- 对称结构的分布式文件系统

在对称结构的分布式文件系统中，没有专门的服务器提供文件系统元数据服务，系统的所有客户端需要协同提供文件系统元数据服务。这类分布式文件系统主要有 DEC 的 VAXCluster[Kronenberg1986]、DEC 的 Frangipani[Thekkath1997]、RedHat 的 GFS [RedHat]、IBM 的 GPFS[Schmuck2002]、Veritas 的 GFS[Veritas]、Polyserve 的 Matrix [Polyserve]、Microsoft 的 FARSITE[Adya2002]和国家智能中心的 COSMOS[Shi2001] [Du2001]等。

对称结构的分布式文件系统要求所有客户端协同提供文件系统元数据服务。它隐含要求：1）所有的客户端同等对待。不论客户端是否需要访问文件系统，都需要参加文件系统的管理工作，在一定程度上导致计算资源的浪费；2）系统对单个客户端的依赖。客户端的性能将影响整个文件系统的性能，系统存在性能“短板”；3）数据访问的并发控制需要很多服务器参与，影响面很大；4）客户端间必须相互信赖，存在安全隐患。

- 非对称结构的分布式文件系统

非对称结构的分布式文件系统由专门的元数据服务器（Metadata Server, MS）提供分布式文件系统元数据服务，主要包括 CMU 的 Andrew File System[Morris1986] [Howard1987]、Duke 的 Slice[Anderson2000-1][Anderson2000-2]、IBM 的 StorageTank [Menon2003]、SGI 的 CXFS[Shepard2003]、ClusterFS 的 Lustre[Braam2002]、Panasas 的 ActiveScale[Panasas2003]、UCSC 的 Lazy Hybrid[Brandt2003]、HP 的 DiFFS[Zhang2001] [Karamanolis2001][Muntz2001-1][Muntz2001-2][Muntz2001-3]、国家高性能计算机工程技术研究中心的 BWFS[Yang2005]和国家智能中心的 DCFS2[Fan2004][Xiong2005]等。

系统可以由单个或多个服务器负责提供元数据服务。由于单个服务器有限的处理能力，多个元数据服务器将成为元数据服务的主流系统架构。在多个元数据服务器的集群环境中，根据服务器间的关系，存在 Active-Failover 和 Active-Active 两种结构。在正常情况下，Active-Failover 方式的系统只有 Active 的元数据服务器提供服务。当 Active 的服务器出现故障时，Failover 服务器接替其工作，继续提供服务。Active-Active 方式的系统则是多个服务器对等工作，服务器间能够做到相互冗余，提供高可用的元数据服务。

1.3 本研究的研究背景

本研究以蓝鲸网络存储系统（BlueWhale Networked Storage System）为平台，其结构如图 1.2 所示。通过蓝鲸集群文件系统（BlueWhale File System, BWFS）的协同，蓝鲸网络存储系统为应用提供文件级数据存储和共享。从图 1.2 的系统结构看，BWFS 采用带外数据（out-of-band）传输方式提供数据服务，用户文件访问的控制流和数据流分离。在获得文件的元数据后，用户直接通过网络块设备访问协议，并发地从存储设备读取数据。

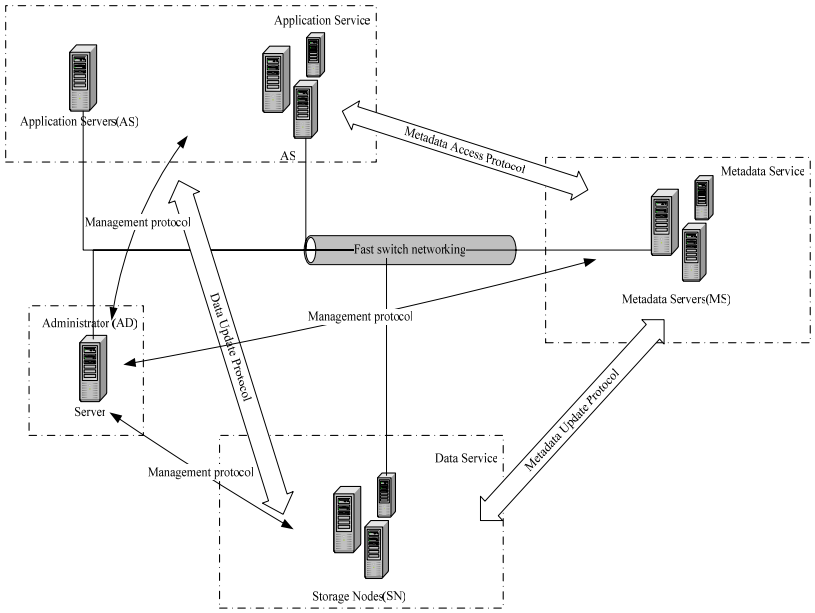

图 1.2 蓝鲸网络存储系统结构

从系统功能划分，系统由五个部分组成。第一部分是通过虚拟化存储技术完成系统存储资源的集中管理，提供存储服务的数据服务(Data Service)部分。它由完成存储资源虚拟化[Glider2003]的虚拟化控制器和多个直接连接到高速网络上的存储设备（storage node，SN）构成。第二个部分是提供文件系统元数据服务的元数据服务（Metadata Service）部分。它主要负责管理虚拟存储子系统提供的逻辑存储资源，并处理分布式文件系统的元数据请求。第三个部分是系统的管理服务器（Administrator，AD）。它通过多种方式监视和控制存储系统的状态，完成系统的单一集中管理功能。最后是应用服务器（Application Server，AS）。它支持多种硬件（x86，MIPS 等）、多种操作系统（Windows，UNIX，Linux 等）、多种应用（E-Mail 服务，文件服务，Web 服务，数据库服务，在线事务处理服务等）构成的异构环境。AS 与元数据服务、数据服务交互，完成文件的访问。这四个部分通过高速交换网络完成互联，构成统一的网络存储系统。

1.4 本研究的关键问题

本研究成果将用于图 1.2 的 BWFS 的元数据服务部分。从逻辑上，分布式文件系统元数据服务由元数据存储服务和元数据请求服务构成。所以，可扩展的分布式文件系统元数据服务需要解决的关键问题包括：

1. 存储资源的组织、管理和使用问题。存储资源的组织、管理和使用是元数据服务的基础问题。存储资源组织对象是存储文件系统元数据和数据的存储资源，其有效性要求能够以较小的代价支持存储资源动态独立地扩展。存储资源管理必需支持存储资源和资源使用者规模的扩展。存储资源的使用方式要求存储资源管理机制能够为资源用户间的共享提供有效支持。

2. 元数据请求的分布管理问题。在满足用户的元数据请求处理效率的前提下，尽可能地平衡各个服务器的负载，避免出现性能瓶颈。
3. 文件系统的一致性问题。元数据请求分布策略可能将更改文件系统请求涉及的多个元数据分布到不同服务器，请求成为跨服务器请求。尽管跨服务器请求的比例非常少[Hendricks2006]，但两阶段提交等传统的请求原子性保证协议，对系统的性能和错误恢复能力的影响都非常大，有必要探讨轻量级协议的可能性。
4. 元数据请求的并发控制问题。在新的存储系统架构下，需要探讨已有的并发控制算法的适应性。分布式文件系统需要代价较小的请求并发控制算法，满足元数据服务扩展的需要。

1.5 本研究的主要贡献

通过分析已有的研究成果，结合用户的元数据访问负载的特点，本研究获得了高扩展能力的存储资源组织、管理和使用，动态高效的元数据请求分布管理，简单有效的文件系统一致性协议，以及高效的元数据请求并发控制算法等研究成果。具体包括：

1. 基于集中虚拟存储模型的存储资源组织、分布式层次化的存储资源管理、完全共享的存储资源使用模式的文件系统元数据存储服务。虚拟存储支持存储资源的独立扩展，层次化的管理机制支持存储资源的有效利用，完全共享的存储资源使用模式解决非常困难的数据放置问题，支持动态灵活的元数据请求服务。在此基础上，元数据的存储和访问彻底分离，元数据的存储不再限制其请求的分布，元数据的存储问题得到有效解决。
2. 动态灵活的元数据请求分布管理机制和策略，支持元数据服务器规模和用户元数据访问负载的动态变化。通过仅管理活跃元数据的请求分布，大幅度缩小请求分布管理的对象集，支持请求分布管理的扩展。根据元数据活跃性管理元数据请求分布，为用户元数据访问负载的动态变化提供有效支持。采用层次化的分布管理机制，支持元数据服务器规模的动态扩展。元数据请求分布策略结合请求的动态性和元数据间客观存在的访问相关性，结合元数据请求的处理效率，尽可能地平衡服务器的负载。
3. 简单高效的更改文件系统的跨服务器请求的原子性保证协议。对称的元数据服务器系统结构、用户元数据访问的统计特征、动态灵活的元数据请求分布管理，为简单高效的元数据请求一致性协议提供足够的支持。通过动态改变活跃元数据在服务器间的分布，完成元数据在服务器间的迁移，将跨服务器请求集中化，利用相关技术保证文件系统一致性，减小协议对系统正常请求处理的影响。
4. 基于系统的请求并发情形分析，提出简单的请求同步控制算法。利用元数据请求分布信息，结合元数据请求语义，充分分析可能的请求并发，根据请求的优

先级进行判定，尽可能地并发元数据请求。

5. 根据上述结果，设计和实现了 1 个应用到 BWFS 的原型系统（蓝鲸多元数据服务器系统，BWMMS），并通过实验验证其扩展能力的有效性。

相对于传统分布式文件系统，BWMMS 的元数据服务基于集中共享的虚拟存储模型，通过动态重分布方式管理活跃元数据的请求分布，能够很好地满足元数据存储和请求服务的扩展需求，支持应用请求负载的动态变化。针对仅约为 0.006%的跨服务器元数据请求[Hendricks2006]的原子性保证，BWMMS 通过动态元数据分布管理机制的支持，改变元数据请求的分布，将元数据集中到单个服务器处理。动态迁移协议不仅在协议延迟上优于传统的协议，同时也在系统的错误恢复上占有一定的优势。

1.6 本文的组织

本章是第一章，引言部分。介绍本研究的目的和意义，相关技术的发展，研究背景和主要贡献，以及本文的组织等。

元数据服务包括元数据存储和元数据访问两个逻辑层次。第二章探讨元数据存储的关键问题，包括存储资源的组织、管理和使用。其目标是解决存储资源管理的可扩展性问题，并为元数据请求服务提供必要的支持。通过集中虚拟存储模型组织系统存储资源，层次化机制管理存储资源的使用，元数据服务器间完全共享的元数据访问方式，元数据的存储问题得到解决，为灵活的元数据请求服务提供支持。

第三章探讨有效的元数据请求分布管理问题。由于用户元数据访问表现出活跃性、动态性和局部性等特点，BWMMS 采用“后端集中决策机制”，结合服务器的负载、元数据的活跃性、文件系统的名字空间结构，动态管理元数据请求的分布。元数据访问的局部性决定元数据分布信息缓存的有效性。第四章探讨元数据分布信息缓存管理的关键问题。

更改文件系统的跨服务器请求的原子性保证是文件系统一致性的关键问题。由于传统的原子性保证协议在性能和错误恢复等方面存在不足，第五章探讨轻量级原子性保证协议的可能。通过元数据的迁移，分布式元数据请求集中到单个服务器处理，借助本地文件系统技术保证文件系统的一致性。动态迁移协议的效率和容错能力同样得到分析。

元数据迁移加剧请求同步问题。第六章介绍 BWMMS 通过分离元数据分布信息管理和文件系统元数据管理的同步机制，根据元数据访问协议分析可能的并发情形，通过判定请求的优先级，尽可能地并发元数据请求。

第七章介绍了系统的原型实现及评估。第八章是本研究工作的总结和下一步的研究建议。本文最后是参考文献、致谢和作者简历。