(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2008 年5 月2 日 (02.05.2008)

PCT

(10) 国际公布号
WO 2008/049353 A1

(51) 国际专利分类号:
H04L 12/56 (2006.01)

(21) 国际申请号: PCT/CN2007/070564

(22) 国际申请日: 2007 年8 月27 日 (27.08.2007)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
200610150325.3
2006 年10 月26 日 (26.10.2006) CN

(71) 申请人 (对除美国外的所有指定国): **阿里巴巴公司(ALIBABA.COM CORPORATION)** [KY/KY]; 开曼群岛英属开曼群岛大开曼岛资本大厦一座四层847号邮箱, Cayman Islands (KY)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **杨金生(YANG, Jinsheng)** [CN/CN]; 中国浙江省杭州市华星路99号东部软件园创业大厦6层, Zhejiang 310099 (CN)。 **汤峥嵘(TANG, Zhengrong)** [CN/CN]; 中国浙江省杭州市华星路99号东部软件园创业大厦6层, Zhejiang 310099 (CN)。 **潘磊(PAN, Lei)** [CN/CN]; 中国浙江省杭州市华星路99号东部软件园创业大厦6层, Zhejiang 310099 (CN)。

(74) 代理人: 北京集佳知识产权代理有限公司(UNITALEN ATTORNEYS AT LAW); 中国北京市朝阳区建国门外大街22号赛特广场7层, Beijing 100004 (CN)。

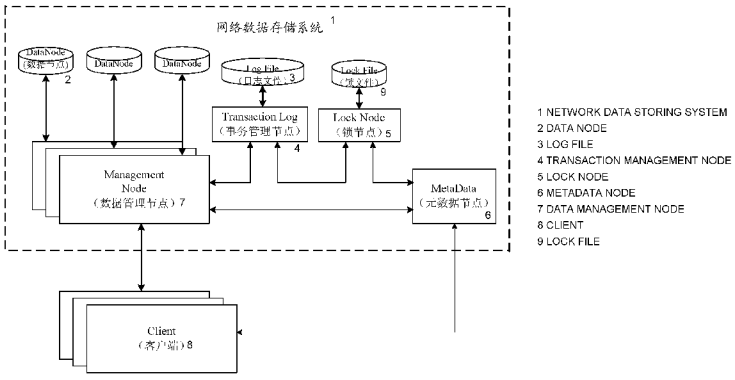
(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告。

(54) **Title:** NETWORK DATA STORING SYSTEM AND DATA ACCESSING METHOD THEREOF

(54) **发明名称:** 网络数据存储系统及其数据访问方法



(57) **Abstract:** A network data storing system and a network data accessing method. The network data storing system comprises: a data node, for storing the data unit; a metadata node, for storing and managing the route information, and providing the route information for the client based on its data processing request; a data management node, for processing the requested data unit in the data node, based on the data accessing request from the client.

[见续页]

WO 2008/049353 A1

(57) 摘要：

一种网络数据存储系统以及网络数据访问方法。所述网络数据存储系统包括：数据节点，用于存储数据单元；元数据节点，用于存储和管理路由信息，并根据客户端数据处理请求向其提供路由信息；数据管理节点，用于根据来自客户端的数据访问请求，处理请求的所述数据节点中的所述数据单元。

网络数据存储系统及其数据访问方法

本申请要求于 2006 年 10 月 26 日提交中国专利局、申请号为 200610150325.3、发明名称为“网络数据存储系统及其数据访问方法”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本发明涉及数据存储与管理领域，尤其涉及一种网络数据存储系统，以及基于该系统的网络数据访问方法。

背景技术

IT 技术的发展历程经历了以计算技术为中心，以处理器的发展为核心动力的过程，发展到以传输技术为中心，并由此促进了计算机网络的发展和普及，使得越来越多的企业信息活动转变为数字形式，从而导致数字化信息爆炸性增长，引发了对 IT 技术的存储技术需求大量增长。数据存储的应用呈现出以下新的特点：

（1）数据成为最宝贵的财富。数据丢失对于企业来讲，损失将是无法估量的，甚至是毁灭性的。

（2）数据总量呈爆炸性的增长。

（3）全天候服务成为大势所趋。在电子商务和大部分网络服务应用中，365×24 小时的全天候服务已是大势所趋，这要求现代数据存储系统具备优异的高可用性。

（4）存储管理和维护要求集中化、自动化、智能化。

（5）存储技术要求平台独立。

传统的存储系统采用的是 DAS（Direct Attached Storage，直接附加存储），即直连方式存储，也可称为 SAS（Server-Attached Storage，服务器附加存储）。在这种方式中，存储设备通过电缆（通常是 SCSI 接口电缆）直接连接到服务器，I/O（输入/输出）请求直接发送到存储设备。这种存储方式依赖于服务器，存储设备本身只是硬件的堆叠，不带有任何存储操作系统。由于受服务器总线技术的限制，DAS 方式系统的可扩展性差，当客户连接数增多时，服务器将成为整个系统的性能瓶颈，这是因为：

（1）主机的带宽限制：虽然计算机技术的发展使主机总线带宽大为增加，

但是仍赶不上现代存储应用对带宽的要求。

（2）主机的内存容量限制：由于主机的内存容量有限，当有连续的大量数据访问请求时，主机的内存容量将很快达到饱和，而不能处理剩下的数据传输请求。

（3）文件系统的管理开销也会增加数据访问时间。

目前大量的企业应用对数据库技术产生的很强的依赖，采用数据库中央服务器集中存储数据，往往成为系统的单点和性能的瓶颈，扩展也很困难，而且成本高，对于海量数据高并发的在线处理尤其困难。因此，传统的集中式的数据存储和管理方式已经无法满足信息日益快速增长的需求。

发明内容

本发明提供一种网络数据存储系统，用以解决现有技术中存在的网络数据存储系统扩展能力差、扩展成本高，以及网络数据访问性能差的问题。

基于相同的技术构思，本发明还提供一种网络数据的访问方法。

本发明提供的网络数据存储系统，包括：

数据节点，用于存储数据单元；

元数据节点，用于存储和管理路由信息，并根据客户端数据处理请求向其提供路由信息；

数据管理节点，用于根据客户端数据访问请求，处理请求的所述数据节点中的所述数据单元。

所述元数据节点、数据管理节点和数据节点以树型连接；

元数据节点为该树型结构的根节点，其下连接有一个或多个数据管理节点；每个数据管理节点下连接有一个或多个数据节点。

根据本发明的上述系统，所述元数据节点存储的路由信息包括：

所述元数据节点到所述数据管理节点的路由信息；

所述数据管理节点到所述数据节点的路由信息。

所述元数据节点还存储有所述数据单元的路由算法；所述路由算法为通过数据单元标识推算存储该数据单元的数据节点标识和该数据单元在所述数据节点中的位置信息的算法。

根据本发明的上述系统，所述数据管理节点中配置有数据访问服务或/和

冗余策略。

根据本发明的上述系统，所述数据节点中存储的所述数据单元为面向业务应用的最小数据集合。

所述数据单元内部由多个文件或/和目录构成。

所述数据单元内部的文件包括数据文件或/和索引文件。

本发明的上述系统，还包括事务管理节点，用于存储日志文件，并提供日志管理服务。

本发明的上述系统，还包括锁节点，用于存储锁文件，并提供锁管理服务。

本发明提供的网络数据访问方法，包括步骤：

客户端向元数据节点发送访问数据单元的请求，并从该元数据节点获取到所述数据管理节点的路由信息；

客户端按照所述路由信息向所述数据管理节点发送访问数据单元的请求；

所述数据管理节点收到请求后，从所述元数据节点获取到存储所述数据单元的数据节点的路由信息，并按照该路由信息和客户端请求的操作对所述数据节点中的所述数据单元进行处理。

根据本发明的上述方法，所述元数据节点接收到客户端发送的访问数据单元的请求后，向该客户端提供所述数据管理节点的路由信息，包括步骤：

所述元数据节点获取数据访问请求中的数据单元信息，并根据数据单元信息和数据单元标识的映射关系获取到该数据单元的标识；

按照数据单元的路由算法，由该数据单元的标识计算出存储该数据单元的数据节点的标识；

根据数据节点标识与数据管理节点标识的映射关系，获取到对应的数据管理节点标识，并将该数据管理节点标识提供给客户端。

根据本发明的上述方法，所述元数据节点向所述数据管理节点提供路由信息，包括步骤：

所述元数据节点从所述数据管理节点的请求中获取数据单元标识，并按照数据单元的路由算法，由该数据单元标识计算出存储该数据单元的数据节点的标识和该数据单元在数据节点中的位置信息，并提供给所述数据管理节点。

根据本发明的上述方法，所述数据管理节点将存储操作与计算操作分离。

所述存储操作以存储工作队列对应的线程池执行，所述计算操作以计算工作队列对应的线程池执行。

本发明的上述方法中，还包括步骤：

所述数据节点接收到所述数据管理节点的数据操作指令后，按照该操作指令，通过该数据节点本地的文件系统对所述数据单元进行操作。

根据本发明的上述方法，所述数据单元具有唯一标识。

所述数据单元标识由该数据单元所在的数据节点的标识和该数据单元在数据节点中的位置信息经过映射计算得出。

根据本发明的上述方法，所述对数据节点中的所述数据单元进行处理，进一步包括步骤：

将待写块的副本提交到日志文件；

在提交日志文件成功后，将该待写块提交到所述数据节点本地的文件系统；

若提交文件系统成功，则将所述日志文件中的所述块副本丢弃；否则保留该块副本。

当系统从异常状态恢复到正常状态时，按照所述日志文件中保留的块副本记录进行数据恢复。

本发明的上述方法中，还包括步骤：

使用所述数据节点本地的文件锁或/和网络文件系统中的文件锁，对数据单元的访问操作进行锁保护。

本发明有益效果如下：

（1）本发明提供的网络数据存储系统，将数据分布式存储在三层结构的网络节点上，并提供统一的访问管理和路由，从而支持线形扩容和升级，比现有技术增强了扩展能力，降低了扩展成本。

（2）本发明提供的网络数据访问机制，基于上述分布式数据存储系统，采用两级路由算法，使得数据文件的位置对客户端透明；采用三层结构的分布式设计，使位于中间层的数据管理节点分担了数据访问处理操作，因而可通过配置合理的三层结构，提高网络数据访问性能。

（3）本发明还采用日志技术支持事务处理功能，提高了网络数据访问的

一致性、完整性。

（4）本发明还采用锁管理功能，解决了网络文件系统下文件锁失效的问题。

附图说明

图 1 为本发明实施例的网络数据存储系统的结构示意图；

图 2 为本发明实施例的网络数据存储系统的树型结构示意图；

图 3 为本发明实施例的网络数据访问过程的示意图。

具体实施方式

下面结合实施例和附图对本发明进行详细描述。

参见图 1，为本发明实施例的网络数据存储系统的结构示意图，该数据存储系统包括:

DataNode: 数据节点，是网络上的节点，用于存储原始数据和索引。这些原始数据以数据单元的形式存储于 DataNode。

Management Node: 数据管理节点，是网络上的节点，作为中间层提供一些通用的服务，如索引、冗余策略等。Management Node 管理一组相关的 DataNode。

MetaNode: 管理数据节点的名称空间和映射关系的元数据节点，是网络上的节点，用于提供基础的路由信息，主要维护两种路由关系: MetaNode 到 Management Node 的路由（一级路由），以及 Management Node 到 DataNode 的路由（二级路由）。

Transaction Log: 这是基于日志技术的事务管理节点，通常部署在 Management Node，其中存储有日志文件，用于完成数据的事务保护。

Lock Node: 这是全局可见的网络节点，其中存储有以文件形式存在的数据锁，以实现对数据访问进行锁管理。

图 1 所示的网络数据存储系统的体系结构按照树型结构组织，如图 2 所示。

参见图 2，为本发明实施例的网络数据存储系统的树型结构示意图。

如图所示，从逻辑上将网络数据存储系统中的节点划分为三层，从底层到上层分别是 DataNode、Management Node 和 MetaNode。MetaNode 作为根节点，其下有多个 Management Node 作为叶子节点，每个 Management Node 下

面又有多个 DataNode 作为叶子节点。

构建如上实施例所示的网络数据存储系统的步骤包括:

步骤 1、确定数据单元，分配数据单元 ID，并将数据单元分布存储到 DataNode 中。

本实施例中的数据单元是在文件系统层次之上的抽象的数据集合，可根据业务特点和业务需要，将可以被单独管理的最小数据集合定义为数据单元。大部分企业数据的请求和处理都具有明显的局部性特征，如在邮件系统中，分类、检索和收发邮件都是在一个固定的命名空间——邮件帐号内实现，因此，可以将邮件帐号作为数据单元。

数据单元内部可以由多个文件或者目录构成，如数据文件和索引文件以及文件目录，这些文件和目录通过数据单元所在的 DataNode 本地的文件系统进行管理。

数据单元的ID唯一标识了该数据单元。数据单元的ID包含了两部分信息:存储该数据单元的 DataNode 的 ID 和该数据单元在 DataNode 中的具体位置信息。可通过数据单元的路由算法，从数据单元 ID 中计算获取到上述两种信息。因此，数据单元的 ID 隐含了该数据单元和存储该数据单元的 DataNode 的对应关系。

步骤 2、确定路由算法和路由信息，并存储到 MetaNode 中。

MetaNode 维护的路由信息包括: MetaNode 到 Management Node 的路由信息（一级路由信息），以及 Management Node 到 DataNode 的路由信息（二级路由信息）。这两种路由信息通过以下映射关系表和算法实现:

建立数据单元信息（如数据单元名称）与数据单元 ID 的映射关系表、DataNode ID 与 Management Node ID 的映射关系表，设置数据单元的路由算法，即通过该算法，可以从数据单元 ID 提取出存储该数据单元的 DataNode 的 ID 和该数据单元在 DataNode 中的具体位置信息。

一级路由的实现过程为: MetaNode 依次根据数据单元信息与数据单元 ID 的映射关系表，数据单元的路由算法，以及 DataNode ID 与 Management Node ID 的映射关系表，得到从 MetaNode 到 Management Node 的路由。

二级路由的实现过程为: Management Node 向 MetaNode 发送请求，

MetaNode 根据所请求的数据单元的路由算法得到 Management Node 到存储相应数据单元的 DataNode 的路由。

步骤 3、部署 Management Node，包括:

在 Management Node 中配置有数据访问服务，如索引服务，还可以配置冗余策略;

在 Management Node 内部实现采用存储（I/O-bound task）与计算（CPU-bound task）分离的技术，将工作分为两个队列，计算工作队列和存储工作队列分别用两个线程池并行来完成工作，充分利用 CPU 和 I/O 资源。

基于上述实施例所描述的网络数据存储系统，网络数据的访问过程如图 3 所示。

参见图 3，为本发明实施例的网络数据访问过程示意图，包括步骤:

S301、客户端向 MetaNode 发送数据访问请求，告知要访问的数据单元描述信息（如数据单元名称）。

S302、MetaNode 向客户端返回一级路由信息，告知负责管理该数据单元的 Management Node 的位置信息。

MetaNode 从客户端的请求中获取到数据单元描述信息，再根据数据单元描述信息和数据单元 ID 的映射关系表，获取到客户端请求的数据单元 ID; 然后，根据数据单元的路由算法，由该数据单元 ID 计算得到存储该数据单元的 DataNode 的 ID; 再根据 DataNode ID 与 Management Node ID 的映射关系表获得管理该 DataNode 的 Management Node 的 ID; MetaNode 将获取到的 Management Node 的 ID 发送到客户端。

S303、客户端根据一级路由信息找到 Management Node，向其发起数据访问请求。

S304、Management Node 根据客户端的身份和请求的数据单元信息，向 MetaNode 请求该数据单元在网络中的位置分布。

在步骤 S302 中，MetaNode 向客户端返回一级路由信息的同时，还可返回客户端所请求的数据单元的 ID。这样，步骤 S303 中，客户端在发送的数据访问请求中携带需要访问的数据单元的 ID; 在步骤 S304 中，Management Node 向 MetaNode 发送的请求中携带该数据单元 ID。

S305、MetaNode 向 Management Node 返回二级路由信息，告知请求的数据单元在网络中的位置。

MetaNode 从 Management Node 发送的请求中获取到数据单元 ID，并根据数据单元的路由算法，由该数据单元 ID 计算得到存储该数据单元的 DataNode ID 和该数据单元在 DataNode 中的具体位置，并将这些信息返回给 Management Node。

S306、Management Node 根据该位置信息找到存储该数据单元的 DataNode 以及该数据单元在 DataNode 的位置，并按照客户端提出的请求，处理数据单元中的数据。

DataNode 按照 Management Node 的操作指令，通过 DataNode 的本地文件系统对数据单元进行操作。

S307、Management Node 根据需要向客户端返回数据处理结果。

在上述数据访问过程中，Management Node 内部采用存储（I/O-bound task）与计算（CPU-bound task）分离的技术，将工作分为两个队列，计算工作队列和存储工作队列分别用两个线程池并行来完成工作。

在上述数据访问过程中，本发明实施例还采用事务处理机制来保证网络数据访问的可靠性，包括日志机制和锁技术。

文件的很多操作都是非原子行为，尤其是跨越多个文件或者多个节点的过程中，数据的一致性和完整性容易遭到破坏，导致异常情况出现，如系统非正常关闭。本发明实施例借鉴数据库和操作系统的文件系统所提供的日志保护机制，为本发明实施例的非数据库结构的数据存取提供了一种事务保护机制。

当对本发明实施例的数据单元进行访问（如存取操作）时，将待写块的一个副本写入日志文件；当发往日志的 I/O 数据传送完成后（即数据成功提交日志文件），再将该块写入 DataNode 本地的文件系统；当发往文件系统的 I/O 数据传送完成后（即数据成功提交文件系统），将日志文件中的块副本丢弃；若发往文件系统的 I/O 数据传送失败，则日志文件保留该块副本。

当系统发生崩溃或需要重新启动时，系统首先读取日志文件，并按照日志文件中的记录的块副本进行恢复，使系统恢复到发生异常前的正常状态。

为了加强事务隔离性，本发明实施例还提供了锁机制。事务隔离性通常采

用锁住事务所访问资源的方式来保证，同时为了保证文件事务的特性和高并发和高可靠性，锁技术是一个非常有利的工具。

本发明实施例将目前广泛采用的应用于本地硬盘单节点的文件锁 Dotlock 和网络文件系统（Network File System，NFS）文件系统中锁采用的 POSIX-compliant 和 BSD-based system 的 Flock() or Fcntl()技术结合使用，具体做法为：

首先获取 DotLock（这一步通常有可能会被多个节点同时获取成功），成功后在尝试获取 Flock()或者 Fcntl()。这些锁以文件的方式存在一个全局可见的节点上。在系统恢复的时候检测并释放那些悬挂的锁。锁的粒度，可以对某一个数据块，也可以对一个文件、目录，甚至可以对一个 DataNode 加锁。

下面以一个为超大容量的邮件系统添加存储能力的例子进行说明。

第一步：进行数据规划，确定系统管理的最小数据单元。

邮件地址（邮件帐号）通常由用户名和域名两部分组成，中间用@符号分割。可以用邮件帐号作为最小单元，也可以用域名作为数据单元，本实施例选择用邮件帐号作为数据单元。

第二步：确定路由算法和路由表。

规划路由算法的目的是解决如何根据用户提供的邮件帐号找到邮箱内容的存储位置。为了支持系统容量的不断扩容，本实施例采用 32 位的地址空间来路由，这样最大可以支持 IG 用户数量（大约 10 亿）。该 32 位的地址空间称为 RID（Route ID），用于唯一标识邮件帐号。本实施例假设一个 DataNode 的最大容量为支持 1M（2^{20} = 1,048,576）的用户量，因此 DataNode 的地址空间大小为 1M，具体地址通常可以用本机文件系统的目录来表示。本实施例用 RID 的低 20 位来映射具体的目录，称作 DataNode 内部地址 DataID。每个 DataNode 有唯一编号 NSN（Node Sequence Number），用 32 位 RID 的高 12 位表示。即 RID>>20 = NSN，RID 右移 20 位得到 NSN。

MetaNode 中存储的路由表如表 1 和表 2 所示：

表 1：邮件帐号与邮件 RID 的映射关系表

Mail Address	RID
xxx@yyy.com	11033234
aaa@bbb.com	1033134
…	…

表 2：DataNode ID 与 Management Node ID 的映射关系表

Node NSN	Management Node
0,1,2	worker-1
3,4,5	Worker-2
…	…

表 2 表明一个 Management Node 管理三个节点，地址 xxx@yyy.com 请求由标识为 worker-1 的 Management Node 负责处理。

第三步：完成数据规划和路由策略制定后，进行容量规划。

在用户较少的时候，可以部署一台 Management Node 和一个 DataNode（编号为 0，负责管理 0-1M 的用户）。随着用户数量的增长，一个 DataNode 不能满足存储需求的时候，可以再增加一个 DataNode，假设编号为 1，则 RID 的高 12 位为 000000000001，负责管理低 20 位 1M-2M 之间的用户，这样如上所述，随着业务的不断发展，系统可以线性不断扩展，实现海量数据存储。

第四步：部署 MetaNode。

如表 1 和表 2 所示的路由信息表可以存储在 MetaNode 的数据库中也可以以文件的形式保存。由于路由信息表不是很大，服务器启动后，可以将整个路由表放在内存中，这样可以快速响应客户端请求。对于不同的应用可以定制不同的策略，如对于简单的应用，数据规则组织良好，则可以简化 MetaNode 为一个根据应用提供的唯一数据 ID 实现两级 Hash 的算法。

第五步：部署 Management Node 服务，指定建立索引的配置。

在 Management Node 上添加数据搜索功能，对需要存储的数据进行索引，将索引文件和数据文件保存在对应的数据节点上，如果需要对数据进行业务相

关的处理，可以将相关逻辑 job 作为服务部署在服务器上，Management Node 采用存储与计算分离的技术，充分发挥系统的能力。

以后随着业务的发展，系统可以不断的按需扩充容量。随着用户量的增长，可以不断添加 DataNode，每添加 3 个 DataNode，部署一台 Management Node，而对于 MetaNode，通常仅需要一台服务器，为了使之不成为系统的单点，可采用备份机制，增设一台用作备份的 MetaNode 服务器。

上述优选实施例针对的是一个比较简单的邮件系统，但本发明实施例提供的网络数据存储系统的应用场合不限此类应用。本发明实施例提供的网络数据存储系统尤其适用于 B2B 电子商务平台和软件。这类应用通常以企业和用户为中心，存在大量的在线事务处理，因此可以将用户或者一家企业用户作为一个数据单元集合，因为这些数据主要是内部私有，不容许其他用户任意访问。把这些数据作为一个数据单元集合来管理，这就从物理上保证了企业数据独立性和隔离性，不会和其他用户交织在一起，同时支持在线搜索和事务处理等。这种方案相对数据库的好处是非常明显的。数据库不可能为每个企业用户创建一套数据库，数据库通常将所有的企业同种应用的数据放在一张表中，物理上没有实现安全隔离，需要应用处理非法访问等问题，同时如果用户量非常大的时候这种数据库方法将会导致很大的性能问题。

显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

权 利 要 求

1、一种网络数据存储系统，其特征在于，包括:

数据节点，用于存储数据单元;

元数据节点，用于存储和管理路由信息，并根据客户端数据处理请求向其提供路由信息;

数据管理节点，用于根据客户端数据访问请求，处理请求的所述数据节点中的所述数据单元。

2、如权利要求 1 所述的系统，其特征在于，所述元数据节点、数据管理节点和数据节点以树型连接;

元数据节点为该树型结构的根节点，其下连接有一个或多个数据管理节点; 每个数据管理节点下连接有一个或多个数据节点。

3、如权利要求 1 所述的系统，其特征在于，所述元数据节点存储的路由信息包括:

所述元数据节点到所述数据管理节点的路由信息;

所述数据管理节点到所述数据节点的路由信息。

4、如权利要求 3 所述的系统，其特征在于，所述元数据节点还存储有所述数据单元的路由算法; 所述路由算法为通过数据单元标识推算存储该数据单元的数据节点标识和该数据单元在所述数据节点中的位置信息的算法。

5、如权利要求 1 所述的系统，其特征在于，所述数据管理节点中配置有数据访问服务或/和冗余策略。

6、如权利要求 1 所述的系统，其特征在于，所述数据节点中存储的所述数据单元为面向业务应用的最小数据集合。

7、如权利要求 6 所述的系统，其特征在于，所述数据单元内部由多个文件或/和目录构成。

8、如权利要求 7 所述的系统，其特征在于，所述数据单元内部的文件包括数据文件或/和索引文件。

9、如权利要求 1 所述的系统，其特征在于，还包括事务管理节点，用于存储日志文件，并提供日志管理服务。

10、如权利要求 1 或 9 所述的系统，其特征在于，还包括锁节点，用于存

储锁文件，并提供锁管理服务。

11、一种网络数据访问方法，其特征在于，包括以下步骤：

客户端向元数据节点发送访问数据单元的请求，并从该元数据节点获取到所述数据管理节点的路由信息；

客户端按照所述路由信息向所述数据管理节点发送访问数据单元的请求；

所述数据管理节点收到请求后，从所述元数据节点获取到存储所述数据单元的数据节点的路由信息，并按照该路由信息和客户端请求的操作对所述数据节点中的所述数据单元进行处理。

12、如权利要求 11 所述的方法，其特征在于，所述元数据节点接收到客户端发送的访问数据单元的请求后，向该客户端提供所述数据管理节点的路由信息，包括步骤：

所述元数据节点获取数据访问请求中的数据单元信息，并根据数据单元信息和数据单元标识的映射关系获取到该数据单元的标识；

按照数据单元的路由算法，由该数据单元的标识计算出存储该数据单元的数据节点的标识；

根据数据节点标识与数据管理节点标识的映射关系，获取到对应的数据管理节点标识，并将该数据管理节点标识提供给客户端。

13、如权利要求 11 所述的方法，其特征在于，所述元数据节点向所述数据管理节点提供路由信息，包括步骤：

所述元数据节点从所述数据管理节点的请求中获取数据单元标识，并按照数据单元的路由算法，由该数据单元标识计算出存储该数据单元的数据节点的标识和该数据单元在数据节点中的位置信息，并提供给所述数据管理节点。

14、如权利要求 11 所述的方法，其特征在于，所述数据管理节点将存储操作与计算操作分离。

15、如权利要求 14 所述的方法，其特征在于，所述存储操作以存储工作队列对应的线程池执行，所述计算操作以计算工作队列对应的线程池执行。

16、如权利要求 11 所述的方法，其特征在于，还包括步骤：

所述数据节点接收到所述数据管理节点的数据操作指令后，按照该操作指令，通过该数据节点本地的文件系统对所述数据单元进行操作。

17、如权利要求 11-16 任一权项所述的方法，其特征在于，所述数据单元具有唯一标识。

18、如权利要求 17 所述的方法，其特征在于，所述数据单元标识由该数据单元所在的数据节点的标识和该数据单元在数据节点中的位置信息经过映射计算得出。

19、如权利要求 11 所述的方法，其特征在于，所述对数据节点中的所述数据单元进行处理，进一步包括步骤:

将待写块的副本提交到日志文件;

在提交日志文件成功后，将该待写块提交到所述数据节点本地的文件系统;

若提交文件系统成功，则将所述日志文件中的所述块副本丢弃；否则保留该块副本。

20、如权利要求 19 所述的方法，其特征在于，当系统从异常状态恢复到正常状态时，按照所述日志文件中保留的块副本记录进行数据恢复。

21、如权利要求 11 或 19 所述的方法，其特征在于，还包括步骤:

使用所述数据节点本地的文件锁或/和网络文件系统中的文件锁，对数据单元的访问操作进行锁保护。

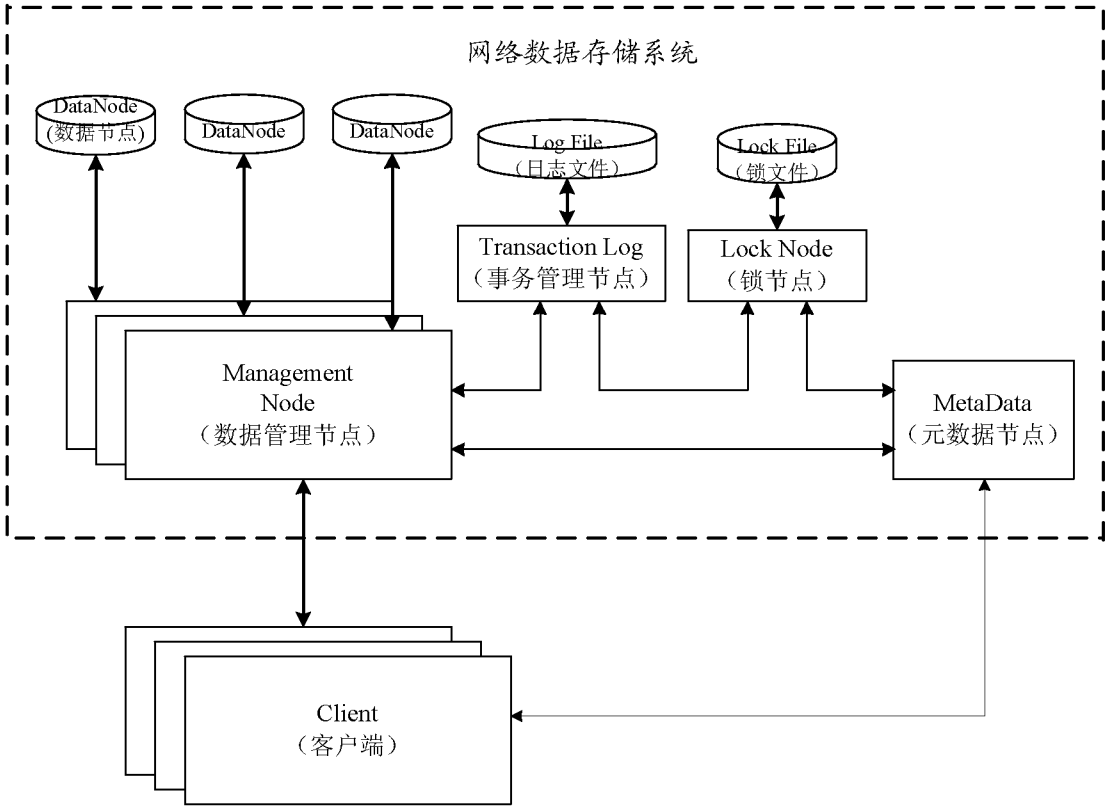

网络数据存储系统
DataNode
（数据节点）
DataNode
DataNode
Log File
（日志文件）
Lock File
（锁文件）
Transaction Log
（事务管理节点）
Lock Node
（锁节点）
Management
Node
（数据管理节点）
MetaData
（元数据节点）
Client
（客户端）

图 1

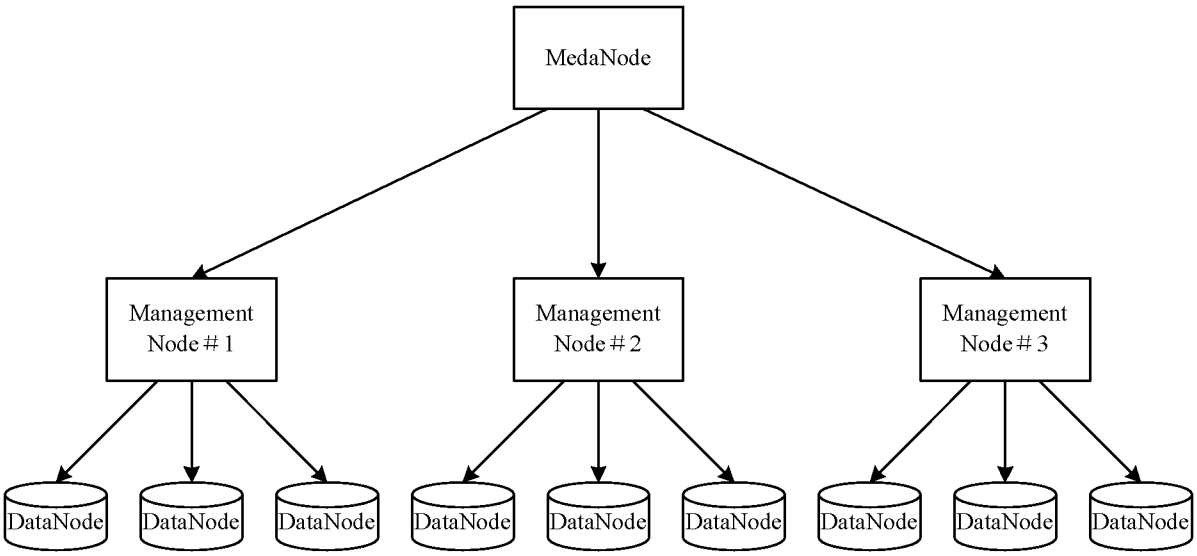

MedaNode
Management
Node＃1
Management
Node＃2
Management
Node＃3
DataNode
DataNode
DataNode
DataNode
DataNode
DataNode
DataNode
DataNode
DataNode

图 2

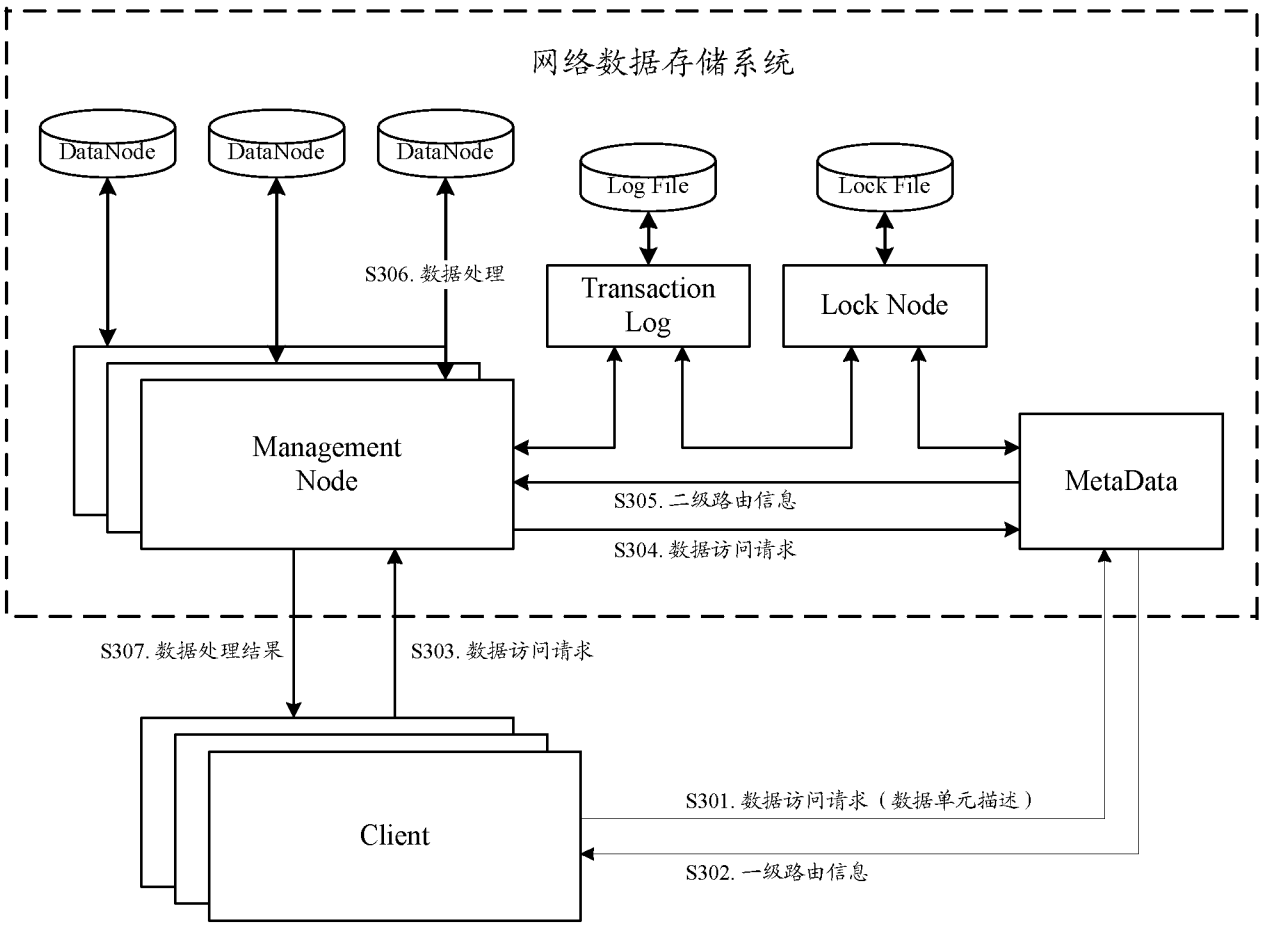

网络数据存储系统
DataNode
DataNode
DataNode
Log File
Lock File
S306. 数据处理
Transaction Log
Lock Node
Management Node
MetaData
S305. 二级路由信息
S304. 数据访问请求
S307. 数据处理结果
S303. 数据访问请求
Client
S301. 数据访问请求（数据单元描述）
S302. 一级路由信息

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2007/070564

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/56(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: H04L 12/-; H04L 29/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

database: WPI, EPODOC, PAJ, CNPAT, CNKI

searched words: data, storing, storage, network, node, route, routing, metadata, client, user, customer, request, unit

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN1463522A (KONINK PHILIPS ELECTRONICS NV) 24 December 2003 (24.12.2003) see claims 1, 2, 12, specification pp. 3, line 4 – pp. 4, line 26, figs. 1, 2	1-21
A	US20030223378A1 (ISHWAR P et al) 04 December 2003 (04.12.2003) see the whole document	1-21
A	CN1333966A (TELEFONAKTIEBOLAGET ERICSSON L M) 30 January 2002 (30.01.2002) see the whole document	1-21

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim (S) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 November 2007 (22.11.2007)	**06 Dec. 2007 (06.12.2007)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer XIE, Zhiyuan Telephone No. (86-10)62085029

Form PCT/ISA/210 (second sheet) (April 2007)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2007/070564

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN1463522A	24.12.2003	WO02093844A2	21.11.2002
		US20020173321A1	21.11.2002
		KR20030020388A	08.03.2003
		EP1393507A2	03.03.2004
		JP2004525587T	19.08.2004
		CN1237762C	18.01.2006
		EP1393507B1	11.10.2006
		DE60215340E	23.11.2006
		DE60215340T2	03.05.2007
		AT342623T	15.11.2006
US20030223378A1	04.12.2003	none	
CN1333966A	30.01.2002	WO0030305A2	25.05.2000
		SE9803869A	13.05.2000
		AU1592200A	05.06.2000
		SE514727C2	09.04.2001
		EP1129549A2	05.09.2001
		TW462167A	01.11.2001
		JP2002530936T	17.09.2002
		US6674757B1	06.01.2004
		CA2351174A1	25.05.2000

国际检索报告

国际申请号
PCT/CN2007/070564

A. 主题的分类

H04L 12/56(2006.01)i

按照国际专利分类表(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: H04L 12/-; H04L 29/-

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

数据库：WPI, EPODOC, PAJ, CNPAT, CNKI

检索词：数据，存储，网络，节点，结点，路由，元数据，客户端，用户，请求，单元

data, storing, storage, network, node, route, routing, metadata, client, user, customer, request, unit

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN1463522A（皇家菲利浦电子有限公司）24. 12 月 2003（24.12.2003） 参见权利要求 1、2、12，说明书第 3 页第 4 行-第 4 页第 26 行，图 1，2	1-21
A	US20030223378A1（ISHWAR P 等）04. 12 月 2003（04.12.2003） 参见全文	1-21
A	CN1333966A（艾利森电话股份有限公司）30. 1 月 2002（30.01.2002） 参见全文	1-21

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
22. 11 月 2007（22.11.2007）	**06.12 月 2007 (06.12.2007)**
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	受权官员 谢志远 电话号码：(86-10) **62085029**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2007/070564

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN1463522A	24.12.2003	WO02093844A2	21.11.2002
		US20020173321A1	21.11.2002
		KR20030020388A	08.03.2003
		EP1393507A2	03.03.2004
		JP2004525587T	19.08.2004
		CN1237762C	18.01.2006
		EP1393507B1	11.10.2006
		DE60215340E	23.11.2006
		DE60215340T2	03.05.2007
		AT342623T	15.11.2006
US20030223378A1	04.12.2003	无	
CN1333966A	30.01.2002	WO0030305A2	25.05.2000
		SE9803869A	13.05.2000
		AU1592200A	05.06.2000
		SE514727C2	09.04.2001
		EP1129549A2	05.09.2001
		TW462167A	01.11.2001
		JP2002530936T	17.09.2002
		US6674757B1	06.01.2004
		CA2351174A1	25.05.2000