第33卷 第10期 计 算 机 工 程 2007年5月
Vol.33 No.10 Computer Engineering May 2007
·开发研究与设计技术· 文章编号：1000—3428(2007)10—0274—03 文献标识码：A 中图分类号：TP393

NFS over Lustre 性能评测与分析

张 媛，卢泽新，刘亚萍
（国防科学技术大学计算机学院，长沙 410073）

摘 要：传统的网络文件系统难以满足高性能计算系统的I/O 需求，基于对象存储的全局并行文件系统Lustre可以有效地解决传统文件系统在可扩展性、可用性和性能上存在的问题。该文介绍了Lustre文件系统的结构及其优势，对NFS over Lustre 进行了性能测试，并将测试结果与Lustre文件系统、NFS网络文件系统及本地磁盘Ext3文件系统的性能进行了比较分析，给出了性能差异的原因，提出了一种可行的解决方法。

关键词：Lustre；NFS；文件系统；性能测试

Performance Evaluation and Analysis of NFS over Lustre

ZHANG Yuan, LU Zexin, LIU Yaping
(School of Computer, National University of Defense Technology, Changsha 410073)

【Abstract】 Traditional network file system can not meet the demand of high-performance computing systems, but object-based global parallel file system——Lustre can resolve the problems of traditional file systems in scalability, availability, performance. This paper introduces the architecture and advantage of the object-based file system——Lustre. It does the test of NFS over Lustre, and compares its result with Lustre, NFS and Ext3. It points out the reason of the differences and gives a feasible method.

【Key words】 Lustre; NFS; File system; Performance test

随着计算机及其相关技术的飞速发展和计算机应用的日益普及，处于计算领域最高端的高性能计算机已经成为本世纪IT领域人们争夺的制高点。目前高性能计算机已经从科学计算和工程计算发展到商业应用和网络信息服务领域。因此，高性能计算机的发展也将面临更多的挑战。网络化是高性能计算机发展的趋势，在网络化的应用中不仅需要较高的运算能力，同时对存储管理也提出了很高的要求。传统的文件系统难以满足高性能计算的I/O需求，基于对象存储(Object-based Storage)的全局并行文件系统凭借其在可扩展性、可用性和性能方面的诸多优势，在高性能计算机存储管理系统的设计中倍受青睐，Lustre文件系统就是其中的典型代表。在高性能计算机网络化研究的过程中，部署基于对象存储的Lustre文件系统与网络服务如何有机地结合已成为当前的研究热点之一。

1 Lustre对象存储文件系统

1.1 对象存储文件系统的关键技术

传统存储结构中的元数据服务器通常提供2个主要功能：(1)为计算结点提供一个存储数据的逻辑视图(Virtual File System, VFS)、文件名列表及目录结构。(2)组织物理存储介质的数据分布(inode层)。对象存储结构[1]将存储数据的逻辑视图与物理视图分开，并将负载分布，避免元数据服务器引起的瓶颈(如NAS系统)。元数据的VFS部分通常是元数据服务器10%的负载，剩下90%的工作(inode部分)是在存储介质块的数据物理分布上完成的。在对象存储结构中，inode工作分布到每个智能化的基于对象存储设备(Object-based Storage Device, OSD)上，每个OSD负责管理数据分布和检索，这样90%的元数据管理工作分布到智能的存储设备，从而提高了系统元数据管理的性能。另外，分布的元数据管理，在增加更多的OSD到系统中的同时，可以增加元数据的性能和系统存储容量。

并发数据访问对象存储体系结构定义了一个新的、更加智能化的磁盘接口OSD。OSD是与网络连接的设备，它自身包含存储介质，如磁盘或磁带，并具有足够的智能可以管理本地存储的数据。计算结点直接与OSD通信，访问它存储的数据，由于OSD具有智能性，因此不需要文件服务器的介入。如果将文件系统的数据分布在多个OSD上，则聚合I/O速率和数据吞吐率将线性增长，对绝大多数Linux集群应用来说，持续的I/O聚合带宽和吞吐率对较多数目的计算结点是非常重要的。对象存储结构提供的性能是目前其它存储结构难以达到的。

1.2 Lustre文件系统

Lustre文件系统是由Cluster File Systems公司开发的一个开源、高性能的分布式并行全局文件系统。Lustre针对大文件的读/写做了优化，可以为集群系统提供高性能的I/O吞吐率、全局数据共享环境、数据存储位置独立性和对节点的失效提供了冗余机制，以及当集群重新配置或者服务器和网关失效时的快速恢复服务，较好地满足了高性能计算集群系统的需要。

Lustre使用了基于对象的存储技术，基于意图的分布式锁管理机制，元数据和存储数据相分离的解决方案，提供了全局的命名空间，并融合了传统分布式文件系统的特色和传

基金项目：国家“973”计划基金资助项目(2003CB3148020)
作者简介：张 媛(1982－)，女，硕士生，主研方向：网络体系结构，高速网络；卢泽新，研究员；刘亚萍，副研究员
收稿日期：2006-05-25 **E-mail：**zhangyuan820111@126.com

统共享存储集群文件系统的设计思想，消除了传统文件系统在可扩展性、可用性和性能上的问题[3]。

Lustre 对象存储文件系统就是由客户端(client)、存储服务器(Object Storage Target, OST)和元数据服务器(MDS)3 个主要部分组成。Lustre 的客户端运行 Lustre 文件系统，它和 OST 进行文件数据 I/O 的交互，和 MDS 进行命名空间操作的交互，基本结构如图 1 所示。

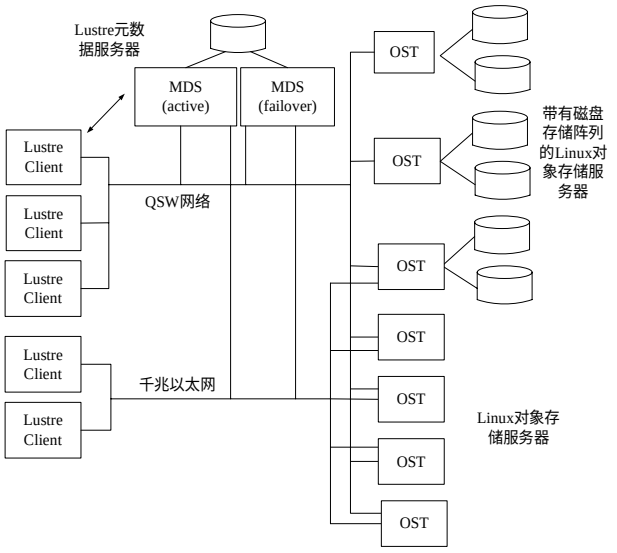

图 1 Lustre 基本结构

Lustre 是一个透明的全局文件系统，客户端可以透明地访问集群文件系统中的数据，而无须知道这些数据的实际存储位置。客户端通过网络读取服务器上的数据，存储服务器负责实际文件系统的读/写操作以及存储设备的连接，元数据服务器负责文件系统目录结构、文件权限和文件的扩展属性以及维护整个文件系统的数据一致性和响应客户端的请求。

Lustre 把文件当作由元数据服务器定位的对象，元数据服务器指导实际的文件 I/O 请求到存储服务器，存储服务器管理在基于对象的磁盘组上的物理存储。高速网络和硬盘技术的发展为集群文件系统的扩展提供了技术保证。

由于采用元数据和存储数据相分离的技术，因此其可以充分分离计算和存储资源，使得客户端计算机可以专注于用户和应用程序的请求；存储服务器和元数据服务器专注于读数据、传输数据和写数据。存储服务器端的数据备份和存储配置以及存储服务器扩充等操作不会影响到客户端，存储服务器和元数据服务器均不会成为性能的瓶颈[2]。

Lustre的全局命名空间为文件系统的所有客户端提供了一个有效的全局唯一的目录树，并将数据条块化，再把数据分配到各个存储服务器上，提供了比传统SAN的“块共享”更为灵活的共享访问方式[4]。全局目录树消除了在客户端的配置信息，并且在配置信息更新时仍然保持有效。

2 测试与分析

2.1 测试环境

测试中软件和硬件配置信息如表 1、表 2 所示。

表 1 软件配置

Lustre 版本	lustre-1.4.5
操作系统	Red Hat Enterprise Linux AS3 2.4.21-4
测试工具	bonnie++-1.03a

表 2 硬件配置

node1 /node2	服务器名称	浪潮英信 NF130
	CPU	Intel Pentium 4 2.66GHz
	内存容量	512MB
	硬盘类型	IDE
node3	服务器名称	浪潮英信 NF180
	CPU	Intel Xeon MP 2.4GHz
	内存容量	512MB
	硬盘类型	Ultra 320 SCSI
网络	千兆以太网	

其中，测试工具 bonnie++-1.03a 是针对“文件并发访问数量较少、单个文件相对较大”的应用而设计的。

2.2 块读/写带宽测试

测试中始终使用 node1 作为实际的客户端，node2 作为读/写的存储设备，并采用单个客户端上运行单个进程的方式来进行测试。如图 2 所示，测试中分别部署了 Lustre 文件系统、NFS 网络文件系统、NFS over Lustre 的测试环境。

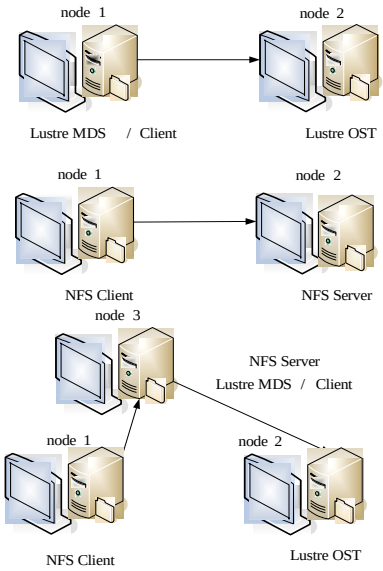

图 2 测试部署

为了消除文件系统缓存策略对系统性能的影响，每组块读/写测试中每个客户端读写文件的长度都均大于节点内存大小的 2 倍，即 2 048MB、4 096MB、8 192MB。在硬盘的基准性能测试中使用 Ext3 本地文件系统，得到的结果的平均值为：块写带宽为 43 705KBps，块读带宽为 47 884KBps。

由于千兆位网络有 100MBps以上的通信带宽，而本地硬盘仅能提供 43MBps～47MBps的数据传输率，因此可以认为，所测得数据的差异能真实反映文件系统的性能[5]。测试得读/写带宽结果如图 3、图 4 所示，横轴坐标表示测试文件的大小，纵轴坐标表示读/写带宽(KBps)。

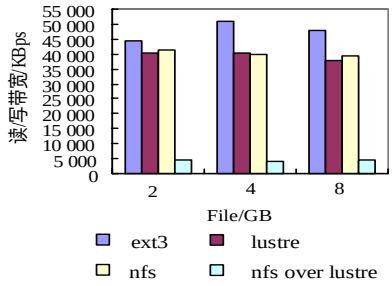

图 3 块方式读带宽

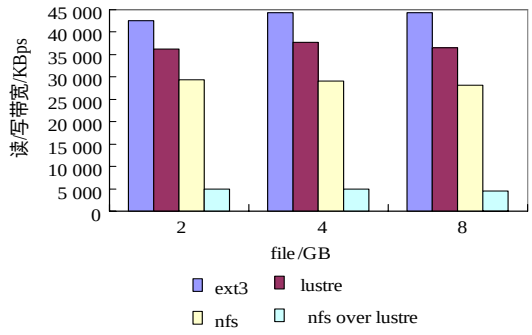

图 4 块方式写带宽

从实验结果中可以看出 NFS over Lustre 带来了很严重的性能损耗，比 Lustre 和 NFS 网络文件系统的读/写带宽降低了很多，而且在多客户、多进程的情况下这种现象更为明显。分析其原因，造成这种现象的本质原因是：Lustre 文件系统和 NFS 网络文件系统的固有机制不同。

Lustre 文件系统是一种对象存储文件系统，虽然从计算节点的角度来看它对存储设备的访问和 NFS 一样都是以文件为单位进行的，但实际上 Lustre 文件系统的访问过程是对块设备进行操作的，当将 NFS 挂载在 Lustre 文件系统之上时，就带来了对文件重新整合并进行存储转发的过程，这个过程经历了 2 次文件系统的操作，与此同时也带来了很大的开销。为了掌握整个过程中报文的存储转发过程，测试中使用了集线器和监听工具 Wild Packets Ether Peek NX，在 NFS over Lustre 测试环境下、小文件读/写过程中 TCP/IP 协议族的报文进行了全程监听。TCP/IP 协议族的报文为分析提供了依据。在监听的过程中 NFS 客户端所访问的文件数据包是在 Lustre 客户端中被整合后，在经 NFS 服务器将其转发到外部的。在这个过程中，文件被再次分包并通过 NFS 网络文件传输协议经 TCP/IP 协议族的再次封装后传输到客户端。这个存储转发的过程正是性能衰减的所在。

2.3 改进方案

分层存储结构如图 5 所示(1.NAS Gateway；2.Lustre Gateway；3.Lutre with no Gateway)。

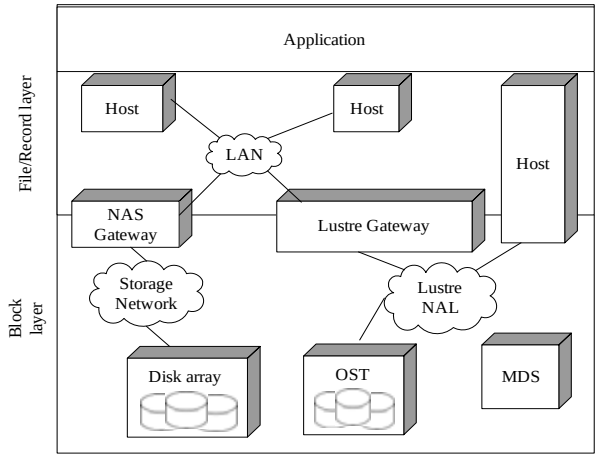

图 5 分层存储结构

在这种传输过程中，由于不同文件系统操作带来的性能衰减，因此本文提出可以借鉴目前国际上流行的 NAS 网关的原理在专用的 Lustre 文件系统上形成一个网关，使其可以直接对内连网上的对象存储设备进行快速访问。网关将 IP 网络连接到高速内联网上，架起了一座连接外部网络与内部 Lustre 世界的桥梁。

图 5 采用了分层的存储结构给出了 NAS 网关，Lustre 网关和没有网关的专用 Lustre 文件系统的存储结构。

NAS 网关的提出是为了有效地结合 NAS 和 SAN 的优点，提出一种比较经济的方式来使用网络中的存储设备，并消除“存储孤岛”的现象，而本文提出的 Lustre 网关是为了使部署了 Lustre 文件系统的高性能计算机能够更好地提供网络化服务，两者虽然目的不同，但要求实现的基本功能都是 block-level 数据到 file-level 数据的转换。其网络结构见图 6。

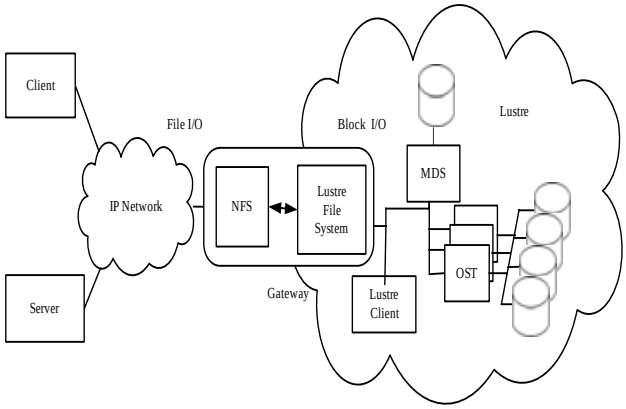

图 6 网络结构

Lustre 网关的工作原理是：网关将 IP 网络与存储区域的内联网连接在一起，使用户可以通过访问网关来访问内联网上对象存储设备中的文件，同时要求网关操作的整个流程对用户透明。其操作流程如下：

(1)外部客户端通过网关请求文件数据；

(2)网关转换客户机请求，并经 Lustre 文件系统通过元数据服务器 MDS 向对象存储设备发出请求；

(3)对象存储设备取出块数据，并将数据返回给网关；

(4)网关将块数据转换为文件数据在转发给外部客户；

(5)外部客户通过网关得到请求文件。

本文也可以通过在网关处为元数据信息建立大容量的高速缓存，并通过采用快速有效的预取策略来实现对存储设备的快速访问，从而提高文件系统传输带宽。

3 结束语

本文分别对 NFS over Lustre、Lustre 文件系统、NFS 网络文件系统及本地磁盘 Ext3 文件系统进行了测试，并将测试结果进行比较分析，给出了性能差异的原因，同时提出相应的解决方案。在今后的工作中，将进一步分析和实现所提出的设想。

参考文献

1 SNIA/T10. Draft OSD Standard[EB/OL]. ftp://ftp.t10.org/t10/drafts/osd/.

2 Braam P J. The Luster Storage Architecture[EB/OL]. 2004. www.lustre.org/docs/lustre.pdf.

3 Schwan P. Lustre: Building a File System for 1 000 Node Clus-Ters[EB/OL]. 2004. http://www.lustre.org /docs/ols2003.pdf.2004.

4 Braam P J, Callahan M J. Lustre: A SAN File System for Linux.pdf[EB/OL]. www.lustre.org/docs/luswhite.pdf.

5 曹立强, 熊 劲. Global File System性能评测与分析[J]. 高性能计算技术, 2003, (6): 2-3.