第二章 文件系统元数据存储服务

逻辑上，文件系统元数据服务由元数据存储和元数据访问构成。元数据存储服务是元数据请求服务的基础。如何合理有效地组织、管理和使用系统的物理存储资源是元数据存储服务的关键问题。本章讨论 BWMMS 的基于集中虚拟化存储技术的物理存储资源组织、分布式层次化的存储资源管理和完全共享的存储资源使用的元数据存储服务，有效解决存储资源的管理问题，为元数据请求管理提供坚实的基础。

2.1 元数据存储服务需求

从系统的逻辑结构来看，文件系统元数据服务包括提供文件系统元数据存储的元数据存储服务、提供文件系统元数据访问的元数据请求服务。元数据请求服务以元数据存储服务为基础，通过元数据存储服务完成元数据的存储和访问管理。文件系统元数据存储服务是文件系统元数据服务的基础问题。

存储资源的组织结构为存储资源的有效管理提供支持。只有通过合理的存储资源组织，才能有效地管理和使用系统的存储资源。在大规模系统环境中，存储资源管理和使用的参与者规模非常庞大，需要通过有效的存储资源管理机制管理，避免出现限制系统扩展的瓶颈。存储资源的使用模式是应用有效共享文件系统元数据的保证。只有通过有效的存储资源使用模式支持，存储资源用户间才能以较低的代价实现元数据的有效共享，提高元数据共享的效率。

所以，如何有效地组织异构的存储资源，加以有效的管理和使用，提供具有较强扩展能力的存储服务是文件系统元数据存储服务的关键问题。只有有效解决存储服务的关键问题，为文件系统元数据服务提供具有较强扩展能力的元数据存储服务，为有效解决元数据请求服务的关键问题提供基础，才能有效地提供文件系统元数据服务，提高文件系统元数据服务的扩展能力。

2.2 相关研究概况

已有研究在存储资源的组织、管理和使用方面进行了大量的研究。物理存储资源的组织结构主要存在分散和集中两种，如图 2.1 所示。

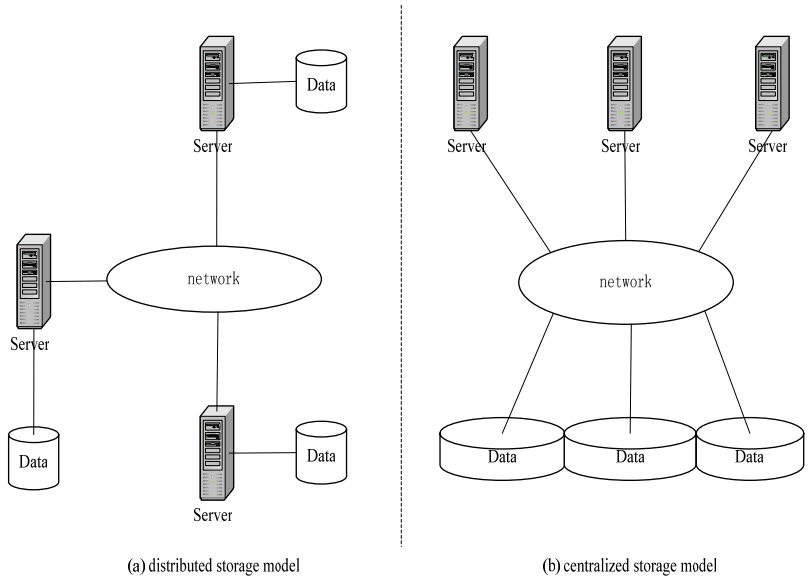

图 2.1 两种主要的存储资源组织结构

在分散的物理存储资源组织结构[Kronenberg1986][Xiong2005-1]中，存储设备和服务器紧耦合在一起，每个服务器通过其 I/O 总线连接一定数量的存储设备，多个服务器之间通过网络完成互联。这种组织结构的好处是，只要服务器支持某种方式的访问，就能很容易地加入系统，且文件系统元数据请求处理的服务器分配策略非常简单，存储位置决定请求服务器。其主要的问题包括：1）存储设备通过 I/O 总线连接到服务器，单个服务器能够支持的存储资源扩展有限；2）由于服务器和存储设备的紧耦合关系，存储设备和服务器不能彼此独立地扩展，容易导致资源的浪费；3）服务器独立管理存储资源，系统资源不容易得到充分利用；4）数据很难充分共享；5）元数据存储位置决定请求服务器，请求分布缺乏灵活性。

集中的物理存储资源组织结构集中管理系统的存储资源，并统一调度资源的使用，能够充分利用系统的存储资源。DEC Petal[Lee1996]、UCSC 的 Lazy Hybrid[Brandt2003][Weil2004]和清华大学的 Dynamic Hashing[Li2006]等，都采用这种组织结构。

存储资源管理机制将影响存储资源管理的有效性。单一集中的存储资源管理机制[Kronenberg1986][Anderson1995][Callaghan1995][Xiong2005-2]集中管理存储资源的分配和释放。其优点是系统资源管理结构简单，请求处理的路径较短，单个的资源请求的处理效率较高。但是，这种结构不能很好满足大规模系统频繁的资源请求，需要合理分布资源管理功能，提高其扩展能力。

存储资源的使用模式将决定系统的数据共享方式，影响元数据请求管理策略。私有使用模式[Morris1986][Xiong2005-1]将存储资源与存储用户紧耦合，两者间形成从属关系。各个用户独立管理私有的存储资源，请求的重新分布过程，要求通过服务器转发、

或数据在存储资源的迁移等才能完成，请求在服务器间的重分布较难完成。在[Yan2004]的互斥使用模型中，存储资源划分成多个可独立装载(mount)的文件系统。同一时刻，每个文件系统只能被一个用户装载。请求重分布过程要求原用户卸载(umount)文件系统，新用户再重新装载。私有模式和互斥使用模式的服务器形成非对称的服务器结构，请求的分布策略很难灵活支持请求服务器规模的扩展。共享使用模式[Brandt2003][Weil2004][Yang2005][Li2006]通过有效的共享控制机制支持，用户的数据共享不需要数据的迁移、或者文件系统的卸载/装载过程，服务器间形成对称的结构。数据可以放置在任意存储位置，请求可以分配到任何服务器完成处理。

2.3 BWMMS 的元数据存储服务

BWMMS 的元数据存储服务，通过虚拟化存储技术集中组织系统的物理存储资源，存储设备形成对称的结构，元数据的存储位置不再影响其访问请求的分布。存储资源管理功能分散到多个部分，形成层次化的资源管理机制，适应系统的规模扩展。元数据采用 64 位的逻辑元数据资源号标识，逻辑元数据资源动态分配、并动态映射到物理存储资源，有效支持存储资源的扩展。文件系统不存在分割，请求服务器可见任意元数据，服务器间形成对称的结构，为元数据请求在服务器间的任意分布提供基础。

2.3.1 集中虚拟化的存储资源组织

已有的采用集中存储资源组织的系统，通常由各个存储资源用户直接管理物理存储资源。当系统规模非常庞大时，存储资源管理任务非常繁重，用户需要为存储资源管理做出很大的开销[Glider2003]。

为了提高系统的扩展能力、容错能力和可管理能力，增强系统各个部分的模块化，BWMMS 采用虚拟化存储技术[Morris2003][Glider2003]管理系统的物理存储资源。它在系统的存储资源提供者和存储资源使用者之间加入存储资源虚拟化管理层（Storage Virtualization Layer，SVL），物理存储资源管理任务从存储资源使用者剥离，由 SVL 独立完成系统的物理存储资源管理。系统的物理存储资源组织和管理形成独立的存储子系统。除了管理物理存储资源，SVL 还需要建立和管理物理存储资源与逻辑存储资源之间的映射关系。对存储资源用户而言，系统的物理存储资源表现为可以通过某种寻址方式访问的逻辑存储空间。

由于 SVL 集中管理系统的物理存储资源，并且物理存储资源的任何变化将被 SVL 屏蔽，系统的存储资源管理具有较强的扩展能力。并且，存储资源用户的加入和退出不需要与大规模的物理存储资源直接进行交互，用户的规模扩展同样得到支持。

通过集中虚拟化的存储资源组织，BWMMS 面对的存储资源是通过虚拟化存储技术提供的逻辑存储空间，它采用 64 位的逻辑存储资源号标识，形成线性存储地址空间。

2.3.2 分布式层次化的存储资源管理

存储资源管理机制是存储服务的重要内容。集中的存储资源管理容易导致系统瓶颈出现，不能满足大规模系统的存储资源管理需要。层次化管理结构在存储资源的提供者和使用者间形成明显的层次关系[Huang2005]，将资源管理功能分散到各个层次，消除可能的瓶颈限制。

BWMMS 通过层次化机制管理元数据存储资源，如图 2.2 所示。为避免瓶颈限制，逻辑资源管理功能分散，层次之间通过批量方式管理资源的分配和释放，层次间有效的缓存机制，降低层次化导致的资源请求时间延迟变长问题。

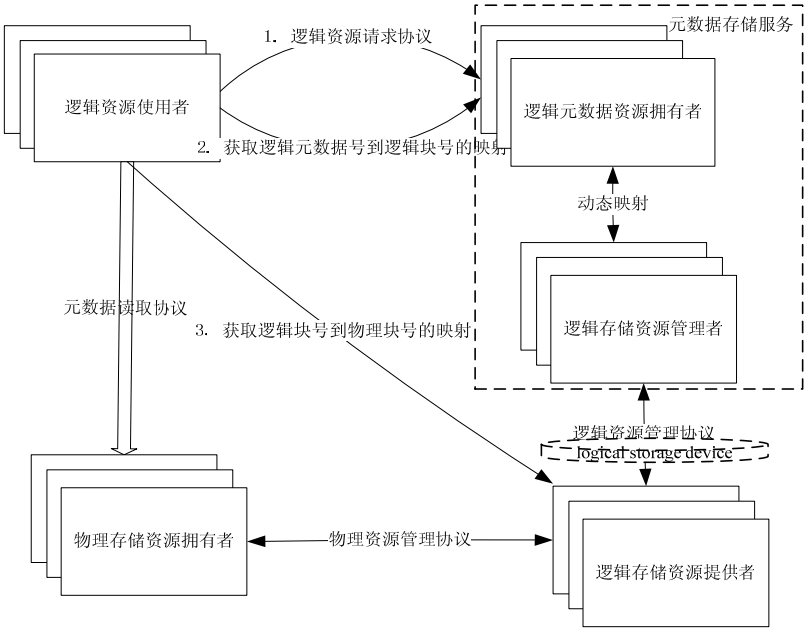

图 2.2 分布式层次化的存储资源管理

元数据存储服务需要管理的逻辑资源包括索引节点、存放元数据的间接块、目录的数据块等，逻辑元数据资源采用 64 位的逻辑元数据资源号标识，支持文件系统规模的扩展。文件系统逻辑元数据资源访问需要经过“逻辑元数据资源〈——〉逻辑存储资源〈——〉物理存储资源”的映射过程。“逻辑存储资源〈——〉物理存储资源”由存储虚拟化层完成，支持物理存储资源的扩展、逻辑存储资源与物理存储资源的动态映射等。为使用扩展的逻辑存储资源，BWMMS 通过动态分配和动态映射方式完成“逻辑元数据资源〈——〉逻辑存储资源”的映射。

逻辑存储资源管理者以批量方式从逻辑资源提供者获取可用的逻辑存储资源信息。元数据的动态分配由逻辑资源使用者驱动，逻辑资源使用者通过元数据访问协议，驱动逻辑元数据资源拥有者分配元数据。逻辑元数据资源拥有者从 64 位线性的逻辑元数据资源进行分配，并动态建立分配的逻辑元数据资源与逻辑存储资源的映射关系。元数据访问通过动态映射完成物理存储资源的定位。逻辑资源使用者首先从逻辑元数据资源拥有

者获得动态建立的逻辑元数据资源与逻辑存储资源的映射关系，然后从逻辑资源提供者获得逻辑存储资源与物理存储资源的映射关系，最后访问物理资源提供者。

逻辑元数据资源的释放同样由逻辑资源使用者动态驱动。逻辑元数据资源拥有者解除逻辑元数据资源和逻辑存储资源的映射，记录可用逻辑元数据资源，释放可用逻辑存储资源给逻辑存储资源管理者，逻辑存储资源管理者以主动或者被动的方式将可用逻辑存储资源释放给逻辑存储资源提供者。

为平衡多个逻辑元数据资源管理者间可用的逻辑元数据资源，逻辑元数据资源拥有者之间通过主动或者被动方式，交换可用逻辑元数据资源信息。所有的逻辑元数据资源拥有者的可用逻辑元数据资源交集为空。逻辑元数据资源可以从任意逻辑元数据资源拥有者分配，在任意逻辑元数据资源拥有者释放，资源管理不会出现问题。逻辑存储资源具有同样的性质。这为元数据请求的灵活分布提供基础。

2.3.3 完全共享的存储资源使用

存储资源使用模式指的是用户对存放在存储资源的数据的共享方式。私有存储将数据与请求服务器紧耦合，数据的访问请求不能灵活地分布到请求服务器。它通常只能通过显式的方式完成数据的共享，存储资源很难共享。已有的某些采用集中方式组织物理存储资源的系统，存在与私有存储相似的逻辑结构。它将逻辑存储空间划分成子空间，每个子空间对应可以独立装载的文件系统。同一时刻，一个子文件系统只能被一个用户装载，子文件系统的使用是互斥的。当用户需要访问其他子文件系统时，它首先要求装载该文件系统的用户卸载，然后自己装载该文件系统，进行请求的处理。其本质依然是通过数据的存储位置决定处理请求的服务器，同样不能很好支持请求服务器规模的扩展。

BWMMS 采用单一的逻辑元数据资源空间，由单一文件系统组织和管理统一的线性逻辑元数据资源空间。出于支持不同特征应用的需要，线性元数据资源空间可以划分成不同使用特性的区域。但是，与私有存储结构和集中互斥存储架构的存储资源使用模式的本质不同是，它没有任何请求分布限制。在逻辑上，存储资源不与请求服务器关联，所有请求服务器装载同一个文件系统。在文件系统的并发控制下，每个请求服务器可见、并能访问文件系统任意的逻辑元数据资源，用户对存储资源的访问不需要请求其他用户授予权限，这种系统称为“集中共享存储架构”。

综合文件系统元数据存储服务的关键问题，BWMMS 采用基于虚拟存储技术的集中化存储资源组织、分布式层次化的存储资源管理和完全共享的存储资源使用，元数据服务器形成对称的服务器结构，如图 2.3 所示。