(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
2022 年 1 月 13 日 (13.01.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/007968 A1

(51) 国际专利分类号:
G06F 3/06 (2006.01) *G06F 11/00* (2006.01)

(21) 国际申请号: PCT/CN2021/105640

(22) 国际申请日: 2021 年 7 月 12 日 (12.07.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010661972.0 2020年7月10日 (10.07.2020) CN
202011148485.0 2020年10月23日 (23.10.2020) CN

(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: 吴 祥 (**WU, Xiang**); 中国广东省深圳市 龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 罗 小 东 (**LUO, Xiaodong**); 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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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,

(54) **Title:** STRIPE MANAGEMENT METHOD, STORAGE SYSTEM, STRIPE MANAGEMENT APPARATUS, AND STORAGE MEDIUM

(54) **发明名称**: 分条管理方法、存储系统、分条管理装置及存储介质

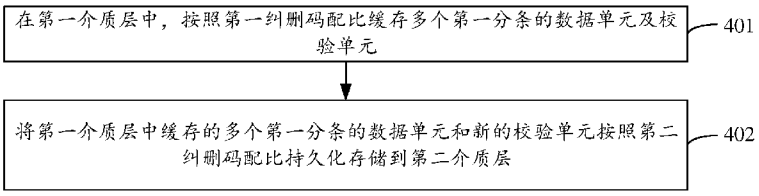

图 4

401 In a first medium layer, cache data units and check units of multiple first stripes according to a first erasure code ratio
402 Persistently store the data units of the multiple first stripes cached in the first medium layer and new check units into a second medium layer according to a second erasure code ratio

(57) **Abstract:** A stripe management method, a storage system, a stripe management apparatus, and a storage medium, which relate to the technical field of data storage. The method comprises: acquiring check units in multiple first stripes, the first stripe complying with a first erasure code ratio; and generating new check units according to the check units of the multiple first stripes, the new check units and data units in the multiple first stripes belonging to a new stripe, the new stripe complying with a second erasure code ratio, and the quantity of data units corresponding to the first erasure code ratio being less than the quantity of data units corresponding to the second erasure code.

(57) **摘要**: 一种分条管理方法、存储系统、分条管理装置及存储介质，属于数据存储技术领域。该方法包括：获取多个第一分条中的校验单元；其中，所述第一分条遵从第一纠删码配比；根据所述多个第一分条的校验单元生成新的校验单元；其中，所述新的校验单元与所述多个第一分条中的数据单元属于新的分条；所述新的分条遵从第二纠删码配比；其中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数。

WO 2022/007968 A1

[见续页]

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)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

分条管理方法、存储系统、分条管理装置及存储介质

本申请要求于2020年7月10日提交中国专利局、申请号为202010661972.0、申请名称为“数据存储方法以及存储设备”的中国专利申请以及于2020年10月23日提交中国专利局、申请号为202011148485.0、申请名称为“分条管理方法、存储系统、分条管理装置及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及数据存储技术领域，特别涉及一种分条管理方法、存储系统、分条管理装置及存储介质。

背景技术

在存储系统中，提升有效存储容量是降低存储成本的有力武器，而纠删码（erasure code，EC）技术就能够提升存储系统的有效存储容量。当前，EC技术被广泛的应用于存储系统中。EC技术主要是通过纠删码算法将数据单元进行编码得到校验单元，并将数据单元和校验单元一并存储起来，以达到容错的目的。存储系统中为了降低成本，采用EC技术时编码时数据单元的个数越大，存储空间利用率越高，但是数量单元的个数较大时凑满EC条带比较困难，从而影响数据存储可靠性。

发明内容

本申请实施例提供了一种分条管理方法、存储系统、分条管理装置及存储介质，能够提高存储系统的存储空间利用率的同时提高数据存储可靠性。所述技术方案如下：

第一方面，提供了一种分条管理方法，应用于存储系统中，所述方法包括：

获取多个第一分条中的校验单元；其中，所述第一分条遵从第一纠删码配比；

根据所述多个第一分条的校验单元生成新的校验单元；其中，所述新的校验单元与所述多个第一分条中的数据单元属于新的分条；所述新的分条遵从第二纠删码配比；其中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数。也就是说，先采用较小的纠删码配比来存储数据，然后再转换成较大纠删码配比较大的纠删码配比来存储数据。通过较小的纠删码配比来存储数据，容易凑满分条，降低了写放大，提高了存储空间利用率。另外，采用较大的纠删码配比来存储数据，能够减少冗余数据在存储空间中的占比，从而提高存储空间利用率。因此，能够提高存储系统的存储空间利用率的同时提高数据存储可靠性。同时，新的分条中的校验单元由多个第一分条的校验单元生成，多个第一分条中的数据单元不参与运算，从而节约了计算资源。

一种情况下，所述新的分条中的校验单元的个数与所述第一分条中的校验单元的个数相同。

在一种实现方式中，所述多个第一分条包含至少一个在所述存储系统中未持久化存储的第一分条和至少一个已经在所述存储系统持久化存储的第一分条；所述获取多个第一分条中的校验单元，具体包括：

读取所述至少一个已经在所述存储系统持久化存储的第一分条中的校验单元；

读取所述至少一个在所述存储系统中未持久化存储的第一分条中的校验单元。

进一步地，所述方法还包括：

持久化存储所述至少一个在所述存储系统中未持久化存储的第一分条中的数据单元和所述新的校验单元。

在一种实现方式中，所述多个第一分条为所述存储系统中未持久化存储的第一分条；所述方法还包括：

持久化存储所述多个第一分条中的数据单元以及所述新的校验单元。

第二方面，提供了一种存储系统，所述存储系统包含一个或多个处理器，所述一个或多个处理器用于实现上述第一方面的数据存储方法。

第二方面所提供的存储系统既可以是一种分布式的存储系统，也可以是一种集中式的存储系统。

第三方面提供一种分条管理装置，所述存储设备应用于存储系统中，所述分条管理装置包含多个单元，所述多个单元用于实现上述第一方面的数据存储方法。

第四方面，提供了一种计算机可读存储介质，

所述计算机可读存储介质包含计算机程序指令，存储系统中的一个或多个中央处理器执行所述计算机程序指令使得所述存储系统执行上述第一方面所述的数据存储方法。

第五方面，提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述第一方面所述的数据存储方法。

上述第二方面、第三方面、第四方面、第五方面和第六方面所获得的技术效果与第一方面中对应的技术手段获得的技术效果近似，在这里不再赘述。

附图说明

图 1 是本申请实施例提供的一种存储系统架构图；

图 2 是本申请实施例提供的另一种存储系统架构图；

图 3 是本申请实施例提供的一种存储设备的系统架构图；

图 4 是本申请实施例提供的一种分条管理方法的流程图；

图 5 是本申请实施例提供的一个第一校验矩阵中的各个单元在第一介质层中的分布示意图；

图 6 是本申请实施例提供的 w 个第一校验矩阵中的单元在第一介质层中的分布示意图；

图 7 是本申请实施例提供第二校验矩阵中的单元在第二介质层中的分布示意图；

图 8 是本申请实施例提供的一种根据 w 个第一校验矩阵合并得到第二校验矩阵的原理示意图；

图 9 是本申请实施例提供的一种获得第二校验矩阵的校验单元的示意图；

图 10 是本申请实施例提供的一种在第二介质层中按照第二纠删码配比写入数据的示意图；

图 11 是本申请实施例提供的一种分条管理装置的结构示意图。

具体实施方式

为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

在对本申请实施例进行详细的解释说明之前，先对本申请实施例涉及的系统架构进行介绍。

图 1 是本申请实施例提供的一种存储系统的结构示意图。如图 1 所示，该存储系统包括计算机节点集群和存储节点集群。其中，计算节点集群包括一个或多个计算节点 10（图 1 中示出了两个计算节点 10，但不限于两个计算节点 10）。计算节点 10 是用户侧的一种计算设备，如服务器、台式计算机等。在硬件层面，计算节点 10 中设置有中央处理器和内存（图 1 中未示出）。在软件层面，计算节点 10 上运行有应用程序（application）101（简称应用）和客户端程序 102（简称客户端）。应用 101 是对用户呈现的各种应用程序的统称。客户端 102 用于接收由应用 101 触发的数据访问请求，并且与存储节点 20 交互，向存储节点 20 发送数据访问请求。客户端 102 还用于接收来自存储节点的数据，并向应用 101 转发该数据。应理解的是，当客户端 102 是软件程序时，客户端 102 的功能由计算节点 10 所包含的中央处理器运行内存中的程序来实现。客户端 102 也可以由位于计算节点 10 内部的硬件组件来实现。计算节点集群中的任意一个客户端 102 可以访问存储节点集群中的任意一个存储节点 20。

存储节点集群包括一个或多个存储节点 20（图 1 中示出了三个存储节点 20，但不限于三个存储节点 20），各个存储节点 20 之间可以互联。存储节点如服务器、台式计算机或者存储阵列的控制器、硬盘框等。在功能上，存储节点 20 主要用于对数据进行计算或处理等。

在硬件上，如图 1 所示，存储节点 20 至少包括存储器、网卡和一个或多个中央处理器。

其中，中央处理器（central processing unit，CPU），用于处理来自存储节点 20 外部的数据，或者存储节点 20 内部生成的数据。

存储器，是指用于存储数据的装置。在本申请实施例中，存储器可以是内存，也可以是硬盘。其中，内存是指与处理器直接交换数据的内部存储器，它可以随时读写数据，而且速度很快，作为操作系统或其他正在运行中的程序的临时数据存储器。内存包括一种或多种类型的存储器，例如内存既可以是随机存取存储器，也可以是只读存储器（Read Only Memory, ROM）。举例来说，随机存取存储器可以是动态随机存取存储器（Dynamic Random Access Memory, DRAM），也可以是存储级存储器（Storage Class Memory, SCM）。DRAM 是一种半导体存储器，与大部分随机存取存储器（Random Access Memory，RAM）一样，属于一种易失性存储器（volatile memory）设备。SCM 是一种同时结合传统储存装置与存储器特性的复合型储存技术，SCM 能够提供比硬盘更快速的读写速度，但运算速度上比 DRAM 慢，在成本上也比 DRAM 更为便宜。需

要说明的是，处理器可以直接访问内存，例如，如图 2 中所示，处理器可以直接访问 DRAM 和 SCM。

然而，DRAM 和 SCM 在本实施例中只是示例性的说明，在一些可能的情况中，内存可以只包含 DRAM 和 SCM 中的其中一种。或者，内存还可以包括其他随机存取存储器，例如静态随机存取存储器（Static Random Access Memory，SRAM）等。而对于只读存储器，举例来说，可以是可编程只读存储器（Programmable Read Only Memory, PROM）、可抹除可编程只读存储器（Erasable Programmable Read Only Memory, EPROM）等。另外，内存还可以是双列直插式存储器模块或双线存储器模块（Dual In-line Memory Module，简称 DIMM），即由动态随机存取存储器（DRAM）组成的模块。在后续的实施例中，均以内存包括一种存储器为例那个说明，但这并不构成对内存包括的存储器类型数量的限制。

硬盘读写数据的速度比内存慢，通常用于持久性地存储数据。以存储节点 20a 为例，其内部设置一个或多个硬盘；或者，在存储节点 20 的外部挂载一个硬盘框（如图 2 所示），在硬盘框中设置多个硬盘。无论哪一种部署方式，这些硬盘都可以视作存储节点 20 所包含的硬盘。其中，这里的硬盘可以是指物理硬盘，也可以是指一个包括多个物理硬盘的逻辑域或故障域，本申请实施例对此不作限定。另外，物理硬盘的类型为固态硬盘、机械硬盘，或者其他类型的硬盘。

需要说明的是，内存包括的存储器与硬盘是完全不同的两种存储介质，二者的性能完全不同。其中，相较于硬盘，内存的数据读取速度更快，时延更小，也即，内存的性能高于硬盘的性能。基于此，在本申请实施例中，如图 1 和 2 中所示，将各个存储节点 20 中的内存称为第一介质层，将各个存储节点 20 中的硬盘称为第二介质层。其中，第一介质层的性能高于第二介质层。当然，也可以将各个存储节点 20 中的内存称为第二介质层，将各个存储节点 20 中的硬盘称为第一介质层。此时，第一介质层的性能低于第二介质层。可选地，当内存包括多种性能不同的存储介质时，也可以将内存中的每种存储介质作为一个介质层，例如，如图 1 和图 2 中，各个存储节点中的 DRAM 组成一个介质层，SCM 组成一个介质层，硬盘组成一个介质层。可选地，在一些可能的情况中，第一介质层可以是固态硬盘（solid state disk，SSD），第二介质层可以为硬盘驱动器（hard disk drvie，HDD）。后续实施例中将以内存中包括一种存储介质，将各个存储节点的内存中的该种存储介质作为第一介质层、硬盘作为第二介质层为例进行说明。

网卡用于与其他存储节点进行通信，或者，用于与该存储节点耦合的硬盘框进行通信。另外，网卡可以直接访问存储节点的内存，如图 2 所示，网卡可以直接访问 DRAM 和 SCM。

在本发明另一实施例中，存储系统不包含计算节点。在本发明另一实施例中，存储系统中的计算节点和存储节点可以在同一个节点。关于存储系统的具体形态，本发明对此不作限定。另外，本发明实施例中的存储节点中的一个或多个 CPU 的功能可以由现场可编程门阵列（Field Programmable Gate Array, FPGA）或者专用集成电路（Application-specific integrated circuit，ASIC）等或者上述多种组合来实现。本发明实施例中将上述各种实现方式统称为由一个或多个处理器实现。

图 3 是本申请实施例提供的另一种存储系统的结构示意图。图 3 所示的存储系统为一个存储阵列，该存储阵列包括至少一个控制器（如图 3 所示的控制器 11）和多个硬盘 22。控制器 11 通过存储区域网络（英文：storage area network，SAN）与主机（图中未示出）连接。控制器 11 可以是一种计算设备，如服务器、台式计算机等等。在控制器 11 上安装有操作系统以及应用程序。控制器 11 可以接收来自主机的输入输出（I/O）请求。控制器 11 还可以存储 I/O 请求中携带的数据（如果有的话），并且将该数据写入硬盘 22 中。其中，硬盘 22 为机械硬盘或固态硬盘，固态硬盘是以闪存（英文：flash memory） 芯片为介质的存储器，又名固态驱动器（Solid State Drive，SSD）。

图 3 仅是示例性说明，在实际应用中存储阵列可包含两个或两个以上控制器，每个控制器的物理结构和功能与控制器 11 类似，并且本实施例并不限定控制器之间，以及任意一个控制器与硬盘 22 之间的连接方式。只要各个控制器之间，以及各个控制器和硬盘 22 之间能够相互通信。另外，在本实施例中，硬盘可以是指物理硬盘，也可以是指一个包括多个物理硬盘的逻辑域或故障域，本申请实施例对此不作限定。

如图 3 所示，控制器 11 包括接口卡 110、一个或多个处理器 112 和接口卡 113。

接口卡 110 用于和主机通信，控制器 11 通过接口卡 110 接收主机的操作指令。处理器 112 可能是中央处理器（英文：central processing unit，CPU）。在本申请实施例中，处理器 112 用于接收来自主机的 I/O 请求、处理所述 I/O 请求。所述 I/O 请求是写数据请求或者读数据请求，处理器 112 还可以将写数据请求中的数据发送给硬盘 22。接口卡 113，用于和硬盘 22 通信，控制器 11 通过接口卡 113 将写数据请求（包括数据、数据的逻辑地址以及数据的虚拟地址）发送给硬盘 22 存储。本发明实施例中的处理器 112 还可以由现场可编程门阵列（Field Programmable Gate Array, FPGA）或者专用集成电路（Application-specific integrated circuit，ASIC）等或者上述多种组合来实现。本发明实施例中将上述各种实现方式统称为由一个或多个处理器实现。

可选地，控制器 11 还包括内存 111。内存 111 用于临时存储从主机接收的数据或从硬盘 22 读取的数据。控制器 11 接收主机发送的多个写数据请求时，可以将多个写数据请求中的数据暂时保存在内存 111 中。当内存 111 的容量达到一定阈值时，将内存 111 存储的数据、数据的虚拟地址以及为数据分配的逻辑地址发送给硬盘 22。硬盘 22 存储接收到的数据。内存 111 包括易失性存储器，闪存芯片或其组合。易失性存储器例如为随机访问存储器(英文：random-access memory，RAM)。闪存芯片例如软盘、硬盘、光盘等各种可以存储程序代码的机器可读介质。内存 111 具有保电功能，保电功能是指系统发生掉电又重新上电时，内存 111 中存储的数据也不会丢失。

需要说明的是，控制器 11 包括的内存 111 与硬盘 22 为完全不同的两种存储介质。其中，相较于硬盘，内存的数据读取速度更快，时延更小，也即，内存的性能高于硬盘的性能。在本申请实施例中，将性能更高的内存 111 称为第一介质层，将性能相对内存而言较低的多个硬盘 22 称为第二介质层，也即，第一介质层的性能高于第二介质层。或者，将性能更高的内存 111 称为第二介质层，将性能相对内存而言较低的多个硬盘 22 称为第一介质层，此时，第一介质层的性能低于第二介质层。

纠删码是一种数据冗余技术，相对于多副本策略，纠删码具有更高的磁盘利用率。例如 Reed-Solomon 码就是一种常见的纠删码。纠删码技术主要是通过纠删码算法将原

始的数据进行编码得到冗余，并将数据和冗余一并存储起来，以达到容错的目的。其基本思想是将 n 块原始的数据元素（数据单元）通过一定的计算，得到 m 块冗余元素（校验单元），磁盘利用率为 n/(n+m)。对于这 n+m 块的元素，当其中任意的 m 块元素出错（包括原始的数据元素和冗余元素）时，均可以通过对应的重构算法恢复出原来的 n 块数据元素。生成校验的过程被成为编码（encoding），恢复丢失的数据元素的过程被称为解码（decoding）。本申请所提到的纠删码配比是指数据元素 n 与冗余元素 m 的比值。基于纠删码技术的 n 块数据元素和 m 块冗余元素属于一个分条；其中，数据元素也称为数据单元，冗余元素也称为校验单元。接下来对本申请实施例提供的数据存储方法进行介绍。

图 4 是本申请实施例提供的一种分条管理方法的流程图。该方法可以应用于图 1 或图 2 所示的存储系统中，也可以应用于图 3 所示的存储系统中。参见图 4，该方法包括以下步骤：

步骤 401：在第一介质层中，按照第一纠删码配比缓存多个第一分条的数据单元及校验单元。

由前述图 1 至图 3 中的介绍可知，存储系统中包括多种不同的存储介质，本申请实施例中以存储系统包括两种不同的存储介质且第一介质层的性能高于第二介质层的性能为例对该数据存储方法进行解释说明。例如，第一介质层是 DRAM，第二介质层是硬盘。在第一介质层中，按照第一纠删码配比存储数据，即遵从第一纠删码配比的分条缓存在 DRAM，即分条中的数据单元和校验单元均缓存在 DRAM。所谓第一介质层的性能高于第二介质层的性能是指第一介质层比第二介质层的读写速度更快、时延更小。

在本申请实施例中，针对第一介质层和第二介质层，可以分别获取第一介质层对应的第一纠删码配比和第二介质层对应的第二纠删码配比，进而在第一介质层中，按照第一纠删码配比缓存第一分条的数据单元及校验单元。即按照第一纠删码配比将接收的数据划分为数据单元，基于纠删码算法得到相应的校验单元，从而得到遵从第一纠删码配比的第一分条。在本发明实施例中，当分条中的数据单元和校验单元缓存在 DRAM 中时，该分条为在存储系统中未持久化存储的分条。当分条中的数据单元和校验单元存储在存储系统的非易失性存储介质上时，该分条为在存储系统中持久化存储的分条。进一步地，本发明实施例中，当第一介质层包含 SCM 时，当分条中的数据单元和校验单元存储在第一介质层的 SCM 时，仍称该分条为在存储系统中未持久化存储的分条，即分条缓存在 SCM 中。另一种实现，当第一介质层不包含 SCM，第二介质层包含 SCM，当分条中的数据单元和校验单元存储在第二介质层的 SCM 时，仍称该分条为在存储系统中持久化存储的分条。

其中，如果该数据存储方法应用于图 1 或图 2 所示的存储系统中，则可以由存储系统中的管理节点获取第一纠删码配比和第二纠删码配比。其中，该管理节点为该存储系统的多个存储节点中的任一存储节点，或者是，该管理节点为该存储系统中独立于存储节点之外用于对各个存储节点进行管理的一个节点。另外，该管理节点可以在初始化时获取第一纠删码配比和第二纠删码配比，也可以在存储系统运行过程中获取第一纠删码配比和第二纠删码配比，本申请实施例对此不做限定。如果该数据存储方

法应用于图 3 所示的存储系统中，则可以由存储系统中的控制器来获取第一纠删码配比和第二纠删码配比。接下来以管理节点获取第一纠删码配比和第二纠删码配比为例来进行介绍。

在一种可能的实现方式中，管理节点根据该存储系统的拓扑结构和容错能力确定第一纠删码配比或第二纠删码配比。其中，第一纠删码配比和第二纠删码配比可以均根据存储系统的拓扑结构和容错能力确定得到，或者，第一纠删码配比根据存储系统的拓扑结构和容错能力确定得到，第二纠删码配比根据第一纠删码配比获得，或者，第二纠删码配比根据存储系统的拓扑结构和容错能力确定得到，第一纠删码配比根据第二纠删码配比获得，本申请实施例对此不作限定。另外，拓扑结构用于指示存储系统所包含的存储节点的数量，容错能力用于指示存储系统容忍出错的存储节点的数量。其中，该存储系统容忍出错的存储节点的数量等于第一纠删码配比对应的校验单元的数量，或者第二纠删码配比对应的校验单元的数量。

其中，管理节点首先获取存储系统的拓扑结构。示例性地，管理节点中可以存储有存储系统的拓扑结构，或者，接收由其他设备发送的该存储系统的拓扑结构，或者，接收用户输入的存储系统的拓扑结构。该拓扑结构能够指示存储系统的组成，例如，该存储系统内所包含的存储节点数量，每个存储节点包括的子节点的数量。其中，当存储节点为一台服务器时，存储节点的子节点的数量是指服务器包括的物理硬盘的数量或者对相应存储节点包括的物理硬盘进行划分得到的硬盘逻辑域的数量。当存储节点为一个机柜时，存储节点的子节点的数量是指机柜内包括的服务器的数量。通常，一个机柜包括多个服务器。

例如，假设存储系统内包括 4 个服务器，每个服务器包括 60 个物理硬盘，其中，每 15 个物理硬盘划分为 1 个硬盘逻辑域，则根据该拓扑结构可知，该存储系统包括 4 个存储节点，每个服务器即为一个存储节点。每个存储节点包括 4 个硬盘逻辑域，也即，每个存储节点包括的子节点的数量为 4。

除了获取存储系统的拓扑结构，管理节点还获取该存储系统的安全级别和容错能力。在一种可能的实现方式中，管理节点上显示有配置界面，该配置界面包括安全级别配置项和容错能力配置选项。用户在该安全级别配置项中输入所需的安全级别，并在容错能力配置选项中输入允许出错的节点数量 t，t 为大于或等于的整数。管理节点获取用户输入的安全级别和允许出错的节点数量 t。其中，安全级别包括服务器级安全、机柜级安全等。其中，服务器级安全用于指示存储系统最多能够容忍 t 个服务器出现故障。机柜级安全用于指示该存储系统最多能够容忍 t 个机柜出现故障。可选地，管理节点也可以根据该存储系统的拓扑结构，按照预设原则确定该存储系统的安全级别，其中，该预设原则是指能够保证该存储系统的可靠性的计算原则，本申请实施例对此不作限定。另外，该存储系统的容错能力也可以是一个系统默认值，本申请实施例对此不作限定。

在获取到存储系统的拓扑结构、容错能力和安全级别之后，管理节点通过下述公式（1）确定第一纠删码配比中的数据单元的数量的取值范围。

$$N \leqslant (k * M) - M \qquad (1)$$

其中，N 为第一纠删码配比对应的数据单元的数量。k 为存储系统中包含的节点

的数量，当安全级别为服务器级安全时，上述节点指服务器，当安全级别为机柜级安全时，上述节点为机柜。M 为容错能力所指示的该存储系统能够容忍出错的节点的数量，也即，第一纠删码配比中的校验单元的数量。需要说明的是，M 可以是默认值，也可以是由用户自定义的值，且 M 为大于或等于 1 的整数，例如，M=2。

在确定出第一纠删码配比中的数据单元的数量的取值范围之后，管理节点根据该取值范围和 M，确定得到多个第一候选纠删码配比，每个候选纠删码配比对应所述取值范围的一个值。之后，从多个第一候选纠删码配比中选择对应的写放大值最小的纠删码配比作为第一纠删码配比。

其中，写放大是指存储节点实际写入的数据量大于从计算节点接收到的数据量。在本申请实施例中，写放大通过写放大值来表征，对于任一个第一候选纠删码配比而言，该第一候选纠删码配比对应的写放大值等于该第一候选纠删码配比的数据单元和校验单元的总数量与数据单元的数量之间的比值。例如，对于纠删码配比 6:2 而言，该纠删码配比用于表征每 6 个数据单元对应 2 个校验单元，如此，该纠删码配比对应的写放大值即为（6+2）/6。

示例性地，假设根据该存储系统的拓扑结构指示该存储系统包括 4 个服务器，用户输入的安全级别为服务器级安全，则 k=4。假设该存储系统能够容忍出错的存储节点的数量为 2 个，也即，M=2。根据上述公式（1）可得，第一纠删码配比的数据单元的数量取值范围为 N≤4*2-2，也即，N≤6。在确定第一纠删码配比的数据单元的数量取值范围之后，根据该取值范围和校验单元的数量可得到常用的多个第一候选纠删码配比分别为 6:2、4:2 和 2:2。由于 6:2 这一配比是三个配比中写放大最小的，因此，将 6:2 这一配比作为第一纠删码配比。

除了第一纠删码配比之外，管理节点还用于根据该存储系统的拓扑结构和容错能力获取第二纠删码配比。具体的，管理节点通过下述公式（2）确定第二纠删码配比中的数据单元的数量取值范围。

$$X \leqslant (i * Y) - Y \qquad (2)$$

其中，X 为第二纠删码配比对应的数据单元的数量，且 X 大于 N。i 为该存储系统包含的节点的子节点的数量，其中，当安全级别为服务器级安全时，i 为该存储系统包含的服务器的子节点的数量，其中，服务器的子节点可以是指该服务器连接的物理硬盘或硬盘逻辑域。当安全级别为机柜级安全时，i 为该存储系统包含的机柜的子节点的数量，其中，机柜的子节点是指该机柜包含的服务器的数量。Y 为容错能力所指示的该存储系统能够容忍出错的节点的数量，也即，Y 为第二纠删码配比对应的校验单元的数量。需要说明的是，Y 可以是默认值，也可以是由用户自定义的值，且 Y 大于或等于 1，例如，Y=2。另外，Y 与 M 可以相等，也可以不相等，本申请实施例对此不作限定。还需要说明的是，安全级别可以由前述介绍的配置方式由用户进行配置，在这种情况下，管理节点直接获取用户配置的安全级别。或者，该安全级别也可以是管理节点根据该存储系统的拓扑结构按照预设原则确定得到的，其中，该预设原则是指能够保证该存储系统的可靠性的计算原则，本申请实施例对此不作限定。

在确定出第二纠删码配比中的数据单元的数量取值范围之后，管理节点根据该取值范围和 Y，确定第二纠删码配比。

例如，仍以前述的包含 4 个服务器的存储系统为例，假设每个服务器包括 4 个硬盘逻辑域，则在安全级别为服务器级安全时，该存储系统中包含的每个服务器下有 4 个子节点，这样，4 个服务器的子节点的总数量为 16。假设容错能力所指示的该存储系统能够容忍出错的节点的数量为 2，也即 Y=2，则根据上述公式（2）可得，X≤（16*2）- 2，也即，X≤30。根据该取值范围，考虑到系统可靠性约束机制，管理节点可选择数据单元的数量为 24，此时，第二纠删码配比即为 24:2。

由上文可知，第一纠删码配比中的 N 和第二纠删码配比中的 X 不相等，且 N 小于 X。另外，第一纠删码配比中的 M 和第二纠删码配比中的 Y 可以相等也可以不等。除此之外，N 和 M 的比值不等于 X 和 Y 的比值。

上述介绍了分别根据存储系统的拓扑结构和容错能力确定第一纠删码配比和第二纠删码配比的实现过程。在一些可能的实现方式中，在参考上述方式确定得到第一纠删码配比之后，管理节点根据第一纠删码配比 N:M 和预设的 w 确定第二纠删码配比 X:Y。其中，X 等于 w*N，Y 等于 M 或大于 M。或者，在参考上述方式确定得到第二纠删码配比之后，管理节点根据第二纠删码配比 X:Y 和预设的 w 确定第一纠删码配比 N:M。其中，N 等于 X/w，M 等于 Y 或小于 Y。

管理节点获得了第一纠删码配比和第二纠删码配比之后，计算第二纠删码配比中的数据单元的数量 X 和第一纠删码配比中的数据单元的数量 N 之间的比值，该比值就等于在第一介质层中按照第一纠删码配比存储的数据包括的第一校验矩阵的个数 w。例如，当第二纠删码配比中的数据单元的数量 X=24，第一纠删码配比中的数据单元的数量 N=6，则可以确定在第一介质层中按照第一纠删码配比存储的数据包括的第一校验矩阵的个数 w=4。由此可见，在前述根据第一纠删码配比获得第二纠删码配比或者是根据第二纠删码配比获得第一纠删码配比的实现方式中，预设的 w 实际上就是在第一介质层中按照第一纠删码配比存储的数据包括的第一校验矩阵的个数。其中，一个第一校验矩阵为一个遵从第一纠删码配比的分条。

本发明实施例的另一种实现方式，可以由管理员通过管理节点配置第一纠删码配比和第二纠删码配比。

在获得第一纠删码配比、第二纠删码配比以及 w 之后，后续，当存储节点接收到计算节点发送的写数据请求时，存储节点按照第一纠删码配比和 w 在第一介质层中写入数据。其中，写数据请求包括待写入的数据。接下来以存储系统中的目标存储节点接收计算节点发送的写数据请求为例对该过程进行说明。

示例性地，目标存储节点接收计算节点发送的写数据请求，当接收的待写入数据的数据量达到 N 个数据单元的尺寸时，目标存储节点将这些待写入数据划分成 N 个数据单元，并根据该 N 个数据单元生成 M 个校验单元。所述 N 个数据单元和 M 个校验单元属于一个子数据，该子数据对应一个第一校验矩阵，其中，该第一校验矩阵包括该 N 个数据单元和 M 个校验单元。之后，目标存储节点将第一校验矩阵包含的 N 个数据单元和 M 个校验单元存储至该存储系统的第一介质层中。与此同时，目标存储节点继续接收计算节点发送的写数据请求，按照上述方式获得另一个第一校验矩阵，并存储至所述第一介质层。如此，当按照上述方式该目标存储节点将该 w 个第一校验矩阵包括的数据单元和校验单元作为写入至第一介质层后，即可以执行后续步骤 402。

例如，第一纠删码配比为 6:2，也即，N=6，M=2，且 w=4，当目标存储节点接收到计算节点发送的待写入数据的数量达到 6 个数据单元的尺寸时，将这些待写入数据划分成 6 个数据单元，根据这 6 个数据单元生成 2 个校验单元，之后，生成包括 6 个数据单元和 2 个校验单元的第一校验矩阵。将第一校验矩阵包括的 8 个单元存储至该存储系统中的各个存储节点的内存中。

具体的，目标存储节点可以将各个第一校验矩阵包括的校验单元分布在同一个存储节点上，对于各个第一校验矩阵包括的数据单元，则可按照平均分布的原则分布在各个存储节点中。

参见图 5，假设该存储系统包括的存储节点为 4 个服务器，这 4 个服务器包括的内存为第一介质层。第一纠删码配比为 6:2，也即，N=6，M=2，并且，w=4。由于 M=2，所以每个服务器的内存中最多允许存储第一校验矩阵的 8 个单元中的 2 个单元，基于此，目标存储节点在获得一个第一校验矩阵之后，将第一校验矩阵包括的 2 个校验单元存储至自身内存中，而将剩余的 6 个数据单元转发至其他 3 个服务器中存储，例如，参见图 5，在其他 3 个服务器中的每个服务器上各存储 2 个数据单元。如此，在将 4 个第一校验矩阵存储至第一介质层后，4 个校验矩阵包含的 32 个单元在存储系统的存储节点的分布如图 6 所示。

图 5 和图 6 仅是本申请实施例示例性的给出的一种第一校验矩阵中各个单元的可能分布。可选地，目标存储节点也可以根据存储系统每个存储节点最多允许分布的单元数量和第一校验矩阵包括的单元的数量，随机将多个单元平均分布在各个存储节点中。也即，不限制校验单元在同一个存储节点上。例如，在上述的示例中，可以将一个第一校验矩阵中的 1 个数据单元和 1 个校验单元存储在目标存储节点上，另外一个校验单元和一个数据单元存储至另外一个存储节点中，这样，还剩余 4 个数据单元，这 4 个数据单元分别存储在剩余的两个存储节点上。进一步地，当第一校验矩阵中的两个校验单元为校验单元 p 和校验单元 q 时，4 个第一校验矩阵中的 4 个校验单元 p 存储在一个存储节点中，4 个校验单元 q 存储在一个存储节点中，且 4 个校验单元 p 所在的存储节点和 4 个校验单元 q 所在的存储节点可以不同。

上述介绍了存储系统为图 1 或图 2 所示的存储系统时，在第一介质层中按照第一纠删码配比缓存数据的过程。可选地，当存储系统为图 3 所示的存储阵列中，可以由存储阵列中的控制器来确定第一纠删码配比、第二纠删码配比和第一介质层中存储的数据包括的第一校验矩阵的个数 w，确定的方法参考前述介绍的方法，本申请实施例在此不再赘述。在确定第一纠删码配比、第二纠删码配比和 w 之后，控制器可以按照第一纠删码配比和 w 在第一介质层中缓存数据，其中，第一介质层即为存该存储系统包括的控制器的内存。需要说明的是，控制器参考前述介绍的方法生成 w 个第一校验矩阵。根据该存储系统包括的控制器的数量和/或每个控制器包括的内存的数量，参考前述的方法将每个第一校验矩阵包括的数据单元和校验单元分布存储至各个控制器的内存中。

步骤 402：将第一介质层中缓存的多个第一分条的数据单元和新的校验单元按照第二纠删码配比持久化存储到第二介质层。

其中，新的校验单元是由多个第一分条的校验单元生成的，多个第一分条的数据

单元和新的校验单元属于遵从第二纠删码配比的分条（第二分条）。

在按照第一纠删码配比在第一介质层中缓存的数据达到设定条件时，存储节点或控制器将达到设定条件的这部分数据，按照第二纠删码配比存储到至第二介质层。例如，设定条件可以是：第一介质层中缓存的遵从第一纠删码配比的多个第一分条中的数据单元的个数达到遵从第二纠删码配比的第二分条中数据单元的个数。即在第一介质层中凑满第二分条。其中，第二纠删码配比为 X:Y，也即，存储第二介质层中的数据包括 X 个数据单元和 Y 个校验单元。

接下来仍以该数据存储方法应用于图 1 或图 2 所示的存储系统中为例来对本步骤进行说明。在该存储系统中，第二介质层包括该存储系统内的存储节点所包括的硬盘。

示例性地，在凑齐 w 个第一校验矩阵，也即在将 w 个第一校验矩阵包括的数据单元和校验单元缓存至第一介质层后，根据第一介质层中存储的数据包括的 w 个第一校验矩阵中每个第一校验矩阵所包含的 N 个数据单元获得 X 个数据单元，X 是 N 的整数倍；计算获得 Y 个校验单元以生成第二校验矩阵，第二校验矩阵包括 X 个数据单元和 Y 个校验单元；将第二校验矩阵写入第二介质层中。其中，上述过程可以通过以下几种不同的实现方式来实现。一个第二校验矩阵为一个遵从第二纠删码配比的分条。

第一种实现方式：目标存储节点在通过上述步骤 401 凑齐 w 个第一校验矩阵之后，根据 w 个第一校验矩阵包括的 w × N 个数据单元计算得到第二校验矩阵中的 Y 个校验单元。之后，目标存储节点将计算得到的 Y 个校验单元存储至第二介质层。对于其他存储节点而言，各个存储节点在满足设定条件时，将自身存储的属于 w 个第一校验矩阵的数据单元存储至第二介质层。如此，存储至第二介质层中的 w × N 个数据单元即为第二校验矩阵的 X 个数据单元，目标存储节点计算得到的 Y 个校验单元即为第二校验矩阵包括的 Y 个校验单元。

其中，目标存储节点在将自身存储的数据单元和 Y 个校验单元存储至第二介质层中时，如果第二介质层包括的硬盘的数量大于第二校验矩阵所包含的单元的总数量，则目标存储节点根据计算得到的校验单元的数量 Y，从自身所包括的多个硬盘中选择 Y 个硬盘。然后目标存储节点将 Y 个校验单元写入至选择的硬盘中，其中，每个硬盘上写入一个单元。可选地，如果目标存储节点上还存储有属于 w 个第一校验矩阵的数据单元，则目标存储节点从自身所包括的硬盘中为每个数据单元选择一个硬盘，并将数据单元写入至选择的硬盘上，其中，每个硬盘上也同样写入一个单元。

可选地，如果第二介质层包括的硬盘的数量不大于第二校验矩阵所包含的单元的总数量，则存储节点根据第二校验矩阵包括的校验单元的数量，确定第二介质层中每个硬盘上允许分布的最大单元数。之后，如果目标存储节点上还存储有属于 w 个第一校验矩阵的数据单元，则按照该最大单元数和自身存储的属于 w 个第一校验矩阵的数据单元的数量和 Y 从自身所包括的硬盘中选择多个硬盘，进而将存储的数据单元和校验单元写入至选择的多个硬盘中。当然，如果目标存储节点中未存储属于 w 个第一校验矩阵的数据单元，则按照最大单元数和 Y 从自身所包括的硬盘中选择多个硬盘，从而将 Y 个校验单元写入至选择的硬盘上。在这种情况下，一个硬盘上分布可能存储第二校验矩阵中的多个单元，但是，存储的单元的数量不超过硬盘允许分布的最大单元数。对于除目标存储节点之外的其他存储节点，均可参考上述方法将自身存储的属于

w 个第一校验矩阵的数据单元写入至第二介质层中。

例如，参见图 7，假设第二介质层包括 16 个硬盘逻辑域，且 16 个硬盘逻辑域分属于 4 个存储节点，w 个第一校验矩阵一共包括 24 个数据单元，也即，第二校验矩阵包括的数据单元的数量为 24。目标存储节点计算得到的校验单元的数量为 2。由于第二校验矩阵包括的校验单元的数量为 2，因此可知每个硬盘逻辑域上允许分布的最大单元数为 2，如此，各个存储节点在按照上述介绍的方式将 24 个数据单元和 2 个校验单元存储至第二介质层时，对于 4 个存储节点中的 3 个存储节点，这 3 个存储节点中每个存储节点的每个硬盘逻辑域内存储 2 个单元，也即，每个存储节点上一共存储 8 个单元，而对于另外一个存储节点，则可以在该存储节点的一个硬盘逻辑域上存储 2 个单元，或者是在该存储节点的两个硬盘逻辑域上各存储一个单元。

第二种实现方式：目标存储节点在凑齐 w 个第一校验矩阵之后，根据 w 个第一校验矩阵包括的 w × M 个校验单元获得第二校验矩阵中的 Y 个校验单元。之后，目标存储节点将计算得到的 Y 个校验单元存储至第二介质层。对于其他存储节点而言，各个存储节点在自身缓存中的数据量达到一定阈值时，将自身存储的属于 w 个第一校验矩阵的数据单元存储至第二介质层。

对于第二种实现方式，分别针对以下几种不同的情况进行说明。

（1）当 w 个第一校验矩阵包括的所有校验单元均存储在目标存储节点中时，目标存储节点获取自身存储的 w × M 个校验单元，进而根据该 w × M 个校验单元获得 Y 个校验单元。

举例来说，当每个第一校验矩阵包括的 M 个校验单元分别为校验单元 p 和校验单元 q 时，目标存储节点对存储的 w 个校验单元 p 进行异或运算或者其他计算方式得到第二校验矩阵中的校验单元 p’，对存储的 w 个校验单元 q 进行异或运算或者其他计算方式得到第二校验矩阵中的校验单元 q’。由此可见，本申请实施例中，通过直接对各个第一校验矩阵包括的 M 个校验单元进行计算即能够得到第二校验矩阵的校验单元，相较于根据 w 个第一校验矩阵中的所有数据单元重新计算校验单元，减少了计算量。并且，由于各个第一校验矩阵中的校验单元均存储在同一个存储节点上，因此，该存储节点能够直接获取存储的校验单元来获得第二校验矩阵中的校验单元，相较于将校验单元分布存储在各个存储节点中的情况，无需跨存储节点中获取校验单元，减少了网络转发量。

在计算获得 Y 个校验单元之后，目标存储节点参考前述第一种实施例中介绍的方法将 Y 个校验单元存储至第二介质层中，或者，将自身存储的属于 w 个校验矩阵的数据单元和 Y 个校验单元存储至第二介质层中。其他存储节点在满足设定条件之后，将各自存储的数据单元在存储至第二介质层中，如此，各个存储节点中存储的 w 个第一校验矩阵则合并为第二校验矩阵存储至了第二介质层中。

图 8 是本申请实施例示出的一种根据 w 个第一校验矩阵合并得到第二校验矩阵的原理示意图。如图 8 所示，w=4，第一个第一校验矩阵中的前六列元素 a1 至 a6 为 6 个数据单元，剩下两列元素 p1 和 q1 为 2 个校验单元。其中，p1 为校验单元 p，q1 为校验单元 q。同理，第二个第一校验矩阵中的 a7 至 a12 为 6 个数据单元，并且，剩下两列元素 p2 和 q2 为 2 个校验单元，以此类推。将每个校验矩阵的前 6 列元素取出，

组成第二校验矩阵的 24 个数据单元。将每个第一校验矩阵中的校验单元 p（也即 p1 至 p4）取出，进行异或运算或者采用其他计算方式得到第二校验矩阵中的校验单元 p'，将每个第一校验矩阵中的校验单元 q（也即 q1 至 q4）取出，进行异或运算，或者采用其他计算方式得到第二校验矩阵中的校验单元 q'。

需要说明的是，图 8 示例性地说明了将 w 个第一校验矩阵合并为第二校验矩阵的过程，在一些可能的应用场景中，例如，当第一介质层的性能低于第二介质层的性能，将第一介质层中的数据读取到第二介质层的过程中，可能需要将第二校验矩阵拆分为 w 个第一校验矩阵，在这种情况下，只需将上述过程逆向执行即能够得到 w 个第一校验矩阵，本申请实施例对此不做限定。

可选地，当每个第一校验矩阵包括的 M 个校验单元为校验单元 r 时，目标存储节点对存储的 w 个校验单元 r 进行异或运算，增量计算得到第二校验矩阵中的校验单元 r'，之后，目标存储节点可以获取自身以及其他存储节点上存储的各个第一校验矩阵中的数据单元，根据获取的 w × N 个数据单元计算得到校验单元 p' 和校验单元 q'。将计算得到的校验单元 r'、校验单元 p' 和校验单元 q' 作为第二校验矩阵中的 Y 个校验单元。由此可见，在该种实现方式中，Y 和 M 不相等。并且，由于根据数据单元计算校验单元 r' 的过程较为复杂，因此，本申请实施例中根据各个第一校验矩阵中的校验单元 r 增量计算得到第二校验矩阵中的校验单元 r'，减小了计算开销。另外，根据 w × N 个数据单元计算得到校验单元 p' 和校验单元 q'，使得第二校验矩阵中包含了 3 个校验单元，提升了第二介质层中存储的数据的冗余度，使得容错能力得以提升。

例如，参见图 9，假设有 3 个第一校验矩阵，每个第一校验矩阵包括 7 个数据单元和 1 个校验单元 r，目标存储节点将存储的 3 个校验单元 r 进行异或运算得到第二校验矩阵中的校验单元 r'。之后，目标存储节点获取自身及其他各个存储节点中存储的 21 个数据单元，根据这 21 个数据单元计算得到校验单元 p' 和校验单元 q'。如此，第二校验矩阵包括的 Y 个校验单元即为生成的校验单元 p'、校验单元 q' 和校验单元 r'。由此可见，通过该种实现方式，能够提升第二介质层中存储的数据的冗余度，使得容错能力得以提升，同时还能够减少部分计算开销。

在计算获得 Y 个校验单元之后，目标存储节点同样参考前述第一种实现方式中介绍的方法将存储的数据单元和 Y 个校验单元存储至第二介质层中，其他存储节点在自身缓存中的数据量达到一定阈值之后，将各自存储的数据单元在存储至第二介质层中，本申请实施例在此不再赘述。

（2）当各个第一校验矩阵包括的 M 个校验单元分散存储在不同的存储节点中时，目标存储节点从各个存储节点中获取存储的校验单元，进而根据获取到的 w × M 个校验单元获得 Y 个校验单元。其中，目标存储节点根据获取到的 w × M 个校验单元获得 Y 个校验单元的实现方式参考上述（1）中的实现方式，本申请实施例不再赘述。在获得 Y 个校验单元之后，目标存储节点参考前述第一种实现方式中介绍的方法将存储的数据单元和 Y 个校验单元存储至第二介质层中，其他存储节点在自身缓存中的数据量达到一定阈值之后，将各自存储的数据单元在存储至第二介质层中，本申请实施例在此不再赘述。

第三种实现方式：目标存储节点在凑齐 w 个第一校验矩阵之后，将自身存储的属

于 w 个第一校验矩阵的单元写入至第二介质层，其他各个存储节点在自身缓存存储的数量达到一定阈值之后，将自身存储的属于 w 个第一校验矩阵的单元也写入至第二介质层中。之后，目标存储节点获取写入至第二介质层中的 w × M 个校验单元，进而根据 w × M 个校验单元计算获得 Y 个校验单元，将计算得到的 Y 个校验单元作为第二校验矩阵的 Y 个校验单元写入至第二介质层中。

其中，如果第二介质层包括的硬盘的数量大于第二校验矩阵所包含的单元的总数量，则各个存储节点在将自身存储的属于 w 个第一校验矩阵的数据单元和校验单元写入至第二介质层时，对于数据单元，可以为每个数据单元选择一个硬盘，并将每个数据单元写入至为相应数据单元选择的硬盘中，其中，为不同的数据单元选择的硬盘也不同。这样，对于第二校验矩阵包括的 X 个数据单元，这 X 个数据单元将被写入至 X 个硬盘中。对于 w 个第一校验矩阵中的校验单元，各个存储节点可以将自身存储的校验单元存储至除上述的 X 个硬盘中的剩余硬盘中。

需要说明的是，在写入校验单元时，可以在每个硬盘上写入一个校验单元，这样，w × M 个校验单元将被写入至 w × M 个硬盘中。或者，所有的校验单元可以写入至一个硬盘中。或者，可以在 M 个硬盘上写入 w × M 个校验单元，其中，M 个硬盘中每个硬盘上写入的校验单元为第一校验矩阵中位于同一列的校验单元。例如，当 M=2 时，两个校验单元中的一个校验单元为校验单元 p，另一个校验单元为校验单元 q，则各个第一校验矩阵中的校验单元 p 写入至一个硬盘上，校验单元 q 写入至另一个硬盘上。

各个存储节点在将自身存储的属于 w 个第一校验矩阵的数据单元和校验单元写入至第二介质层中之后，目标存储节点从第二介质层中获取 w × M 个校验单元。其中，如果 w × M 个校验单元将被写入至 w × M 个硬盘中，则目标存储节点从这 w × M 个硬盘中读取 w × M 个校验单元。如果所有的校验单元被写入至一个硬盘中，则目标存储节点从该硬盘中一次性获取 w × M 个校验单元，这样，能够减少网络通信次数，节省带宽资源。如果 w × M 个校验单元被写入至 M 个硬盘中，且 M 个硬盘中每个硬盘上写入的校验单元为第一校验矩阵中位于同一列的校验单元，则目标存储节点从各个硬盘上读取位于相同列的校验单元，从而得到 w × M 个校验单元。如此，在一定程度上也可以减少网络通信次数，节省带宽资源。

在获取到 w × M 个校验单元之后，目标存储节点参考前述第一种实现方式中介绍的方法，根据该 w × M 个校验单元计算获得 Y 个校验单元，进而将这 Y 个计算单元分别写入至 Y 个硬盘，其中，每个硬盘上写入一个校验单元。并且，写入 Y 个校验单元的 Y 个硬盘不为前述写入数据单元的 X 个硬盘中的硬盘。

可选地，如果第二介质层包括的硬盘的数量不大于第二校验矩阵所包含的单元的总数量，则各个存储节点可参考前述第一种实现方式中介绍的方法，在一个硬盘上写入两个或两个以上的单元，只要不超出允许存储的最大单元数即可。同样的，在这种情况下，各个第一校验矩阵包括的 M 个校验单元中可以存储在同一个存储节点下的硬盘上，或者，各个第一校验矩阵包括的 M 个校验单元中位于同一列的校验单元可以存储在一个存储节点下的硬盘中，例如，存储在一个存储节点下的同一个硬盘逻辑域上，或者是存储在一个存储节点下的一个物理硬盘上，以此来减少计算第二校验矩阵中 Y 个校验单元时所需的网络转发次数。

举例说明，图 10 是本申请实施例提供的一种在第二介质层中按照第二纠删码配比写入数据的示意图。如图 10 所示，第一校验矩阵的个数 w=4，每个第一校验矩阵包括 6 个数据单元和 2 个校验单元。2 个校验单元分别为校验单元 p 和校验单元 q。第二介质层包括分布在 4 个存储节点（a1 至 a4）上的 16 个硬盘逻辑域，每个存储节点上分布有 4 个硬盘逻辑域，每个硬盘逻辑域内包括 16 个物理硬盘。由于校验单元的数量为 2，因此，每个硬盘逻辑域上最多允许分布第二校验矩阵中的 2 个单元。基于此，首先从 16 个硬盘逻辑域中选择两个硬盘逻辑域，例如，选择的两个硬盘逻辑域分别为存储节点 a2 上的硬盘逻辑域 a21 和存储节点 a4 上的硬盘逻辑域 a41。之后，存储有各个第一校验矩阵的校验单元 p 的各个存储节点将自身存储的校验单元 p 写入至硬盘逻辑域 a21 的物理硬盘中，例如，每个第一校验矩阵中的校验单元 p 均写入至硬盘逻辑域 a21 的第一个物理硬盘中。同理，存储有各个第一校验矩阵的校验单元 q 的各个存储节点将自身存储的校验单元 q 写入至硬盘逻辑域 a41 的第一个物理硬盘中。之后，对于硬盘逻辑域 a21，目标存储节点或存储节点 a2 对写入至硬盘逻辑域 a21 上的 4 个第一校验矩阵中的 4 个校验单元 p 进行异或运算，得到第二校验矩阵中的校验单元 p’，将该校验单元 p’存储在硬盘逻辑域 a21 上。由于每个硬盘逻辑域上最多允许存储第二校验矩阵中的两个单元，所以，在将计算得到的第二校验矩阵中的校验单元 p’写入至硬盘逻辑域 a21 上之后，该硬盘逻辑域 a21 上最多还能再存储第二校验矩阵中的一个数据单元。同理，对于硬盘逻辑域 a41，同样可以由目标存储节点或存储节点 a4 对其中存储的 4 个校验单元 q 进行异或运算，增量计算得到第二校验矩阵中的校验单元 q’，并将其存储在硬盘逻辑域 a41 中，如此，硬盘逻辑域 a41 上最多也只能存储第二校验矩阵中的一个数据单元。之后，对于每个第一校验矩阵包括的 6 个数据单元，各个存储节点根据每个硬盘逻辑域上最多分布第二校验矩阵中的两个单元的原则，将 24 个数据单元分布在 4 个存储节点包括的 16 个硬盘逻辑域上。

上述介绍了存储系统为图 1 或图 2 所示的存储系统时，在第一介质层中按照第一纠删码配比缓存数据的过程。可选地，当存储系统为图 3 所示的存储阵列中，则上述存储节点执行的操作可以由控制器来执行，从而将第一介质层中数据包括的 w 个第一校验矩阵中的数据单元和校验单元合并为第二校验矩阵写入至第二介质层中，本申请实施例对此不再赘述。

在按照上述的数据存储方法存储数据后，当第一介质层包括的节点或第二介质层包括的硬盘发生故障时，如果数据已经存入至了第二介质层，也即，第二校验矩阵已经生成，则根据故障点的个数、故障位置以及第二校验矩阵中各个单元的分布位置，从第二介质层中读取除故障点之外的其他位置上的数据单元和校验单元进行重构，从而恢复出故障点中的数据。可选地，如果数据已存入至第一介质层，但是还未存入至第二介质层，则根据故障点的个数、故障位置以及各个第一校验矩阵中各个单元的分布位置，从第一介质层中读取未发生故障的位置上的数据单元和校验单元进行重构，从而恢复出第一介质层中故障点中的数据。

在本申请实施例中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数，也即，第一纠删码配比为小比例的配比，而第二纠删码配比为大比例的配比。在此基础上，在第一介质层中按照第一纠删码配比缓存数据，在第二介

质层中按照第二纠删码配比存储数据，而第一介质层的性能高于第二介质层，也就是说，在高性能介质层采用较小的纠删码配比来存储数据，而在低性能介质层采用较大的纠删码配比来存储数据。由于高性能介质层接收到的IO粒度比较小，所以，在高性能介质层通过较小的纠删码配比来存储数据时，每当接收到数据的尺寸达到该纠删码配比对应的N个数据单元的尺寸时，即可以凑满一个分条（N个数据单元和M个校验单元即可组成一个分条），相较于较大的纠删码配比而言，小的纠删码配比更容易凑满分条，从而使得分条中补0的数据量减少，降低了写放大，提高了存储空间利用率。例如，在高性能介质层中采用6:2的纠删码配比来存储数据，相较于采用24:2来存储数据，在指定的时间段内，根据接收到的小粒度的IO请求，凑齐6个数据单元比凑齐24个数据单元更为容易，这样，就不必在凑不齐24个数据单元时进行补0，也即，使得分条中补0的数据量减少，降低了分条中冗余数据量的占比，降低了写放大，提高了存储空间利用率。另外，在低性能介质层采用较大的纠删码配比来存储数据，能够减少冗余数据在存储空间中的占比，从而提高存储空间利用率。

另外，在本申请实施例中，在第一介质层中按照第一纠删码配比缓存的一份数据能够直接转换为符合第二纠删码配比的一份数据，进而存储至第二介质层，在提高存储系统的存储空间利用率的同时提高数据存储可靠性。同时上述转换过程中，多个第一分条中的数据单元不再需要参与运算，从而节省了存储系统的计算资源。

上述实施例主要介绍了第一介质层中缓存的遵从第一纠删码配比的多个第一分条中的数据单元的个数达到遵从第二纠删码配比的第二分条中数据单元的个数，即缓存在第一介质层的多个第一分条可以凑满第二分条时，将多个第一分条的数据单元以及第二分条的校验单元存储到第二介质层。

本发明另一实施例，为了提高数据存储的可靠性，可以将遵从第一纠删码配比的第一分条在存储系统中进行持久化存储，即存储到第二介质层，这样可以提高数据可靠性，防止存储系统故障导到第一介质层中没有持久化存储的第一分条数据丢失。当第一介质层中非持久化存储的第一分条与已经持久化存储的第一分条能够凑满第二分条，即第一介质层中非持久化存储的第一分条中的数据单元的个数与已经持久化存储的第一分条中的数据单元的个数等于第二分条中数据单元的个数时，一方面，读取已经持久化存储的第一分条中的校验单元，另一方面，读取第一介质层中非持久化存储的第一分条中的校验单元，生成第二分条的校验单元，具体实现方式可以参考前述实施例的描述，本发明实施例不再赘述。将非持久化存储的第一分条中的数据单元以及第二分条的校验单元持久化存储到第二介质层，实现第二分条在存储系统中持久化存储。

本发明上述实施例另外一种实现方式，可以由存储系统中的接口卡实现，例如由主机总线适配器（Host Bus Adapter, HBA）、网络接口卡（Network Interface Card, NIC）或扩展器(Expander)等实现，本发明对此不再赘述。

接下来对本申请实施例提供的数据存储装置进行介绍。

参见图11，本申请实施例提供了一种分条管理装置，该分条管理装置应用于前述介绍图1或图2所示的存储系统中的任意一个存储节点，也可以应用于图3所示存储阵列等。该分条管理装置包括:

获取单元 1101，用于获取多个第一分条中的校验单元；其中，所述第一分条遵从第一纠删码配比；

生成单元 1102，用于根据所述多个第一分条的校验单元生成新的校验单元；其中，所述新的校验单元与所述多个第一分条中的数据单元属于新的分条；所述新的分条遵从第二纠删码配比；其中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数。本发明实施例上述各单元的实现可以参考本发明前述实施例中的描述。

可选地，新的分条中的校验单元的个数与所述第一分条中的校验单元的个数相同。

可选地，多个第一分条包含至少一个在所述存储系统中未持久化存储的第一分条和至少一个已经在所述存储系统持久化存储的第一分条；获取单元 1101，具体用于：

读取所述至少一个已经在所述存储系统持久化存储的第一分条中的校验单元；

读取所述至少一个在所述存储系统中未持久化存储的第一分条中的校验单元。

进一步地，所述分条管理装置还包括存储单元 1103；所述存储单元 1103，用于持久化存储所述至少一个在所述存储系统中未持久化存储的第一分条中的数据单元和所述新的校验单元。

可选地，所述多个第一分条为所述存储系统中未持久化存储的第一分条；所述分条管理装置还包括存储单元 1103；所述存储单元 1103，用于持久化存储所述多个第一分条中的数据单元以及所述新的校验单元。

综上所述，在本申请实施例中，第一介质层和第二介质层的性能不同，基于此，在第一介质层和第二介质层中按照不同的纠删码配比来进行数据存储。由于不同的纠删码配比对应的写放大不同，所导致的存储空间利用率也不同，因此，根据介质层的性能的不同选取不同的纠删码配比进行数据存储能够更好的发挥相应介质层的存储性能，有效的提高存储空间利用率。

需要说明的是：上述实施例提供的分条管理装置在进行数据存储时，仅以上述各功能单元的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的分条管理装置与分条管理方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意结合来实现。当使用软件实现时，可以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程序指令。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。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机程序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，所述计算机程序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、数据用户线（Digital Subscriber Line，DSL））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：数字通用光盘（Digital Versatile Disc，DVD））、或者半导体介质（例如：固态硬盘（Solid State Disk，SSD））等。

本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

应当理解的是，本文提及的“至少一个”是指一个或多个，“多个”是指两个或两个以上。在本文的描述中，除非另有说明，“/”表示或的意思，例如，A/B 可以表示 A 或 B；本文中的“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，为了便于清楚描述本申请实施例的技术方案，在本申请的实施例中，采用了“第一”、“第二”等字样对功能和作用基本相同的相同项或相似项进行区分。本领域技术人员可以理解“第一”、“第二”等字样并不对数量和执行次序进行限定，并且“第一”、“第二”等字样也并不限定一定不同。

权 利 要 求 书

1、一种分条管理方法，其特征在于，应用于存储系统，所述方法包括：

获取多个第一分条中的校验单元；其中，所述第一分条遵从第一纠删码配比；

根据所述多个第一分条的校验单元生成新的校验单元；其中，所述新的校验单元与所述多个第一分条中的数据单元属于新的分条；所述新的分条遵从第二纠删码配比；其中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数。

2、根据权利要求1所述的方法，其特征在于，所述新的分条中的校验单元的个数与所述第一分条中的校验单元的个数相同。

3、根据权利要求1或2所述的方法，其特征在于，所述多个第一分条包含至少一个在所述存储系统中未持久化存储的第一分条和至少一个已经在所述存储系统持久化存储的第一分条；所述获取多个第一分条中的校验单元，具体包括：

读取所述至少一个已经在所述存储系统持久化存储的第一分条中的校验单元；

读取所述至少一个在所述存储系统中未持久化存储的第一分条中的校验单元。

4、根据权利要求3所述的方法，其特征在于，所述方法还包括：

持久化存储所述至少一个在所述存储系统中未持久化存储的第一分条中的数据单元和所述新的校验单元。

5、根据权利要求1或2所述的方法，所述多个第一分条为所述存储系统中未持久化存储的第一分条；所述方法还包括：

持久化存储所述多个第一分条中的数据单元以及所述新的校验单元。

6、一种存储系统，其特征在于，所述存储系统包含一个或多个处理器，所述一个或多个处理器用于：

获取多个第一分条中的校验单元，其中，所述第一分条遵从第一纠删码配比；

根据所述多个第一分条的校验单元生成新的校验单元；其中，所述新的校验单元与所述多个第一分条中的数据单元属于新的分条；所述新的分条遵从第二纠删码配比；其中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数。

7、根据权利要求6所述的存储系统，其特征在于，所述新的分条中的校验单元的个数与所述第一分条中的校验单元的个数相同。

8、根据权利要求6或7所述的存储系统，其特征在于，所述多个第一分条包含至少一个在所述存储系统中未持久化存储的第一分条和至少一个已经在所述存储系统持久化存储的第一分条；所述一个或多个处理器具体用于：

读取所述至少一个已经在所述存储系统持久化存储的第一分条中的校验单元；

读取所述至少一个在所述存储系统中未持久化存储的第一分条中的校验单元。

9、根据权利要求8所述的存储系统，其特征在于，所述一个或多个处理器还用于：

持久化存储所述至少一个在所述存储系统中未持久化存储的第一分条中的数据单元和所述新的校验单元。

10、根据权利要求6或7所述的存储系统，其特征在于，所述多个第一分条为所述存储系统中未持久化存储的第一分条；所述一个或多个处理器还用于：

持久化存储所述多个第一分条中的数据单元以及所述新的校验单元。

11、一种分条管理装置，其特征在于，所述分条管理装置应用于存储系统中，所述数据存储装置包含获取单元和生成单元；其中，

所述获取单元，用于获取多个第一分条中的校验单元，其中，所述第一分条遵从第一纠删码配比；

所述生成单元，用于根据所述多个第一分条的校验单元生成新的校验单元；其中，

所述新的校验单元与所述多个第一分条中的数据单元属于新的分条；所述新的分条遵从第二纠删码配比；其中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数。

12、根据权利要求 11 所述的分条管理装置，其特征在于，所述新的分条中的校验单元的个数与所述第一分条中的校验单元的个数相同。

13、根据权利要求 11 或 12 所述的分条管理装置，其特征在于，所述多个第一分条包含至少一个在所述存储系统中未持久化存储的第一分条和至少一个已经在所述存储系统持久化存储的第一分条；所述获取单元具体用于：

读取所述至少一个已经在所述存储系统持久化存储的第一分条中的校验单元；

读取所述至少一个在所述存储系统中未持久化存储的第一分条中的校验单元。

14、根据权利要求 13 所述的分条管理装置，其特征在于，所述分条管理装置还包括存储单元；所述存储单元，用于持久化存储所述至少一个在所述存储系统中未持久化存储的第一分条中的数据单元和所述新的校验单元。

15、根据权利要求 11 或 12 所述的分条管理装置，其特征在于，所述多个第一分条为所述存储系统中未持久化存储的第一分条；所述分条管理装置还包括存储单元；所述存储单元，用于持久化存储所述多个第一分条中的数据单元以及所述新的校验单元。

16、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质包含计算机程序指令，存储系统中的一个或多个中央处理器执行所述计算机程序指令使得所述存储系统执行：

获取多个第一分条中的校验单元，其中，所述第一分条遵从第一纠删码配比；

根据所述多个第一分条的校验单元生成新的校验单元；其中，所述新的校验单元与所述多个第一分条中的数据单元属于新的分条；所述新的分条遵从第二纠删码配比；其中，第一纠删码配比对应的数据单元的个数小于第二纠删码对应的数据单元的个数。

17、根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述新的分条中的校验单元的个数与所述第一分条中的校验单元的个数相同。

18、根据权利要求 16 或 17 所述的计算机可读存储介质，其特征在于，所述多个第一分条包含至少一个在所述存储系统中未持久化存储的第一分条和至少一个已经在所述存储系统持久化存储的第一分条；所述获取多个第一分条中的校验单元，具体包括：

读取所述至少一个已经在所述存储系统持久化存储的第一分条中的校验单元；

读取所述至少一个在所述存储系统中未持久化存储的第一分条中的校验单元。

19、根据权利要求 18 所述的计算机可读存储介质，其特征在于，所述一个或多个中央处理器执行所述计算机程序指令使得所述存储系统还执行：

持久化存储所述至少一个在所述存储系统中未持久化存储的第一分条中的数据单元和所述新的校验单元。

20、根据权利要求 16 或 17 所述的计算机可读存储介质，所述多个第一分条为所述存储系统中未持久化存储的第一分条；所述一个或多个中央处理器执行所述计算机程序指令使得所述存储系统还执行：

持久化存储所述多个第一分条中的数据单元以及所述新的校验单元。

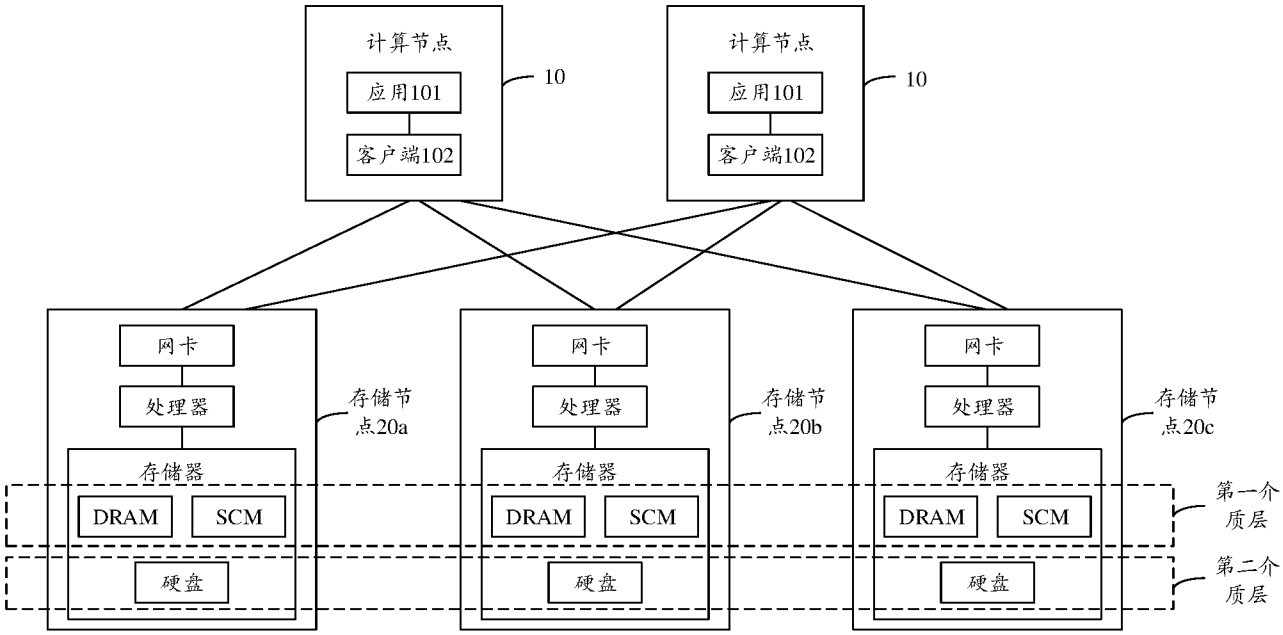
计算节点
应用101
客户端102
10
计算节点
应用101
客户端102
10
网卡
处理器
存储器
DRAM
SCM
硬盘
存储节点20a
网卡
处理器
存储器
DRAM
SCM
硬盘
存储节点20b
网卡
处理器
存储器
DRAM
SCM
硬盘
存储节点20c
第一介质层
第二介质层

图 1

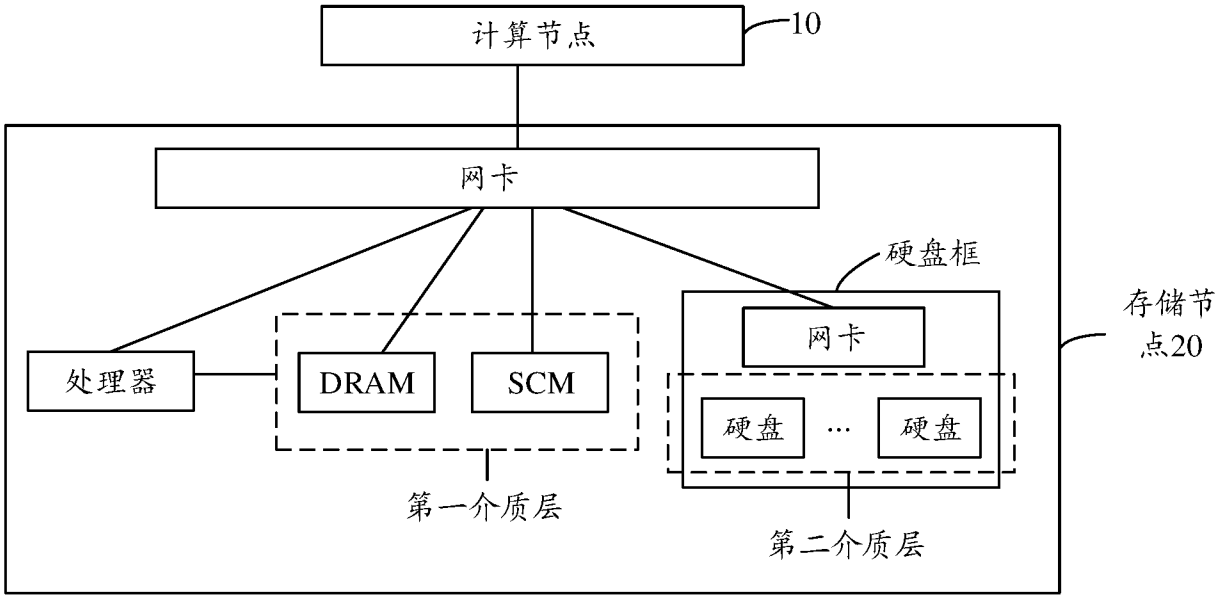
计算节点
10
网卡
硬盘框
处理器
DRAM
SCM
网卡
硬盘
…
硬盘
存储节点20
第一介质层
第二介质层

图 2

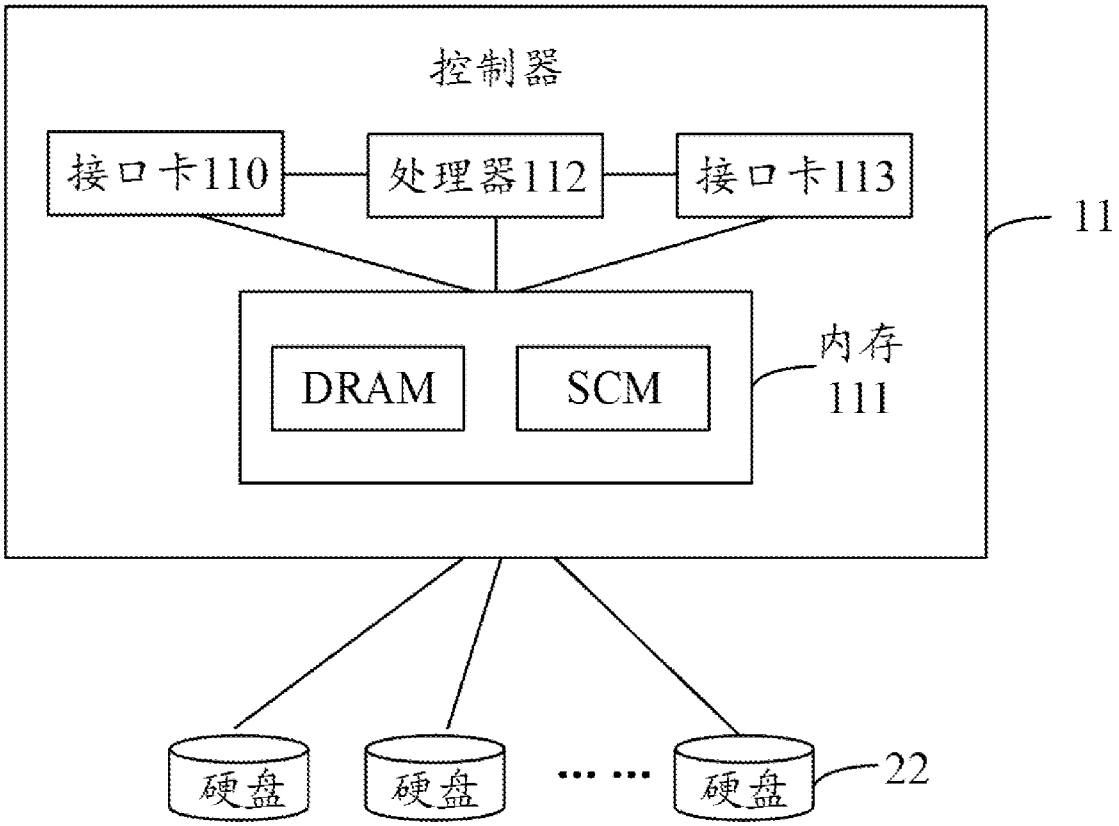
控制器
接口卡110
处理器112
接口卡113
11
DRAM
SCM
内存
111
硬盘
硬盘
硬盘
22

图 3

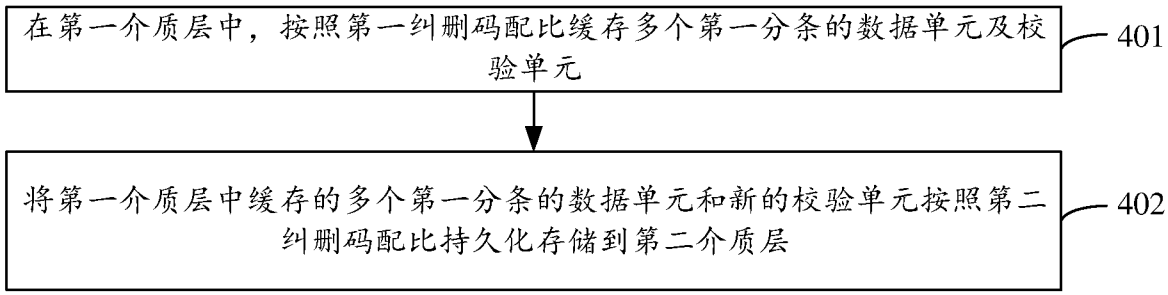
在第一介质层中，按照第一纠删码配比缓存多个第一分条的数据单元及校验单元
401
将第一介质层中缓存的多个第一分条的数据单元和新的校验单元按照第二纠删码配比持久化存储到第二介质层
402

图 4

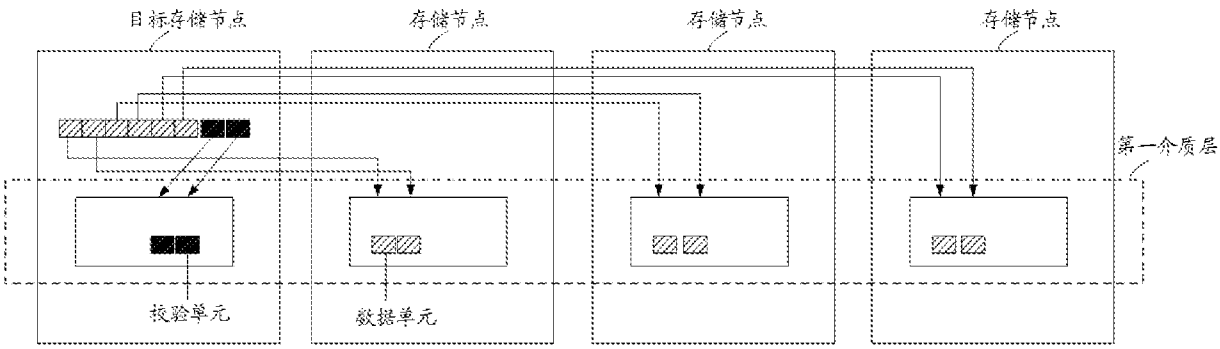
目标存储节点
存储节点
存储节点
存储节点
第一介质层
校验单元
数据单元

图 5

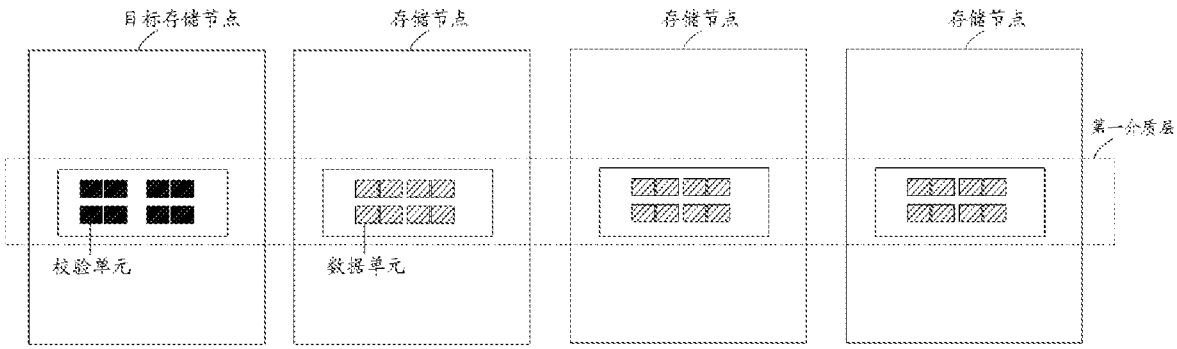
目标存储节点
存储节点
存储节点
存储节点
第一介质层
校验单元
数据单元

图 6

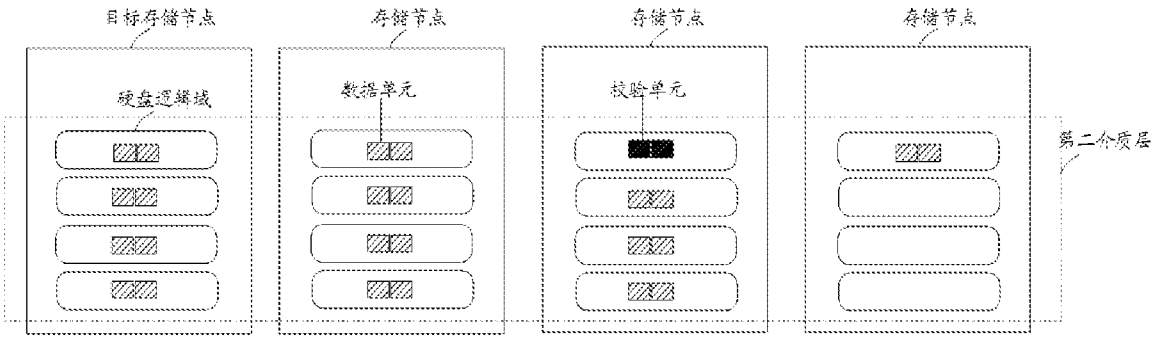
目标存储节点
存储节点
存储节点
存储节点
硬盘逻辑域
数据单元
校验单元
第二介质层

图 7

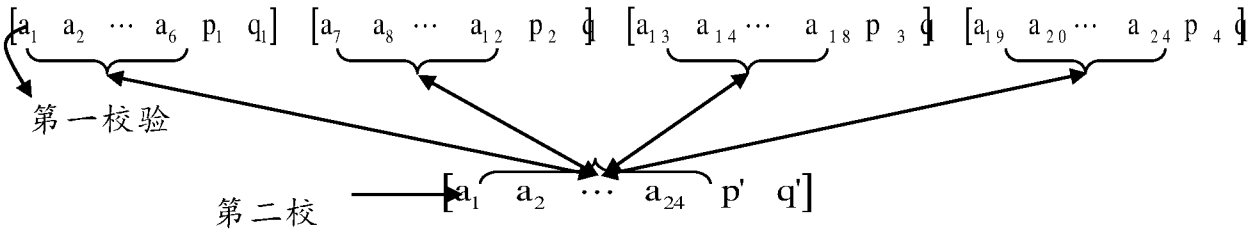
[a1 a2 … a6 p1 q1] [a7 a8 … a12 p2 q2] [a13 a14 … a18 p3 q3] [a19 a20 … a24 p4 q4]
第一校验
第二校
[a1 a2 … a24 p' q']

图 8

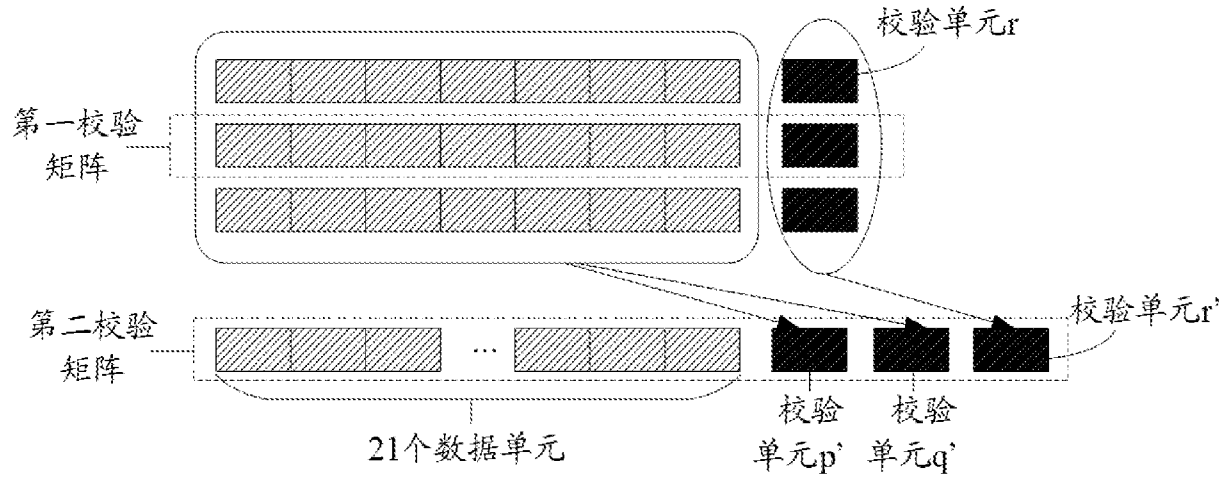
校验单元r
第一校验
矩阵
第二校验
矩阵
校验单元r'
21个数据单元
校验
单元p'
校验
单元q'

图 9

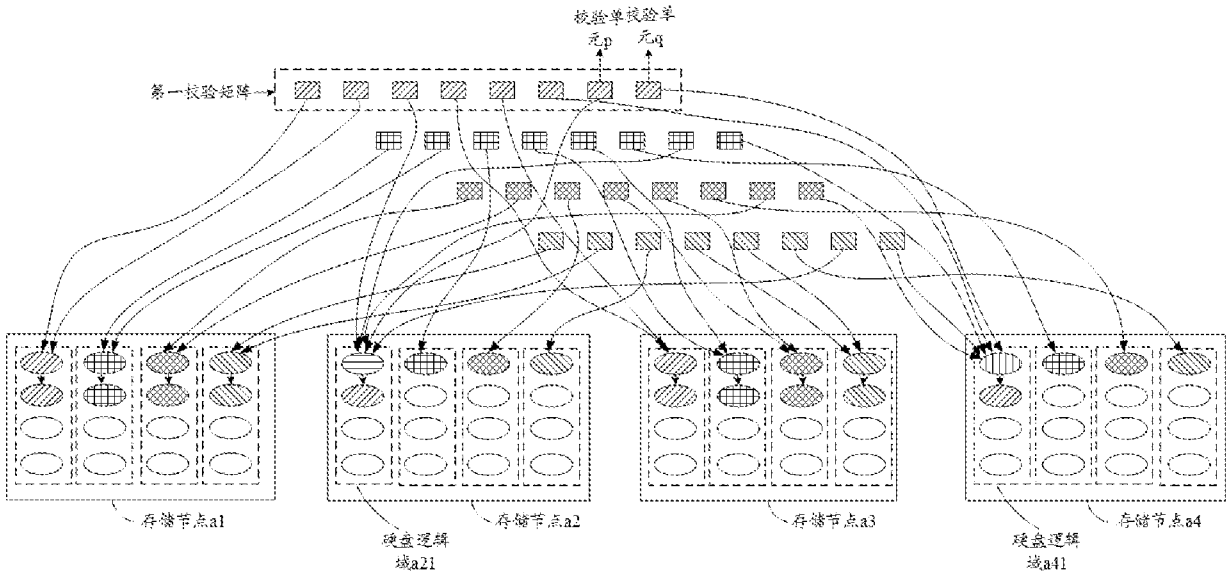
校验单校验单元p
元q
第一校验矩阵
存储节点a1
硬盘逻辑域a21
存储节点a2
存储节点a3
硬盘逻辑域a41
存储节点a4

图 10

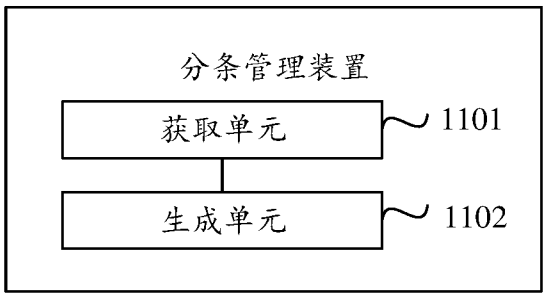
分条管理装置
获取单元
1101
生成单元
1102

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/105640

A. CLASSIFICATION OF SUBJECT MATTER

G06F 3/06(2006.01)i; G06F 11/00(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, EPODOC, WPI, CNKI, 百度学术: 纠删码, 校验单元, 分条, 条带, 冗余比, 配比, 生成 3d 新, 缓存, 硬盘, EC, erasure code, check unit, stripe, redundancy, ratio, proportion, cache, disk

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2017173623 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 12 October 2017 (2017-10-12) description, pages 7-10, figures 1-2	1-20
A	US 2015378820 A1 (QUANTUM CORPORATION) 31 December 2015 (2015-12-31) entire document	1-20
A	CN 109189326 A (HUAWEI TECHNOLOGIES CO., LTD.) 11 January 2019 (2019-01-11) entire document	1-20
A	CN 106201766 A (SHENZHEN CHINA BLOG INFORMATION TECHNOLOGY CO., LTD.) 07 December 2016 (2016-12-07) entire document	1-20
A	CN 107436733 A (HUAWEI TECHNOLOGIES CO., LTD.) 05 December 2017 (2017-12-05) entire document	1-20
A	CN 105630423 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 01 June 2016 (2016-06-01) entire 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
22 September 2021	**15 October 2021**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103577274 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 12 February 2014 (2014-02-12) entire document	1-20
A	CN 110865901 A (HUAWEI TECHNOLOGIES CO., LTD.) 06 March 2020 (2020-03-06) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/105640

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2017173623	A1	12 October 2017	CN	107484427	B	06 November 2020
				EP	3336706	A4	20 March 2019
				US	2018239671	A1	23 August 2018
				EP	3336706	A1	20 June 2018
				CN	107484427	A	15 December 2017
US	2015378820	A1	31 December 2015	US	9465692	B2	11 October 2016
CN	109189326	A	11 January 2019	CN	109189326	B	08 September 2020
CN	106201766	A	07 December 2016	WO	2018018827	A1	01 February 2018
				CN	106201766	B	20 March 2018
CN	107436733	A	05 December 2017	EP	3617867	A1	04 March 2020
				WO	2019000950	A1	03 January 2019
				US	2020125286	A1	23 April 2020
				CN	112328168	A	05 February 2021
				EP	3617867	A4	27 May 2020
				CN	107436733	B	06 November 2020
CN	105630423	A	01 June 2016	CN	105630423	B	27 November 2018
CN	103577274	A	12 February 2014	US	9229810	B2	05 January 2016
				US	2014040702	A1	06 February 2014
				US	2016011941	A1	14 January 2016
				CN	103577274	B	06 July 2016
				US	9417963	B2	16 August 2016
CN	110865901	A	06 March 2020	CN	110865901	B	04 May 2021

国际检索报告

国际申请号
PCT/CN2021/105640

A. 主题的分类

G06F 3/06(2006.01)i; G06F 11/00(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, EPODOC, WPI, CNKI, 百度学术：纠删码，校验单元，分条，条带，冗余比，配比，生成 3d 新，缓存，硬盘，EC, erasure code, check unit, stripe, redundancy, ratio, proportion, cache, disk

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	WO 2017173623 A1 (华为技术有限公司) 2017年 10月 12日 (2017-10-12) 说明书第7-10页，附图1-2	1-20
A	US 2015378820 A1 (QUANTUM CORPORATION) 2015年 12月 31日 (2015-12-31) 全文	1-20
A	CN 109189326 A (华为技术有限公司) 2019年 1月 11日 (2019-01-11) 全文	1-20
A	CN 106201766 A (深圳市中博科创信息技术有限公司) 2016年 12月 7日 (2016-12-07) 全文	1-20
A	CN 107436733 A (华为技术有限公司) 2017年 12月 5日 (2017-12-05) 全文	1-20
A	CN 105630423 A (华中科技大学) 2016年 6月 1日 (2016-06-01) 全文	1-20
A	CN 103577274 A (国际商业机器公司) 2014年 2月 12日 (2014-02-12) 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
2021年 9月 22日	2021年 10月 15日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局(ISA/CN) 中国 北京市海淀区蓟门桥西土城路6号 100088	刘梦瑶
传真号 (86-10)62019451	电话号码 86-(10)-53961396

国际检索报告

国际申请号
PCT/CN2021/105640

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 110865901 A (华为技术有限公司) 2020年 3月 6日 (2020 - 03 - 06) 全文	1-20

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/105640

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)
WO	2017173623	A1	2017年 10月 12日	CN	107484427	B	2020年 11月 6日
				EP	3336706	A4	2019年 3月 20日
				US	2018239671	A1	2018年 8月 23日
				EP	3336706	A1	2018年 6月 20日
				CN	107484427	A	2017年 12月 15日
US	2015378820	A1	2015年 12月 31日	US	9465692	B2	2016年 10月 11日
CN	109189326	A	2019年 1月 11日	CN	109189326	B	2020年 9月 8日
CN	106201766	A	2016年 12月 7日	WO	2018018827	A1	2018年 2月 1日
				CN	106201766	B	2018年 3月 20日
CN	107436733	A	2017年 12月 5日	EP	3617867	A1	2020年 3月 4日
				WO	2019000950	A1	2019年 1月 3日
				US	2020125286	A1	2020年 4月 23日
				CN	112328168	A	2021年 2月 5日
				EP	3617867	A4	2020年 5月 27日
				CN	107436733	B	2020年 11月 6日
CN	105630423	A	2016年 6月 1日	CN	105630423	B	2018年 11月 27日
CN	103577274	A	2014年 2月 12日	US	9229810	B2	2016年 1月 5日
				US	2014040702	A1	2014年 2月 6日
				US	2016011941	A1	2016年 1月 14日
				CN	103577274	B	2016年 7月 6日
				US	9417963	B2	2016年 8月 16日
CN	110865901	A	2020年 3月 6日	CN	110865901	B	2021年 5月 4日