

一种高性能对象存储系统

刘群¹, 冯丹², 王芳²

(1. 华中科技大学网络与计算中心, 武汉 430074; 2. 华中科技大学信息存储系统教育部重点实验室, 武汉 430074)

摘要: 当大量用户并行访问对象存储设备(OSD)时, OSD可能成为限制系统性能的瓶颈。该文提出一种基于网络磁盘阵列的对象存储系统, 网络磁盘阵列包括外设和网络2个通道, 可直接与客户端进行数据传输。该系统具有基于对象存储系统的特点, 且系统容量可扩展性和整体性能较高。

关键词: 网络磁盘阵列; 存储对象; 基于网络磁盘阵列的对象存储系统

High Performance Object Storage System

LIU Qun¹, FENG Dan², WANG Fang²

(1. Network and Computer Center, Huazhong University of Science and Technology, Wuhan 430074;

2. Key Laboratory of Data Storage System, Ministry of Education, Huazhong University of Science and Technology, Wuhan 430074)

【Abstract】 Object-based Storage Device(OSD) is possible to be the bottleneck which restricts the system performance during a large number parallel access to the same OSD. This paper proposes an object storage system based on Net-RAID. Net-RAID has peripheral channel and network channel, and data directly transfer between Net-RAID and clients. The system has characteristics of object-based storage system, better capacity extendibility and performance.

【Key words】 Net-RAID; storage object; Object Storage System Based on Net-RAID(OSSBNR)

1 概述

传统直接存储系统(Direct Access Storage, DAS)采用块接口, 系统总线的限制导致服务器成为系统性能瓶颈。为了提高DAS的可扩展性, 存储区域网(Storage Area Network, SAN)通过高速、专用的数据存储网向外界提供块服务, 但其共享粒度大, 元数据服务器(MetaData Server, MDS)管理系统中的元数据容易降低系统性能。附网存储(Network Attached Storage, NAS)通过网络接口把存储设备直接连入网络并提供文件接口, 实现细粒度数据共享和跨平台文件共享, 但文件服务器处于数据访问路径上。

基于对象存储(Object-Based Storage, OBS)^[1]技术是新一代存储模式, 是存储领域的发展方向, 它克服了NAS和SAN的不足, 采用对象作为接口, 具有块接口的快速和文件接口便于共享的优点, 且在扩展性、安全性、系统性能和跨平台数据共享等方面优于SAN及NAS。

CMU并行数据实验室的NASD项目使OBS得到实质性发展^[2]。存储网络工业协会在此基础上成立了OBS工作组(OBS TWG), 定义SCSI OBS Device Commands-2^[3], 并将其纳入SCSI-3协议框架。

许多公司和实验室研究并实现了基于对象的存储系统。Cluster file system公司实现的对象访问协议文件系统——Lustre^[4]被应用于美国能源部和Lawrence Livermore。Panasas公司实现了ActiveScale文件系统^[5]。UCSC存储系统研究中心研究开发了支持10 000个客户机、100 GB/s吞吐量、2 PB系统总容量的并发访问对象存储系统^[6]。

基于对象存储系统(Object-Based Storage System, OBSS)是将数据的逻辑视图(约占负载的20%)和物理视图(约占负载

的80%)分别交给MDS和对象存储设备(Object-based Storage Device, OSD)管理。OSD是一个智能设备, 由处理器、内存、网络接口和物理存储设备组成, 负责存储空间的分配和数据组织, 维护所有对象的空间分配及所有与空闲空间有关的元数据, 实现对象到存储块设备的映射。当大量用户同时并行访问OSD时, OSD可能影响系统性能。因此, 本文以具有外设与网络的双通道网络磁盘阵列(Net-RAID)作为物理存储设备, 提出一种新的基于网络磁盘阵列的对象存储系统(Object Storage System Based on Net-RAID, OSSBNR), 消除OBSS的潜在瓶颈, 实现对象控制与数据存储分离, 提高了系统整体性能。

2 基于网络磁盘阵列的对象存储系统

OSSBNR由元数据服务器(Metadata Server cluster, MDS)集群、存储对象(Storage Object, SO)、客户端(Client)及高速网络组成, 如图1所示。SO是一个智能存储实体, 由存储对象服务(Storage Object Service, SOS)和Net-RAID组成, 是OSSBNR的基本单位。它与OBSS中的OSD有明显不同, 其本身也是对象, 包含数据、属性和方法, 具有接口与状态, 由128 bit的唯一对象ID(UID)识别。SOS包括处理器、内

基金项目: 国家“973”计划基金资助项目“下一代互联网信息存储的组织模式和核心技术研究”(2004CB318201); 国家自然科学基金资助项目“树结构并行存储系统——磁盘树”(60503059); 校实验技术研究基金资助项目“计算机网络实验教学的改革与研究”

作者简介: 刘群(1969-), 女, 工程师、博士, 主研方向: 计算机网络, 高性能网络存储; 冯丹, 教授、博士; 王芳, 教授、博士

收稿日期: 2008-02-10 **E-mail:** liliuqun@mail.hust.edu.cn

存、网络接口、外设通道(SCSI 或 Fibre)、自带永久性存储介质磁盘及对象控制软件。SOS 负责数据对象存储空间的管理和分配。Net-RAID 是数据存储的物理设备^[7],具有 2 类通道:(1)连接主机的主从外设通道,通过该通道磁盘阵列向主机提供标准存储设备的块服务;(2)直接接入网络的网络通道,它使磁盘阵列以对等身份与所有网上设备通信,形成不同于外设通道协议且物理上分离的数据通道。

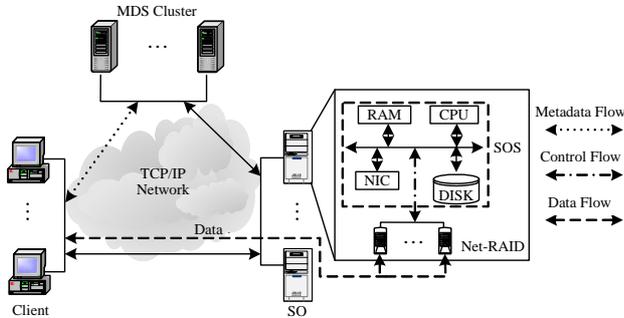


图 1 OSSBNR 的体系结构

SO 通过外设通道控制所属的网络磁盘阵列,分配存储空间,根据对象 ID 映射为逻辑块,利用 CPU 计算能力优化数据分布,并建立对象属性的索引机制,将对象的属性存储于 SOS 的本地磁盘中,数据对象则存储在 Net-RAID 中,根据对象属性调整存储策略。因此,在体系结构上它不同于 DAS, NAS, SAN, NAS-SAN 混合式及 OBSS,将元数据流、控制流、数据流进行分离。

3 OSSBNR 的扩展方式

目前,存储系统扩展的唯一方法是增加存储设备,虽然 OBSS 有良好的可扩展性,但当单个 OSD 内有多个 RAID 时,若大量数据通过外设通道存储转发,将降低系统性能。

OSSBNR 的扩展方式有对象内扩展和对象间扩展。对象内扩展方式通过增加外设通道,在一个 SO 内挂接新的 Net-RAID,其扩展能力由 SOS 外设通道适配器的支持能力决定。可以由 Fibre 将带有网络通道的 RAID 组成 SAN,随 Net-RAID 的增加,系统的 I/O 带宽随之动态扩展。对象间扩展方式通过增加 SO 来实现,这种方式与在 OBSS 中增加 OSD 相似,但 OBSS 中每增加一个 OSD,只增加一个网络带宽,而 OSSBNR 中每增加一个 OS,就增加了含有 Net-RAID 数目的聚合带宽。

假设 SO 有 2 个 Net-RAID,都具有 400 Mb/s 的传输能力,若数据分条状跨 2 个 SO,则 Client 直接与 4 个 Net-RAID 进行数据传输,聚合带宽可达 1 600 Mb/s;若每个 SO 包含 3 个 Net-RAID,则聚合带宽可达 2 400 Mb/s,使 OSSBNR 具有系统容量和系统性能同步扩展的特点。

4 性能分析与测试

4.1 性能分析

吞吐率是存储系统的重要性能指标,指存储系统在单位时间内完成的任务量,有 2 种形式:(1)IOPS,即每秒 I/O 访问的次数,通常用于事务处理;(2)数据传输率,即每秒传输的数据量。本文的性能分析和实验均采用数据传输率。

由于 MDS 在 OBSS 与 OSSBNR 中的作用相同,在相同情况下相同请求花费的时间相同,因此本文忽略 MDS 对系统性能的影响,只考虑 OSD 或 SO 与网络对系统性能的影响。引入的一些系统参数如表 1 所示。

表 1 系统参数

参数说明	OSD	SO
CPU 时钟速率	S'	S
CPU 时钟周期	C'	C
单个 RAID(或者 Net-RAID)带宽	B'_{RAID}	$B_{Net-RAID}$
BUS(或者网络)带宽	B'_{BUS}	B_{Net}
吞吐率	T'	T
系统吞吐率	ST'	ST
RAID(或者 Net-RAID)的数目	m	
OSD(或者 SO)的数目	n	

假设 OSD 或 SO 处理请求均采用 3 个流水阶段:底层 I/O 调度,数据拷贝和返回 Client,若系统请求能充满流水线,即系统中有大量数据等待处理,则此时效率最高。因此,OBSS 和 OSSBNR 的性能如下:

$$ST' = n \times T' = n \times \min(S'/C', m \times B'_{RAID}, \lambda_i \times B'_{BUS})$$

$$ST = n \times T = n \times \min[(m \times S)/C, m \times B_{Net-RAID}, \mu_j \times B_{Net}]$$

其中, $i, j = 1, 2, \dots, m$; λ 和 μ 分别为 OBSS 和 OSSBNR 处理能力系数,数值在 (0, 1) 之间。

由上述分析可知,若 OSD 与 SO 各种硬件配置相同,即 $S' = S, C' = C, B'_{RAID} = B_{Net-RAID}, \lambda = \mu$,则增加 RAID 的网络通道使其成为 Net-RAID。由于 $B'_{BUS} < B_{Net}$,因此在单个 SO 中增加 Net-RAID 或 SO, OSSBNR 性能将都显著超过 OBSS。

4.2 性能测试

OSSBNR 的实验硬件配置如下:(1)一个 MDS,其 CPU 为 Xeon3.0,内存为 1 GB,操作系统为 Linux2.4,硬盘配置为 Maxtor 80 GB, NIC 为集成双 10/100/1 000 以太网卡。(2)2 个 SOS,其中每个包括 2 个 Net-RAID,其 CPU 为 Xeon3.0,内存为 512 MB,操作系统为 Linux2.4,硬盘配置为 Maxtor 80 GB, NIC 为 SK98lin10/100/1 000。(3)2 个 Net-RAID,其 CPU 为 Celeron2.66,内存为 512 MB,操作系统为 Linux2.2,硬盘配置为 Maxtor 80 GB \times 2, NIC 为 Yukon10/100/100。

网络环境配置为 1 000 Mb/s 交换以太网,交换机为 Cisco3750G-24T-S。OBSS 与 OSSBNR 硬件配置基本相同,但去除 OSSBNR 中的 Net-RAID 的网络接口,使之成为 OSD 中存储设备 RAID。采用 FTP 应用,测试 1 台 ~ 4 台客户端同时上传或下载文件大小约为 1 GB 时系统聚合带宽。

图 2 表示 2 种系统在单个 SO 或单个 OSD 环境中上传和下载文件的聚合吞吐率曲线,图 3 显示 2 个 SO 或 2 个 OSD 进行上传和下载文件时 2 个系统的聚合吞吐率。

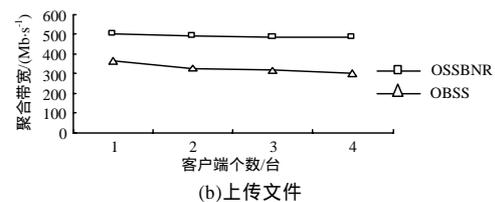
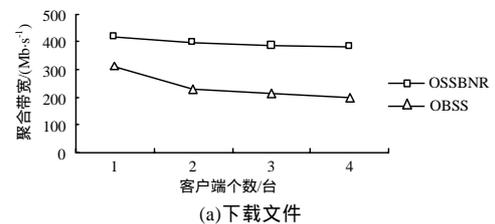


图 2 OSSBNR 中单个 SO 与 OBSS 中单个 OSD 的性能比较

(下转第 74 页)