(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106095700 B
(45)授权公告日 2020.04.21

(21)申请号 201610200533.3

(22)申请日 2016.03.31

(65)同一申请的已公布的文献号
申请公布号 CN 106095700 A

(43)申请公布日 2016.11.09

(30)优先权数据
14/699,594 2015.04.29 US

(73)专利权人 EMC公司
地址 美国马萨诸塞州

(72)发明人 J·S·邦威克

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 王艳春

(51)Int.Cl.
G06F 12/16(2006.01)
G06F 11/10(2006.01)

(56)对比文件
US 2014156966 A1,2014.06.05,
CN 102667738 A,2012.09.12,
US 8861123 B1,2014.10.14,
CN 104272274 A,2015.01.07,
CN 103339609 A,2013.10.02,
CN 104272261 A,2015.01.07,
US 7228352 B1,2007.06.05,
CN 102157202 A,2011.08.17,
CN 102681789 A,2012.09.19,
WO 2014113175 A1,2014.07.24,

审查员 赵识谦

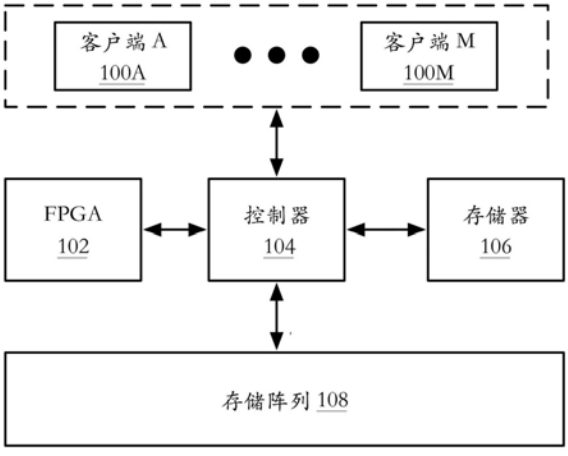
权利要求书3页 说明书14页 附图9页

(54)发明名称

在存储系统中复制和使用网格层级元数据的方法和系统

(57)摘要

一般地,本技术的实施例涉及一种保护持久性储存器中的数据的方法和系统。更具体地,本技术的各种实施例涉及使用不同的复制方案来保护持久性储存器的不同类型的数据。



CN 106095700 B

1.一种用于存储数据的方法,包括:

(a)产生第一多个标签,所述第一多个标签中的每一个包括用于RAID网格中的第一区的第一P/E计数和用于所述第一区的第一坏位置信息;

(b)擦除第一组RAID网格位置中的每一个,其中,所述第一组RAID网格位置中的每一个包括与所述第一区的第一侧面相关联的块;

(c)将所述第一多个标签中的一个写入到所述第一组RAID网格位置中的每一个;

(d)擦除第二组RAID网格位置中的每一个,其中,所述第二组RAID网格位置中的每一个包括与所述第一区的第二侧面相关联的块,其中,(c)是在(d)之前执行的;

(e)产生第二多个标签,所述第二多个标签中的每一个包括用于所述RAID网格中的所述第一区的第二P/E计数和用于所述第一区的第二坏位置信息,其中所述第二多个标签中的内容至少部分地基于(b)和(d);

(f)将所述第二多个标签中的一个写入到所述第二组RAID网格位置中的每一个;以及

(g)将用户数据写入到所述第一组RAID网格位置中的每一个和所述第二组RAID网格位置中的每一个。

2.根据权利要求1所述的方法,其还包括:

(h)产生第三多个标签,所述第三多个标签中的每一个包括用于所述RAID网格中的第二区的第三P/E计数和用于所述第二区的第三坏位置信息;

(i)擦除第三组RAID网格位置中的每一个,其中,所述第三组RAID网格位置中的每一个包括与所述第二区的第一侧面相关联的块;

(j)将所述第三多个标签中的一个写入到所述第三组RAID网格位置中的每一个;

(k)擦除第四组RAID网格位置中的每一个,其中,所述第四组RAID网格位置中的每一个包括与所述第二区的第二侧面相关联的块;

(l)产生第四多个标签,所述第四多个标签中的每一个包括用于所述RAID网格中的所述第二区的第四P/E计数和用于所述第二区的第四坏位置信息;

(m)将所述第四多个标签中的一个写入到所述第四组RAID网格位置中的每一个;以及

(n)将用户数据写入到所述第三组RAID网格位置中的每一个和所述第四组RAID网格位置中的每一个。

3.根据权利要求2所述的方法,其中,所述第一多个标签中的每一个还包括时间戳,并且其中,所述第三多个标签中的每一个还包括所述时间戳。

4.根据权利要求1所述的方法,其中,所述用户数据包括选自由块层级元数据、客户端数据以及使用客户端数据生成的奇偶值组成的组中的至少一个。

5.根据权利要求4所述的方法,其中,所述奇偶值包括选自由P奇偶值、Q奇偶值以及交集奇偶值组成的组中的至少一个。

6.根据权利要求4所述的方法,其中,用RAID方案来保护所述RAID网格中的所述用户数据,其中,用复制方案来保护所述RAID网格中的所述第一多个标签。

7.根据权利要求1所述的方法,其中,用于所述第一区的所述第二坏位置信息包括选自由坏存储模块信息、坏块信息以及坏页信息组成的组中的至少一个。

8.根据权利要求1所述的方法,其中,所述第二坏位置信息不同于所述第一坏位置信息。

9.根据权利要求1所述的方法,其中,在(g)之前执行(c)和(f)。

10.根据权利要求1所述的方法,其中,所述第一多个标签中的每一个包括用于所述RAID网格的网格几何结构,并且其中,所述网格几何结构包括所述RAID网格中的奇偶位置。

11.根据权利要求10所述的方法,其中,所述第二多个标签中的每一个包括用于所述RAID网格的网格几何结构。

12.一种包括计算机可读程序代码的非临时计算机可读介质,所述计算机可读程序代码在被计算机处理器执行时使得计算机处理器能够:

(a)产生第一多个标签,所述第一多个标签中的每一个包括用于RAID网格中的第一区的第一P/E计数和用于所述第一区的第一坏位置信息;

(b)擦除第一组RAID网格位置中的每一个,其中,所述第一组RAID网格位置中的每一个包括与所述第一区的第一侧相关联的块;

(c)将所述第一多个标签中的一个写入到所述第一组RAID网格位置中的每一个;

(d)擦除第二组RAID网格位置中的每一个,其中,所述第二组RAID网格位置中的每一个包括与所述第一区的第二侧面相关联的块,其中,(c)是在(d)之前执行的;

(e)产生第二多个标签,所述第二多个标签中的每一个包括用于所述RAID网格中的所述第一区的第二P/E计数和用于所述第一区的第二坏位置信息,其中所述第二多个标签中的内容至少部分地基于(b)和(d);

(f)将所述第二多个标签中的一个写入到所述第二组RAID网格位置中的每一个;以及

(g)将用户数据写入到所述第一组RAID网格位置中的每一个和所述第二组RAID网格位置中的每一个。

13.根据权利要求12所述的非临时计算机可读介质,还包括计算机可读程序代码,所述计算机可读程序代码在被计算机处理器执行时使得计算机处理器能够:

(h)产生第三多个标签,所述第三多个标签中的每一个包括用于所述RAID网格中的第二区的第三P/E计数和用于所述第二区的第三坏位置信息;

(i)擦除第三组RAID网格位置中的每一个,其中,所述第三组RAID网格位置中的每一个包括与所述第二区的第一侧相关联的块;

(j)将所述第三多个标签中的一个写入到所述第三组RAID网格位置中的每一个;

(k)擦除第四组RAID网格位置中的每一个,其中,所述第四组RAID网格位置中的每一个包括与所述第二区的第二侧面相关联的块;

(l)产生第四多个标签,所述第四多个标签中的每一个包括用于所述RAID网格中的所述第二区的第四P/E计数和用于所述第二区的第四坏位置信息;

(m)将所述第四多个标签中的一个写入到所述第四组RAID网格位置中的每一个;以及

(n)将第二用户数据写入到所述第三组RAID网格位置中的每一个和所述第四组RAID网格位置中的每一个。

14.根据权利要求13所述的非临时计算机可读介质,其中,所述第一多个标签中的每一个还包括时间戳,并且其中,所述第三多个标签中的每一个还包括所述时间戳。

15.根据权利要求12所述的非临时计算机可读介质,其中,所述用户数据包括选自由块层级元数据、客户端数据以及使用客户端数据生成的奇偶值组成的组的至少一个,并且其中,所述奇偶值包括选自由P奇偶值、Q奇偶值、以及交集奇偶值组成的组中的至少一个。

16.根据权利要求12的非临时计算机可读介质,其中,至少部分地使用(b)和(d)的结果来确定所述第二坏位置信息,并且其中,用于所述第一区的所述第二坏位置信息包括选自由坏存储模块信息、坏块信息、以及坏页信息组成的组中的至少一个。

17.根据权利要求12所述的非临时计算机可读介质,其中,在(g)之前执行(c)和(f)。

18.根据权利要求12所述的非临时计算机可读介质,其中,所述第一多个标签中的每一个包括用于所述RAID网格的网格几何结构,其中,所述第二多个标签中的每一个包括用于所述RAID网格的网格几何结构,并且其中,所述网格几何结构包括所述RAID网格中的奇偶位置。

19.一种用于存储数据的系统,包括:

控制器;

非临时介质,其被可操作地耦合到所述控制器;

持久性储存器,其被可操作地连接到所述控制器并包括多个存储模块,其中,所述多个存储模块中的每一个包括固态存储器;

其中,所述非临时计算机可读介质包括指令,所述指令在被控制器执行时执行一种方法,该方法包括:

(a)产生第一多个标签,所述第一多个标签中的每一个包括用于RAID网格中的第一区的第一P/E计数和用于所述第一区的第一坏位置信息;

(b)擦除第一组RAID网格位置中的每一个,其中,所述第一组RAID网格位置中的每一个包括与所述第一区的第一侧相关联的块,其中,所述块中的每一个位于所述多个存储模块中的一个上;

(c)将所述第一多个标签中的一个写入到的所述第一组RAID网格位置中的每一个;

(d)擦除第二组RAID网格位置中的每一个,其中,所述第二组RAID网格位置中的每一个包括与所述第一区的第二侧面相关联的块,其中,(c)是在(d)之前执行的;

(e)产生第二多个标签,所述第二多个标签中的每一个包括用于所述RAID网格中的所述第一区的第二P/E计数和用于所述第一区的第二坏位置信息,其中所述第二多个标签中的内容至少部分地基于(b)和(d);

(f)将所述第二多个标签中的一个写入到所述第二组RAID网格位置中的每一个;以及

(g)将用户数据写入到所述第一组RAID网格位置中的每一个和所述第二组RAID网格位置中的每一个。

在存储系统中复制和使用网格层级元数据的方法和系统

背景技术

[0001] 为了防止存储系统中的潜在数据丢失,实现复制方案常常是有利的。

附图说明

[0002] 图1示出了根据本技术的一个实施例的系统。

[0003] 图2示出了根据本技术的一个实施例的RAID网格层。

[0004] 图3示出了根据本技术的一个实施例的RAID立方体和RAID立方体的各种视图。

[0005] 图4A—4D示出了根据本技术的一个或多个实施例的RAID立方体的示例。

[0006] 图5A—5C示出了根据本技术的一个或多个实施例的块。

[0007] 图6示出了根据本技术的一个或多个实施例的示例。

[0008] 图7示出了根据本技术的一个实施例的数据结构。

[0009] 图8—9示出了根据本技术的一个实施例的流程图。

具体实施方式

[0010] 现在将参考附图来详细地描述本技术的特定实施例。在本技术的实施例的以下详细描述中,阐述了许多特定细节以便提供本技术的更透彻理解。然而,对于本领域的技术人员而言将显而易见的是可在没有这些特定细节的情况下实施本技术。在其它情况下,并未详细地描述众所周知的特征以避免不必要地使本描述变得复杂。

[0011] 在图1—9的以下描述中,在本技术的各种实施例中相对于图所述的任何部件可等价于相对于任何其它图所述的一个或多个相同名称部件。为了简便起见,将不会相对于每个图重复这些部件的描述。因此,每个图的部件的每个实施例是通过引用而结合的,并且被假设为可选地存在于具有一个或多个相同命名的部件的每个其它图内。另外,根据本技术的各种实施例,应将图的部件的任何描述解释为相对于任何其它图中的相应相同命名部件所描述的实施例额外添加、与之相结合或作为其替代可实现的可选实施例。

[0012] 一般地,本技术的实施例涉及一种保护持久性储存器中的数据的方法和系统。更具体地,本技术的实施例涉及使用不同的复制方案来保护持久性储存器中的不同类型的数据。在本技术的一个实施例中,使用多维RAID方案(例如,2D RAID方案、3D RAID方案等)来保护用户数据(例如,客户端数据、块层级元数据以及奇偶数据),并且使用复制方案(例如,标签)来保护网格层级元数据(例如,网格几何结构、坏位置信息以及P/E计数)。

[0013] 在本技术的一个实施例中,使用2D RAID方案,当在给定RAID条带中存在超过两个错误时,可恢复实现此类RAID方案的RAID网格内存储的用户数据。相似地,使用3D RAID方案,当在给定RAID条带中存在超过两个错误时,可恢复实现此类RAID方案的RAID立方体内存储的用户数据。此外,在本技术的各种实施例中,当在超过一个独立故障域(IFD)中存在故障时可恢复所有用户数据。在本技术的一个实施例中,要求网格层级元数据以实现多维RAID方案。更具体地,网格层级元数据可包括但不限于关于网格尺寸、在多维RAID方案中使用的奇偶值的数目以及奇偶值的位置的信息。可要求上述信息以实现多维RAID方案。

[0014] 在本技术的一个或多个实施例中，IFD对应于故障模式，该故障模式导致给定位置处的数据不可访问。每个IFD对应于存储阵列中的故障的独立模式。例如，如果数据被存储在NAND闪存中，其中NAND闪存是存储模块(其在某些实施例中也可称为闪存模块)(其包括多个NAND管芯)的一部分，则IFD可以是(i)存储模块、(ii)通道(即，被存储模块中的闪存控制器(未示出)用来向NAND闪存写入数据的通道)、以及(iii)NAND管芯。

[0015] 出于本技术的目的，如本文所使用的术语“RAID”指代“独立磁盘冗余阵列”。虽然“RAID”指代独立磁盘的任何阵列，但可使用可基于本技术的实施方式将RAID网格位置跨一个或多个持久性存储设备分布的任何类型的持久性存储设备来实现本技术的实施方式。

[0016] 图1示出了根据本技术的一个实施例的系统。如图1中所示，该系统包括一个或多个客户端100A和100M、控制器104、存储器106、FPGA 102(其可以可选地存在)以及存储阵列108。

[0017] 在本技术的一个实施例中，客户端100A、100M是在包括用以向控制器104发布读请求或写请求的功能的系统上执行的任何系统或过程。在本技术的一个实施例中，客户端100A、100M的每个可以包括处理器(未示出)、存储器(未示出)以及持久性储存器(未示出)。在本技术的一个实施例中，控制器104被配置成实现图8—9中所示的方法。此外，控制器包括用以根据多维RAID方案来存储用户数据(参见例如图5A，501)的功能，其包括以与多维RAID方案一致的方式向存储阵列写入数据(参见例如图2—4D)和以与多维RAID方案一致的方式从存储阵列读取数据(包括重构数据)(参见例如图2—4D)。在本技术的一个实施例中，控制器104包括被配置成执行用以实现本技术的一个或多个实施例的指令的处理器，其中，该指令被存储在位于控制器104内或被可操作地连接到控制器104的非临时计算机可读介质(未示出)上。可替换地，可使用硬件来实现控制器104。本领域的技术人员将认识到可使用软件和/或硬件的任何组合来实现控制器104。

[0018] 在本技术的一个实施例中，控制器104被可操作地连接到存储器106。存储器106可以是任何易失性存储器或非易失性存储器，包括但不限于动态随机存取存储器(DRAM)、同步DRAM、SDR SDRAM以及DDR SDRAM。在本技术的一个实施例中，存储器106被配置成在临时地存储各种数据之后(包括奇偶数据)将此类数据存储在存储阵列中(参见例如图7中所述的数据)。

[0019] 在本技术的一个实施例中，FPGA 102(如果存在的话)包括出于将数据存储在存储阵列108中的目的而计算P和/或Q奇偶值的功能和/或执行恢复使用多维RAID方案存储的已毁坏或漏失数据所需的各种计算的功能。在本技术的一个实施例中，FPGA可包括用以执行图8和9中所述的方法的全部或一部分的功能。根据本技术的一个或多个实施例，控制器104可使用FPGA102来卸载各种数据的处理。

[0020] 在本技术的一个实施例中，存储阵列108包括许多单独的持久性存储设备，包括但不限于：磁性存储器设备、光学存储器设备、固态存储器设备、相变存储器设备、任何其它适当类型的持久性存储器设备或其任何组合。在本技术的一个实施例中，每个存储阵列108可包括许多存储模块，其中，每个存储模块包括固态存储器和存储模块控制器。在此类实施例中，存储模块控制器包括用以从控制器接收页并将该页写入到固态存储器中的响应物理位置的功能。此外，存储模块控制器可包括用以在页被写入到固态存储器之前生成用于每个页的纠错码(ECC)。另外，存储模块控制器可包括用以根据多维RAID方案来重构页的功能。

[0021] 本领域的技术人员将认识到虽然图1示出了FPGA，但可在没有FPGA的情况下实现本技术。此外，本领域的技术人员将认识到在不脱离本技术的情况下可使用其它部件代替FPGA。例如，可使用ASIC、图形处理单元(GPU)、通用处理器、能够出于将数据存储在存储阵列中和/或执行恢复使用多维RAID方案存储的已毁坏数据所需的各种计算的目的计算P和/或Q奇偶值的任何其它硬件设备、包括被配置成出于将数据存储在存储阵列108中和/或执行恢复使用多维RAID方案存储的已毁坏数据所需的各种计算的目的计算P和/或Q奇偶值的硬件、固件和/或软件的组合的任何设备或其任何组合来实现本技术。

[0022] 在本技术的一个实施例中，如果控制器实现2D RAID方案或3D RAID方案(参见图3)，则控制器以RAID网格来存储数据，其中，RAID网格包括一组RAID网格层(参见例如图2，200)。RAID网格包括一组RAID网格位置，其中，每个RAID网格位置与块相关联。此外，每个块与一组页相关联。

[0023] 例如，考虑其中存在4×4RAID网格且RAID网格中的每个RAID网格位置与包括256个页的块相关联的情形。在这种情形中，RAID网格可由最多达255个RAID网格层构成(参见例如图2，200)，其中，每个RAID网格层包括16个页(即，来自与RAID网格相关联的每个块一个页)。16个块(即，与RAID网格相关联的块)中的每一个中的剩余页被用来存储标签(参见例如图5，508)。

[0024] 继续本示例，如果RAID网格与第一维度上的第一IFD＝存储模块(SM)和第二维度上的第二IFD＝通道(CH)相关联，则可如下表示用于给定RAID网格层中的16个页中的每一个的物理地址：

[0025] 表1：RAID网格层中的物理地址

[0026]

RAID 网格层中的 RAID 网格层位置	存储池中的物理地址
0	<SM0, CH0, CE, LUN, 平面, 块, 页>
1	<SM1, CH0, CE, LUN, 平面, 块, 页>
2	<SM2, CH0, CE, LUN, 平面, 块, 页>
3	<SM3, CH0, CE, LUN, 平面, 块, 页>
4	<SM0, CH1, CE, LUN, 平面, 块, 页>
5	<SM1, CH1, CE, LUN, 平面, 块, 页>
6	<SM2, CH1, CE, LUN, 平面, 块, 页>
7	<SM3, CH1, CE, LUN, 平面, 块, 页>
8	<SM0, CH2, CE, LUN, 平面, 块, 页>
9	<SM1, CH2, CE, LUN, 平面, 块, 页>
10	<SM2, CH2, CE, LUN, 平面, 块, 页>
11	<SM3, CH2, CE, LUN, 平面, 块, 页>
12	<SM0, CH3, CE, LUN, 平面, 块, 页>
13	<SM1, CH3, CE, LUN, 平面, 块, 页>
14	<SM2, CH3, CE, LUN, 平面, 块, 页>
15	<SM3, CH3, CE, LUN, 平面, 块, 页>

[0027] 如上表中所示，芯片启用(CE)、逻辑单元(LUN)、平面、块以及页对于给定RAID网格层中的每页而言是相同的。此外，在给定RAID网格的不同RAID网格层内，芯片启用(CE)、逻辑单元(LUN)、平面以及块保持恒定，而页号改变。例如，RAID网格中的第一RAID网格层中的页可对应于包括＜块37，页1＞的物理地址的页，而同一RAID网格中的第二RAID网格层中的页可包括包含＜块37，页2＞的物理地址。换言之，在本技术的一个实施例中，针对RAID网格中的所有页在物理地址中指定的块是相同的，而针对RAID网格内的页在物理地址中指定的页对于给定RAID网格层中的所有页而言是相同的，但是对于与其它RAID网格层相关联的页而言是不同的。

[0028] 图2示出了根据本技术的一个实施例的RAID网格层。更具体地，图2示出了根据本技术的一个或多个实施例的RAID网格层的概念部分。RAID网格层200包括许多RAID网格层位置，其中，每个RAID网格层位置最终对应于与物理地址相关联的存储阵列中的页。

[0029] 关于RAID网格层的结构，每个RAID网格层200包括：(i)数据网格202，其包括存储从客户端接收到的客户端数据(即，客户端已命令控制器写入到存储阵列的数据)的RAID网格层位置；(ii)行P奇偶组204，其包括在其中存储使用一行中的RAID网格层位置上的数据计算的P奇偶值的RAID网格层位置(下面描述)；(iii)行Q奇偶组206，其包括在其中存储使用一行中的RAID网格层位置上的数据计算的Q奇偶值的RAID网格层位置(下面描述)；(iv)列P奇偶组208，其包括在其中存储使用一列中的RAID网格层位置上的数据计算的P奇偶值

的RAID网格层位置(下面描述);(v)列Q奇偶组210,其包括在其中存储使用一列中的RAID网格层位置上的数据计算的Q奇偶值的RAID网格层位置(下面描述);以及(vi)交集奇偶组212,其包括使用(a)来自行P奇偶组204中的RAID网格层位置的数据、(b)来自行Q奇偶组206中的RAID网格层位置的数据、(c)来自列P奇偶组208中的RAID网格层位置的数据、以及(d)来自列Q奇偶组210中的RAID网格层位置的数据(下面描述)计算的奇偶值。

[0030] 参考行214,在本技术的一个实施例中,通过对包括数据(例如,$P_{r2}=f_P(D_1,D_2,D_3,D_4)$的行214中的所有RAID网格层位置应用P奇偶函数来计算存储在行214中的表示为P_{r2}的RAID网格层位置上的奇偶值。同样地,在本技术的一个实施例中,通过对包括数据(例如,$Q_{r2}=f_Q(D_1,D_2,D_3,D_4)$的行214中的所有RAID网格层位置应用Q奇偶函数来计算存储在行214中的表示为Q_{r2}的RAID网格层位置上的奇偶值。

[0031] 参考列216,在本技术的一个实施例中,通过对包括数据(例如,$P_{C6}=f_P(D_5,D_2,D_6,D_7)$的列216中的所有RAID网格层位置应用P奇偶函数来计算存储在列216中的表示为P_{c6}的RAID网格层位置上的奇偶值。同样地,在本技术的一个实施例中,通过对包括数据(例如,$Q_{C6}=f_Q(D_5,D_2,D_6,D_7)$的列216中的所有RAID网格层位置应用Q奇偶函数来计算存储在列216中的表示为Q_{C6}的RAID网格层位置上的数据。

[0032] 参考交集奇偶组212,在本技术的一个实施例中,可通过对行P奇偶组204中的所有RAID网格层位置应用P奇偶函数或者对列P奇偶组208中的所有RAID网格层位置应用P奇偶函数来计算存储在表示为I_{r1}的RAID网格层位置上的数据。例如,$I_{r1}=f_P(P_{r1},P_{r2},P_{r3},P_{r4})$或$I_{r1}=f_P(P_{c5},P_{c6},P_{c7},P_{c8})$。

[0033] 在本技术的一个实施例中,可通过对行Q奇偶组204中的所有RAID网格层位置应用P奇偶函数或者对列P奇偶组208中的所有RAID网格层位置应用Q奇偶函数来计算存储在表示为Ir2的RAID网格层位置上的数据。例如,$I_{r2}=f_P(Q_{r1},Q_{r2},Q_{r3},Q_{r4})$或$I_{r2}=f_Q(P_{c5},P_{c6},P_{c7},P_{c8})$。

[0034] 在本技术的一个实施例中,可通过对列Q奇偶组210中的所有RAID网格层位置应用P奇偶函数或者对行P奇偶组204中的所有RAID网格层位置应用Q奇偶函数来计算存储在表示为I_{r3}的RAID网格层位置上的数据。例如,$I_{r3}=f_P(Q_{c5},Q_{c6},Q_{c7},Q_{c8})$或$I_{r3}=f_Q(P_{r1},P_{r2},P_{r3},P_{r4})$。

[0035] 在本技术的一个实施例中,可通过对列Q奇偶组210中的所有RAID网格层位置应用Q奇偶函数或者对行Q奇偶组206中的所有RAID网格层位置应用Q奇偶函数来计算存储在表示为I_{r4}的RAID网格层位置上的数据。例如,$I_{r4}=f_Q(Q_{r1},Q_{r2},Q_{r3},Q_{r4})$或$I_{r4}=f_Q(Q_{c5},Q_{c6},Q_{c7},Q_{c8})$。

[0036] 在本技术的一个实施例中,用来计算用于所有奇偶组的值的P和Q奇偶函数可对应于被用来实现RAID 6的任何P和Q奇偶函数。

[0037] 如上文所讨论的,图2中所示的RAID网格层200表示RAID网格层的概念布局。然而,当单独RAID网格层位置被写入到存储阵列时,各种RAID网格层位置的相对位置可跨行和/或列改变。例如,参考行214,当行214内的RAID网格层位置被写入到存储阵列时,包括用户数据(用“D”表示)的RAID网格层位置和包括奇偶数据的RAID网格层位置(即,表示为“P_r”和“Q_r”的RAID网格层位置)的相对位置可如下:<D_1,D_2,P_{r2},D_3Q_{r2},D_4>、<P_{r2},Q_{r2},D_1,D_2,D_3,D_4>或行内214的任何其它布置。同样地,参考列216,当行214内的RAID网格层位置被写入到存储

阵列时，包括用户数据(用“D”表示)的RAID网格层位置和包括奇偶数据的RAID网格层位置(即，表示为“P_c”和“Q_c”的RAID网格层位置)的相对位置可如下：<D_5,D_2,D_6,P_{c6},D_7,Q_{c6}>、<P_{c6},D_5,D_2,Q_{c6},D_6,D_7>或行列216的任何其它布置。

[0038] 在本技术的一个实施例中，(i)行P奇偶组，(ii)行Q奇偶组，(iii)列P奇偶组以及(iv)列Q奇偶组中的每一个的位置可基于本技术的实施方式而改变。此外，在此类实施例中，基于上述奇偶组的位置来确定交集奇偶组的位置。

[0039] 继续图2的讨论，控制器(或系统中的另一实体)可确定与每个RAID网格位置相关联的数据被写入到存储阵列中的哪个物理地址。此确定可在从客户端接收到用于特定RAID网格(或RAID网格层)的任何客户端数据(其是表示为“D”的用户数据的一部分)之前进行。可替换地，该确定可在将与RAID网格层位置相关联的任何数据写入到存储阵列之前进行。

[0040] 在本技术的一个实施例中，使用如图2中所述的2D RAID方案来保护构成RAID网格的每个RAID网格层的页内的用户数据(参见例如图5，501)(或其部分)。

[0041] 本领域的技术人员将认识到虽然图2示出了6×6的RAID网格层，但在不脱离本技术的情况下可使用任何其它尺寸来实现RAID网格层。此外，虽然图2仅示出了RAID网格的单个RAID网格层，但构成RAID网格的每个RAID网格层的RAID尺寸是相同的。例如，RAID网格可由全部具有相同尺寸的255个RAID网格层构成。此外，单独RAID网格层内的奇偶数据的位置可以跨RAID网格内的所有RAID网格层是相同的，或者可替换地，奇偶数据可在RAID网格内的不同RAID网格层中的不同位置上。

[0042] 在本技术的一个实施例中，P奇偶值是里德—所罗门特征群(syndrome)，并且同样地，P奇偶函数可对应于可以产生里德—所罗门特征群的任何函数。在本技术的一个实施例中，P奇偶函数是XOR函数。

[0043] 在本技术的一个实施例中，Q奇偶值是里德-所罗门特征群，并且同样地，Q奇偶函数可对应于可以产生里德-所罗门特征群的任何函数。在本技术的一个实施例中，Q奇偶值是里德—所罗门代码。在本技术的一个实施例中，$Q=g^0 \cdot D_0+g^1 \cdot D_1+g^2 \cdot D_2+\cdots+g^{n-1} \cdot D_{n-1}$，其中，Q对应于相对于图2定义的Q个奇偶值中的任何一个，g是字段的发生器，并且D的值对应于数据(其可包括来自数据网格的值和/或来自包括P或Q奇偶值的一个或多个行或列的值)。

[0044] 本领域的技术人员将认识到虽然图2中所示的RAID网格层包括用于每个行和列的P和Q个奇偶值，但在不脱离本技术的情况下可使用更多或更少的奇偶值来实现本技术的实施例。例如，每个行和列可仅包括P奇偶值。在另一实施例中，每个行和列可包括三个奇偶值。上述示例并不意图限制本技术的范围。在本技术的一个实施例中，无论在本技术的实施方式中使用的奇偶值的数目如何，每个奇偶值都是里德—所罗门特征群。

[0045] 图3示出了根据本技术的一个实施例的RAID立方体和RAID立方体的各种视图。如图3中所示，RAID立方体300对应于RAID网格302的概念堆栈。如上文所讨论的，控制器(或系统中的另一实体)选择用于每个RAID网格位置的数据(包括用户数据和标签)将存储到其中的存储阵列内的物理地址。在本技术的一个实施例中，可根据RAID网格(或RAID立方体)被设计成针对其进行保护的IFD来确定物理地址的选择。换言之，可以以将针对一个或多个IFD中的故障进行保护的方式来选择物理地址。例如，如图3中所示，与用于给定RAID网格302、304的每个RAID网格位置(例如，块中的数据，参见图5，500)相关联的数据被写入到使

用来自IFD 1和IFD 2的唯一一对值选择的存储阵列(未示出)中的一组物理地址(或者将被写入到设定物理地址),但是对于IFD 3而言具有相同的值。例如,如果存储阵列中的数据(即,用户数据和标签)被存储在NAND闪存中,其中,NAND闪存是存储模块(其包括多个NAND管芯)的一部分,则IFD可如下:(i)IFD 1=存储模块,(ii)IFD 2=通道,并且(iii)IFD 3=NAND管芯。因此,在给定RAID网格中,与每个RAID网格位置相关联的数据被写入到存储模块IFD 1和通道IFD 2的唯一组合,但是被写入到同一NAND管芯(在每个存储模块上)。本领域的技术人员将认识到本技术不限于上述三个独立故障域。此外,本领域的技术人员将认识到本技术不限于包括NAND闪存的存储阵列。

[0046] 继续图3,如上文所讨论的,RAID立方体300是RAID网格的概念堆栈。更具体地,在本技术的一个实施例中,RAID立方体300可包括(i)数据部分316,其包括两个或更多RAID网格304、306、308、310和奇偶部分318,其包括P奇偶RAID网格312和Q奇偶RAID网格314。

[0047] 在本技术的一个实施例中,数据部分316中的RAID网格304、306、308、310包括奇偶数据(例如,P奇偶或Q奇偶值),其允许仅使用RAID网格内的数据(包括奇偶数据)来恢复RAID网格内的数据。在本技术的一个实施例中,RAID立方体被布置成使得可使用来自其它RAID网格(即,数据部分316和奇偶部分318两者中的RAID网格)的数据(包括奇偶数据)来恢复与给定RAID网格304、306、308、310中的给定RAID网格位置相关联的数据。在本技术的一个实施例中,RAID立方体的奇偶部分318使得能够实现此类机制。

[0048] 在本技术的一个实施例中,P奇偶RAID网格312与底层RAID网格304、306、308、310是相同尺寸,其中,通过对来自数据部分316中的RAID网格中的块的数据(包括奇偶数据)应用P奇偶函数(例如,XOR函数)来计算存储在与P奇偶RAID网格内的每个RAID网格位置相关联的块中的数据(参见例如图4A—4D)。同样地,Q奇偶RAID网格314与底层RAID网格304、306、308、310是相同尺寸,其中,通过对来自数据部分316中的RAID网格的数据(包括奇偶数据)应用Q奇偶函数来计算存储在与Q奇偶RAID网格内的每个RAID网格位置相关联的块中的数据(参见例如图4A—4D)

[0049] 图4A—4D示出了根据本技术的一个或多个实施例的填充RAID立方体的示例。本示例并不意图限制本技术的范围。

[0050] 考虑在图4D中描绘的RAID立方体,其包括RAID网格A 400、RAID网格B402、RAID网格C 404、P奇偶RAID网格406以及Q奇偶RAID网格408。此外,RAID立方体中的每个RAID网格400、402、404、406、408包括跨IFD 1和IFD 2写入的RAID网格位置,但是具有IFD 3的恒定值。因此,在本技术的一个实施例中,可使用以下各项来恢复与RAID网格中的RAID网格位置(“目标RAID网格位置”)相关联的块中的数据(在一个或多个页上):(i)仅存储在与目标RAID网格位于其中的行或列中的RAID网格位置相关联的块中的数据;(ii)使用与目标RAID网格位置位于其中的RAID网格内的任何RAID网格位置相关联的块中的数据;或者(iii)使用与目标RAID网格位置位于其中的RAID立方体内的任何RAID网格位置相关联的块中的数据。换言之,在本技术的一个实施例中,RAID网格和/或RAID立方体内的数据和奇偶值的布置允许当在目标RAID网格位置位于其中的行和列中的每一个中存在超过两个错误时恢复与目标RAID网格位置相关联的块中的数据(在一个或多个页上)。

[0051] 参考图4A,图4A包括三个RAID网格400、402、404,去构成RAID立方体的数据部分。每个RAID网格400、402、404中的每个RAID网格位置包括定义RAID网格位置中的数据被写入

其中的存储阵列中的位置的3元组。在本示例中，3元组中的元素对应于如下IFD：<IFD1，IFD2，IFD3>。3元组图示出如何跨各种IFD选择存储阵列中的位置。特别地，RAID网格A中的每个RAID网格位置包括IFD1和IFD2的唯一组合，但是对于IFD 3而言具有相同的值。例如，如果IFD1是存储模块，IFD2是通道，并且IFD3是NAND管芯，则3元组<4,2,1>指示与特定RAID网格位置相关联的块中的数据(在一个或多个页上)将被使用通道2写入到位于存储模块4中的NAND管芯中的物理位置。同样地，3元组<2,3,1>指示与特定RAID网格位置相关联的块中的数据(在一个或多个页上)将被使用通道3写入到存储模块2中的NAND 1中的物理地址。

[0052] RAID网格B 402和RAID网格C 404被以与RAID网格A 400类似的方式布置。然而，用于针对RAID网格B 402中的RAID网格位置的3元组中的IFD 3的值不同于用于针对RAID网格A 400的RAID网格位置的3元组中的IFD3的值。此外，用于针对RAID网格C 404中的RAID网格位置的3元组中的IFD3的值不同于用于针对RAID网格A 400且针对RAID网格B 402的RAID网格位置的3元组中的IFD3的值。

[0053] 参考图4B，与P奇偶RAID网格406中的每个RAID网格位置相关联的块中的数据(在一个或多个页上)被以与RAID网格A 400、RAID网格B 402)以及RAID网格C 404类似的方式布置。此外，如上所述，使用RAID立方体(即，RAID网格A 400、RAID网格B 402、RAID网格C 404)中的每个数据网格中的一个RAID网格位置相关联的块中的数据(在一个或多个页上)来计算与P奇偶RAID网格406中的每个RAID网格相关联的块中的数据(在一个或多个页上)。例如，通过对与以下RAID网格位置中的一个相关联的块中的数据(在一个或多个页上)应用P奇偶函数(例如，XOR函数)来确定与P奇偶RAID网格406中的RAID网格位置<1,1,4>相关联的块中的数据(在一个或多个页上)：(i)与RAID网格A(400)<1,1,1>相关联的块中的数据(在一个或多个页上)，(ii)来自RAID网格B 402<1,1,2>的数据以及(iii)与RAID网格C 404<1,1,3>相关联的块中的数据(在一个或多个页上)。以类似方式计算与P奇偶RAID网格406中的其它RAID网格位置相关联的块中的数据(在一个或多个页上)。

[0054] 参考图4C，与Q奇偶RAID网格408中的每个RAID网格位置相关联的块中的数据(在一个或多个页上)被以与RAID网格A 400、RAID网格B 402以及RAID网格C 404类似的方式布置。此外，如上所述，使用RAID立方体(即，RAID网格A 400、RAID网格B 402、RAID网格C 404)中的每个数据网格中的一个RAID网格位置相关联的块中的数据(在一个或多个页上)来计算与Q奇偶RAID网格408中的每个RAID网格相关联的块中的数据(在一个或多个页上)。例如，通过对与以下RAID网格位置中的一个相关联的块中的数据(在一个或多个页上)应用Q奇偶函数(如上所述)来确定与Q奇偶RAID网格408中的RAID网格位置<1,1,5>相关联的块中的数据(在一个或多个页上)：(i)与RAID网格A 400<1,1,1>相关联的块中的数据(在一个或多个页上)，(ii)与RAID网格B 402<1,1,2>相关联的块中的数据(在一个或多个页上)，以及(iii)与RAID网格C 404<1,1,3>相关联的块中的数据(在一个或多个页上)。以类似方式计算与Q奇偶RAID网格408中的其它RAID网格位置相关联的块中的数据(在一个或多个页上)。

[0055] 图5A—5C示出了根据本技术的一个或多个实施例的块500。在本技术的一个实施例中，RAID网格(上文讨论)中的每个RAID位置被配置成存储块500，其中，该块包括一组页。参考图5A，块500至少包括包含用户数据501的一个或多个页和包含标签508的至少一个页。在本技术的一个实施例中，用户数据501可包括客户端数据502、块层级元数据506和奇偶数据504。用户数据501被存储在块500内的一个或多个页中。在本技术的一个实施例中，用户

数据502对应于从客户端接收到的任何数据。在本技术的一个实施例中，块层级元数据506包括用于存储在块500中的客户端数据502的元数据。在本技术的一个实施例中，块层级元数据对应于如在美国专利号8,370,567中描述的内容表条目，其被通过引用结合到本文中。继续图5A的讨论，奇偶数据504包括根据如上文在图2—4D中所述和在被通过引用结合到本文中的美国专利号8,327,185中所述的多维RAID方案生成的奇偶值。在本技术的一个实施例中，标签508对应于用于区(即，RAID网格的一部分)(参见例如图6)和/或用于块500位于其中的RAID网格的元数据。标签508可被存储在块500内的单独页中。换言之，每个块可包括其中存储在页中的唯一内容是标签(或其一部分)的至少一个页。在图5B、5C和6中描述了关于标签508的附加细节。

[0056] 图5B示出了根据本技术的一个或多个实施例的用于块500的标签508。如图5B中所示，标签508可包括：(i)存储模块ID 510、用于区512的P/E计数、网格几何结构514、时间戳516以及用于区(518)的坏位置信息。下面描述这些部件中的每一个。

[0057] 在本技术的一个实施例中，存储模块ID 510指定标签位于其上面的存储阵列108内的存储模块。更具体地，标签最初被存储在存储模块(未示出)上的块500中，存储模块ID对应于此存储模块。在本技术的一个实施例中，使用整数来表示存储模块ID 508字段。

[0058] 在本技术的一个实施例中，用于区512的程序/擦除(P/E)计数对应于在特定时间点的用于该区的P/E计数。P/E计数可表示：(i)已对块500内的页执行的P/E循环的数目或(ii)P/E循环范围(例如，5,000—9,999P/E循环)，其中，已在块内的页上执行的P/E循环的数目在P/E循环范围内。在本技术的一个实施例中，P/E循环是数据到擦除块中的一个或多个页的写入(即，用于擦除操作的最小可寻址单元，通常是一组的多个页)和该块的擦除，按照任一顺序。在本技术的一个实施例中，控制模块包括用以跟踪用于存储池中的每个块的P/E计数的功能。

[0059] 在本技术的一个实施例中，网格几何结构514指定关于RAID网格的几何结构的信息。在图5C中描述了关于网格几何结构514的附加细节。

[0060] 在本技术的一个实施例中，时间戳516对应于标签508被写入到存储阵列中的块500内的页的时间。时间戳的精度可基于本技术的实施方式而改变。此外，在不脱离本技术的情况下可使用序号来代替时间戳。

[0061] 在本技术的一个实施例中，用于区518的坏位置信息可包括：(i)坏存储模块信息；(ii)坏块信息；和/或(iii)坏页信息。坏存储模块信息指定哪些存储模块(被用来存储用于特定RAID网格的用户数据)可用于或不可用于执行写请求。坏块信息指定上述存储模块内的哪些块不应被用来存储任何数据(包括用户数据和标签)。如果不能成功地从块内的大多数(某个阈值数目)的页擦除、向其写入和/或从其检索数据，则可将给定块视为坏块。例如，如果不能从块中的75％的页擦除、向块中的75％的页写入和/或检索块中的75％的页，则可将特定块视为坏块。在本技术的另一实施例中，可在块内的数据不可检索或块不能成功地存储数据之前基于关于块(或块内的页)的其它信息的分析来将给定块视为坏块。例如，可使用诸如总P/E循环、原始位出错率等信息来主动地将给定块标记为坏块。

[0062] 在本技术的一个实施例中，坏页信息指定被用来存储用于RAID网格的数据的上述块内的哪些页不应被用来存储任何数据(包括用户数据和标签)。如果超过时间的阈值值(例如，90％)不能成功地从给定页擦除、向给定页写入和/或从给定页检索数据，则可将该

页视为坏页。

[0063] 可使用一个或多个位图来将上述坏位置信息编码。例如，可存在用于坏存储模块信息的位图、用于坏块信息的一组位图以及用于坏页信息的一组位图。该位图可使用“0”来表示坏存储模块、坏块或坏页，并且可使用“1”来表示所有其它存储模块、块以及页。在不脱离本技术的情况下可使用其它位图编码方案。

[0064] 参考图5C，网格几何结构(514)可包括：(i)网格尺寸字段520、(ii)网格位置522字段、以及(iii)一个或多个奇偶位置524、526。下面描述这些部件中的每一个。

[0065] 在本技术的一个实施例中，网格尺寸520信息可包括RAID网格中的行和列的数目和与每个RAID网格尺寸相关联的IFD。在本技术的一个实施例中，网格位置522字段可包括块500内的网格的位置。

[0066] 在本技术的一个实施例中，网格几何结构包括用于每个维度上的每种奇偶的一个奇偶位置524、526。例如，如果RAID网格在两个维度上都包括P和Q奇偶，则网格几何结构将包括P奇偶行组、Q奇偶行组、P奇偶列组以及Q奇偶列组的奇偶位置。在本技术的一个实施例中，基于每个维度(例如，行和列)对每个奇偶类型(例如，P、Q等)指定奇偶位置。本领域的技术人员将认识到在不脱离本技术的情况下可使用更多(例如，使用P、Q以及R奇偶)或更少的奇偶值(例如，使用P奇偶)来实现本技术的实施例。此外，本领域的技术人员将认识到在不脱离本技术的情况下每个维度可包括不同数目的奇偶值。

[0067] 虽然图5A—5C示出了存储在块中的各种类型的数据，但在不脱离本技术的情况下可用包括附加(或不同)数据和/或不同数据排序的块来实现本技术的实施例。此外，在不脱离本技术的情况下可使用任何编码方案对标签内的各种字段中的值进行编码。

[0068] 图6示出了根据本技术的一个或多个实施例的示例。更具体地，图6示出了用来保护存储在RAID网格内的块中的标签(参见图5，508)的复制方案的示例。关于图6，图6示出了包括48个RAID网格位置(C)的示例性6×8RAID网格600，其中，第一维度基于IFD1且第二维度基于IFD2。在本示例中，假设IFD1对应于存储模块且IFD2对应于通道。此外，假设针对标签指定的复制量(R)是16。因此，RAID网格600中的区的数目是3(即，在本示例中C/R或48/16)。此外，由于每个区具有两个侧面(表示为侧面A和侧面B)，所以区内的每个侧面中的RAID网格位置的数目是16/2＝8。

[0069] 如上文所讨论的，每个RAID网格位置与标签相关联。然而，标签的内容可跨与给定RAID网格相关联的标签而改变。更具体地，在本技术的一个实施例中，标签包括标签特定元数据(例如，存储模块ID)、侧面层级元数据(例如，时间戳)、区层级元数据(例如，用于区的P/E计数、用于区的坏位置信息)以及网格层级元数据(例如，网格几何结构)。因此，参考图6，在RAID网格内存在网格几何结构的48个拷贝且在给定区内存在时间戳的16个拷贝。此外，在给定区内，存在用于区的P/E计数的8—16个拷贝和用于区的坏位置信息的8—16个拷贝。相对于区的P/E计数和用于区的坏位置信息，此数据在区内被基于侧面更新(参见图8)。因此，在某些时间，给定区中的两个侧面具有用于区的P/E计数的相同内容和用于区的相同坏位置信息，而在其它时间，给定区内的两个侧面具有针对用于区的P/E计数的不同内容和针对用于区的坏位置信息的不同内容。下面所述的图8提供了在各种时间的关于标签内容的附加细节。

[0070] 本领域的技术人员将认识到给定区可包括任何量的复制(R)，条件是R≤C(即网格

位置的数目)，使得区内的每个侧面包括在R/2与R-1之间的RAID网格位置。

[0071] 在图6中示出并在上文描述的示例并不意图限制本技术的范围。

[0072] 图7示出了根据本技术的一个实施例中的系统内的各种部件之间的关系。在本技术的一个实施例中，控制器包括一个或多个数据结构以跟踪关于各种部件的信息和/或关于一个或多个部件之间的关系的信息。

[0073] 在本技术的一个实施例中，每个RAID网格702包括一个或多个RAID网格位置704。此外，在控制器实现3D RAID方案的情况下，RAID网格702可与RAID立方体700相关联。此外，每个RAID网格位置704与块708相关联，其中，每个块进一步与一个或多个页710相关联。每个页710与物理地址712相关联。此外，虽然在图7中未示出，但每个页710还与RAID网格层相关联(参见例如图2)。

[0074] 在本技术的一个实施例中，控制器跟踪由客户端提供的数据与存储阵列中的此类数据的物理地址之间的映射。在本技术的一个实施例中，控制器使用从客户端的角度出发识别数据的逻辑地址(例如，<对象，偏移>714)与识别存储阵列内的数据位置的物理地址712之间的映射来跟踪上述信息。在本技术的一个实施例中，映射可以是在从对<对象，偏移>应用散列函数(例如，MD5、SHA 1)得到的散列值与相应物理地址712之间。本领域的技术人员将认识到在不脱离本技术的情况下可使用任何形式的逻辑地址。在本技术的一个实施例中，使用块层级元数据来确定物理地址712到逻辑地址714映射。在本技术的一个实施例中，根据在美国专利号8,370,567中描述的方法来确定上述映射。

[0075] 在本技术的一个实施例中，控制器跟踪哪个RAID网格702(包括数据部分和奇偶部分中的RAID网格)与哪个RAID立方体700相关联(假设控制器正在实现3D RAID方案)以及哪些RAID网格位置(704)与每个RAID网格(702)相关联。

[0076] 在本技术的一个实施例中，控制器跟踪每个RAID网格位置706的状态716。在本技术的一个实施例中，可将RAID网格位置的状态716设定为已填充(表示块已被写入到RAID网格位置)或空(表示没有东西被写入到RAID网格位置)。在本技术的一个实施例中，如果控制器已识别到要写入到RAID网格位置的数据，控制器页可将RAID网格位置的状态设置成已填充(参见图8)。当最初创建RAID网格时，控制器可在最初将每个RAID网格位置的状态设置成空的。

[0077] 在本技术的一个实施例中，控制器跟踪每个RAID网格位置(706)被关联到哪个<区，侧面>(718)。在本技术的一个实施例中，控制器跟踪与每个RAID网格位置相关联的每个标签(720)(参见例如图5A—5C)的内容。

[0078] 图8—9示出了根据本技术的一个或多个实施例的流程图。更具体地，虽然按顺序提出并描述了流程图中的各种步骤，但本领域的技术人员将认识到某些或所有步骤可按照不同顺序执行，可被组合或省略，并且某些或所有步骤可并行地执行。在本技术的一个实施例中，可并行地执行图8—9中所示的方法。

[0079] 参考图8，图8示出了根据本技术的一个或多个实施例的用于将数据存储在存储阵列中的方法。

[0080] 在步骤800中，获得用于RAID网格的客户端数据和块层级元数据。在本技术的一个实施例中，可通过来自客户端的一系列写请求来获得客户端数据。此外，例如根据美国专利号8,370,567，获得块层级元数据可包括在接收到客户端数据之后生成块层级元数据。

[0081] 在步骤802中，使用在步骤800中获得的块层级元数据和客户端数据来生成奇偶值。可根据多维RAID方案(诸如上文所述的)来生成奇偶值。

[0082] 在步骤804中，生成针对用于RAID网格中的每个区中的侧面A中的每个RAID网格位置的标签。该标签包括如例如上文相对于图5A—5C和6所述的内容。在步骤804中产生的标签是基于针对用于区的至少P/E计数的当前值和用于区的坏位置信息，该坏位置信息是由控制器维护的。更具体地，在数据(包括用户数据和标签)最后一次被写入到区内的块期间确定用于区的P/E计数和用于区的坏位置信息。

[0083] 在步骤806中，擦除与用于每个区的侧面A的RAID网格位置相关联的所有块的内容。

[0084] 在步骤808中，使用擦除操作的结果来更新由控制器维护的一个或多个数据结构。更具体地，可更新针对所有区的用于侧面A的P/E计数和针对RAID网格中的所有区的用于侧面A的坏块信息。例如，擦除操作可导致P/E计数的更新，并且在对于给定块而言擦除操作失败的情况下可更新坏块信息。

[0085] 在步骤810中，向与RAID网格内的所有区中的侧面A相关联的RAID网格位置发布写。更具体地，包括相应标签的一页被写入到与用于RAID网格中的每个区的侧面A相关联的每个RAID网格位置。

[0086] 在步骤812中，擦除与用于每个区的侧面B的RAID网格位置相关联的所有块的内容。

[0087] 在步骤814中，使用擦除操作的结果来更新由控制器维护的一个或多个数据结构。更具体地，可更新针对所有区的用于侧面B的P/E计数和针对RAID网格中的所有区的用侧面B的坏块信息。例如，擦除操作可导致P/E计数的更新，并且在对于给定块而言擦除操作失败的情况下可更新坏块信息。在此阶段，用于区的当前P/E计数和用于区的当前坏块信息可用于控制器。换言之，在步骤808中，用于每个区的当前P/E计数的仅一部分是已知的，因为用于与每个区的侧面A相关联的RAID网格位置的仅已更新信息可用(即，因为仅对与用于给定区的侧面A相关联的RAID网格位置执行擦除操作)。

[0088] 在步骤816中，生成针对用于RAID网格中的每个区中的侧面B中的每个RAID网格位置的标签。该标签包括如例如上文关于图5A-5C和6所述的内容。在步骤804中产生的标签是基于由控制器维护的至少用于区的P/E计数的当前值和用于区的坏位置信息(即，至少部分地在步骤808和814中获得的已更新内容)。

[0089] 在步骤818中，向与RAID网格内的所有区中的侧面B相关联的RAID网格位置发布写。更具体地，包括相应标签的一个页被写入到与用于RAID网格中的每个区的侧面B相关联的每个RAID网格位置。

[0090] 在步骤820中，将相应的用户数据(例如，图5中的501)写入到RAID网格。

[0091] 参考图9，图9示出了根据本技术的一个或多个实施例的用于从RAID网格导入数据的方法。更具体地，图9示出了用于确定来自存储模块的先前存储内容是否有效，且在有效的情况下然后将内容的适当部分导入到存储控制器的存储器中的方法。图9中所示的方法在控制器(和/或系统中的其它部件)失去电源且然后其随后被用备用电源供电时发生。

[0092] 转到图9，假设可向系统恢复电源。一旦电源已经恢复，则该过程可前进至(直接地或在已执行重新启动系统有关的其它步骤(未示出)之后)步骤900。在步骤900中，选择RAID

网格。在步骤902中，选择RAID网格内的区。

[0093] 在步骤904中，从所选区的侧面B中的每个块获得有效标签。在本技术的实施例中，可不从所选区的侧面B中的每个块获得有效标签。在本技术的一个实施例中，有效标签对应于可从存储阵列成功地读取的标签。

[0094] 在步骤906中，进行关于至少每个有效标签中的时间戳是否一致的确定。如果标签(和/或标签内的其它内容)(例如，网格几何结构、坏位置信息等)是不一致(即，相同)的，则过程前进至步骤908；可替换地，过程前进至步骤912。

[0095] 在步骤908中，选择来自在步骤904中获得的有效标签组的标签。在步骤910中，进行关于在RAID网格中是否存在要处理的附加区的确定。如果在RAID网格中存在要处理的附加区，则过程前进至步骤902；否则，过程前进至步骤922。

[0096] 关于步骤906，如果至少每个有效标签中的时间戳是不一致的，则过程前进至步骤912。在步骤912中，选择RAID网格中的区。在步骤914中，从已从步骤912选择的区的侧面A中的每个块获得有效标签。在本技术的一个实施例中，可不从所选区的侧面A中的每个块获得有效标签。

[0097] 在步骤916中，进行关于至少每个有效标签中的时间戳是否一致的确定。如果标签(和/或标签内的其它内容)(例如，网格几何结构、坏位置信息等)是不一致(即，相同)的，则过程前进至步骤924；可替换地，过程前进至步骤918。

[0098] 在步骤918中，选择来自在步骤914中获得的有效标签组的标签。

[0099] 在步骤920中，进行关于在RAID网格中是否存在要处理的附加区的确定。更具体地，步骤920中的确定是基于在RAID网格中是否存在对于其而言与区的侧面A相关联的标签未被处理的其它区。如果在RAID网格中存在要处理的附加区，则方法前进至步骤912；否则，过程前进至步骤922。在本技术的一个实施例中，如果(i)在步骤902—910中针对每个区已识别到来自侧面B的有效标签，或者(ii)在步骤912—920中针对每个区已识别到来自侧面B的有效标签，则过程可仅前进至步骤922。

[0100] 在步骤922中，进行关于至少每个所选标签(在步骤908或918中获得)中的时间戳是否一致的确定。如果标签(和/或标签内的其它内容)(例如，网格几何结构、坏位置信息等)是不一致(即，相同)的，则过程前进至步骤924；可替换地，过程前进至步骤926。

[0101] 在本技术的一个实施例中，步骤902—910和912—920确保特定区的给定侧面内的标签是一致的，而步骤922确保跨整个RAID网格的标签是一致的。在本技术的一个实施例中，由于用于RAID网格中的所有区中的给定侧面的标签应具有至少相同的时间戳(基于其被写入到存储阵列的方式，参见例如图8)，所以可至少部分地基于这样的事实来确定标签的有效性，即针对RAID网格中的所有区的用于给定侧面的标签应具有相同时间戳。如果其不具有，则标签是不一致的的。此外，由于每个标签仅包括网格层级元数据的一部分(参见图5A—5C)，以便重构用于RAID网格的网格层级元数据，所以必须从RAID网格中的每个区获得有效标签，并且这些标签中的每一个必须具有相同的时间戳。至少时间戳的一致性确保标签的内容是适当可组合的(即，每个所选标签的内容对应于同一RAID网格(即，在特定时间的RAID网格)的一部分)。

[0102] 参考图9，在步骤926中，完成RAID网格的导入。RAID网格的导入可包括从标签导入信息，以便确定例如RAID网格和各种奇偶值在RAID网格内的位置。导入还可包括生成关于

图6所述的数据结构中的一个或多个。在本技术的一个实施例中，在没有来自标签的信息的情况下，控制器将不具有关于用户数据如何被存储在存储阵列中的足够信息，并且同样地将不能检索此信息和/或实现多维RAID方案。

[0103] 关于图9，如果未获得一致的一组有效标签，则在步骤924中，不存在用以导入RAID网格的足够信息，并且同样地向适当的个体、实体等发布错误通知。

[0104] 本技术的实施例提供了一种用于使用两个不同的机制—分别是复制和多维RAID方案—在存储池中存储网格层级元数据(即，标签)和用户数据的机制。此外，用来在存储阵列中存储上述数据的方式是自描述的。具体地，标签提供网格层级和区层级元数据，其使得能够恢复至少实现多维RAID方案所需的网格几何结构(及其它信息)。此外，在每个块内，存在包括用户数据(包括块层级元数据)的一个或多个页，其中，块层级元数据被用来填充一个或多个数据结构。这些数据结构然后可被用来获得存储在存储阵列内的用户数据。

[0105] 本领域的技术人员将认识到已相对于沿着IFD在存储阵列中存储数据和/或在NAND闪存中存储数据描述了本技术的各种示例，但在不脱离本技术的情况下可在任何多维磁盘阵列上实现本技术的实施例。例如，可使用存储设备(磁性、光学、固态或任何其它类型的存储设备)的二维阵列来实现本技术的一个或多个实施例，其中，用于RAID网格中的每个RAID网格位置的数据被存储在单独磁盘上。

[0106] 此外，在本技术的一个实施例中，在控制器在使用三维磁盘阵列来实现3D RAID方案时，控制器可使用以下n元组来存储用于每个RAID网格位置的数据：＜磁盘x，磁盘y，磁盘z，逻辑块地址(LAB)a＞，其中，x、y和z是磁盘阵列的维度。

[0107] 用于使用二维磁盘阵列来实现本技术的实施例的上述示例并不意图限制本技术的范围。

[0108] 本领域的技术人员将认识到已关于2D RAID方案和3D RAID方案描述了本技术，但可将本技术的实施例扩展至任何多维RAID方案。

[0109] 可使用由系统中的一个或多个处理器执行的指令来实现本技术的一个或多个实施例。此外，此类指令可对应于存储在一个或多个非临时计算机可读介质上的计算机可读指令。

[0110] 虽然已相对于有限数目的实施例描述了本技术，但受益于本公开的本领域的技术人员将认识到可以发明不脱离如在本文中公开的本技术的范围的其它实施例。因此，本技术的范围将仅仅受到所附权利要求的限制。

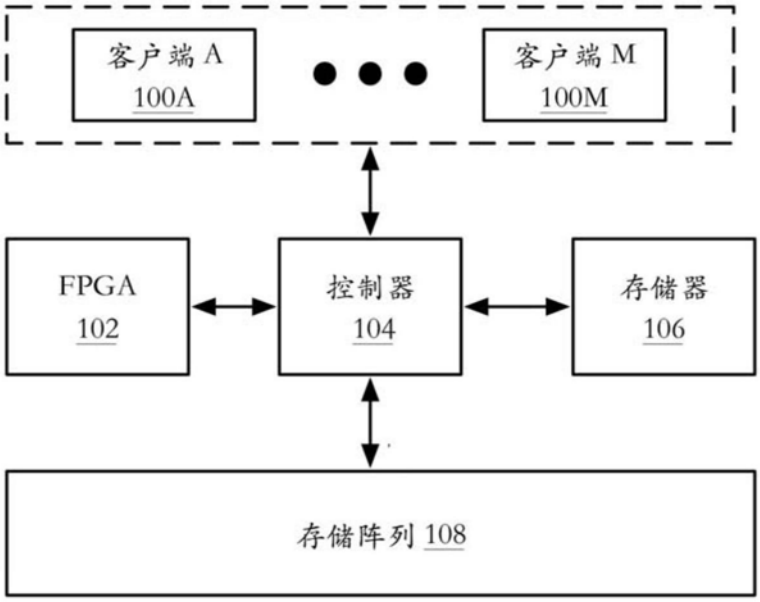

客户端 A
100A
客户端 M
100M
FPGA
102
控制器
104
存储器
106
存储阵列 108

图1

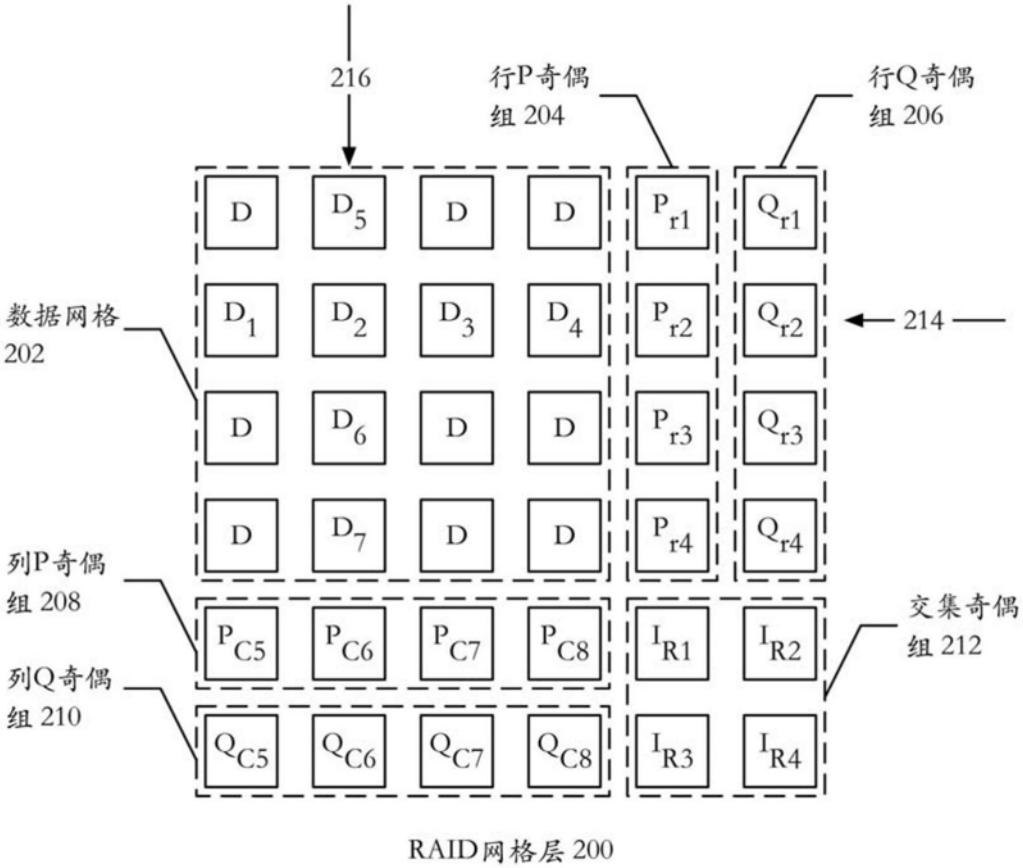
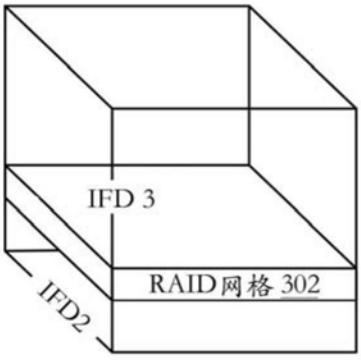

216
行P奇偶组 204
行Q奇偶组 206
数据网格 202
214
列P奇偶组 208
列Q奇偶组 210
交集奇偶组 212
D
D5
D1
D2
D3
D4
D6
D7
Pr1
Pr2
Pr3
Pr4
Qr1
Qr2
Qr3
Qr4
PC5
PC6
PC7
PC8
QC5
QC6
QC7
QC8
IR1
IR2
IR3
IR4
RAID网格层 200

图2



RAID立方体 300

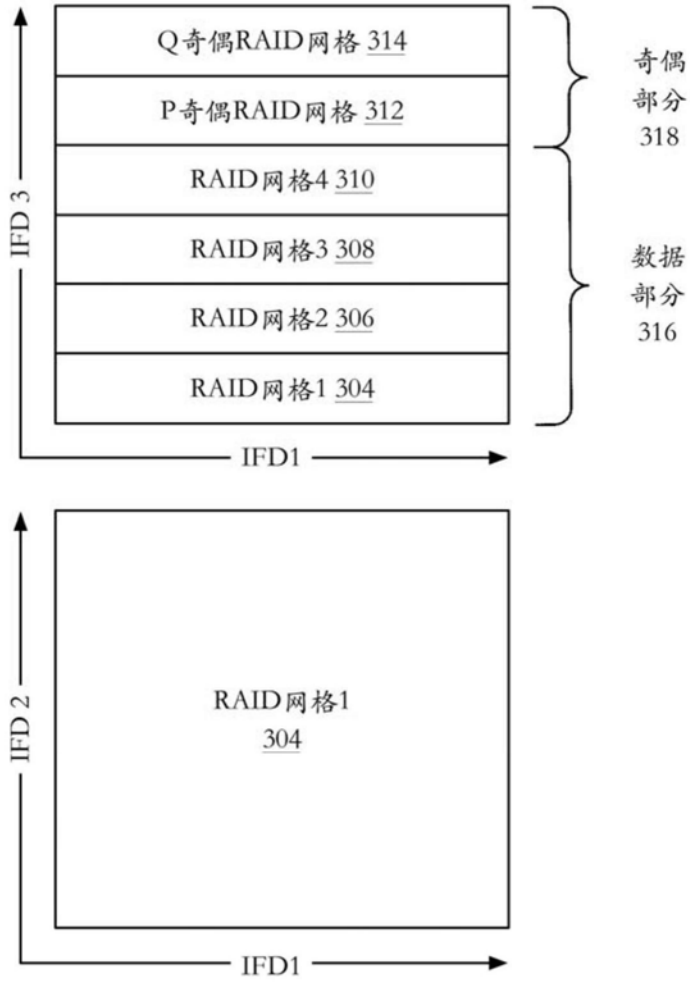


图3

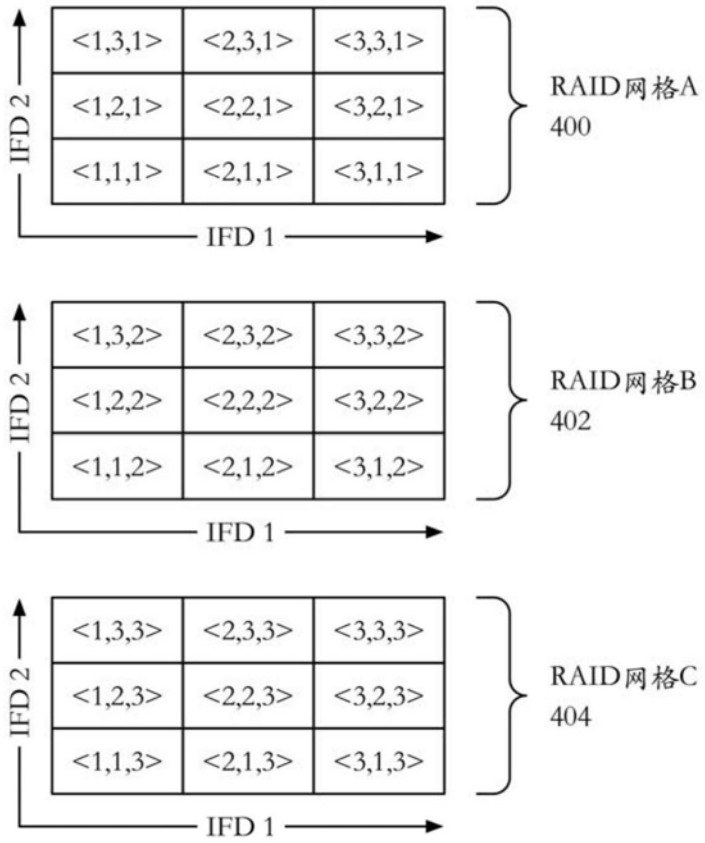

<1,3,1> <2,3,1> <3,3,1>
<1,2,1> <2,2,1> <3,2,1>
<1,1,1> <2,1,1> <3,1,1>
IFD 2
IFD 1
RAID网格A
400
<1,3,2> <2,3,2> <3,3,2>
<1,2,2> <2,2,2> <3,2,2>
<1,1,2> <2,1,2> <3,1,2>
IFD 2
IFD 1
RAID网格B
402
<1,3,3> <2,3,3> <3,3,3>
<1,2,3> <2,2,3> <3,2,3>
<1,1,3> <2,1,3> <3,1,3>
IFD 2
IFD 1
RAID网格C
404

图4A

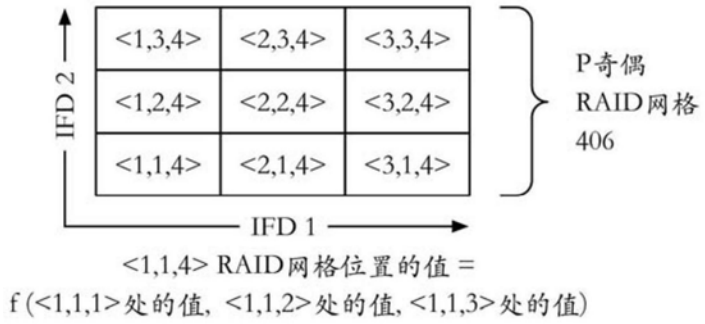
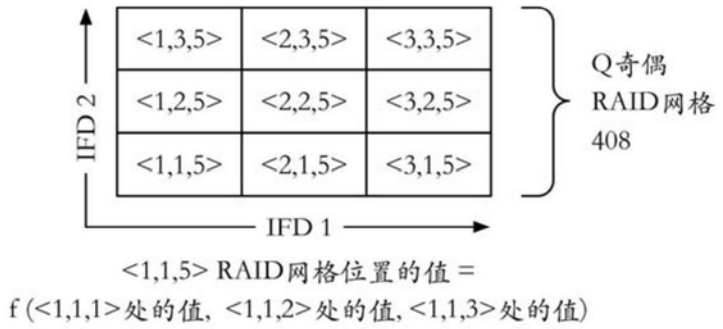

<1,3,4> <2,3,4> <3,3,4>
<1,2,4> <2,2,4> <3,2,4>
<1,1,4> <2,1,4> <3,1,4>
IFD 2
IFD 1
P奇偶
RAID网格
406
<1,1,4> RAID网格位置的值 =
f(<1,1,1>处的值, <1,1,2>处的值, <1,1,3>处的值)

图4B



<1,1,5> RAID网格位置的值 =
f (<1,1,1>处的值, <1,1,2>处的值, <1,1,3>处的值)

图4C

图4D

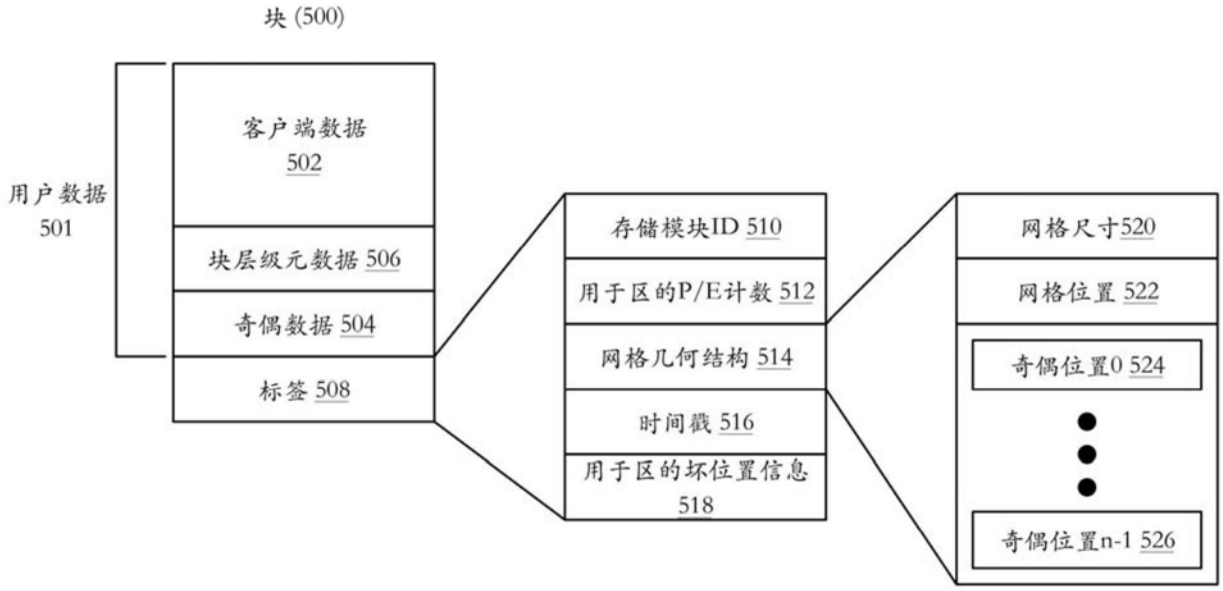

块(500)
用户数据
501
客户端数据
502
块层级元数据 506
奇偶数据 504
标签 508
存储模块ID 510
用于区的P/E计数 512
网格几何结构 514
时间戳 516
用于区的坏位置信息
518
网格尺寸520
网格位置 522
奇偶位置0 524
奇偶位置n-1 526

图5A 图5B 图5C

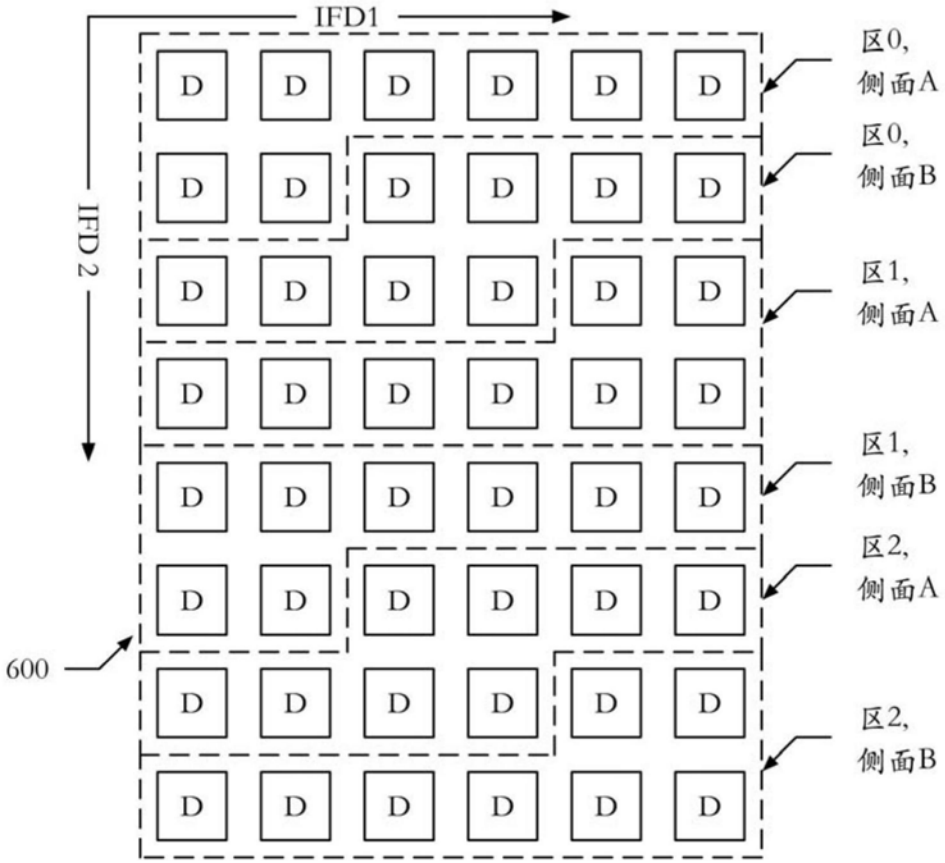


图6

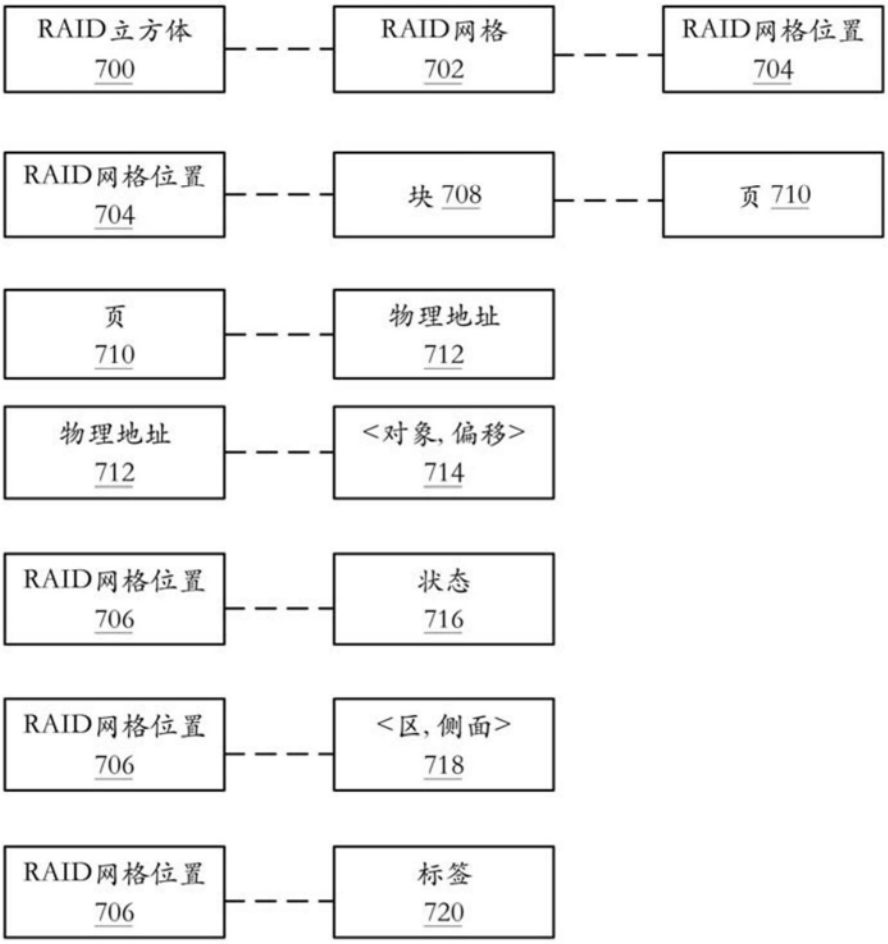

RAID立方体
700
RAID网格
702
RAID网格位置
704
RAID网格位置
704
块 708
页 710
页
710
物理地址
712
物理地址
712
<对象, 偏移>
714
RAID网格位置
706
状态
716
RAID网格位置
706
<区, 侧面>
718
RAID网格位置
706
标签
720

图7

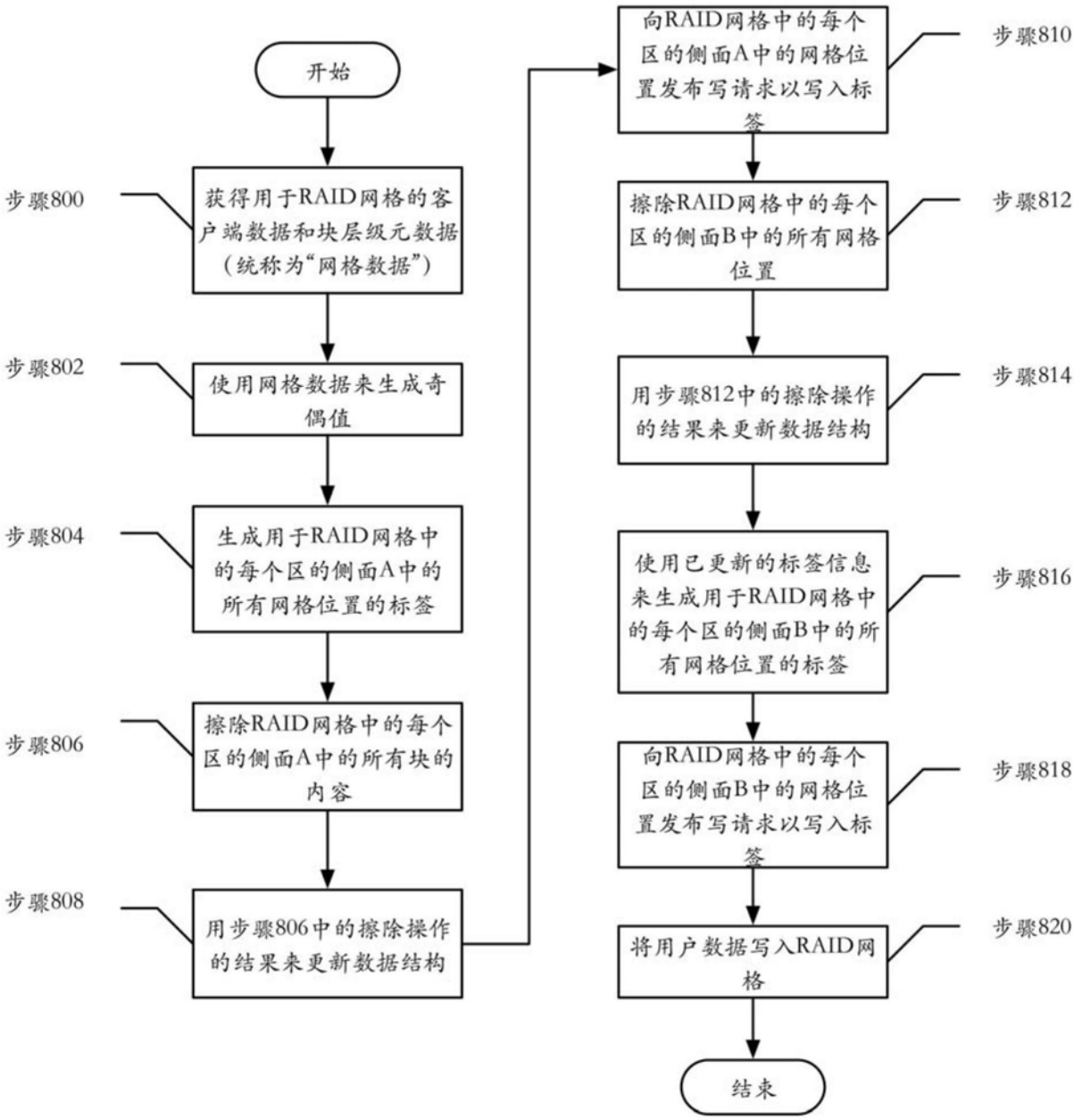

开始
步骤800
获得用于RAID网格的客户端数据和块层级元数据（统称为“网格数据”）
步骤802
使用网格数据来生成奇偶值
步骤804
生成用于RAID网格中的每个区的侧面A中的所有网格位置的标签
步骤806
擦除RAID网格中的每个区的侧面A中的所有块的内容
步骤808
用步骤806中的擦除操作的结果来更新数据结构
步骤810
向RAID网格中的每个区的侧面A中的网格位置发布写请求以写入标签
步骤812
擦除RAID网格中的每个区的侧面B中的所有网格位置
步骤814
用步骤812中的擦除操作的结果来更新数据结构
步骤816
使用已更新的标签信息来生成用于RAID网格中的每个区的侧面B中的所有网格位置的标签
步骤818
向RAID网格中的每个区的侧面B中的网格位置发布写请求以写入标签
步骤820
将用户数据写入RAID网格
结束

图8

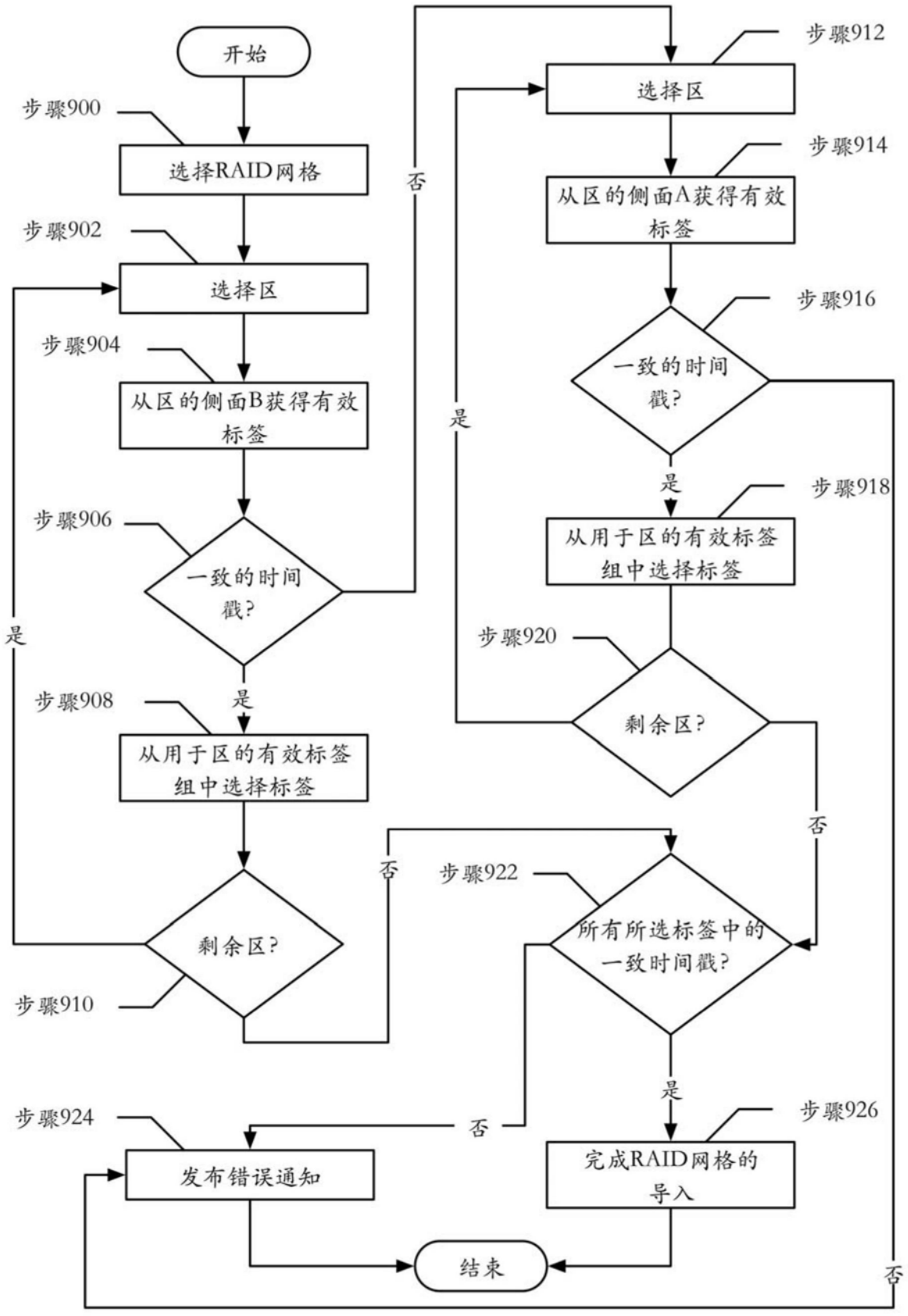

开始
步骤900
选择RAID网格
步骤902
选择区
步骤904
从区的侧面B获得有效标签
步骤906
一致的时间戳?
否
是
步骤908
从用于区的有效标签组中选择标签
剩余区?
步骤910
是
否
步骤912
选择区
步骤914
从区的侧面A获得有效标签
步骤916
一致的时间戳?
是
否
步骤918
从用于区的有效标签组中选择标签
步骤920
剩余区?
是
否
步骤922
所有所选标签中的一致时间戳?
是
否
步骤924
发布错误通知
步骤926
完成RAID网格的导入
结束

图9