(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 108632029 B
(45)授权公告日 2022.05.17

(21)申请号 201810226662.9
(22)申请日 2018.03.19
(65)同一申请的已公布的文献号
申请公布号 CN 108632029 A
(43)申请公布日 2018.10.09
(30)优先权数据
62/474,039 2017.03.20 US
62/561,625 2017.09.21 US
62/562,219 2017.09.22 US
15/876,028 2018.01.19 US
(73)专利权人 三星电子株式会社
地址 韩国京畿道水原市灵通区三星路129号
(72)发明人 奇亮奭
(74)专利代理机构 华进联合专利商标代理有限公司 44224
专利代理师 刘培培 黄隶凡
(51)Int.Cl.
H04L 9/08(2006.01)
(56)对比文件
US 8504535 B1,2013.08.06
US 2015363269 A1,2015.12.17
CN 102449613 A,2012.05.09
CN 102713851 A,2012.10.03
CN 102405460 A,2012.04.04
US 8504535 B1,2013.08.06
审查员 李常亮

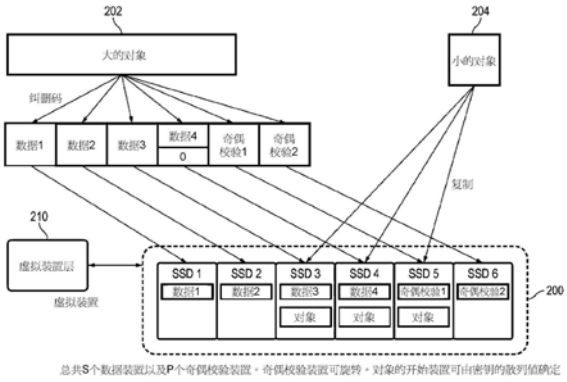
权利要求书2页 说明书11页 附图7页

(54)发明名称

密钥值固态驱动器

(57)摘要

一种存储装置及密钥值固态驱动器。存储装置包括:多个存储器装置,被配置成利用无状态数据保护的虚拟密钥值固态驱动器装置;以及虚拟装置层,被配置成管理所述虚拟装置以通过以下方式存储对象:根据所述对象各自的大小向所述对象中的一些对象应用纠删编码、以及向所述对象中的其他对象应用复制。



CN 108632029 B

1.一种存储装置,其特征在于,包括:

多个存储器装置,被配置成利用无状态数据保护的虚拟装置;以及

虚拟装置层,被配置成管理所述虚拟装置以通过以下方式存储对象:根据所述对象的第一数据保护的第一空间开销以及第二数据保护的第二空间开销向所述对象中的部份对象应用所述第一数据保护、以及向所述对象中的其他对象应用所述第二数据保护,其中所述第一数据保护的所述第一空间开销以及所述第二数据保护的所述第二空间开销是基于所述对象各自的大小。

2.根据权利要求1所述的存储装置,其特征在于,所述存储器装置被配置成一个或多个数据装置以及一个或多个奇偶校验装置。

3.根据权利要求2所述的存储装置,其特征在于,所述第一数据保护包括纠删编码,且所述第二数据保护包括复制。

4.根据权利要求3所述的存储装置,其特征在于,当所述对象中的对应一者被分类为大的对象时利用所述纠删编码进行数据保护。

5.根据权利要求4所述的存储装置,其特征在于,当((P+1)*0>(S+P)*m且0>=S*m)时,所述对象中的所述对应一者被分类为所述大的对象,其中0是指对象大小;P是指奇偶校验装置的数目;S是指数据装置的数目;且m是指容许的最小大小值。

6.根据权利要求3所述的存储装置,其特征在于,当所述对象中的对应一者被分类为小的对象时利用所述复制进行数据保护。

7.根据权利要求6所述的存储装置,其特征在于,当((P+1)*0=<(S+P)*m))时,所述对象中的所述对应一者被分类为所述小的对象,其中0是指对象大小;P是指奇偶校验装置的数目;S是指数据装置的数目;且m是指容许的最小大小值。

8.根据权利要求3所述的存储装置,其特征在于,当所述对象中的对应一者既不被分类为大的对象也不被分类为小的对象时,基于性能指标及数据使用特性利用所述纠删编码或所述复制进行数据保护。

9.根据权利要求8所述的存储装置,其特征在于,当((P+1)*0>(S+P)*m)>S*m>0)时,所述对象中的所述对应一者被分类为中等对象,其中0是指对象大小;P是指奇偶校验装置的数目;S是指数据装置的数目;且m是指容许的最小大小值。

10.根据权利要求2所述的存储装置,其特征在于,当存储一个或多个大的对象时,所述奇偶校验装置是固定的。

11.根据权利要求2所述的存储装置,其特征在于,当存储一个或多个大的对象时,所述奇偶校验装置旋转。

12.根据权利要求1所述的存储装置,其特征在于,所述存储器装置包括固态驱动器。

13.一种利用虚拟装置层在包括多个存储器装置的虚拟装置中存储对象的方法,其特征在于,所述方法包括:

由所述虚拟装置层判断所述对象中的对应一者是大还是小;

如果所述对象中的所述对应一者被分类为大的对象:

确定用于纠删编码的块大小以及所述对象中的所述对应一者的一个或多个数据块的填充量;

使用纠删编码来计算P个奇偶校验块;

确定用以存储所述数据块及奇偶校验块的所述存储器装置;以及

将所述数据块及奇偶校验块写入所述存储器装置;且

如果所述对象中的所述对应一者被分类为小的对象:

确定用于数据及通过复制而产生的复制物的所述存储器装置;以及

将所述数据及所述复制物写入所述存储器装置。

14.根据权利要求13所述的方法,其特征在于,当所述对象中的所述对应一者既不是大的对象也不是小的对象时,所述对象中的所述对应一者被分类为中等对象,且基于性能指标及数据使用特性应用所述复制或所述纠删编码。

15.根据权利要求13所述的方法,其特征在于,对应于所述对象中的至少两者的所述奇偶校验块存储在所述存储器装置的固定子集上。

16.根据权利要求13所述的方法,其特征在于,对应于所述对象中的不同者的所述奇偶校验块不存储在所述存储器装置的固定子集上。

17.根据权利要求13所述的方法,其特征在于,对应于所述对象中的至少两者的所述数据及所述复制物存储在所述存储器装置中的不同者上。

18.根据权利要求13所述的方法,其特征在于,所述数据块中的至少一者被填充以零。

19.一种由虚拟装置层利用密钥从包括多个存储器装置的虚拟装置读取对象的方法,其特征在于,所述方法包括:

由所述虚拟装置层向所有所述存储器装置发送读取请求;以及

由所述虚拟装置层接收来自所述存储器装置的回应,其中

根据所述对象的纠删编码的第一空间开销以及复制的第二空间开销判断所述对象为大的对象或小的对象,其中所述纠删编码的所述第一空间开销以及所述复制的所述第二空间开销是基于所述对象的大小,

如果所述对象是大的对象,由所述虚拟装置层接收数据块及奇偶校验块以利用所述纠删编码重建所述对象,且

如果所述对象是小的对象,所述数据块是所述对象或是所述对象的复制物。

20.根据权利要求19所述的方法,其特征在于,所述密钥包括用于从所述多个存储器装置中确定开始装置或主装置的散列值。

密钥值固态驱动器

[0001] 相关申请的交叉参考

[0002] 本申请主张在2017年3月20号提出申请且标题为密钥值固态驱动器的美国临时专利申请第62/474,039号、在2017年9月21号提出申请且标题为用于密钥值存储的混合无状态数据保护方法及系统的美国临时专利申请第62/561,625号、以及在2017年9月22号提出申请且标题为用于密钥值存储的混合无状态数据保护方法及系统的美国临时专利申请第62/562,219号的优先权及权利,所述三个美国临时专利申请的全部内容均并入本案供参考。

技术领域

[0003] 本发明概念涉及密钥值存储系统。

背景技术

[0004] 传统固态驱动器(solid state drive,SSD)通常仅使用区块接口(block interface)并通过独立盘的冗余阵列(redundant array of independent disks,RAID)、纠删编码、或复制来提供数据可靠性。随着对象格式变得在大小方面可变且变得非结构化,期望在对象与区块级接口之间进行有效的数据转换。此外,可取的是在保持空间效率及快速访问时间特性的同时确保数据可靠性。

发明内容

[0005] 根据本发明的示例性实施例涉及与区块装置(block device)不同的密钥值存储系统(例如,密钥值固态驱动器)。

[0006] 本发明的一些示例性实施例涉及如何针对密钥值固态驱动器实现数据可靠性。将基于空间开销(space overhead)的复制与纠删编码的混合应用至一组密钥值固态驱动器,所述密钥值固态驱动器可实现对象的无状态可变长度纠删码。

[0007] 本发明的一些示例性实施例具有一个或多个以下特性:1)针对每一可变对象而不针对每一固定区块提供可靠性;2)可将复制与纠删编码混合以针对单一磁盘组实现对象的目标可靠性;3)空间效率是主要指标且性能是次要指标以为对象确定正确的技术;4)与独立盘的冗余阵列(RAID)类似,机制是无状态的;5)对复制或纠删编码来说,不需要存储额外的信息;以及6)无论对象大小如何,不需要读-修改-写(read-modify-write)用于更新。

[0008] 本发明的一些示例性实施例提供一种实现一组密钥值固态驱动器的可靠性的方法。此外,示例性实施例可避免针对由于根据示例性实施例在对区块内的一部分数据进行更新的情形中的区块装置发生的读-修改-写,可靠性针对每一对象(例如,可变对象)而不是针对每一区块(例如,固定区块)来提供。

[0009] 根据本发明的示例性实施例,一种存储装置包括:多个存储器装置,被配置成利用无状态数据保护的虚拟装置;以及虚拟装置层,被配置成管理所述虚拟装置以通过以下方式存储对象:根据所述对象各自的大小向所述对象中的一些对象应用第一数据保护、以及

向所述对象中的其他对象应用第二数据保护。

[0010] 所述存储器装置可被配置成一个或多个数据装置以及一个或多个奇偶校验装置。

[0011] 所述第一数据保护可包括纠删编码，且所述第二数据保护可包括复制。

[0012] 当所述对象中的对应一者被分类为大的对象时可利用所述纠删编码进行数据保护。

[0013] 当((P+1)*O>(S+P)*m且O>＝S*m)时，所述对象中的所述对应一者可被分类为所述大的对象，其中O是指对象大小；P是指奇偶校验装置的数目；S是指数据装置的数目；且m是指容许的最小大小值。

[0014] 当所述对象中的对应一者被分类为小的对象时可利用所述复制进行数据保护。

[0015] 当((P+1)*O＝<(S+P)*m))时，所述对象中的所述对应一者可被分类为所述小的对象，其中O是指对象大小；P是指奇偶校验装置的数目；S是指数据装置的数目；且m是指容许的最小大小值。

[0016] 当所述对象中的对应一者既不被分类为大的对象也不被分类为小的对象时，可基于性能指标及数据使用特性利用所述纠删编码或所述复制进行数据保护。

[0017] 当((P+1)*O>(S+P)*m)>S*m>O)时，所述对象中的所述对应一者可被分类为中等对象，其中O是指对象大小；P是指奇偶校验装置的数目；S是指数据装置的数目；且m是指容许的最小大小值。

[0018] 当存储一个或多个大的对象时，所述奇偶校验装置可以是固定的。

[0019] 当存储一个或多个大的对象时，所述奇偶校验装置可旋转。

[0020] 所述存储器装置可包括固态驱动器。

[0021] 根据本发明的另一示例性实施例，提供一种利用虚拟装置层在包括多个存储器装置的虚拟装置中存储对象的方法。所述方法包括：由所述虚拟装置层判断所述对象中的对应一者是大还是小；如果所述对象中的所述对应一者被分类为大的对象：确定用于纠删编码的块(chunk)大小以及所述对象中的所述对应一者的数据块(data chunk)的填充量；使用纠删编码来计算P个奇偶校验块；确定用以存储所述数据块及奇偶校验块的所述存储器装置；以及将所述数据块及奇偶校验块写入所述存储器装置；且如果所述对象中的所述对应一者被分类为小的对象：确定用于数据及复制物的所述存储器装置；以及将所述数据及所述复制物写入所述存储器装置。

[0022] 当所述对象中的所述对应一者既不是大的对象也不是小的对象时，所述对象中的所述对应一者可被分类为中等对象，且可基于性能指标及数据使用特性应用所述复制或所述纠删编码。

[0023] 对应于所述对象中的至少两者的所述奇偶校验块可存储在所述存储器装置的固定子集上。

[0024] 对应于所述对象中的不同者的所述奇偶校验块可不存储在所述存储器装置的固定子集上。

[0025] 对应于所述对象中的至少两者的所述数据及所述复制物可存储在所述存储器装置中的不同者上。

[0026] 所述数据块中的至少一者可被填充以零。

[0027] 根据本发明的另一示例性实施例，提供一种由虚拟装置层利用密钥从包括多个存

储器装置的虚拟装置读取对象的方法。所述方法包括:由所述虚拟装置层向所有所述存储器装置发送读取请求;以及由所述虚拟装置层接收来自所述存储器装置的回应,其中如果所述对象是大的对象,那么由所述虚拟装置层接收数据块及奇偶校验块以利用纠删编码重建所述对象,且如果所述对象是小的对象,那么所述数据块是所述对象或是所述对象的复制物。

[0028] 所述密钥可包括用于从所述多个装置中确定开始装置(start device)或主装置(primary device)的散列(密钥)。

附图说明

[0029] 以下,将参照附图更详细地阐述示例性实施例,在所有附图中相同的参考编号指代相同的元件。然而,本发明可实现为各种不同形式,而不应被视为仅限于本文中所说明的实施例。确切来说,提供这些实施例作为实例是为了使公开内容将透彻及完整,并将向所属领域中的技术人员充分传达本发明的方面及特征。因此对所属领域中的普通技术人员完全理解本发明的方面及特征来说非必要的工艺、元件及技术可不再进行阐述。除非另有说明,否则在所有附图及书面说明通篇中相同的参考编号表示相同的元件,且因此将不再对其予以重复赘述。在图式中,为清晰起见可夸大各元件、层及区的相对大小。

[0030] 尽管已说明并阐述了本发明的某些实施例,但所属领域中的普通技术人员将理解,在不背离由以上权利要求书及其等效范围界定的本发明的精神及范围的条件下,可对所述实施例作出某些修改及变化。举例来说,如所属领域中的技术人员可理解,在不背离本发明的精神及范围的条件下,各种图式中的示例性实施例的特征可进行组合。

[0031] 图1是根据本发明示例性实施例的密钥值(key value,KV)固态驱动器(SSD)的示意图。

[0032] 图2是根据示例性实施例说明包括一组装置的虚拟装置以及在所述虚拟装置中对对象的存储的概念图。

[0033] 图3是根据本发明示例性实施例将对象写入虚拟装置的流程图。

[0034] 图4是根据示例性实施例说明以共享奇偶校验方式(shared parity manner)将大的对象存储在图2所示的虚拟装置中的概念图。

[0035] 图5是根据示例性实施例说明以专用奇偶校验方式(dedicated parity manner)将大的对象存储在图2所示的虚拟装置中的概念图。

[0036] 图6是根据示例性实施例说明将小的对象存储在图2所示的虚拟装置中的概念图。

[0037] 图7是根据本发明示例性实施例从虚拟装置读取对象的流程图。

[0038] [符号的说明]

[0039] 10:固态驱动器(SSD);

[0040] 15:密钥值应用程序接口;

[0041] 20:用户密钥值装置驱动器;

[0042] 200:虚拟装置;

[0043] 202:大的对象;

[0044] 204:小的对象;

[0045] 210:虚拟装置层;

[0046] 242、244:大的对象;

[0047] 262、264:小的对象;

[0048] 300、302、304、306、308、310、312、314、316、318、320:方框

[0049] 700、702、704、706、708、710、712、714、716、718、720、722、724、726、728、730、732、734:方框。

具体实施方式

[0050] 应理解,尽管本文中可能使用“第一(first)”、“第二(second)”、“第三(third)”等用语来阐述各种元件、组件、区、层及/或区段,然而这些元件、组件、区、层及/或区段不应受这些用语限制。这些用语仅用于区分元件、组件、区、层或区段与另一元件、组件、区、层或区段。因此,在不背离本发明的精神及范围的条件下,以下阐述的第一元件、组件、区、层、或区段亦可被称为第二元件、组件、区、层、或区段。

[0051] 应理解,当称一个元件或层位于另一元件或层“上(on)”,“连接到(connected to)”或“耦合到(coupled to)”另一元件或层时,所述元件或层可直接位于所述另一元件或层上,直接连接到或耦合到所述另一元件或层,抑或可存在一个或多个中间元件或层。另外,还应理解,当称一个元件或层“位于”两个元件或层“之间(between)”时,所述元件或层可为所述两个元件或层之间的唯一元件或层,抑或也可存在一个或多个中间元件。

[0052] 本文中所使用的术语仅是为了阐述特定实施例而并非旨在限制本发明。除非上下文清楚地另外指明,否则本文中所使用的单数形式“一(a及an)”旨在也包括复数形式。应进一步理解,当在本说明书中使用用语“包括(comprises及comprising)”及“包含(includes及including)”时,是指明所陈述特征、整数、步骤、操作、元件及/或组件的存在,但不排除一个或多个其他特征、整数、步骤、操作、元件、组件及/或其群组的存在或添加。本文中所使用的用语“及/或(and/or)”包括相关所列项其中一个或多个项的任意及所有组合。当例如“...中的至少一个(at least one of)”等表达位于一系列元件之前时,是修饰整个系列元件而不是修饰所述一系列元件中的个别元件。

[0053] 本文所使用的用语“大体上(substantially)”、“大约(about)”及类似用语用作近似用语而非程度用语,并且旨在考虑所属领域中的普通技术人员将认识到的测量值或计算值的固有偏差。此外,在阐述本发明的实施例时使用的“可(may)”是指“本发明的一个或多个实施例”。本文中所使用的用语“使用(use)”、“正使用(using)”、及“被使用(used)”可被视为分别与用语“利用(utilize)”、“正利用(utilizing)”、及“被利用(utilized)”同义。此外,用语“示例性”旨在指实例或例证。

[0054] 根据本文中所述的本发明的实施例,电子装置或电气装置及/或任意其他相关的装置或组件(例如,主机、固态驱动器、存储器装置、及虚拟装置层)可利用任意适当的硬件、韧件(例如,专用集成电路)、软件、或软件、韧件及硬件的适当组合来实现。举例来说,这些装置的各种组件(例如,密钥值固态驱动器、主机、固态驱动器、存储器装置、及虚拟装置层)可形成在一个集成电路(integrated circuit,IC)芯片上或形成在单独的集成电路芯片上。此外,这些装置的各种组件可在柔性印刷电路膜、载带封装(tape carrier package,TCP)、印刷电路板(printed circuit board,PCB)上实现,或可形成在一个衬底上。此外,这些装置的各种组件可为在一个或多个计算装置中在一个或多个处理器上运行的进程

(process)或线程(thread),用于执行计算机程序指令并与其他系统组件交互作用以执行本文中所述的各种功能。所述计算机程序指令存储在存储器中,所述存储器可在使用标准存储器装置(例如,随机存取存储器(random access memory,RAM)或快闪存储器(例如,NAND快闪存储器)装置)的计算装置中实现。所述计算机程序指令也可被存储在其他非暂时性计算机可读介质(例如,CD-ROM、或快闪驱动器等)中。此外,所属领域中的技术人员应认识到,在不背离本发明的示例性实施例的精神及范围的条件下,各种计算装置的功能可组合或集成到单个计算装置中,或特定计算装置的功能可分布在一个或多个其他计算装置上。

[0055] 除非另外定义,否则本文中所用的所有用语(包括技术及科学用语)的意义均与本发明所属领域中的普通技术人员所通常理解的意义相同。应进一步理解,用语(例如在常用字典中所定义的用语)应被解释为具有与其在相关技术的上下文及/或本说明书中的含义一致的含义,且除非在本文中明确定义,否则不应将其解释为具有理想化或过于正式的意义。

[0056] 图1是根据本发明示例性实施例的密钥值(key value,KV)固态驱动器(SSD)10的示意图。根据示例性实施例的存储系统(或存储装置)包括一个或多个密钥值固态驱动器,例如图1中所示的一个密钥值固态驱动器,但本发明并非仅限于此。

[0057] 根据本发明的示例性实施例,密钥值固态驱动器10中的密钥值应用程序接口15与不需要传统区块映射的用户密钥值装置驱动器20一起运作。

[0058] 根据本发明的示例性实施例,包括密钥值固态驱动器10的存储系统通过以下方式利用混合无状态数据保护方法:根据对象各自的大小向一些对象应用第一数据保护(例如,纠删编码),并向其他对象应用第二数据保护(例如,复制),以实现所需的可靠性(例如,目标可靠性)。如此一来,可在不牺牲可靠性的情况下提供空间高效型(space-efficient)解决方案。在根据一些实施例密钥值固态驱动器10自身可执行混合无状态数据保护方法时,当所述混合无状态数据保护方法是由例如存储系统执行时,对驱动器(固态驱动器)的管理(例如,虚拟装置层的操作)可变得更容易。

[0059] 根据本发明的示例性实施例,可对对象进行分类以获得空间效率,可基于大小对所述对象进行分类,且针对每一大小等级可使用不同的备份(backup)方法。

[0060] 如果对一个对象进行纠删编码的空间开销小于对所述对象进行复制的空间开销,那么所述对象可被视为大的对象。在此种情形中,由于纠删编码具有较小的空间占用区域(space footprint),因此可期望使用纠删编码。换句话说,当对象满足以下不等式(((P+1)*O>(S+P)*m且O>=S*m))时,可将对象视为大的对象,其中O是对象值大小。在本文中且在以下不等式中,O=对象大小(即,对象的大小);P=奇偶校验装置计数(即,在虚拟装置中奇偶校验装置的数目);S=数据装置计数(即,虚拟装置中数据装置的数目);且m=容许的最小大小值(即,个别装置的所有最小值大小中的最大值)。举例来说,根据示例性实施例的“容许的最小大小值”是指在不违反系统中任意装置的最小值大小要求的情况下可被存储到系统中任意装置的值大小。每一装置具有装置支持的最小对象大小。由于根据示例性实施例对象被分成用于所有装置的相等大小,因此所述大小应大于装置支持的任意最小大小。如果尝试存储大小小于m的对象,那么至少一个装置无法存储所述对象。

[0061] 换句话说,当满足以下条件中的两者时将对象视为大的对象:1)对象的大小O乘以

奇偶校验装置的数目加一(P+1)大于容许的最小大小值m乘以数据装置S的数目与奇偶校验装置P的数目的总和S+P;以及2)对象的大小O大于或等于数据装置S的数目乘以容许的最小大小值m。

[0062] 根据示例性实施例,可对大的对象进行纠删编码。也就是说,可将对象分成S个块(即,数据块或S个部分),且使用所述S个块来计算奇偶校验块(即,奇偶校验部分)。如本文中其他地方所述,S个块及P个块中的每一者可存储在对应的装置中。

[0063] 当对一个对象进行复制的空间开销小于对所述对象进行纠删编码的空间开销时,所述对象可被视为小的对象。在此种情形中,由于复制提供更好的读取性能且可比相对复杂的纠删编码更好地处理更新,因此可期望使用复制。从应用元数据(application metadata)常常为小的观察结果来看,此也是合理的。换句话说,如果对象满足以下不等式(((P+1)*O=<(S+P)*m)),那么所述对象可被视为小的对象且可被复制。

[0064] 换句话说,当对象大小O乘以奇偶校验装置的数目加一(P+1)小于数据装置S的数目与奇偶校验装置P的数目的总和S+P乘以容许的最小大小值m时,可将对象视为小的对象。

[0065] 可存在一些其中对象可被分类为小的对象或大的对象的灰色区域。举例来说,当对象满足以下不等式(((P+1)*O>(S+P)*m)>S*m>0)时,所述对象可被视为中等对象,且可基于性能指标(例如,空间对(vs.)访问时间)及/或数据使用特性(例如,更新频率)来使用复制或纠删码。

[0066] 换句话说,当对象大小O乘以奇偶校验装置的数目加一(P+1)大于数据装置(S)的数目与奇偶校验装置(P)的数目的总和乘以容许的最小大小值m、数据装置(S)的数目与奇偶校验装置(P)的数目的总和乘以容许的最小大小值m大于数据装置的数目乘以容许的最小大小值、且数据装置的数目乘以容许的最小大小值大于对象大小O时,可将对象视为中等对象。

[0067] 举例来说,如果性能更为重要且对象被频繁更新,那么复制可为更好的选择。在此种情形中,可将中等对象分类为小的对象。举例来说,在以下不等式(((P+1)*O=<(S+P)*m)或((P+1)*O>(S+P)*m)且S*m>0,即如果((P+1)*O<=(S+P)*m或0<S*m))的情形中,根据示例性实施例可将对象分类为小的对象。

[0068] 另举例来说,如果空间效率更为重要,那么可使用纠删编码。在此种情形中,可将中等对象分类为大的对象。举例来说,在满足以下不等式(((P+1)*O>(S+P)*m且O>=S*m)或((P+1)*O>(S+P)*m)>S*m>0=((P+1)*O>(S+P)*m),即如果((P+1)*O>(S+P)*m))的情形中,根据示例性实施例可将对象分类为大的对象。

[0069] 图2是说明包括一组装置(固态驱动器1、固态驱动器2、固态驱动器3、固态驱动器4、固态驱动器5、固态驱动器6)的虚拟装置200以及在虚拟装置200中对对象(大的对象202及小的对象204)的存储的概念图。在所述装置中,固态驱动器1、固态驱动器2、固态驱动器3及固态驱动器4被配置成数据装置,且固态驱动器5及固态驱动器6是奇偶校验装置。尽管出于说明目的在图2中仅示出了四个数据装置固态驱动器1、固态驱动器2、固态驱动器3及固态驱动器4以及两个奇偶校验装置固态驱动器5及固态驱动器6,但虚拟装置200中数据装置及奇偶校验装置的数目并非仅限于此。此外,不同的固态驱动器(SSD)可被配置成数据装置及奇偶校验装置。

[0070] 举例来说,虚拟装置200可包括总共S个数据装置以及P个奇偶校验装置,所述装置

中的奇偶校验装置可为固定的或可旋转(因此,我们可参照图2看出使用示例性S值4及示例性P值2)。例如,当奇偶校验装置可旋转时,并非不同的大的对象的所有奇偶校验块(或奇偶校验部分)都可存储在相同的奇偶校验装置中,且一些数据装置可用作用于一个或多个大的对象的奇偶校验装置。换句话说,当奇偶校验装置是固定的时,对应于对象的"P"个奇偶校验块存储在存储器装置的同一集合"P"上,而当奇偶校验装置可旋转时,对应于对象的"P"个奇偶校验块没有必要存储在存储器装置的同一集合上。此外,所述装置可以平面方式或层级形式进行组织。在多个装置上扩展或复制的对象的开始装置可由密钥的散列值(hash value)确定。

[0071] 此外,根据需求及/或用户的设计选择可将数据装置重新配置成奇偶校验装置,或可将奇偶校验装置重新配置成数据装置。举例来说,虚拟装置200的装置的数目可基于可靠性目标进行配置。对于纠删编码来说,为了容忍P次故障,装置的总数目可为数据装置的数目(S)与奇偶校验装置的数目(P)的总和。对于复制来说,可容忍P次故障的装置的总数目可为P+1。装置的容量可为彼此相同或类似。

[0072] 根据本发明的示例性实施例,虚拟装置200中(或对应于虚拟装置200)的装置的集合构成作为可靠性管理的单元的群组。所述群组的装置(固态驱动器1、固态驱动器2、固态驱动器3、固态驱动器4、固态驱动器5及固态驱动器6)可存在于单个服务器或机架(rack)内、或存在于多个服务器或机架上,且所述装置可被结构化为具有层级架构或平面架构。

[0073] 包括所述一组装置的虚拟装置200可由被称为虚拟装置层210的层管理,使得所述一组装置可被呈现为单个虚拟装置。虚拟装置层210可为无状态的。虚拟装置层210可在运行时间缓存并保持装置的最小元数据信息,例如对象的数目及/或可用容量等。应注意,根据示例性实施例的虚拟装置层210不需要保持密钥信息(例如,没有针对密钥的映射)。虚拟装置200的容量可由所有装置容量中的最小装置容量(例如,图2中固态驱动器1、固态驱动器2、固态驱动器3、固态驱动器4、固态驱动器5及固态驱动器6的容量中的最少者)乘以群组中装置的数目来确定。

[0074] 虚拟装置层210可知晓每一装置可处理的最小值大小及最大值大小。虚拟装置层210可确定虚拟装置200的最小值大小及最大值大小。举例来说,根据示例性实施例,个别装置的所有最小值大小中的最大值(m_i)可被定义为虚拟装置200的最小值大小(m),而个别装置的所有最大值大小中的最小值(M_i)可被定义为虚拟装置200的最大值大小(M)。在其他实施例中,虚拟装置的最大值大小(M)可由个别装置的所有最大值大小中的最小值(M_i)乘以数据装置的数目(S)定义。

[0075] 根据本发明一些示例性实施例的虚拟装置200可利用所属领域中的技术人员已知的任意适当纠删编码算法,且可使用可用的最大距离可分(maximum distance separable,MDS)算法,例如里德-所罗门(Reed-Solomon,RS)码。如可在图2中所看到,将具有奇偶校验值二的纠删码(erasure code,EC)(即,纠删编码算法)应用至大的对象202,使得使用奇偶校验(Parity)1及奇偶校验2。

[0076] 根据示例性实施例,可将对象(例如,大的对象202)分成S个块并编码(具有相同大小并分布在数据装置及奇偶校验装置(即,S+P个装置)上)。举例来说,已经纠删编码的大的对象202可被分成数据(Data)1、数据2、数据3、数据4、奇偶校验1以及奇偶校验2。根据示例性实施例,对象占据的实际存储空间可被称为频带(band)。对于纠删编码来说频带可跨越S

+P个装置，而对于复制来说频带可跨越P+1个装置。举例来说，对小的对象204应用复制。频带可完全包含对象(即，整个对象可被存储在频带中)。在一些实施例中，频带可跨越其中存储有对象的S个块的S个装置。

[0077] 当对象大小未对齐到装置的分配或对齐单元时，可对频带中为对象分配的额外的空间进行填充(例如，可以0进行填充)。举例来说，图2示出大的对象202的数据4已被填充以0，以占据对存储数据4的所有数据位来说非必要的额外的空间。此外，根据本发明的实施例频带大小可为可变的。

[0078] 图3是根据本发明示例性实施例将对象写入虚拟装置(例如，图2及图4到图6所示的虚拟装置200)的流程图。图4是根据本发明示例性实施例说明以共享奇偶校验方式将大的对象242及244存储在虚拟装置200中的概念图。图5是根据本发明示例性实施例说明以专用奇偶校验方式将大的对象242及244存储在虚拟装置200中的概念图。图6是根据本发明示例性实施例说明将小的对象262及264存储在虚拟装置200中的概念图。

[0079] 如可在图3中看到，在方框300中，虚拟装置层(例如，图2及图4到图6所示的虚拟装置层210)接收(例如，自主机接收)指令或命令以利用密钥将大小为0的对象写入虚拟装置(例如，图2及图4到图6所示的虚拟装置200)。在其他实施例中，写入指令或命令可由虚拟装置层响应于由主机提供的写入指令产生。

[0080] 在方框302中，虚拟装置层使用上述不等式判断对象是否为大的对象。举例来说，当((P+1)*O>(S+P)*m且O>＝S*m)时，可将对象视为大的对象，其中O＝对象大小;P＝奇偶校验装置数目;S＝数据装置数目;且m＝容许的最小大小值(即，个别装置的所有最小值大小中的最大值)。

[0081] 当将对象分类为大的对象时，如在方框312中所示，虚拟装置层确定用于纠删编码的数据块的大小以及一个或多个数据块的填充(例如，以零进行填充)量。然后，将对象分成具有相同大小的S个块，考虑通过填充进行对齐，且然后如在方框314中所示，利用所属领域中的技术人员已知的适当的纠删编码算法从S个块产生(例如，计算)P个码块(即，P个奇偶校验块)。

[0082] 然后在方框316中，虚拟装置层基于分布策略(distribution policy)确定用于存储数据块及奇偶校验块的装置(即，S个装置及P个装置)。举例来说，所述分布策略可涉及通过密钥的散列值来确定对象的开始装置及/或将数据块及/或奇偶校验块存储在固定的装置及/或点上。在方框318中，将数据块及奇偶校验块写入对应的装置。举例来说，将S+P个块分布且存储在S+P个装置(例如，图2所示的固态驱动器1、固态驱动器2、固态驱动器3、固态驱动器4、固态驱动器5及固态驱动器6)中。对于图4所示的旋转奇偶校验装置来说，例如，在利用密钥的散列确定的装置处开始数据写入，且依次写入每一区块(即，块(以及奇偶校验区块，即奇偶校验块))，从第一装置上的第一数据开始。对于如图5所示的固定奇偶校验装置，例如，将所有数据区块及奇偶校验区块(即，块)存储在预先指定的装置中。此处，开始装置也被预先指定。对于小的对象，也通过使密钥散列而确定开始装置及复制装置。

[0083] 如可在图4中看到，奇偶校验装置可被共享(即，旋转)。换句话说，单个装置可根据被存储的大的对象而被用作用于存储数据块的数据装置或用于存储奇偶校验块的奇偶校验装置两者。举例来说，对象242可被分成数据1、数据2、数据3、数据4(被填充以0)、奇偶校验1以及奇偶校验2，且对象244也可被分成数据1、数据2、数据3、数据4(被填充以0)、奇偶校

验1以及奇偶校验2。可在图4中看到，在大的对象242的数据1、数据2、数据3以及数据4分别被存储在虚拟装置200的固态驱动器1、固态驱动器2、固态驱动器3以及固态驱动器4中时，大的对象244的数据1、数据2、数据3以及数据4分别被存储在虚拟装置200的固态驱动器6、固态驱动器1、固态驱动器2以及固态驱动器3中。

[0084] 此外，当对象242的奇偶校验1及奇偶校验2分别被存储在虚拟装置200的固态驱动器5以及固态驱动器6中时，对象244的奇偶校验1及奇偶校验2分别被存储在虚拟装置200的固态驱动器4以及固态驱动器5中。因此，当存在总共S个数据装置及P个奇偶校验装置时，所述奇偶校验装置可旋转，使得没有专用奇偶校验装置。

[0085] 不同于图4中所绘示的实例，图5说明利用专用奇偶校验装置(即，虚拟装置200的固态驱动器5及固态驱动器6)的实施方式。举例来说，大的对象242以及大的对象244两者的数据1、数据2、数据3、数据4(被填充以0)、奇偶校验1以及奇偶校验2分别被存储在虚拟装置200的固态驱动器1、固态驱动器2、固态驱动器3、固态驱动器4、固态驱动器5及固态驱动器6中。

[0086] 对于旋转奇偶校验实施例且对于小的对象来说，对象的开始装置可由密钥的散列值确定。举例来说，在图4所示的共享奇偶校验装置情形中，开始装置可通过Hash(key)％(S+P)进行确定。然后，后续的数据块及奇偶校验块(即，S+P个块)被依序写入(Hash(key)+1)％(S+P)、(Hash(key)+2)％(S+P)、...、(Hash(key)+S+P-1)％(S+P)。如果存在专用奇偶校验装置，那么使用S个装置代替(S+P)。

[0087] 在将数据块及奇偶校验块写入对应的装置后，在方框320中结束(完成)大的对象写入进程。

[0088] 当在方框302中未将对象确定为大的对象时，进程进入方框304，在方框304中，判断对象是否为小的对象(即，((P+1)*0＝<(S+P)*m)是否成立)。如果确定对象为小的对象，那么虚拟装置层继续执行复制并在方框308中基于分布策略确定利用哪些装置来存储数据及复制物。举例来说，所述分布策略可涉及通过密钥的散列值来确定对象的开始装置及/或将数据及/或复制物存储在固定的装置及/或点上。然后在方框310中，将数据及复制物写入对应的装置。

[0089] 根据示例性实施例，可为对象生成P+1个复制物(包括一个数据副本及P个奇偶校验副本)，考虑通过填充进行对齐，且所述P+1个复制物可被分布到P+1个装置上。如图6所示，例如，对象1 262被复制三次(包括数据及2个复制物)，且副本分别被存储在虚拟装置200的固态驱动器1、固态驱动器2以及固态驱动器3中。类似地，对象2 264被复制三次(包括数据及2个复制物)，且副本分别被存储在虚拟装置200的固态驱动器3、固态驱动器4以及固态驱动器5中。在图6所示的实例中，虚拟装置200包括总共S个数据装置以及P个奇偶校验装置。此外，由于对象1 262及对象2 264两者均为小的对象，因此在图6所示的实例中不使用纠删编码。

[0090] 可利用密钥的散列值在S+P个装置中选择主装置。可基于存储组织、及/或性能等确定地选择P个复制物。举例来说，当数据可被存储在主装置中时，复制物可被存储在(Hash(key)+1)％(S+P)、(Hash(key)+2)％(S+P)、...、(Hash(key)+P)％(S+P)上，或存储在不同的节点、机架上，而无论是否使用专用奇偶校验装置。

[0091] 现在返回图3，当在方框304中确定对象不是小的对象时(即，当对象既不是大的对

象(参见方框302)也不是小的对象(参见方框304)时),将对象确定为中等对象(即,(P+1)*0>(S+P)*m)>S*m>0),且进程进入方框306以判断是否将把中等对象视为小的对象。如果将把所述对象视为小的对象,那么进程进入方框308以起始小的对象存储进程,且如果将把所述对象视为大的对象,那么进程进入方框312以起始大的对象存储进程。

[0092] 图7说明根据本发明示例性实施例从虚拟装置(例如,图2及图4到图6所示的虚拟装置200)读取对象的进程。虚拟装置层(例如,图2及图4到图6所示的虚拟装置层210)不知道将读取的对象是小还是大,因为所述虚拟装置层不保存对象元数据,例如密钥及值大小。因此,虚拟装置层通过使用对象的用户密钥向所有实体装置(即,S+P个装置)发送读取请求而起始读取进程(700),其中如在方框702中所示向所有实体装置发送子读取请求。在方框704中,虚拟装置层从装置接收回应。当用户(例如,主机)请求的对象是大的对象时,如果没有错误(此在方框706中确定),那么所有S+P个装置都向具有用户密钥的请求返回相应的回应。

[0093] 举例来说,在没有错误时,如果将读取的对象是大的对象,那么所有装置(即,S+P个装置)将作出回应。然而,当N个装置具有错误时,那么仅有S+P-N个装置可作出回应。只要虚拟装置层接收具有相同大小的任意S个块(即,与数据块S的总数目相等的数据块S与奇偶校验块P的任意组合),那么便可重建用户对象。换句话说,只要不超过装置的奇偶校验数目(即,数目等于P的装置)发生故障,那么在大的对象的情形中数据便可被重建。

[0094] 如果所接收的块的总数目小于S或块的大小不相同,那么存在错误。在所有装置返回不存在(NON_EXIST)错误的情形中,可能是对不存在的对象的读取,或可能已发生了不可恢复的错误。

[0095] 最初虚拟装置层不知道对象类型,因此虚拟装置层将所述类型初始化为无(NONE)。当对象如在方框708中所确定是大的对象时,在方框718中确定类型。如果所述类型是无(NONE),那么在方框720中将对象类型设定为大。如果在方框718中没有将类型确定为无(NONE),那么在方框732中,虚拟装置层检查所述类型是否为大。如果在方框732中所述类型并非为大,那么如在方框734中所示确定错误。在方框720中将对象类型设定为大之后或如果在方框732中确定对象类型为大,那么在方框722中虚拟装置层判断其是否具有所有数据块。

[0096] 如果虚拟装置层确定已接收到所有数据块,那么如在方框730中所示结束(即,完成)读取进程。如果没有接收到所有数据块,那么虚拟装置层在方框724中判断是否已从所有装置接收到回应。如果所有装置均以作出回应,那么在方框726中虚拟装置层判断其是否具有数据的至少S个块(将所有数据块及奇偶校验块计算在内)。如果已接收到少于S个块,那么如在方框734中所示,虚拟机器层确定存在错误。如果已准确地接收到至少S个块(将所有所接收的数据块及奇偶校验块计算在内),那么虚拟装置层在方框728中利用纠删编码算法以S个块重建对象,且在方框730中读取进程结束。一个或多个装置有可能不作出回应,例如,在一个或多个装置意外离线的情形中。因此,在一些示例性实施例中,即使并非所有的装置均作出回应,只要已接收到至少S个块,虚拟装置层便如在方框728中所示继续进行重建对象。

[0097] 如果虚拟装置层在方框708中确定对象不是大的对象,那么进程进入方框710中以判断类型是否为无(NONE)。如果所述类型是无(NONE),那么在方框712中将对象类型设定为

小。如果所述类型不是无(NONE),那么在方框716中判断所述类型是否为小。此处,如果所述类型不是小,那么如在方框734中所示发现错误。在方框712中将对象类型设定为小之后或如果在方框716中虚拟装置层确定所述类型是小,那么在方框714中虚拟装置层判断所接收的块是否有效。如果所接收的块是有效的,那么如在方框730中所示结束(即,完成)读取进程。

[0098] 当用户(例如,主机)请求的对象是小的对象时,如果没有错误,那么具有复制物(即,主副本及复制物中的一者)的P+1个装置将返回,而其他的返回通知对象不存在的错误。只要在方框714中虚拟装置层接收到任意有效块,那么其便具有对象。如果所有装置均返回不存在(NON_EXIST)错误,那么便不存在此对象(或存在错误)。如果不是所有的装置均返回但所有返回的装置均报告不存在(NON_EXIST),那么如在方框734中所示已发生了不可恢复的错误。

[0099] 如果在方框714中虚拟装置层确定所述块是无效的,那么在方框724中判断是否已从所有装置接收到回应。如果没有从所有装置接收到回应,那么在方框704中虚拟装置层继续进行从所有装置取得回应,并继续所述进程以在方框706中判断是否存在错误等等,如图7所示。

[0100] 根据本发明的示例性实施例,在读取失败的情形中,虚拟装置层可要求每一装置列举所有的对象密钥并具有所有密钥的总序,用于在概念上进行再建。虚拟装置层可按次序逐个检查所述密钥。

[0101] 在不使用固定奇偶校验装置的情形中,如果所述对象是大的对象,那么虚拟装置层可使用Hash(key)确定密钥的开始装置,并基于开始装置信息确定应生成哪一块(数据块或码块)。在使用奇偶校验装置的情形中,哪一块必须被再建是显而易见的。类似于大的对象读取情形,利用有效块对用于新装置的块进行再建。

[0102] 如果所述对象是小的对象,那么虚拟装置层可使用Hash(key)确定密钥的主装置,并基于主装置信息确定哪些装置具有复制物。如果新装置必须具有复制物,那么将所述对象写入所述新装置。重复此过程直到已拜访了装置上的所有对象并再建了故障装置。

[0103] 因此,根据本发明的一个或多个示例性实施例,基于空间开销利用纠删编码与复制的无状态混合。此外,中等大小对象可例如基于访问图案而在纠删编码与复制之间切换。另外,每个对象的块大小是可变的。此外,可能没有必要进行因与其他对象共享空间而产生的读-修改-写。

[0104] 应理解,本文中所述的实施例应仅以阐述性意义进行考虑而并非用于限制目的。在每一实施例内对特征或方面的阐述通常应被视为可用于其它实施例中的其他类似特征或方面。尽管已参照图式阐述了一个或多个实施例,但所属领域中的普通技术人员应理解,在不背离由以上权利要求书及其等效范围所界定的精神及范围的条件下,可作出各种形式及细节上的变化。

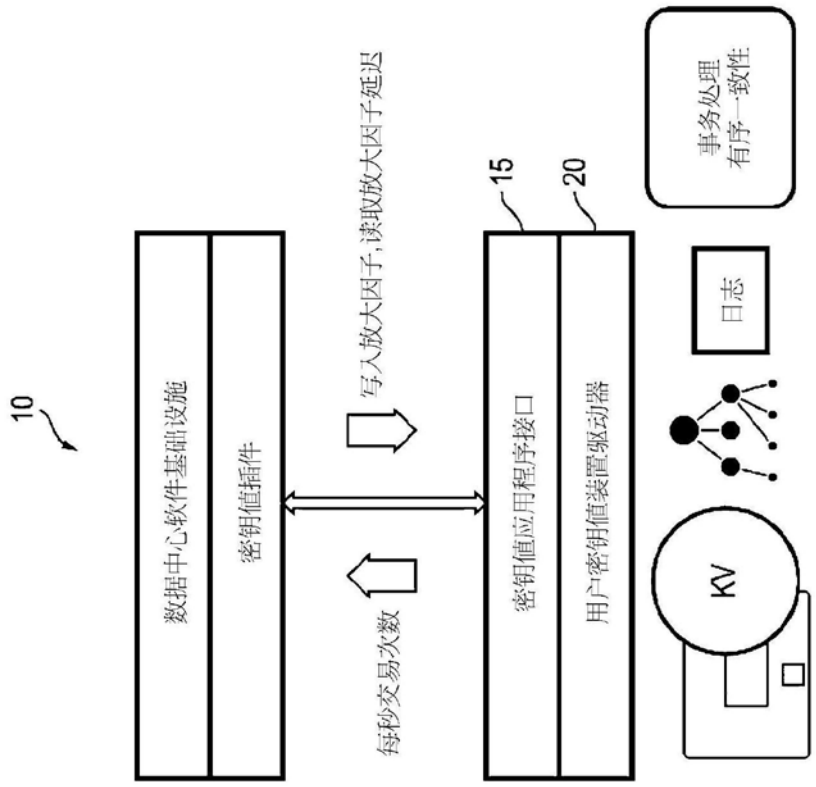

10
数据中心软件基础设施
密钥值插件
写入放大因子，读取放大因子延迟
每秒交易次数
15
20
密钥值应用程序接口
用户密钥值装置驱动器
事务处理
有序一致性
日志
KV

图1

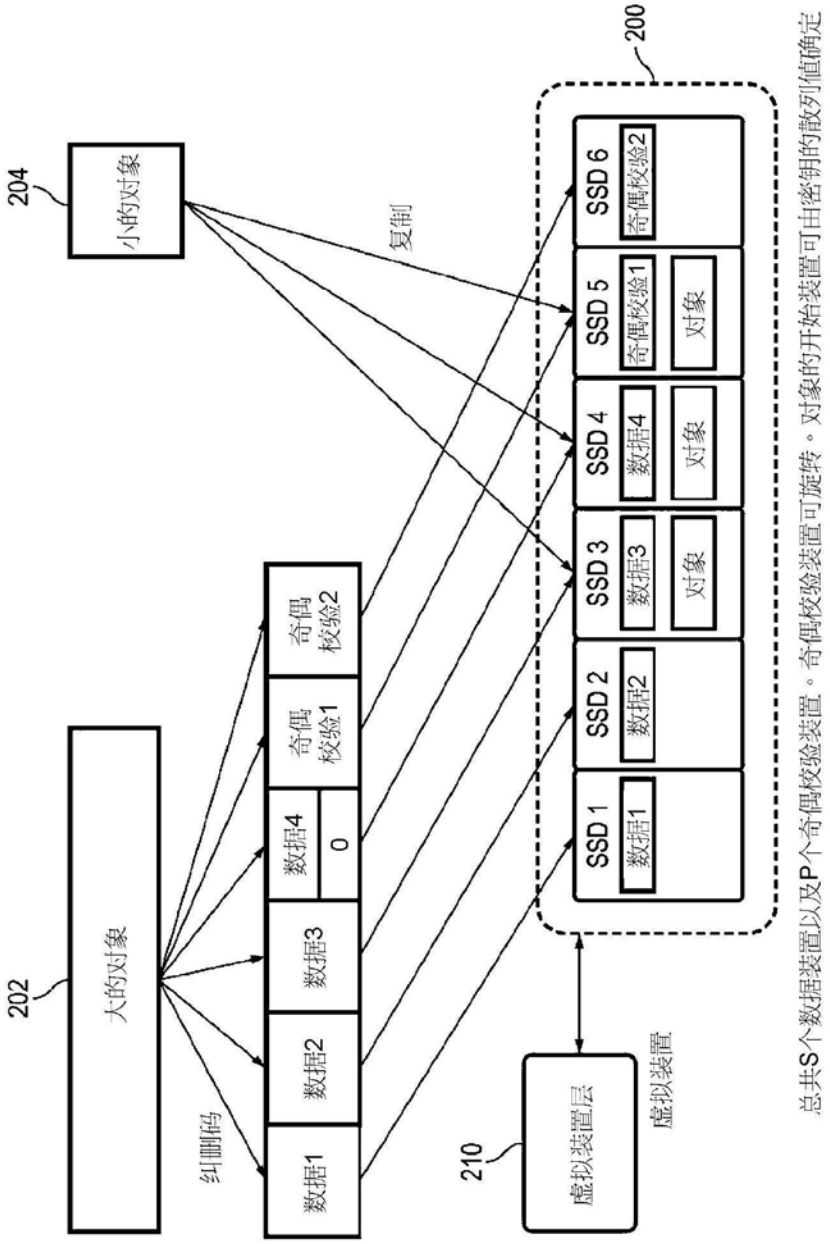

200
204
小的对象
复制
202
大的对象
纠删码
数据1
数据2
数据3
数据4
0
奇偶校验1
奇偶校验2
SSD 1
数据1
SSD 2
数据2
SSD 3
数据3
对象
SSD 4
数据4
对象
SSD 5
奇偶校验1
对象
SSD 6
奇偶校验2
210
虚拟装置层
虚拟装置
总共S个数据装置以及P个奇偶校验装置。奇偶校验装置可旋转。对象的开始装置可由密钥的散列值确定

图2

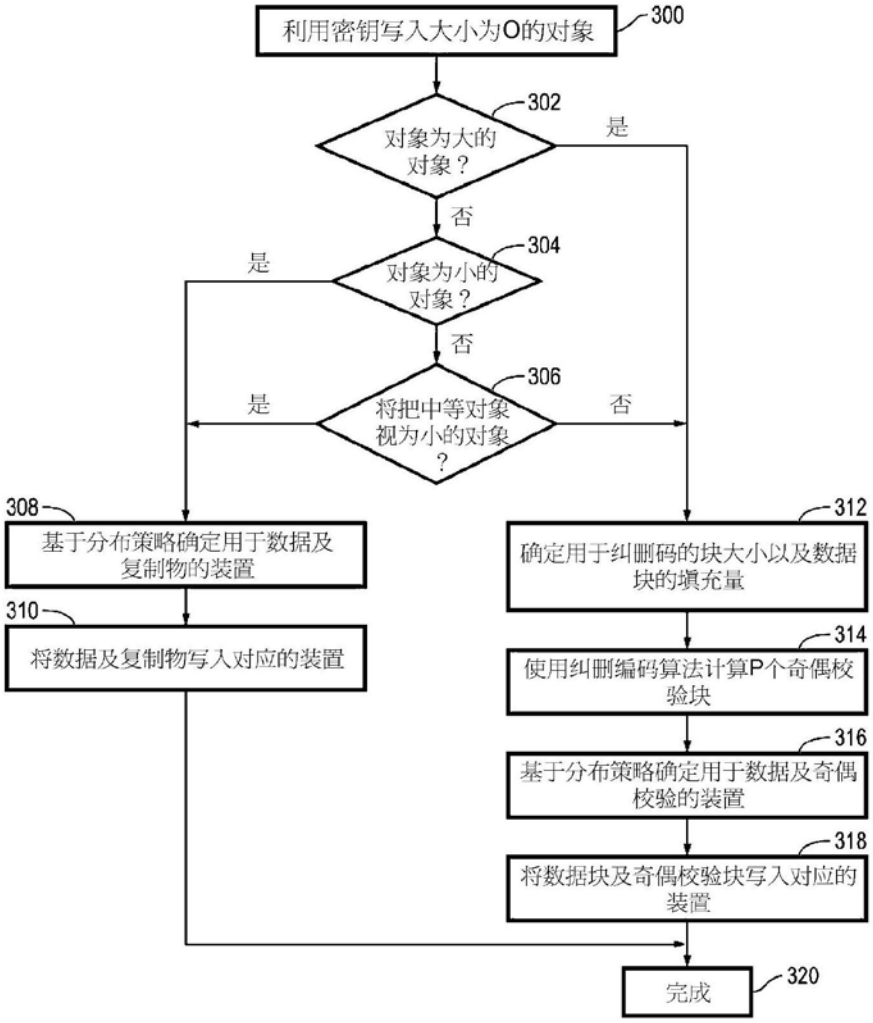

利用密钥写入大小为O的对象
300
对象为大的对象？
302
是
否
对象为小的对象？
304
是
否
将把中等对象视为小的对象？
306
是
否
基于分布策略确定用于数据及复制物的装置
308
将数据及复制物写入对应的装置
310
确定用于纠删码的块大小以及数据块的填充量
312
使用纠删编码算法计算P个奇偶校验块
314
基于分布策略确定用于数据及奇偶校验的装置
316
将数据块及奇偶校验块写入对应的装置
318
完成
320

图3

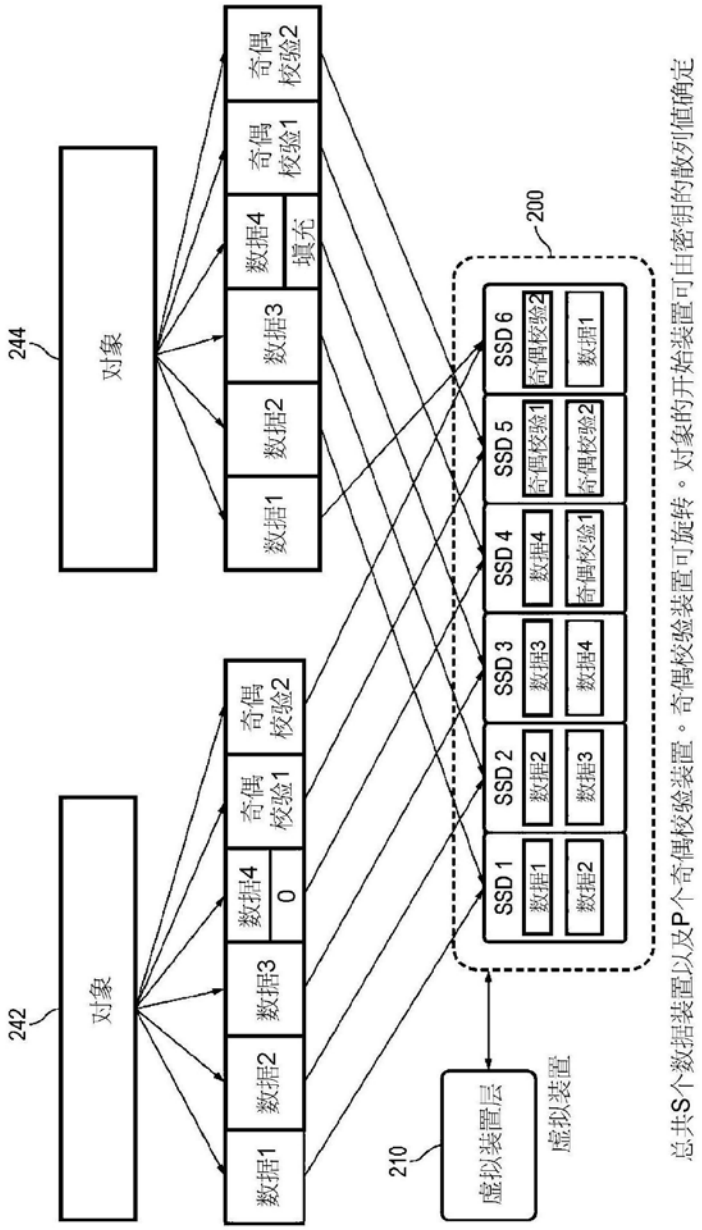

242
对象
数据1
数据2
数据3
数据4
0
奇偶校验1
奇偶校验2
244
对象
数据1
数据2
数据3
数据4
填充
奇偶校验1
奇偶校验2
200
SSD 1
数据1
数据2
SSD 2
数据2
数据3
SSD 3
数据3
数据4
SSD 4
数据4
奇偶校验1
SSD 5
奇偶校验1
奇偶校验2
SSD 6
奇偶校验2
数据1
210
虚拟装置层
虚拟装置
总共S个数据装置以及P个奇偶校验装置。奇偶校验装置可旋转。对象的开始装置可由密钥的散列值确定

图4

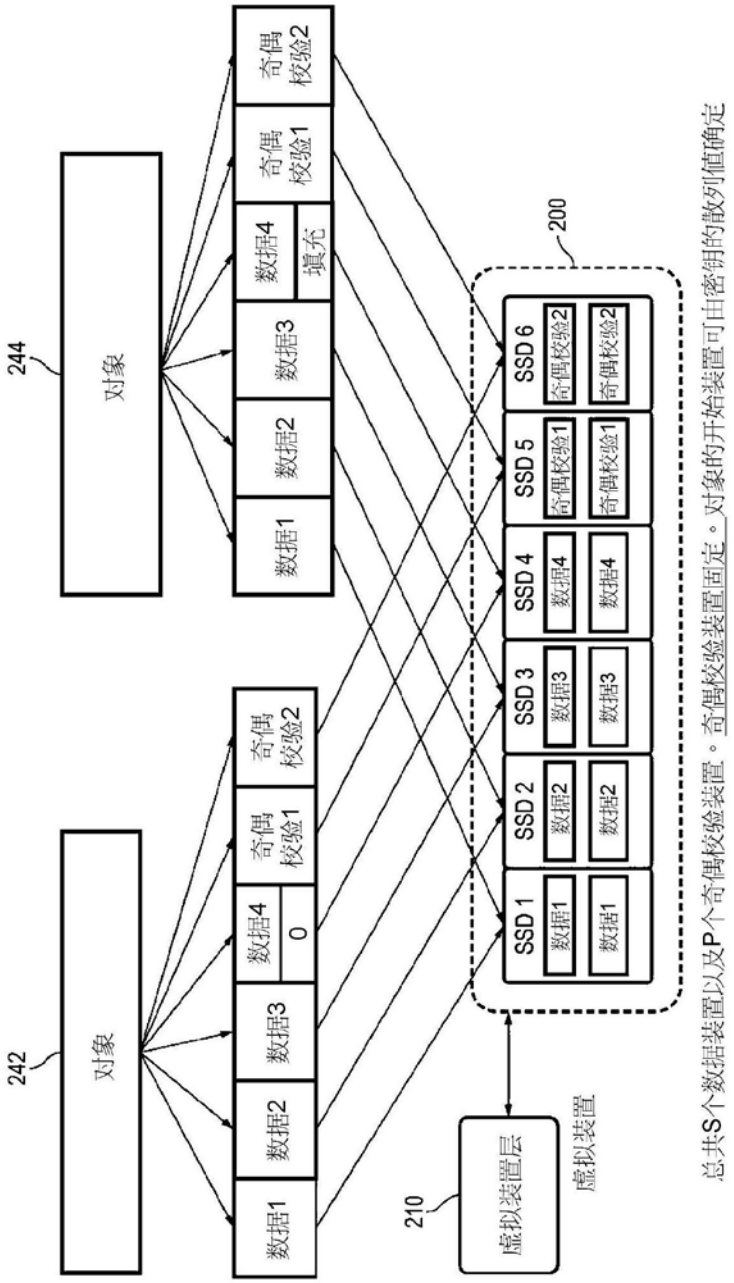

242
对象
数据1
数据2
数据3
数据4
0
奇偶校验1
奇偶校验2
244
对象
数据1
数据2
数据3
数据4
填充
奇偶校验1
奇偶校验2
200
SSD 1
数据1
数据1
SSD 2
数据2
数据2
SSD 3
数据3
数据3
SSD 4
数据4
数据4
SSD 5
奇偶校验1
奇偶校验1
SSD 6
奇偶校验2
奇偶校验2
210
虚拟装置层
虚拟装置
总共S个数据装置以及P个奇偶校验装置。奇偶校验装置固定。对象的开始装置可由密钥的散列值确定

图5

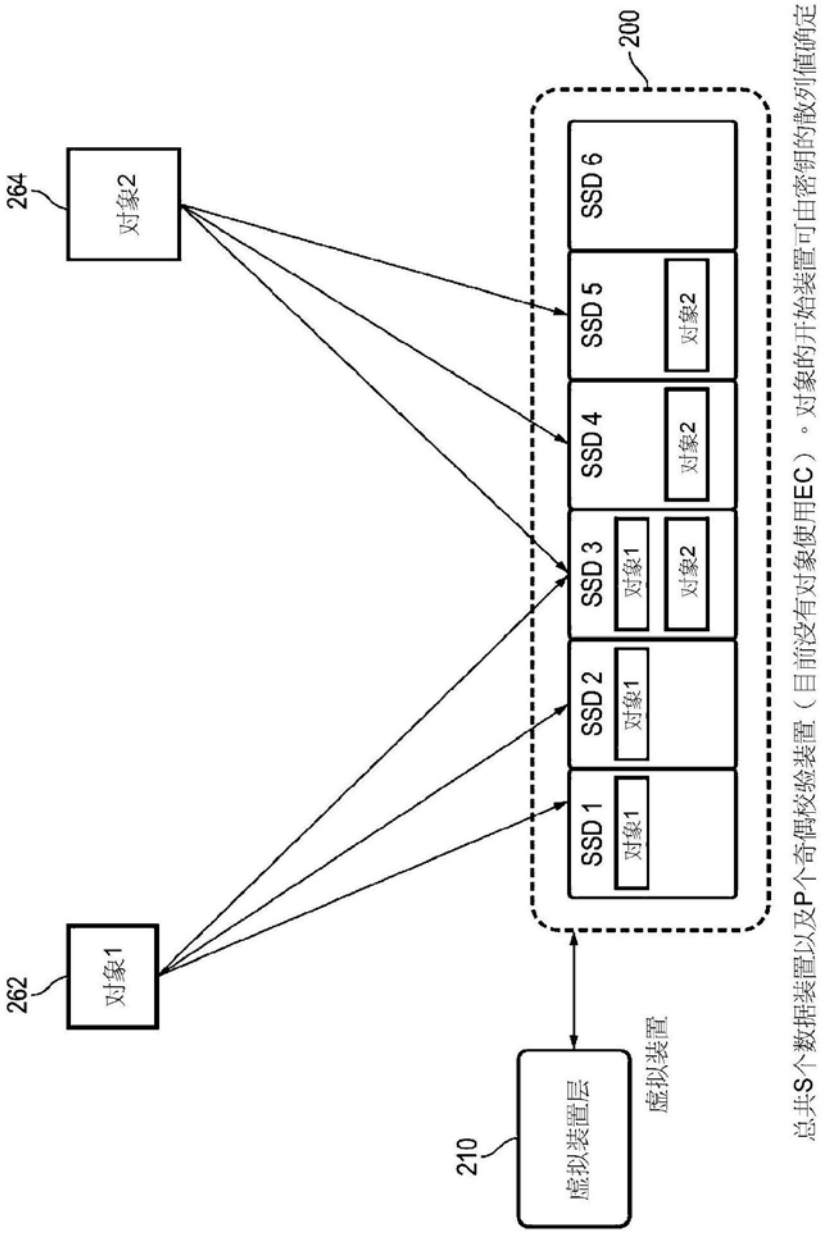

200
264
对象2
262
对象1
SSD 1
对象1
SSD 2
对象1
SSD 3
对象1
对象2
SSD 4
对象2
SSD 5
对象2
SSD 6
210
虚拟装置层
虚拟装置
总共S个数据装置以及P个奇偶校验装置（目前没有对象使用EC）。对象的开始装置可由密钥的散列值确定

图6

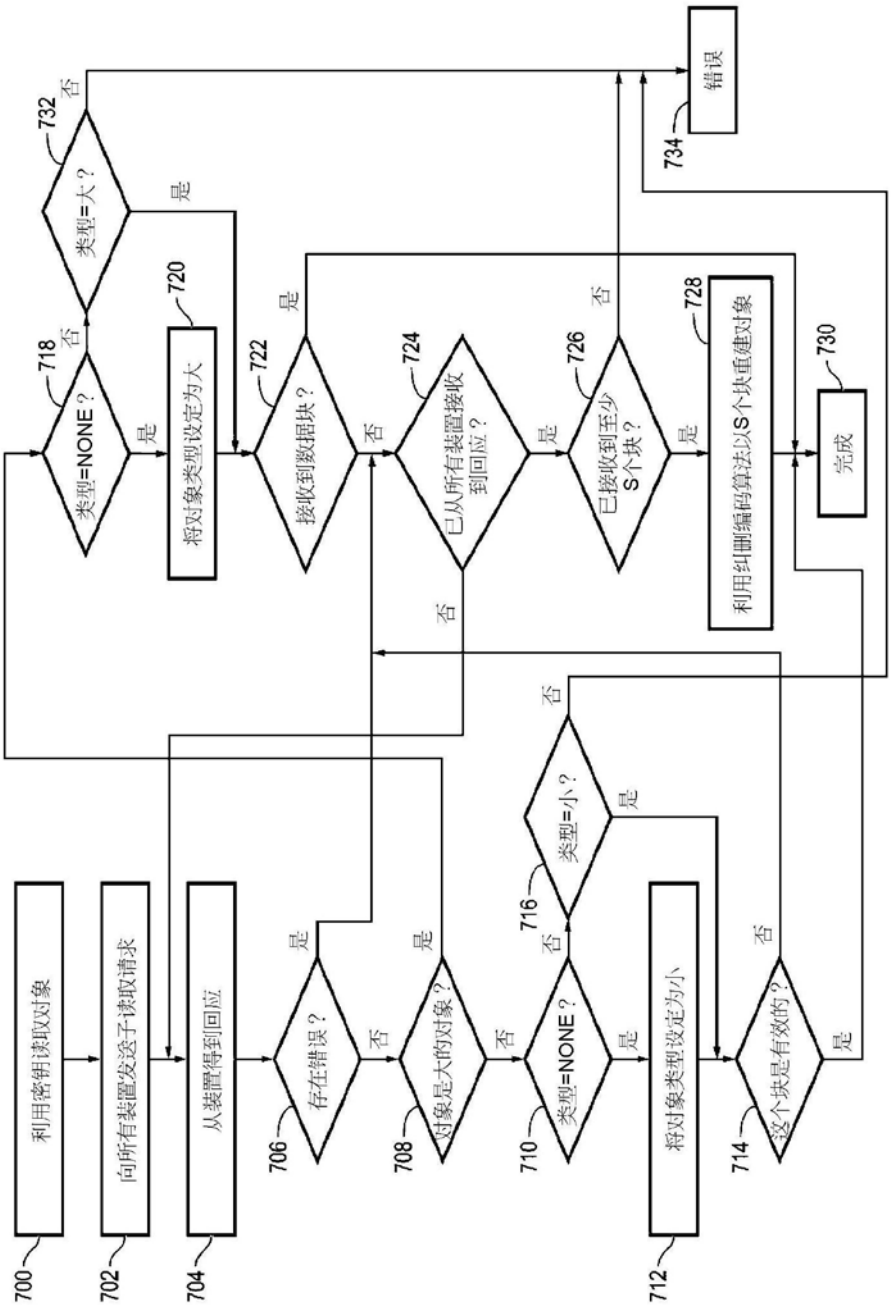

利用密钥读取对象
向所有装置发送子读取请求
从装置得到回应
存在错误？
对象是大的对象？
类型=NONE？
类型=小？
将对象类型设定为小
这个块是有效的？
类型=NONE？
将对象类型设定为大
类型=大？
接收到数据块？
已从所有装置接收到回应？
已接收到至少S个块？
利用纠删编码算法以S个块重建对象
完成
错误
700
702
704
706
708
710
712
714
716
718
720
722
724
726
728
730
732
734
是
否

图7