(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
2017 年 7 月 6 日 (06.07.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/113276 A1

(51) 国际专利分类号:
G06F 11/10 (2006.01)

(21) 国际申请号: PCT/CN2015/100078

(22) 国际申请日: 2015 年 12 月 31 日 (31.12.2015)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: **华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.)** [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: **曾永强 (ZENG, Yongqiang)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (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, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, 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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则 4.17 的声明:

— 关于申请人有权申请并被授予专利(细则 4.17(ii))

[见续页]

(54) **Title**: DATA RECONSTRUCTION METHOD, APPARATUS AND SYSTEM IN DISTRIBUTED STORAGE SYSTEM

(54) **发明名称** : 分布式存储系统中的数据重建的方法、装置和系统

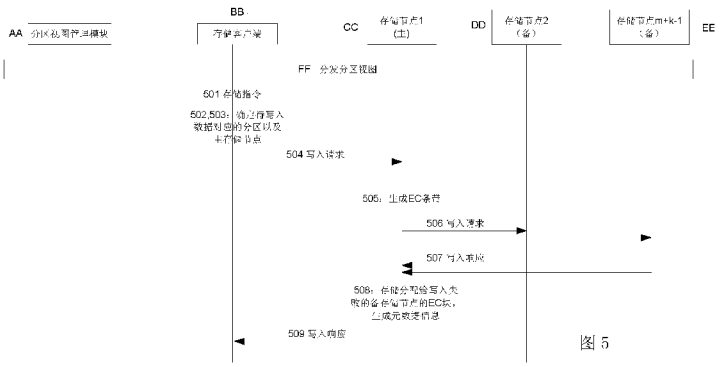

图 5

501 STORAGE AN INSTRUCTION
502, 503 DETERMINE A PARTITION CORRESPONDING TO DATA TO BE WRITTEN, AND A PRIMARY STORAGE NODE
504, 506 WRITE A REQUEST
505 GENERATE AN EC STRIPE
507 WRITE A RESPONSE
508 STORE EC BLOCKS ALLOCATED TO STANDBY STORAGE NODES FAILING IN WRITING AND GENERATING METADATA INFORMATION
509 WRITE A RESPONSE
AA PARTITIONED VIEW MANAGEMENT MODULE
BB STORAGE CLIENT
CC STORAGE NODE 1 (PRIMARY)
DD STORAGE NODE 2 (STANDBY)
EE STORAGE NODE M+K-1 (STANDBY)
FF DISTRIBUTION AND PARTITION VIEW

(57) **Abstract**: A data reconstruction method, apparatus and system in a distributed storage system. The method comprises: a primary storage node in a distributed storage system performs EC on data to be written to generate an EC stripe, stores EC blocks in the EC stripe on storage nodes, locally stores, when some storage nodes fail in writing the EC blocks due to a failure, the EC blocks allocated to the storage nodes failing in writing the EC blocks, and generates metadata information required for data reconstruction; the primary storage node resends, after the storage nodes recover from the failure, to the storage nodes the stored EC blocks allocated to the storage nodes failing in writing the EC blocks, and the metadata information corresponding to the EC blocks so that the storage nodes recovering from the failure complete data reconstruction. By means of the data reconstruction solution in a distributed storage system provided in the present application, when a failure occurs to some storage nodes, data on the nodes to which the failure occurs is cached and allocated by the primary storage node for the EC blocks allocated to the nodes to which the failure occurs without executing reverse EC to recover the data, and after the nodes recover from the failure, the cached EC blocks are resent to the nodes to which the failure occurs, for data reconstruction. The computing resource consumption caused by the execution of reverse EC during data reconstruction when the storage nodes recover from the failure is avoided, and the network resource consumption caused by transfer of a large amount of data during the execution of reverse EC is also avoided.

(57) **摘要**:

[见续页]

WO 2017/113276 A1

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种分布式存储系统中的数据重建的方法、装置和系统。分布式存储系统中的主存储节点对待写入数据进行 EC 编码，生成 EC 条带，将 EC 条带中的各个 EC 块分别存储在各个存储节点上，当部分存储节点由于故障导致写入 EC 块失败时，主存储节点将分配给写入失败的存储节点的 EC 块存储在本地，并生成数据重建所需的元数据信息，当存储节点故障恢复后，主存储节点将存储的分配给写入失败的存储节点的 EC 块以及该 EC 块对应的元数据信息重新发送给该存储节点，以使得故障恢复后的该存储节点完成数据重建。本申请提供的分布式存储系统中的数据重建的方案，当部分存储节点故障时，无需执行 EC 反编码以恢复故障节点上的数据，而是由主存储节点缓存分配给故障节点的 EC 块，再故障节点恢复后再将缓存的 EC 块重新发送给故障节点进行数据重建。避免了存储节点故障恢复进行数据重建时执行 EC 反编码带来的计算资源消耗，同时也避免了执行 EC 反编码时传递大量数据带来的网络资源消耗。

分布式存储系统中的数据重建的方法、装置和系统

技术领域

本发明涉及 IT 技术领域，尤其涉及分布式存储系统中的数据重建的方法、装置和系统。

背景技术

存储系统中，为了保证数据的安全，通常使用多副本存储技术来实现数据的冗余备份。多副本冗余技术就是对一份数据同时存储多份相同的副本，当一份数据丢失时，可以通过其他副本的数据将丢失的数据恢复出来，从而降低数据丢失的概率。副本个数的增加将会大大增加系统存储空间和网络带宽的消耗，从而增加数据存储的成本。如两副本情况下，用户真正可用空间是整个系统总存储空间的 50%，而在三副本的情况下，用户真正可用空间则只有 33%。

由于多副本存储技术存在存储空间浪费的缺点，现阶段的分布式存储系统越来越多的采用纠删码（EC，Erasure Code）技术对数据进行存储。目前在存储领域广泛应用的是 Reed-Solomen 类纠删码，具体原理是，将数据分割成 m 个数据块，采用冗余算法对 m 个数据块进行校验编码，用编码矩阵和 m 个数据块做乘法运算，从而生成 k 个校验块，该 m 个数据块与 k 个校验块组成一个 EC 条带。由于矩阵运算是可逆，当 EC 条带中的 m+k 个块中小于或等于 k 个块丢失时，均可以还原丢失的块中的数据。

相对副本而言，纠删码的编码技术无疑对存储空间利用率带来很大提升，但由于引入额外的编码、解码运算，对分布式存储系统的计算能力带来额外的要求。

发明内容

本申请描述了一种分布式存储系统中的数据重建的方法、装置和系统，解决了故障节点恢复后重建故障节点上的数据的问题，无需使用 EC 反编码的方式来恢复数据，降低了计算资源以及网络资源的消耗。

一方面，本发明实施例提供了一种分布式存储系统中数据重建的方法，第一存储节点获取待写入数据，确定所述待写入数据的 key 值，刷新所述 key 值对应的版本号，对所述待写入数据进行 EC 编码，生成 EC 条带，所述 EC 条带包括 m+k 个 EC 块，其中 m 个 EC 块为数据块，k 个 EC 块为校验块，m 为大于等于 2 的正整数，k 为自然数；第一存储节点查询分区视图，确定所述待写入数据所在的分区对应的第二存储节点，所述第一存储节点为所述分区对

应的主存储节点，所述第二存储节点为所述分区对应的备存储节点；所述第一存储节点向所述第二存储节点发送写入请求，所述写入请求携带所述待写入数据所在的分区标识（ID）、所述待写入数据的 key 值和版本号，以及分配给所述第二存储节点的 EC 块的数据、块内数据偏移和块内数据长度；第一存储节点确定所述第二存储节点写入失败，存储所述分配给所述第二存储节点的 EC 块的数据，生成与所述分配给所述第二存储节点的 EC 块对应的元数据信息，所述元数据信息包括所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号，以及分配给所述第二存储节点的 EC 块的块内数据偏移和块内数据长度；当所述第二存储节点故障恢复后，所述第一存储节点将存储的所述分配给所述第二存储节点的 EC 块的数据以及所述元数据信息发送给所述第二存储节点，以使得所述第二存储节点进行数据重建。

分布式存储系统中的主存储节点对待写入数据进行 EC 编码，生成 EC 条带，将 EC 条带中的各个 EC 块分别存储在各个存储节点上，当部分存储节点由于故障导致写入 EC 块失败时，主存储节点将分配给写入失败的存储节点的 EC 块存储在本地，并生成数据重建所需的元数据信息，当存储节点故障恢复后，主存储节点将存储的分配给写入失败的存储节点的 EC 块以及该 EC 块对应的元数据信息重新发送给该存储节点，以使得故障恢复后的该存储节点完成数据重建。本申请提供的分布式存储系统中的数据重建的方案，当部分存储节点故障时，无需执行 EC 反编码以恢复故障节点上的数据，而是由主存储节点缓存分配给故障节点的 EC 块，再故障节点恢复后再将缓存的 EC 块重新发送给故障节点进行数据重建。通过上述方案，避免了存储节点故障恢复进行数据重建时执行 EC 反编码带来的计算资源消耗，同时也避免了执行 EC 反编码时传递大量数据带来的网络资源消耗。示例性的，当 EC 编码为 4+2 时，恢复 1 份数据需要 4 份数据，而本申请中，只是在故障节点恢复后，由主存储节点向故障节点重新发送 1 份数据，明显降低了网络资源消耗；进一步的，分布式存储系统中，丢失 1 份数据的概率远远大于丢失两份数据，而丢失 1 份数据时，仍然可以还原出原始数据，因此，无需立刻执行 EC 反编码将丢失的数据还原，当故障节点恢复后，采用本申请提出的数据重建的方案即可将故障节点上的数据与其他存储节点进行同步。

在一种可能的实施方式中，所述待写入数据的 key 值用于表示存储所述待写入数据的逻辑卷的地址范围。

在一种可能的实施方式中，所述待写入数据的 key 值由存储所述待写入数据的逻辑卷的卷标识和数据偏移标识组成，所述数据偏移标识表示所述待写入数据在所述逻辑卷中的地址范围。

在一种可能的实施方式中，存储客户端根据存储所述待写入数据的逻辑卷的卷标识和数据偏移标识，确定所述待写入数据的key值，所述数据偏移标识表示所述待写入数据在所述逻辑卷中的地址范围。可以将逻辑卷的卷标识和数据偏移标识合并后得到的字符串作为所述待写入数据的key值。所述待写入数据的key值可以用来唯一区分该待写入数据存储的地址范围。

在一种可能的实施方式中，存储客户端使用一致性哈希算法计算所述待写入数据的 key值对应的哈希值，确定所述哈希值所属的分区的分区ID。存储客户端确定该分区对应的主存储节点为第一存储节点时，向第一存储节点发送写入请求，所述写入请求携带所述待写入数据、所述key值、分区ID，以及待写入数据的数据偏移和数据长度。

在一种可能的实施方式中，所述分区视图包括分区ID、主存储节点标识以及备存储节点标识。分区视图由分区视图管理模块维护并分发给各存储节点。通过统一的分区视图管理，可以实现待写入数据尽量均衡的写入到各个分区，且各个分区尽量均衡的分布在分布式存储系统的各个存储节点上，实现数据的冗余备份。

在一种可能的实施方式中，第一存储节点接收分区视图管理模块发送的故障节点通知消息，所述故障节点通知消息中携带所述第二存储节点的标识，确定所述第二存储节点写入失败。需要说明的是，当第一存储节点接收到备存储节点返回的写入失败响应确定备存储节点写入失败，不一定认为备存储节点必然会写入失败，由于发送备存储节点的数据可能丢失，此时，第一存储节点会向备存储节点再次发送写入请求，备存储节点重试写入EC块。因此，优选地，以分区视图管理模块发送的故障节点通知消息为准确定存储节点写入失败。

当第二存储节点故障恢复后，第二存储节点向分区视图管理模块请求最新的分区视图，确定自身存储的分区对应的主存储节点，向分区对应的主存储节点请求数据同步。故障节点上可能存在多个分区，每个分区的主存储节点可能不同，因此，针对不同的分区，故障节点需要向不同的主存储节点请求数据同步。

具体的，所述第一存储节点接收所述第二存储节点发送的数据同步请求，所述数据同步请求中携带所述分区ID;所述第一存储节点从所述第二存储节点获取所述第二存储节点中记录的所述分区ID对应的一个或多个key值，以及与所述key值的版本号；所述第一存储节点将自身记录的所述分区ID对应的key值以及各key值的版本号，与从所述第二存储节点获取的所述分区ID对应的一个或多个key值，以及与所述key值的版本号，进行比对，根据比对结果确定需要进行数据重建；所述第一存储节点根据所述元数据信息，将存储的需要进行数据重建的key值对应的EC块的数据以及所述EC块的元数据信息发送给所述第二存储

节点进行数据重建。

当第一存储节点接收到第二存储节点的数据同步请求时，执行比对操作，判断需要针对那些 key 值对应的数据进行同步，所述比对包括以下至少一种：

当所述第一存储节点记录的 key 值对应的版本号与从所述第二存储节点获取的所述 key 值对应的版本号一致时，无需进行数据重建；

当所述第一存储节点记录的 key 值中不包含从所述第二存储节点获取的 key 值时，则通知所述第二存储节点删除所述第一存储节点不包含的所述 key 值对应的数据；

当所述第一存储节点记录的 key 值对应的版本号大于从所述第二存储节点获取的所述 key 值对应的版本号时，执行数据重建操作；或，

当所述第一存储节点记录的 key 值中不包含在从所述第二存储节点获取的 key 值中时，则通知所述第二存储节点重建所述第二存储节点不包含的所述 key 值对应的数据。

在一种可能的实施方式中，当第二存储节点在故障恢复后接收到第一存储节点缓存的 EC 块以及对应的元数据信息时，所述第二存储节点根据所述分配给所述第二存储节点的 EC 块的块内数据偏移和块内数据长度，将所述 EC 块的数据写入到磁盘中，更新所述待写入数据的 key 值对应的版本号，从而完成该 EC 块的数据重建。

在一种可能的实现方式中，本申请以 m 个 EC 块的大小为粒度将所述逻辑卷的存储地址进行等分，得到多个存储单元，为所述多个存储单元分配数据偏移标识。进一步的，所述存储客户端接收上层应用发送的存储指令，所述存储指令中携带待存储的数据，以及存储所述待存储的数据的所述逻辑卷的卷标识、数据偏移和数据长度；所述存储客户端确定存储所述待存储的数据的地址范围对应的至少一个存储单元，将每个存储单元对应的部分待存储的数据作为一次写入操作中需要写入到分布式存储系统的所述待写入数据。

当一次存储的数据较大时，待存储的数据可能需要分成多段分几次写入到分布式存储系统中。本申请提供了一种对待存储数据进行分段写入的方法，一次写入的数据被称为待写入数据。

更进一步的，第一存储节点使用一致性哈希算法计算所述待写入数据的 key 值对应的哈希值，确定所述哈希值所属的分区的分区 ID。通过采用一致性哈希的方式，将数据均匀的分布到各个分区上。

在一种可能的实施方式中，第一存储节点可以在接收到写入失败响应后，确定发送写入

失败响应的存储节点故障；或者，第一存储节点可以根据分区视图中记录的存储节点的状态信息确定故障节点。

在一种可能的实施方式中，本申请还提供了第一存储节点缓存分配给写入失败的故障节点的 EC 块的方法，具体的，所述第一存储节点分配空闲的存储空间作为日志卷，用于存储分配给写入失败的存储节点的 EC 块，所述日志卷由至少一个日志块组成，所述日志块的大小与所述 EC 块的大小相同。

通过设定日志块的大小与 EC 块的大小相同，即实现了 EC 块粒度的数据重建，重建方法复杂度低。

另一方面，本发明实施例提供了一种存储节点，该存储节点具体实现上述方法中第一存储节点的功能。所述功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括一个或多个与上述功能相对应的模块。

在一个可能的设计中，存储节点的结构中包括处理器和存储器，所述处理器被配置为支持存储节点执行上述系统中相应的功能。所述存储节点还可以包括存储器，所述存储器用于与处理器耦合，其保存存储节点执行上述功能所必要的程序指令和数据。

又一方面，本发明实施例提供了一种分布式存储系统，该分布式存储系统具体实现上述方法中第一存储节点、第二存储节点、存储客户端以及分区视图管理模块的功能。所述功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括一个或多个与上述功能相对应的模块。

再一方面，本发明实施例提供了一种计算机存储介质，用于储存为第一存储节点所用的计算机软件指令，其包含用于执行上述方面所设计的程序。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍。显而易见地，下面附图中反映的仅仅是本发明的一部分实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得本发明的其他实施方式。而所有这些实施例或实施方式都在本发明的保护范围之内。

图 1 为实现本发明的一种可能的分布式存储系统的结构示意图；

图 2 为所示为本发明实施例提供的主存储节点上存储 EC 块的结构示意图；

图 3 为本发明实施例提供的一种日志块的管理逻辑示意图；

图 4 为所示为本发明实施例提供的计算机设备示意图；

图 5 为本发明实施例提供的一种写数据的流程示意图；

图 6 为本发明实施例提供的一种逻辑卷存储空间的等分示意图；

图 7 为本发明实施例提供的 DHT 哈希环的示意图；

图 8 为本发明实施例提供的一种故障节点恢复后的数据重建方法流程示意图；

图 9 为本发明实施例提供的一种存储节点的结构示意图；

图 10 为本发明实施例提供的一种分布式存储系统的结构示意图。

具体实施方式

下面将结合附图，对本发明实施例中的技术方案进行清楚、完整地描述。显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有付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。

本发明实施例描述的网络架构以及业务场景是为了更加清楚的说明本发明实施例的技术方案，并不构成对于本发明实施例提供的技术方案的限定，本领域普通技术人员可知，随着网络架构的演变和新业务场景的出现，本发明实施例提供的技术方案对于类似的技术问题，同样适用。

分布式存储系统使用 EC 技术进行数据存储时，会根据待存储的数据的大小生成一条或多条 EC 条带，并将每个 EC 条带的 m 个数据块及 k 个校验块分发给分布式存储系统的 m+k 个存储节点进行存储，EC 条带中的每个数据块或校验块也可以称为一个纠删码块 EC block。当分布式存储系统中有节点故障时，只要故障节点的数量小于 k，就可以根据非故障节点上的同一个 EC 条带的 EC 块将故障节点上存储的 EC 块恢复出来。恢复的方法为，首先获取至少 m 个存储节点上存储的属于同一个 EC 条带的 m 个 EC 块，针对 m 个 EC 块中的数据执行 EC 反编码，即可将故障节点上的同一个 EC 条带的 EC 块恢复出来。因此，采用 EC 技术存储数据的分布式存储系统具有很高的可靠性。

但是，EC 反编码技术对分布式存储系统的计算能力要求较高，且恢复故障节点上的 1 个 EC 块的数据需要传递 m 个 EC 块数据，对网络传输也造成了较大的负担。为克服上述问题，本发明实施例提供了一种分布式存储系统中的数据重建的方法，当出现存储节点故障的时候，将本来要写入故障节点的 EC 块临时存储到另一节点存储（优选的，存储到主存储节点上）。当故障节点恢复后，将转存到另一存储节点的 EC 块再重新写回到已恢复的故障节点上，从而实现该 EC 块中的数据在故障节点上的数据重建。采用本申请提出的方案，由于未进行 EC 反编码达到了节省计算资源的目的，同时数据重建仅需要将待重建的数据写回已恢复的故障节点，降低了网络传输的带宽消耗。在一种可能的实施方式中，由主存储节点存储分配给故障节点的数据，具体的，在主存储节点上分配一块空闲的存储空间，当某个存储节点出现故障时，将本来要写入故障节点的数据暂时存储在主存储节点上的上述空闲的存储空间中。等故障的节点恢复时，直接将主节点上保存的数据发送给故障节点。

如图 1 所示，为本发明实施例提供的一种分布式存储系统的结构示意图。分布式存储系统中包括多个存储节点，多个存储客户端以及分区视图（Partition View，PT View）管理模块。存储客户端接收上层应用发送的数据存储请求，采用纠删码技术，将待写入的数据存储到多个存储节点进行冗余备份存储，图 1 中的各组件的功能如下所述：

PT view 管理模块：可以部署在一台服务器上，也部署在一台存储节点上。其主要功能包括：

1. 监控各存储节点的状态。具体的，PT view 管理模块与各存储节点之间存在心跳连接，通过心跳连接监控各存储节点的状态：

存储节点的状态为正常，则表示存储节点存储的数据状态正常，存储节点未出现异常掉电或者在异常掉电后存储节点已完成数据重建；存储节点的状态为故障，表示存储节点存储的数据状态异常，存储节点出现异常掉电，或者未完成数据重建，与其他存储节点的数据不同步。

2. 生成及更新分区视图：基于存储节点状态及个数，根据 Partition 分区分配算法生成或更新分区视图，每个分区对应的存储节点包括一个主存储节点以及若干备份存储节点，每个分区对应的存储节点的总个数为 EC 条带包含的 EC 块的个数，即为 m+k。如下所示，为本发明实施例给出的 PT view 的示意图，为每个分区分配 m+k 个存储节点，指定其中一个存储节点为主存储节点，其他存储节点为备份存储节点，分区视图中包括各存储节点当前的状态信息。

分区标识（ID）	存储节点标识	节点身份	状态信息
****	1	主节点	正常
	2	备节点	正常
	……	……	……
	m+k	备节点	故障

分区视图可以由管理员进行手动设置，也可以由管理服务器进行分配，只要尽可能离散地将各分区分配给各存储节点即可，本发明实施例并不限定分区视图的建立方式。

3. 将分区视图发布给各存储客户端及存储节点.

存储客户端：部署在服务器上的逻辑功能模块。负责接收外部主机发送的数据存储请求，并将数据存储请求转换成对存储节点的键值对（key-Value）形式的 IO 请求；

存储节点：可以是物理存储节点，也可以是由物理存储节点划分的多个逻辑存储节点。主要功能有：数据相关处理（例如，EC 编码），完成分布式事务（两阶段事务）处理，并将 IO 最终转换为对盘的读写请求处理；

优选的，在各分区的主存储节点上需要分配一块空闲的存储空间，用于存放分配给写入失败的故障存储节点的 EC 块的数据，该空闲的存储空间可以称为日志卷。将日志卷的存储空间划分为若干个日志块，每个日志块的大小与 EC 块的大小相同，每个日志块对应一个日志块标识（identity，ID）。日志块在被分配使用的时候，会生成一个元数据信息，元数据信息中记录了分区 ID、写入该日志块的数据的 Key 值、该数据的版本号，以及该 EC 块的块内数据偏移和块内数据长度，进一步的，元数据信息还可以包括该日志块的 ID。

在一种可能的实现方式中，可以采用空间队列来管理空闲的日志块，使用 HashMap 方式来管理已被使用的日志块，Hash 的键值为分区 ID,这样相同分区的日志块放在一个队列中。当需要向日志卷写入新数据时，首先在空闲块列表中申请一个空闲的块，根据待写入数据的分区 ID 插入到对应的 HashMap 中。当该日志块使用完后，将该日志块重新放回空闲块列表中。

如图 1 所示的存储客户端、存储节点及分区视图管理模块可以采用硬件/软件实现，示例性的，如图 4 所示，为本发明实施例提供的计算机设备示意图。计算机设备 200 包括至少一个处理器 201，通信总线 202，存储器 203 以及至少一个通信接口 204。

处理器 201 可以是一个通用中央处理器（CPU），微处理器，特定应用集成电路

（application-specific integrated circuit，ASIC），或一个或多个用于控制本发明方案程序执行的集成电路。

通信总线 202 可包括一通路，在上述组件之间传送信息。所述通信接口 304，使用任何收发器一类的装置，用于与其他设备或通信网络通信，如以太网，无线接入网（RAN），无线局域网(Wireless Local Area Networks，WLAN)等。

存储器 203 可以是只读存储器（read-only memory，ROM）或可存储静态信息和指令的其他类型的静态存储设备，随机存取存储器（random access memory，RAM）或者可存储信息和指令的其他类型的动态存储设备，也可以是电可擦可编程只读存储器（Electrically Erasable Programmable Read-Only Memory，EEPROM）、只读光盘（Compact Disc Read-Only Memory，CD-ROM）或其他光盘存储、光碟存储（包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等）、磁盘存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质，但不限于此。存储器可以是独立存在，通过总线与处理器相连接。存储器也可以和处理器集成在一起。

其中，所述存储器 203 用于存储执行本发明方案的应用程序代码，并由处理器 201 来控制执行。所述处理器 201 用于执行所述存储器 203 中存储的应用程序代码。

在具体实现中，作为一种实施例，处理器 201 可以包括一个或多个 CPU，例如图 2 中的 CPU0 和 CPU1。

在具体实现中，作为一种实施例，计算机设备 200 可以包括多个处理器，例如图 2 中的处理器 201 和处理器 208。这些处理器中的每一个可以是一个单核（single-CPU）处理器，也可以是一个多核（multi-CPU）处理器。这里的处理器可以指一个或多个设备、电路、和/或用于处理数据（例如计算机程序指令）的处理核。

在具体实现中，作为一种实施例，计算机设备 200 还可以包括输出设备 205 和输入设备 206。输出设备 205 和处理器 201 通信，可以以多种方式来显示信息。例如，输出设备 205 可以是液晶显示器（liquid crystal display，LCD），发光二级管（light emitting diode，LED)显示设备，阴极射线管（cathode ray tube，CRT）显示设备，或投影仪（projector）等。输入设备 206 和处理器 201 通信，可以以多种方式接受用户的输入。例如，输入设备 206 可以是鼠标、键盘、触摸屏设备或传感设备等。

上述的计算机设备 200 可以是一个通用计算机设备或者是一个专用计算机设备。在具体实现中，计算机设备 200 可以是台式机、便携式电脑、网络服务器、掌上电脑（Personal

Digital Assistant，PDA)、移动手机、平板电脑、无线终端设备、通信设备、嵌入式设备或有图 4 中类似结构的设备。本发明实施例不限定计算机设备 200 的类型。

图 1 中的存储客户端、存储节点及分区视图管理模块可以为图 4 所示的设备，存储器中存储了一个或多个软件模块，用于实现存储客户端、存储节点及分区视图管理模块的功能(例如：存储节点的日志块的存储功能等)。存储客户端、存储节点及分区视图管理模块可以通过处理器以及存储器中的程序代码来实现虚拟机间的应用拓扑发现的方法。

需要说明的是，图 4 所示的计算机设备仅仅是给出了分布式存储系统中各部分的可能的硬件实现方式，根据系统各部分功能的不同或者变化，可以对计算机设备的硬件组件进行增删，以使得与系统各部分的功能进行匹配。

如图 5 所示，为本发明实施例给出的一种存在故障节点时的写数据的流程示意图，本发明实施例以 m=2，k=1 为例进行说明，此时，针对每个 EC 条带，存在 3 个 EC 块，需要将 3 个 EC 块分别存储到 3 个存储节点中，EC 块块的大小假定为 1M，示例性的，本发明实施例以存储节点 2 的状态为故障为例进行说明。

在本发明实施例中，针对逻辑卷的存储空间，以 m 个 EC 块的大小为粒度（下图以 m=2，EC 块块大小为 1M 为例)，将逻辑卷进行等分，如下所示，为逻辑卷存储空间的等分示意图。从地址 0 开始，等分后的逻辑卷的各个部分的数据偏移 ID 分别为 0,1,2……n。

步骤 501：存储客户端接收外部主机发送的存储指令，所述存储指令携带待存储的数据的逻辑卷的卷标识、数据偏移、数据长度，以及待存储的数据，所述存储客户端根据待存储的数据的数据偏移及数据长度生成每次写入操作时需要写入的待写入数据的数据偏移 ID。每次写入操作对应的待写入数据对应一个数据偏移 ID。

进一步的，所述存储客户端将所述逻辑卷的卷标识与所述待写入数据的数据偏移 ID 合并，生成待写入的数据的 key 值。当需要将待写入数据分成多次写入过程进行数据存储时，生成与每次写入过程的待写入数据对应的 key 值。

具体的，存储客户端对接收到的带存储的数据进行处理，划分成多次写入对应的待写入数据的过程如下：

示例性的，假如所述待存储的数据的数据长度为 3M，数据偏移为 2.5M，则待存储的数据存储的逻辑卷的地址范围为 2.5M 至 5.5M，该待存储的数据将会落到等分后的逻辑卷的第 2,3 个部分，两个部分分别对应的数据偏移 ID 为 1 和 2，数据偏移 ID 为 1 的存储空间存储的是地址范围为 2.5M 至 4M 的第一段待写入数据、数据偏移为 2 的存储空间存储的是地址范围为

4M 至 5.5M 的第二段待写入数据。此时，存储客户端会将上述待存储的数据分成上述两段待写入的数据，执行两次写入操作，将两端待写入数据分两次写入到存储节点中。第一段待写入数据的 key 值为所述逻辑卷的卷标识和数据偏移 ID1 合并生成的字符串；第二段待写入数据的 key 值为所述逻辑卷的卷标识和数据偏移 ID2 合并生成的字符串，由此可以看出，待写入数据的 key 值是由待写入数据的地址范围决定的，与待写入数据的具体内容无关。

需要说明的是，本发明实施例以写入第一段待写入数据为例进行说明，此时，待写入数据的地址范围为 2.5M 至 4M。本领域技术人员可以理解的是，当待存储的数据分成多段数据，需要分别执行多次写入操作时，多次写入的过程类似。鉴于不同写入过程中的待写入数据的数据偏移 ID 不同，所以不同写入过程中的待写入数据的 key 值不同，每次写入操作的待写入数据的数据偏移、数据长度发生变化而已。

在本发明实施例中，待存储数据为存储客户端从上层应用或者外部主机接收到的存储指令中所包含的待存储的数据；待写入数据为存储客户端在一次写入操作中写入到各存储节点的数据，一次写入过程中的待写入数据以一个 EC 条带的形式写入到各存储节点中。

步骤 502：存储客户端计算所述待写入数据的 key 值的哈希值，确定计算得到的哈希值对应的分区 ID。

在一种可能的实施方式中，可以采用一致性哈希的方式计算待写入数据的 key 值对应的分区 ID。如图 5 所示，为 DHT 哈希环的示例，在系统初始化的时候，会对[0, 2^32-1]这个大范围的整数区间进行分段，分成多个区间大小相等的分区(partition)，每个分区内的 Hash 整数个数一样，代表了相同长度的 Hash 空间。根据待写入数据的 key 值的哈希值，确定该哈希值落到的分区的分区 ID。

步骤 503：存储客户端根据所述分区 ID 查询分区视图，确定处理所述待写入数据的主存储节点。

分布式存储系统中的分区视图管理模块会维护分区与存储节点（存储节点可以是物理节点，也可以是由物理节点逻辑划分的多个逻辑节点）的对应关系，该对应关系即为分区视图。

进一步的，当待存储的数据较大，需要分成多个待写入的数据经过多次写入过程写入到存储节点时，第一段待写入数据的数据偏移与待存储数据的数据偏移相同，后续各段待写入数据的数据偏移为 0。当待存储数据可以在一次写入过程中写入时，待写入数据的数据偏移即为待存储数据的数据偏移。

步骤 504：存储客户端向主存储节点发送写入请求，所述写入请求携带所述待写入数据，

以及所述待写入数据对应的 key 值、数据偏移、数据长度以及分区 ID 等等。

本发明实施例以存储节点 1 为主存储节点，存储节点 2,3 为备存储节点为例进行说明。

步骤 505：主存储节点对待写入数据执行 EC 编码，生成 m 个数据块以及 k 个校验块，m+k 个 EC 块构成一个 EC 条带，由 m+k 个存储节点分别进行处理（包括主存储节点）。主存储节点生成或刷新所述待写入数据的版本号，并存储 key 值与版本号的对应关系。

具体的，主存储节点将待写入数据以 EC 块的粒度进行切分，形成 m 个数据块，执行 EC 编码，生成与 m 个数据块对应的 k 个校验块。主存储节点对所述待写入数据进行上述处理，生成了 m+k 个 EC 块。主存储节点生成针对每个 EC 块的数据偏移以及数据长度。示例性的，待写入数据的地址范围为 2.5M 至 4M，EC 块的大小为 1M，因此，待写入数据被划分为 2M-3M 以及 3M-4M 这两个 EC 块。其中，第一个 EC 块的块内数据偏移为 0.5M，块内数据长度为 0.5M；第二个 EC 块的块内数据偏移为 0M，块内数据长度为 1M；校验块的数据偏移为各个 EC 块数据偏移的最小值，校验块的数据长度为各 EC 块数据长度的地址范围的叠加，具体的，在本例中，k=1，即生成 1 个校验块，此时，校验块的块内数据偏移为 0，校验块的数据长度为 1M。在一种可能的实施方式中，在主存储节点记录校验块的块内数据偏移和块内数据长度，在各备存储节点以整个块为粒度进行校验块的写入。

需要说明的是，计算 EC 块的块内数据偏移以及块内数据长度可以采用现有技术中 EC 编码的常用方式，本发明实施例对此并不进行限定。

在一种可能的实施方式中，当逻辑卷的存储空间被第一次写入数据时，此时，在写入操作过程中，主存储节点生成该存储空间存储的待写入数据的版本号，当该存储空间中存储的数据被刷新的过程中，主存储节点刷新上述版本号，可以采用每次刷新将版本号加 1 的方式。

步骤 506：主存储节点根据分区视图确定该分区 ID 对应的备存储节点，向备存储节点发送写入请求，写入请求中携带待写入数据的 key 值和版本号，以及分配给该备存储节点的 EC 块的数据内容、块内数据偏移以及块内数据长度。

在一种可能的实施方式中，所述写入请求可以使 prepare 请求。主存储节点处理 1 个 EC 块，将剩余的 m+k-1 个 EC 块分别发送给分区视图中与分区 ID 对应的 m+k-1 个备存储节点。主存储节点分配 EC 块可以采用随机分配的方式或者按照存储节点的标识顺序分配的方式，本发明实施例对此并不进行限定。

步骤 507：备存储节点接收所述写入请求，根据所述写入请求将 EC 块中的数据内容按照写入到块内数据偏移以及块内数据长度对应的磁盘地址空间中，备存储节点记录待写入数据

的 key 以及版本号。写入操作完成后，备存储节点向主存储节点返回写入成功响应。如果备存储节点发生故障导致写入失败时，备存储节点向所述主存储节点返回写入失败响应。

在一种可能的实施方式中，所述写入响应为 prepare log 消息，其中携带成功或失败的指示标识。

在一种可能的实施方式中，主存储节点可以根据分区视图中存储节点的状态信息确定备存储节点是否故障，如果备存储节点故障，主存储节点可以直接将分配给故障的备存储节点的 EC 块存储在本地，不执行步骤 506 和 507。

步骤 508：主存储节点接收各备存储节点返回的写入响应，当确定写入成功的备存储节点数量大于或等于 m 个时，主存储节点确定本次写入操作成功，存储分配给写入失败的备存储节点的 EC 块，生成元数据信息，所述元数据信息包括待写入数据的分区 ID 、key 值和版本号，以及写入失败的 EC blcok 的块内数据偏移和块内数据长度。

在一种可能的方式中，主存储节点可以在接收到备存储节点返回的写入成功响应时，确定发送该写入成功响应的备存储节点已经成功的将待写入数据写入磁盘。另外，主存储节点可以从分区视图管理模块接收节点状态通知消息，确定故障节点写入失败，或者，主存储节点接收到写入失败响应，确定发送写入失败响应的备存储节点写入失败。

元数据信息中：所述 key 值为步骤 504 中主存储节点接收到的待写入数据的 key 值，版本号为步骤 505 中生成的与所述 key 值对应的版本号，块内数据偏移和块内数据长度为步骤 506 中所述主存储节点向写入失败的备存储节点发送的写入请求中包含的，分区 ID 为步骤 504 中存储客户端向主存储节点发送的所述待写入数据所属的分区 ID。

主存储节点存储分配给写入失败的备存储节点的待写入数据的具体方式可以如图 3 所示的方式。

在一种可能的实施方式中，当写入失败的备存储节点的个数大于 k 个时，主存储节点确定写入失败，向存储客户端返回写入失败响应。

步骤 509：主存储节点向写入成功的备存储节点返回确认消息，向存储客户端返回写入成功响应。

在一种可能的实施方式中，所述确认消息具体为 commit 消息。

如图 8 所示，本发明实施例还提供了一种故障节点恢复后的数据重建方法流程示意图，包括：

步骤 801：存储节点 2 在故障恢复以后，从分区视图管理模块获取分区视图，所述分区视图包括分区 ID，以及与所述分区 ID 对应的主存储节点和备存储节点。

分区视图管理模块在收到存储节点 2 发送的分区视图获取请求时，将与所述存储节点 2 相关分区的分区视图返回给所述存储节点 2。与所述存储节点 2 相关的分区为：主存储节点或备存储节点为存储节点 2 的分区。

在一种可能的实施方式中，当存储节点故障时，该故障节点可能是某些分区 ID 对应的主存储节点，此时，分布式存储系统的管理模块会将该故障节点降格为备存储节点，并刷新分区视图。

在一种可能的实施方式中，故障节点可能对应于多个分区 ID，为多个分区 ID 的备存储节点，故障节点以分区为粒度进行数据重建，本发明实施例以重建故障节点上一个分区的数据为例进行说明。存储节点 2 向分区 ID 对应的主存储节点请求同步数据，本领域技术人员可以理解的是，当存储节点上有多个分区的数据需要同步时，可以分别向各个分区 ID 对应的主存储节点请求同步数据。

步骤 802：存储节点 2 向主存储节点发送数据同步请求，所述数据同步请求携带分区 ID、key 值以及与所述 key 值对应的版本号；

需要说明的是，在一种可能的实施方式中，存储节点 2 可以先向主存储节点发送分区 ID，主存储节点接收到分区 ID 后，再从存储节点 2 获取该分区 ID 下的所有 key 值以及各 key 值对应的版本号。本领域技术人员可以理解的是，也可以分多次重建一个分区下的数据，本发明实施例对此并不进行限定。

步骤 803：主存储节点将本节点记录的所述分区 ID 对应的各 key 值以及各 key 值对应的版本号，与所述存储节点 2 上报的所述分区 ID 对应的所有 key 值以及与所述所有 key 值对应的版本号进行比对：

情况一、当主存储节点上记录的所述分区 ID 对应的 key 值不包含在所述存储节点 2 上报的所述分区 ID 对应的 key 值中时，说明在存储节点 2 故障期间，有新的数据（与未包含的 key 值对应）写入，主存储节点已成功写入该新的数据对应的 EC 块，但由于存储节点 2 故障，分配给存储节点 2 的 EC 块暂时存储在主存储节点上，因此，该 key 值对应的数据（具体是分配给存储节点 2 的 EC 块）未写入到存储节点 2 中，所述主存储节点将所述分配给存储节点 2 的 EC 块以及元数据信息发送给所述存储节点 2，存储节点 2 重建所述 key 值对应的数据；

情况二、所述 key 值对应的版本号与存储节点 2 上报的所述 key 值对应的版本号，当版

本号一致时，则说明存储节点 2 上所述 key 值对应的数据无需更新，即在存储节点 2 故障期间，所述 key 值对应的数据没有发生更新，版本号没有发生变化；

情况三、当主存储节点未查询到所述 key 值时，说明主存储节点上的所述 key 值对应的数据已经被删除，所述主存储节点通知所述存储节点 2 删除所述 key 值对应的数据；

情况四、当主存储节点上的所述 key 值对应的版本号大于所述存储节点 2 上报的 key 值对应的版本号时，说明存储节点 2 在故障期间，所述 key 值对应的数据发生了更新，所述主存储节点将所述 key 值对应的数据（具体是分配给存储节点 2 的 EC 块）以及元数据信息发送给所述存储节点 2，存储节点 2 重建所述 key 值对应的数据；

在情况一和情况四，需要进行数据重建，数据重建的具体过程包括：

步骤 804：主存储节点根据所述分区 ID 和待重建的数据对应的 key 值，查找元数据信息，确定待重建的 EC 块对应的日志卷以及元数据信息，所述主存储节点将日志卷中记录的待重建的 EC 块以及元数据信息中包含的块内数据偏移、块内数据长度以及版本号发送给存储节点 2。

步骤 805：存储节点 2 进行数据重建，根据所述分配给所述第二存储节点的 EC 块的块内数据偏移和块内数据长度，将所述 EC 块的数据写入到磁盘中，更新所述待写入数据的 key 值对应的版本号。。数据重建完成后，向所述主存储节点返回数据重建成功消息。

步骤 806：主存储节点删除已完成重建的数据的日志卷和元数据信息，所述主存储节点回收所述日志卷，放入空闲队列。

在一种可能的实施方式中，当存储节点 2 正在进行数据重建时，如果数据恢复过程中，有新的写请求要写入到正在恢复的故障节点，当写的位置时已经恢复完成的，则可以直接写；如果还没恢复或正在恢复的位置，则等待，等数据恢复完以后再写。

本申请提供了一种分布式存储系统中的数据重建的方法、装置和系统，分布式存储系统中的主存储节点对待写入数据进行 EC 编码，生成 EC 条带，将 EC 条带中的各个 EC 块分别存储在各个存储节点上，当部分存储节点由于故障导致写入 EC 块失败时，主存储节点将分配给写入失败的存储节点的 EC 块存储在本地，并生成数据重建所需的元数据信息，当存储节点故障恢复后，主存储节点将存储的分配给写入失败的存储节点的 EC 块以及该 EC 块对应的元数据信息重新发送给该存储节点，以使得故障恢复后的该存储节点完成数据重建。本申请提供的分布式存储系统中的数据重建的方案，当部分存储节点故障时，无需执行 EC 反编码以恢复故障节点上的数据，而是由主存储节点缓存分配给故障节点的 EC 块，再故障节点恢

复后再将缓存的 EC 块重新发送给故障节点进行数据重建。通过上述方案，避免了存储节点故障时执行 EC 反编码带来的计算资源消耗，同时也避免了执行 EC 反编码时传递大量数据带来的网络资源消耗。示例性的，当 EC 编码为 4+2 时，恢复 1 份数据需要 4 份数据，而本申请中，只是在故障节点恢复后，由主存储节点向故障节点重新发送 1 份数据，明显降低了网络资源消耗；进一步的，分布式存储系统中，丢失 1 份数据的概率远远大于丢失两份数据，而丢失 1 份数据时，仍然可以还原出原始数据，因此，无需立刻执行 EC 反编码将丢失的数据还原，当故障节点恢复后，采用本申请提出的数据重建的方案即可将故障节点上的数据与其他存储节点进行同步。

如前述分布式存储系统中数据重建的方法实施例相对应，如图 9 所示，本发明实施例还提供了一种存储节点，所述存储节点包括：

获取单元 901，用于获取待写入数据以及所述待写入数据的 key 值，刷新所述 key 值对应的版本号，对所述待写入数据进行 EC 编码，生成 EC 条带，所述 EC 条带包括 m+k 个 EC 块，其中 m 个 EC 块为数据块，k 个 EC 块为校验块，m 为大于等于 2 的正整数，k 为自然数；

处理单元 902，用于查询分区视图，确定所述待写入数据所在的分区对应的备存储节点，其中，所述第一存储节点为所述分区对应的主存储节点，第二存储节点为所述分区对应的其中一个备存储节点；

发送单元 903，用于向各备存储节点分别发送写入请求，所述写入请求携带所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号，以及分配给各备存储节点的 EC 块的数据；

所述处理单元 902，用于确定所述第二存储节点写入失败时，存储所述分配给所述第二存储节点的 EC 块的数据，生成与所述分配给所述第二存储节点的 EC 块对应的元数据信息，所述元数据信息包括所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号；

当所述第二存储节点故障恢复后，所述发送单元 903，还用于将存储的所述分配给所述第二存储节点的 EC 块的数据以及所述元数据信息发送给所述第二存储节点，以使得所述第二存储节点进行数据重建。

进一步的，所述处理单元 902，还用于确定写入成功的存储节点的数量大于等于 m。

所述获取单元 901，还用于接收所述第二存储节点发送的数据同步请求，所述数据同步请求中携带所述分区 ID，从所述第二存储节点获取所述第二存储节点中记录的所述分区 ID 对应的一个或多个 key 值，以及与所述 key 值的版本号；

所述处理单元 902，还用于将自身记录的所述分区 ID 对应的 key 值以及各 key 值的版

本号，与从所述第二存储节点获取的所述分区 ID 对应的一个或多个 key 值，以及与所述 key 值的版本号，进行比对，根据比对结果确定需要进行数据重建；

所述发送单元 903，具体用于根据所述元数据信息，将存储的需要进行数据重建的 key 值对应的 EC 块的数据以及所述 EC 块的元数据信息发送给所述第二存储节点进行数据重建。

所述处理单元 902，具体用于执行以下至少一种比对处理：

当本存储节点记录的 key 值对应的版本号与从所述第二存储节点获取的所述 key 值对应的版本号一致时，无需进行数据重建；

当本存储节点记录的 key 值中不包含从所述第二存储节点获取的 key 值时，则通知所述第二存储节点删除所述本存储节点不包含的所述 key 值对应的数据；

当所述本存储节点记录的 key 值对应的版本号大于从所述第二存储节点获取的所述 key 值对应的版本号时，执行数据重建操作；或，

当所述本存储节点记录的 key 值中不包含在从所述第二存储节点获取的 key 值中时，则通知所述第二存储节点重建所述第二存储节点不包含的所述 key 值对应的数据。

所述处理单元 902，还用于以 m 个 EC 块的大小为粒度将所述逻辑卷的存储地址进行等分，得到多个存储单元，为所述多个存储单元分配数据偏移标识。

所述获取单元 901，具体用于使用一致性哈希算法计算所述待写入数据的 key 值对应的哈希值，确定所述哈希值所属的分区的分区 ID；或者，

所述获取单元 901，具体用于从存储客户端发送的写入请求中获取所述待写入数据对应的分区的分区 ID。

所述获取单元 901，还用于接收所述第二存储节点返回的写入失败响应，确定所述第二存储节点写入失败；或者，

所述获取单元 901，还用于根据所述分区视图，确定所述第二存储节点的状态为故障，其中，所述分区视图中包含存储节点的状态信息。

所述处理单元 902，具体用于分配空闲的存储空间作为日志卷，用于存储分配给写入失败的存储节点的 EC 块，所述日志卷由至少一个日志块组成，所述日志块的大小与所述 EC 块的大小相同。

如图 10 所示，本发明实施例还提供了一种分布式存储系统，包括第一存储节点 1001 和第二存储节点 1002，

所述第一存储节点 1001，用于获取待写入数据以及所述待写入数据的键 key 值，刷新所述 key 值对应的版本号，对所述待写入数据进行纠删码 EC 编码，生成 EC 条带，所述 EC 条带包括 m+k 个 EC 块，其中 m 个 EC 块为数据块，k 个 EC 块为校验块，m 为大于等于 2 的正

整数，k 为自然数；

所述第一存储节点 1001，还用于查询分区视图，确定所述待写入数据所在的分区对应的备存储节点，其中，所述第一存储节点 1001 为所述分区对应的主存储节点，第二存储节点 1002 为所述分区对应的其中一个备存储节点；

所述第一存储节点 1001，还用于向各备存储节点发送写入请求，所述写入请求携带所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号，以及分配给各备存储节点的 EC 块的数据；

当所述第二存储节点 1002 写入失败时，所述第一存储节点 1001，还用于存储所述分配给所述第二存储节点 1002 的 EC 块的数据，生成与所述分配给所述第二存储节点 1002 的 EC 块对应的元数据信息，所述元数据信息包括所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号；

当所述第二存储节点 1002 故障恢复后，所述第一存储节点 1001，还用于将存储的所述分配给所述第二存储节点 1002 的 EC 块的数据以及所述元数据信息发送给所述第二存储节点 1002；

所述第二存储节点 1002，用于根据所述元数据信息写入所述分配给所述第二存储节点 1002 的 EC 块的数据。

分布式存储系统中的主存储节点对待写入数据进行 EC 编码，生成 EC 条带，将 EC 条带中的各个 EC 块分别存储在各个存储节点上，当部分存储节点由于故障导致写入 EC 块失败时，主存储节点将分配给写入失败的存储节点的 EC 块存储在本地，并生成数据重建所需的元数据信息，当存储节点故障恢复后，主存储节点将存储的分配给写入失败的存储节点的 EC 块以及该 EC 块对应的元数据信息重新发送给该存储节点，以使得故障恢复后的该存储节点完成数据重建。本申请提供的分布式存储系统中的数据重建的方案，当部分存储节点故障时，无需执行 EC 反编码以恢复故障节点上的数据，而是由主存储节点缓存分配给故障节点的 EC 块，再故障节点恢复后再将缓存的 EC 块重新发送给故障节点进行数据重建。通过上述方案，避免了存储节点故障恢复进行数据重建时执行 EC 反编码带来的计算资源消耗，同时也避免了执行 EC 反编码时传递大量数据带来的网络资源消耗。示例性的，当 EC 编码为 4+2 时，恢复 1 份数据需要 4 份数据，而本申请中，只是在故障节点恢复后，由主存储节点向故障节点重新发送 1 份数据，明显降低了网络资源消耗；进一步的，分布式存储系统中，丢失 1 份数据的概率远远大于丢失两份数据，而丢失 1 份数据时，仍然可以还原出原始数据，因此，无需立刻执行 EC 反编码将丢失的数据还原，当故障节点恢复后，采用本申请提出的数据重建的方案即可将故障节点上的数据与其他存储节点进行同步。

在图 9 和 10 对应的实施例中，存储节点是以功能单元/功能模块的形式来呈现。这里的“单元/模块”可以指特定应用集成电路（application-specific integrated circuit，ASIC），电路，执行一个或多个软件或固件程序的处理器和存储器，集成逻辑电路，和/或其他可以提供上述功能的器件。在一个简单的实施例中，本领域的技术人员可以想到存储节点可以采用图 2 所示的形式。例如，获取单元 901、处理单元 902 或发送单元 903 可以通过图 2 的处理器和存储器来实现。

本发明实施例还提供了一种计算机存储介质，用于储存为上述图 9 和 10 所示的设备所用的计算机软件指令，其包含用于执行上述方法实施例所设计的程序。通过执行存储的程序，可以实现应用分布式存储系统中数据重建的方法。

尽管在此结合各实施例对本发明进行了描述，然而，在实施所要求保护的本发明过程中，本领域技术人员通过查看所述附图、公开内容、以及所附权利要求书，可理解并实现所述公开实施例的其他变化。在权利要求中，“包括”（comprising）一词不排除其他组成部分或步骤，“一”或“一个”不排除多个的情况。单个处理器或其他单元可以实现权利要求中列举的若干项功能。相互不同的从属权利要求中记载了某些措施，但这并不表示这些措施不能组合起来产生良好的效果。

本领域技术人员应明白，本发明的实施例可提供为方法、装置（设备）、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。计算机程序存储/分布在合适的介质中，与其它硬件一起提供或作为硬件的一部分，也可以采用其他分布形式，如通过 Internet 或其它有线或无线电信系统。

本发明是参照本发明实施例的方法、装置（设备）和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中

指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

尽管结合具体特征及其实施例对本发明进行了描述，显而易见的，在不脱离本发明的精神和范围的情况下，可对其进行各种修改和组合。相应地，本说明书和附图仅仅是所附权利要求所界定的本发明的示例性说明，且视为已覆盖本发明范围内的任意和所有修改、变化、组合或等同物。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. 一种分布式存储系统中数据重建的方法，其特征在于，包括：

第一存储节点获取待写入数据以及所述待写入数据的键 key 值，刷新所述 key 值对应的版本号，对所述待写入数据进行纠删码 EC 编码，生成 EC 条带，所述 EC 条带包括 m+k 个 EC 块，其中 m 个 EC 块为数据块，k 个 EC 块为校验块，m 为大于等于 2 的正整数，k 为自然数；

所述第一存储节点查询分区视图，确定所述待写入数据所在的分区对应的备存储节点，其中，所述第一存储节点为所述分区对应的主存储节点，第二存储节点为所述分区对应的其中一个备存储节点；

所述第一存储节点向各备存储节点发送写入请求，所述写入请求携带所述待写入数据所在的分区的分区标识 ID、所述待写入数据的 key 值和版本号，以及分配给各备存储节点的 EC 块的数据；

当所述第二存储节点写入失败时，所述第一存储节点存储所述分配给所述第二存储节点的 EC 块的数据，生成与所述分配给所述第二存储节点的 EC 块对应的元数据信息，所述元数据信息包括所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号；

当所述第二存储节点故障恢复后，所述第一存储节点将存储的所述分配给所述第二存储节点的 EC 块的数据以及所述元数据信息发送给所述第二存储节点，以使得所述第二存储节点进行数据重建。

2. 如权利要求 1 所述的方法，其特征在于，在所述第一存储节点存储所述分配给所述第二存储节点的 EC 块的数据之前，所述方法还包括：

所述第一存储节点确定写入成功的存储节点的数量大于等于 m。

3. 如权利要求 1 所述的方法，其特征在于，所述第一存储节点向各备存储节点发送的写入请求中还携带有分配给各备存储节点的 EC 块的块内数据偏移和块内数据长度，以使得各存储节点根据接收到的所述写入请求中携带的所述数据偏移和块内数据长度写入所述 EC 块的数据。

4. 如权利要求 1 所述的方法，其特征在于，在所述第一存储节点将存储的所述分配给所述第二存储节点的 EC 块的数据以及所述元数据信息发送给所述第二存储节点之前，所述方法还包括：

所述第一存储节点接收所述第二存储节点发送的数据同步请求，所述数据同步请求中携带所述分区 ID；

所述第一存储节点从所述第二存储节点获取所述第二存储节点中记录的所述分区 ID 对

应的 key 值以及与所述 key 值的版本号；

所述第一存储节点将自身记录的所述分区 ID 对应的 key 值以及 key 值的版本号，与从所述第二存储节点获取的所述分区 ID 对应的 key 值以及所述 key 值的版本号，进行比对，根据比对结果确定需要进行数据重建；

所述第一存储节点根据所述元数据信息，将存储的需要进行数据重建的 key 值对应的 EC 块的数据以及所述 EC 块的元数据信息发送给所述第二存储节点进行数据重建。

5. 如权利要求 4 所述的方法，其特征在于，所述比对包括以下至少一种：

当所述第一存储节点记录的 key 值对应的版本号与从所述第二存储节点获取的所述 key 值对应的版本号一致时，无需进行数据重建；

当所述第一存储节点记录的 key 值中不包含从所述第二存储节点获取的 key 值时，则通知所述第二存储节点删除所述第一存储节点不包含的所述 key 值对应的数据；

当所述第一存储节点记录的 key 值对应的版本号大于从所述第二存储节点获取的所述 key 值对应的版本号时，执行数据重建操作；或，

当所述第一存储节点记录的 key 值中不包含在从所述第二存储节点获取的 key 值中时，则通知所述第二存储节点重建所述第二存储节点不包含的所述 key 值对应的数据。

6. 如权利要求 3 所述的方法，其特征在于，在所述第二存储节点故障恢复后，还包括：

所述第二存储节点根据所述分配给所述第二存储节点的 EC 块的块内数据偏移和块内数据长度，将所述 EC 块的数据写入到磁盘中，更新所述待写入数据的 key 值对应的版本号。

7. 如权利要求 2 所述的方法，其特征在于，还包括：

以 m 个 EC 块的大小为粒度将所述逻辑卷的存储地址进行等分，得到多个存储单元，为所述多个存储单元分配数据偏移标识。

8. 如权利要求 7 所述的方法，其特征在于，在所述第一存储节点获取待写入数据之前，所述方法还包括：

所述存储客户端接收上层应用发送的存储指令，所述存储指令中携带待存储的数据，以及存储所述待存储的数据的所述逻辑卷的卷标识、数据偏移和数据长度；

所述存储客户端确定存储所述待存储的数据的地址范围对应的至少一个存储单元，将每个存储单元对应的部分待存储的数据作为一次写入操作中需要写入到分布式存储系统的所述待写入数据。

9. 如权利要求 1 所述的方法，其特征在于，在第一存储节点查询分区视图之前，还包括：

第一存储节点使用一致性哈希算法计算所述待写入数据的 key 值对应的哈希值，确定所述哈希值所属的分区的分区 ID；或者，

第一存储节点从存储客户端获取所述待写入数据对应的分区的分区 ID。

10. 如权利要求 1 所述的方法，其特征在于，第一存储节点确定所述第二存储节点写入失败包括：

所述第一存储节点接收所述第二存储节点返回的写入失败响应，确定所述第二存储节点写入失败；或者，

所述第一存储节点根据所述分区视图，确定所述第二存储节点的状态为故障，其中，所述分区视图中包含存储节点的状态信息。

11. 如权利要求 1 所述的方法，其特征在于，所述第一存储节点存储所述分配给所述第二存储节点的 EC 块的数据包括：

所述第一存储节点分配空闲的存储空间作为日志卷，用于存储分配给写入失败的存储节点的 EC 块，所述日志卷由至少一个日志块组成，所述日志块的大小与所述 EC 块的大小相同。

12. 一种存储节点，其特征在于，包括：

获取单元，用于获取待写入数据以及所述待写入数据的 key 值，刷新所述 key 值对应的版本号，对所述待写入数据进行 EC 编码，生成 EC 条带，所述 EC 条带包括 m+k 个 EC 块，其中 m 个 EC 块为数据块，k 个 EC 块为校验块，m 为大于等于 2 的正整数，k 为自然数；

处理单元，用于查询分区视图，确定所述待写入数据所在的分区对应的备存储节点，其中，所述第一存储节点为所述分区对应的主存储节点，第二存储节点为所述分区对应的其中一个备存储节点；

发送单元，用于向各备存储节点分别发送写入请求，所述写入请求携带所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号，以及分配给各备存储节点的 EC 块的数据；

所述处理单元，用于确定所述第二存储节点写入失败时，存储所述分配给所述第二存储节点的 EC 块的数据，生成与所述分配给所述第二存储节点的 EC 块对应的元数据信息，所述元数据信息包括所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号；

当所述第二存储节点故障恢复后，所述发送单元，还用于将存储的所述分配给所述第二存储节点的 EC 块的数据以及所述元数据信息发送给所述第二存储节点，以使得所述第二存储节点进行数据重建。

13. 如权利要求 12 所述的存储节点，其特征在于，所述处理单元，还用于确定写入成功的存储节点的数量大于等于 m。

14. 如权利要求 12 所述的存储节点，其特征在于，

所述获取单元，还用于接收所述第二存储节点发送的数据同步请求，所述数据同步请求中携带所述分区 ID，从所述第二存储节点获取所述第二存储节点中记录的所述分区 ID 对应的一个或多个 key 值，以及与所述 key 值的版本号；

所述处理单元，还用于将自身记录的所述分区 ID 对应的 key 值以及各 key 值的版本号，与从所述第二存储节点获取的所述分区 ID 对应的一个或多个 key 值，以及与所述 key 值的版本号，进行比对，根据比对结果确定需要进行数据重建；

所述发送单元，具体用于根据所述元数据信息，将存储的需要进行数据重建的 key 值对应的 EC 块的数据以及所述 EC 块的元数据信息发送给所述第二存储节点进行数据重建。

15. 如权利要求 14 所述的存储节点，其特征在于，所述处理单元，具体用于执行以下至少一种比对处理：

当本存储节点记录的 key 值对应的版本号与从所述第二存储节点获取的所述 key 值对应的版本号一致时，无需进行数据重建；

当本存储节点记录的 key 值中不包含从所述第二存储节点获取的 key 值时，则通知所述第二存储节点删除所述本存储节点不包含的所述 key 值对应的数据；

当所述本存储节点记录的 key 值对应的版本号大于从所述第二存储节点获取的所述 key 值对应的版本号时，执行数据重建操作；或，

当所述本存储节点记录的 key 值中不包含在从所述第二存储节点获取的 key 值中时，则通知所述第二存储节点重建所述第二存储节点不包含的所述 key 值对应的数据。

16. 如权利要求 12 所述的存储节点，其特征在于，

所述处理单元，还用于以 m 个 EC 块的大小为粒度将所述逻辑卷的存储地址进行等分，得到多个存储单元，为所述多个存储单元分配数据偏移标识。

17. 如权利要求 12 所述的存储节点，其特征在于，

所述获取单元，具体用于使用一致性哈希算法计算所述待写入数据的 key 值对应的哈希值，确定所述哈希值所属的分区的分区 ID；或者，

所述获取单元，具体用于从存储客户端发送的写入请求中获取所述待写入数据对应的分区的分区 ID。

18. 如权利要求 12 所述的存储节点，其特征在于，

所述获取单元，还用于接收所述第二存储节点返回的写入失败响应，确定所述第二存储节点写入失败；或者，

所述获取单元，还用于根据所述分区视图，确定所述第二存储节点的状态为故障，其中，所述分区视图中包含存储节点的状态信息。

19. 如权利要求 12 所述的存储节点，其特征在于，

所述处理单元，具体用于分配空闲的存储空间作为日志卷，用于存储分配给写入失败的存储节点的 EC 块，所述日志卷由至少一个日志块组成，所述日志块的大小与所述 EC 块的大小相同。

20. 一种分布式存储系统，其特征在于，包括第一存储节点和第二存储节点，

所述第一存储节点，用于获取待写入数据以及所述待写入数据的键 key 值，刷新所述 key 值对应的版本号，对所述待写入数据进行纠删码 EC 编码，生成 EC 条带，所述 EC 条带包括 m+k 个 EC 块，其中 m 个 EC 块为数据块，k 个 EC 块为校验块，m 为大于等于 2 的正整数，k 为自然数；

所述第一存储节点，还用于查询分区视图，确定所述待写入数据所在的分区对应的备存储节点，其中，所述第一存储节点为所述分区对应的主存储节点，第二存储节点为所述分区对应的其中一个备存储节点；

所述第一存储节点，还用于向各备存储节点发送写入请求，所述写入请求携带所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号，以及分配给各备存储节点的 EC 块的数据；

当所述第二存储节点写入失败时，所述第一存储节点，还用于存储所述分配给所述第二存储节点的 EC 块的数据，生成与所述分配给所述第二存储节点的 EC 块对应的元数据信息，所述元数据信息包括所述待写入数据所在的分区的分区 ID、所述待写入数据的 key 值和版本号；

当所述第二存储节点故障恢复后，所述第一存储节点，还用于将存储的所述分配给所述第二存储节点的 EC 块的数据以及所述元数据信息发送给所述第二存储节点；

所述第二存储节点，用于根据所述元数据信息写入所述分配给所述第二存储节点的 EC 块的数据。

说 明 书 附 图

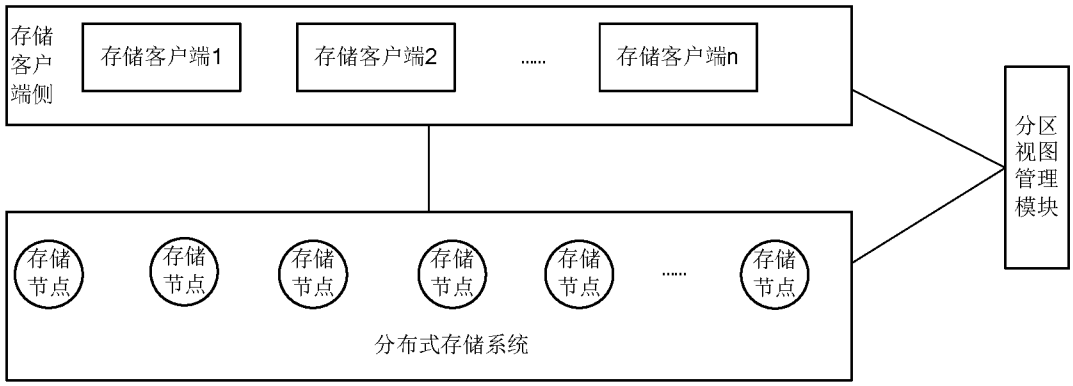

图 1

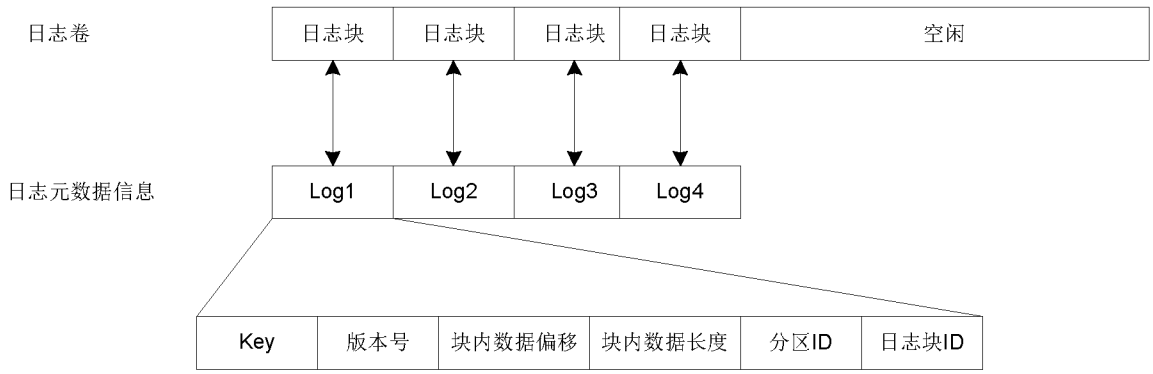

图 2

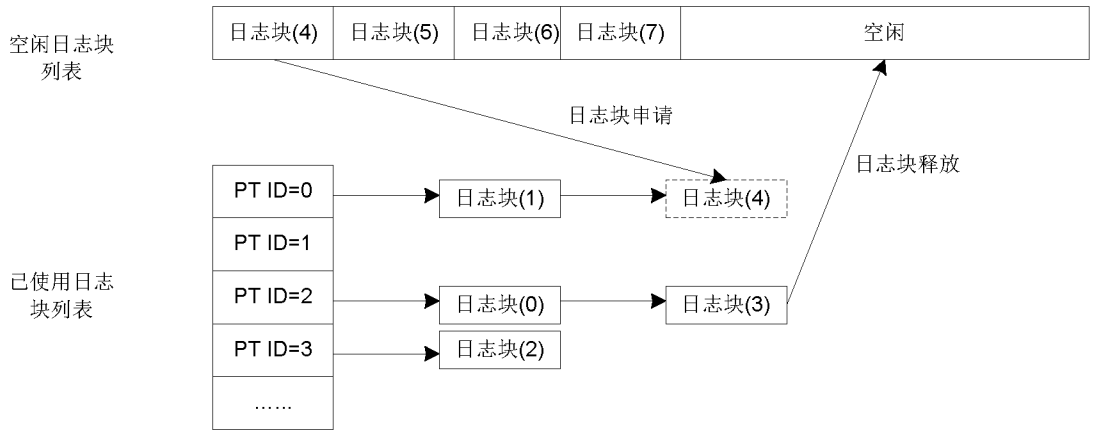

图 3

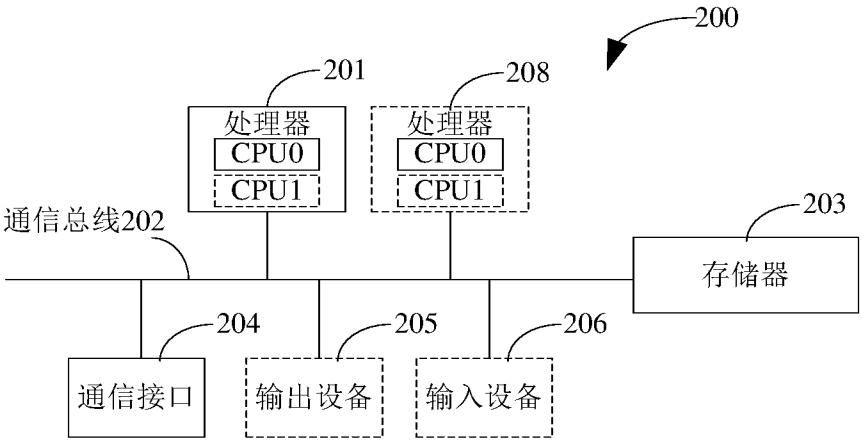
200
201
208
处理器
CPU0
CPU1
处理器
CPU0
CPU1
通信总线202
203
存储器
204
205
206
通信接口
输出设备
输入设备

图 4

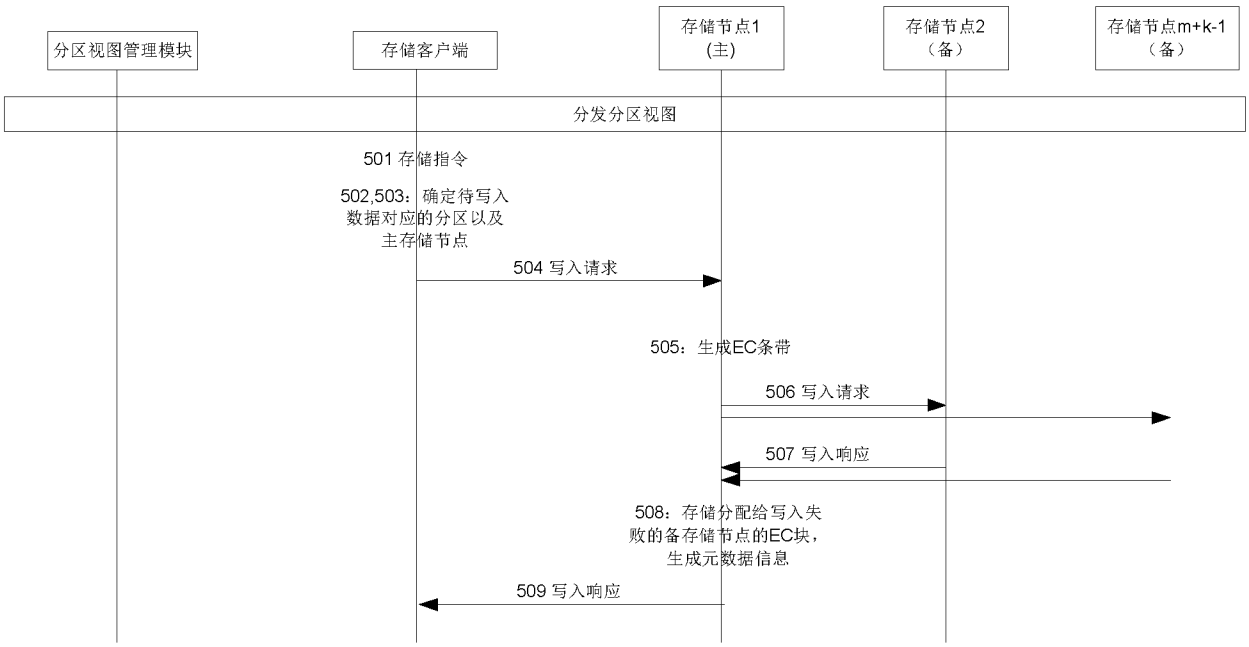
分区视图管理模块
存储客户端
存储节点1
(主)
存储节点2
（备）
存储节点m+k-1
（备）
分发分区视图
501 存储指令
502,503：确定待写入
数据对应的分区以及
主存储节点
504 写入请求
505：生成EC条带
506 写入请求
507 写入响应
508：存储分配给写入失
败的备存储节点的EC块，
生成元数据信息
509 写入响应

图 5


2M
逻辑卷
.....
.....
0
2M
4M
6M
8M
2nM

图 6

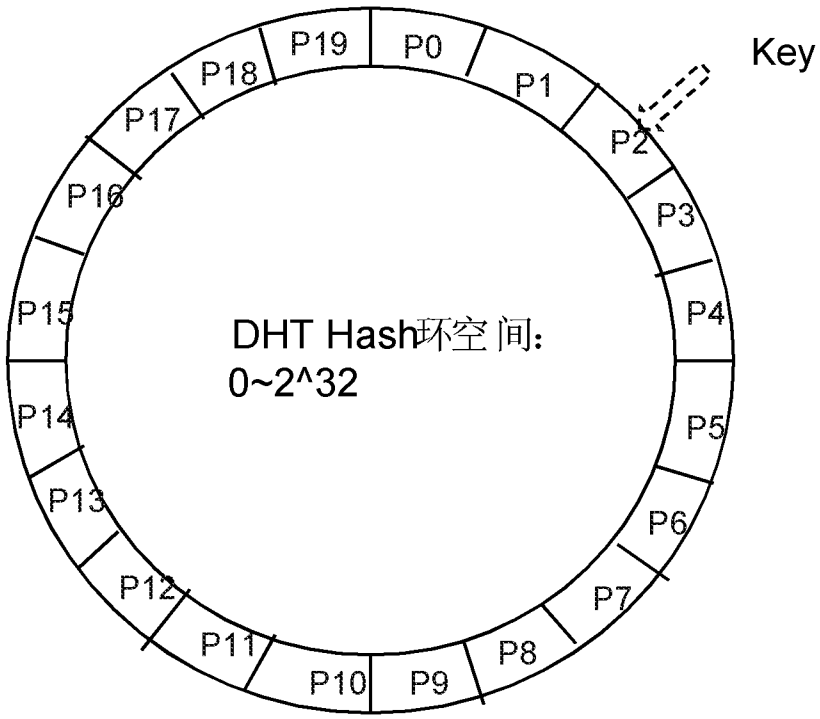
Key
P0
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
P17
P18
P19
DHT Hash环空间：
0~2^32

图 7

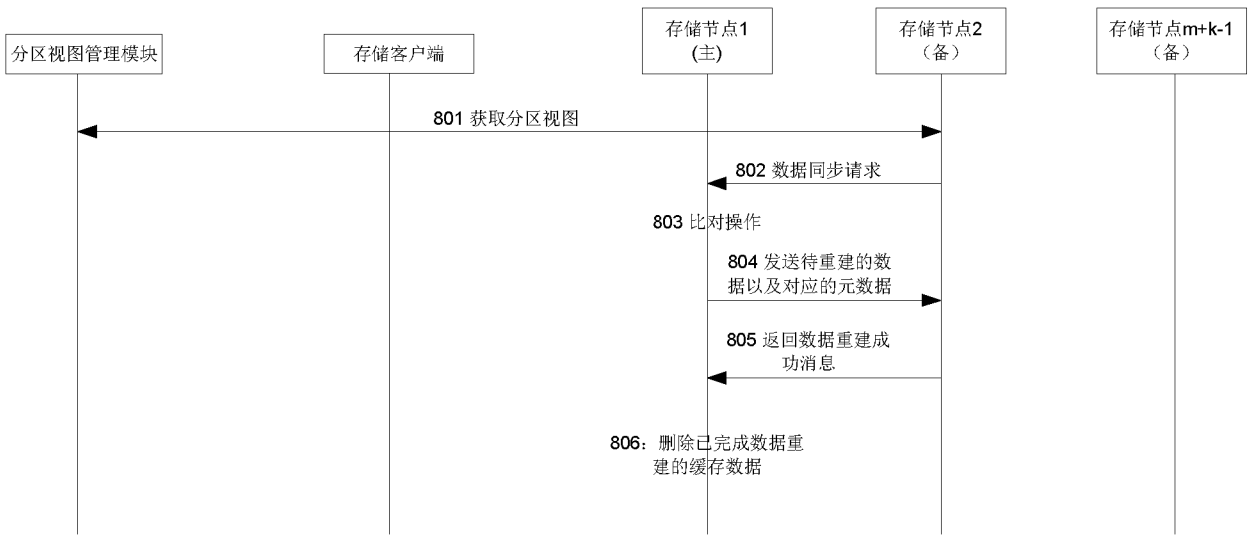
分区视图管理模块
存储客户端
存储节点1
(主)
存储节点2
（备）
存储节点m+k-1
（备）
801 获取分区视图
802 数据同步请求
803 比对操作
804 发送待重建的数据以及对应的元数据
805 返回数据重建成功消息
806：删除已完成数据重建的缓存数据

图 8

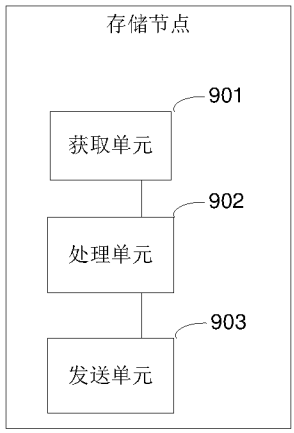

图 9

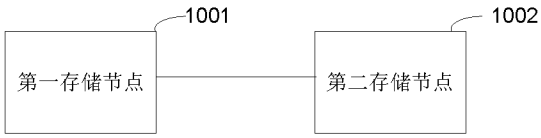

图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/100078

A. CLASSIFICATION OF SUBJECT MATTER

G06F 11/10 (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)

G06F

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)

CNPAT, CNKI, WPI, EPODOC, IEEE: data segment, erasure code, prepare, identify, key, EC, distribut+, memory, storage, segment, node, coding, resume, rebuild, checkout, verify, bake, subarea, ID, edition

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105095013 A (HUAWEI TECHNOLOGIES CO., LTD.), 25 November 2015 (25.11.2015), abstract, claims 5 and 13, and description, paragraphs 146-197	1-20
A	CN 104580324 A (HUAWEI TECHNOLOGIES CO., LTD.), 29 April 2015 (29.04.2015), the whole document	1-20
A	CN 103729352 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.), 16 April 2014 (16.04.2014), the whole document	1-20
A	WO 2009049023 A2 (BLUEARC UK LIMITED), 16 April 2009 (16.04.2009), the whole document	1-20

☐ 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
25 July 2016 (25.07.2016)	**19 August 2016 (19.08.2016)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **ZHANG, Qian** Telephone No.: (86-10) **62413681**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/100078

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105095013 A	25 November 2015	None	
CN 104580324 A	29 April 2015	None	
CN 103729352 A	16 April 2014	None	
WO 2009049023 A2	16 April 2009	EP 2212812 A2	04 August 2010
		US 2009100110 A1	16 April 2009
		US 2009183056 A1	16 July 2009
		US 2009182785 A1	16 July 2009

国际检索报告

国际申请号
PCT/CN2015/100078

A. 主题的分类

G06F 11/10(2006.01)i

按照国际专利分类(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)

G06F

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））

CNPAT,CNKI,WPI,EPODOC,IEEE:分布式，存储，数据段，节点，编码，恢复，重建，纠删码，校验，备，分区，标识，版本，key,EC,distribut+,memory,storage,segment,node,coding,resume,rebuild,checkout,verify,bake,subarea,ID,edition

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
A	CN 105095013 A (华为技术有限公司) 2015年 11月 25日 (2015-11-25) 摘要，权利要求5，13，说明书第146-197段	1-20
A	CN 104580324 A (华为技术有限公司) 2015年 4月 29日 (2015-04-29) 全文	1-20
A	CN 103729352 A (腾讯科技深圳有限公司) 2014年 4月 16日 (2014-04-16) 全文	1-20
A	WO 2009049023 A2 (BLUEARC UK LIMITED) 2009年 4月 16日 (2009-04-16) 全文	1-20

☐ 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（如具体说明的）

“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
2016年 7月 25日	2016年 8月 19日

ISA/CN的名称和邮寄地址	受权官员
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	张千
传真号 (86-10)62019451	电话号码 (86-10)62413681

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2015/100078

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	105095013	A	2015年 11月 25日	无			
CN	104580324	A	2015年 4月 29日	无			
CN	103729352	A	2014年 4月 16日	无			
WO	2009049023	A2	2009年 4月 16日	EP	2212812	A2	2010年 8月 4日
				US	2009100110	A1	2009年 4月 16日
				US	2009183056	A1	2009年 7月 16日
				US	2009182785	A1	2009年 7月 16日