第 34 卷 第 5 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.5 Computer Engineering March 2008
·软件技术与数据库· 文章编号：1000—3428(2008)05—0071—03 文献标识码：A 中图分类号：N945

基于 OBS 的分布并行海量地形数据服务系统

郑 胜[1]，喻占武[2]，李忠民[2]
(1. 武汉大学电子信息学院，武汉 430079；2. 武汉大学测绘遥感信息工程国家重点实验室，武汉 430079)

摘 要：海量地形数据的存储与管理是大规模地形实时漫游系统的关键。该文提出一种基于对象存储的分布式并行地形数据服务系统(DPTSS)，采用自治的存储对象存储和管理地形块数据，实现了控制路径和数据路径分离。通过元数据集群提供高效率和高可用的元数据服务，以及基于对象的存储集群实现并行的地形数据块传输服务，提供高吞吐率和高带宽的地形数据服务。对比实验表明，DPTSS 在较低的 TCO 情况下能提供高性能的地形数据服务。

关键词：基于对象存储；海量；分布；并行；地形数据

Distributed Parallel Service System for Massive Terrain Data Based on OBS

ZHENG Sheng[1], YU Zhan-wu[2], LI Zhong-min[2]
(1. School of Electronic Information, Wuhan University, Wuhan 430079;
2. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079)

【Abstract】The storage and management of massive terrain data is the key of large-scale terrain roaming. This paper proposes a distributed parallel mechanism based on object storage——Distributed Parallel Terrain Service System(DPTSS), which adopts self-rule storage object to store and manage terrain block data and realizes the division of control path and data path. Divided metadata cluster provides high-effective and high-available metadata service, and OSD cluster with parallel terrain data transmission provides high throughout and bandwidth. A prototype system for DPTSS is established. Experimental results show that DPTSS in lower TCO enables good terrain data distributed service.

【Key words】Object-Based Storage(OBS); massive; distributed; parallel; terrain data

随着空间探测技术的快速发展，地形数据获取的数量迅速增长，大规模甚至超大规模三维地形实时漫游系统的应用越来越广泛，如数字地球、大规模军事仿真等。这类系统对地形绘制的效率有很苛刻的要求，需要高效的地形绘制算法和数据调度策略。由于它属于数据密集型的应用，对存储系统的存储容量、I/O性能也有很高的要求。国内外针对地形的实时绘制开展了大量的研究工作，并取得了突破性的进展，如视点相关的LOD和可见性剔除等技术，有效提高了地形绘制的效率。但是海量地形数据存储和管理方面的研究还比较少，根据“木桶原理”，系统的存储瓶颈直接制约了系统整体性能的提高，进而影响了地形漫游的效果。面对地形实时漫游系统对存储容量和性能不断增加的需求，如何提供一个高性能、可扩展、高可用、安全的、高性价比的存储系统必将成为大规模地形漫游系统需要解决的问题。本文通过分析目前各主流存储系统的优缺点，提出了基于新一代网络存储技术——对象存储系统(Object Storage System)[1]的分布式并行地形服务系统(Distributed Parallel Terrain Service System, DPTSS)，并分析了地形数据分布策略和数据读取性能。

1 对象存储系统

近年来，基于对象存储(Object-based Storage, OBS)是存储领域研究的一个新热点，将面向对象技术引入存储领域，结合NAS和SAN两者的优点，具有高性能、可扩展、高可用和可靠的安全性[2]。对象存储系统主要由对象(object)、对象存储设备(Object Storage Device, OSD)、元数据服务器(Meta-Data Server, MDS)和网络 4 部分组成：

(1)对象是对象存储系统中的关键组件，是对象存储系统中数据存取的基本单元。区别于以往的数据存储单位——文件或者块，一个存储对象是指存储设备内一组逻辑相关的数据块的集合，它是应用数据和存储访问属性的组合。传统的块存储系统中，存储系统必须跟踪系统中每个块的全部属性，而对象存储系统则不同，对象包含了允许数据自治和自我管理的属性信息，分担了存储系统的部分数据管理工作，简化了系统的管理任务，提高了系统的管理效率和灵活性。

(2)对象存储设备是存储对象的载体，自带 CPU、内存、网络接口和磁盘系统，对于本地存储的数据具有充分的自主性和自治性。对象存储设备主要有 3 个功能特征：1)数据存储。OSD 与基于块的设备的主要区别不是介质，而是接口。OSD 对外提供基于对象的设备接口，通过对象 ID、偏移来读写数据，对内实现存储对象到物理存储介质的映射。基于对象的接口使得计算节点在取得安全证书的情况下，能够直接访问 OSD 存取数据，而不必使用文件服务器或数据库服务器作为数据存取的中介，避免了文件或数据库服务器作为中介时可能出现的瓶颈。2)智能分布。OSD 能通过自身的 CPU、内存以及对象的属性信息优化数据分布，支持数据的预取和

基金项目：国家“973”计划基金资助项目(2004CB318206)
作者简介：郑 胜(1980－)，男，博士研究生，主研方向：分布式计算体系和方法；喻占武，教授、博士生导师；李忠民，博士研究生
收稿日期：2007-05-20 **E-mail：**zsh.whu@gmail.com

缓存，进而优化磁盘的性能。3)对象元数据管理。OSD 管理其内部存储对象的元数据，包括对象的数据块和对象的长度等信息。

(3)元数据服务器控制计算节点与 OSD 对象的交互，主要有以下几个功能：1)共享认证机制。MDS 和 OSD 共享访问认证协议，只有 OSD 获得 MDS 的信任状后，才能加入存储系统、扩充容量。2)对象存储访问。MDS 构造、管理描述每个文件分布的视图，允许直接访问对象。MDS 为应用服务器提供访问该文件所含对象的能力，OSD 接收到每个请求时先验证该能力，然后才能访问。3)文件和目录访问管理。MDS 在存储系统上构建一个文件结构，包括限额控制、目录和文件的创建和删除、访问控制等。4)缓存一致性。为了提高性能，通常支持应用服务器端对象缓存机制。MDS 提供对象缓存更新通告机制，防止了服务器端对象的缓存不一致。

(4)网络结构：计算节点、MDS 和 OSD 之间可以采用任何一种基于 TCP/IP 的网络连接，如以太网、InfiniBand 和 Myrinet。通常采用千兆以太网，因为千兆以太网可以提供 FC 的传输性能，还可以降低成本及实施、管理的技术难度。

2 基于 OBS 的 DPTSS

2.1 DPTSS 的体系结构

在地形漫游系统中，为了降低计算机内存消耗、加快计算机的处理时间、缩短网络传输时间，通常将地形数据以分层分块的瓦片金字塔模型来组织。根据地形组织的这种特点和对象存储系统的相关特性，本文基于 OBS 的分布并行地形数据存储系统包括地形分发服务器集群、元数据集群和 OSD 集群 3 大主要组成部分：地形分发服务器集群提供并行的数据传输；元数据集群提供并行的元数据检索；OSD 集群提供并行的数据存取机制。其系统结构如图 1 所示。

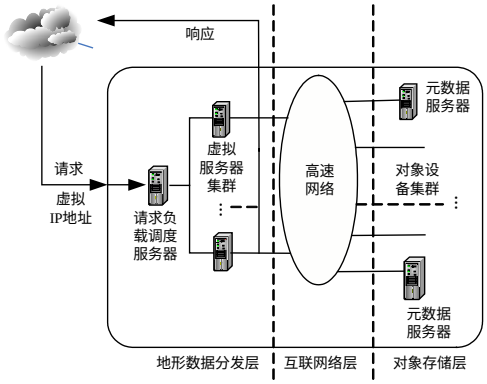

图 1 DPTSS 的结构

(1)数据分布服务器集群

由于 OBS 系统采用 OSD-2 协议传输数据，目前通用文件系统的客户端无法访问 OBS 系统的存储资源，因此必须有一类类似于网关的节点完成系统内外的交互与通信，其主要功能包括：

1)与远端客户进行连接交互，实现访问控制和身份认证功能；

2)协议转换与适配功能，解析来自 Internet 基于 TCP/IP 协议的地形数据请求，并将其转换为基于 iSCSI 协议的对象数据请求，实现 OSD 的远程、并行访问；

3)与 OSD 和元数据服务器的交互功能以及与 OSD 协同完成多级缓存的管理与调度功能；

4)负载均衡的能力。

(2)元数据服务器集群

1TB的地形数据按照每块 64 KB划分将产生 2 097 152 个块文件，管理上千万甚至过亿的文件不论是对文件系统还是数据库都是一个挑战。海量的文件将导致元数据非常庞大，从而使读取和检索元数据的响应时间过大，以致于地形漫游系统的远程用户无法忍受。目前常用的地形数据管理方法是聚合一定量的小块瓦片数据作为一大块数据，存入文件，在文件里建立索引。这种方式能够有效减少小文件的数目，减轻元数据管理的负载，但是牺牲了块数据的并行读取性能和系统的响应时间。对于OBS而言，将占元数据管理 90%工作量的物理视图元数据管理分布到多个OSD上[2]，MDS只须提供逻辑视图元数据和认证功能，因此，其负载能力得到极大的提高，有能力支持高性能海量数据文件的检索和高并发的用户访问。此外，MDS只须提供数据逻辑视图，便于利用目录子树分割、散列分割等方法将元数据分布到多个元数据服务器上，形成元数据服务集群，从而具有高效的元数据访问、动态负载平衡、失败接管和方便扩展的能力。其主要功能有：1)提供统一的逻辑视图，建立地形数据块编码与OSD的ID映射的统一目录树型逻辑结构；2)提供基于信任状的访问控制，使存储在OSD内的地形数据安全得到有力保障； 3)根据OSD反馈的负载信息，对热点数据实施相应的复制算法和数据迁移策略，均衡OSD集群的负载，提高系统吞吐率。

(3)OSD 集群

在 DPTSS 中，由多个 OSD 组成的存储节点群是地形数据和纹理的最终存放场所。一组 OSD 能够并发地为用户提供地形数据服务，从而满足大规模地形漫游对存储容量和性能不断增加的需要。单个 OSD 只负责管理与维护存储于本地的数据而不必关心其他 OSD 存放的数据，因此，在 DPSS 系统中 OSD 集群的功能主要包括：

1)响应来自授权数据分发服务器的读写请求，完成远程节点对地形数据的访问服务；

2)与数据分发服务器协同完成多级缓存的管理与调度；

3)定期和元数据服务器交互，报告自身的负载以及存储资源的使用情况，以便元数据服务器合理地复制热点地形数据和有效地调度任务，使系统中各 OSD 负载均衡，提供更好的数据访问性能。

2.2 地形数据访问

DPTSS 中的地形数据访问模型如图 2 所示。

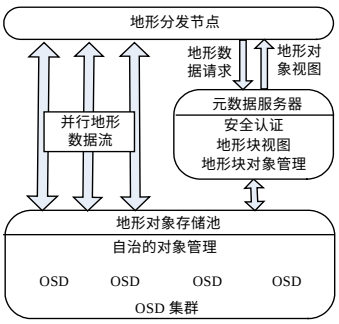

图 2 DPTSS 读数据事务处理模型

数据分发服务器将远程地形漫游客户端发送逻辑瓦片请求(通常是一个查询窗口内包括的数块瓦片数据)转发给元数据服务器；元数据服务器对客户端进行认证，产生信任状后，将逻辑瓦片请求映射成存储对象 ID 列表，然后将信任状和对象 ID 列表发回数据分发服务器 数据分发服务器与一组 OSD

建立连接，OSD 验证信任状后，数据分发服务器将和多个 OSD 建立连接通道，并行传输数据；数据分发服务器和远程客户端建立多线程机制，并行传输瓦片数据。

2.3 数据分布策略

数据布局策略是提高计算性能的关键因素，数据分布是并行数据库中提供高性能查询的常用技术，但是并行数据库价格昂贵，磁盘阵列提供的数据并行存储，通常只能按照某一 RAID 级别机械地进行，缺乏灵活性。OBS 作为一种分布式存储系统，支持大数据的分布。目前的 OBS 系统中，通常采用顺序分片映射或哈希的方式存储，但是地形漫游系统通常需要提取一个查询窗口内相邻数个瓦片的数据，为了提高响应时间，应使彼此相邻或相近的瓦片尽可能地分布到不同的 OSD 上。

针对二维数据分布，人们提出了许多算法，包括磁盘模数分布、域异域分布和希尔伯特曲线等。这些方法虽然都能够实现分开二维平面中相邻瓦片数据的目的，但是在磁盘数据较大的情况下，系统的响应时间不能提供下限保证，这会导致实时地形漫游系统出现用户等待时间过长。文献[3]提出了矢量基对分布方法(Vector-based Declustering)。该算法对于给定的磁盘数 M，产生一对线形无关的矢量 $\boldsymbol{u}=(a,b)$ 和 $\boldsymbol{v}=(c,d)$，并假定：

(1)瓦片 $(tx_1,ty_1),(tx_2,ty_2)$ 记为矢量 $\boldsymbol{T}_1$ 和 $\boldsymbol{T}_2$；

(2)由 $\boldsymbol{U}$, $\boldsymbol{V}$ 张成的线形空间记为

$$S(\boldsymbol{u},\boldsymbol{v})=\{\boldsymbol{w}\mid\forall m,n\in Z,\ \boldsymbol{w}=m\boldsymbol{u}+n\boldsymbol{v}\}$$

(3)S 空间中非零矢量间最短的距离记为

$$L(r)=\min\{r\mid r=|\boldsymbol{w}_2-\boldsymbol{w}_1|,\ \boldsymbol{w}_1,\boldsymbol{w}_2\in S\}$$

(4)如果 $\boldsymbol{T}_2=\boldsymbol{T}_1+\boldsymbol{w}$，则 $\boldsymbol{T}_1$ 和 $\boldsymbol{T}_2$ 同处一个磁盘。

该算法能够保证查询窗口半径小于 $L(r)/2$，每个磁盘最多被访问一次。其他分布方法不能提供这样的保证。矢量基对分布方法的这种性质对实时地形漫游系统来说非常重要，它能够使查询窗口在满足一定要求的情况下，响应时间恒定，使系统运行比较稳定。因为矢量基对分布方法能够满足地形实时漫游系统对数据分布的要求，所以 DPTSS 选择其作为数据分布策略。

3 系统原型与对比分析

在分析对比 DPTSS 性能时，对于对象存储系统所具有的可扩展性、高带宽等性能，这里不作验证，而是主要将 DPTSS 和目前的地形数据存储系统进行对比，分析其数据读取效率。并建立了 DPTSS 原型实验系统和 2 类具有代表性的地形数据存储系统。对比系统 1 采用光纤磁盘阵列+文件方式，系统 2 采用光纤磁盘阵列+数据库方式。DPTSS 与对比系统的硬件配置如下：

(1)DPTSS

1) 4 台对象存储设备(OSD)：CPU 为 P4 2.4 GHz；内存为 256 MB；硬盘为 Seagate 400 GB，SATA 接口；网卡为 1 000 Mb/s。

2)元数据服务器(MDS)：型号为 IBM 346；内存为 1 GB；网卡为 1 000 Mb/s。

3)数据分发服务器：型号为 IBM 346；内存为 1 GB；网卡为 1 000 Mb/s。

4)交换机为华为 12 口全千兆交换机。

(2)对比系统

1)光纤磁盘阵列：型号为 proware SB-3163FA；阵列级别为 RAID 0；外接主机通道为 2 个光纤通道；内置 4 块 Seagate 400 GB SATA 接口硬盘。

2)数据库或文件服务器：型号为 IBM 346；CPU 为 Xeon 3.0 GHz；内存为 1 GB；网卡为 1 000 Mb/s；操作系统为 Window 2003 Server。

实验数据包括：(1)最高分辨率为 90 m 的全球 SRTM 地形数据，数据块大小为 150×150，金字塔层数为 8 层，采用 7Z-ZIP 压缩，数据块总数为 3 538 890；(2)最高分辨率为 30 m 的全球 Landsat 7 卫星影像数据，数据块大小为 512×512，金字塔层数为 5 层，数据块总数为 4 364 800，数据块格式为 JPEG。

图 3 是 DPTSS、对比系统 S1 和对比系统 S2 在单用户条件下块数据读性能的比较。从中可以看出， S1 和 S2 在相同硬件配置的情况下，S2 的数据读取时间为 S1 的 10 倍左右，主要是因为对于海量数目文件的检索，数据库的效率远高于文件系统。另外，DPTSS 的 TCO(Total Cost Ownership)在远低于对比系统的条件下，效率略优于 S2。可以预见，在提高 TCO 的条件下，如果增加 OSD 和建立元数据服务器集群、提高系统的聚合带宽和元数据负载均衡，系统吞吐率和带宽将得到进一步的提高。

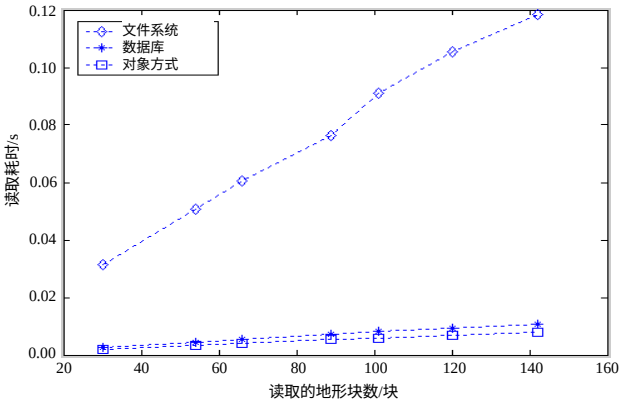

图 3 块数据读性能比较

4 结束语

海量地形数据的存储与管理是数字地球等大规模地形可视化应用系统的关键。本文提出一种基于对象存储的分布式并行海量地形存储系统，它能够以较低的 TCO 满足大规模地形可视化应用的需要。DPTSS 以 OSD 设备为数据的存储载体，以地形块对象为基本的数据存取单位，具有优良的可扩展性，能满足海量地形数据增长的需要。实验表明，在较低的 TCO 条件下，DPTSS 能够提供接近于光纤磁盘阵列和 Oracle 数据库管理的地形数据系统的性能。由于 OBS 系统的带宽随着 OSD 集群数量的增加呈线性增长，在投入增加的情况下，DPTSS 将能够为数字地球、数字城市等大规模地形漫游系统提供高吞吐率、高带宽的地形数据服务。

参考文献

[1] Mesnier M, Ganger G R, Riedel E. Object-based Storage[J]. IEEE Communications Magazine, 2003, 41(8): 84-90.

[2] Panasas. Panasas White Paper: Object Storage Architecture[Z]. [2006-01-20]. http://www.panasas.com.

[3] Chen Ling, Rotem T D, Seshadri S. Declustering Databases on Heterogeneous Disk Systems[C]//Proc. of the 21st International Conference on Very Large Data Bases. Switzerland: [s. n.], 1995-09.