(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107948248 A
(43)申请公布日 2018.04.20

(21)申请号 201711060240.0

(22)申请日 2017.11.01

(71)申请人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区八卦岭工业区平安大厦六楼

(72)发明人 李芳 王建明 肖京

(74)专利代理机构 深圳市沃德知识产权代理事务所(普通合伙) 44347
代理人 于志光 郭梦霞

(51)Int.Cl.
H04L 29/08(2006.01)

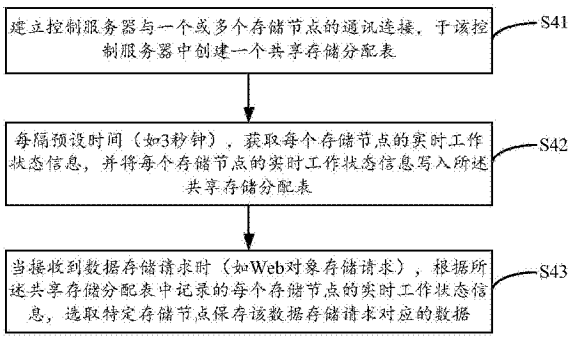
权利要求书2页 说明书10页 附图3页

(54)发明名称

分布式存储方法、控制服务器及计算机可读存储介质

(57)摘要

本发明公开了一种分布式存储方法，该方法包括步骤：建立控制服务器与一个或多个存储节点的通讯连接，于该控制服务器中创建一个共享存储分配表；每隔预设时间，获取每个存储节点的实时工作状态信息，并将每个存储节点的实时工作状态信息写入所述共享存储分配表；当接收到数据存储请求时，根据所述共享存储分配表中记录的每个存储节点的实时工作状态信息，选取特定存储节点保存该数据存储请求对应的数据。本发明可以提高分布式存储的效率和可靠性。



CN 107948248 A

1.一种控制服务器，其特征在于，所述控制服务器包括存储器及处理器，所述存储器上存储有可在所述处理器上运行的分布式存储系统，所述分布式存储系统被所述处理器执行时实现如下步骤：

建立控制服务器与一个或多个存储节点的通讯连接，于该控制服务器中创建一个共享存储分配表；

每隔预设时间，获取每个存储节点的实时工作状态信息，并将每个存储节点的实时工作状态信息写入所述共享存储分配表；及

当接收到数据存储请求时，根据所述共享存储分配表中记录的每个存储节点的实时工作状态信息，选取特定存储节点保存该数据存储请求对应的数据。

2.如权利要求1所述的控制服务器，其特征在于，每个存储节点包括一个主服务器和一个从服务器；

所述共享存储分配表包括存储节点存储的数据、存储节点数组、及数据被分配在存储节点的位置，其中，所述存储节点数组用于记录存储节点的状态信息；及

所述存储节点的状态信息包括存储节点的存储数据量、节点状态、节点是否活着、上一个节点和下一个节点、总容量、及负载因子。

3.如权利要求2所述的控制服务器，其特征在于，所述选取特定的存储节点保存该数据存储请求对应的数据包括：

根据每个存储节点的剩余存储容量和负载因子大小，从所有存储节点中选取剩余存储容量满足该数据存储请求且负载因子最小的特定存储节点，其中，每个存储节点的剩余存储容量等于每个存储节点的总容量减去存储数据量；及

将该数据存储请求对应的数据保存至该特定存储节点的主服务器存储单元中，并在所述共享存储分配表中记录该数据存储请求对应的数据在该特定存储节点的主服务器的存储单元地址。

4.如权利要求3所述的控制服务器，其特征在于，所述分布式存储系统被所述处理器执行时还用于实现如下步骤：

当该数据存储请求对应的数据保存至该特定存储节点的主服务器后，控制该特定存储节点开启数据同步进程，将该数据存储请求对应的数据复制到该特定存储节点的从服务器存储单元中，并在所述共享存储分配表中记录该复制数据在该特定存储节点的从服务器存储单元地址。

5.如权利要求2所述的控制服务器，其特征在于，所述分布式存储系统被所述处理器执行时还用于实现如下步骤：

当一个存储节点的主服务器停止工作时，在所述共享存储分配表中更新该存储节点的主服务器状态信息，并将该存储节点的备份服务器作为主服务器添加到所述共享存储分配表；及

当一个存储节点的主服务器和从服务器都停止工作时，将该存储节点从所述共享存储分配表中删除，并将该存储节点的上一个节点连接至下一个节点。

6.一种分布式存储方法，应用于控制服务器，其特征在于，所述方法包括：

建立控制服务器与一个或多个存储节点的通讯连接，于该控制服务器中创建一个共享存储分配表；

每隔预设时间，获取每个存储节点的实时工作状态信息，并将每个存储节点的实时工作状态信息写入所述共享存储分配表；及

当接收到数据存储请求时，根据所述共享存储分配表中记录的每个存储节点的实时工作状态信息，选取特定存储节点保存该数据存储请求对应的数据。

7.如权利要求6所述的分布式存储方法，其特征在于，每个存储节点包括一个主服务器和一个从服务器；

所述共享存储分配表包括存储节点存储的数据、存储节点数组、及数据被分配在存储节点的位置，其中，所述存储节点数组用于记录存储节点的状态信息；及

所述存储节点的状态信息包括存储节点的存储数据量、节点状态、节点是否活着、上一个节点和下一个节点、总容量、及负载因子。

8.如权利要求7所述的分布式存储方法，其特征在于，所述选取特定的存储节点保存该数据存储请求对应的数据包括：

根据每个存储节点的剩余存储容量和负载因子大小，从所有存储节点中选取剩余存储容量满足该数据存储请求且负载因子最小的特定存储节点，其中，每个存储节点的剩余存储容量等于每个存储节点的总容量减去存储数据量；及

将该数据存储请求对应的数据保存至该特定存储节点的主服务器存储单元中，并在所述共享存储分配表中记录该数据存储请求对应的数据在该特定存储节点的主服务器的存储单元地址。

9.如权利要求8所述的分布式存储方法，其特征在于，该方法还包括：

当该数据存储请求对应的数据保存至该特定存储节点的主服务器后，控制该特定存储节点开启数据同步进程，将该数据存储请求对应的数据复制到该特定存储节点的从服务器存储单元中，并在所述共享存储分配表中记录该复制数据在该特定存储节点的从服务器存储单元地址；

当一个存储节点的主服务器停止工作时，在所述共享存储分配表中更新该存储节点的主服务器状态信息，并将该存储节点的备份服务器作为主服务器添加到所述共享存储分配表；及

当一个存储节点的主服务器和从服务器都停止工作时，将该存储节点从所述共享存储分配表中删除，并将该存储节点的上一个节点连接至下一个节点。

10.一种计算机可读存储介质，所述计算机可读存储介质存储有分布式存储系统，所述分布式存储系统可被至少一个处理器执行，以使所述至少一个处理器执行如权利要求6-9中任一项所述的分布式存储方法的步骤。

分布式存储方法、控制服务器及计算机可读存储介质

技术领域

[0001] 本发明涉及计算机信息技术领域，尤其涉及一种分布式存储方法、控制服务器及计算机可读存储介质。

背景技术

[0002] 目前，针对Web对象的存储大多使用的是传统的集中式存储服务器，在集中式存储服务器的存储模式下，存储服务器容易成为系统的瓶颈，一旦服务器发生故障，可能导致整个系统的瘫痪，并且无法实现大规模存储。分布式存储是目前解决大规模数据存储的有效途径，但是，现有的分布式存储方案不能很好地满足Web对象的自身特点和特定应用场景。因此，急需针对Web对象设计符合特定要求的分布式存储方案。故，现有技术中的分布式存储方法设计不够合理，亟需改进。

发明内容

[0003] 有鉴于此，本发明提出一种分布式存储方法、控制服务器及计算机可读存储介质，通过设置共享存储分配表和分布式存储架构，提高了分布式存储的存储效率。

[0004] 首先，为实现上述目的，本发明提出一种控制服务器，所述控制服务器包括存储器及处理器，所述存储器上存储有可在所述处理器上运行的分布式存储系统，所述分布式存储系统被所述处理器执行时实现如下步骤：

[0005] 建立控制服务器与一个或多个存储节点的通讯连接，于该控制服务器中创建一个共享存储分配表；

[0006] 每隔预设时间，获取每个存储节点的实时工作状态信息，并将每个存储节点的实时工作状态信息写入所述共享存储分配表；及

[0007] 当接收到数据存储请求时，根据所述共享存储分配表中记录的每个存储节点的实时工作状态信息，选取特定存储节点保存该数据存储请求对应的数据。

[0008] 优选地，每个存储节点包括一个主服务器和一个从服务器；

[0009] 所述共享存储分配表包括存储节点存储的数据、存储节点数组、及数据被分配在存储节点的位置，其中，所述存储节点数组用于记录存储节点的状态信息；及

[0010] 所述存储节点的状态信息包括存储节点的存储数据量、节点状态、节点是否活着、上一个节点和下一个节点、总容量、及负载因子。

[0011] 优选地，所述选取特定的存储节点保存该数据存储请求对应的数据包括：

[0012] 根据每个存储节点的剩余存储容量和负载因子大小，从所有存储节点中选取剩余存储容量满足该数据存储请求且负载因子最小的特定存储节点，其中，每个存储节点的剩余存储容量等于每个存储节点的总容量减去存储数据量；及

[0013] 将该数据存储请求对应的数据保存至该特定存储节点的主服务器存储单元中，并在所述共享存储分配表中记录该数据存储请求对应的数据在该特定存储节点的主服务器的存储单元地址。

[0014] 优选地，所述分布式存储系统被所述处理器执行时还用于实现如下步骤：

[0015] 当该数据存储请求对应的数据保存至该特定存储节点的主服务器后，控制该特定存储节点开启数据同步进程，将该数据存储请求对应的数据复制到该特定存储节点的从服务器存储单元中，并在所述共享存储分配表中记录该复制数据在该特定存储节点的从服务器存储单元地址。

[0016] 优选地，所述分布式存储系统被所述处理器执行时还用于实现如下步骤：

[0017] 当一个存储节点的主服务器停止工作时，在所述共享存储分配表中更新该存储节点的主服务器状态信息，并将该存储节点的备份服务器作为主服务器添加到所述共享存储分配表；及

[0018] 当一个存储节点的主服务器和从服务器都停止工作时，将该存储节点从所述共享存储分配表中删除，并将该存储节点的上一个节点连接至下一个节点。

[0019] 此外，为实现上述目的，本发明还提供一种分布式存储方法，该方法应用于控制服务器，所述方法包括：

[0020] 建立控制服务器与一个或多个存储节点的通讯连接，于该控制服务器中创建一个共享存储分配表；

[0021] 每隔预设时间，获取每个存储节点的实时工作状态信息，并将每个存储节点的实时工作状态信息写入所述共享存储分配表；及

[0022] 当接收到数据存储请求时，根据所述共享存储分配表中记录的每个存储节点的实时工作状态信息，选取特定存储节点保存该数据存储请求对应的数据。

[0023] 优选地，每个存储节点包括一个主服务器和一个从服务器；

[0024] 所述共享存储分配表包括存储节点存储的数据、存储节点数组、及数据被分配在存储节点的位置，其中，所述存储节点数组用于记录存储节点的状态信息；及

[0025] 所述存储节点的状态信息包括存储节点的存储数据量、节点状态、节点是否活着、上一个节点和下一个节点、总容量、及负载因子。

[0026] 优选地，所述选取特定的存储节点保存该数据存储请求对应的数据包括：

[0027] 根据每个存储节点的剩余存储容量和负载因子大小，从所有存储节点中选取剩余存储容量满足该数据存储请求且负载因子最小的特定存储节点，其中，每个存储节点的剩余存储容量等于每个存储节点的总容量减去存储数据量；及

[0028] 将该数据存储请求对应的数据保存至该特定存储节点的主服务器存储单元中，并在所述共享存储分配表中记录该数据存储请求对应的数据在该特定存储节点的主服务器的存储单元地址。

[0029] 优选地，该方法还包括：

[0030] 当该数据存储请求对应的数据保存至该特定存储节点的主服务器后，控制该特定存储节点开启数据同步进程，将该数据存储请求对应的数据复制到该特定存储节点的从服务器存储单元中，并在所述共享存储分配表中记录该复制数据在该特定存储节点的从服务器存储单元地址；

[0031] 当一个存储节点的主服务器停止工作时，在所述共享存储分配表中更新该存储节点的主服务器状态信息，并将该存储节点的备份服务器作为主服务器添加到所述共享存储分配表；及

[0032] 当一个存储节点的主服务器和从服务器都停止工作时，将该存储节点从所述共享存储分配表中删除，并将该存储节点的上一个节点连接至下一个节点。

[0033] 进一步地，为实现上述目的，本发明还提供一种计算机可读存储介质，所述计算机可读存储介质存储有分布式存储系统，所述分布式存储系统可被至少一个处理器执行，以使所述至少一个处理器执行如上述的分布式存储方法的步骤。

[0034] 相较于现有技术，本发明所提出的控制服务器、分布式存储方法及计算机可读存储介质，通过设置共享存储分配表和分布式存储架构(控制服务器和存储节点的存储服务器分开)，提供了一种基于共享存储分配表的分布式存储方案，提高了分布式存储的存储效率、可靠性、和容错性，更优于传统的集中式数据存储方案。

附图说明

[0035] 图1是本发明控制服务器与存储节点之间的系统架构示意图；

[0036] 图2是本发明控制服务器一可选的硬件架构的示意图；

[0037] 图3是本发明控制服务器中分布式存储系统一实施例的程序模块示意图；

[0038] 图4为本发明分布式存储方法一实施例的实施流程示意图。

[0039] 附图标记：

[0040]

控制服务器	2
共享存储分配表	24
存储节点	4、5、6
主服务器	41、51、61
从服务器	42、52、62
存储器	21
处理器	22
网络接口	23
分布式存储系统	20
创建模块	201
写入模块	202
存储模块	203
流程步骤	S41-S43

[0041] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0042] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0043] 需要说明的是，在本发明中涉及“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第

一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本发明要求的保护范围之内。

[0044] 进一步需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0045] 参阅图1所示，是本发明控制服务器与存储节点之间的系统架构示意图。本实施例中，控制服务器2与一个或多个存储节点(如存储节点4-6)建立通讯连接，该控制服务器2中创建有一个共享存储分配表24。进一步地，存储节点4包括主服务器41和从服务器42，存储节点5包括主服务器51和从服务器52，存储节点6包括主服务器61和从服务器62。以下通过图2至图4的描述进一步说明本发明的技术方案。

[0046] 首先，本发明提出一种控制服务器2。

[0047] 参阅图2所示，是本发明控制服务器2一可选的硬件架构的示意图。本实施例中，所述控制服务器2可包括，但不限于，可通过系统总线相互通信连接存储器21、处理器22、网络接口23。需要指出的是，图2仅示出了具有组件21-23的控制服务器2，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

[0048] 其中，所述控制服务器2可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器等计算设备，该控制服务器2可以是独立的服务器，也可以是多个服务器所组成的服务器集群。

[0049] 所述存储器21至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器(例如，SD或DX存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器21可以是所述控制服务器2的内部存储单元，例如该控制服务器2的硬盘或内存。在另一些实施例中，所述存储器21也可以是所述控制服务器2的外部存储设备，例如该控制服务器2上配备的插接式硬盘，智能存储卡(Smart Media Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)等。当然，所述存储器21还可以既包括所述控制服务器2的内部存储单元也包括其外部存储设备。本实施例中，所述存储器21通常用于存储安装于所述控制服务器2的操作系统和各类应用软件，例如所述分布式存储系统20的程序代码等。此外，所述存储器21还可以用于暂时地存储已经输出或者将要输出的各类数据。

[0050] 所述处理器22在一些实施例中可以是中央处理器(Central Processing Unit，CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器22通常用于控制所述控制服务器2的总体操作，例如执行与所述控制服务器2进行数据交互或者通信相关的控制和处理等。本实施例中，所述处理器22用于运行所述存储器21中存储的程序代码或者处理数据，例如运行所述的分布式存储系统20等。

[0051] 所述网络接口23可包括无线网络接口或有线网络接口，该网络接口23通常用于在

所述控制服务器2与其他电子设备之间建立通信连接。例如，所述网络接口23用于通过网络将所述控制服务器2与外部数据平台(如主服务器41或从服务器42)相连，在所述控制服务器2与外部数据平台之间的建立数据传输通道和通信连接。所述网络可以是企业内部网(Intranet)、互联网(Internet)、全球移动通讯系统(Global System of Mobile communication，GSM)、宽带码分多址(Wideband Code Division Multiple Access，WCDMA)、4G网络、5G网络、蓝牙(Bluetooth)、Wi-Fi等无线或有线网络。

[0052] 至此，已经详细介绍了本发明各个实施例的应用环境和相关设备的硬件结构和功能。下面，将基于上述应用环境和相关设备，提出本发明的各个实施例。

[0053] 参阅图3所示，是本发明控制服务器2中分布式存储系统20一实施例的程序模块图。本实施例中，所述的分布式存储系统20可以被分割成一个或多个程序模块，所述一个或者多个程序模块被存储于所述存储器21中，并由一个或多个处理器(本实施例中为所述处理器22)所执行，以完成本发明。例如，在图3中，所述的分布式存储系统20可以被分割成创建模块201、写入模块202、以及存储模块203。本发明所称的程序模块是指能够完成特定功能的一系列计算机程序指令段，比程序更适合于描述所述分布式存储系统20在所述控制服务器2中的执行过程。以下将就各程序模块201-203的功能进行详细描述。

[0054] 所述创建模块201，用于建立控制服务器2与一个或多个存储节点(如存储节点4-6)的通讯连接，于该控制服务器2中创建一个共享存储分配表24。在本实施例中，每个存储节点设置为一个存储服务器组，每个存储服务器组包括一个主服务器和一个从服务器，用于数据同步和备份。例如，存储节点4包括主服务器41和从服务器42，存储节点5包括主服务器51和从服务器52，存储节点6包括主服务器61和从服务器62。

[0055] 优选地，在本实施例中，所述共享存储分配表24存储于控制服务器2的存储器21，整个系统中各个存储节点共享一个存储分配表，并且由控制服务器2维护该共享存储分配表24的状态。进一步地，所述共享存储分配表24包括，但不限于，存储节点存储的数据、存储节点数组、及数据被分配在存储节点的位置(如主服务器的存储单元地址或从服务器的存储单元地址)。其中，所述存储节点数组用于记录存储节点的状态信息等数据。

[0056] 进一步地，在本实施例中，所述存储节点的状态信息包括，但不限于如下信息：存储节点的存储数据量(size)、节点状态(开启和关闭)、节点是否活着(isAvlie)、上一个节点和下一个节点、总容量(total)、及负载因子(存储数据量/总容量＝size/total)等。其中，在本实施例中，所述节点状态包括主服务器的状态和从服务器的状态。

[0057] 举例而言，所述共享存储分配表24的数据结构可以定义为如下格式：

[0058]

```
public class SharedEntryList<T>{
    T data;        //存储的数据
    int length;    //表长度
    StoreNode[] entryList;   //存储节点数组，维护存储节点的状态
    int index = data.hashCode() % length; //数据被分配在存储节点的位置
    /*
     *存储节点的数据结构
     */
    private class StoreNode{
      T data;        //数据
      int size;        //存储数据量
      String status;   //状态（开启和关闭）
      Boolean isAlive;   //是否活着
       StoreNode next;     //下一个节点
       StoreNode previous;   //上一个节点
       int total;                  //总容量
       float loadFactor;          //负载因子，即size/total
       }
    }
```

[0059]

[0060] 所述写入模块202，用于每隔预设时间(如3秒钟)，获取每个存储节点的实时工作状态信息，并将每个存储节点的实时工作状态信息写入所述共享存储分配表24，即实时更新所述共享存储分配表24中存储的节点信息。其中，每个存储节点的实时工作状态信息包括，但不限于，存储数据量(size)、节点状态、节点是否活着、上一个节点和下一个节点、总容量、及负载因子等。

[0061] 优选地，在本实施例中，控制服务器2和存储节点之间采用心跳机制进行通讯。具体而言，本实施例中采用的心跳机制是在存储节点中添加一个函数keepAlive()，该函数每隔预设时间(如3秒钟)发送一次请求消息(如http请求)到控制服务器2，该请求消息包括存储节点的实时工作状态信息(如status，isAlive，size等信息)。控制服务器2收到存储节点的实时工作状态信息后，将实时更新所述共享存储分配表24中存储的节点信息，并返回

一个应答消息(如ACK应答)至存储节点。

[0062] 所述存储模块203,用于当接收到数据存储请求时(如Web对象存储请求),根据所述共享存储分配表24中记录的每个存储节点的实时工作状态信息,选取特定存储节点保存该数据存储请求对应的数据。

[0063] 优选地,在本实施例中,所述选取特定的存储节点保存该数据存储请求对应的数据包括:

[0064] 根据每个存储节点的剩余存储容量和负载因子大小,从所有存储节点中选取剩余存储容量满足该数据存储请求且负载因子最小的特定存储节点,将该数据存储请求对应的数据保存至该特定存储节点的主服务器存储单元中,并在所述共享存储分配表24中记录该数据存储请求对应的数据在该特定存储节点的主服务器的存储单元地址。其中,每个存储节点的剩余存储容量等于每个存储节点的总容量(total)减去存储数据量(size)。由于在本实施例中根据每个存储节点的存储数据量(size)进行负载均衡操作,从而保证了数据能被均匀地分布存储在各个存储节点上。

[0065] 进一步地,当该数据存储请求对应的数据保存至该特定存储节点的主服务器后,控制该特定存储节点开启数据同步进程,将该数据存储请求对应的数据复制到该特定存储节点的从服务器存储单元中,并在所述共享存储分配表24中记录该复制数据在该特定存储节点的从服务器存储单元地址,实现数据同步和备份。也就是说,同一份数据,在该特定存储节点上有两个拷贝,当主服务器停止工作时,还可以在从服务器中找到相应数据。

[0066] 进一步地,在其它实施例中,所述分布式存储系统20还用于:

[0067] 当一个存储节点的主服务器停止工作时,在所述共享存储分配表24中更新该存储节点的主服务器状态信息(如更新为关闭),并将该存储节点的备份服务器(如从服务器)作为主服务器添加到所述共享存储分配表24,以保证系统的可用性;

[0068] 当一个存储节点的主服务器和从服务器都停止工作时,将该存储节点从所述共享存储分配表24中删除,并将该存储节点的上一个节点连接至下一个节点。

[0069] 通过上述程序模块201-203,本发明所提出的分布式存储系统20,通过设置共享存储分配表和分布式存储架构(控制服务器和存储节点的存储服务器分开),提供了一种基于共享存储分配表的分布式存储方案,提高了分布式存储的存储效率、可靠性、和容错性,更优于传统的集中式数据存储方案。

[0070] 此外,本发明还提出一种分布式存储方法。

[0071] 参阅图4所示,是本发明分布式存储方法一实施例的实施流程示意图。在本实施例中,根据不同的需求,图4所示的流程图中的步骤的执行顺序可以改变,某些步骤可以省略。

[0072] 步骤S41,建立控制服务器2与一个或多个存储节点(如存储节点4-6)的通讯连接,于该控制服务器2中创建一个共享存储分配表24。在本实施例中,每个存储节点设置为一个存储服务器组,每个存储服务器组包括一个主服务器和一个从服务器,用于数据同步和备份。例如,存储节点4包括主服务器41和从服务器42,存储节点5包括主服务器51和从服务器52,存储节点6包括主服务器61和从服务器62。

[0073] 优选地,在本实施例中,所述共享存储分配表24存储于控制服务器2的存储器21,整个系统中各个存储节点共享一个存储分配表,并且由控制服务器2维护该共享存储分配表24的状态。进一步地,所述共享存储分配表24包括,但不限于,存储节点存储的数据、存储

节点数组、及数据被分配在存储节点的位置(如主服务器的存储单元地址或从服务器的存储单元地址)。其中,所述存储节点数组用于记录存储节点的状态信息等数据。

[0074] 进一步地,在本实施例中,所述存储节点的状态信息包括,但不限于如下信息:存储节点的存储数据量(size)、节点状态(开启和关闭)、节点是否活着(isAvlie)、上一个节点和下一个节点、总容量(total)、及负载因子(存储数据量/总容量=size/total)等。其中,在本实施例中,所述节点状态包括主服务器的状态和从服务器的状态。

[0075] 举例而言,所述共享存储分配表24的数据结构可以定义为如下格式:

[0076]
```
public class SharedEntryList<T>{
T data;        //存储的数据
int length;   //表长度
StoreNode[] entryList;   //存储节点数组, 维护存储节点的状态
int index = data.hashCode() % length; //数据被分配在存储节点的位置
/*
 *存储节点的数据结构
 */
private class StoreNode{
```

[0077]
```
  T data;        //数据
  int size;        //存储数据量
  String status;   //状态(开启和关闭)
  Boolean isAlive;   //是否活着
  StoreNode next;     //下一个节点
  StoreNode previous;   //上一个节点
  int total;                  //总容量
  float loadFactor;          //负载因子, 即size/total
  }
}
```

[0078] 步骤S42,每隔预设时间(如3秒钟),获取每个存储节点的实时工作状态信息,并将每个存储节点的实时工作状态信息写入所述共享存储分配表24,即实时更新所述共享存储分配表24中存储的节点信息。其中,每个存储节点的实时工作状态信息包括,但不限于,存储数据量(size)、节点状态、节点是否活着、上一个节点和下一个节点、总容量、及负载因子

等。

[0079] 优选地，在本实施例中，控制服务器2和存储节点之间采用心跳机制进行通讯。具体而言，本实施例中采用的心跳机制是在存储节点中添加一个函数keepAlive()，该函数每隔预设时间(如3秒钟)发送一次请求消息(如http请求)到控制服务器2，该请求消息包括存储节点的实时工作状态信息(如status，isAlive，size等信息)。控制服务器2收到存储节点的实时工作状态信息后，将实时更新所述共享存储分配表24中存储的节点信息，并返回一个应答消息(如ACK应答)至存储节点。

[0080] 步骤S43，当接收到数据存储请求时(如Web对象存储请求)，根据所述共享存储分配表24中记录的每个存储节点的实时工作状态信息，选取特定存储节点保存该数据存储请求对应的数据。

[0081] 优选地，在本实施例中，所述选取特定的存储节点保存该数据存储请求对应的数据包括：

[0082] 根据每个存储节点的剩余存储容量和负载因子大小，从所有存储节点中选取剩余存储容量满足该数据存储请求且负载因子最小的特定存储节点，将该数据存储请求对应的数据保存至该特定存储节点的主服务器存储单元中，并在所述共享存储分配表24中记录该数据存储请求对应的数据在该特定存储节点的主服务器的存储单元地址。其中，每个存储节点的剩余存储容量等于每个存储节点的总容量(total)减去存储数据量(size)。由于在本实施例中根据每个存储节点的存储数据量(size)进行负载均衡操作，从而保证了数据能被均匀地分布存储在各个存储节点上。

[0083] 进一步地，当该数据存储请求对应的数据保存至该特定存储节点的主服务器后，控制该特定存储节点开启数据同步进程，将该数据存储请求对应的数据复制到该特定存储节点的从服务器存储单元中，并在所述共享存储分配表24中记录该复制数据在该特定存储节点的从服务器存储单元地址，实现数据同步和备份。也就是说，同一份数据，在该特定存储节点上有两个拷贝，当主服务器停止工作时，还可以在从服务器中找到相应数据。

[0084] 进一步地，在其它实施例中，所述分布式存储方法还包括如下步骤：

[0085] 当一个存储节点的主服务器停止工作时，在所述共享存储分配表24中更新该存储节点的主服务器状态信息(如更新为关闭)，并将该存储节点的备份服务器(如从服务器)作为主服务器添加到所述共享存储分配表24，以保证系统的可用性；

[0086] 当一个存储节点的主服务器和从服务器都停止工作时，将该存储节点从所述共享存储分配表24中删除，并将该存储节点的上一个节点连接至下一个节点。

[0087] 通过上述步骤S41-S43及其它相关步骤，本发明所提出的分布式存储方法，通过设置共享存储分配表和分布式存储架构(控制服务器和存储节点的存储服务器分开)，提供了一种基于共享存储分配表的分布式存储方案，提高了分布式存储的存储效率、可靠性、和容错性，更优于传统的集中式数据存储方案。

[0088] 进一步地，为实现上述目的，本发明还提供一种计算机可读存储介质(如ROM/RAM、磁碟、光盘)，所述计算机可读存储介质存储有分布式存储系统20，所述分布式存储系统20可被至少一个处理器22执行，以使所述至少一个处理器22执行如上所述的分布式存储方法的步骤。

[0089] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方

法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件来实现，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本发明各个实施例所述的方法。

[0090] 以上参照附图说明了本发明的优选实施例，并非因此局限本发明的权利范围。上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。另外，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0091] 本领域技术人员不脱离本发明的范围和实质，可以有多种变型方案实现本发明，比如作为一个实施例的特征可用于另一实施例而得到又一实施例。凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

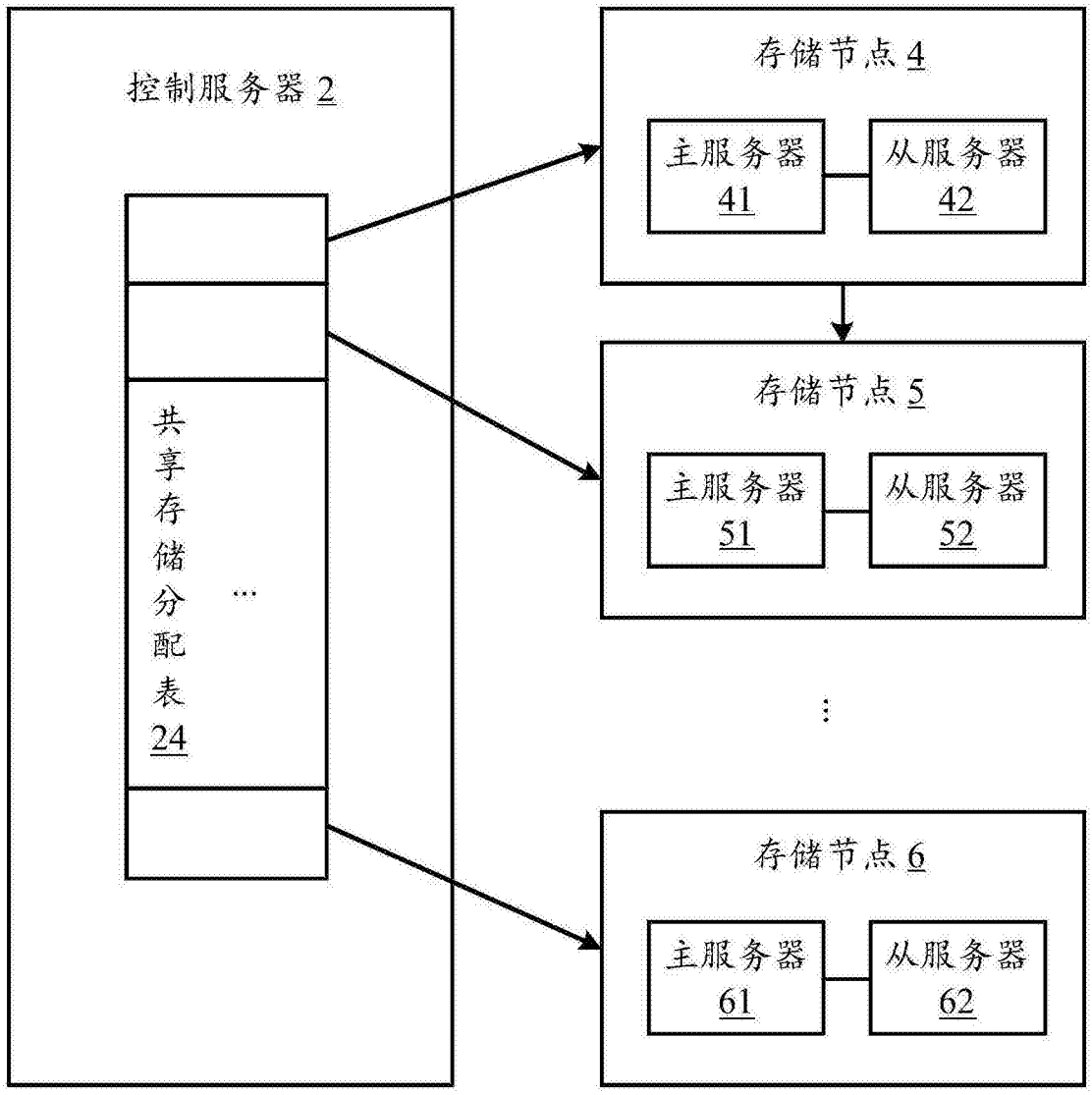

控制服务器 2
共享存储分配表 24
...
存储节点 4
主服务器 41
从服务器 42
存储节点 5
主服务器 51
从服务器 52
存储节点 6
主服务器 61
从服务器 62

图1

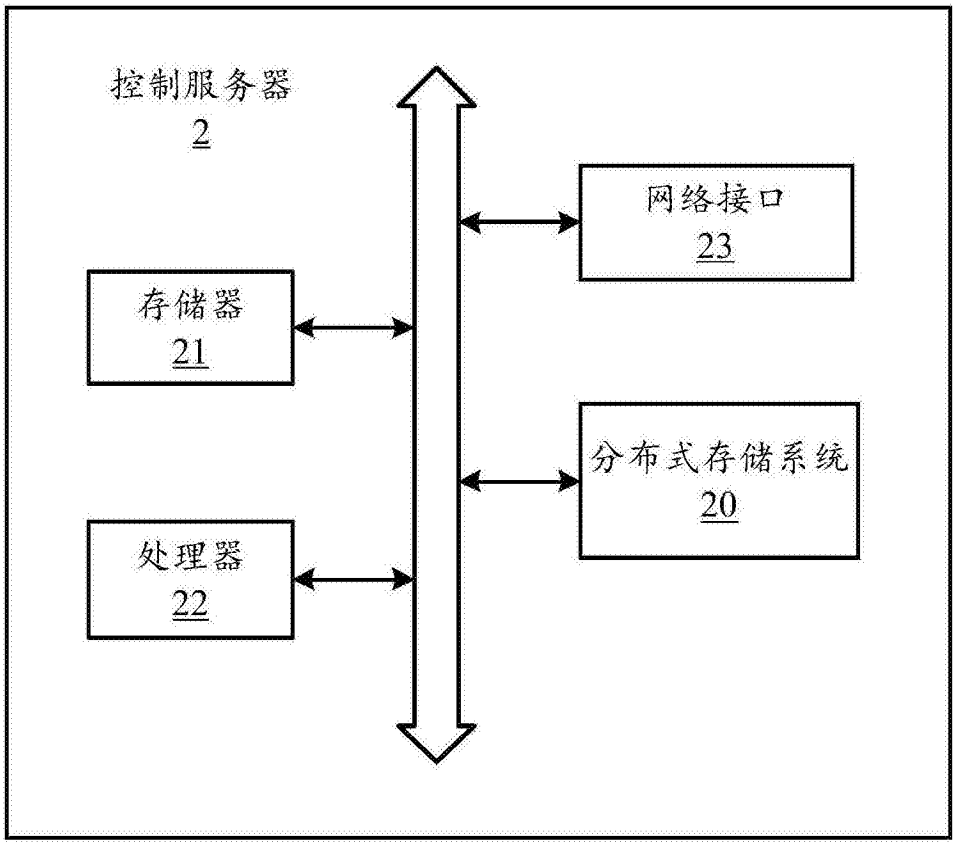


图2

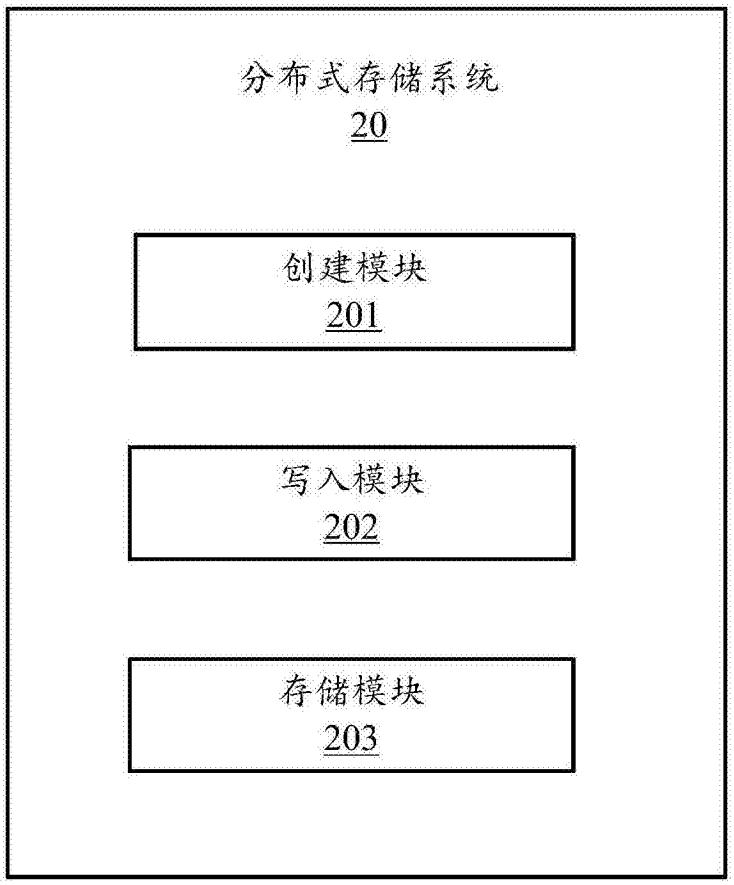

分布式存储系统
20
创建模块
201
写入模块
202
存储模块
203

图3

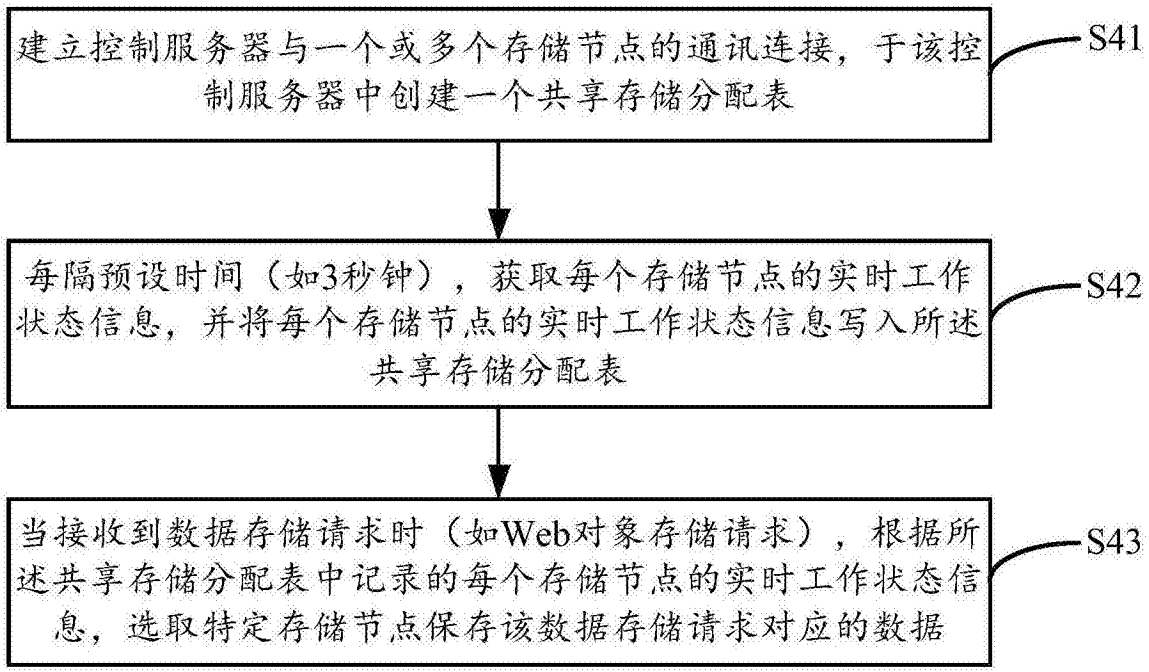

建立控制服务器与一个或多个存储节点的通讯连接，于该控制服务器中创建一个共享存储分配表
S41
每隔预设时间（如3秒钟），获取每个存储节点的实时工作状态信息，并将每个存储节点的实时工作状态信息写入所述共享存储分配表
S42
当接收到数据存储请求时（如Web对象存储请求），根据所述共享存储分配表中记录的每个存储节点的实时工作状态信息，选取特定存储节点保存该数据存储请求对应的数据
S43

图4