表目录

第一章 引言

随着服务器技术、存储技术和网络技术的发展，网络存储逐渐成为网络服务器系统的主要存储架构。网络存储系统通过集中管理存储资源，为应用提供有效的存储资源共享。大型分布式文件系统，是解决大规模集群亟需的海量集中存储、文件级共享、高并发带宽、高可扩展性的关键技术。

文件系统元数据是描述文件系统和文件的数据，文件访问包括文件元数据和数据的访问。应用规模的扩大和新应用的出现，对分布式文件系统元数据服务的性能和扩展能力提出更高的要求。在新的存储系统架构下，提供高性能、可扩展、动态平衡负载、高可用的元数据访问服务，是大型分布式文件系统研究的热点问题。

本论文主要研究大型分布式文件系统中元数据服务的可扩展性问题。期望通过深入探讨影响分布式文件系统元数据服务扩展能力的关键问题，解决分布式文件系统元数据服务的可扩展性问题，满足应用动态变化的需求，推进网络存储技术的发展。

1.1 研究的目标和意义

文件系统为应用提供数据存储服务，支持应用间的文件级数据共享。通用的 UNIX® 文件系统模型[Bovet2002][Daley1965]包括描述文件系统属性的超级块、描述文件属性的索引节点、存放目录项的目录数据块、存放数据对应的设备块号等信息的元数据块、以及存放文件数据的数据块等，除文件数据外的所有信息统称为“文件系统元数据”，简称“元数据(metadata)”。文件系统通过称为“目录”的特殊文件，组织成倒置的树形结构。从根目录出发，通过目录树遍历，应用可以定位需要访问的文件。应用能够访问到的文件名字构成文件系统名字空间（file system namespace）。

文件访问包括文件的元数据和数据访问两个步骤。在获得文件的元数据后，才能根据元数据定位文件数据的存储位置，进行数据访问。所以，文件访问时间由元数据访问时间和数据访问时间组成，即：

$$\begin{aligned} &Time(file) = Time(metadata) + Time(data), \\ &Ratio(metadata) = Time(metadata) / Time(file) \end{aligned}$$

在大文件应用中，由于一次元数据访问可以支持大量的数据访问，Ratio(metadata) 很小，元数据的访问效率对请求的影响不很明显。但是，随着系统规模的不断扩大，大文件应用的聚合元数据请求数目也将非常可观，要求有效的元数据服务。比如，ASCI[Lustre-SGSRFP2001]要求文件系统能够支持 $1.8*10^7$ 个目录、$4.5*10^8$ 到 $1.0*10^{10}$ 个文件的文件系统规模，这将导致规模非常庞大的文件系统元数据请求。

同时，在企业应用环境中，web 服务、E-Mail、在线事务处理等不断涌现的新的应用需要通过分布式文件系统管理数据的存储和共享。这些应用的主要特征是数量庞大的小文件[Williams2005]，Ratio(metadata)将变大。并且，这些应用的规模扩展同样将增加元数据服务的压力。

所以，深入探讨分布式文件系统元数据服务的关键问题，满足已有应用不断增强的扩展需求，并支持特征多样的新应用，推动网络存储技术在更宽广领域的发展，具有非常积极的理论和实际意义。

本研究主要面向局域环境的网络存储系统，目标是通过集群方式的系统服务器架构，管理元数据的存储和访问，满足应用动态变化的分布式文件系统元数据服务需求。本研究主要关注提供分布式文件系统元数据服务的服务器部分的可扩展性，不包括客户端部分相关技术的研究。本研究的分布式文件系统元数据服务的扩展性包括：

1. 支持客户端数目增加带来的系统负载的增加。系统能够支持客户端聚合元数据请求吞吐率随客户端数目的扩展。
2. 支持在既定客户端数目的前提下，客户端请求的变化带来的系统负载的动态变化。系统需要支持客户端负载变化导致的聚合元数据请求吞吐率的扩展要求。

总之，分布式文件系统需要提供可透明扩展的元数据服务，满足系统服务器规模的动态扩展和应用的元数据服务需求的动态变化，支持数据的有效存储和共享。

1.2 相关技术的发展

1.2.1 存储设备

存储设备主要包括光盘、闪存、磁带和磁盘等[Hennessy2002]。光盘主要用于移动存储。由于其相对磁介质具有更长的使用寿命，可以用来更长时间的保存归档数据。但光盘容量有限，访问速度较慢，不适合用作在线存储设备。闪存是消费电子的移动存储，其容量很小。磁带作为磁盘的后端支持设备，主要用作归档数据的存储。磁盘是主要的在线存储设备，用来存储活跃的生产数据。磁盘技术发展非常迅速，根据Seagate®[Kryder2006]的数据，2006 年 3.5 英寸硬盘的存储能力达到 500GB，带宽达到 1,000Mb/s，读寻道时间为 8 毫秒。预计到 2009 年，3.5 英寸硬盘将拥有 2,000GB 的存储能力，2,000Mb/s 的访问带宽和 7.2 毫秒的读寻道时间。到 2013 年，这些指标分别为 8,000GB，5,000Mb/s 和 6.5 毫秒。磁盘性能的增强将弱化磁带在存储系统中的作用，将取代磁带成为归档数据的主要存储介质。

1.2.2 存储系统

从存储系统结构来看，服务器的存储系统经历了直连存储（Direct-Attached Storage, DAS），附网存储(Network-Attached Storage, NAS)和存储区域网络(Storage Area Network,

SAN)等三个阶段的发展[Morris2003]。如图 1.1 所示。

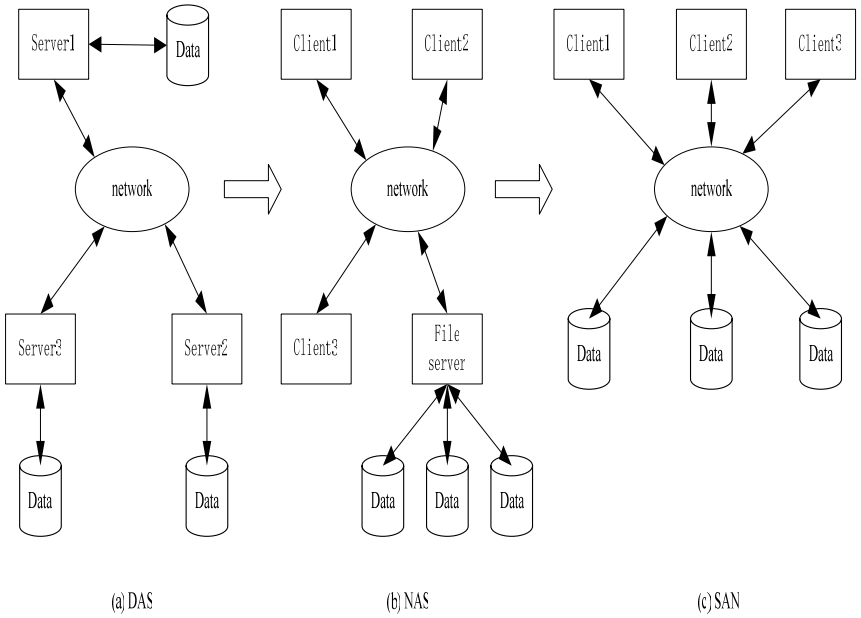

图 1.1 存储系统演进图

在 DAS 系统中，各个服务器通过 I/O 总线连接存储设备，并管理存储资源的使用。服务器不能直接使用其他服务器的存储设备，只能通过 FTP、SCP 等完成数据的共享。DAS 具有明显的缺点，包括：1）各个服务器独立管理其存储资源的使用，存储资源共享困难。2）存储设备通过 I/O 总线连接，扩展能力受限于 I/O 总线的能力。存储设备很难独立于服务器扩展，限制系统的扩展能力。3）存储资源扩展要求计算资源的同步扩展，导致系统管理成本的剧增。4）数据共享困难。

NAS 将 DAS 的存储设备集中，由专门的服务器管理，系统区分为客户端和文件服务器。各个客户端通过网络将数据读写请求提交给文件服务器，文件服务器转发客户端的数据请求。相对于 DAS 而言，NAS 能够解决存储和数据的共享问题。但文件服务器管理存储资源和转发数据读写请求的文件访问方式，将限制存储资源和数据读写性能的扩展。

SAN 使用高速网络直接连接存储设备，并通过虚拟化存储技术集中管理存储资源。客户端直接与存储设备读写数据，提高系统的存储和数据共享能力。为方便系统的管理，SAN 需要 SAN 文件系统[Menon2003]为数据共享控制提供支持。

1.2.3 分布式文件系统

分布式文件系统的研究开始于 20 世纪 80 年代，其典型代表包括 NFS[Sandberg1985][SUN1989][Callaghan1995][Shepler1999]和 CIFS[Hertel2003]等。从提供文件系统元数据服务的服务器结构看，分布式文件系统主要可以分为对称结构和非对称结构两大类[Welch2004]。