



(12) 发明专利

(10) 授权公告号 CN 110580204 B

(45) 授权公告日 2022.12.06

(21) 申请号 201910460383.3

(51) Int.CI.

(22) 申请日 2019.05.30

G06F 11/10 (2006.01)

(65) 同一申请的已公布的文献号

G06F 11/08 (2006.01)

申请公布号 CN 110580204 A

(56) 对比文件

(43) 申请公布日 2019.12.17

WO 2017041233 A1, 2017.03.16

(30) 优先权数据

审查员 陈学元

62/682,763 2018.06.08 US

16/103,907 2018.08.14 US

(73) 专利权人 三星电子株式会社

地址 韩国京畿道水原市

(72) 发明人 瑞卡·皮塔楚玛尼 奇亮奭

(74) 专利代理机构 北京铭硕知识产权代理有限公司 11286

专利代理人 方成 张川绪

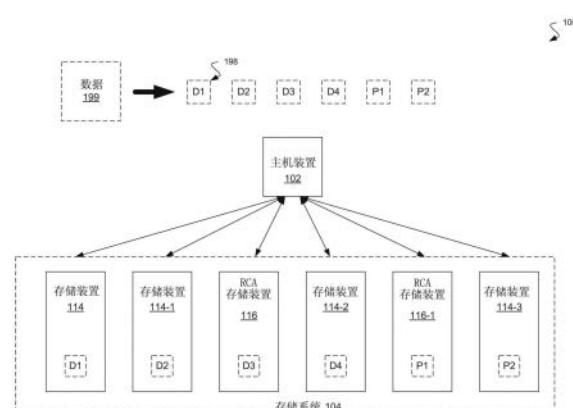
权利要求书3页 说明书12页 附图7页

(54) 发明名称

数据存储设备和数据存储系统

(57) 摘要

提供了数据存储设备和数据存储系统。根据一个总体方面，一种设备可包括被配置为计算用于数据纠错的至少一种类型的数据再生码的再生码感知(RCA)存储装置。RCA存储装置可包括被配置为以组块存储数据的存储器，其中，组块又包括数据块。RCA存储装置可包括被配置为当被外部主机装置请求时基于所选数量的数据块来计算数据再生码的处理器。RCA存储装置可包括被配置为将数据再生码发送到外部主机装置的外部接口。



1.一种数据存储设备,包括:

再生码感知存储装置,被配置为计算用于数据纠错的至少一种类型的数据再生码;

再生码感知存储装置包括:

存储器,被配置为以包括数据块的组块存储数据;

处理器,被配置为:基于与外部主机装置相关联的请求,基于所选数量的数据块来计算数据再生码;以及

外部接口,被配置为:

从外部主机装置接收命令,其中,所述命令开启处理器的用于计算数据再生码的能力,以及

将数据再生码发送到外部主机装置,其中,外部主机装置被配置为基于数据再生码重建出错的数据。

2.根据权利要求1所述的数据存储设备,其中,再生码感知存储装置还包括:

代码存储器,被配置为存储一个或多个指令集,所述一个或多个指令集被配置为生成不同的数据再生码;以及

其中,处理器被外部主机装置配置为选择所述一个或多个指令集中的一个指令集来计算数据再生码。

3.根据权利要求2所述的数据存储设备,其中,代码存储器被配置为具有由外部主机装置写入代码存储器的所述一个或多个指令集。

4.根据权利要求1所述的数据存储设备,其中,所述命令基于所选数量的数据块启用数据再生码的生成。

5.根据权利要求1所述的数据存储设备,其中,外部接口被配置为:

从外部主机装置接收指示请求修复数据并指示数据再生码将被计算的修复命令;以及将数据再生码返回到外部主机装置,其中,数据再生码的大小小于数据集的大小。

6.根据权利要求1所述的数据存储设备,其中,处理器被配置为:当被外部主机装置请求时,经由数据再生技术计算不同版本的数据再生码;以及

其中,由外部主机装置确定由处理器计算的版本。

7.一种数据存储系统,包括:

主机装置,被配置为:

将数据作为数据的组块存储在分布式存储系统之中;

检测数据的组块是否与错误相关联;以及

响应于检测到错误,经由数据再生技术,基于数据的组块来重构与错误相关联的数据的组块;以及

分布式存储系统,包括:

存储装置,被配置为存储数据的相应的组块,

其中,存储装置包括:至少一个再生码感知存储装置,被配置为内部地计算至少一种类型的数据再生码,

其中,主机装置被配置为:基于数据再生码和数据的组块来重构与错误相关联的数据的组块。

8.根据权利要求7所述的数据存储系统,其中,再生码感知存储装置包括:

存储器,被配置为以组块存储数据,其中,组块又包括数据块;

处理器,被配置为:基于所选数量的数据块来计算数据再生码;以及
外部接口,被配置为将数据再生码发送到主机装置。

9.根据权利要求7所述的数据存储系统,其中,主机装置被配置为:
确定能够内部地计算数据再生码的存储装置;

向存储装置请求数据的组块或数据的部分,并通过主机装置至少部分地基于数据的组块或数据的部分来计算数据再生码。

10.根据权利要求9所述的数据存储系统,其中,主机装置被配置为:至少部分地基于以下因素中的一个或多个来确定将数据再生码的计算卸载到存储装置:

存储装置可用的数据再生技术;

与分布式存储系统相关联的可用带宽的量;

与数据的组块或数据的部分的大小进行比较的数据再生码的大小;以及
主机装置内可用的计算能力的量。

11.根据权利要求9所述的数据存储系统,其中,主机装置被配置为:基于由主机装置计算的数据再生码和由相应的存储装置计算的数据再生码,重构错误的组块。

12.根据权利要求9所述的数据存储系统,其中,主机装置被配置为:

经由第一协议与能够内部地计算数据再生码的第一存储装置通信;以及
经由第二协议与不能内部地计算数据再生码的第二存储装置通信。

13.根据权利要求7所述的数据存储系统,其中,主机装置被配置为:

检测能够内部地计算相应的数据再生码的存储装置;以及

将与数据再生技术相关联的指令存储在相应的存储装置上,使得存储装置被配置为经由数据再生技术计算数据再生码。

14.根据权利要求13所述的数据存储系统,其中,主机装置被配置为:

通过至少部分地检测能够经由主机装置所选择的数据再生技术来计算数据再生码的存储装置,来检测能够内部地计算相应的数据再生码的存储装置。

15.一种数据存储系统,包括:

主机装置,被配置为:

将数据以组块存储在存储系统之中;

检测组块与错误相关联;以及

响应于检测到错误,经由数据再生技术,至少部分地基于组块来纠正错误;以及

存储系统,包括:

存储装置,被配置为存储数据的相应的组块,

其中,存储装置包括:至少一个再生码感知存储装置,被配置为内部地计算至少一种类型的数据再生码,并且,其中,再生码感知存储装置包括:

存储器,被配置为以包括数据块的组块存储数据;

处理器,被配置为:根据与主机装置相关联的请求,基于所选数量的数据块来计算数据再生码;

代码存储器,被配置为存储一个或多个指令集,其中,所述一个或多个指令集被配置为生成不同的数据再生码;以及

外部接口,被配置为将数据再生码发送到主机装置,

其中,主机装置被配置为:至少部分地基于组块和基于数据再生码来纠正错误。

16.根据权利要求15所述的数据存储系统,其中,主机装置被配置为:

将一个指令集写入再生码感知存储装置的代码存储器,其中,所述一个指令集被配置为便于所述再生码感知存储装置经由数据再生技术进行计算。

17.根据权利要求15所述的数据存储系统,其中,主机装置被配置为:通过至少部分地将一个或多个数据再生码的计算动态地卸载到一个或多个相应的再生码感知存储装置来纠正错误。

18.根据权利要求17所述的数据存储系统,其中,主机装置被配置为:通过由主机装置针对由存储系统内的存储装置存储的数据的一个或多个组块计算数据再生码来纠正错误,并且,其中,由主机装置进行的计算包括从存储装置发送数据的组块的至少一部分;以及

其中,由再生码感知存储装置计算并发送到主机装置的数据再生码的大小小于从存储装置发送到主机装置的数据的组块的至少一部分的大小。

19.根据权利要求15所述的数据存储系统,其中,存储装置包括非再生码感知存储装置,以及

其中,主机装置被配置为:

经由第一协议与再生码感知存储装置通信;以及

经由第二协议与非再生码感知存储装置通信。

数据存储设备和数据存储系统

技术领域

[0001] 本描述涉及数据存储,更具体地,涉及用于存储装置辅助的低带宽数据修复的系统、装置和/或方法。

背景技术

[0002] 在编码理论中,擦除码是在比特擦除(而不是比特错误)的假设下的前向纠错(FEC)码,其中,前向纠错码将k个符号的消息转换成具有n个符号的较长消息(码字),使得原始消息可从n个符号的子集被恢复。分数 $r = k/n$ 被称为码率。分数 k'/k 被称为接收效率,其中, k' 表示恢复所需的符号的数量。

[0003] 再生码解决了从现有的编码段重建(也称为修复)丢失的编码段的问题。更详细地说,再生码是一类旨在减少修复期间的下载量同时保持传统最大距离可分(MDS)码的存储效率的码。这样的问题发生在分布式存储系统中,其中,在分布式存储系统中,维护编码冗余的通信是问题。

[0004] 分布式存储系统通常是计算机网络,其中,在该计算机网络中,信息通常以复制的方式存储在一个以上的节点或装置上。分布式存储系统通常用于表示分布式数据库或计算机网络,其中,在分布式数据库中,用户将信息存储在多个节点上,在计算机网络中,用户将信息存储在多个对等网络节点上。分布式存储系统通常使用错误检测和纠正技术。当原始文件的部分被损坏或不可用时,一些分布式存储系统使用前向纠错技术来恢复该文件、组块(chunk)或二进制大对象(blob)。其他分布式存储系统再次尝试从不同的镜像下载该文件。

发明内容

[0005] 根据一个总体方面,一种设备可包括被配置为计算用于数据纠错的至少一种类型的数据再生码的再生码感知(RCA)存储装置。RCA存储装置可包括被配置为以组块存储数据的存储器,其中,组块又包括数据块。RCA存储装置可包括被配置为当被外部主机装置请求时基于所选数量的数据块来计算数据再生码的处理器。RCA存储装置可包括被配置为将数据再生码发送到外部主机装置的外部接口。

[0006] 根据另一总体方面,一种系统可包括主机装置和分布式存储系统。主机装置可被配置为将数据作为多个组块存储在分布式存储系统之中,检测至少一个组块何时与错误相关联,并且响应于所述错误,经由数据再生技术,至少部分地基于所述数据的所述多个组块来重构与所述错误相关联的组块。分布式存储系统可包括多个存储装置,其中,每个存储装置被配置为至少存储相应的组块,并且,其中,所述多个存储装置包括至少一个再生码感知(RCA)存储装置,其中,每个RCA存储装置被配置为内部地计算至少一种类型的数据再生码。

[0007] 根据另一总体方面,一种系统可包括主机装置,其中,主机装置被配置为:将数据以多个组块存储在存储系统之中,检测至少一个组块何时与错误相关联,并且响应于所述错误,经由数据再生技术,至少部分地基于所述数据的所述多个组块来纠正所述错误。所述

系统可包括存储系统，其中，存储系统包括多个存储装置，其中，每个存储装置被配置为至少存储所述数据的相应的组块，并且，其中，所述多个存储装置包括至少一个再生码感知(RCA)存储装置，其中，每个RCA存储装置被配置为内部地计算至少一种类型的数据再生码。所述RCA存储装置可包括：存储器，被配置为以组块存储数据，其中，每个组块包括数据块；处理器，被配置为当被主机装置请求时基于所选数量的数据块来计算数据再生码；代码存储器，被配置为存储多个指令集，其中，每个指令集生成不同的数据再生码；以及外部接口，被配置为将数据再生码发送到主机装置。

[0008] 一个或多个实施方式的细节在附图和下面的描述中阐述。从说明书和附图以及从权利要求书中，其他特征将是清楚的。

[0009] 一种用于数据存储的系统和/或方法，更具体地说，一种用于存储装置辅助的低带宽数据修复的系统和/或方法基本上如结合附图中的至少一个所示和/或所述，如权利要求中更完整地阐述。

附图说明

- [0010] 图1是根据公开的主题的系统的示例实施例的框图。
- [0011] 图2A是根据公开的主题的系统的示例实施例的框图。
- [0012] 图2B是根据公开的主题的系统的示例实施例的框图。
- [0013] 图2C是根据公开的主题的系统的示例实施例的框图。
- [0014] 图2D是根据公开的主题的系统的示例实施例的框图。
- [0015] 图3是根据公开的主题的技术的示例实施例的流程图。
- [0016] 图4是可包括根据公开的主题的原理形成的装置的信息处理系统的示意性框图。
- [0017] 各种附图中的相同的附图标记指示相同的元件。

具体实施方式

[0018] 在下文中，将参照示出了一些示例实施例的附图对各种示例实施例进行更加全面的描述。然而，本公开的主题可以以多种不同的形式来实现，并且不应该被解释为限于这里阐述的示例实施例。相反，提供这些示例实施例使得本公开将是彻底的和完整的，并且将本公开的主题的范围全面地传达给本领域技术人员。在附图中，为了清楚，可能夸大层和区域的大小以及相对大小。

[0019] 将理解，当元件或层被称为“在”另一个元件或层“上”、“连接到”或“结合到”另一个元件或层时，所述元件或层可以直接在所述另一个元件或层上、直接连接到或结合到所述另一个元件或层，或者可存在中间元件或层。相反，当元件被称为“直接在”另一个元件或层“上”、“直接连接到”或“直接结合到”另一个元件或层时，不存在中间元件或层。相同的标号始终指代相同的元件。如这里使用的，术语“和/或”包括一个或多个关联的所列项的任何组合和全部组合。

[0020] 将理解，尽管可在这里使用术语第一、第二、第三等描述各种元件、组件、区域、层和/或部分，但是这些元件、组件、区域、层和/或部分不应该受这些术语限制。这些术语只是用于将一个元件、组件、区域、层或部分与另一个区域、层或部分区分开来。因此，在不脱离本公开的主题的教导的情况下，以下讨论的第一元件、组件、区域、层或部分可被称为第二

元件、组件、区域、层或部分。

[0021] 为了描述的方便,空间相对术语(诸如“在…以下”、“在…下面”、“低于”、“在…上面”、“上面的”等)可在这里使用以描述在附图中示出的一个元件或特征与另外的元件或特征的关系。将理解,空间相对术语意图包含除了在附图中描述的方位之外的装置在使用或操作中的不同方位。例如,如果附图中的装置被翻转,则被描述为在其他元件或特征“下面”或“以下”的元件将会被定位为在其他元件或特征“上面”。因此,示例性术语“在…下面”可包含上面和下面的两个方位。装置可被另外定位(旋转90度或其他方位),并且这里使用的与空间相对描述符被相应地解释。

[0022] 同样,为了描述的方便,电气术语(诸如,“高”、“低”、“上拉”、“下拉”、“1”、“0”等)可在这里使用以描述如附图中示出的相对于其他电压电平或相对于另外的元件或特征的电压电平或电流。将理解,电气相对术语意图包含除了在附图中描述的电压或电流之外的装置在使用或操作中的不同参考电压。例如,如果附图中的装置或信号被翻转,或者使用其他参考电压、电流或电荷,则与新的参考电压或电流相比,被描述为“高”或“上拉”的元件将会是“低”或“下拉”。因此,示例性术语“高”可包含相对低或相对高的电压或电流二者。装置可另外基于参考的不同电气框架和相应地解释的在这里使用的电气相对描述符。

[0023] 这里使用的术语仅是用于描述特定的示例实施例的目的,而不意图限制本公开的主题。如这里所使用的,除非上下文明确地另有指示,否则单数形式也意图包括复数形式。还将理解,当在本说明书中使用术语“包括”和/或“包含”时,表明存在叙述的特征、整体、步骤、操作、元件和/或组件,但不排除存在或添加一个或多个其他特征、整体、步骤、操作、元件、组件和/或它们的组。

[0024] 在这里参照作为理想化的示例实施例(和中间结构)的示意图的截面图对示例实施例进行描述。这样,由于例如制造技术和/或公差导致的示图的形状的变化将是预期的。因此,示例实施例不应被解释为限于这里示出的区域的特定形状,而是将包括由于例如制造引起的形状上的偏差。例如,示出为矩形的注入区域通常将在其边缘处具有圆形的或弯曲的特征和/或注入浓度的梯度,而不是从注入区域到非注入的区域的二元变化。同样地,通过注入形成的埋区可导致在埋区与发生注入的表面之间的区域中的一些注入。因此,在附图中示出的区域在本质上是示意性的,并且所述区域的形状不意图示出装置的区域的实际形状,并且不意图不限制本公开的主题的范围。

[0025] 除非另有定义,否则这里使用的所有术语(包括技术术语和科学术语)具有和本公开的主题所属领域的普通技术人员通常理解的含义相同的含义。还将理解,除非在这里明确地如此定义,否则术语(诸如在通用字典中定义的术语)应该被解释为具有与它们在相关领域的语境中的含义一致的含义,而将不被解释为理想化或过于形式化的意义。

[0026] 在下文中,将参照附图详细说明示例实施例。

[0027] 图1是根据公开的主题的系统100的示例实施例的框图。在示出的实施例中,系统100可包括分布式存储系统104,其中,分布式存储系统104在多个节点或存储装置之间存储数据。

[0028] 分布式存储系统通常用于提供大规模的可靠存储。通常,这是通过在大量节点或存储装置之间扩展冗余或纠错(例如,奇偶校验)来实现的。然而,当节点或存储装置例如由于网络错误、硬件故障等而离线时,数据变得被怀疑为可能损坏,或者至少具有降低的冗余

级别。存储系统分布得越广,这就发生得越频繁。

[0029] 可采用许多技术(例如,镜像、里德-所罗门(Reed-Solomon)编码)来防止这种发生,而公开的主题关注再生编码。在这样的实施例中,使用基于剩余数据段的公式来再生或重构丢失的数据段(组块)。

[0030] 在示出的实施例中,系统100可包括被配置为管理分布式存储系统104的一个或多个主机装置102。主机装置102可包括从存储系统104读取和向存储系统104写入的计算装置(例如,计算机、服务器、虚拟机)。当发生错误(例如,丢失的数据组块)时,主机装置102通常负责检测,并且如果可能的话修复该错误。

[0031] 在示出的实施例中,每个数据集199可被主机装置102分解或分段成多个较小的数据段或组块198。在示出的实施例中,数据199被划分成组块198D1、D2、D3和D4。此外,在各种实施例中,主机装置102可将一些形式的冗余应用于数据组块198(诸如,奇偶校验组块P1和P2(由于它们也是组块,所以也被编号为198))。

[0032] 在本领域的说法中,原始数据组块198(D1、D2、D3和D4)的数量被描述为变量K或k。同样,冗余数据组块198(P1和P2)的数量被描述为变量R或r。使得组块198的总数是K+R。在示出的实施例中,K等于4,R等于2,并且K+R等于6;然而,应当理解,以上仅仅是公开的主题不限于的说明性示例。

[0033] 在示出的实施例中,主机装置102将这些组块198(原始和冗余二者)中的每个存储在存储系统104的各个节点或存储装置上。在示出的实施例中,存储装置114存储组块198D1,存储装置114-1存储组块198D2,存储装置116(即,RCA存储装置116)存储组块198D3,存储装置114-2存储组块198D4,存储装置116-1(即,RCA存储装置116-1)存储组块198P1,存储装置114-3存储组块198P2。在各种实施例中,存储装置114/116的数量可不等于组块198的数量。

[0034] 在各种实施例中,组块198可能丢失(例如,网络或硬件故障),或者可能另外与错误相关联。在示出的实施例中,让我们假设组块198D3(和存储装置116)突然变得不可用。主机装置102在检测到错误时可尝试重建组块198D3或另外纠正错误。

[0035] 在这样的实施例中,如果一个组块故障(例如,组块198D3),并且在原始数据199中存在总共K个(例如,4个)组块,则至少K个(例如,4个)节点或存储装置114/116必须向主机装置102发送信息以恢复故障的组块(例如,组块198D3)。注意,这些K个(例如,4个)组块可来自K+R个(例如,6个)组块中的任意组块。例如,组块198D1、D2、D4和P1可用于重建组块198D3。

[0036] 再生码通过从D个节点发送少于完整组块大小信息的信息来减小修复带宽,其中,通常,D>K。换句话说,通过使用巧妙的公式,主机装置102能够通过不使用完整的组块198D1、D2、D4和P1,而是通过仅使用198D1、D2、D4、P1和P2的一部分来重建丢失的组块198D3。再生码通常从更多的存储装置114/116获取信息,但是与非再生码相比,再生码从每个存储装置114/116获取更少的信息。

[0037] 例如,如果6个数据组块被使用(K=6)并且6个冗余组块被使用(R=6并且K+R=12),且每个组块大小为16MB,则标准里德-所罗门(Reed-Solomon,RS)纠错编码方案会要求:6(K)个16MB组块被发送到主机装置或者96MB数据被发送以纠正丢失的16MB组块。相反,如果使用再生技术,则会读取全部12个(在这种情况下,K+R或D个)组块的部分,但是由于仅

使用每个组块的部分(例如,2.7MB),所以发送到主机装置的总量可较低(例如,29.7MB)。

[0038] 通常,再生码具有存储和带宽权衡。在各种实施例中,通常存在两类或两组再生码。如果存储开销最小,则再生码被称为最小存储再生(MSR)码。如果修复带宽对于增加的存储开销是最小的,则再生码被称为最小带宽再生(MBR)码。在这些宽泛的类别中,可采用各种特定技术或公式来执行再生码。应当理解,以上仅仅是公开的主题不限于的一些说明性示例。

[0039] 返回图1,在示出的实施例中,存储系统104可包括多个存储装置114/116。每个存储装置114/116可被配置成以组块或另外方式存储数据。在示出的实施例中,存储装置114可以是相对传统的存储装置(诸如,硬盘驱动器、固态驱动器,或者甚至是易失性存储器)。

[0040] 然而,在示出的实施例中,存储系统104还可包括再生码感知(RCA)存储装置116。在这样的实施例中,与传统或非RCA存储装置114不同,RCA存储装置116可被配置为帮助计算数据再生码并包括允许RCA存储装置116帮助计算数据再生码的组件。如稍后更详细讨论的,主机装置102能够将数据再生码的计算中的一些计算动态地卸载到RCA存储装置116。在各种实施例中,这可减少在主机装置102与存储系统104之间来回发送的消息量、在主机装置102与存储系统104之间传送的数据量和/或主机装置102上的计算负载。在各种实施例中,如果存储系统104的相应的存储装置能够内部地计算相应数据再生码,则主机装置102可至少部分地基于以下因素中的一个或多个来确定是否将相应数据再生码的计算卸载到相应的存储装置:相应的存储装置可用的数据再生技术;与分布式存储系统(例如,存储系统104)相关联的可用带宽的量;与数据的组块或数据的部分的大小相比的数据再生码的大小;以及主机装置102内可用的计算能力的量。应当理解,以上仅仅是公开的主题不限于的一些说明性示例。

[0041] 在各种实施例中,RCA存储装置116可以是可编程的,从而主机装置102能够使用最新的或期望的再生码公式或技术来更新RCA存储装置116。在这样的实施例中,RCA存储装置116能够存储多种再生技术,并且使多种再生技术中的一种被主机装置102动态地或半静态地选择。在这样的实施例中,主机装置102可选择在给定时刻应该采用哪种再生技术。

[0042] 在各种实施例中,存储系统104可以是分布式的。在这样的实施例中,存储装置114/116可在物理上彼此远离并且经由网络协议进行通信。在另一实施例中,存储装置114/116可以是相对本地化的(例如,在服务器农场或同一建筑物中),但是仍然经由网络协议进行通信。在又一实施例中,存储系统104可以不是分布式的。在这样的实施例中,公开的主题可用于不使用网络协议(例如,USB、SATA)的本地装置(例如,同一机器)。应当理解,以上仅仅是公开的主题不限于的一些说明性示例。

[0043] 在各种实施例中,再生码感知(RCA)存储装置116可包括计算不同类型或版本的再生码的能力。在这样的实施例中,可由主机装置102动态地选择期望的类型或版本的再生码。在一些实施例中,RCA存储装置116能够将数据分割成较小的块或数据包,计算擦除码或擦除码的部分,处理一个或多个数据组块以修复另一故障组块等。

[0044] 在各种实施例中,通信协议可存在于主机装置102与RCA存储装置116之间,以使用任何再生码或技术来实现数据可靠性。在这样的实施例中,协议可考虑选择再生技术、传递输入、指导期望技术的操作以及检索任何输出。在一些实施例中,协议可定义当协议在包含RCA和非RCA存储装置116/114两者混合环境中工作时的主机行为以及如何与两者交互。

在各种实施例中，主机系统102可使用协议来设置RCA存储装置116、编码/读取/写入用户数据并在数据修复期间卸载计算、减少数据流量并使用RCA存储装置116的能力加速计算和重建原始数据。

[0045] 图2A是根据公开的主题的系统201的示例实施例的框图。在示出的实施例中，系统201示出了主机装置210与存储装置212之间的交互，以计算第一种(类型1)再生码。在各种实施例中，系统201可用于传统或非RCA存储装置，并且如果RCA存储装置的RCA功能未被采用，则系统201甚至可用于该RCA存储装置。

[0046] 在示出的实施例中，系统201可包括主机装置210和存储装置212。在这样的实施例中，主机装置210可包括用于执行指令并执行计算的处理器232、用于至少临时存储数据或数据段的存储器234以及用于与存储装置212或更一般地与存储系统(未示出)通信的接口236。在这样的实施例中，存储装置212可包括被配置为存储数据的存储器224。在各种实施例中，这个存储器224可以是非易失性的或易失性的。

[0047] 在示出的实施例中，组块214被细分为块216。在这样的实施例中，主机装置210可以从存储在存储装置212上的一个或多个组块214(以及从存储在其他存储装置上的K-1个组块)获取块216，并计算再生码218(R1)。

[0048] 在这种再生码技术(类型1)中，块216由较小的数据包(未示出)组成。对于每个节点或存储装置212，主机装置210使用各种数据包计算奇偶校验数据包或再生码218。每个存储装置的相应的再生码218用于重构丢失的或出错的组块。通常，对于类型1再生码技术，计算是线性的并且依赖于故障的组块。发回的数据量取决于子分包化级别(sub-packetization level)和功能。

[0049] 在示出的实施例中，一旦主机装置210检测到错误，主机装置210就可向存储装置212发送数据读取请求或命令242A。数据读取命令242A可包括将要读取哪个组块214(例如，组块214C)。然后，存储装置212经由数据读取响应或消息244A向主机装置210发送期望的组块214。在各种实施例中，这可以全部使用传统的主机到存储装置212协议(例如，SATA)来完成。

[0050] 在通过接口236接收到期望的组块214C时，主机装置210可将组块214C或块216存储在存储器234中。然后，处理器232可执行期望的再生码技术287。虽然再生码技术287被示出为简单加法或布尔异或，但是应当理解，以上仅仅是公开的主题不限于的一些说明性示例。如上所述，在各种实施例中，这可包括将块216细分为更小的数据包。再生码技术287可计算或生成再生码218(R1)，然后，再生码218(R1)和与其他组块或存储装置相关联的再生码可用于重构或修复出错的组块。

[0051] 图2B是根据公开的主题的系统203的示例实施例的框图。在示出的实施例中，系统203示出了主机装置210与RCA存储装置252之间的交互，以计算第一种(类型1)再生码。在各种实施例中，系统203可仅用于RCA存储装置而不用于非RCA存储装置。

[0052] 在示出的实施例中，系统201可包括主机装置210和RCA存储装置252。在这样的实施例中，主机装置210可包括用于执行指令和执行计算的处理器232、用于至少临时存储数据或数据段的存储器234以及用于与RCA存储装置252或更一般地与存储系统(未示出)通信的接口236。

[0053] 在这样的实施例中，RCA存储装置252可包括被配置为存储数据的存储器224。在各

种实施例中,该存储器224可以是非易失性的或易失性的。此外,在各种实施例中,RCA存储装置252可包括处理器222,其中,处理器222被配置为:当由主机装置210(通常在存储装置外部)请求时,基于所选数量的数据块216来计算数据再生码218。在各种实施例中,处理器222可包括可编程门阵列(例如,FGPA)、图形处理器(GPU)、通用处理器(例如,CPU)、控制器处理器或片上系统(SOC)。应当理解,以上仅仅是公开的主题不限于的一些说明性示例。RCA存储装置252可包括被配置为存储多个指令集229的代码存储器228,其中,每个指令集229生成不同的数据再生码或关于如何执行不同的再生码技术的指令。在各种实施例中,指令集229可被预配置到存储装置252中或在运行时间期间(例如,由主机装置210)动态添加/调整,或者采用二者组合的方式。RCA存储装置252可包括被配置为至少与主机装置210通信的外部接口226。

[0054] 在示出的实施例中,主机装置210可确定:存储装置252是否能够内部地计算数据再生码,或者通常是否是RCA存储装置。如果是,则主机装置210可确定:RCA存储装置252是否可执行期望的再生码技术,或者RCA存储装置252是否可被编程(经由代码存储器228)为这样做。如果不是,则可采用图2A中所示的技术。

[0055] 如果RCA存储装置252能够执行期望的再生码技术,则主机装置210可发布用于修复的读取(Read for Repair)命令242B。在各种实施例中,用于修复的读取命令242B可包括或指示以下项中的一个或多个:期望的再生或修复技术的指示、期望的数据包或块大小、期望的再生或修复技术的任何参数、数据或组块地址以及故障的组块号。应当理解,以上仅仅是公开的主题不限于的一些说明性示例。

[0056] 响应于命令242B,处理器222可检索期望的块216或组块214C。处理器222还可检索与期望的再生或修复技术相关联的指令集229。处理器222可执行期望的再生技术287并计算数据再生码(DRC)218(R1)。

[0057] 然后,RCA存储装置252可经由接口226向主机装置210发送数据再生码218(R1)(消息244B)。在这样的实施例中,数据再生码218(R1)可具有比经由图2A的消息244A发送的数据更小的大小或消耗更少的带宽。

[0058] 在示出的实施例中,消息242B和244B可需要与用于消息242A和244A的协议不同的协议。虽然消息242A和244A可被传统存储装置协议允许,但是消息242B和244B可能需要附加的和不同的信息,并因此需要新的消息协议或至少需要新的命令。

[0059] 在示出的实施例中,主机装置210随后可使用数据再生码218(R1)以及由其他RCA存储装置(未示出)提供的或由主机装置210自身生成的任何附加数据再生码来重建出错的数据组块。

[0060] 图2C是根据公开的主题的系统205的示例实施例的框图。在示出的实施例中,系统205示出主机装置210与存储装置212之间的交互以计算第二种(类型2)再生码。在各种实施例中,系统205可用于传统或非RCA存储装置,并且如果RCA存储装置的RCA功能未被采用,则系统205甚至可用于RCA存储装置。

[0061] 在示出的实施例中,系统205可包括主机装置210和存储装置212。主机装置210和存储装置212二者都可包括上面示出和描述的组件。

[0062] 在这样的再生码技术(类型2)中,计算数据再生码,使得必须读取更少的数据包(未示出)或块216。然而,这通常意味着期望的块216或数据包是提前完全已知的,但是随着

计算的进行而被逐个请求。虽然这种再生技术在理论上减少了网络带宽和数据读取二者，但是这种再生技术将一个大的读取转换为多个较小的读取，这对性能是不利的。

[0063] 在示出的实施例中，主机装置210使用期望的再生技术的部分288计算：如果块E1与错误相关联，则将需要块B1和B3（或块B1和B3的数据包）来修复出错的块E1。在这样的实施例中，一旦主机装置210检测到需要块B1，主机装置210就可向存储装置212发送数据读取请求或命令242C。数据读取命令242C可指示将要读取哪个块216（例如，块B1）。然后，存储装置212经由数据读取响应或消息244C向主机装置210发送期望的块216B1。在各种实施例中，这可全部使用传统的主机到存储装置协议（例如，SATA）来完成。

[0064] 在这样的实施例中，一旦主机装置210检测到需要块B3，主机装置210就可向存储装置212发送数据读取请求或命令246C。这通常作为与请求块B1的数据请求分开的第二数据请求来完成。数据读取命令246C可指示将要读取哪个块216（例如，现在是块B3）。然后，存储装置212经由数据读取响应或消息248C向主机装置210发送期望的块216B3。在各种实施例中，这可全部使用传统的主机到存储装置协议（例如，SATA）来完成。

[0065] 在通过接口236接收到期望的块216时，主机装置210可将块216存储在存储器234中。然后，处理器232可执行期望的再生码技术（由部分289示出）。再生码技术（或部分289）可计算或生成再生码219（R1），然后，再生码219（R1）和与其他组块或存储装置相关联的再生码可用于重构或修复出错的组块。

[0066] 图2D是根据公开的主题的系统207的示例实施例的框图。在示出的实施例中，系统207示出了主机装置210与RCA存储装置252之间的交互，以计算第二种（类型2）再生码。在各种实施例中，系统207可仅用于RCA存储装置而不用于非RCA存储装置。

[0067] 在示出的实施例中，系统207可包括主机装置210和存储装置252。主机装置210和存储装置252二者都可包括上面示出和描述的组件。

[0068] 在示出的实施例中，主机装置210可确定：存储装置252是否能够内部地计算数据再生码，或者通常是否是RCA存储装置。如果是，则主机装置210可确定：RCA存储装置252是否可执行期望的再生码技术，或者RCA存储装置252是否可被编程（经由代码存储器228）为这样做。如果不是，则可采用图2C中所示的技术。

[0069] 如果RCA存储装置252能够执行期望的再生码技术，则主机装置210可发布用于修复的读取命令242D。在各种实施例中，用于修复的读取命令242D可包括或指示以下项中的一个或多个：期望的再生或修复技术的指示、期望的数据包或块大小、期望的再生或修复技术的任何参数、数据或组块地址以及故障的组块号（例如，块E1）。应当理解，以上仅仅是公开的主题不限于的一些说明性示例。

[0070] 响应于命令242D，处理器222可检索与期望的再生或修复技术相关联的指令集229。处理器222可执行期望的再生技术或期望的再生技术的部分288。在这样的实施例中，处理器222可计算出期望的块是B1和B3。在这样的实施例中，这些块B1和B3可包括在由RCA存储装置252计算的数据再生码中。在这样的实施例中，这些块可被认为仅仅是对用于修复的读取命令242D的响应的部分。

[0071] 然后，RCA存储装置252可经由接口226向主机装置210发送期望的块B1和B3（消息244D）。在这样的实施例中，数据再生码或期望的块B1和B3可具有比经由图2C的消息244C和248C发送的数据更小的大小或消耗更少的带宽，或至少包括更少的消息并因此减少开销。

[0072] 在示出的实施例中,消息242D和244D可需要与用于消息242C、244C、246C和248C的协议不同的协议。虽然消息242C、244C、246C和248C可被传统存储装置协议允许,但是消息242D和244D可能需要附加的和不同的信息,因此需要新的消息协议或至少需要新的命令。

[0073] 在示出的实施例中,主机装置210随后可使用数据再生码或块B1和B3以及由其他RCA存储装置(未示出)提供的或由主机装置210自身生成的任何附加数据再生码或数据来重建出错的数据(E1)。

[0074] 图3是根据公开的主题的技术300的示例实施例的流程图。在各种实施例中,技术300可由诸如图1、图2A、图2B、图2C和图2D的系统的系统使用或产生。但是,应当理解,以上仅仅是公开的主题不限于的一些说明性示例。应当理解,公开的主题不限于由技术300所示的动作的顺序或数量。

[0075] 在示出的实施例中,为了简单起见,技术300示出了存储系统的所有装置是RCA存储装置或非RCA存储装置(即,同构存储系统)的示例。对于混合或异构存储系统,本领域技术人员将理解可如何扩展简化的技术300以基于单个存储装置进行应用。

[0076] 框302示出:在一个实施例中,可检测到与数据组块相关联的错误。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0077] 框304示出:如上所述,在一个实施例中,可对数据再生码(DRC)是由主机装置计算还是由各个RCA存储装置计算做出确定。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0078] 框306示出:如上所述,在一个实施例中,如果将由主机以更传统的方式计算DRC,则可对是否存在足够的现有数据来计算DRC做出确定。在一个这样的实施例中,这可包括确定在K+R个数据组块之中是否有K个组块可用。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0079] 框399示出:如上所述,在一个实施例中,如果没有足够的无错的组块存在以计算DRC,则除了出错的数据组块的重建之外,可能发生一些其他形式的错误处理。在各种实施例中,这可以仅仅是数据损坏或不可用的报告。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0080] 框308示出:如上所述,在一个实施例中,可从各种(例如,K+R个)存储装置读取所需数量的组块(例如,K个组块)。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0081] 框310示出:如上所述,在一个实施例中,主机装置可使用无错的组块(例如,K个组块)来重构或重建出错的组块。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0082] 框350示出:如上所述,在一个实施例中,可对是否存在足够的无错的组块(例如,D个组块)来计算DRC做出确定。如果不是,则在各种实施例中,技术300可采取尝试以框306开始的非RCA装置路径。否则,技术300可继续到框352。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0083] 框352示出:如上所述,在一个实施例中,可将用于修复的读取命令发布到全部(例

如,K+R个)存储装置中的所需数量(例如,D)的存储装置。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0084] 框354示出:如上所述,在一个实施例中,可对将要使用多个版本或类型的DRC技术中的哪一个做出确定。如上所述,在示出的实施例中,DRC技术的版本或类型被概括为上述类型1和2的技术,但是应当理解,这些类型仅仅是公开的主题不限于的一些说明性示例,并且此外,在那些宽泛的类型内可存在许多子类型。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0085] 框356示出:如上所述,在一个实施例中,如果选择了类型1DRC技术,则可将修复功能应用于RCA存储装置内的组块。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0086] 框358示出:如上所述,在一个实施例中,如果选择了类型2DRC技术,则可计算修复所需的块(或诸如数据包的其他子部分)。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0087] 框360示出:如上所述,在一个实施例中,一旦计算了DRC或所需的块,则可将DRC或块发送到主机装置。如上所述,在各种实施例中,这可包括比非RCA路径更小大小的数据或更小数量的消息。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0088] 框362示出:如上所述,在一个实施例中,主机装置可使用DRC或返回的块来重构或重建出错的组块。如上所述,在各种实施例中,由这个框示出的一个或多个动作中的一个或多个可由图1、图2A、图2B、图2C或图2D的设备或系统执行。

[0089] 在各种实施例中,主机装置可向RCA存储装置提供命令,其中,命令开启或关闭RCA存储装置的(或其中的处理器的)用于计算数据再生码的能力。在一个示例中,命令可基于所选数量的数据块开启或关闭RCA存储装置的(或其中的处理器的)用于计算数据再生码的能力。响应于该命令,RCA存储装置可作为图2B或图2D中所示的RCA存储装置或者图2A或图2C中所示的传统或非RCA存储装置来操作。

[0090] 图4是信息处理系统400的示意性框图,其中,信息处理系统400可包括根据公开的主题的原理形成的半导体装置。

[0091] 参照图4,信息处理系统400可包括根据公开的主题的原理构造的一个或多个装置。在另一实施例中,信息处理系统400可根据公开的主题的原理采用或执行一种或多种技术。

[0092] 在各种实施例中,例如,信息处理系统400可包括计算装置(诸如,膝上型计算机、台式机、工作站、服务器、刀片服务器、个人数字助理、智能电话、平板电脑和其他适当的计算机或它们的虚拟机或虚拟计算装置)。在各种实施例中,信息处理系统400可由用户(未示出)使用。

[0093] 根据公开的主题的信息处理系统400还可包括中央处理器(CPU)、逻辑或处理器410。在一些实施例中,处理器410可包括一个或多个功能单元块(FUB)或组合逻辑块(CLB)415。在这样的实施例中,组合逻辑块可包括各种布尔(Boolean)逻辑运算(例如,与非、或非、非、异或)、稳定逻辑器件(例如,触发器、锁存器)、其他逻辑器件或它们的组合。这些组

合逻辑运算可以以简单或复杂的方式配置,以处理输入信号来实现期望的结果。应当理解,虽然描述了同步组合逻辑运算的一些说明性示例,但是公开的主题不被这样限制,并且可包括异步运算或它们的混合。在一个实施例中,组合逻辑运算可包括多个互补金属氧化物半导体(CMOS)晶体管。在各种实施例中,这些CMOS晶体管可布置成执行逻辑运算的门;但是应当理解,其他技术可被使用并且在公开的主题的范围内。

[0094] 根据公开的主题的信息处理系统400还可包括易失性存储器420(例如,随机存取存储器(RAM))。根据公开的主题的信息处理系统400还可包括非易失性存储器430(例如,硬盘驱动器、光学存储器、NAND或闪存)。在一些实施例中,易失性存储器420、非易失性存储器430或者它们的组合或部分可被称为“存储介质”。在各种实施例中,易失性存储器420和/或非易失性存储器430可被配置为以半永久或基本上永久的形式存储数据。

[0095] 在各种实施例中,信息处理系统400可包括一个或多个网络接口440,其中,一个或多个网络接口440被配置为允许信息处理系统400成为通信网络的部分并经由通信网络进行通信。Wi-Fi协议的示例可包括但不限于电气和电子工程师协会(IEEE)802.11g、IEEE802.11n。蜂窝协议的示例可包括但不限于:IEEE802.16m(又名,高级无线MAN(城域网))、高级长期演进(LTE)、GSM(全球移动通信系统)演进(EDGE)的增强的数据速率、演进高速数据包接入(HSPA+)。有线协议的示例可包括但不限于IEEE802.3(又名,以太网)、光纤通道、电力线通信(例如,HomePlug,IEEE1901)。应当理解,以上仅仅是公开的主题不限于的一些说明性示例。

[0096] 根据公开的主题的信息处理系统400还可包括用户接口单元450(例如,显示适配器、触觉接口、人机接口装置)。在各种实施例中,这个用户界面单元450可被配置为从用户接收输入和/或向用户提供输出。也可使用其他种类的装置来提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感觉反馈(例如,视觉反馈、听觉反馈或触觉反馈);并且来自用户的输入可以以包括声学、语音或触觉输入的任何形式被接收。

[0097] 在各种实施例中,信息处理系统400可包括一个或多个其他装置或硬件组件460(例如,显示器或监视器、键盘、鼠标、相机、指纹读取器、视频处理器)。应当理解,以上仅仅是公开的主题不限于的一些说明性示例。

[0098] 根据公开的主题的信息处理系统400还可包括一个或多个系统总线405。在这样的实施例中,系统总线405可被配置为通信地结合处理器410、易失性存储器420、非易失性存储器430、网络接口440、用户接口单元450和一个或多个硬件组件460。由处理器410处理的数据或从信息处理系统400输入的数据可存储在非易失性存储器430或易失性存储器420中。

[0099] 在各种实施例中,信息处理系统400可包括或执行一个或多个软件组件470。在一些实施例中,软件组件470可包括操作系统(OS)和/或应用程序。在一些实施例中,OS可被配置为向应用程序提供一个或多个服务,并且管理或者充当信息处理系统400的应用程序和各种硬件组件(例如,处理器410、网络接口440)之间的中介。在这样的实施例中,信息处理系统400可包括一个或多个本地应用程序,其中,一个或多个本地应用程序可安装在本地(例如,在非易失性存储器430内)并且被配置为由处理器410直接执行并且与OS直接交互。在这样的实施例中,本地应用程序可包括预编译的机器可执行代码。在一些实施例中,本地应用程序可包括被配置为将源代码或目标代码转换成其后由处理器410执行的可执行代码。

的脚本解释器(例如,C shell(CSH)、AppleScript、AutoHotkey)或虚拟执行机(VM)(例如,Java虚拟机、微软公共语言运行时(Microsoft Common Language Runtime))。

[0100] 上述半导体装置可使用各种封装技术来封装。例如,根据公开的主题的原理构造的半导体装置可使用以下技术中的任何一种来封装:层叠封装件(PoP)技术、球栅阵列(BGA)技术、芯片级封装件(CSP)技术、塑料引线芯片载体(PLCC)技术、塑料双列直插式封装件(PDIP)技术、华夫封装件中裸片技术、晶圆形式裸片技术、板上芯片(COB)技术、陶瓷双列直插式封装件(CERDIP)技术、塑料公制四方扁平封装件(PMQFP)技术、塑料四方扁平封装件(PQFP)技术、小外形封装件(SOIC)技术、收缩型小外形封装件(SSOP)技术、薄型小外形封装件(TSOP)技术、薄型四方扁平封装件(TQFP)技术、系统级封装件(SIP)技术、多芯片封装件(MCP)技术、晶圆级制造封装件(WFP)技术、晶圆级处理堆叠封装件(WSP)技术或本领域技术人员已知的其他技术。

[0101] 方法步骤可由执行计算机程序的一个或多个可编程处理器执行,以通过对输入数据进行操作并生成输出来执行功能。方法步骤还可由专用逻辑电路(例如,FPGA(现场可编程门阵列)或ASIC(专用集成电路))执行,并且设备可被实现为专用逻辑电路(例如,FPGA(现场可编程门阵列)或ASIC(专用集成电路))。

[0102] 在各种实施例中,计算机可读介质可包括当被执行时使装置执行方法步骤中的至少一部分的指令。在一些实施例中,计算机可读介质可被包括在磁性介质、光学介质、其他介质或它们的组合(例如,CD-ROM、硬盘驱动器、只读存储器、闪存驱动器)中。在这样的实施例中,计算机可读介质可以是有形和非暂时体现的制品。

[0103] 虽然已经参照示例实施例描述了公开的主题的原理,但是对于本领域技术人员来说将清楚的是,在不脱离这些公开的构思的精神和范围的情况下,可对其进行各种改变和修改。因此,应当理解,上述实施例不是限制性的,而仅是说明性的。因此,公开的构思的范围将通过对所附权利要求及其等同物的最广泛允许的解释来确定,并且不应受到前述描述的限制或限定。因此,应当理解,所附权利要求旨在覆盖落入实施例的范围内的所有这样的修改和改变。

100

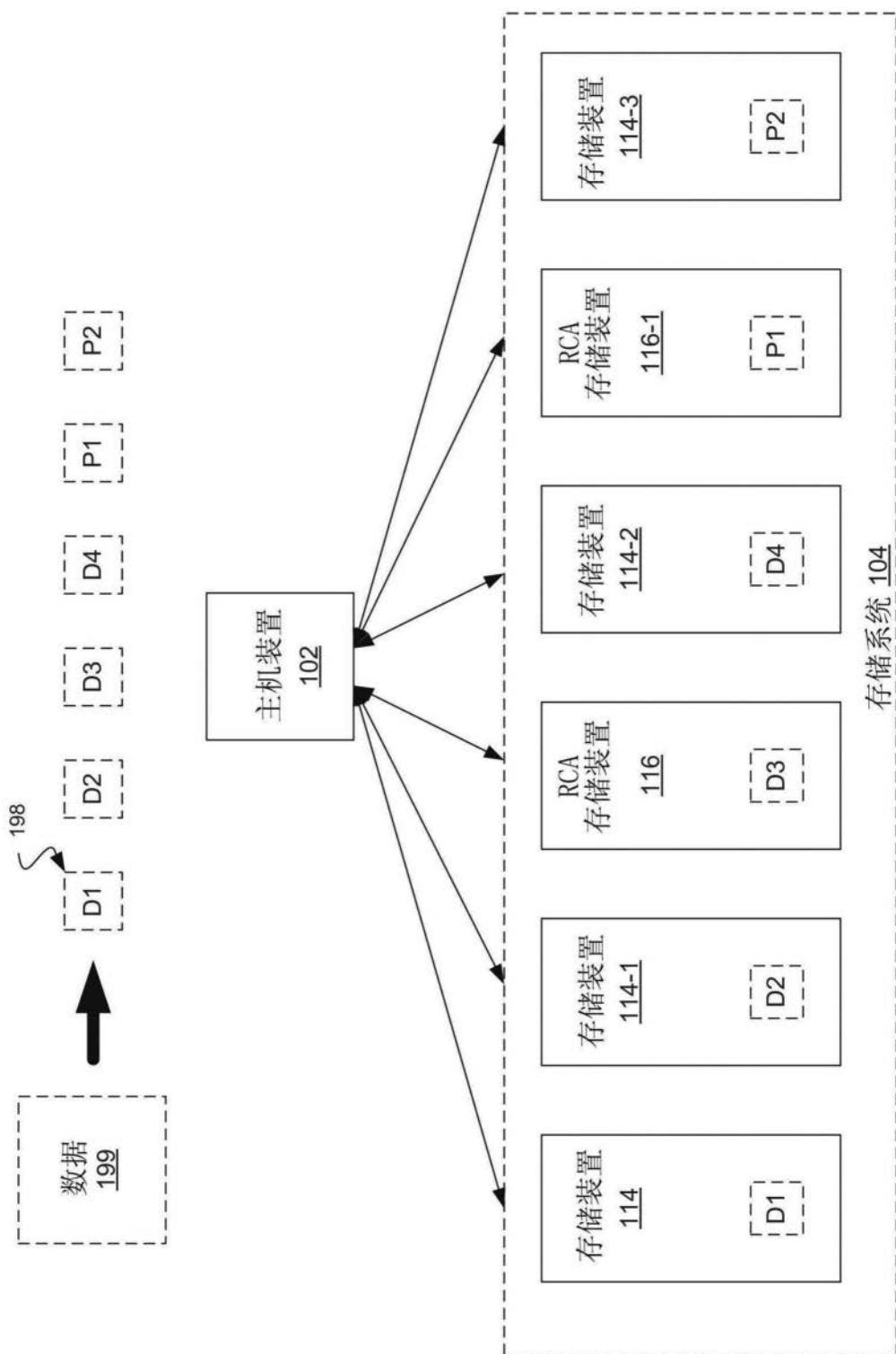


图1

201

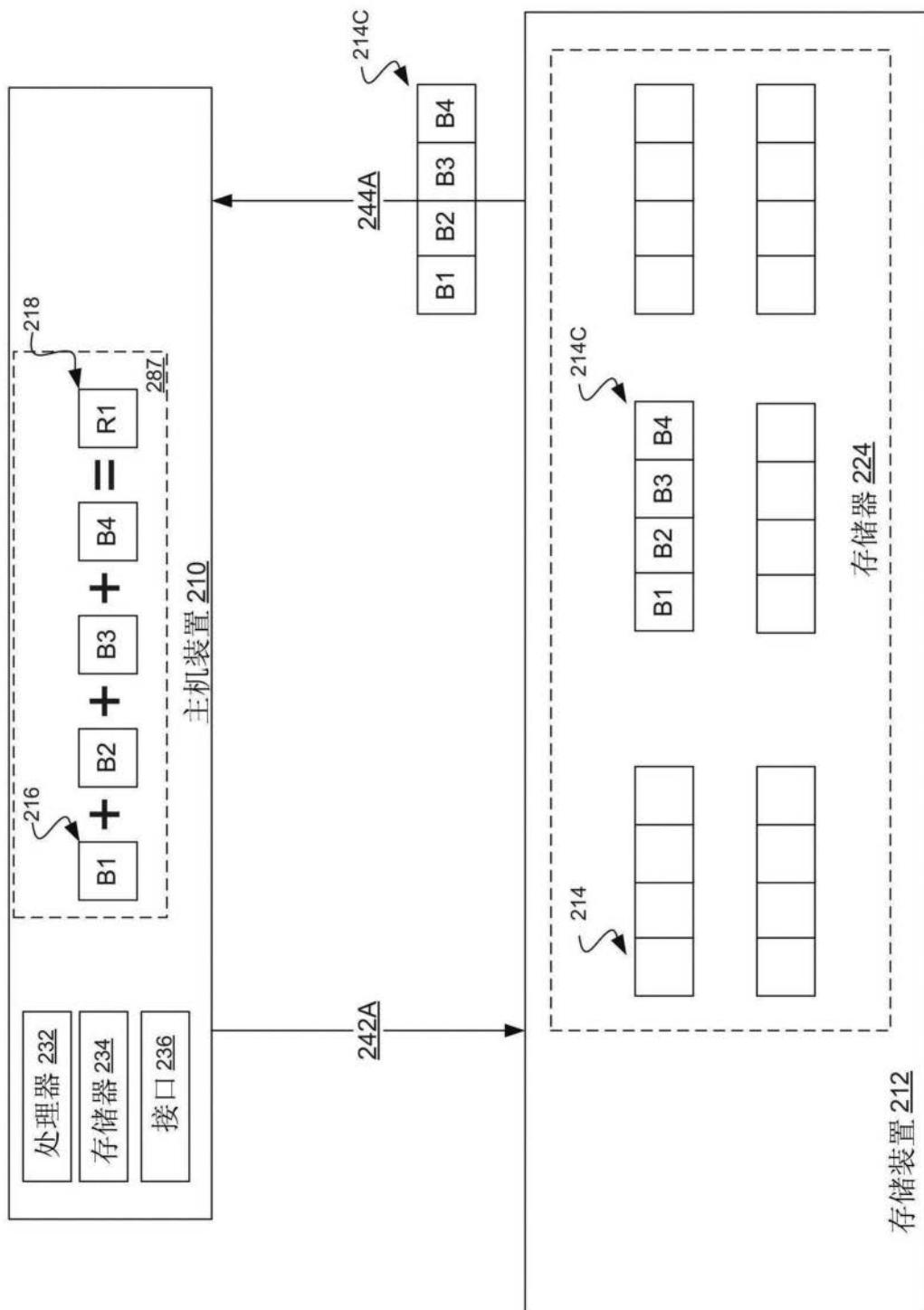


图2A

203

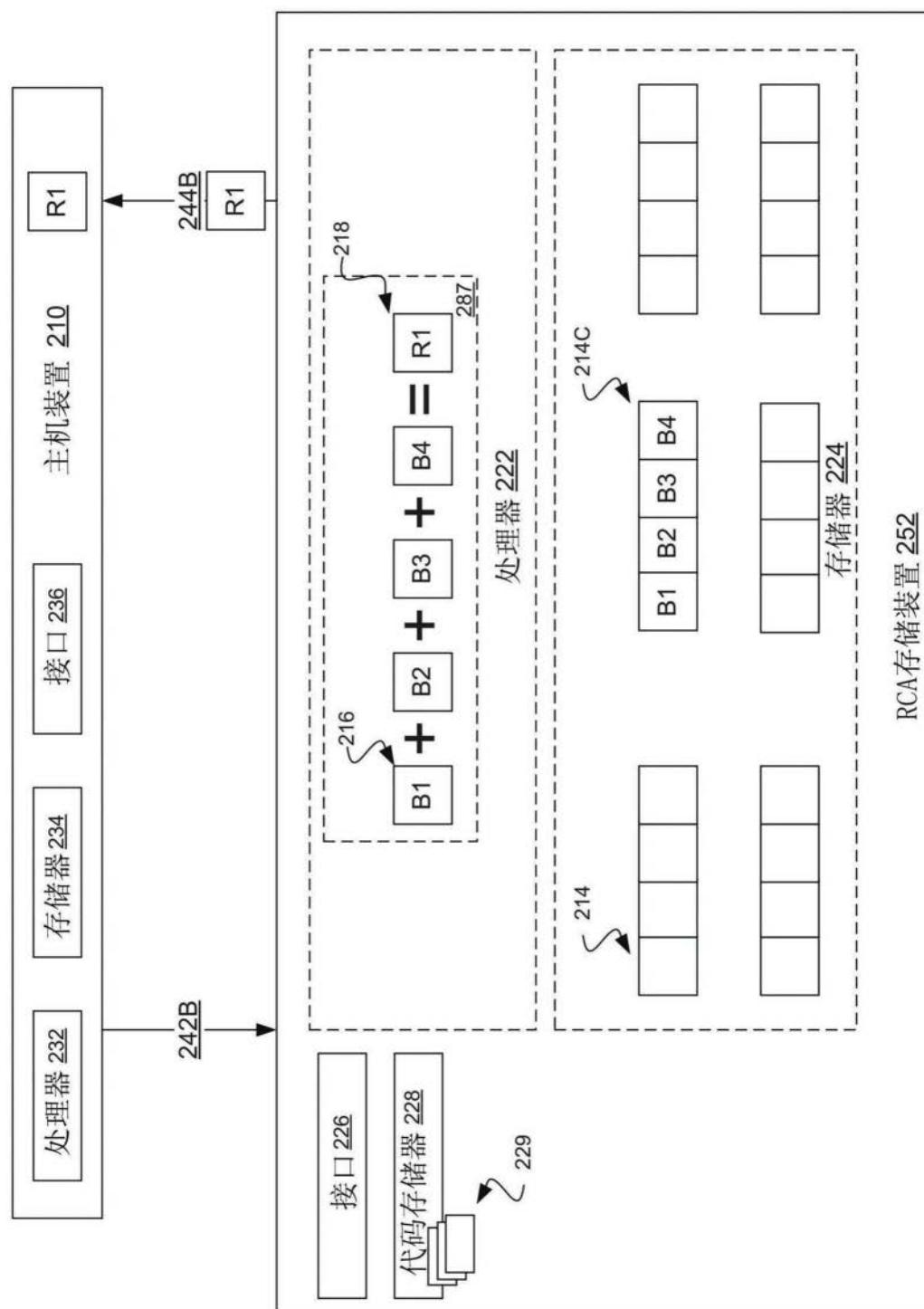


图2B

205

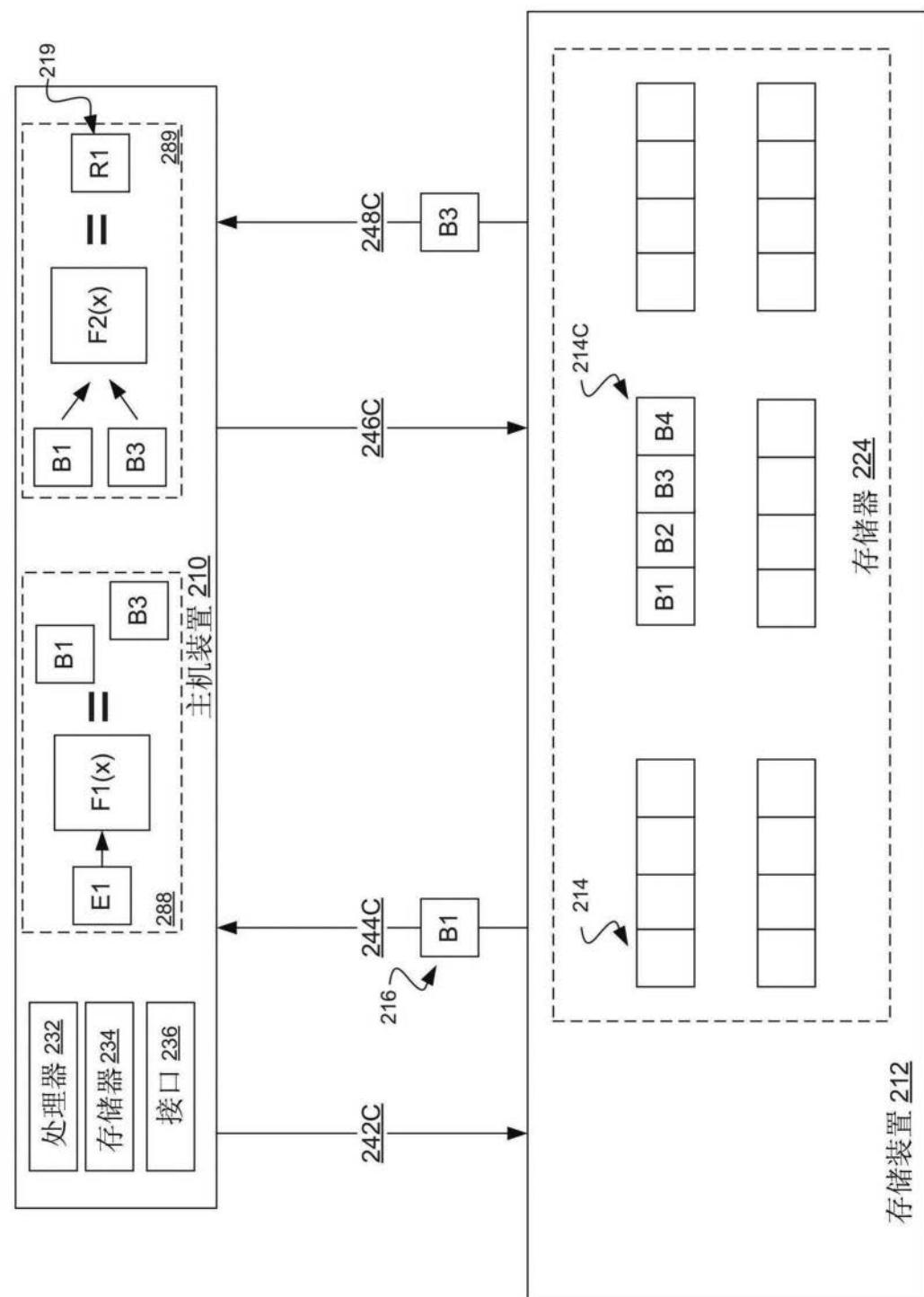


图2C

207

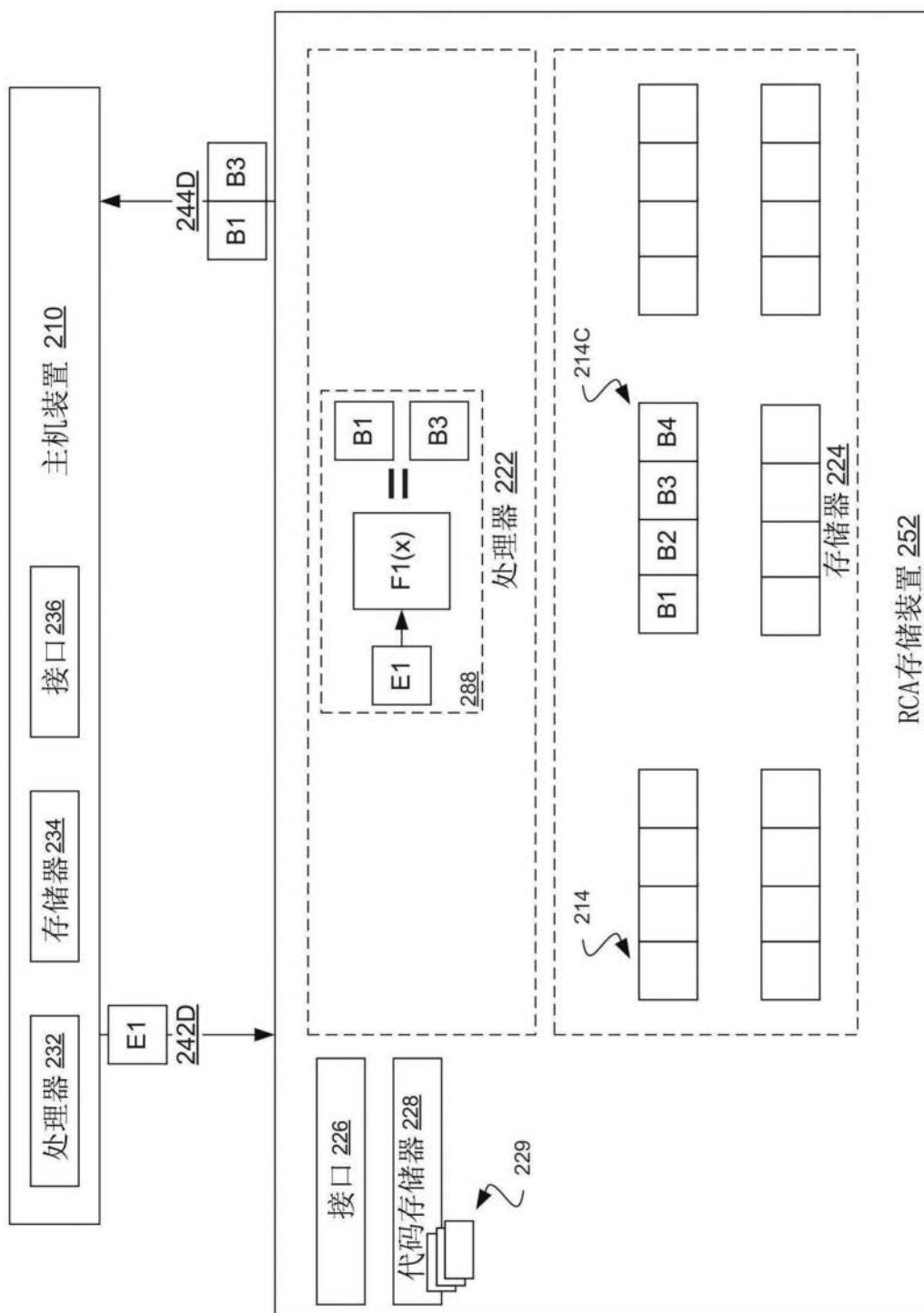


图2D

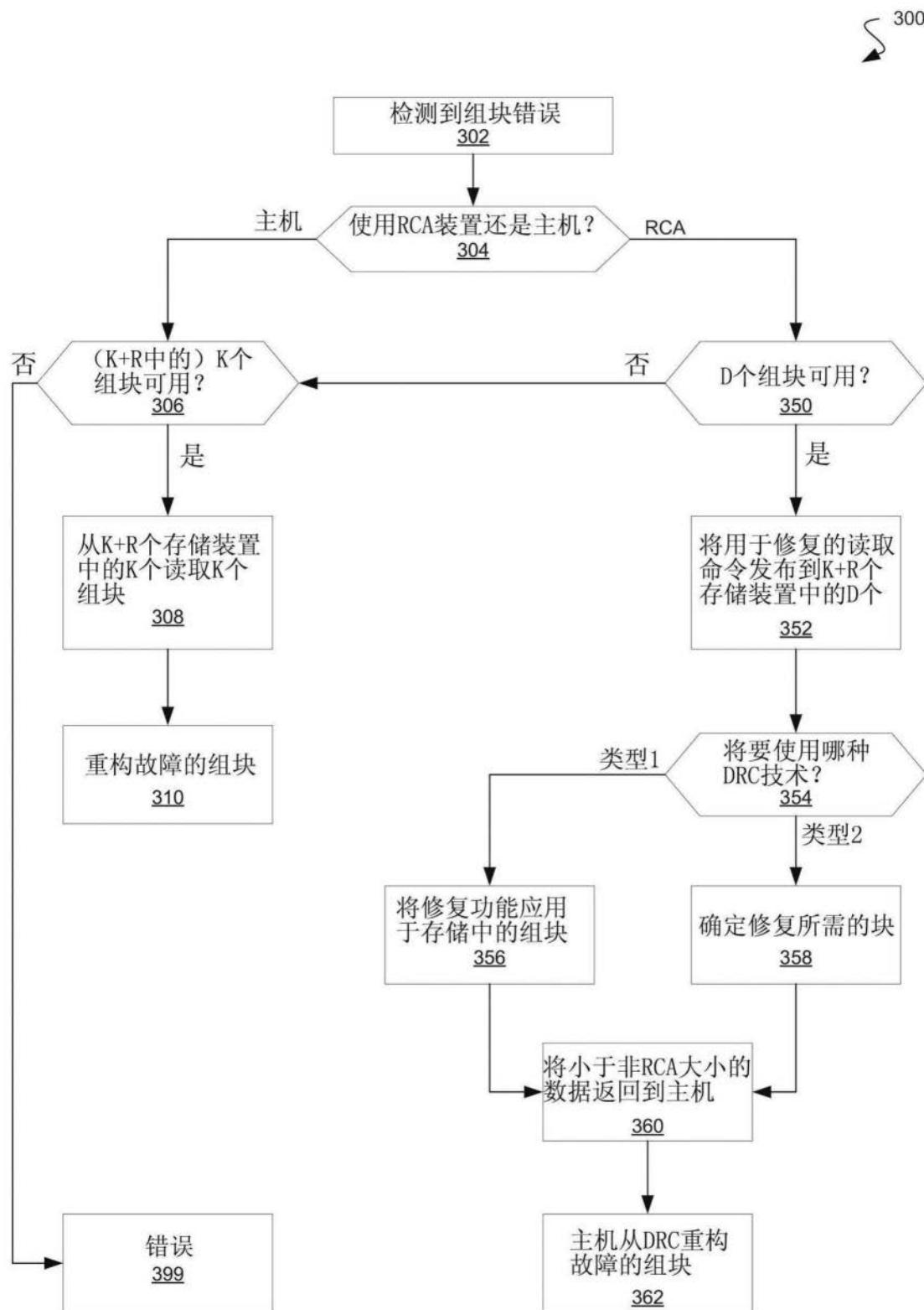


图3

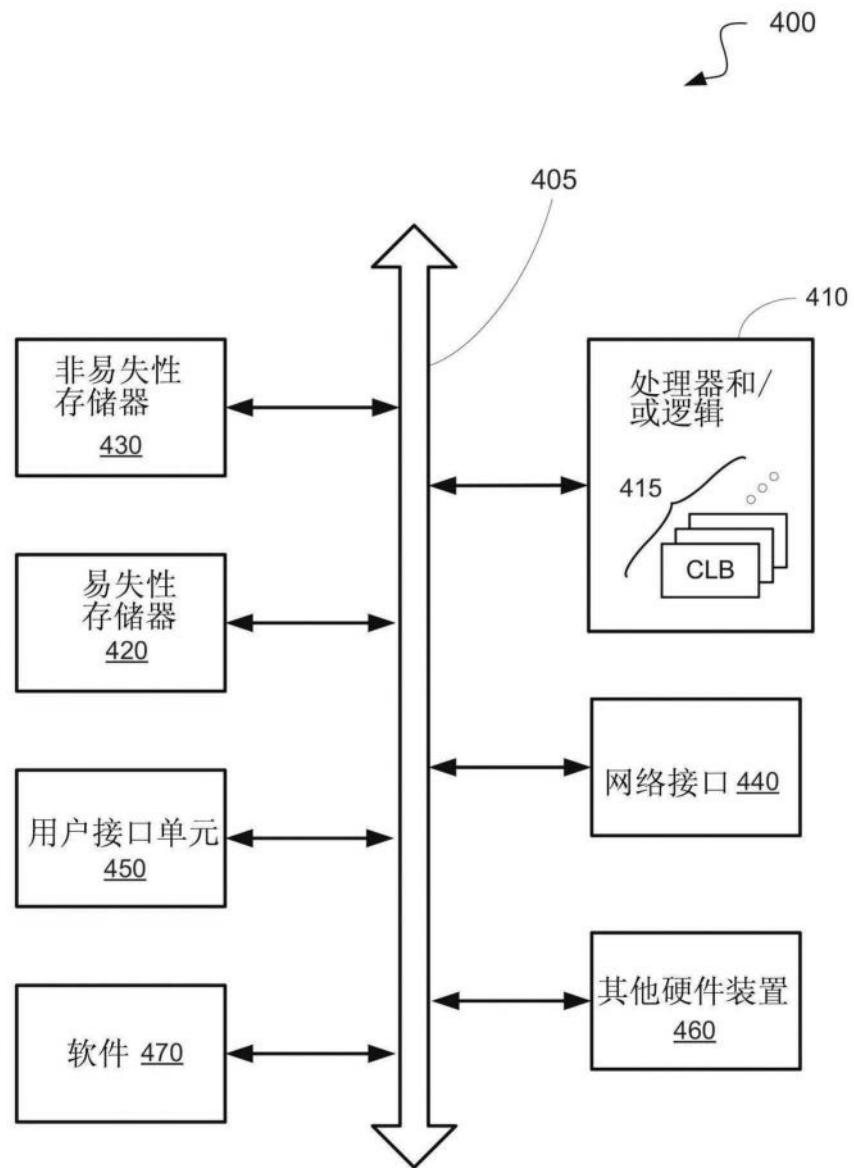


图4