

分布式存储系统的访问方法和相关装置和相关系统

技术领域

本申请涉及了计算机技术领域，具体涉及了分布式存储系统的访问方法和相关装置和相关系统。

5 背景技术

传统网络存储系统采用集中的存储服务器存放所有数据，这种架构下存储服务器成为系统性能瓶颈，也是可靠性和安全性的薄弱点，难以满足大规模存储应用的需要。分布式存储系统则是将数据分散存储在多台独立设备（即存储节点）上。分布式存储系统采用可扩展的系统结构，利用多台存储节点分担存储负荷，分布式存储系统有利于提高系统的可靠性、可用性、存取效率和扩展性。

随着存储容量的不断增大，分布式存储系统中的存储节点数量和硬盘数量也在不断的扩大。越来越多的分布式存储系统使用纠删码（英文：erasure code，缩写：EC）技术将数据打散存放在不同硬盘上，以提高数据可靠性。EC技术在尽量保证系统可靠性的同时，还可有效地提升系统性能。当然，除EC技术之外，其他类似的冗余校验技术也可能被分布式存储系统使用。

分条（stripe）（例如EC分条）包括多个条带（英文：strip），具体可包括N个数据条带和M个校验条带（例如包括2个数据strip和1个校验strip），这M个校验条带基于这N个数据条带计算得到，数据条带的长度可达1Mb。传统技术中，即使读操作所欲读取的目标数据仅占分条的一小部分，每次读操作都需读取出目标数据所属分条中的所有数据条带和校验条带，这样才能进行数据的正确性校验，从而确保目标数据的正确性。但由于分条的长度往往比较大，这使得在例如随机读等场景下，会加重硬盘负担，大大降低系统性能。

发明内容

本申请实施例提供数据访问方法和相关装置及系统。

25 第一方面，本申请实施例提供一种数据读取方法，所述方法应用于包括m个存储节点的分布式存储系统，每个所述存储节点包括至少1个存储器（存储器例如可为硬盘或其他形式的存储器），所述m为大于1的整数。所述方法可包括：例如当终端需从分布式存储系统请求读取第一数据，所述m个存储节点之中的第一存储节点可接收来自终端的第一读数据请求。所述第一读数据请求携带有第一位置信息，所述第一位置信息用于描述第一数据在其所属数据体中的位置。在所述第一存储节点基于所述第一位置信息确定所述第一数据所在的存储器为第一存储器，且所述第一存储器归属于所述第一存储节点的情况之下，所述
30 第一存储节点确定状态信息集合体（所述第一存储节点缓存的状态信息集合体）中当前记录的所述第一存储器所处状态（所述第一存储器例如可能处于信任访问状态或者非信任访问状态）。当确定出所述第一存储器处于信任访问状态，所述第一存储节点从所述第一存储器读取所述第一数据。第一存储节点向所述终端发送用于响应所述第一读数据请求的第
35

一数据响应，所述第一数据响应携带有读取到的所述第一数据。所述第一数据为所述第一数据所属条带中的部分或全部数据。

本申请各实施例中的数据体例如可为文件、对象、文件分段、数据库记录或者数据段等等。

- 5 其中，本申请各实施例中的状态信息集合体用于记录硬盘状态等，状态信息集合体的格式例如可为文件或其他数据格式。如果状态信息集合体的格式为文件，那么状态信息集合体也可称为“状态文件”。

可以看出，上述方案中，由于在分布式存储系统中，引入了用于记录存储器状态的状态信息集合体，故而可以利用状态信息集合体来记录管理分布式存储系统的存储器的访问状态，对于状态信息集合体记录的处于信任访问状态的存储器，可直接从中读取被请求数据来向终端反馈，而在这种信任访问状态的情况下，由于无需执行传统方式中先读取出相关分条的全部条带来进行正确性校验的繁琐步骤，通过减少数据读取量就有利于降低存储器负担，进而有利于提升系统性能。

- 15 此外，在本申请一些可能实施方式中，在所述第一存储节点确定状态信息集合体中当前记录的所述第一存储器所处状态之前，所述第一存储节点可先确定是否缓存了状态信息集合体，如果已缓存了状态信息集合体，那么所述第一存储节点可确定缓存的状态信息集合体中当前记录的所述第一存储器所处状态；如果未缓存状态信息集合体，且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，所述第一存储节点先从管理节点获取状态信息集合体进行缓存，而后再确定缓存的状态信息集合体中当前记录的所述第一存储器所处状态。

- 20 此外，在本申请一些可能实施方式中，在所述第一存储节点确定状态信息集合体中当前记录的所述第一存储器所处状态之前，所述第一存储节点可先确定是否缓存了状态信息集合体，如果已经缓存了状态信息集合体，那么所述第一存储节点可确定缓存的状态信息集合体的获得时长是否超过时长阈值（时长阈值例如可为2分钟、10分钟或30分钟等，通常来说，当前缓存的状态信息集合体的获得时长越长，那么这个状态信息集合体失效的可能性越大）。在缓存的状态信息集合体的获得时长超过时长阈值，且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，所述第一存储节点先从管理节点获取状态信息集合体，利用最新获取的状态信息集合体更新此前已缓存的状态信息集合体，而后再确定缓存的状态信息集合体中当前记录的所述第一存储器所处状态。在缓存的状态信息集合体的获得时长未超过，或者所述第一存储节点与用于发布状态信息集合体的管理节点之间未处于连接态的情况下，那么所述第一存储节点确定缓存的状态信息集合体中当前记录的所述第一存储器所处状态。

- 30 此外，在本申请一些可能实施方式中，所述方法还可包括：在所述第一存储器归属于所述m个存储节点中的第二存储节点，且确定出所述第一存储器处于信任访问状态（具体例如第一存储节点基于状态信息集合体确定出所述第一存储器处于信任访问状态，即第一存储节点缓存的状态信息集合体中记录：所述第一存储器处于信任访问状态），且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下（当所述第一

存储节点与用于发布状态信息集合体的管理节点之间处于连接态，那么第一存储节点缓存的状态信息集合体是管理节点最新发布的状态信息集合体的可能性相对很大，因此第一存储节点当前缓存的这个状态信息集合体当前有效的可能性相对很大。故而可认为，第一存储节点基于缓存的这个状态信息集合体来确定出所述第一存储器处于信任访问状态，很大可能是符合所述第一存储器的实际情况的），所述第一存储节点向所述第二存储节点转发所述第一读数据请求。所述第二存储节点在接收到来自所述第一存储节点的所述第一读数据请求之后，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

在本申请的一些可能实施方式中，所述方法还可包括：在所述第一存储器归属于所述m个存储节点之中的第二存储节点，且所述第一存储器处于信任访问状态（具体例如第一存储节点当前缓存的状态信息集合体中记录：所述第一存储器处于信任访问状态）的情况下，或者，在所述第一存储器归属于所述m个存储节点之中的第二存储节点，且所述第一存储器处于信任访问状态，且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下（当第一存储节点与管理节点之间处于连接态，那么第一存储节点获得管理节点最新发布的状态信息集合体的可能性就非常大），第一存储节点可在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识，向所述第二存储节点发送添加了所述状态信息集合体的标识的所述第一读数据请求。

所述第二存储节点在接收到来自所述第一存储节点的所述第一读数据请求之后，可将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下（当所述第一读数据请求携带的所述状态信息集合体的标识与第二存储节点自身当前缓存的状态信息集合体的标识相同，那就表示所述第一存储节点和所述第二存储节点当前缓存的状态信息集合体相同，这种情况下所述第一存储节点当前缓存的状态信息集合体当前有效的可能性非常大。在这种情况下，如果所述第一存储节点与管理节点之间处于连接态，那么，第一存储节点当前缓存的状态信息集合体当前有效的可能性进一步加大，因此所述第一存储节点基于其当前缓存的状态信息集合体确定所述第一存储器处于信任访问状态，很大可能是符合所述第一存储器的实际情况的），从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

本申请各实施例中，状态信息集合体的标识例如包括状态信息集合体的版本号和/或状态信息集合体的数据摘要等。可以理解，某版本的状态信息集合体的标识具体可以是能代表这个版本的状态信息集合体的任何信息，状态信息集合体的版本号和/或数据摘要只是这个状态信息集合体的标识的一些可能举例。

在一般情况下，通常是由某个特定的节点（例如分布式存储系统中的某特定存储节点）来承担管理节点的角色，而在一些特别情况下（例如当前管理节点故障或处理能力不足等

情况下),也可能允许承担管理节点这个角色的存储节点发生变化,例如在某个时段是由存储节点A来承担管理节点这个角色(此时,管理节点的节点标识也即为存储节点A的节点标识),但在另一个时段存储节点A不再承担管理节点这个角色,而变化为由存储节点B来承担管理节点这个角色(此时管理节点的节点标识为存储节点B的节点标识)。也就是说在5 一些特别的情况下,由于承担管理节点这个角色的存储节点可能变化,因此管理节点的节点标识(即承担管理节点这个角色的存储节点的节点标识)存在变化的可能性。分布式存储系统中的各存储节点可缓存当前被其视为最新管理节点的节点标识,当存储节点发现承担管理节点这个角色的存储节点发生变化,那么就将使用其发现的最新的管理节点的节点标识来更新其此前缓存的管理节点的节点标识。

10 在本申请一些可能实施方式中,所述方法还可包括:在所述第一存储器归属于所述m个存储节点中的第二存储节点,且所述第一存储器处于信任访问状态(具体例如第一存储节点缓存的状态信息集合体中记录所述第一存储器处于信任访问状态),并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下,所述第一存储节点在所述第一读数据请求中添加所述第一存储节点当前缓存的用于发布状态信息集合体的15 管理节点的节点标识,向所述第二存储节点发送添加了所述管理节点的节点标识的所述第一读数据请求。

那么,所述第二存储节点在接收到来自所述第一存储节点的所述第一读数据请求的情况下,或者在接收到来自所述第一存储节点的所述第一读数据请求,且所述第二存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下,所述第二存储节点将所20 述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识进行比较。其中,在比较出所述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识相同的情况之下(当所述第一存储节点和第二存储节点缓存的管理节点的节点标识相同,也就说明第一存储节点和第二存储节点将同一个存储节点视为管理节点,那么被第一存储节点和第二存储节点均视为管理节点的这个存储节点,是分布25 式存储系统中的当前最新管理节点的可能性就非常大,在这种情况下,如果所述第一存储节点与管理节点之间处于连接态,那么,所述第一存储节点缓存的状态信息集合体当前有效的可能性非常大,因此可以认为,所述第一存储节点基于其缓存的这个状态信息集合体来确定出所述第一存储器处于信任访问状态,很大可能是符合所述第一存储器的实际情况的),从所述第一存储器读取所述第一数据,向所述第一存储节点或所述终端发送用于30 响应所述第一读数据请求的第一读数据响应,其中,所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中,所述方法还可包括:在所述第一存储器归属于所述m个存储节点中的第二存储节点,且所述第一存储器处于信任访问状态(具体例如第一存储节点缓存的状态信息集合体中记录所述第一存储器处于信任访问状态)的情况下,或者在35 所述第一存储器归属于所述m个存储节点中的第二存储节点,且所述第一存储器处于信任访问状态(例如第一存储节点缓存的状态信息集合体中记录所述第一存储器处于信任访问状态),并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的

情况下, 所述第一存储节点在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识和用于发布状态信息集合体的管理节点的节点标识, 向所述第二存储节点发送添加了所述所述状态信息集合体的标识和所述管理节点的节点标识的所述第一读数据请求。

5 那么, 所述第二存储节点在接收到来自所述第一存储节点的所述第一读数据请求的情况下, 或者在接收到来自所述第一存储节点的所述第一读数据请求, 且所述第二存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下, 所述第二存储节点将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较, 并且将所述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识进行比较, 其中, 在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同, 且比较出所述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识相同的情况之下 (当所述第一存储节点和第二存储节点缓存的管理节点的节点标识相同, 也就说明第一存储节点和第二存储节点将同一个存储节点视为管理节点, 那么被第一存储节点和第二存储节点均视为管理节点的这个存储节点, 是分布式存储系统中的当前最新管理节点的可能性就非常大, 在这种情况下, 如果第一存储节点和第二存储节点当前缓存的状态信息集合体相同, 那么第一存储节点当前缓存的状态信息集合体当前有效的可能性非常大, 进一步的, 如果第一存储节点和/或第一存储节点与管理节点之间处于连接态, 那么第一存储节点当前缓存的状态信息集合体当前有效的可能性就更大了, 因此所述第一存储节点基于其当前缓存的状态信息集合体确定所述第一存储器处于信任访问状态, 很大可能是符合所述第一存储器的实际情况的), 从所述第一存储器读取所述第一数据, 向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应, 所述第一读数据响应携带有读取到的所述第一数据。

20 在本申请的一些可能实施方式中, 所述方法还包括: 在所述第一存储节点确定所述第一存储器处于非信任访问状态的情况下, 所述第一存储节点确定所述第一数据所属的分条, 所述第一存储节点确定所述分条的各条带所在的N个存储器。

 在所述N个存储器均属于所述第一存储节点的情况下, 所述第一存储节点从所述N个存储器读取所述分条的各条带, 所述第一存储节点基于读取到的所述分条的各条带进行校验运算以得到所述第一数据 (具体的, 所述第一存储节点先基于读取到的所述分条的各条带进行校验运算以得到所述第一数据所属条带, 而后从所述第一数据所属条带中得到所述第一数据), 向所述终端发送用于响应所述第一读数据请求的第一读数据响应, 所述第一读数据响应携带有得到的所述第一数据。

30 在本申请的一些可能的实施方式中, 当所述第一数据所属的分条的各条带所在的N个存储器中的x个存储器 (x个存储器包括所述第一存储器) 属于所述第一存储节点, 且所述N个存储器中的N-x个存储器属于不同于所述第一存储节点的y个存储节点, 所述方法还可包括: 在所述第一存储节点确定所述第一存储器处于非信任访问状态, 且确定所述N个存储器中除所述第一存储节点之外的其它存储器处于信任访问状态的情况下 (或者在所述第

一存储节点确定所述第一存储器处于非信任访问状态，并且确定所述N个存储器中除所述第一存储节点之外的其它存储器处于信任访问状态，并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下），所述第一存储节点向所述y个存储节点发送携带所述分条的分条标识和所述第一存储节点所缓存状态信息集合体的标识的读数据请求。相应的，所述y个存储节点中的每个存储节点可在接收到来自所述第一存储节点的读数据请求之后，将所述读数据请求携带的状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，若比较出所述读数据请求携带的状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同（当读数据请求携带的状态信息集合体的标识与接收这个读数据请求的存储节点自身当前缓存的状态信息集合体的标识相同，那就表示所述第一存储节点和接收这个读数据请求的存储节点当前缓存的状态信息集合体相同，这种情况下，所述第一存储节点当前缓存的状态信息集合体当前有效的可能性非常大，在这种情况下，如果第一存储节点与管理节点之间处于连接态，那么第一存储节点当前缓存的状态信息集合体当前有效的可能性进一步加大，因此所述第一存储节点基于其当前缓存的状态信息集合体确定所述第一存储器处于非信任访问状态，很大可能是符合所述第一存储器的实际情况的），则可从其包括的相应存储器中读取所述分条的相应条带，向第一存储节点发送读取到的所述分条的相应条带。

相应的，所述第一存储节点基于从所述y个存储节点和所述第一存储节点收集到的所述分条的各条带（此处不包括第一数据所属条带）进行校验运算以得到第一数据，向所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有得到的所述第一数据。

在本申请的一些可能的实施方式之中，所述m个存储节点中的其中1个存储节点为管理节点。所述方法还可包括：当第一存储器所处状态由第一状态变更为第二状态，所述第一存储节点向管理节点发送存储器状态变更报告，其中，所述存储器状态变更报告指示第一存储器处于第二状态；其中，所述第一状态和第二状态不同，所述第一状态和所述第二状态包括如下状态中的任意一个：离线状态、数据重构状态和信任访问状态。

相应的，所述管理节点可以在接收到了来自所述第一存储节点的存储器状态变更报告之后，将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态，并更新所述管理节点缓存的所述状态信息集合体的版本号。所述管理节点向所述m个存储节点之中除所述管理节点之外的其他存储节点发送更新后的状态信息集合体。

相应的，所述第一存储节点用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

在本申请一些可能实施方式中，所述方法还包括：例如当终端需向分布式存储系统请求写入第二数据，所述第一存储节点可接收来自终端的第一写数据请求，所述第一写数据请求携带有第二数据和第二位置信息，所述第二位置信息用于描述所述第二数据在其所属数据体中的位置。所述第一存储节点基于所述第二位置信息确定写所述第二数据所涉及到的W个存储器。所述第一存储节点将所述第二数据切分为W-T个数据条带，利用所述W-T个数据条带计算得到T个校验条带。其中，所述T个校验条带和W-T个数据条带形成包括W

个条带的1个分条，所述W个存储器与所述W个条带之间一一对应，所述T和W为正整数且所述T小于所述W。

所述第一存储节点确定其缓存的状态信息集合体之中当前记录的所述W个存储器所处状态，当确定出所述W个存储器中的W1个存储器处于非离线状态（例如并且所述W个存储器中的W2个存储器处于离线状态），向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带和所述状态信息集合体的标识的写数据请求。

相应的，所述y2个存储节点中的每个存储节点在接收到来自所述第一存储节点的写数据请求之后，将所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，例如在比较出所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下，向其包括的相应存储器中写入所述分条的相应条带；在比较出所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识不相同的情况下，拒绝向其包括的相应存储器中写入所述分条的相应条带。

又如在本申请一些可能实施方式中，所述方法还可包括：例如当终端需向分布式存储系统请求写入第三数据，所述第一存储节点接收来自终端的第二写数据请求，所述第二写数据请求携带有第三数据和第三位置信息，所述第三位置信息用于描述所述第三数据在其所属数据体中的位置。所述第一存储节点基于所述第三位置信息确定写所述第三数据所涉及到的W个存储器；所述第一存储节点将所述第三数据切分为W-T个数据条带，利用所述W-T个数据条带计算得到T个校验条带，其中，所述T个校验条带和W-T个数据条带形成包括W个条带的1个分条，所述W个存储器与所述W个条带之间一一对应，所述T和W为正整数且所述T小于所述W。

所述第一存储节点确定其缓存的状态信息集合体之中当前记录的所述W个存储器所处状态，在确定出所述W个存储器中的W1个存储器处于非离线状态（例如并且所述W个存储器中的W2个存储器处于离线状态），并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带的写数据请求。相应的，所述y2个存储节点之中的每个1个存储节点在接收到来自所述第一存储节点的写数据请求之后，便可向其包括的相应存储器中写入所述分条的相应条带。

又例如，在本申请一些可能实施方式中，所述方法还可包括：所述第一存储节点接收来自终端的第三写数据请求，所述第三写数据请求携带有第四数据和第四位置信息，所述第四位置信息用于描述所述第四数据在其所属数据体中的位置。所述第一存储节点基于所述第四位置信息确定写所述第四数据涉及的W个存储器之后；将所述第四数据切分为W-T个数据条带，利用所述W-T个数据条带计算得到T个校验条带，所述T个校验条带和W-T个数据条带形成包括W个条带的分条，所述W个存储器与所述W个条带之间一一对应，所述T和W为正整数且所述T小于所述W。当所述第一存储节点缓存的状态信息集合体中当前记录所述W个存储器中的W1个存储器处于非离线状态（例如并且所述W个存储器中的W2个存储器处于离线状态），并且所述第一存储节点与用于发布状态信息集合体的管理节点之间

处于连接态的情况下,所述第一存储节点向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带和用于发布状态信息集合体的管理节点的节点标识的写数据请求。

相应的,所述y2个存储节点中的每个存储节点在接收到来自所述第一存储节点的写数据请求之后,可将所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识进行比较,其中,在比较出所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识相同的情况下,向其包括的相应存储器中写入所述分条的相应条带,此外,在比较出所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识不相同的情况下,可拒绝向其包括的相应存储器中写入所述分条的相应条带。

在本申请一些可能实施方式中,当分配给某存储器(例如第二存储器)的条带因这个存储器当时离线等原因而未成功写入这个存储器,那么在这个存储器重新在线之后可通过重构机制来进行相关条带重构。例如所述W1小于所述W,其中,所述第一存储节点缓存的状态信息集合体中当前记录所述W个存储器中的W2个存储器处于离线状态,所述方法还可包括:所述第一存储节点生成第一重构日志,所述第一重构日志中记录了所述W2个存储器中的第二存储器的存储器标识,所述第一重构日志中还进一步记录了所述W个条带中的与所述第二存储器所对应的第一条带的条带标识,所述第一重构日志中还记录了所述分条的分条标识;所述第二存储器为所述W2个存储器中的任意一个存储器。

相应的,当所述W2个存储器中的所述第二存储器重新上线之后,并且所述第二存储器所归属的第二存储节点收集到了在所述第二存储器离线期间产生的所述第一重构日志,第二存储节点获取第一重构日志中记录的需写入所述第二存储器的第一条带的标识;确定所述分条所包括的所述第一条带之外其他W-T个条带所在的W-T个存储器,从所述W-T个存储器中读取所述W-T个条带,利用所述W-T个条带进行校验运算以重构出需写入所述第二存储器的所述第一条带,将重构出的所述第一条带写入所述第二存储器。

可以看出,上述举例的方案中给出了可在一定程度上验证状态信息集合体有效性的若干种可能实施方式,这些可能实施方式有利于满足多种可靠性等级的验证需求,进而有利于满足多种读写访问效率需求。

第二方面,本申请实施例还提供一种分布式存储系统,所述分布式存储系统包括m个存储节点,其中,每个所述存储节点包括至少1个存储器,每个所述存储器包含非易失性存储介质,所述m为大于1的整数。

所述m个存储节点中的第一存储节点用于:接收来自终端的第一读数据请求,所述第一读数据请求携带有第一位置信息,其中,所述第一位置信息用于描述第一数据在其所属数据体中的位置;基于所述第一位置信息确定所述第一数据所在的第一存储器。

在所述第一存储器归属于所述第一存储节点的情况下,当状态信息集合体中当前记录所述第一存储器处于信任访问状态,从所述第一存储器读取所述第一数据,向所述终端发送用于响应所述第一读数据请求的第一读数据响应,所述第一读数据响应携带有读取到的所述第一数据。所述第一数据例如为所述第一数据所属条带中的部分或全部数据。

在本申请一些可能实施方式中,所述分布式存储系统还包括用于发布状态信息集合体

的管理节点。

其中，所述第一存储节点还用于，在所述第一存储器归属于所述m个存储节点之中的第二存储节点，并且所述第一存储器处于信任访问状态，并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，向所述第二存储节点转发所述第一读数据请求。

其中，所述第二存储节点用于，在接收到来自所述第一存储节点的所述第一读数据请求后，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中，所述第一存储节点还用于，在所述第一存储器归属于所述m个存储节点之中的第二存储节点，并且，所述第一存储器处于信任访问状态的情况之下，在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识，向所述第二存储节点发送添加了所述所述状态信息集合体的标识的所述第一读数据请求。

其中，所述第二存储节点用于，接收到来自所述第一存储节点的所述第一读数据请求之后，将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中，所述第一存储节点还用于，在所述第一存储器归属于所述m个存储节点之中的第二存储节点，并且，所述第一存储器处于信任访问状态的情况之下，在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识和用于发布状态信息集合体的管理节点的节点标识，向所述第二存储节点发送添加了所述状态信息集合体的标识和所述管理节点的节点标识的所述第一读数据请求。

相应的，所述第二存储节点用于，在接收到来自所述第一存储节点的所述第一读数据请求之后，将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，且将所述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识进行比较，在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同，且比较出所述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识相同的情况之下，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中，所述第一存储节点还用于：在确定所述第一存储器处于非信任访问状态，并且所述第一数据所属的分条的各条带所在的N个存储器中的x个存储器属于所述第一存储节点，且所述N个存储器中的N-x个存储器属于不同于所述第一存储节

点的 y 个存储节点的情况下,向所述 y 个存储节点发送携带所述分条的分条标识和所述状态信息集合体的标识的读数据请求。

所述 y 个存储节点中的每个存储节点用于,在接收到来自所述第一存储节点的读数据请求之后,将所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较,若比较出所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同,从其包括的相应存储器中读取所述分条的相应条带,向第一存储节点发送读取到的所述分条的相应条带。

所述第一存储节点还用于,基于从所述 y 个存储节点和所述第一存储节点收集到的所述分条的各条带进行校验运算以得到第一数据,向所述终端发送用于响应所述第一读数据请求的第一读数据响应,所述第一读数据响应携带有得到的所述第一数据。

在本申请一些可能实施方式中,例如所述 m 个存储节点之中的其中1个存储节点为管理节点。其中,所述第一存储节点还用于,当所述第一存储器所处状态由第一状态变更为第二状态,向管理节点发送存储器状态变更报告,其中,所述存储器状态变更报告指示第一存储器处于第二状态。其中,所述第一状态和第二状态不同,所述第一状态和所述第二状态包括如下状态中的任意一个:离线状态、数据重构状态和信任访问状态。

其中,所述管理节点用于,在接收到了来自所述第一存储节点的存储器状态变更报告之后,将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态,并更新所述管理节点缓存的所述状态信息集合体的版本号;所述管理节点向所述 m 个存储节点中除所述管理节点之外的其他存储节点发送更新后的状态信息集合体。

相应的,所述第一存储节点还用于,用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。相应的,其他存储节点也用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

在本申请一些可能实施方式中,所述第一存储节点还用于,接收来自终端的第一写数据请求,所述第一写数据请求携带有第二数据和第二位置信息,所述第二位置信息用于描述所述第二数据在其所属数据体中的位置。

其中,所述第一存储节点还用于,基于所述第二位置信息确定写所述第二数据涉及的 W 个存储器;将第二数据切分为 $W-T$ 个数据条带,利用所述 $W-T$ 个数据条带计算得到 T 个校验条带。其中,所述 T 个校验条带和 $W-T$ 个数据条带形成包括 W 个条带的分条,所述 W 个存储器与所述 W 个条带之间一一对应,所述 T 和 W 为正整数且所述 T 小于所述 W 。

所述第一存储节点还用于,当所述第一存储节点缓存的状态信息集合体中当前记录所述 W 个存储器中的 W_1 个存储器处于非离线状态且所述 W 个存储器中的 W_2 个存储器处于离线状态,向所述 W_1 个存储器所属的 y_2 个存储节点发送携带所述分条的条带和所述状态信息集合体的标识的写数据请求。

相应的,所述 y_2 个存储节点中的每个存储节点用于,在接收到来自所述第一存储节点的写数据请求之后,将所述写数据请求携带状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较,在比较出所述写数据请求携带的状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下,向其包括的相应存储器中写入所述分

条的相应条带；在比较出所述写数据请求携带的状态信息集合体的标识与自身当前缓存的状态信息集合体的标识不相同的情况下，拒绝向其包括的相应存储器中写入所述分条的相应条带。

在本申请一些可能实施方式中，所述第一存储节点还用于，接收来自终端的第二写数据请求，所述第二写数据请求携带有第三数据和第三位置信息，所述第三位置信息用于描述所述第三数据在其所属数据体中的位置。

所述第一存储节点还用于，基于所述第三位置信息确定写所述第三数据涉及的W个存储器；将所述第三数据切分为W-T个数据条带，利用所述W-T个数据条带计算得到T个校验条带，其中，所述T个校验条带和W-T个数据条带形成包括W个条带的分条，所述W个存储器与所述W个条带之间一一对应，所述T和W为正整数且所述T小于所述W。

所述第一存储节点还用于，确定其缓存的状态信息集合体中当前记录的所述W个存储器所处状态，在确定出所述W个存储器中的W1个存储器处于非离线状态（例如并且所述W个存储器中的W2个存储器处于离线状态），并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带的写数据请求。

相应的，所述y2个存储节点中的每个1个存储节点用于，在接收到来自所述第一存储节点的写数据请求之后，向其包括的相应存储器中写入所述分条的相应条带。

在本申请一些可能实施方式中，所述第一存储节点还用于，接收来自终端的第三写数据请求，所述第三写数据请求携带有第四数据和第四位置信息，所述第四位置信息用于描述所述第四数据在其所属数据体中的位置；

所述第一存储节点还用于，基于所述第四位置信息确定写所述第四数据涉及的W个存储器；将所述第四数据切分为W-T个数据条带，利用所述W-T个数据条带计算得到T个校验条带，其中，所述T个校验条带和W-T个数据条带形成包括W个条带的分条，所述W个存储器与所述W个条带之间一一对应，所述T和W为正整数且所述T小于所述W。

所述第一存储节点还用于，当所述第一存储节点缓存的状态信息集合体中当前记录所述W个存储器中的W1个存储器处于非离线状态（例如并且所述W个存储器中的W2个存储器处于离线状态），并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带和用于发布状态信息集合体的管理节点的节点标识的写数据请求。

相应的，所述y2个存储节点中的每个存储节点用于，在接收到来自所述第一存储节点的写数据请求之后，将所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识进行比较，在比较出所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识相同的情况下，向其包括的相应存储器中写入所述分条的相应条带；在比较出所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识不相同的情况下，拒绝向其包括的相应存储器中写入所述分条的相应条带。

在本申请一些可能实施方式中，第一存储节点还用于生成第一重构日志，所述第一重构日志中记录了所述w2个存储器中的第二存储器的存储器标识，所述第一重构日志中还记

录了所述W个条带中的与所述第二存储器所对应的第一条带的条带标识,所述第一重构日志中还记录了所述分条的分条标识;其中,所述第二存储器为所述W2个存储器中的任意一个存储器。

相应的,例如当所述W2个存储器中的所述第二存储器重新上线之后,并且所述第二存储器所归属的第二存储节点收集到了在所述第二存储器离线期间产生的所述第一重构日志,第二存储节点获取第一重构日志中记录的需写入所述第二存储器的第一条带的标识;确定所述分条所包括的所述第一条带之外其他W-T个条带所在的W-T个存储器,从所述W-T个存储器中读取所述W-T个条带,利用所述W-T个条带进行校验运算以重构出需写入所述第二存储器的所述第一条带,将重构出的所述第一条带写入所述第二存储器。

第三方面,本申请实施例还提供一种存储节点,所述存储节点为分布式存储系统所包括的m个存储节点的其中1个存储节点,其中,每个所述存储节点包括至少1个存储器,每个所述存储器包含非易失性存储介质,所述m为大于1的整数。所述存储节点包括:通信单元和处理控制单元。

所述通信单元,用于接收来自终端的第一读数据请求,所述第一读数据请求携带有第一位置信息,所述第一位置信息用于描述第一数据在其所属数据体中的位置。

处理控制单元,用于在基于所述第一位置信息确定所述第一数据所在的存储器为的第一存储器,且所述第一存储器归属于所述存储节点的情况下,当状态信息集合体中当前记录所述第一存储器处于信任访问状态,从所述第一存储器读取所述第一数据。

其中,所述通信单元还用于,向所述终端发送用于响应所述第一读数据请求的第一读数据响应,所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中,所述通信单元还用于,在所述第一存储器归属于所述m个存储节点之中的第二存储节点,且所述第一存储器处于信任访问状态,且所述存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下,向所述第二存储节点转发所述第一读数据请求。

所述第一读数据请求用于触发所述第二存储节点在接收到了来自所述第一存储节点的所述第一读数据请求之后,从所述第一存储器读取所述第一数据,向所述存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应,所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中,所述通信单元还用于,在所述第一存储器归属于所述m个存储节点之中的第二存储节点,并且所述第一存储器处于信任访问状态的情况下,在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识,向所述第二存储节点发送添加了所述状态信息集合体的标识的所述第一读数据请求。

所述第一读数据请求用于触发所述第二存储节点接收到来自所述存储节点的所述第一读数据请求之后,将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较,在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下,从所述第一存储器读取所述第一数据,向所述存储节点或所述终端发送用于响应所述第一读数据请求的第一

读数据响应, 其中, 所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中, 所述 m 个存储节点中其中1个存储节点为管理节点。

所述通信单元还用于, 当所述第一存储器所处状态由第一状态变更为第二状态, 向管理节点发送存储器状态变更报告, 其中, 所述存储器状态变更报告指示第一存储器处于第二状态; 所述第一状态和第二状态不同, 所述第一状态和所述第二状态包括如下状态中的任意一个: 离线状态、数据重构状态和信任访问状态。其中, 所述存储器状态变更报告用于触发所述管理节点在接收到来自所述第一存储节点的存储器状态变更报告后, 将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态, 并且更新所述管理节点缓存的所述状态信息集合体的版本号。向所述 m 个存储节点中除所述管理节点之外的其他存储节点发送更新后的状态信息集合体。

所述处理控制单元用于, 采用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

第四方面, 本申请实施例提供了一种数据访问方法, 应用于第一存储节点, 所述第一存储节点位于包括 m 个存储节点的分布式存储系统中, 其中, 每个所述存储节点包括至少1个存储器, 每个所述存储器包含非易失性存储介质, 所述 m 为大于1的整数,

所述方法包括: 所述第一存储节点接收(来自终端)第一读数据请求, 其中, 所述第一读数据请求携带有第一位置信息, 所述第一位置信息用于描述第一数据在其所属数据体中的位置; 在所述第一存储节点基于所述第一位置信息确定所述第一数据所在的存储器为第一存储器, 且所述第一存储器归属于所述第一存储节点的情况下, 当状态信息集合体中当前记录所述第一存储器处于信任访问状态, 所述第一存储节点从所述第一存储器读取所述第一数据, 发送用于响应所述第一读数据请求的第一读数据响应, 其中, 所述第一读数据响应携带有读取到的所述第一数据。

第五方面, 本申请实施例提供了一种存储节点, 所述存储节点为分布式存储系统所包括的 m 个存储节点的其中1个存储节点, 每个所述存储节点包括至少1个存储器, 每个所述存储器包含非易失性存储介质, 所述 m 为大于1的整数,

所述存储节点包括相互耦合的处理器和通信接口; 所述处理器用于执行以上各方面中由第一存储节点(或其它存储节点)所执行方法的部分或全部步骤。

例如, 所述处理器用于, 通过所述通信接口接收第一读数据请求, 所述第一读数据请求携带有第一位置信息, 其中, 所述第一位置信息用于描述第一数据在其所属数据体中的位置; 在基于所述第一位置信息确定所述第一数据所在的存储器为第一存储器, 且所述第一存储器归属于所述第一存储节点的情况下, 当状态信息集合体中当前记录所述第一存储器处于信任访问状态, 从所述第一存储器读取所述第一数据, 通过所述通信接口发送用于响应所述第一读数据请求的第一读数据响应, 其中, 所述第一读数据响应携带有读取到的所述第一数据。

第六方面, 本申请实施例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了程序代码。所述程序代码包括用于执行第一方面中的任意一存储节点(例如第一存储节点或第二存储节点)所执行方法的部分或全部步骤的指令。

第七方面，本申请实施例提供了一种包括指令的计算机程序产品，当所述计算机程序产品在计算机（计算机例如为存储节点）上运行时，使得所述计算机执行以上各方面中由任意一存储节点（例如第一存储节点或第二存储节点等）所执行方法的部分或全部步骤。

第八方面，本申请实施例提供了一种存储节点，包括：相互耦合的处理器、通信接口和存储器；所述处理器用于执行以上各方面中由任意一存储节点所执行方法的部分或全部步骤。

第九方面，本申请实施例还提供了一种业务系统，可以包括：分布式存储服务系统和终端，分布式存储服务系统和终端之间通信连接；所述分布式存储服务系统为如权本申请实施例提供的任意一种分布式存储服务系统。

10 附图说明

- 图1-A和图1-B是本申请实施例举例的一些分布式存储系统的架构示意图；
图1-C是本申请实施例举例提供的存储器的状态迁移示意图；
图1-D是本申请实施例举例提供的状态信息集合体的组织形式示意图；
图1-E和图1-F是本申请实施例举例提供的一些分条的组织形式示意图；
15 图2-A和图2-B是本申请实施例举例提供的一些文件的组织形式示意图；
图3-A是本申请实施例举例提供的一种状态更新方法的流程示意图；
图3-B和图3-C是本申请实施例举例提供的一些状态信息集合体的组织形式示意图；
图4-A是本申请实施例举例提供的另一种状态更新方法的流程示意图；
图4-B和图4-C是本申请实施例举例提供的一些状态信息集合体的组织形式示意图；
20 图5-A是本申请实施例举例提供的另一种数据访问方法的流程示意图；
图5-B是本申请实施例举例提供的一种重构日志的组织形式示意图；
图6是本申请实施例举例提供的另一种数据访问方法的流程示意图；
图7-A是本申请实施例举例提供的一种数据重构方法的流程示意图；
图7-B和图7-C是本申请实施例举例提供的一些状态信息集合体的组织形式示意图；
25 图8是本申请实施例举例提供的另一种数据访问方法的流程示意图；
图9是本申请实施例举例提供的另一种数据访问方法的流程示意图；
图10是本申请实施例举例提供的另一种数据访问方法的流程示意图；
图11是本申请实施例举例提供的一种分布式存储系统的示意图；
图12是本申请实施例举例提供的一种存储节点的示意图；
30 图13是本申请实施例举例提供的另一种存储节点的示意图。

具体实施方式

本申请的说明书和权利要求书以及上述附图中的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包括。例如包括了一系列步骤或单元的过程、方法、系统或产品或设备没有限定于已列出的步骤或单元，而是可选地还可以包括没有列出的步骤或者单元，或者可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。另外来说，术语“第一”、“第二”、“第三”和“第四”等是用于区别不同对象，而不是用于描述特定

顺序。

下面首先结合附图介绍分布式存储系统的一些可能架构。

参见图1-A和图1-B，本申请实施例的一些技术方案可基于1-A或图1-B举例所示架构的分布式存储系统或其变形架构具体实施。分布式存储系统包括通过互联网络连接的多个存储节点（在某些场景下，存储节点也可称为“存储服务节点”）。终端可通过网络对分布式存储系统进行例如读写等访问。

可以理解的是，本申请的实施例的终端（terminal）的产品形态例如可以为移动互联网设备、笔记本电脑、服务器、平板电脑、掌上电脑、台式电脑、手机或其他产品形态的可发出读数据请求或写数据请求等数据访问请求的终端设备。在本申请各个实施例中，终端也可以被主机（host）代替。

其中，存储节点包括1个或多个存储器。本申请实施例中的存储器包含非易失性存储介质，包含非易失性存储介质的存储器也可称为非易失性存储器。本实施例中，存储器例如可为硬盘、可擦写光盘和固态硬盘等存储介质，图1-A和图1-B中是以存储器是硬盘为例进行介绍。

存储节点还可包括网卡、内存、处理器和扩展卡等。处理器例如通过扩展卡与存储器互联。存储节点可通过其网卡与其他存储节点或其他外部设备互联。网卡可看作是存储节点的一种通信接口。

其中，处理器中可包括若干个功能单元，这些功能单元可以是硬件（例如处理器/现场可编程门阵列（英文：field programmable gate array，缩写：FPGA）电路，或者处理器/FPGA电路与其他辅助硬件的组合），也可以是软件。例如图1-B举例所示架构中，处理器中可以运行例如客户端（英文：client）、数据服务（英文：data service，缩写：DS）单元和状态管理器（英文：monitor）等功能。在运行这些功能时，可以认为这些功能是处理器中的一个模块。在不运行这些功能时，实现这些功能的程序可以是位于内存中的程序，在这种情况下，也可以认为这些模块位于内存中。总的来说，这些功能是位于各个存储节点，具体而言，位于各个存储节点中处理器和内存组成的集合之中。为了方便描述，在图1-B中把这些模块画在了处理器中。在一些特殊情况下，例如在系统级芯片（system on chip，SoC）之中，处理器本身有存储功能，程序代码可以直接烧录在处理器中，这种情况下内存不再是必须的器件。

其中，client为相应存储节点的一种系统输入输出（英文：input output，缩写：IO）入口，主要负责接收其他设备（例如终端或其他存储节点）发送的访问请求，访问请求也称IO请求，访问请求例如为读数据请求、写数据请求等。DS单元主要负责接收client发送的访问请求，基于访问请求对本地存储器进行读写操作，并将读写操作结果返回给client。monitor可负责监控相应存储节点所包含存储器的状态，当然monitor也还可负责监控相应存储节点的状态等。

例如可在分布式存储系统中的每个存储节点上部署monitor，每个存储节点中的monitor则可用于监控这个存储节点包括的存储器的状态等信息。进一步的，可将所有存储节点的monitor连接在一起形成monitor集群。并且通过管理员指定的方式或者通过选举算法，从

monitor集群中选举出1个“主monitor”，其他monitor则为“从monitor”。主monitor所在的存储节点可称“管理节点”。例如每个从monitor可将自身所在存储节点所包括的各存储器当前所处状态发送给主monitor。主monitor则可基于收集到的各存储节点所包括的存储器所处的状态等信息生成状态信息集合体。其中，状态信息集合体例如可简称“节点地图（英文：node map）”或“状态文件”等。状态信息集合体中记录了各存储器的状态等信息。进一步的，例如从monitor可将自身所在存储节点当前所处状态发送给主monitor。状态信息集合体还可以进一步记录各存储节点的状态等。如前所述，monitor是处理器通过运行内存中的程序所实现的功能。总的来说，可以认为monitor是处理器+内存所具有的功能；或者说monitor是存储节点所具有的功能。因此，下面的主monitor的功能可以由主处理器执行，或者主处理器+处理器+内存执行，或者主存储节点执行；从monitor的功能可以由从处理器执行，或者从处理器+处理器+内存执行，或者从存储节点执行。

其中，主monitor可主动的将最新的状态信息集合体推送给其他存储节点，或者其他存储节点也可主动的从主monitor请求最新的状态信息集合体。其他的存储节点上的monitor（即从monitor）可将接收到的状态信息集合体缓存在其内存之中，缓存在内存中的这个状态信息集合体便为相应存储节点（相应monitor）本地维护的状态信息集合体。每当从monitor接收到来自主monitor的状态信息集合体，那么，这个从monitor便可用最新接收到的来自主monitor的这个状态信息集合体，来更新其当前缓存的状态信息集合体。即从monitor保持当前缓存的状态信息集合体是其接收到的最新状态信息集合体。

可以理解的是，当从monitor与主monitor之间处于连接态，那么从monitor当前缓存的状态信息集合体通常就会是主monitor所发布的最新的状态信息集合体；而当从monitor与主monitor之间未处于连接态，那么从monitor有可能遗漏接收主monitor所发布的最新的状态信息集合体，因此从monitor当前缓存的状态信息集合体可能就不是主monitor所发布的最新的状态信息集合体。当然，如果在从monitor与主monitor之间未处于连接态的期间，主monitor并未发布新的状态信息集合体，那么从monitor当前缓存的状态信息集合体也有可能还是主monitor所发布的最新的状态信息集合体。处于连接态的设备之间可以直接通信或者间接的通信。出于这样的原因，和管理节点处于连接态的存储节点可以从管理节点获得最新版本的状态信息集合体。

其中，存储器的状态（访问状态）包括在线（英文：online）状态和离线（英文：offline）状态。其中，在线状态包括：数据重构状态和信任访问状态。其中，数据重构状态也可称追赶（英文：catchup）状态；信任访问状态也可称正常（英文：normal）状态。

可以理解的是，在需进行数据重构的情况下，存储器在数据重构完成后可进入信任访问状态，或者在一些无需进行数据重构的情况下，例如，存储器可直接从离线状态进入信任访问状态。总的来说，离线状态和数据重构状态都可看作是非信任访问状态。

存储器的状态反映出存储器中数据的有效性（也即是否可信任）。存储器处于信任访问状态，则表示分配给这个存储器的数据被全部成功写入这个存储器了。也就是说，被分配的数据被全部写入其中的这个存储器处于信任访问状态。例如，当存储器处于信任访问状态，那可认为处于这种状态的存储器上的数据是有效的。例如，当存储器处于数据重构

状态，那么表示存储器此时正在对离线期间丢失的相关数据进行重构，那么可认为这种情况下存储器上的数据很可能不是有效的，也就是说存储器处于非信任访问状态。当存储器处于离线状态，那么可认为这种情况下存储器上的数据不是有效的，也就是说存储器处于非信任访问状态。可以理解的是，处于离线状态的存储器无法被访问，不可从存储器中读数据也不可从存储器中写数据。

参见图1-C，图1-C举例了存储器的不同状态之间的可能变迁方式，存储器可从在线状态迁移到离线状态，也可从离线状态迁移到在线状态。具体例如，存储器可从离线状态变更为数据重构状态或信任访问状态，可从数据重构状态变更为信任访问状态，可从数据重构状态或信任访问状态变更为离线状态。

存储器offline可能由多种因素引发，例如，当存储节点中的存储器无法被这个存储节点检测到，或当存储节点接收到了来自某存储器的故障状态报告（故障状态报告指示存储器offline），那么存储节点可将相应存储器的状态设置为offline状态。又例如当管理节点（主monitor）连续多次接收不到某个存储节点的心跳消息，这就表示这个存储节点处于offline状态，那么管理节点可以将这个存储节点包括的所有存储器设定为offline状态。

其中，状态信息集合体中可记录各存储器的存储器ID和状态，还可进一步记录各存储器所属存储节点的节点ID和状态。参见图1-D，图1-D举例了状态信息集合体的一种可能表现形式，状态信息集合体中包括状态信息集合体的版本号、各存储器（图中以存储器为硬盘为例）的标识和状态（状态例如为离线状态、数据重构状态或信任访问状态）。

由于存储器的状态可能是随着时间变化而变化，因此状态信息集合体也会因存储器状态变化而适应性更新。例如当管理节点（主monitor）接收到存储器状态变更报告（存储器状态发生变化，例如存储器被拔出下电或者重新上电）后或判定存储节点离线（例如在一定时长内未收到相应存储节点的心跳消息），那么管理节点认为状态信息集合体更新条件满足，管理节点进行状态信息集合体更新，而更新后的状态信息集合体对应1个新的版本号，更新后的状态信息集合体中记录了变更后的存储器状态。管理节点将更新后的状态信息集合体推送给其他存储节点（从monitor）。状态信息集合体的版本号可由主monitor基于全局变量生成。具体例如，状态信息集合体每更新1次，版本号则在当前基础上递增+1或+2等。因此，不同版本号可以表示不同的状态信息集合体。通过比较两个状态信息集合体的版本号，便可判定这两个状态信息集合体是否相同。

分布式存储系统的存储器中可能存储各种数据体（数据体是指可以被存储器所存储的数据，数据体例如为文件、对象或记录或其他格式的数据）。一般来说，考虑到系统可靠性等原因，同1个数据体的不同条带分别存储至不同存储器上。其中，数据条带进行冗余校验运算（如EC运算等）可以得到校验条带。校验冗余配置（例如EC配置）用于描述分条中的数据条带和校验条带的比例关系。例如文件的校验冗余配置可表示为A:T，那么，A为分条包括的数据条带的数量，T（T表示冗余度）则为分条包括的校验条带的数量。例如假设校验冗余算法为EC算法，那么当校验冗余配置（即EC配置）为2:1，那么2为分条（EC分条）包括的数据条带的数量，1为分条包括的校验条带数量，即，在这种情况下，分条包括2个数据条带和1个校验条带，分条包括的2个数据条带和1个校验条带将分别存储至不同的存

存储器。也就是说需要3个存储器来存储分条包括的2个数据条带和1个校验条。又假设当校验冗余配置（即EC配置）为16:4，那么16为分条（EC分条）包括的数据条带的数量，4为分条包括的校验条带的数量，即这种情况下，分条包括16个数据条带和4个校验条带，这种情况下则需20个存储器来分别存储分条包括的这20个条带。其他情况以此类推。图1-E举例示出了1种包括2个数据条带和1个校验条带的分条。图1-F示出了1种包括多个数据条带和多个校验条带的分条。

进一步的，一个分布式存储系统可支持一种或多种校验冗余配置，每种校验冗余配置可对应一个或多个校验成员组，其中，每个校验成员组包括多个存储器。当然，不同的校验成员组所包括的存储器有可能完全不同；也有可能部分相同，也就是说每个存储器可以作为一个或多个校验成员组服务。例如校验冗余配置为2:1，那么这个校验冗余配置对应的每个校验成员组都包括3个（ $2+1=3$ ）存储器。例如校验冗余配置为4:2，那么这个校验冗余配置对应的每个校验成员组都包括6个（ $4+2=6$ ）存储器。例如校验冗余配置为16:4，那么这个校验冗余配置对应的每个校验成员组都包括20个（ $16+4=20$ ）存储器，以此类推。

此外，当某些文件长度很大时（如文件长度超过1Gb），那么有时为便于管理，可能将文件划分为若干文件分段，例如文件分段不超过500M或1Gb等。参见图2-A，图2-A为将文件划分为若干文件分段的示意图。文件的每个文件分段可对应相同或不同的校验冗余配置，不同文件分段的校验成员组也可能相同或者不同。在这种情况下，文件分段可被当作文件来对待的。例如参见图2-B，图2-B为未将文件划分为若干文件分段的示意图。

下面通过一些实施例对本申请的一些相关方法进行描述。

实施例一

下面先结合附图举例介绍一种状态更新方法，其中，状态更新方法应用于分布式存储系统。例如分布式存储系统包括m个存储节点，其中，所述m个存储节点例如可以包括存储节点Nd2、Nd1、Nd3，……，Nd4等，例如存储节点Nd1为管理节点，分布式存储系统包括更多数量或更少数量的存储节点的情况下的实施方式可以此类推。下面举例的方案主要由非管理节点来主动触发管理节点更新和发布状态信息集合体。

请参见图3-A，图3-A是本申请一实施例提供的一种状态更新方法的流程示意图。参见图3-A，本申请一实施例提供的一种状态更新方法可以包括：

301、当存储节点Nd2中的存储器Nd2-D1的状态发生变化，那么存储节点Nd2向管理节点Nd1发送存储器状态变更报告Report1。存储器状态变更报告Report1用于指示出存储器Nd2-D1变更后的最新状态。

其中，存储器状态变更报告Report1中可以携带存储器Nd2-D1的存储器标识和状态标识等等。举例来说，当存储器Nd2-D1变更后的最新状态为offline状态，那么存储器状态变更报告Report1携带的存储器Nd2-D1的状态标识指示出offline状态。又例如，当存储器Nd2-D1变更后的最新状态为数据重构状态，那么存储器状态变更报告Report1携带的存储器Nd2-D1的状态标识指示出数据重构状态。又举例来说，当存储器Nd2-D1变更后的最新状态为信任访问状态，那么存储器状态变更报告Report1携带的存储器Nd2-D1的状态标识指示出信任访问状态。以此类推。

302、管理节点Nd1接收来自存储节点Nd2的所述存储器状态变更报告Report1。管理节点Nd1基于所述存储器状态变更报告Report1更新自身当前缓存的状态信息集合体。

具体例如，所述管理节点Nd1将其当前缓存的状态信息集合体中记录的存储器Nd2-D1的状态，更新为存储器状态变更报告Report1所指示出的存储器Nd2-D1的状态，并且将这个状态信息集合体中的版本号（即这个状态信息集合体自身的版本号）更新。

举例来说，在各存储节点中都部署了monitor的情况下，那么具体可由位于存储节点Nd2当中的从monitor向存储节点Nd1发送存储器状态变更报告Report1。相应的可以由位于管理节点Nd1当中的主monitor接收来自存储节点Nd2的存储器状态变更报告Report1。

管理节点Nd1更新前的状态信息集合体例如图3-B举例所示，管理节点Nd1更新后的状态信息集合体例如图3-C举例所示，当然实际应用中并不限于这样的举例形式。

303、管理节点Nd1向其它的存储节点发送更新后的最新的状态信息集合体。

相应的，存储节点Nd2、Nd3和Nd4等分别接收来自管理节点Nd1的更新后的状态信息集合体。存储节点Nd2、Nd3和Nd4分别利用来自管理节点Nd1的状态信息集合体更新各自当前缓存的状态信息集合体。这样有利于各存储节点（Nd1、Nd2、Nd3和Nd4等等）所缓存的状态信息集合体就可保持相对同步。

上述方案提供了一种主要由非管理节点来主动的触发管理节点更新和发布状态信息集合体，具体的，当某存储节点中的存储器的状态发生变化，那么这个存储节点主动向管理节点上报存储器状态变更报告，进而触发管理节点基于这个存储器状态变更报告来及时更新和发布状态信息集合体。这种机制有利于尽量使得分布式存储系统中的各存储节点同步状态信息集合体的机制，有利于使得各存储节点可以相对准确的知晓分布式存储系统中的各存储器的状态，为后续基于存储器的状态进行读写操作奠定了一定基础。

实施例二

下面结合附图举例介绍另一种状态更新方法，其中，状态更新方法应用于分布式存储系统。例如分布式存储系统可包括m个存储节点，其中，m个存储节点例如可包括存储节点Nd2、Nd1、Nd3，……，Nd4，例如存储节点Nd1为管理节点，分布式存储系统包括更多数量或更少数量的存储节点的情况下的实施方式可以此类推。下面举例的方案主要由管理节点基于对其他存储节点的心跳监控结果来更新和发布状态信息集合体。

请参见图4-A，图4-A是本申请的另一个实施例所提供的另一种状态更新方法的流程示意图。参见图4-A，本申请的另一个实施例提供的另一种状态更新方法可以包括：

401、管理节点Nd1接收各存储节点周期性发送的心跳消息。

402、当管理节点Nd1在设定时长范围内（其中，设定时长范围例如可为5分钟、10分钟、20分钟、1分钟或其他时长）未接收到存储节点Nd2发送的心跳消息，那么可认定存储节点Nd2处于离线状态。管理节点Nd1在认定存储节点Nd2处于离线状态的情况下，更新管理节点Nd1当前缓存的状态信息集合体。

具体的，管理节点Nd1将其当前缓存的状态信息集合体中记录的存储节点Nd2所包括的所有存储器的状态均更新为离线状态，并且将这个状态信息集合体中的集合体版本号更新。

举例来说，在各存储节点中都部署了monitor的情况下，那么具体可由管理节点Nd1中

的主monitor接收各存储节点中的从monitor周期性发送的心跳消息。

例如存储节点Nd2包括存储器Nd2-D1、Nd2-D2和Nd2-D3。那么管理节点Nd1更新前的状态信息集合体例如图4-B举例所示，而管理节点Nd1更新后的状态信息集合体例如图4-C举例所示，当然实际应用中并不限于这样的举例形式。

5 403、管理节点Nd1向其它的存储节点发送更新后的最新的状态信息集合体。

相应的，所述存储节点Nd3和Nd4等分别接收来自管理节点Nd1的更新后的状态信息集合体。所述存储节点Nd3和Nd4分别利用来自所述管理节点Nd1的状态信息集合体更新掉各自当前已缓存的状态信息集合体。这样管理节点Nd1与其他各存储节点（例如Nd3和Nd4）所缓存的状态信息集合体就可保持相对同步，而存储节点Nd2由于处于离线状态，因此就
10 无法接收到管理节点Nd1发布的最新状态信息集合体，因此存储节点Nd2在离线期间就无法保持与管理节点直接的状态信息集合体同步。当存储节点Nd2重新上线之后，存储节点Nd2可以主动向管理节点请求最新的状态信息集合体来对其缓存的状态信息集合体进行更新，当然存储节点Nd2也可被动等待管理节点再次发布最新的状态信息集合体，利用管理节点再次发布的最新状态信息集合体来对其缓存的状态信息集合体进行更新。

15 上述方案提供了一种主要由管理节点基于对其他存储节点的心跳监控结果来更新和发布状态信息集合体的机制，这种机制有利于尽量使得分布式存储系统中的各存储节点同步状态信息集合体，进而使得各存储节点可以相对准确的知晓分布式存储系统中的各存储器的状态，为后续基于存储器的状态进行读写操作奠定了一定基础。

可以理解，在一些可能实施方式中，图3-A和图4-A所描述的状态更新方法可能均被引
20 用于同一分布式存储系统。也就是说，管理节点基于对其他存储节点的心跳监控结果来更新和发布状态信息集合体，同时，非管理节点也可主动的触发管理节点更新和发布状态信息集合体。

实施例三

下面再结合附图举例介绍一种数据访问方法，主要针对终端从分布式存储系统读数据
25 的一种可能场景，这种数据访问方法应用于分布式存储系统。例如分布式存储系统包括m个存储节点，m个存储节点包括存储节点Nd1和Nd2。存储器Nd2-D1、存储器Nd2-D2和存储器Nd2-D3均位于存储节点Nd1之中。其中，本实施例中主要是以相应校验冗余配置具体为2:1为例来说明的，即分条包括2个数据条带和1个校验条带，其他校验冗余配置的情况可以此类推。

30 参见图5-A，图5-A是本申请一实施例提供的一种数据访问方法的流程示意图。如图5-A举例所示，本申请一实施例提供的一种数据访问方法可以包括：

501、当终端需向分布式存储系统写入数据，终端向分布式存储系统发送携带有待写入数据的写数据请求，为了下面便于引述，本实施例中把这个写数据请求命名为Wq3。本步骤中，存储节点Nd1接收来自终端的写数据请求Wq3。

35 写数据请求Wq3可携带数据Data1（数据Data1为待写入数据）、数据Data1所属文件的文件标识和数据Data1的位置信息。其中，数据Data1的位置信息用于描述数据Data1在数据Data1所属文件中的位置。数据Data1的位置信息例如包括数据Data1的文件偏移地址（数据

Data1的文件偏移地址表示了数据Data1在文件中的起始位置)和数据Data1的长度等。

502、存储节点Nd1将数据Data1切分为2个数据条带(为了下面便于引述,本实施例中把这两个数据条带命名为数据条带Pd1和数据条带Pd2)。存储节点Nd1利用所述2个数据条带计算得到1个校验条带(为了下面便于引述,本实施例之中把这个校验条带命名为校验条带Pj1),所述1个校验条带和2个数据条带形成包括3个条带的1个分条。例如数据条带Pd1
5 和数据条带Pd2的长度可相等或不等。

503、存储节点Nd1基于数据Data1的位置信息确定写数据Data1所涉及的存储器。

例如存储节点Nd1确定出写数据Data1涉及的存储器为存储器Nd2-D1、存储器Nd2-D2
和存储器Nd2-D3。

10 一般来说,在为文件数据分配存储器时,通常是基于文件数据在文件中的位置来确定这段数据文件将具体分配给那些存储器存储。因此,基于数据Data1的位置信息可确定写数据Data1所涉及的存储器。存储节点

上述3个存储器(Nd2-D1、Nd2-D2和Nd2-D3)与所述3个条带之间一一对应。具体例如校验条带Pj1对应存储器Nd2-D3,数据条带Pd1对应存储器Nd2-D1,数据条带Pd2对应存
15 储器Nd2-D2。

504、存储节点Nd1确定其本地所缓存的状态信息集合体中当前所记录的所述存储器Nd2-D1、存储器Nd2-D2和存储器Nd2-D3的状态。

其中,下面以存储节点Nd1本地缓存的状态信息集合体中当前记录存储器Nd2-D2和Nd2-D3处于在线状态(信任访问状态或数据重构状态),而存储器Nd2-D1处于离线状态的情况为例来进行说明。
20

505、当确定出存储器Nd2-D2和存储器Nd2-D3处于在线状态且所述存储器Nd2-D1处于离线状态,那么,存储节点Nd2将数据条带Pd2写入存储器Nd2-D2,存储节点Nd2将校验条带Pj1写入存储器Nd2-D3。

本步骤中,之所以在存在离线存储节点情况下,仍然对余下的非离线存储节点正常写入数据,是由于:按照EC校验算法,即使有一定数量的条带没有成功写入,也可以通过写入成功的条带把写入不成功的条带重构恢复出来,并不会造成数据丢失。只要不成功写入的条带的数量不超过校验条带的数量,EC校验算法就可以正常运算。当然,对于其他校验算法也是类似的。
25

506、由于存储节点Nd2中的存储器Nd2-D1处于离线状态,因此数据条带Pd1写入存储器Nd2-D1失败,因此存储节点Nd1或存储节点Nd2生成重构日志log1。
30

其中,重构日志log1中记录写入失败的条带(数据条带Pd1)的条带标识、写入失败的存储器(存储器Nd2-D1)的存储器标识。重构日志log1中还可记录写入失败的条带(数据条带Pd1)所属分条的分条标识等。所述重构日志log1用于后续当存储器Nd2-D1重新上线之后重构数据条带Pd1。图5-B举例示出了一种重构日志log1的可能表现形式,当然也可能采用其他表现形式。
35

可以理解,如果存储器Nd2-D1也处于信任访问状态,那么存储节点Nd2可直接将数据条带Pd1写入存储器Nd2-D1。在这种情况下,存储节点Nd1或存储节点Nd2就无需执行生成

重构日志log1的相关步骤了。

在上述方案主要针对的写数据场景，是分条中的各条带所对应存储器均位于同一个存储节点的一种可能场景，在这种可能场景下，由于在分布式存储系统中，引入了用于记录存储器状态的状态信息集合体，故而可以利用状态信息集合体来记录管理分布式存储系统的存储器的访问状态。对于处于在线状态的存储器，可直接向其中写入数据；对于处于离线状态的存储器，无需在尝试写入失败后再生成相关重构日志，而可以直接生成相关重构日志。在本实施例例举的场景下，现有技术是先尝试数据写入，在尝试写入失败之后再生成相关重构日志。本实施例方案由于可降低无效尝试，进而提升数据写入效率，进而有利于提升系统性能。

实施例四

下面再结合附图举例介绍另一种数据访问方法，主要针对终端从分布式存储系统读数据的另一可能场景。在上一个实施例中，条带涉及的存储器位于同一个存储节点；在本实施例的场景中，这些存储器位于不同存储节点。

例如分布式存储系统包括m个存储节点，m个存储节点包括存储节点Nd2、Nd1、Nd3和Nd4等。存储器Nd2-D1属于存储节点Nd2，存储器Nd3-D1属于存储节点Nd3，存储器Nd4-D1属于存储节点Nd4。本实施例中主要以校验冗余配置据具体为2:1为例，即分条包括2个数据条带和1个校验条带，其他校验冗余配置的情况可以此类推。

请参见图6，图6是本申请另一实施例提供的另一种数据访问方法的流程示意图。如图6举例所示，本申请另一个实施例提供的另一种数据访问方法可以包括：

601~602，其中，步骤601~602与步骤501~502相同，因此，相关描述可以相应参考步骤501~502，此处不再赘述。

603、存储节点Nd1基于数据Data1的位置信息确定写数据Data1所涉及的存储器。

例如存储节点Nd1确定出写数据Data1涉及的存储器为存储器Nd2-D1、存储器Nd3-D1和存储器Nd4-D1。其中，上述3个存储器（Nd2-D1、Nd3-D1和Nd4-D1）与所述3个条带之间一一对应。具体例如，校验条带Pj1对应存储器Nd4-D1，数据条带Pd1对应存储器Nd2-D1，数据条带Pd2对应存储器Nd3-D1。

604、存储节点Nd1确定其本地所缓存的状态信息集合体之中当前所记录的存储器Nd2-D1、存储器Nd3-D1和存储器Nd4-D1的状态。其中，下面以存储节点Nd1本地缓存的状态信息集合体中当前记录的是：存储器Nd3-D1和Nd4-D1处于在线状态，而存储器Nd2-D1处于离线状态的情况为例来进行说明。

具体的，存储节点Nd1可在其与管理节点之间处于连接态的情况下，确定其本地所缓存的状态信息集合体之中当前所记录的存储器Nd2-D1、存储器Nd3-D1和存储器Nd4-D1的状态。

605、当确定出存储器Nd3-D1和存储器Nd4-D1处于在线状态且所述存储器Nd2-D1处于离线状态，存储节点Nd1向所述存储器Nd3-D1所属的存储节点Nd3发送写数据请求Wq1，并且，存储节点Nd1向所述存储器Nd4-D1所属的存储节点Nd4发送写数据请求Wq2。

例如写数据请求Wq1携带数据条带Pd2和存储节点Nd1当前本地缓存的状态信息集合

体的版本号。进一步的，写数据请求Wq1还可以携带数据条带Pd2的长度（长度例如可为110Kb）、数据条带Pd2的文件偏移地址、数据条带Pd2所属分条的分条标识、数据条带Pd2的条带标识、数据条带Pd2所属文件的文件标识等等。

又例如，写数据请求Wq2可携带校验条带Pj1和存储节点Nd1当前本地缓存的状态信息集合体的版本号。进一步的，写数据请求Wq2还可携带所述校验条带Pj1的长度、校验条带Pj1所属分条的分条标识、校验条带Pj1的条带标识等等。

可以理解，存储节点Nd1将自身缓存的状态信息集合体的版本号携带在发送给其他存储节点的写数据请求中，可以便于其他存储节点通过版本号比较来验证存储节点Nd1缓存的状态信息集合体是否有效。

606、存储节点Nd3在接收到来自存储节点Nd1的写数据请求Wq1之后，将写数据请求Wq1携带的状态信息集合体的版本号，与存储节点Nd3自身当前缓存的状态信息集合体的版本号进行比较。

在比较出写数据请求Wq1携带的状态信息集合体的版本号与存储节点Nd3自身当前缓存的状态信息集合体的版本号相同的情况下，并且存储节点Nd3与管理节点之间处于连接态的情况下（即存储节点Nd3中的从monitor与管理节点中的主monitor连接正常），而存储节点Nd3中的从monitor与管理节点中的主monitor连接正常，那么就可认为存储节点Nd3中的从monitor可以正常及时的接收到管理节点中的主monitor所发布的最新版本的状态信息集合体。因此这种情况下可认为存储节点Nd3当前缓存的状态信息集合体是主monitor维护的最新版的状态信息集合体。而写数据请求Wq1携带的状态信息集合体的版本号与存储节点Nd3自身当前缓存的状态信息集合体的版本号相同，也就说明，存储节点Nd1当前缓存的状态信息集合体也是主monitor维护的最新版的状态信息集合体，那么存储节点Nd1当前缓存的状态信息集合体中记录的存储器Nd3-D1的状态（信任访问状态）也应该是符合实际情况的），存储节点Nd3将数据条带Pd2写入存储器Nd3-D1。

此外，在比较出写数据请求Wq1携带的状态信息集合体的版本号与自身当前缓存的状态信息集合体的版本号不同的情况下，可认为Nd1所缓存的Nd3-D1的状态并不准确，换句话说，虽然在Nd1中所记录的Nd3-D1的状态是信任访问状态，但是并不能确保Nd3-D1的真实状态是信任访问状态。由于Nd3-D1可能处于不信任的访问状态，因此存储节点Nd3可拒绝将数据条带Pd2写入存储器Nd3-D1。

特别说明的是，当存储节点与管理节点之间处于连接态，那么这个存储节点缓存的状态信息集合体是管理节点最新发布的状态信息集合体的可能性相对很大。然而，在一些比较特殊的情况下，也有可能即使存储节点与管理节点之间处于连接态，但如果在管理节点发布的状态信息集合体那个时刻恰好存储节点出现了短暂性故障，或在管理节点发布的状态信息集合体的那个时刻，恰好管理节点与存储节点直接的通信质量极差，那么就可能导致存储节点未能获得管理节点最新发布的状态信息集合体，也就是说，即使存储节点与管理节点之间处于连接态，也有很小的可能出现：存储节点缓存的状态信息集合体不是管理节点最新发布的状态信息集合体。因此，在存储节点与管理节点之间处于连接态的情况下进一步比较各自当前缓存的状态信息集合体的版本号，如果版本号相同，可认为各自当前

缓存的状态信息集合体有效。这种做法可以进一步提高可靠性。

进一步的，存储节点Nd3向存储节点Nd1发送用于响应写数据请求Wq1的写数据响应Wr1，写数据响应Wr1携带数据条带Pd2的写操作结果。具体的，当存储节点Nd3成功将数据条带Pd2写入存储器Nd3-D1，那么写操作结果为写入成功。当存储节点Nd2未成功将数据条带Pd2写入存储器Nd3-D1（例如存储节点Nd2拒绝将数据条带写入存储器Nd3-D1），那么相应的写操作结果为写入失败。

607、存储节点Nd4在接收到来自存储节点Nd1的写数据请求Wq2之后，进行和步骤606类似的状态信息集合体的版本号判断（区别在于：存储节点Nd3变为存储节点Nd4，写数据请求Wq1变为写请求Wq1），以确定是否把校验条带Pj1写入存储器Nd4-D1。步骤607的具体执行过程此处不再赘述。

步骤606和步骤607可同时执行，也可以任意一个先执行。

608、存储节点Nd1接收来自所述存储节点Nd3的所述写数据响应Wr1。所述存储节点Nd1接收来自所述存储节点Nd4的所述写数据响应Wr2。

进一步的，若写数据响应指示相应写入失败，那么存储节点Nd1可以重新发送相应写数据请求来请求重新写入。例如，如果写数据响应Wr1指示相应写入失败，那么存储节点Nd1可以重新向存储节点Nd3发送写数据请求Wq1。

609、由于存储节点Nd2中的存储器Nd2-D1处于离线状态，数据条带Pd1写入存储器Nd2-D1失败，因此存储节点Nd1生成重构日志log1，所述存储节点Nd1向成功写入条带的存储节点（例如所述存储节点Nd3和/或存储节点Nd4）发送重构日志log1。相应的，存储节点Nd3和/或存储节点Nd4可接收并且存储所述重构日志log1，所述重构日志log1用于后续当存储器Nd2-D1重新上线之后重构数据条带Pd1。

其中，重构日志log1中记录写入失败的条带（数据条带Pd1）的条带标识、写入失败的存储器（存储器Nd2-D1）的存储器标识。重构日志log1还可记录写入失败的条带（数据条带Pd1）所属分条的分条标识等。

上述方案中主要针对的写数据场景，是分条中的各条带所对应存储器并未位于同一个存储节点的一种可能场景，在这种可能的场景下，由于在分布式存储系统中，引入了用于记录存储器状态的状态信息集合体，故而可利用状态信息集合体来记录管理分布式存储系统的存储器的访问状态。

具体来说，本实施例介绍了对待写数据的写入方法，待写数据所对应分条中的各条带所对应存储器并未位于同一个存储节点，当存储节点Nd1与管理节点处于连接态，且存储节点Nd1缓存的状态信息集合体中记录存储器Nd3-D1和Nd4-D1在线，那么存储节点Nd1初步认为可向存储器Nd3-D1和Nd4-D1成功写入，为了验证存储节点Nd1缓存的状态信息集合体的有效性，存储节点Nd1将自身缓存的状态信息集合体的版本号携带在发送给其他存储节点（存储节点Nd3和存储节点Nd4）的写数据请求中，其他存储节点将写数据请求携带的状态信息集合体的版本号和自身缓存的状态信息集合体的版本号进行比较，当写数据请求携带的状态信息集合体的版本号与自身当前缓存的状态信息集合体的版本号相同的情况下，可认为存储节点Nd1所缓存的状态信息集合体有效，因此这个状态信息集合体中记录的相关存储

器的状态应当准确，这个情况下进行相应数据写入通常成功，因此就直接向相关存储器写入数据。反之，可认为存储节点Nd1所缓存的状态信息集合体可能是无效的，因此这个状态信息集合体中记录的相关存储器的状态很可能不准确，这个情况下进行相应数据写入很可能失败，因此这个时候拒绝写入是相对合理做法。

- 5 本实施例和前一实施例相似的效果是：对于处于离线状态的存储器，可直接生成相关重构日志。重构日志无需按照传统方式在尝试写入失败之后再生成相关重构日志，进而有利于提升数据写入效率，进而有利于提升系统性能。

实施例五

10 下面再结合相关附图举例介绍一种数据重构方法，这种数据重构方法应用于分布式存储系统。例如分布式存储系统可包括m个存储节点，其中，所述m个存储节点包括存储节点Nd1、Nd2、Nd3和Nd4等。存储节点Nd1为发布状态信息集合体的管理节点（即为主monitor所在的存储节点）。其中，例如存储器Nd2-D1属于存储节点Nd2，存储器Nd3-D1属于存储节点Nd3，存储器Nd4-D1属于存储节点Nd4。这种数据重构方法例如可执行于图6-A举例所示方案之后。

- 15 参见图7-A，图7-A是本申请一实施例提供的一种数据重构方法的流程示意图。如图7-A举例所示，本申请的一实施例提供的一种数据重构方法可包括：

701、当存储节点Nd2确定存储器Nd2-D1上线之后，存储节点Nd2可向管理节点Nd1发送存储器状态变更报告P2，所述存储器状态变更报告P2可指示存储器Nd2-D1的状态由离线变更为数据重构状态。具体例如，存储器状态变更报告P2中可携带存储器Nd2-D1的存储器标识和状态标识（状态标识为数据重构状态的状态标识，即状态标识表示的状态为数据重构状态）。

20

702、管理节点Nd1接收来自存储节点Nd2的所述存储器状态变更报告P2，所述管理节点Nd1更新缓存的状态信息集合体。具体的，所述存储节点Nd1将其缓存的状态信息集合体中记录的存储器Nd2-D1的状态，更新为存储器状态变更报告P2所指示出的存储器Nd2-D1的状态，并且将这个状态信息集合体记录的状态信息集合体的版本号更新。

25

举例来说，在各存储节点中都部署了monitor的情况下，那么具体可以由存储节点Nd2中的从monitor向管理节点Nd1发送存储器状态变更报告P2，相应的可由所述管理节点Nd1中的主monitor接收来自存储节点Nd2的所述存储器状态变更报告P2。

30 举例来说，管理节点Nd1更新前的状态信息集合体例如图3-C举例所示，管理节点Nd1更新后的状态信息集合体例如图7-B举例所示，当然实际应用中，状态信息集合体并不限于这样的举例形式。

703、存储节点Nd1向其它存储节点发送更新后的状态信息集合体。

相应的，存储节点Nd2、Nd3和Nd4分别接收来自管理节点Nd1的更新后的状态信息集合体，所述存储节点Nd2、Nd3和Nd4用来自所述管理节点Nd1的状态信息集合体更新掉其当前缓存的状态信息集合体。这样，各个存储节点（Nd1、Nd2、Nd3和Nd4）所维护的状态信息集合体就可保持相对同步。

35

704、存储节点Nd1通知存储节点Nd2针对存储器Nd2-D1进行数据重构。

可以理解,管理节点Nd1可以通过发送更新的状态信息集合体来通知存储节点Nd2针对存储器Nd2-D1进行数据重构。或者,管理节点Nd1可通过其他消息(例如数据重构通知消息)来通知存储节点Nd2针对存储器Nd2-D1进行数据重构。

705、存储节点Nd2在接收到来自管理节点Nd1的针对存储器Nd2-D1进行数据重构的通知后,所述存储节点Nd2向所述存储节点Nd3和/或Nd4收集存储器Nd2-D1离线期间的相关重构日志。

706、存储节点Nd2基于收集到的存储器Nd2-D1离线期间的相关重构日志重构数据。

具体例如,存储节点Nd2收集到了存储器Nd2-D1离线期间的重构日志log1,重构日志log1记录了存储器Nd2-D1离线期间写入失败的条带(数据条带Pd1)的条带标识、写入失败的存储器(存储器Nd2-D1)的存储器标识。并且,所述重构日志log1还可以记录写入失败的条带(数据条带Pd1)所属分条的分条标识等。

其中,存储节点Nd2可以基于数据条带Pd1所属分条的分条标识,确定数据条带Pd1所属分条,进而获取数据条带Pd1所属分条的其它条带(数据条带Pd2和校验条带Pj1),存储节点Nd2基于数据条带Pd2和校验条带Pj1进行冗余校验运算以重构出数据条带Pd1,存储节点Nd2将重构出的数据条带Pd1写入存储器Nd2-D1。

可以理解,当存储节点Nd2收集到存储器Nd2-D1离线期间的多个相关重构日志,那么所述存储节点Nd2可先对收集到的这多个重构日志进行去重处理,而后基于去重后的各个重构日志分别进行数据重构。基于重构日志进行数据重构的方式可如上述举例。

707、当针对存储器Nd2-D1的数据重构完成后,存储节点Nd2将存储器Nd2-D1的状态设置为信任访问状态。进一步的,存储节点Nd2可向管理节点Nd1发送存储器状态变更报告P3,所述存储器状态变更报告P3可指示存储器Nd2-D1的状态由数据重构状态变更为信任访问状态。

708、管理节点Nd1接收来自存储节点Nd2的所述存储器状态变更报告P3,所述管理节点Nd1更新缓存的状态信息集合体。具体的,所述存储节点Nd1将其缓存的状态信息集合体中记录的存储器Nd2-D1的状态,更新为存储器状态变更报告P3所指示出的存储器Nd2-D1的状态,并且将这个状态信息集合体记录的状态信息集合体的版本号更新。

举例来说,在各存储节点中都部署了monitor的情况下,那么具体可以由存储节点Nd2中的从monitor向管理节点Nd1发送存储器状态变更报告P3,相应的可由所述管理节点Nd1中的主monitor接收来自存储节点Nd2的所述存储器状态变更报告P3。

例如,管理节点Nd1更新前的状态信息集合体例如图7-B举例所示,管理节点Nd1更新后的状态信息集合体例如图7-C举例所示,当然实际应用中,状态信息集合体并不限于这样的举例形式。

709、管理节点Nd1向其它存储节点(例如Nd2、Nd3和Nd4等)发送更新后的状态信息集合体。

相应的,存储节点Nd2、Nd3和Nd4可分别接收来自存储节点Nd1的更新后的状态信息集合体,存储节点Nd2、Nd3和Nd4可分别用来自管理节点Nd1的状态信息集合体更新掉其当前缓存的状态信息集合体。这样各个存储节点(Nd1、Nd2、Nd3和Nd4)所缓存的状态

信息集合体就可尽量保持相对的同步。

上述方案主要针对的是离线存储器重新在线之后的一种数据重构场景，给出在这种可能的场景下的一种可能的数据重构机制。原本离线的存储器重新上线后，其存储节点通知管理节点，管理节点对自己记录的这个存储器的状态更新为“在线重构状态”，然后管理节点向其他存储节点发布最新的状态信息集合体，进而通知其他存储节点这个存储器的状态已被更新为“在线重构状态”。接着，相关存储节点把之前未能写入这个存储器的条带进行重构。重构完成后，重新上线的存储器所在存储节点把这个重新上线的存储器的状态更新为“信任访问状态”，并通知管理节点把这个重新上线的存储器的状态更新为“信任访问状态”，接着，管理节点再次向其他存储节点发布最新的状态信息集合体，进而通知余下的存储节点这个重新上线的存储器的状态被更新为“信任访问状态”。

本实施例引入了用于记录存储器状态的状态信息集合体，故而可利用状态信息集合体来记录管理分布式存储系统的存储器的访问状态，便于分布式存储系统中的存储节点较为准确了解存储器的状态切换（从“离线状态”切换到“在线重构状态”，再从“在线重构状态”切换到“信任访问状态”），进而有利于降低尝试读写的失败几率，进而有利于提升系统性能。

实施例六

下面再结合相关附图举例介绍另一种数据访问方法，主要针对终端从分布式存储系统读取数据的一种可能的场景。这种数据访问方法应用于分布式存储系统，例如分布式存储系统包括m个存储节点，所述m个存储节点包括存储节点Nd2、Nd1、Nd3和Nd4等。存储器Nd2-D1属于所述存储节点Nd2，存储器Nd3-D1属于所述存储节点Nd3，存储器Nd4-D1属于所述存储节点Nd4。本实施例中以数据体的格式为文件为例进行描述。

请参见图8，图8是本申请的一实施例提供的一种数据访问方法的流程示意图。如图8所示，本申请的一实施例提供的一种数据访问方法可包括：

801、当终端需要从分布式存储系统读取数据，那么终端可向分布式存储系统发送读数据请求，为了下面便于引述，本实施例中把这个读数据请求命名为Rq1。本步骤中，例如分布式存储系统中的存储节点Nd1接收来自终端的读数据请求Rq1。

例如，读数据请求Rq1携带数据Data2（数据Data2为待读取数据）所属文件的文件标识和数据Data2的位置信息。其中，数据Data2的位置信息用于描述数据Data2在其所属文件中的位置。数据Data2的位置信息例如可包括的数据Data2的文件偏移地址（数据Data2的文件偏移地址表示了数据Data2在文件中的起始位置或结束位置）和数据Data2的长度等。

802、存储节点Nd1基于数据Data2的位置信息确定数据Data2所在存储器。

例如存储节点Nd1确定出数据Data2所在存储器为存储器Nd2-D1。例如：Data2所属文件在存储时被划分为多个数据条带（以及若干校验条带），其中一个数据条带存储在存储器Nd2-D1，而所述数据Data2是这个条带的一部分。

803、存储节点Nd1确定其本地缓存的状态信息集合体之中当前记录的存储器Nd2-D1的状态。其中，本实施例中，存储节点Nd1与管理节点之间可能处于连接态，也可能处于未处于连接态。

例如, 当存储节点Nd1当前缓存的状态信息集合体中记录: 存储器Nd2-D1处于信任访问状态, 那么执行步骤804。当存储节点Nd1当前维护的状态信息集合体记录: 存储器Nd2-D1处于数据重构状态或离线状态, 那么执行步骤步骤807。

804、由于确定出存储器Nd2-D1处于信任访问状态, 那么所述存储节点Nd1可在读数据请求Rq1中添加存储节点Nd1当前本地缓存的状态信息集合体的版本号以得到更新的读数据请求Rq1, 所述存储节点Nd1向所述存储器Nd2-D1所属的存储节点Nd2发送更新的读数据请求Rq1。

也就是说, 更新的读数据请求Rq1可携带存储节点Nd1当前本地缓存的状态信息集合体的版本号。进一步的, 读数据请求Rq1还可携带数据Data2的位置信息, 数据条带Pd2所属文件的文件标识等等。

可以理解, 存储节点Nd1将自身缓存的状态信息集合体的版本号携带在发送给其他存储节点的读数据请求中, 主要是为便于其他存储节点通过版本号比较来验证存储节点Nd1缓存的状态信息集合体是否有效。

805、存储节点Nd2在接收到来自自存储节点Nd1的读数据请求Rq1之后, 可将读数据请求Rq1携带的状态信息集合体的版本号, 与其自身当前维护的状态信息集合体的版本号进行比较。

其中, 在比较出读数据请求Rq1携带的状态信息集合体的版本号与自身当前缓存的状态信息集合体的版本号相同, 并且存储节点Nd2与管理节点之间处于连接态的情况下 (即存储节点Nd2中的从monitor与管理节点中的主monitor连接正常, 在这种情况下, 可以认为存储节点Nd2当前缓存的状态信息集合体极大可能是主monitor维护的最新版本的状态信息集合体), 存储节点Nd2便可从存储器Nd2-D1读取数据Data2。此外, 在比较出读数据请求Rq1携带的状态信息集合体的版本号, 与自身当前缓存的状态信息集合体的版本号不相同的情况之下, 存储节点Nd2则可拒绝从存储器Nd2-D1读取数据。此外, 若存储节点Nd2与管理节点之间未处于连接态 (即存储节点Nd2中的从monitor与管理节点中的主monitor直接没有连接正常, 在这种情况下, 可以认为存储节点Nd2当前缓存的状态信息集合体可能是主monitor维护的最新版本的状态信息集合体, 也可能不是主monitor维护的最新版本的状态信息集合体), 因此在这种情况下, 存储节点Nd2也可拒绝从存储器Nd2-D1读取数据。当然在某些验证级别较低的场景下, 也可能选择无论存储节点Nd2与管理节点之间是否处于连接态, 在数据请求Rq1携带的状态信息集合体的版本号, 与自身当前缓存的状态信息集合体的版本号相同的情况之下, 存储节点Nd2便可从存储器Nd2-D1读取数据Data2。

当当存储节点Nd2成功从存储器Nd1-D1读取到数据Data2, 存储节点Nd2向存储节点Nd1发送用于响应读数据请求Rq1的读数据响应Rr1, 读数据响应Rr1携带读取到的数据Data2。当存储节点Nd2未成功从存储器Nd1-D1读取到数据Data2 (例如存储节点Nd2拒绝从存储器Nd2-D1读取数据), 那么读数据响应Rr1可携带读操作结果 (读操作结果为读取失败)。

806、存储节点Nd1接收来自所述存储节点Nd2的所述读数据响应Rr1。若存储节点Nd2读取数据Data2成功, 则所述存储节点Nd1可从读数据响应Rr1中获得数据Data2。

存储节点Nd1向终端发送用于响应读数据请求Rq1的读数据响应Rr1, 读数据响应Rr1携带数据Data2。然后结束本流程。

807、由于确定出存储器Nd2-D1处于非信任访问状态, 所述存储节点Nd1确定数据Data1所属的数据条带Pd1, 所述存储节点Nd1确定数据条带Pd1所属分条, 所述分条包括数据条带Pd1、数据条带Pd2和校验条带Pj1。所述存储节点Nd1确定数据条带Pd2和校验条带Pj1所在的存储器。

例如数据条带Pd2所在存储器表示为存储器Nd3-D1, 存储器Nd3-D1所属存储节点表示为存储节点Nd3。例如校验条带Pj1所在存储器表示为存储器Nd4-D1, 存储器Nd4-D1所属存储节点表示为存储节点Nd4。

10 本步骤中, 由于待读数据Data2所属于条带所在的存储器处于非信任访问状态, 因此本步骤中, 收到读请求的存储节点Nd1确定出分条中余下条带所在的存储器, 以便后续利用余下条带使用校验算法恢复出Data2所在的条带, 进而获得Data2。

808、存储节点Nd1向存储器Nd3-D1所属的存储节点Nd3发送读数据请求Rq3。存储节点Nd1向所述存储器Nd4-D1所属的所述存储节点Nd4发送读数据请求Rq4。

15 例如读数据请求Rq3携带数据条带Pd2的条带标识和存储节点Nd1当前缓存的状态信息集合体的版本号。进一步的, 读数据请求Rq3还可携带数据条带Pd2的长度(长度如为110Kb或其他长度)、数据条带Pd2的文件偏移地址、数据条带Pd2所属分条的分条标识、数据条带Pd2所属文件的文件标识, 甚至还可能携带数据条带Pd2所属文件片段的片段标识。

20 例如读数据请求Rq4携带校验条带Pj1的条带标识和存储节点Nd1当前本地缓存的状态信息集合体的版本号。进一步的, 读数据请求Rq4还可携带校验条带Pj1的长度和校验条带Pj1所属分条的分条标识等。

809、存储节点Nd3在接收到来自存储节点Nd1的读数据请求Rq3之后, 将读数据请求Rq3携带的状态信息集合体的版本号, 与自身当前缓存的状态信息集合体的版本号进行比较, 在比较出读数据请求Rq3携带的状态信息集合体的版本号与自身当前缓存的状态信息集合体的版本号相同的情况下, 并且存储节点Nd3与管理节点之间处于连接态的情况下(即存储节点Nd3中的从monitor与管理节点中的主monitor连接正常, 这种情况下可认为存储节点Nd3当前缓存的状态信息集合体是主monitor维护的最新版本的状态信息集合体, 而读数据请求Rq1携带的状态信息集合体的版本号与存储节点Nd3自身当前缓存的状态信息集合体的版本号相同, 也就说明, 存储节点Nd1当前缓存的状态信息集合体也是主monitor维护的最新版本的状态信息集合体, 那么存储节点Nd1当前缓存的状态信息集合体中记录的存储器Nd3-D1的状态(信任访问状态)也是准确的), 存储节点Nd3从存储器Nd3-D1读取数据条带Pd2。此外, 在比较出读数据请求Rq3携带的状态信息集合体的版本号与自身当前维护的状态信息集合体的版本号不相同的情况下, 存储节点Nd3可拒绝从存储器Nd3-D1读取数据条带Pd2。

35 进一步, 存储节点Nd3向存储节点Nd1发送用于响应读数据请求Rq3的读数据响应Rr3, 读数据响应Rr3携带数据条带Pd2的读操作结果。具体的, 当存储节点Nd3成功从存储器Nd3-D1读取到数据条带Pd2, 那么, 读数据响应Rr3携带的读操作结果为读取成功, 并且读

数据响应Rr3携带数据条带Pd2。当存储节点Nd3未成功从存储器Nd3-D1读取数据条带Pd2（例如存储节点Nd3拒绝从存储器Nd3-D1读取数据条带Pd2），那么读数据响应Rr3携带的读操作结果为读取失败。

810、存储节点Nd4在接收到来自存储节点Nd1的读数据请求Rq4之后，执行和存储节点Nd3类似操作（例如，进行状态信息集合体版本的比较，以及发送读数据响应Rr4给存储节点Nd1）。由于本步骤可以参考步骤809，故此处不做赘述。

811、存储节点Nd1接收来自存储节点Nd3的所述读数据响应Rr3。所述存储节点Nd1接收来自所述存储节点Nd4的所述读数据响应Rr4。

进一步的，若读数据响应指示相应读入失败，那么存储节点Nd1可以重新发送相应读数据请求来请求重新读取。例如，如果读数据响应Rr3指示相应读取失败，那么存储节点Nd1可重新向存储节点Nd3发送写数据请求来重新请求读取数据条带Pd2。又例如，如果读数据响应Rr4指示相应读取操作失败，那么存储节点Nd1可重新发送读数据请求Wq2来重新请求读取校验条带Pj1。以此类推。

812、当存储节点Nd1获得数据条带Pd2和校验条带Pj1，存储节点Nd1利用数据条带Pd2和校验条带Pj1进行校验运算以得到数据条带Pd1，从数据条带Pd1中获得数据Data2，存储节点Nd1向终端发送用于响应读数据请求Rq1的读数据响应Rr1，读数据响应Rr1携带数据Data2。

而当存储节点Nd1从来自所述存储节点Nd2的读数据响应Rr1中获得数据Data2，存储节点Nd1向终端发送用于响应读数据请求Rq1的读数据响应Rr1，读数据响应Rr1携带数据Data2。

可以看出，上述方案中，由于在分布式存储系统中，引入了用于记录存储器状态的状态信息集合体，故而可以利用状态信息集合体来记录管理分布式存储系统的存储器的访问状态。具体引入了接收读数据请求的存储节点，基于这个读数据请求携带的状态信息集合体的版本号和其自身当前缓存的读数据请求携带的状态信息集合体的版本号之间的比较结果来进行相应读数据决策的机制，通过比较版本号有利于较好确定所使用的状态信息集合体的有效性。对于验证发现处于信任访问状态的存储器，可直接从中读取被请求数据来向终端反馈，而在这种信任访问情况下，由于无需执行传统方式中先读取出相关分条的全部条带来进行正确性校验的繁琐步骤，通过减少数据读取量就有利于降低存储器负担，进而有利于提升系统性能。

本实施例介绍了对待读数据的读取方法，待读数据是条带的一部分或整个条带。如果条带所在的存储节点是可以信任的，那么直接返回待读数据给发出请求的设备（终端或者主机）；如果条带所在的存储节点不可信（处于重构状态或者离线状态），那么通过待读数据所在的分条的其他条带（除了待读数据所在条带）进行校验计算，从而获得待读数据所在条带，以便从中获得待读数据。

存储节点Nd1将自身缓存的状态信息集合体的版本号携带在发送给其他存储节点的读数据请求中，主要是为便于其他存储节点通过版本号比较来验证存储节点Nd1缓存的状态信息集合体是否有效。当存储节点Nd2与管理节点连接正常，在这种情况下，若在数据请

求Rq1携带的状态信息集合体的版本号，与存储节点Nd2自身当前缓存的状态信息集合体的版本号相同，那么可认为存储节点Nd1与Nd2缓存了相同的且最新的状态信息集合体，那么存储节点Nd2便可从存储器Nd2-D1读取数据Data2。此外，当存储节点Nd2与管理节点并未连接正常，在这种情况下，可以认为存储节点Nd2当前缓存的状态信息集合体可能是管理节点发布的最新版本的状态信息集合体，也可能不是管理节点发布的最新版本的状态信息集合体，因此由于难以确定存储节点Nd2当前缓存的状态信息集合体的有效性，因此存储节点Nd2便可拒绝从存储器Nd2-D1读取数据Data2。当然在某些验证级别较低场景下，也可能选择无论存储节点Nd2与管理节点之间是否处于连接态，在数据请求Rq1携带的状态信息集合体的版本号，与自身当前缓存的状态信息集合体的版本号相同的情况之下（当两个版本号相同，那么存储节点Nd1所缓存状态信息集合体有效的可能性也比较大），存储节点Nd2便可从存储器Nd2-D1读取数据Data2。其中，由于引入了状态信息集合体有效性的一些验证机制，因此有利于较好确定所使用的状态信息集合体的有效性。

实施例七

下面再结合附图举例介绍另一种数据访问方法。这种数据访问方法应用于分布式存储系统，例如分布式存储系统包括m个存储节点，其中，m个存储节点包括存储节点Nd1、Nd2、Nd3和Nd4等。存储器Nd2-D1属于存储节点Nd2。本实施例中以数据体为文件为例进行描述。

请参见图9，图9是本申请另一实施例提供的另一种数据访问方法的流程示意图。如图9所示，本申请的另一实施例提供的另一种数据访问方法可包括：

901、当终端需要从分布式存储系统读取数据，那么终端可向分布式存储系统发送读数据请求，为了下面便于引述，本实施例中把这个读数据请求命名为Rq1。本步骤中，例如分布式存储系统中的存储节点Nd2接收来自终端的读数据请求Rq1。

例如步骤901中的读数据请求Rq1携带的相关信息，与步骤801中的读数据请求Rq1携带的相关信息相同。

902、存储节点Nd2基于数据Data2的位置信息确定数据Data2所在存储器。根据位置信息确定一个数据所在的存储器的方案，其他实施例已有描述，所以此处不再描述。

例如存储节点Nd2确定出数据Data2所在存储器为存储器Nd2-D1。

903、存储节点Nd2确定其本地缓存的状态信息集合体之中记录的所述存储器Nd2-D1的状态。

例如，存储节点Nd2当前缓存的状态信息集合体记录存储器Nd2-D1处于信任访问状态。并且，存储节点Nd2与管理节点之间处理连接态（即，存储节点Nd2中的从monitor与管理节点中的主monitor连接正常，这种情况下可认为存储节点Nd2当前缓存的状态信息集合体是主monitor发布的最新版本的状态信息集合体）。

904、由于确定出存储器Nd2-D1处于信任访问状态，那么存储节点Nd2可从存储器Nd2-D1读取数据Data2。存储节点Nd2向终端发送用于响应读数据请求Rq1的读数据响应Rr1，读数据响应Rr1携带数据Data2。

此外，当确定出存储器Nd2-D1处于非信任访问状态，那么可参考图8举例所示场景的实施方式，收集数据Data2所属分条的其他条带，利用其他条带来得到数据Data2。具体实

施方式此处不再赘述。

可以看出,上述方案中,由于在分布式存储系统中,引入了用于记录存储器状态的状态信息集合体,故而可以利用状态信息集合体来记录管理分布式存储系统的存储器的访问状态。具体引入了接收读数据请求的存储节点,基于其缓存的发布状态信息集合体中记录的相关存储器的状态,及其与发布状态信息集合体的管理节点之间的连接状态来进行相应读数据决策的机制,这样有利于较好确定所使用的状态信息集合体的有效性。对于验证处于信任访问状态的存储器,可直接从中读取被请求数据来向终端反馈,而在这种信任访问情况下,由于无需执行传统方式中先读取相关分条的全部条带来进行正确性校验的繁琐步骤,通过减少数据读取量就有利于降低存储器负担,进而有利于提升系统性能。

本方案中,待读数据(数据Data2)位于接收到终端/主机的读数据请求的存储节点(Nd2)的存储器(ND2-D1),这个存储节点(Nd2)和管理节点处于连接态,那么可以认为这个存储节点(Nd2)其本地缓存的状态信息集合体中记录的存储器状态是相对可靠的。当状态信息集合体中记录中记录的这个存储节点(Nd2)是信任访问状态,那么可以直接从中读取待读数据(数据Data2),然后返回给终端/主机。

实施例八

下面再结合附图举例介绍另一种数据访问方法。数据访问方法应用于分布式存储系统,例如分布式存储系统包括m个存储节点,其中,m个存储节点包括存储节点Nd2、Nd1、Nd3和Nd4等。存储器Nd2-D1属于存储节点Nd2,存储器Nd3-D1属于存储节点Nd3,存储器Nd4-D1属于存储节点Nd4。本实施例中以数据体为文件为例进行描述。

参见图10,图10是本申请另一实施例提供的另一种数据访问方法的流程示意图。如图10所示,本申请的另一实施例提供另一种数据访问方法可包括:

1001~1002,其中,步骤1001~1002与步骤801~802相同,因此,相关描述可以相应参考步骤801~802,此处不再赘述。

1003、当存储节点Nd1与管理节点之间处于连接态,存储节点Nd1确定其本地缓存的状态信息集合体中记录的存储器Nd2-D1的状态。

例如,当存储节点Nd1当前缓存的状态信息集合体记录存储器Nd2-D1处于信任访问状态,那么执行步骤1004。当存储节点Nd1当前维护的状态信息集合体记录存储器Nd2-D1处于数据重构状态或离线状态,那么执行步骤1007。

其中,当存储节点Nd1与管理节点之间未处于连接态,那么存储节点Nd1可将读数据请求Rq1转发给管理节点,或将读数据请求Rq1转发给与管理节点之间处于连接态的其他存储节点,由收到Rq1的其他存储节点确定其本地缓存的状态信息集合体中记录的存储器Nd2-D1的状态,并据此进行后续的相关操作。具体的后续操作可以参照本实施例1003~步骤1011。本实施例中以存储节点Nd1与管理节点之间处于连接态的情况为例进行描述。

1004、由于确定出存储器Nd2-D1处于信任访问状态,那么存储节点Nd1向所述存储器Nd2-D1所属的存储节点Nd2发送读数据请求Rq1。

1005、存储节点Nd2在接收到来自存储节点Nd1的读数据请求Rq1后,存储节点Nd2可从存储器Nd2-D1读取数据Data2。存储节点Nd2进一步向存储节点Nd1发送用于响应读数据

请求Rq1的读数据响应Rr1, 其中, 读数据响应Rr1携带读取到的数据Data2。

1006、存储节点Nd1接收来自所述存储节点Nd2的所述读数据响应Rr1。若存储节点Nd2读取数据Data2成功, 则所述存储节点Nd1可从读数据响应Rr1中获得数据Data2。执行完本步骤后退出流程, 不再执行下面的步骤1007及其他步骤。

- 5 1007、由于确定出存储器Nd2-D1处于非信任访问状态, 存储节点Nd1确定数据Data1所属的数据条带Pd1, 存储节点Nd1确定数据条带Pd1所属分条, 所述分条包括数据条带Pd1、数据条带Pd2和校验条带Pj1。所述存储节点Nd1确定数据条带Pd2和校验条带Pj1所在的存储器。

- 10 例如数据条带Pd2所在存储器表示为存储器Nd3-D1, 存储器Nd3-D1所属存储节点表示为存储节点Nd3。例如校验条带Pj1所在存储器表示为存储器Nd4-D1, 存储器Nd4-D1所属存储节点表示为存储节点Nd4。存储节点Nd1确定其本地缓存的状态信息集合体中记录的存储器Nd3-D1和Nd4-D1的状态。

- 15 下面以存储节点Nd1当前缓存的状态信息集合体记录: 存储器Nd3-D1和Nd4-D1处于信任访问状态为例进行描述, 当存储节点Nd1当前维护的状态信息集合体记录: 存储器Nd3-D1和Nd4-D1处于非信任访问状态, 那么可直接反馈数据读取失败, 或者在等待设定时长之后, 重新查看当前缓存的状态信息集合体记录存储器Nd1-D1、Nd3-D1和Nd4-D1的状态。

1008、存储节点Nd1向存储器Nd3-D1所属的存储节点Nd3发送读数据请求Rq3。存储节点Nd1向所述存储器Nd4-D1所属的存储节点Nd4发送读数据请求Rq4。

- 20 例如读数据请求Rq3携带数据条带Pd2的条带标识的版本号。读数据请求Rq3还可以携带数据条带Pd2的长度(例如110Kb)、数据条带Pd2的文件偏移地址、数据条带Pd2所属分条的分条标识、数据条带Pd2所属文件的文件标识, 甚至还可能携带数据条带Pd2所属文件片段的片段标识。

例如读数据请求Rq4携带校验条带Pj1的条带标识。进一步的, 读数据请求Rq4还可携带校验条带Pj1的长度和校验条带Pj1所属分条的分条标识等。

- 25 1009、存储节点Nd3在接收到来自存储节点Nd1的读数据请求Rq3后, 存储节点Nd3从存储器Nd3-D1读取数据条带Pd2。

- 进一步的, 存储节点Nd3向存储节点Nd1发送用于响应读数据请求Rq3的读数据响应Rr3, 读数据响应Rr3携带数据条带Pd2的读操作结果。具体的, 当存储节点Nd3成功从存储器Nd3-D1读取到数据条带Pd2, 那么, 读数据响应Rr3携带的读操作结果为读取成功, 并且
30 读数据响应Rr3携带数据条带Pd2。当存储节点Nd3未成功从存储器Nd3-D1读取数据条带Pd2, 那么读数据响应Rr3携带的读操作结果为读取失败。

1010、存储节点Nd4在接收到来自存储节点Nd1的读数据请求Rq4之后, 存储节点Nd4从存储器Nd4-D1读取校验条带Pj1。

- 进一步的, 存储节点Nd4向存储节点Nd1发送用于响应读数据请求Rq4的读数据响应
35 Rr4, 读数据响应Rr4携带校验条带Pj1的读操作结果。具体的, 当存储节点Nd4成功从存储器Nd4-D1读取到校验条带Pj1, 那么, 读数据响应Rr4携带的读操作结果为读取成功, 并且读数据响应Rr3携带校验条带Pj1。当存储节点Nd4未成功从存储器Nd4-D1读取到校验条带

Pj1, 那么读数据响应Rr4携带的读操作结果为读取失败。

1011、存储节点存储节点存储节点存储节点存储节点存储节点存储节点存储节点存储节点存储节点存储节点~1012。其中, 步骤1011~1012与步骤811~812相同。

5 本实施中, 收到读请求的是一个存储节点(Nd1), 需读取的数据(数据Data1)位于另外一个存储节点(Nd2)的存储器(Nd2-D1)。如果Nd2和管理节点处于连接态, 且Nd2本地缓存的状态信息集合体中记录的存储器Nd2-D1的状态是信任访问状态, 那么可以直接从Nd2-D1读出需要读取的数据。如果Nd2和管理节点不处于连接态, 那么, 判断需要读取的数据所在的条带中, 位于其他存储器的分条(数据条带Pd2和校验条带Pj1)所在的存储器(Nd3-D1和Nd4-D1)。如果位于其他存储器的分条(数据条带Pd2和校验条带Pj1)所在的存储器(Nd3-D1和Nd4-D1)的状态是信任访问状态, 并且和管理节点处于连接态。那么就读出位于其他的存储器的分条, 通过校验算法得到需要读取的数据(数据Data1)所在的条带(Pd1), 从Pd1中得到需要读取的数据(数据Data1)返回给终端/主机。

15 可以看出, 上述方案中, 由于在分布式存储系统中, 引入了用于记录存储器状态的状态信息集合体, 故而可以利用状态信息集合体来记录管理分布式存储系统的存储器的访问状态。具体引入了接收和中转读数据请求的存储节点, 基于其缓存的发布状态信息集合体中记录的相关存储器的状态, 以及其与发布状态信息集合体的管理节点之间的连接状态来进行相应读数据请求中转决策的机制, 这样有利于较好确定所使用的状态信息集合体的有效性。对于处于信任访问状态的存储器, 可直接触发相关存储节点从中读取被请求数据来向终端反馈, 而在这种信任访问情况下, 由于无需执行传统方式中先读取出相关分条的全部条带来进行正确性校验的繁琐步骤, 通过减少数据读取量就有利于降低存储器负担, 进而有利于提升系统性能。

下面还提供用于实施上述方案的相关装置。

25 参见图11, 本申请实施例提供一种分布式存储系统1100, 所述分布式存储系统包括m个存储节点, 其中, 每个所述存储节点包括至少1个存储器, 每个所述存储器包含非易失性存储介质, 所述m为大于1的整数。所述m个存储节点之中的第一存储节点1110用于: 接收来自终端的第一读数据请求, 所述第一读数据请求携带有第一位置信息, 其中, 所述第一位置信息用于描述第一数据在其所属数据体中的位置; 基于所述第一位置信息确定所述第一数据所在的第一存储器。

30 在所述第一存储器归属于所述第一存储节点的情况下, 当状态信息集合体中当前记录所述第一存储器处于信任访问状态, 从所述第一存储器读取所述第一数据, 向所述终端发送用于响应所述第一读数据请求的第一读数据响应, 所述第一读数据响应携带有读取到的所述第一数据。所述第一数据例如为所述第一数据所属条带中的部分或全部数据。

35 在本申请一些可能实施方式中, 所述分布式存储系统还包括用于发布状态信息集合体的管理节点。

其中, 所述第一存储节点还用于, 在所述第一存储器归属于所述m个存储节点之中的第二存储节点, 并且所述第一存储器处于信任访问状态, 并且所述第一存储节点与用于发

布状态信息集合体的管理节点之间处于连接态的情况下，向所述第二存储节点转发所述第一读数据请求。

其中，所述第二存储节点1120用于，在接收到来自所述第一存储节点的所述第一读数据请求后，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中，所述第一存储节点1110还用于，在所述第一存储器属于所述m个存储节点之中的第二存储节点，并且，所述第一存储器处于信任访问状态的情况之下，在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识，向所述第二存储节点发送添加了所述所述状态信息集合体的标识的所述第一读数据请求。

其中，所述第二存储节点1120用于，接收到来自所述第一存储节点的所述第一读数据请求之后，将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中，所述第一存储节点1110还用于，在所述第一存储器属于所述m个存储节点之中的第二存储节点，并且，所述第一存储器处于信任访问状态的情况之下，在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识和用于发布状态信息集合体的管理节点的节点标识，向所述第二存储节点发送添加了所述状态信息集合体的标识和所述管理节点的节点标识的所述第一读数据请求。

相应的，所述第二存储节点1120用于，在接收到来自所述第一存储节点的所述第一读数据请求之后，将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，且将所述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识进行比较，在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同，且比较出所述第一读数据请求携带的所述管理节点的节点标识与自身当前缓存的管理节点的节点标识相同的情况之下，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中，所述第一存储节点1110还用于：在确定所述第一存储器处于非信任访问状态，并且所述第一数据所属的分条的各条带所在的N个存储器中的x个存储器属于所述第一存储节点，且所述N个存储器中的N-x个存储器属于不同于所述第一存储节点的y个存储节点的情况下，向所述y个存储节点发送携带所述分条的分条标识和所述状态信息集合体的标识的读数据请求。

所述y个存储节点中的每个存储节点用于，在接收到来自所述第一存储节点的读数据请

求之后，将所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，若比较出所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同，从其包括的相应存储器中读取所述分条的相应条带，向第一存储节点发送读取到的所述分条的相应条带。

- 5 所述第一存储节点还用于，基于从所述 y 个存储节点和所述第一存储节点收集到的所述分条的各条带进行校验运算以得到第一数据，向所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有得到的所述第一数据。

10 在本申请一些可能实施方式中，例如所述 m 个存储节点之中的其中1个存储节点为管理节点。其中，所述第一存储节点还用于，当所述第一存储器所处状态由第一状态变更为第二状态，向管理节点发送存储器状态变更报告，其中，所述存储器状态变更报告指示第一存储器处于第二状态。其中，所述第一状态和第二状态不同，所述第一状态和所述第二状态包括如下状态中的任意一个：离线状态、数据重构状态和信任访问状态。

15 其中，所述管理节点用于，在接收到了来自所述第一存储节点的存储器状态变更报告之后，将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态，并更新所述管理节点缓存的所述状态信息集合体的版本号；所述管理节点向所述 m 个存储节点中除所述管理节点之外的其他存储节点发送更新后的状态信息集合体。

相应的，所述第一存储节点还用于，用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。相应的，其他存储节点也用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

- 20 在本申请一些可能实施方式中，所述第一存储节点还用于，接收来自终端的第一写数据请求，所述第一写数据请求携带有第二数据和第二位置信息，所述第二位置信息用于描述所述第二数据在其所属数据体中的位置。

25 其中，所述第一存储节点还用于，基于所述第二位置信息确定写所述第二数据涉及的 W 个存储器；将第二数据切分为 $W-T$ 个数据条带，利用所述 $W-T$ 个数据条带计算得到 T 个校验条带，其中，所述 T 个校验条带和 $W-T$ 个数据条带形成包括 W 个条带的分条，所述 W 个存储器与所述 W 个条带之间一一对应，所述 T 和 W 为正整数且所述 T 小于所述 W 。

30 所述第一存储节点还用于，当所述第一存储节点缓存的状态信息集合体中当前记录所述 W 个存储器中的 $W1$ 个存储器处于非离线状态且所述 W 个存储器中的 $W2$ 个存储器处于离线状态，向所述 $W1$ 个存储器所属的 $y2$ 个存储节点发送携带所述分条的条带和所述状态信息集合体的标识的写数据请求。

35 相应的，所述 $y2$ 个存储节点中的每个存储节点用于，在接收到来自所述第一存储节点的写数据请求之后，将所述写数据请求携带状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，在比较出所述写数据请求携带的状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下，向其包括的相应存储器中写入所述分条的相应条带；在比较出所述写数据请求携带的状态信息集合体的标识与自身当前缓存的状态信息集合体的标识不相同的情况下，拒绝向其包括的相应存储器中写入所述分条的相应条带。

在本申请一些可能实施方式中,所述第一存储节点还用于,接收来自终端的第二写数据请求,所述第二写数据请求携带有第三数据和第三位置信息,所述第三位置信息用于描述所述第三数据在其所属数据体中的位置。

所述第一存储节点还用于,基于所述第三位置信息确定写所述第三数据涉及的W个存储器;将所述第三数据切分为W-T个数据条带,利用所述W-T个数据条带计算得到T个校验条带,其中,所述T个校验条带和W-T个数据条带形成包括W个条带的分条,所述W个存储器与所述W个条带之间一一对应,所述T和W为正整数且所述T小于所述W。

所述第一存储节点还用于,确定其缓存的状态信息集合体中当前记录的所述W个存储器所处状态,在确定出所述W个存储器中的W1个存储器处于非离线状态(例如并且所述W个存储器中的W2个存储器处于离线状态),并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下,向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带的写数据请求。

相应的,所述y2个存储节点中的每个1个存储节点用于,在接收到来自所述第一存储节点的写数据请求之后,向其包括的相应存储器中写入所述分条的相应条带。

在本申请一些可能实施方式中,所述第一存储节点还用于,接收来自终端的第三写数据请求,所述第三写数据请求携带有第四数据和第四位置信息,所述第四位置信息用于描述所述第四数据在其所属数据体中的位置;

所述第一存储节点还用于,基于所述第四位置信息确定写所述第四数据涉及的W个存储器;将所述第四数据切分为W-T个数据条带,利用所述W-T个数据条带计算得到T个校验条带,其中,所述T个校验条带和W-T个数据条带形成包括W个条带的分条,所述W个存储器与所述W个条带之间一一对应,所述T和W为正整数且所述T小于所述W。

所述第一存储节点还用于,当所述第一存储节点缓存的状态信息集合体中当前记录所述W个存储器中的W1个存储器处于非离线状态(例如并且所述W个存储器中的W2个存储器处于离线状态),并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下,向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带和用于发布状态信息集合体的管理节点的节点标识的写数据请求。

相应的,所述y2个存储节点中的每个存储节点用于,在接收到来自所述第一存储节点的写数据请求之后,将所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识进行比较,在比较出所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识相同的情况下,向其包括的相应存储器中写入所述分条的相应条带;在比较出所述写数据请求携带的管理节点的节点标识与自身当前缓存的管理节点的节点标识不相同的情况下,拒绝向其包括的相应存储器中写入所述分条的相应条带。

在本申请一些可能实施方式中,第一存储节点还用于生成第一重构日志,所述第一重构日志中记录了所述w2个存储器中的第二存储器的存储器标识,所述第一重构日志中还记录了所述W个条带中的与所述第二存储器存储器所对应的第一条带的条带标识,所述第一重构日志中还记录了所述分条的分条标识;其中,所述第二存储器为所述W2个存储器中的任意一个存储器。

相应的,例如当所述W2个存储器中的所述第二存储器重新上线之后,并且所述第二存储器所归属的第二存储节点收集到了在所述第二存储器离线期间产生的所述第一重构日志,第二存储节点获取第一重构日志中记录的需写入所述第二存储器的第一条带的标识;确定所述分条所包括的所述第一条带之外其他W-T个条带所在的W-T个存储器,从所述W-T个存储器中读取所述W-T个条带,利用所述W-T个条带进行校验运算以重构出需写入所述第二存储器的所述第一条带,将重构出的所述第一条带写入所述第二存储器。

可以理解,本实施例中的分布式存储系统1100的功能,可用于基于上述方法实施例的方案来实施,一些未描述的部分可参考上述实施例。

参见图12,本申请实施例还提供一种存储节点1200,所述存储节点为分布式存储系统所包括的m个存储节点的其中1个存储节点,每个所述存储节点包括至少1个存储器,每个所述存储器包含非易失性存储介质,所述m为大于1的整数。所述存储节点包括:通信单元和处理控制单元。

所述通信单元1210,用于接收来自终端的第一读数据请求,所述第一读数据请求携带有第一位置信息,所述第一位置信息用于描述第一数据在其所属数据体中的位置。

处理控制单元1220,用于在基于所述第一位置信息确定所述第一数据所在的存储器为的第一存储器,且所述第一存储器归属于所述存储节点的情况下,当状态信息集合体中当前记录所述第一存储器处于信任访问状态,从所述第一存储器读取所述第一数据。

其中,所述通信单元1210还用于,向所述终端发送用于响应所述第一读数据请求的第一读数据响应,所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中,所述通信单元1210还用于,在所述第一存储器归属于所述m个存储节点之中的第二存储节点,且所述第一存储器处于信任访问状态,且所述存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下,向所述第二存储节点转发所述第一读数据请求。

所述第一读数据请求用于触发所述第二存储节点在接收到了来自所述第一存储节点的所述第一读数据请求之后,从所述第一存储器读取所述第一数据,向所述存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应,所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中,所述通信单元还用于,在所述第一存储器归属于所述m个存储节点之中的第二存储节点,并且所述第一存储器处于信任访问状态的情况下,在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识,向所述第二存储节点发送添加了所述状态信息集合体的标识的所述第一读数据请求;

所述第一读数据请求用于触发所述第二存储节点接收到来自所述存储节点的所述第一读数据请求之后,将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较,在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下,从所述第一存储器读取所述第一数据,向所述存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应,其中,所述第一读数据响应携带有读取到的所述第一数据。

在本申请一些可能实施方式中，所述m个存储节点中其中1个存储节点为管理节点。

所述通信单元还用于，当所述第一存储器所处状态由第一状态变更为第二状态，向管理节点发送存储器状态变更报告，其中，所述存储器状态变更报告指示第一存储器处于第二状态；所述第一状态和第二状态不同，所述第一状态和所述第二状态包括如下状态中的
5 任意一个：离线状态、数据重构状态和信任访问状态。

其中，所述存储器状态变更报告用于触发所述管理节点在接收到来自所述第一存储节点的存储器状态变更报告后，将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态，并且更新所述管理节点缓存的所述状态信息集合体的版本号。向所述m个存储节点中除所述管理节点之外的其他存储节点发送更新后的状态信息集合
10 体。

所述处理控制单元用于，采用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

可以理解，本实施例中的存储节点1200的功能，可基于上述方法实施例的方案的具体实现，一些未描述的部分可参考上述实施例。

参见图13，本申请实施例提供了一种存储节点1300，所述存储节点为分布式存储系统所包括的m个存储节点的其中1个存储节点，每个所述存储节点包括至少1个存储器，每个
15 所述存储器包含非易失性存储介质，所述m为大于1的整数。

所述存储节点包括相互耦合的处理器1310和通信接口1320。处理器用于执行以上各方法实施例中由第一存储节点或其他存储节点所执行方法的部分或全部步骤。

所述存储器1330用于存储指令和数据，所述处理器1310用于执行所述指令，所述通信接口1320用于在所述处理器1310的控制下与其他设备进行通信。当所述处理器1310在执行
20 所述指令时可根据所述指令执行本申请上述实施例中的任意一种方法中由任一存储节点执行的部分或全部步骤。

处理器1310还称中央处理单元（英文：Central Processing Unit，缩写：CPU）。具体的
25 应用中存储节点各组件例如通过总线系统耦合在一起。其中，总线系统除了可包括数据总线之外，还可包括电源总线、控制总线和状态信号总线等。但是为清楚说明起见，在图中将各种总线都标为总线系统1340。其中，上述本申请实施例揭示的方法可应用于处理器1310中，或由处理器1310实现。其中，处理器1310可能是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述方法的各步骤可通过处理器1310中的硬件的集成逻辑电路
30 或者软件形式的指令完成。上述处理器1310可以是通用处理器、数字信号处理器、专用集成电路、现成可编程门阵列或其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。处理器1310可实现或执行本申请实施例中公开的各方法、步骤及逻辑框图。通用处理器1310可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的
35 硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等等本领域成熟的存储介质之中。该存储介质位于存储器1330，例如处理器1310可读取存储器1320中的信息，结合其硬件完成上

述方法的步骤。

例如所述处理器1310用于，通过所述通信接口1320接收第一读数据请求，所述第一读数据请求携带有第一位置信息，所述第一位置信息用于描述第一数据在其所属数据体中的位置；在基于所述第一位置信息确定所述第一数据所在的存储器为第一存储器，且所述第一存储器归属于所述第一存储节点的情况下，当状态信息集合体中当前记录所述第一存储器处于信任访问状态，从所述第一存储器读取所述第一数据，通过所述通信接口发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

本申请实施例提供一种计算机可读存储介质，其中，所述计算机可读存储介质存储了程序代码。所述程序代码包括用于执行第一方面中的任意一存储节点（例如第一存储节点或第二存储节点）所执行方法的部分或全部步骤的指令。

此外，本申请实施例提供了一种包括指令的计算机程序产品，当所述计算机程序产品在计算机（计算机例如为存储节点）上运行时，使得所述计算机执行以上各方面中由任意一存储节点（例如第一存储节点或第二存储节点等）所执行方法的部分或全部步骤。

此外，本申请实施例还提供一种业务系统，包括：分布式存储服务系统和终端，分布式存储服务系统和终端之间通信连接；所述分布式存储服务系统为如权本申请实施例提供的任意一种分布式存储服务系统。

在上述实施例中，可全部或部分地通过软件、硬件、固件或其任意组合来实现。当使用软件实现时，可全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可为通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线（例如同轴电缆、光纤、数字用户线）或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质（例如软盘、硬盘、磁带）、光介质（例如光盘）、或者半导体介质（例如固态硬盘）等。在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述的部分，可以参见其他实施例的相关描述。

需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可能可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本申请所必须的。

在本申请所提供的几个实施例中，应该理解到，所揭露的装置，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如上述单元的划分，仅仅为一

种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或者讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性或其它的形式。

5 上述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可位于一个地方，或者也可以分布到多个网络单元上。可根据实际的需要选择其中的部分或全部单元来实现本实施例方案的目的。

10 另外，在本申请各实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

15 上述集成的单元若以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可获取的存储器中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或者部分，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器中，包括若干请求用以使得一台计算机设备（可以为个人计算机、服务器或者网络设备等，具体可以是计算机设备中的处理器）执行本申请的各个实施例上述方法的全部或部分步骤。

20 以上所述，以上实施例仅用以说明本申请的技术方案而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，然而本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权 利 要 求

1、一种数据访问方法，其特征在于，所述方法应用于分布式存储系统，所述分布式存储系统包括m个存储节点，所述m个存储节点包括第一存储节点，每个所述存储节点包括至少1个存储器，每个所述存储器包含非易失性存储介质，所述m为大于1的整数，所述方法包括：

所述第一存储节点接收来自终端第一读数据请求，其中，所述第一读数据请求携带有第一位置信息，所述第一位置信息用于描述第一数据在其所属数据体中的位置；

在所述第一存储节点基于所述第一位置信息确定所述第一数据所在的存储器为第一存储器，且所述第一存储器归属于所述第一存储节点的情况下，当状态信息集合体中当前记录所述第一存储器处于信任访问状态，所述第一存储节点从所述第一存储器读取所述第一数据，向所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

2、根据权利要求1所述的方法，其特征在于，所述方法还包括：

在所述第一存储器归属于所述m个存储节点之中的第二存储节点，并且所述第一存储器处于信任访问状态，并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，所述第一存储节点向所述第二存储节点转发所述第一读数据请求；

所述第二存储节点在接收到来自所述第一存储节点的所述第一读数据请求之后，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

3、根据权利要求1所述的方法，其特征在于，所述方法还包括：在所述第一存储器归属于所述m个存储节点之中的第二存储节点，并且所述第一存储器处于信任访问状态的情况下，或者，在所述第一存储器归属于所述m个存储节点之中的第二存储节点，并且所述第一存储器处于信任访问状态，并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，所述第一存储节点在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识，向所述第二存储节点发送添加了所述状态信息集合体的标识的所述第一读数据请求；

所述第二存储节点在接收到了来自所述第一存储节点的所述第一读数据请求之后，在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下，从所述第一存储器读取所述第一数据，向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

4、根据权利要求1所述的方法，其特征在于，所述方法还包括：在存储节点所述第一存储器处于非信任访问状态的情况下，并且所述第一数据所属的分条的各条带所在的N个存储器中的x个存储器属于所述第一存储节点，且所述N个存储器中的N-x个存储器属于不同于所述第一存储节点的y个存储节点，所述第一存储节点向所述y个存储节点发送携带所述分条的分条标识和所述状态信息集合体的标识的读数据请求；

所述y个存储节点中的每个存储节点在接收到来自所述第一存储节点的读数据请求之

后, 将所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较, 若比较出所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同, 从其包括的相应存储器中读取所述分条的相应条带, 向第一存储节点发送读取到的所述分条的相应条带;

- 5 所述第一存储节点基于从所述 y 个存储节点和所述第一存储节点收集到的所述分条的各条带进行校验运算以得到第一数据, 向所述终端发送用于响应所述第一读数据请求的第一读数据响应, 其中, 所述第一读数据响应携带有得到的所述第一数据。

5、根据权利要求1-4中任一项所述的方法, 其特征在于, 所述 m 个存储节点中其中1个存储节点为管理节点;

- 10 所述方法还包括: 当第一存储器所处状态由第一状态变更为第二状态, 所述第一存储节点向管理节点发送存储器状态变更报告, 其中, 所述存储器状态变更报告指示第一存储器处于第二状态; 其中, 所述第一状态和第二状态不同, 所述第一状态和所述第二状态包括如下状态中的任意一个: 离线状态、数据重构状态和信任访问状态;

- 15 所述管理节点在接收到来自所述第一存储节点的存储器状态变更报告后, 将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态, 并更新所述管理节点缓存的所述状态信息集合体的版本号; 所述管理节点向所述 m 个存储节点中除所述管理节点之外的其他存储节点发送更新后的状态信息集合体;

所述第一存储节点用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

- 20 6、根据权利要求1-5中任一项所述的方法, 其特征在于, 所述方法还包括: 所述第一存储节点接收来自终端的第一写数据请求, 所述第一写数据请求携带有第二数据和第二位置信息, 所述第二位置信息用于描述所述第二数据在其所属数据体中的位置;

- 25 所述第一存储节点基于所述第二位置信息确定写所述第二数据所涉及到的 W 个存储器; 将所述第二数据切分为 $W-T$ 个数据条带, 利用所述 $W-T$ 个数据条带计算得到 T 个校验条带, 其中, 所述 T 个校验条带和 $W-T$ 个数据条带形成包括 W 个条带的分条, 所述 W 个存储器与所述 W 个条带之间一一对应, 所述 T 和 W 为正整数, 所述 T 小于所述 W ;

当所述第一存储节点缓存的状态信息集合体当前记录所述 W 个存储器中的 $W1$ 个存储器处于非离线状态, 向所述 $W1$ 个存储器所属的 $y2$ 个存储节点发送携带所述分条的条带和所述状态信息集合体的标识的写数据请求;

- 30 所述 $y2$ 个存储节点之中的每个存储节点在接收到来自所述第一存储节点的写数据请求之后, 将所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较, 在比较出所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下, 向其包括的相应存储器中写入所述分条的相应条带; 在比较出所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识不相同的情况下, 拒绝向其包括的相应存储器中写入所述分条的相应条带。
- 35

7、根据权利要求1-5中任一项所述的方法, 其特征在于, 所述方法还包括: 所述第一

存储节点接收来自终端的第二写数据请求，所述第二写数据请求携带有第三数据和第三位置信息，所述第三位置信息用于描述所述第三数据在其所属数据体中的位置；

所述第一存储节点基于所述第三位置信息确定所述第三数据所涉及的 W 个存储器之后；将所述第三数据切分为 $W-T$ 个数据条带，利用所述 $W-T$ 个数据条带计算得到 T 个校验条带，其中，所述 T 个校验条带和 $W-T$ 个数据条带形成包括 W 个条带的分条，所述 W 个存储器与
5 所述 W 个条带之间一一对应，所述 T 和 W 为正整数，所述 T 小于所述 W ；

在所述第一存储节点缓存的状态信息集合体中当前记录所述 W 个存储器中的 W_1 个存储器处于非离线状态，并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况之下，向所述 W_1 个存储器所属的 y_2 个存储节点发送携带有所述分条的条带的写数据请求；
10

所述 y_2 个存储节点之中的每个存储节点在接收到来自所述第一存储节点的写数据请求之后，向其包括的相应存储器中写入所述分条的相应条带。

8、根据权利要求6或7所述的方法，其特征在于，所述 W_1 小于所述 W ，其中，所述第一存储节点缓存的状态信息集合体中当前记录所述 W 个存储器中的 W_2 个存储器处于离线状态；
15

所述方法还包括：所述第一存储节点生成第一重构日志，所述第一重构日志中记录了所述 W_2 个存储器中的第二存储器的存储器标识，所述第一重构日志中还记录了所述 W 个条带中的与所述第二存储器所对应的第一条带的条带标识，所述第一重构日志中还记录了所述分条的分条标识；所述第二存储器为所述 W_2 个存储器中的任意一个存储器；

当所述 W_2 个存储器中的所述第二存储器重新上线之后，并且所述第二存储器所归属的第二存储节点收集到了在所述第二存储器离线期间产生的所述第一重构日志，第二存储节点获取第一重构日志中记录的需写入所述第二存储器的第一条带的标识；确定所述分条所包括的所述第一条带之外其他 $W-T$ 个条带所在的 $W-T$ 个存储器，从所述 $W-T$ 个存储器中读取所述 $W-T$ 个条带，利用所述 $W-T$ 个条带进行校验运算以重构出需写入所述第二存储器的所述
20 第一条带，将重构出的所述第一条带写入所述第二存储器。

9、一种分布式存储系统，其特征在于，包括 m 个存储节点，每个所述存储节点包括至少1个存储器，每个所述存储器包含非易失性存储介质，所述 m 为大于1的整数，

所述 m 个存储节点之中的第一存储节点用于：

接收来自终端的第一读数据请求，所述第一读数据请求携带有第一位置信息，所述第一位置信息用于描述第一数据在其所属数据体中的位置；基于所述第一位置信息确定所述
30 第一数据所在的第一存储器；

在所述第一存储器归属于所述第一存储节点的情况下，当状态信息集合体中当前记录所述第一存储器处于信任访问状态，从所述第一存储器读取所述第一数据，向所述终端发送用于响应所述第一读数据请求的第一读数据响应，其中，所述第一读数据响应携带有读取到的所述
35 第一数据。

10、根据权利要求9所述的分布式存储系统，其特征在于，所述分布式存储系统还包括用于发布状态信息集合体的管理节点；

其中, 所述第一存储节点还用于, 在所述第一存储器归属于所述 m 个存储节点之中的第二存储节点, 并且所述第一存储器处于信任访问状态, 并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下, 向所述第二存储节点转发所述第一读数据请求;

5 其中, 所述第二存储节点用于, 在接收到来自所述第一存储节点的所述第一读数据请求后, 从所述第一存储器读取所述第一数据, 向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应, 所述第一读数据响应携带有读取到的所述第一数据。

11、根据权利要求9所述的分布式存储系统, 其特征在于,

10 所述第一存储节点还用于, 在所述第一存储器归属于所述 m 个存储节点之中的第二存储节点, 并且所述第一存储器处于信任访问状态的情况下, 在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识, 向所述第二存储节点发送添加了所述所述状态信息集合体的标识的所述第一读数据请求;

15 其中, 所述第二存储节点用于, 接收到来自所述第一存储节点的所述第一读数据请求之后, 将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较, 在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下, 从所述第一存储器读取所述第一数据, 向所述第一存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应, 所述第一读数据响应携带有读取到的所述第一数据。

20 12、根据权利要求9所述的分布式存储系统, 其特征在于,

所述第一存储节点还用于: 在确定所述第一存储器处于非信任访问状态, 并且所述第一数据所属的分条的各条带所在的 N 个存储器中的 x 个存储器属于所述第一存储节点, 且所述 N 个存储器中的 $N-x$ 个存储器属于不同于所述第一存储节点的 y 个存储节点的情况下, 向所述 y 个存储节点发送携带所述分条的分条标识和所述状态信息集合体的标识的读数据请求;

25 所述 y 个存储节点中的每个存储节点用于, 在接收到来自所述第一存储节点的读数据请求之后, 将所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较, 若比较出所述读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同, 从其包括的相应存储器中读取所述分条的相应条带, 向第一存储节点发送读取到的所述分条的相应条带。

30 所述第一存储节点还用于, 基于从所述 y 个存储节点和所述第一存储节点收集到的所述分条的各条带进行校验运算以得到第一数据, 向所述终端发送用于响应所述第一读数据请求的第一读数据响应, 其中, 所述第一读数据响应携带有得到的所述第一数据。

13、根据权利要求9至12任意一项所述的分布式存储系统, 其特征在于, 所述 m 个存储节点中其中1个存储节点为管理节点;

其中, 所述第一存储节点还用于, 当所述第一存储器所处状态由第一状态变更为第二状态, 向管理节点发送存储器状态变更报告, 其中, 所述存储器状态变更报告指示第一存

存储器处于第二状态；其中，所述第一状态和第二状态不同，所述第一状态和所述第二状态包括如下状态中的任意一个：离线状态、数据重构状态和信任访问状态；

其中，所述管理节点用于，在接收到了来自所述第一存储节点的存储器状态变更报告之后，将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态，并更新所述管理节点缓存的所述状态信息集合体的版本号；所述管理节点向所述m个存储节点中除所述管理节点之外的其他存储节点发送更新后的状态信息集合体；

所述第一存储节点还用于，用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

14、根据权利要求9至12任一项所述的分布式存储系统，其特征在于，所述第一存储节点还用于，接收来自终端的第一写数据请求，所述第一写数据请求携带有第二数据和第二位置信息，所述第二位置信息用于描述所述第二数据在其所属数据体中的位置；

所述第一存储节点还用于，基于所述第二位置信息确定写所述第二数据涉及的W个存储器；将第二数据切分为W-T个数据条带，利用所述W-T个数据条带计算得到T个校验条带，其中，所述T个校验条带和W-T个数据条带形成包括W个条带的分条，所述W个存储器与所述W个条带之间一一对应，所述T和W为正整数且所述T小于所述W；

所述第一存储节点还用于，当所述第一存储节点缓存的状态信息集合体当前记录所述W个存储器中的W1个存储器处于非离线状态，向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带和所述状态信息集合体的标识的写数据请求；

所述y2个存储节点中的每个存储节点用于，在接收到来自所述第一存储节点的写数据请求之后，将所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，在比较出所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下，向其包括的相应存储器中写入所述分条的相应条带；在比较出所述写数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识不相同的情况下，拒绝向其包括的相应存储器中写入所述分条的相应条带。

15、根据权利要求9至12任一项所述的分布式存储系统，其特征在于，所述第一存储节点还用于，接收来自终端的第二写数据请求，所述第二写数据请求携带有第三数据和第三位置信息，所述第三位置信息用于描述所述第三数据在其所属数据体中的位置；

所述第一存储节点还用于，基于所述第三位置信息确定所述第三数据所涉及的W个存储器之后；将所述第三数据切分为W-T个数据条带，利用所述W-T个数据条带计算得到T个校验条带，其中，所述T个校验条带和W-T个数据条带形成包括W个条带的分条，所述W个存储器与所述W个条带之间一一对应，所述T和W为正整数，所述T小于所述W；

所述第一存储节点还用于，确定其缓存的状态信息集合体中当前记录的所述W个存储器所处状态，在确定出所述W个存储器中的W1个存储器处于非离线状态，并且所述第一存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下，向所述W1个存储器所属的y2个存储节点发送携带所述分条的条带的写数据请求；

所述y2个存储节点中的每个1个存储节点用于，在接收到来自所述第一存储节点的写数

据请求之后,向其包括的相应存储器中写入所述分条的相应条带。

16、根据权利要求14至15任一项所述的分布式存储系统,其特征在于,所述W1小于所述W,其中,所述第一存储节点缓存的状态信息集合体中当前记录所述W个存储器中的W2个存储器处于离线状态

5 所述第一存储节点还用于,生成第一重构日志,所述第一重构日志中记录了所述W2个存储器中的第二存储器的存储器标识,所述第一重构日志中还记录了所述W个条带中的与所述第二存储器存储器所对应的第一条带的条带标识,所述第一重构日志中还记录了所述分条的分条标识;其中,所述第二存储器为所述W2个存储器中的任意一个存储器;

10 当所述W2个存储器中的所述第二存储器重新上线之后,并且所述第二存储器所归属的第二存储节点收集到了在所述第二存储器离线期间产生的所述第一重构日志,第二存储节点获取第一重构日志中记录的需写入所述第二存储器的第一条带的标识;

确定所述分条所包括的所述第一条带之外其他W-T个条带所在的W-T个存储器,从所述W-T个存储器中读取所述W-T个条带,利用所述W-T个条带进行校验运算以重构出需写入所述第二存储器的所述第一条带,将重构出的所述第一条带写入所述第二存储器。

15 17、一种存储节点,其特征在于,所述存储节点为分布式存储系统所包括的m个存储节点的其中1个存储节点,其中,每个所述存储节点包括至少1个存储器,每个所述存储器包含非易失性存储介质,所述m为大于1的整数,

所述存储节点包括:

20 通信单元,用于接收来自终端的第一读数据请求,所述第一读数据请求携带有第一位置信息,所述第一位置信息用于描述第一数据在其所属数据体中的位置;

处理控制单元,用于在基于所述第一位置信息确定所述第一数据所在的存储器为的第一存储器,且所述第一存储器归属于所述存储节点的情况下,当状态信息集合体中当前记录所述第一存储器处于信任访问状态,从所述第一存储器读取所述第一数据;

25 所述通信单元还用于,向所述终端发送用于响应所述第一读数据请求的第一读数据响应,其中,所述第一读数据响应携带有读取到的所述第一数据。

18、根据权利要求17所述的存储节点,其特征在于,

所述通信单元还用于,在所述第一存储器归属于所述m个存储节点之中的第二存储节点,且所述第一存储器处于信任访问状态,且所述存储节点与用于发布状态信息集合体的管理节点之间处于连接态的情况下,向所述第二存储节点转发所述第一读数据请求;

30 所述第一读数据请求用于触发所述第二存储节点在接收到了来自所述第一存储节点的所述第一读数据请求之后,从所述第一存储器读取所述第一数据,向所述存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应,所述第一读数据响应携带有读取到的所述第一数据。

35 19、根据权利要求17所述的存储节点,其特征在于,所述通信单元还用于,在所述第一存储器归属于所述m个存储节点之中的第二存储节点,并且所述第一存储器处于信任访问状态的情况下,或在所述第一存储器归属于所述m个存储节点中的第二存储节点,并且所述第一存储器处于信任访问状态,并且所述第一存储节点与用于发布状态信息集合体的

管理节点之间处于连接态的情况下，在所述第一读数据请求中添加所述第一存储节点当前缓存的所述状态信息集合体的标识，向所述第二存储节点发送添加了所述状态信息集合体的标识的所述第一读数据请求；

所述第一读数据请求用于触发所述第二存储节点接收到来自所述第一存储节点的所述第一读数据请求之后，将所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识进行比较，在比较出所述第一读数据请求携带的所述状态信息集合体的标识与自身当前缓存的状态信息集合体的标识相同的情况下，从所述第一存储器读取所述第一数据，向所述存储节点或所述终端发送用于响应所述第一读数据请求的第一读数据响应，其中，所述第一读数据响应携带有读取到的所述第一数据。

20、根据权利要求17至19任一项所述的存储节点，其特征在于，所述m个存储节点中其中1个存储节点为管理节点；

所述通信单元还用于，当所述第一存储器所处状态由第一状态变更为第二状态，向管理节点发送存储器状态变更报告，其中，所述存储器状态变更报告指示第一存储器处于第二状态；所述第一状态和第二状态不同，所述第一状态和所述第二状态包括如下状态中的任意一个：离线状态、数据重构状态和信任访问状态；其中，所述存储器状态变更报告用于触发所述管理节点在接收到来自所述第一存储节点的存储器状态变更报告后，将所述管理节点缓存的状态信息集合体中记录的所述第一存储器的状态更新为第二状态，并且更新所述管理节点缓存的所述状态信息集合体的版本号；向所述m个存储节点中除所述管理节点之外的其他存储节点发送更新后的状态信息集合体；

所述处理控制单元用于，采用来自所述管理节点的状态信息集合体更新其当前缓存的状态信息集合体。

21、一种数据访问方法，应用于第一存储节点，其特征在于，所述第一存储节点位于包括m个存储节点的分布式存储系统中，每个所述存储节点包括至少1个存储器，每个所述存储器包含非易失性存储介质，所述m为大于1的整数，所述方法包括：

所述第一存储节点接收第一读数据请求，其中，所述第一读数据请求携带有第一位置信息，所述第一位置信息用于描述第一数据在其所属数据体中的位置；

在所述第一存储节点基于所述第一位置信息确定所述第一数据所在的存储器为第一存储器，且所述第一存储器归属于所述第一存储节点的情况下，当状态信息集合体中当前记录所述第一存储器处于信任访问状态，所述第一存储节点从所述第一存储器读取所述第一数据，发送用于响应所述第一读数据请求的第一读数据响应，所述第一读数据响应携带有读取到的所述第一数据。

22、一种存储节点，其特征在于，所述存储节点为分布式存储系统所包括的m个存储节点的其中1个存储节点，其中，每个所述存储节点包括至少1个存储器，每个所述存储器包含非易失性存储介质，所述m为大于1的整数，

所述存储节点包括相互耦合的处理器和通信接口；

其中，所述处理器用于，通过所述通信接口接收第一读数据请求，所述第一读数据请求携带有第一位置信息，其中，所述第一位置信息用于描述第一数据在其所属数据体中的

位置；在基于所述第一位置信息确定所述第一数据所在的存储器为第一存储器，且所述第一存储器归属于所述第一存储节点的情况下，当状态信息集合体中当前记录所述第一存储器处于信任访问状态，从所述第一存储器读取所述第一数据，通过所述通信接口发送用于响应所述第一读数据请求的第一读数据响应，其中，所述第一读数据响应携带有读取到的
5 所述第一数据。

10

15