值, 分别为 2.8GB/s 与 1.0GB/s, 对并行规模参数优化有参考价值.

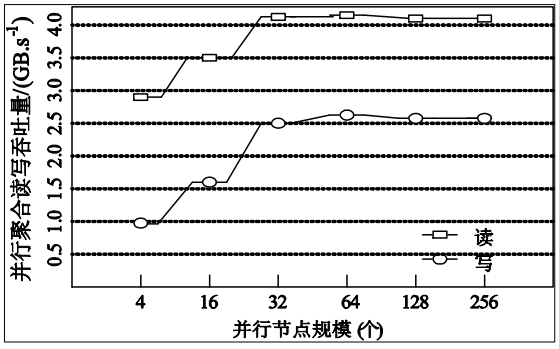

图 7　不同并行规模读写聚合吞吐量

3) 并行访问共享粗细粒度.

并行访问共享存储数据时, 并行 I/O 向一个或多个共享文件系统读写数据时, 共享文件系统个数的粗细粒度影响并行存储的等待时间. 图 8 为记录块大小为 4MB,SB 为 1MB, 并行规模 32 节点环境下, 不同共享文件系统个数的吞吐量变化, 图 8 结果数据表明在大于等于 6 个文件系统时达到读写峰值分别为 4.3GB/s 与 2.6GB/s, 在 1 个文件系统时达到读写谷值分别为 1.5GB/s 与 1.0GB/s. 共享粒度越粗, I/O 并行性越好, 但超过 6 个文件系统时, I/O 并行性基本饱和. 图 6 至图 8 测试结果数据证实了本文设计的基于并行分布共享数据访问思想的存储组织方式与存储访问机制是可行、有效的.

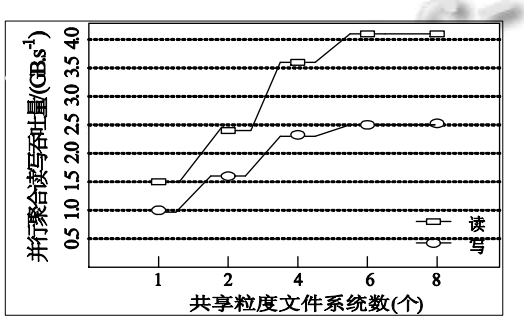

图 8 不同共享粒度读写聚合吞吐量

3.2 应用测试

采用某海域三维地震资料主要信息如表 4, 使用 Omega 地震资料处理软件对该数据进行叠前深度偏移测试. 并行度为 64 节点, 输入与输出地震数据各跨越三个并行文件系统, 某一时刻 I/O 节点的端口聚合带宽如图 9 所示. I/O 节点数与聚合带宽基本成线性关系, 可知多 I/O 节点是并行读写、共享存储数据、I/O 负载均衡运行, 达到预期效果.

4 系统优化建议与方法

根据性能测试与实际地震资料处理的应用运行, 对系统优化方法与建议如下:

1) 通过多个存储参数的组合优化, 提高吞吐量. 各个相关存储访问参数虽然独立设置, 但参数之间的匹配关系影响存储吞吐量性能.

2) 优化并行作业提交方式, 挖掘系统性能峰值. 控制选取适度的并行规模, 利用并行机制优势. 地震数据跨越多并行文件系统分布, 高效利用多路径、多 I/O 主机硬件资源.

表 4 某海域三维地震资料信息

覆盖面积	150Km2	线束	50
总炮数	23500 炮	接受线	60
覆盖次数	60	采样间隔	10ms
面元	25*25m	偏移深度	1000m

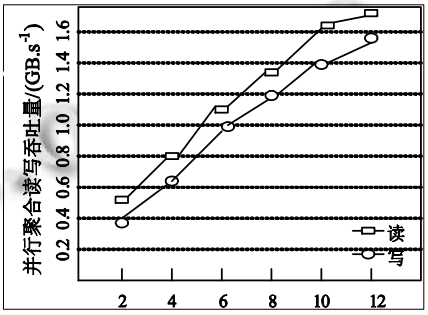

图 9 不同 I/O 节点数读写聚合吞吐量

5 结论

设计部署的地震资料处理并行共享存储系统, 通过测试表明了其优势, 该系统和本文提出的优化方法与建议已经在石油物探行业国内外某些深水海洋矿权区块地震资料处理项目上得到了广泛应用, 取得了良好的效果.

参考文献

1 Quantum. Stor Next File System Installation Guide. Seattle USA: Quantum Press, 2010.

2 Quantum. Stor Next File System User's Guide. Seattle USA: Quantum Press, 2010.

3 HDS. Hitachi Storage Navigator Modular 2 User's Guide. Santa Clara USA: Hitachi Data Systems Press, 2010.

4 HDS. Hitachi Dynamic Link Manager Software User's Guide for Linux. Santa Clara USA: Hitachi Data Systems Press, 2010.

5 Iozone Filesystem Benchmark. http://wwwiozone.org/ docs/ IOzone_msword_98.pdf.

6 Quantum. Stor Next File System Tuning Guide. Seattle USA: Quantum Press, 2010.