

[19] 中华人民共和国国家知识产权局



[12] 发明专利申请公布说明书

[21] 申请号 200780050973.4

[43] 公开日 2010 年 1 月 6 日

[51] Int. Cl.
G06F 12/08 (2006.01)
G06F 3/06 (2006.01)

[11] 公开号 CN 101622606A

[22] 申请日 2007.12.6

[21] 申请号 200780050973.4

[30] 优先权

[32] 2006.12.6 [33] US [31] 60/873,111

[32] 2007.9.22 [33] US [31] 60/974,470

[86] 国际申请 PCT/US2007/025049 2007.12.6

[87] 国际公布 WO2008/070173 英 2008.6.12

[85] 进入国家阶段日期 2009.8.6

[71] 申请人 弗森多系统公司 (dba 弗森 - 艾奥)

地址 美国犹他州

[72] 发明人 大卫·弗林 约翰·斯特拉瑟

乔纳森·撒切尔 迈克尔·扎佩

[74] 专利代理机构 北京安信方达知识产权代理有限公司

代理人 韩龙 阎斌斌

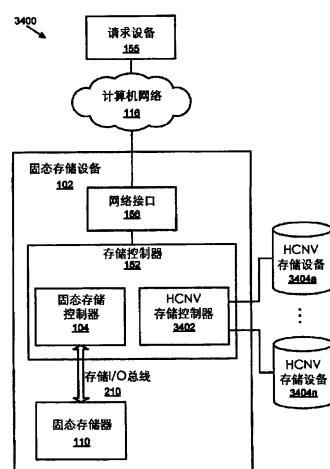
权利要求书 4 页 说明书 107 页 附图 20 页

[54] 发明名称

用于作为大容量、非易失性存储器的高速缓存的固态存储器的装置、系统和方法

[57] 摘要

本发明公开了用于作为大容量、非易失性存储设备的高速缓存的固态存储器的装置、系统、和方法。所述装置、系统、和方法具有有多个模块，包括高速缓存前端模块和高速缓存后端模块。高速缓存前端模块管理与存储请求并联的数据传送。数据传送在请求设备和作为一个或多个 HCNV 存储设备的高速缓存的固态存储器之间进行，并且数据传送可包括一个或多个数据、元数据、和元数据索引。固态存储器可包括非易失性的，固态的数据存储元件的阵列。高速缓存后端模块管理固态存储器与一个或多个 HCNV 存储设备之间的数据传送。



1、一种用于管理一个或多个大容量、非易失性（“HCNV”）存储设备上的数据存储的装置，所述装置包括：

高速缓存前端模块，用于管理与存储请求关联的数据传送，所述数据传送在请求设备和作为一个或多个HCNV存储设备的高速缓存的固态存储器之间进行，所述数据传送包括一个或多个数据、元数据、和元数据索引，所述固态存储器包括非易失性的、固态的数据存储元件的阵列；和

高速缓存后端模块，用于管理所述固态存储器与所述一个或多个HCNV存储设备之间的数据传送。

2、如权利要求1所述的装置，其中，所述高速缓存前端模块和所述高速缓存后端模块与管理所述固态存储器的固态存储控制器共处在一起。

3、如权利要求2所述的装置，其中，所述高速缓存前端模块、所述高速缓存后端模块和所述固态存储控制器独立于请求设备自主运行。

4、如权利要求1所述的装置，其中，所述固态存储控制器还包括对象存储控制器模块，该对象存储控制器模块服务来自一个或多个请求设备的对象请求并且管理所述固态存储器内的所述对象请求的对象。

5、如权利要求1所述的装置，还包括HCNV RAID模块，用于将固态存储器中缓存的数据存储在与RAID级别一致的独立驱动器冗余阵列（“RAID”）中的两个或更多个HCNV存储设备，其中数据对于请求设备作为整体呈现。

6、如权利要求1所述的装置，其中，所述固态存储器和所述一个或多个HCNV存储设备包括配置为RAID群组的混合存储设备集内的混合存储设备，其中，在所述固态存储器中缓存并且随后存储在HCNV设备上的数据段包括条带的N个数据段之一或者所述条带的奇偶校验数据段，其中混合存储设备从一个或多个客户端接收独立于RAID条带的数据段的存储请求。

7、如权利要求6所述的装置，其中，所述混合存储设备是共享的、前端分布式RAID群组中的存储设备，该存储设备从两个或更多个客户端接收两个或更多个同时的存储请求。

8、如权利要求1所述的装置，其中，所述HCNV存储设备是硬盘驱动器（“HDD”）、光盘驱动器、和磁带存储器之一。

9、如权利要求1所述的装置，其中，所述固态存储器和所述一个或多个HCNV存储设备包括混合存储设备，并且还可包括标准设备模拟模块，该标准设备模拟模块通过在一个或多个请求设备加载所述混合存储设备的操作的专用代码之前，模拟附属于一个或多个请求设备的标准设备，提供对混合存储设备的访问，所述标准设备由工业标准的BIOS来支持。

10、如权利要求1所述的装置，其中，所述固态存储设备可分区为两个或更多个区域，其中一个或多个区域可被用作为独立于作为所述HCNV存储设备的高速缓存的固态存储器的固态存储器。

11、如权利要求1所述的装置，其中，一个或多个客户端向所述高速缓存前端模块和所述高速缓存后端模块发送高速缓存控制消息，以管理存储在所述固态存储设备和所述一个或多个HCNV存储设备内的一个或多个文件或对象的状态。

12、如权利要求11所述的装置，其中高速缓存控制消息包括下面的一个或多个：

使得所述高速缓存后端模块扣牢固态存储器中的对象或文件的一部分的控制消息；

使得所述高速缓存后端模块释放固态存储器中的对象或文件的一部分的控制消息；

使得所述高速缓存后端模块将来自固态存储器的对象或文件的一部分清洗到所述一个或多个HCNV存储设备的控制消息；

使得所述高速缓存后端模块从所述一个或多个HCNV存储设备向固态存储器预加载对象或文件的一部分的控制消息；

使得所述高速缓存后端模块将来自所述固态存储器的一个或多个对象或文件的一部分或多个部分卸载到所述一个或多个HCNV存储设备，以便释放所述固态存储器中的预定量存储空间的控制消息。

13、如权利要求11所述的装置，其中，所述高速缓存控制消息通过所述对象或文件的元数据（“高速缓存控制元数据”）来传送。

14、如权利要求13所述的装置，其中，所述高速缓存控制元数据是持久的。

15、如权利要求13所述的装置，其中，所述高速缓存控制元数据在创建

所述文件或对象时通过属性集来建立。

16、如权利要求13所述的装置，其中，所述高速缓存控制元数据从文件或对象管理系统获得。

17、如权利要求1所述的装置，还包括易失性高速缓存存储元件，并且其中所述高速缓存前端模块和所述高速缓存后端模块还包括在易失性高速缓存存储元件中存储数据并且管理固态存储器和易失性高速缓存存储元件中存储的数据，并且其中所述后端存储模块还管理所述易失性高速缓存存储元件、所述固态存储器和所述HCNV存储设备之间的数据传送。

18、如权利要求17所述的装置，其中，在所述固态存储设备和所述易失性高速缓存存储元件内保存存储在所述HCNV存储设备中的对象和文件的一个或多个元数据和索引元数据。

19、如权利要求1所述的装置，其中，所述HCNV存储设备中存储的对象和文件的一个或多个元数据和索引元数据保存在固态存储设备内。

20、如权利要求1所述的装置，其中，所述固态存储器和所述一个或多个HCNV存储设备包括存储设备，以使得对连接到所述存储设备的客户端来说隐藏了所述HCNV存储设备。

21、一种用于管理一个或多个大容量、非易失性（“HCNV”）存储设备上的数据存储的系统，所述系统包括：

 固态存储器，该固态存储器包括非易失性的、固态的数据存储元件的阵列；

 一个或多个HCNV存储设备；和

 存储控制器，包括：

 固态存储控制器；

 HCNV存储设备控制器；

 高速缓存前端模块，用于管理与存储请求关联的数据传送，所述数据传送通常在请求设备和作为一个或多个HCNV存储设备的高速缓存的固态存储器之间进行，所述数据传送包括一个或多个数据、元数据、和元数据索引；和

 高速缓存后端模块，用于管理固态存储器和一个或多个HCNV存储设备之间的数据传送。

22、如权利要求21所述的系统，还包括连接到所述存储控制器的网络接口，所述网络接口通过计算机网络促进所述请求设备和所述固态存储控制器之间的数据传送。

23、如权利要求1所述的系统，还包括服务器，其中所述服务器包括所述固态存储器、所述一个或多个HCNV存储设备和所述存储控制器。

24、如权利要求1所述的系统，其中，所述一个或多个HCNV存储设备通过存储区域网络（“SAN”）连接到所述存储控制器。

25、一种包括计算机可读介质的计算机程序产品，该计算机可读介质具有可执行以完成操作的计算机可用程序代码，所述操作用于管理一个或多个大容量、非易失性（“HCNV”）存储设备上的数据存储，所述计算机程序产品的操作包括：

管理与存储请求关联的数据传送，所述数据传送在请求设备和作为一个或多个HCNV存储设备的高速缓存的固态存储器之间进行，所述数据传送包括一个或多个数据、元数据、和元数据索引，所述固态存储器包括非易失性的、固态的数据存储元件的阵列；和

管理所述固态存储器和所述一个或多个HCNV存储设备之间的数据传送。

用于作为大容量、非易失性存储器的高速缓存的固态存储器的装置、系统和方法

发明背景

相关申请的交叉引用

本申请要求下述申请的优先权：David Flynn等人于2006年12月6日提交的题为“Elemental Blade System”的美国临时专利申请（申请号为：60/873,111）；David Flynn等人于2007年9月22日提交的题为“Apparatus, System, and Method for Object-Oriented Solid-State Storage”的美国临时专利申请（申请号为：60/974,470）。上述申请通过引用并入本文中。

技术领域

本发明涉及管理数据，更具体地，涉及使用作为大容量、非易失性存储设备的高速缓存的固态存储器。

背景技术

通常，高速缓存是有利的，因为经常存取的或者作为应用程序或操作系统的一部分载入的数据可存储在高速缓存中，相比于必须通过大容量、非易失性（“HCNV”）存储设备访问数据的情况，后续的存取操作更快速，所述大容量、非易失性存储设备例如硬盘驱动器（“HDD”）、光盘驱动器、磁带存储器等。高速缓存通常包括在计算机内。

发明内容

某些存储设备和系统在HCNV存储设备中包括高速缓存。某些存储设备包含非易失性固态高速缓存；这些提供了减少访问时间的好处，但是仅仅可提供与HCNV存储设备接口的通常受限能力一致的性能。存在通常位于主板上的

某些非易失性固态高速缓存存储设备；这些设备不能用于多客户端环境中，因为没有提供高速缓存一致性。某些HCNV设备的控制器也包括高速缓存。在多个客户端共享冗余HCNV高速缓存控制器的情况下，需要复杂的高速缓存一致性算法来确保不破坏数据。

通常，在DRMA中实现高速缓存，得到额外的高速缓存能力，并且需要相对高的性能功率比。如果支持非易失性高速缓存的功率失去，高速缓存中存储的数据丢失。通常，某些后备电池用于避免功率故障时的数据丢失，在后备电池故障之前，有足够的能力将高速缓存清洗到非易失性存储器。另外，后备电池系统消耗功率，需要冗余，消极地影响可靠性并且占据空间。电池也必须基于规则来服务并且后备电池相对昂贵。

如上所述，显而易见，存在使用作为高速缓存的固态存储器管理数据的装置、系统和方法的需求。有利地是，这种装置、系统和方法提供了消耗很少功率、提供显著更大的能力并且不需要后备电池来保持高速缓存中存储的数据的非易失性高速缓存。

本发明是针对现有技术的现况开发出来的，具体地，是针对现有技术中通过现有的管理数据存储的系统并未完全解决的问题和需要。因此，本发明已经被开发出来以提供克服现有技术中的上述多数或全部缺陷的管理一个或多个大容量、非易失性（“HCNV”）存储设备上的数据存储的装置、系统和方法。

在一种实施方式中，装置具有包括高速缓存前端模块和高速缓存后端模块的多个模块。高速缓存前端模块管理与存储请求关联的数据传送。所述数据传送在请求设备和作为一个或多个HCNV存储设备的高速缓存的固态存储器之间进行，并且所述数据传送可包括一个或多个数据、元数据和元数据索引。固态存储器可包括非易失性的，固态的数据存储元件的阵列。高速缓存后端模块管理固态存储器与一个或多个HCNV存储设备之间的数据传送。

在装置的一种实施方式中，高速缓存前端模块和高速缓存后端模块与管理固态存储器的固态存储控制器共处在一起。在另一种实施方式中，高速缓存前端模块、高速缓存后端模块和固态存储控制器独立于请求设备运行。

在一种实施方式中，装置包括HCNV RAID模块，该模块将固态存储器中缓存的数据存储在与RAID级别一致的独立驱动器冗余阵列（“RAID”）中的

两个或更多个HCNV存储设备。数据对于请求设备作为整体呈现。在另一种实施方式中，固态存储器和一个或多个HCNV存储设备可包括配置为RAID群组的混合存储设备组内的混合存储设备。固态存储器中缓存并且随后存储在HCNV设备上的数据段可包括条带的N个数据段之一或者该条带的奇偶校验数据段。混合存储设备通常从一个或多个客户端接收独立于RAID条带的数据段的存储请求。在另一种实施方式中，混合存储设备可以是共享的、前端分布式RAID群组中的存储设备，该存储设备从两个或更多个客户端接收两个或更多个同时的存储请求。

在装置的另外实施方式中，HCNV存储设备可以是硬盘驱动器(“HDD”)、光盘驱动器或磁带存储器。在另一种实施方式中，固态存储器和一个或多个HCNV存储设备可以是混合存储设备。在一种实施方式中，装置还可包括标准设备模拟模块，该标准设备模拟模块通过在一个或多个请求设备加载所述混合存储设备的操作的专用代码之前，模拟附属于一个或多个请求设备的标准设备，提供对混合存储设备的访问。标准设备通常由工业标准的BIOS来支持。

在另一种实施方式中，固态存储设备可分区为两个或更多个区域，其中一个或多个分区可被用作为独立于作为HCNV存储设备的高速缓存的固态存储器的固态存储器。在又另一种实施方式中，一个或多个客户端向高速缓存前端模块和高速缓存后端模块发送高速缓存控制消息，以管理存储在固态存储设备和一个或多个HCNV存储设备内的一个或多个文件或对象的状态。

在装置的一种实施方式中，高速缓存控制消息可包括一个或多个控制消息。控制消息的各个实施方式可包括使得高速缓存后端模块扣牢固态存储器中的对象或文件的一部分的控制消息，或者使得高速缓存后端模块释放固态存储器中的对象或文件的一部分的控制消息。控制消息的其他实施方式可包括使得高速缓存后端模块将来自固态存储器的对象或文件的一部分清洗到一个或多个HCNV存储设备的控制消息，或者使得高速缓存后端模块从一个或多个HCNV存储设备向固态存储器预加载对象或文件的一部分的控制消息。控制消息的又另一种实施方式可以是使得高速缓存后端模块将来自固态存储器的一个或多个对象或文件的一部分或多部分卸载到一个或多个HCNV存储设备，以便释放固态存储器中的预定量存储空间的控制消息。在一种实施方式中，高速缓存控制消息通过对象或文件的元数据(“高速缓存控制元数据”)来

传送。在又另一种实施方式中，高速缓存控制元数据可以是持久的。在另一种实施方式中，高速缓存控制元数据在创建所述文件或对象时通过属性集来建立。在又另一种实施方式中，高速缓存控制元数据可从文件或对象管理系统获得。

在装置的一种实施方式中，装置可包括易失性高速缓存存储元件，其中高速缓存前端模块和高速缓存后端模块在易失性高速缓存存储元件中存储数据并且管理固态存储器和易失性高速缓存存储元件中存储的数据。后端存储模块还可管理易失性高速缓存存储元件、固态存储器和HCNV存储设备之间的数据传送。在另一种实施方式中，可在固态存储设备和易失性高速缓存存储元件内保存存储在HCNV存储设备中的对象和文件的元数据和/或索引元数据。

在装置的另外实施方式中，HCNV存储设备中存储的对象和文件的元数据和/或索引元数据可保存在固态存储设备内。在另一种实施方式中，固态存储器和一个或多个HCNV存储设备可包括存储设备，以使得对连接到所述存储设备的客户端来说隐藏了所述HCNV存储设备。

还提出了本发明的一种系统。系统大体上包括上面关于装置描述的模块和实施方式。在一种实施方式中，系统包括固态存储器，该固态存储器包括非易失性的，固态的数据存储元件的阵列。系统还包括一个或多个HCNV存储设备和存储控制器。在一种实施方式中，存储控制器可包括固态存储控制器和HCNV存储设备控制器。存储控制器还可包括高速缓存前端模块和高速缓存后端模块。高速缓存前端模块管理与存储请求关联的数据传送。数据传送通常在请求设备和作为一个或多个HCNV存储设备的高速缓存的固态存储器之间进行。数据传送可包括一个或多个数据、元数据、和元数据索引。高速缓存后端模块管理固态存储器和一个或多个HCNV存储设备之间的数据传送。

在一种实施方式中，系统包括连接到存储控制器的网络接口，其中网络接口通过计算机网络促进请求设备和固态存储控制器之间的数据传送。在另一种实施方式中，系统包括服务器，该服务器包括固态存储器、一个或多个HCNV存储设备、和存储控制器。在又另一种实施方式中，一个或多个HCNV存储设备通过存储区域网络（“SAN”）连接到存储控制器。

还提出了本发明的一种方法，用于在多个主机之间共享设备。在公开的

实施方式中的该方法大体上包括实现上述与所述装置和系统的运行有关的功能的必要步骤。在一种实施方式中，该方法包括管理与存储请求关联的数据传送，其中数据传送在请求设备和用作为一个或多个HCNV存储设备的高速缓存的固态存储器之间进行。数据传送可包括一个或多个数据、元数据、和元数据索引。固态存储器可包括非易失性的、固态的数据存储元件的阵列。方法还可包括管理固态存储器和一个或多个HCNV存储设备之间的数据传送。

本说明书全文所提到的特征、优点或者类似措辞并不意味着可在本发明包含在本发明的任一单独的实施方式中的情况下实现所有的特征和优点。当然，涉及特征和优点的措辞被理解为意味着：与实施方式一起描述的特定的特征、优点或者特点包括在本发明的至少一种实施方式中。因此，在本说明书中，关于特征、优点和类似措辞的讨论可（但未必）涉及同一实施方式。

此外，描述的本发明的特征、优点和特点可采用任何合适的方式与一个或多个实施方式结合。相关领域的技术人员可意识到本发明可在不具备特定实施方式的一个或多个具体特征或优点的情况下被实施。在其他例子中，可意识到附加特征和优点出现在某些实施方式中，而不是在本发明的所有实施方式中都出现。

通过下面的说明和附加的权利要求，本发明的这些特征和优点将变得更加充分的显而易见，或者可以通过按下文所阐述的实施本发明的方法而获悉。

附图说明

为了使本发明的优点更加容易理解，会参考附图中示出的特定实施方式给出上面简要描述的本发明的更具体的说明。在理解到这些附图仅描述了本发明的一般实施方式并且并不因此认为本发明限于此范围的情况下，将通过使用附图并结合更多的具体特征和细节描述和解释本发明，附图中：

图1A是示意性框图，示出了根据本发明的用于固态存储设备内的数据管理的系统的一种实施方式；

图1B是示意性框图，示出了根据本发明的用于存储设备内的对象管理的系统的一种实施方式；

图1C是示意性框图，示出了根据本发明的用于服务器内存存储区域网络的

系统的一种实施方式；

图2A是示意性框图，示出了根据本发明的用于存储设备内的对象管理的装置的一种实施方式；

图2B是示意性框图，示出了根据本发明的位于固态存储设备内的固态存储设备控制器的一种实施方式；

图3是示出了根据本发明的位于固态存储设备内的固态存储设备控制器的一种实施方式的示意性框图，该固态存储设备控制器具有写入数据管道和读取数据管道；

图4A是示意性框图，示出了根据本发明的位于固态存储控制器内的内存库交错控制器的一种实施方式；

图4B是示意性框图，示出了根据本发明的位于固态存储控制器内的内存库交错控制器的一种替代实施方式；

图5A是示意性流程图，示出了根据本发明的在固态存储设备内采用数据管道管理数据的方法的一种实施方式；

图5B是示意性流程图，示出了根据本发明的用于服务器内SAN的方法的一种实施方式；

图6是示意性流程图，示出了根据本发明的在固态存储设备内采用数据管道管理数据的方法的再一种实施方式；

图7是示意性流程图，示出了根据本发明的在固态存储设备内采用内存库交错管理数据的方法的一种实施方式；

图8是示意性框图，示出了根据本发明的用于固态存储设备中无用存储单元收集的装置的一种实施方式；

图9是示意性流程图，示出了根据本发明的用于固态存储设备中无用存储单元收集的方法的一种实施方式；

图10是示意性框图，示出了根据本发明的用于渐进式RAID、前端分布式RAID、和共享的前端分布式RAID的系统的一种实施方式；

图11是示意性框图，示出了根据本发明的用于前端分布式RAID的装置的一种实施方式；

图12是示意性流程图，示出了根据本发明的用于前端分布式RAID的方法的一种实施方式；

图13是示意性框图，示出了根据本发明的用于共享的前端分布式RAID的装置的一种实施方式；

图14是示意性流程图，示出了根据本发明的用于共享的前端分布式RAID的方法的一种实施方式；

图15是示意性框图，示出了根据本发明的具有固态存储器的系统的一种实施方式，该固态存储器作为大容量、非易失性存储设备的高速缓存；

图16是示意性框图，示出了根据本发明的具有固态存储器的装置的一种实施方式，该固态存储器作为大容量、非易失性存储设备的高速缓存；并且

图17是示意性流程图，示出了根据本发明的具有固态存储器的方法的一种实施方式，该固态存储器作为大容量、非易失性存储设备的高速缓存。

具体实施方式

为了更显著地强调功能性单元运行的独立性，在本说明书中描述的许多功能性单元已被标示为模块。例如，模块可作为硬件电路来实施，所述硬件电路包括自定义VLSI电路、门阵列或成品半导体（例如逻辑芯片、晶体管或其他分立元件）。模块也可在可编程硬件设备（如现场可编程门阵列、可编程阵列逻辑、可编程逻辑设备或类似设备）内实施。

模块还可在由不同类型的处理器运行的软件中实施。例如，可执行代码的识别模块可以包括一个或多个计算机指令物理块或逻辑块，该计算机指令被作为对象、程序或函数来组织。然而，识别模块的可执行文件不必在物理上位于一起，但是可包括存储在不同位置的不同命令，当这些命令在逻辑上连接在一起时，所述命令包括所述模块并实现所述模块的指定目标。

当然，可执行代码的模块可以为一个或许多指令，并且甚至可以分布在若干不同的代码段中、分布在不同的程序中并可分布在多个存储设备中。类似地，可以在此在模块内识别并示出运算数据，并且可以以任何合适的形式体现所述运算数据并在任意合适类型的数据结构中组织所述运算数据。所述运算数据可作为单数据集收集，或者可以分布在不同的位置（包括不同的存储设备），并且可在系统或网络中至少部分地仅作为电信号存在。当模块或模块的部分在软件中实施时，软件部分被存储在一个或多个计算机可读媒体上。

本说明书全文所提到的“一种实施方式”、“实施方式”或类似的措辞意味着与实施方式一起描述的特定的特征、结构或特点包括在本发明的至少一种实施方式中。因此，在本说明书全文中，短语“在一种实施方式中”、“在实施方式中”及类似措辞的出现可（但未必）涉及同一实施方式。

提及信号承载媒介可采取任何能够生成信号、导致信号生成或者导致在数字处理设备上执行机器可读命令程序的形式。信号承载媒介可通过下述设备体现：传输线、光盘、数字视频光盘、磁带、伯努利驱动器、磁盘、穿孔卡、闪存、集成电路或其他数字处理装置存储设备。

此外，描述的本发明的特征、结构或特点可以以任何合适的方式合并在一种或多种实施方式中。在下文的说明中，提供了大量的具体细节以全面理解本发明的实施方式，所述具体细节比如编程、软件模块、用户选择、网络事务、数据库查询、数据库结构、硬件模块、硬件电路、硬件芯片等等的实例。然而，相关技术领域的技术人员可认识到：本发明可在不具备一个或多个具体实施方式的具体细节的情况下被实施，或者本发明可结合其他方法、组件、材料等实施。在其他例子中，并没有显示或描述公知的结构、材料或操作以使本发明变得清晰。

此处包括的示意性流程图大体上是作为逻辑流程图来列举的。就这点而言，描述的顺序和标记的步骤是本方法的一种实施方式的指示性说明。可设想其他在功能上、逻辑上或效果上与图示方法的一个或多个步骤（或其中部分）相同的步骤和方法。此外，使用的格式和符号被用于解释方法的逻辑步骤并被理解为不限制本方法的范围。尽管在流程图中可使用不同的箭头类型和线条类型，但这些箭头类型和线条类型被理解为不限制相应方法的范围。的确，一些箭头或其他连接器可用于仅表示方法的逻辑流程。例如，箭头可表示描述的方法的列举的步骤之间的未指明期间的等待或监测时期。此外，特定方法的步骤的顺序可或可不严格依照所示的对应步骤的顺序。

固态存储系统

图1A是示意性框图，示出了根据本发明的用于固态存储设备内的数据管理的系统100的一种实施方式。系统100包括固态存储设备102、固态存储控制器104、写入数据管道106、读取数据管道108、固态存储器110、计算机112、客户端114和计算机网络116，这些装置描述如下。

系统100包括至少一个固态存储设备102。在另一种实施方式中，系统100包括两个或更多个固态存储设备102，每个固态存储设备102可包括非易失性的、固态的存储器110，所述非易失性的、固态的存储器例如纳米随机存取存储器（“纳米RAM”或者“NRAM”）、磁电阻式RAM（“MRAM”）、动态RAM（“DRAM”）、相变RAM（“PRAM”）闪存等等。结合图2和图3更详细地描述了固态存储设备102。固态存储设备102被描述成位于通过计算机网络116与客户端114相连的计算机112内。在一种实施方式中，固态存储设备102位于计算机112内部并且采用系统总线连接，所述系统总线例如快速外围组件互连（“PCI-e”）总线、串行高级技术附件（“串行ATA”）总线或类似总线。在另一种实施方式吧，固态存储设备102位于计算机112外部，并且通过通用串行总线（“USB”）、电气与电子工程师协会（“IEEE”）1394总线（“火线”）或类似总线连接。在其他实施方式中，固态存储设备102采用下述方式与计算机112相连接：外围组件互连（“PCI”）express总线、外部电或光总线扩展或者总线网络解决方案，所述总线网络解决方案例如无限带宽或快速PCI高级交换（“PCIe-AS”）或类似技术。

在不同的实施方式中，固态存储设备102可以是双列直插式内存模块（“DIMM”）、子卡或微型模块的形式。在另一种实施方式中，固态存储设备102是位于机架式刀片内的元件。在另一种实施方式中，固态存储设备102包含在直接集成到高级集成装置（如主板、笔记本电脑、图形处理器）的封装内。在另一种实施方式中，包括固态存储设备102的单独元件直接集成到高级集成装置上而不经过中间封装。

固态存储设备102包括一个或多个固态存储控制器104，每个固态存储控制器104可包括写入数据管道106和读取数据管道108，而且，每个固态存储控制器104还包括固态存储器110，这将在下文中结合图2和图3详细说明。

系统100包括一台或多台连接到固态存储设备102的计算机112。计算机112可以是主机、服务器、存储区域网络（“SAN”）的存储控制器、工作站、个人计算机、笔记本电脑、手持式计算机、超级计算机、计算机集群、网络交换机、路由器或设备、数据库或存储设备、数据采集或数据采集系统、诊断系统、测试系统、机器人、便携式电子设备、无线设备或类似设备。在另一种实施方式中，计算机112可以是客户端，并且固态存储设备102自主运行

以应答发送自计算机112的数据请求。在这种实施方式中，计算机112和固态存储设备102可采用下列方式连接：计算机网络、系统总线或其他适于在计算机112和自主固态存储设备102之间连接的通信手段。

在一种实施方式中，系统100包括一个或多个客户端114，所述一个或多个客户端114通过一个或多个计算机网络116连接到一台或多台计算机112。客户端114可以是主机、服务器、SAN的存储控制器、工作站、个人计算机、笔记本电脑、手持式计算机、超级计算机、计算机集群、网络交换机、路由器或设备、数据库或存储设备、数据采集或数据采集系统、诊断系统、测试系统、机器人、便携式电子设备、无线设备或类似设备。计算机网络116可包括因特网、广域网（“WAN”）、城域网（“MAN”）、局域网（“LAN”）、令牌环网、无线网络、光纤通道网络、SAN、网络附属存储（“NAS”）、ESCON或类似网络、或者是网络的任意组合。计算机网络116还可包括来自IEEE802系列网络技术中的网络，如以太网、令牌环网、WiFi、WiMax及类似网络。

计算机网络116可包括服务器、交换机、路由器、电缆、无线电和其他用于促进计算机112和客户端114的网络连接的设备。在一种实施方式中，系统100包括通过计算机网络116进行对等通信的多台计算机112。在另一种实施方式中，系统100包括通过计算机网络116进行对等通信的多个固态存储设备102。本领域技术人员可认识到其他计算机网络116可包括一个或多个计算机网络116以及相关设备，所述相关设备具有一个或多个客户端114、其他计算机或与一台或多台计算机112相连的一个或多个固态存储设备102之间的单个或冗余连接，所述其他计算机具有一个或多个固态存储设备102。在一种实施方式中，系统100包括两个或更多个通过计算机网络118连接到客户端116的固态存储设备102，而不包括计算机112。

存储控制器管理的对象

图1B是示意性框图，示出了根据本发明的用于存储设备内的对象管理的系统101的一种实施方式。系统101包括一个或多个存储设备150（每一个存储设备150都具有存储控制器152和一个或多个数据存储设备154）和一个或多个请求设备155。存储设备152联网在一起并与一个或多个请求设备155连接。请求设备155将对象请求发给存储设备150a。对象请求可以是创建对象的请求、向对象写入数据的请求、从对象读取数据的请求、删除对象的请求、检查对

象的请求、复制对象的请求及类似请求。本领域技术人员会认识到其他对象请求。

在一种实施方式中，存储控制器152和数据存储设备154是分离的设备。在另一种实施方式中，存储控制器152和数据存储设备154集成到一个存储设备150上。在另一种实施方式中，数据存储设备154为固态存储器110，而存储控制器为固态存储设备控制器202。在其他实施方式中，数据存储设备154可以为硬盘驱动器、光驱动器、磁带存储器或类似存储设备。在另一种实施方式中，存储设备150可包括两个或更多个不同类型的数据存储设备154。

在一种实施方式中，数据存储设备154为固态存储器110，并且被布置为固态存储元件216、218、220的阵列。在另一种实施方式中，固态存储器110被布置在两个或更多个内存库(bank)214a-n内。下文结合图2B更详细地描述了固态存储器110。

存储设备150a-n可联网在一起并且可作为分布式存储设备运行。与请求设备155连接的存储设备150a控制发送到所述分布式存储设备的对象请求。在一种实施方式中，存储设备150和关联的存储控制器152管理对象并对请求设备155来说表现为分布式对象文件系统。在这种情况下，一类分布式对象文件系统的实例是并行对象文件系统。在另一种实施方式中，存储设备150和关联的存储控制器152管理对象并对请求设备155来说表现为分布式对象文件服务器。在这种情况下，一类分布式对象文件服务器的实例是并行对象文件服务器。在这些和其他实施方式中，请求设备155可只管理对象或者与存储设备150结合而参与管理对象，这通常并不将存储设备150的功能限制在为其他客户端114充分管理对象的范围内。在退化情况下，每个分布式存储设备、分布式对象文件系统和分布式对象文件服务器能作为单个设备独立运行。联网的存储设备150a-n可作为分布式存储设备、分布式对象文件系统、分布式对象文件服务器和它们的任意组合运行，所述组合具有一个或多个为一个或多个请求设备155配置的这些功能。例如，存储设备150可配置为：为第一请求设备155a作为分布式存储设备运行，而请求设备155b作为分布式存储设备和分布式对象文件系统为运行。当系统101包括一个存储设备150a时，存储设备150a的存储控制器152a管理对象并对请求设备155来说表现为对象文件系统或对象文件服务器。

在一种实施方式中，其中，存储设备150作为分布式存储设备联网在一起，存储设备150充当由一个或多个分布式存储控制器152管理的独立驱动器冗余阵列（“RAID”）。例如，写入对象数据段的请求导致所述数据段根据RAID级别在数据存储设备154a-n中被条带化为具有奇偶校验条带的条带。这种布置的一个好处是这种对象管理系统可在单独的存储设备150（无论是存储控制器152、数据存储设备154或存储设备150的其他组件）出现故障时继续使用。

当冗余网络用于互连存储设备150和请求设备155时，所述对象管理系统可在出现网络故障的情况下（只要网络中的一个仍在运行）继续使用。具有一个存储设备150a的系统101还可包括多个数据存储设备154a，而存储设备150a的存储控制器152a可作为RAID控制器运行并在存储设备150a的数据存储设备154a间分割数据段，存储设备150a的存储控制器152a可包括根据RAID级别的奇偶校验条带。

在一种实施方式中，其中，一个或多个存储设备150a-n是具有固态存储设备控制器202和固态存储器110的固态存储设备102，固态存储设备102可配置为DIMM配置、子卡、微型模块等，并保留在计算机112内。计算机112可以是服务器或具有固态存储设备102的类似设备，固态存储设备102联网在一起并作为分布式RAID控制器运行。有利地是，存储设备102可采用PCI-e、PCIe-AS、无限带宽或其他高性能总线、交换总线、网络总线或网络连接，并且可提供极致密型、高性能的RAID存储系统，在该系统中，单独的或分布式固态存储控制器202自主地在固态存储器110a-n间条带化数据段。

在一种实施方式中，请求设备155用于与存储设备150通信的同一网络可被对等存储设备150a使用，以与对等存储设备150b-n通信以实现RAID功能。在另一种实施方式中，可为了RAID的目的而在存储设备150间使用单独的网络。在另一种实施方式中，请求设备155可通过向存储设备150发送冗余请求而参与RAID进程。例如，请求设备155可向第一存储设备150a发送第一对象写入请求，而向第二存储设备150b发送具有相同数据段的第二对象写入请求以实现简单的镜像。

当具有在存储设备102内进行对象处理的能力时，只有存储控制器152具有采用一个RAID级别存储一个数据段或对象的能力，而采用不同的RAID级别或不采用RAID条带化来存储另一数据段或对象。这些多个RAID群组可与存储

设备 150 内的多个分区相关联。可同时在各种 RAID 群组间支持 RAID0、RAID1、RAID5、RAID6 和复合 RAID 类型 10、50、60，所述 RAID 群组包括数据存储设备 154a-n。本领域技术人员可认识到还可同时支持的其他 RAID 类型和配置。

而且，由于存储控制器 152 像 RAID 控制器一样自主运行，所述 RAID 控制器能够执行渐进式 RAID 并能够将在数据存储设备 154 间条带化的具有一个 RAID 级别的对象或对象的某些部分转换为另一 RAID 级别，转换时请求设备 155 不受影响、不参与或者甚至不探测 RAID 级别的变化。在优选实施方式中，促进 RAID 配置从一个级别变为另一级别可在对象或甚至在包基上自主实现，并且可由运行在存储设备 150 或存储控制器 152 中的一个上的分布式 RAID 控制模块初始化。通常，RAID 渐进是从高性能和低效率的存储配置（如 RAID1）转换为低性能和高存储效率的存储配置（如 RAID5），其中，转换是基于读取频率被动态地初始化。但是，可以发现，从 RAID5 到 RAID1 的渐进也是可能的。可配置其他用于初始化 RAID 渐进的进程，或者可由客户端或外部代理（如存储系统管理服务器请求）请求该进程。本领域技术人员可认识到具有存储控制器 152 的存储设备 102 的其他特征和优点，该存储控制器 152 自主管理对象。

具有服务器内 SAN 的固态存储设备

图 1C 是示意性框图，示出了根据本发明的用于服务器内存储区域网络（“SAN”）的系统 103 的一种实施方式。系统 103 包括计算机 112，计算机 112 通常被配置为服务器（“服务器 112”）。每个服务器 112 包括一个或多存储设备 150，其中，服务器 112 和存储设备 150 分别连接到共享网络接口 155。每个存储设备 150 包括存储控制器 152 和相应的数据存储设备 154。系统 103 包括客户端 114、114a、114b，客户端 114、114a、114b 位于服务器 112 的内部或者外部。客户端 114、114a、114b 可通过一个或多个计算机网络 116 与每个服务器 112 和每个存储设备 150 通信，所述一个或多个计算机网络 116 大体上与上述的计算机网络类似。

存储设备 150 包括 DAS 模块 158、NAS 模块 160、存储通信模块 162、服务器内 SAN 模块 164、通用接口模块 166、网络代理模块 170、虚拟总线模块 172、前端 RAID 模块 174 及后端 RAID 模块 176，这些模块将在下文中描述。模块 158-176 图示为位于存储设备 150 内，模块 158-176 中的每一个的全

部或部分可位于存储设备 150、服务器 112、存储控制器 152 内或位于其他位置。

服务器 112（如与服务器内 SAN 结合使用的）是具有服务器功能的计算机。服务器 112 至少包括一项服务器功能（如文件服务器功能），而且还可包括其他服务器功能。服务器 112 可以是服务器群的一部分并可服务其他客户端 114。在其他实施方式中，服务器 112 还可以是个人计算机、工作站或其他包括存储设备 150 的计算机。服务器 112 可像访问直接附加存储（“DAS”）、SAN 附加存储或者网络附加存储（“NAS”）那样访问服务器 112 内的一个或多个存储设备 150。参与服务器内 SAN 或 NAS 的存储控制器 150 可位于服务器 112 的内部或外部。

在一种实施方式中，服务器内 SAN 装置包括 DAS 模块 158，该 DAS 模块 158 将由服务器 112 的存储控制器 152 控制的至少一个数据存储设备 154 的至少一部分配置为附属于服务器 112 的 DAS 设备，以服务从至少一个客户端 114 传送到服务器 112 的存储请求。在一种实施方式中，第一数据存储设备 154a 被配置为第一服务器 112a 的 DAS，而且，第一数据存储设备 154a 还被配置为第一服务器 112a 的服务器内 SAN 存储设备。在另一种实施方式中，第一数据存储设备 154a 被分割，以使得一个分区为 DAS 而另一个分区为服务器内 SAN。在另一种实施方式中，第一数据存储设备 154a 内的存储空间的至少一部分被配置为第一服务器 112a 的 DAS，而第一服务器 112a 的存储空间的同一部分被配置为第一服务器 112a 的服务器内 SAN。

在另一种实施方式中，服务器内 SAN 装置包括 NAS 模块 160，该 NAS 模块 160 将存储控制器 152 配置为用于至少一个客户端 114 的 NAS 设备并服务来自客户端 114 的文件请求。存储控制器 152 还可被配置为用于第一服务器 112a 的服务器内 SAN 设备。存储设备 150 可通过共享网络接口 155 直接连接到计算机网络 116，共享网络接口 155 独立于存储设备 150 位于其内的服务器 112。

在一种基本的形式中，用于服务器内 SAN 的装置包括第一服务器 112a 内的第一存储控制器 152a，其中，第一存储控制器 152a 控制至少一个存储设备 154a。第一服务器 112a 包括由第一服务器 112a 和第一存储控制器 152a 共享的网络接口 156，所述服务器内 SAN 装置包括存储通信模块 162，该存储

通信模块 162 促进第一存储控制器 152a 和位于第一服务器 112a 外部的至少一个设备之间的通信，以使得第一存储控制器 152a 和外部设备之间的所述通信独立于第一服务器 112a。存储通信模块 162 可允许第一存储控制器 152a 独立地访问网络接口 156a 以进行外部通信。在一种实施方式中，存储通信模块 162 访问网络接口 156a 中的交换机以管理第一存储控制器 152a 和外部设备之间的网络流量。

服务器内 SAN 装置还包括服务器内 SAN 模块 164，该服务器内 SAN 模块 164 利用网络协议和总线协议中的一个或两个服务存储请求。服务器内 SAN 模块 164 服务独立于第一服务器 112a 的所述存储请求，并且所述服务请求接收自内部或外部客户端 114a、114。

在一种实施方式中，位于第一服务器 112a 外部的设备是第二存储控制器 152b。第二存储控制器 152b 控制至少一个数据存储设备 154b。服务器内 SAN 模块 164 利用第一和第二存储控制器 152a、152b 之间、通过网络接口 156a 的通信服务所述存储请求，第一和第二存储控制器 152a、152b 独立于第一服务器 112a。第二存储控制器 152b 可位于第二服务器 112b 内或位于一些其他设备内。

在另一种实施方式中，第一服务器 112a 外部的设备是客户端 114，并且所述存储请求源于外部客户端 114，其中，第一存储控制器被配置为 SAN 的至少一部分，并且服务器内 SAN 模块 164 通过独立于第一服务器 112a 的网络接口 156a 服务所述存储请求。外部客户端 114 可位于第二服务器 112b 内或可位于第二服务器 112b 的外部。在一种实施方式中，即使当第一服务器 112a 不可用时，服务器内 SAN 模块 164 也能够服务来自外部客户端 114 的存储请求。

在另一种实施方式中，生成所述存储请求的客户端 114a 位于第一服务器 112a 的内部，其中，第一存储控制器 152a 被配置为 SAN 的至少一部分，并且服务器内 SAN 模块 164 通过一个或多个网络接口 156a 和系统总线服务所述存储请求。

传统的 SAN 配置允许像按直接附加存储（“DAS”）访问服务器 112 内的存储设备一样访问远离服务器 112 的存储设备，以使得远离服务器 112 的所述存储设备表现为块存储设备。通常，按 SAN 连接的存储设备需要 SAN 协

议，所述 SAN 协议例如光纤通道、互联网小型计算机系统接口（“iSCSI”）、HyperSCSI、光纤连通性（“FICON”）、通过以太网的高级技术附件（“ATA”）等。服务器内 SAN 包括服务器 112 内的存储控制器 152，同时仍然允许存储控制器 152a 和远程存储控制器 152b 或外部客户端之间利用网络协议和/或总线协议的网络连接。

通常，SAN 协议是网络协议的形式，并且，出现了更多的网络协议，例如可允许存储控制器 150a 和关联的数据存储设备 154a 被配置为 SAN 并与外部外部客户端 114 或第二存储控制器 152b 通信的无限带宽。在另一种实例中，第一存储控制器 152a 可利用以太网与外部客户端 114 或第二存储控制器 152b 通信。

存储控制器 152 可通过总线与内部存储控制器 152 或客户端 114a 通信。例如，存储控制器 152 可通过使用 PCI-e 的总线通信，所述 PCI-e 可支持 PCI 快速输入/输出虚拟化（“PCIe-IOV”）。其他新出现的总线协议允许系统总线扩展超出计算机或服务器 112 并可允许存储控制器 152a 被配置为 SAN。一种这样的总线协议是 PCIe-AS。本发明并不仅限于 SAN 协议，还可利用新出现的网络和总线协议服务存储请求。外部设备（无论是客户端 114 的形式还是外部存储控制器 152b 的形式）可通过扩展系统总线或计算机网络 116 通信。正如此处所使用的，存储请求包括写入数据、读取数据、擦除数据、查询数据的请求等等，并且所述存储请求可包括对象数据、元数据、管理请求以及块数据请求。

传统的服务器 112 通常具有控制访问服务器 112 内的设备的根联合体。通常，服务器 112 的所述根联合体具有网络接口 156，从而使得服务器 112 控制任何通过网络接口 156 的通信。然而，在服务器内 SAN 装置的优选实施方式中，存储控制器 152 能够独立地访问网络接口 156，从而使得客户端 114 可直接地与形成 SAN 的第一服务器内 112a 内的一个或多个存储控制器 152a 通信，或者使得一个或多个第一存储控制器 152a 可与第二存储控制器 152b 或其他远程存储控制器 152 联网在一起以形成 SAN。在这种优选实施方式中，远离第一服务器 112a 的设备可通过单独的、共享的网络地址访问第一服务器 112a 或第一存储控制器 152a。在一种实施方式中，服务器内 SAN 装置包括通用接口模块 166，该通用接口模块 166 配置网络接口 156、存储控制器 152 和

服务器 112，以使得可使用共享网络地址访问服务器 112 和存储控制器 152。

在其他实施方式中，服务器 112 包括两个或更多个网络接口 156。例如，服务器 112 可通过一个网络接口 156 通信，而存储设备 150 可通过另一个接口通信。在另一个实例中，服务器 112 包括多个存储设备 150，每个存储设备 150 具有网络接口 156。本领域技术人员会认识到具有一个或多个存储设备 150 和一个或多个网络接口 156 的服务器 112 的其他配置，其中，一个或多个存储设备 150 访问独立于服务器 112 的网络接口 156。本领域技术人员还可认识到扩展这些不同的配置的方法以支持网络冗余并提高可用性。

有利地是，服务器内 SAN 装置大大降低了传统 SAN 的复杂性和花费。例如，典型的 SAN 需要具有外部存储控制器 152 和关联的数据存储设备 154 的服务器 112。这占用了机架上的额外空间并且需要电缆、交换机等。配置传统的 SAN 所需的电缆、交换机和其他的开销占用了空间、降低了带宽并且昂贵。服务器内 SAN 装置允许存储控制器 152 和关联的存储器 154 适合服务器 112 的形体尺寸，并因此减少了需要的空间和费用。服务器内 SAN 还允许通过内部和外部高速数据总线使用相对高速的通信的连接。

在一种实施方式中，存储设备 150 为固态存储设备 102，存储控制器 152 为固态存储控制器 104，而数据存储设备 154 为固态存储器 110。由于此处所述的固态存储设备 102 的速度，这种实施方式是有利的。此外，固态存储设备 102 可被配置为位于 DIMM 中，所述 DIMM 可方便地装配在服务器 112 内并仅需要少量的空间。

服务器 112 中的一个或多个内部客户端 114a 还可通过服务器的网络接口 156 连接到计算机网络 116，并且服务器 112 通常控制所述客户端的连接。这种方法具有一些好处。客户端 114a 可直接地本地访问或远程访问存储设备 150，并且客户端 114a 可初始化客户端 114a 的存储器和存储设备 150 之间的本地或远程直接存储器存取（“DMA”，“RDMA”）数据的传送。

在另一种实施方式中，当利用本地连接的存储设备 150 作为 DAS 设备、网络连接的存储设备 150、网络连接的固态存储设备 102（这些设备作为服务器内 SAN、外部 SAN 和混合 SAN 的一部分）时，位于服务器 112 内部或外部的客户端 114、114a 可通过一个或多个网络 116 对客户端 114 起文件服务器的作用。存储设备 150 可同时参与 DAS、服务器内 SAN、SAN、NAS 等（及

其中的任意的组合)。此外，每个存储设备 150 可以以如下方式被分割：第一分区使存储设备 150 作为 DAS 可用，第二分区使存储设备 150 作为服务器内 SAN 内的元件可用，第三分区使存储设备 150 作为 NAS 可用，第四分区使存储设备 150 作为 SAN 的元件可用，等等。类似地，存储设备 150 可被分割为符合安全性和存取控制要求。本领域技术人员会认识到可以构建和支持下述设备或系统的任意组合和排列：存储设备、虚拟存储设备、存储网络、虚拟存储网络、专用存储器、共享存储器、平行文件系统、平行对象文件系统、块存储设备、对象存储设备、存储装置、网络装置及类似设备。

此外，通过将存储设备 150 直接地连接到计算机网络 116，存储设备 150 彼此之间能够互相通信并能够起服务器内 SAN 的作用。通过计算机网络 116 连接的服务器 112 内的客户端 114a 和客户端 114 可像访问 SAN 那样访问存储设备 150。通过将存储设备 150 移到服务器 112 内并使其具备将存储设备 150 配置为 SAN 的能力，服务器 112/存储设备 150 的结合使得在常规 SAN 中不需要专用的存储控制器、光纤通道网络和其他设备。服务器内 SAN 系统 103 具有能够使存储设备 150 与客户端 114 和计算机 112 共享共用的资源(如电源、制冷、管理和物理空间)的优点。例如，存储设备 150 可插在服务器 112 的空的插槽中并提供 SAN 或 NAS 的所有工作性能、可靠性和可用性。本领域技术人员会认识到服务器内 SAN 系统 103 的其他特征和优点。

在另一种配置中，多个服务器内 SAN 存储设备 150a 被布置在单独的服务器 112a 基础架构内。在一种实施方式中，服务器 112a 由一个或多个利用 PCI 快速 IOV 互连的内部刀片服务器客户端 114a 组成，而没有外部网络 156、外部客户端 114、114b 或外部存储设备 150b。

此外，服务器内 SAN 存储设备 150 可通过一个或多个计算机网络 116 与对等存储设备 150 通信，所述对等存储设备 150 位于计算机 112(图 1 中的每一台计算机)内，或者不通过计算机 112 而直接连接到计算机网络 116 以形成同时具有 SAN 和服务器内 SAN 的全部功能的混合 SAN。这种灵活性具有以下优点：简化了扩展性和在不同的可能的固态存储网络实施方案之间的移植。本领域技术人员会认识到放置和互连固态控制器 104 的其他组合、配置、实施方案和布局。

当网络接口 156a 仅能被运行在服务器 112a 内的一个代理控制时，运行在

所述代理中的链路建立模块 168 能够通过连接到外部存储设备 150b 和客户端 114、114b 的网络接口 156a 建立内部客户端 114a 和存储设备 150a/第一存储控制器 152a 之间的通信通路。在优选的实施方式中，一旦建立了所述通信通路，单独的内部存储设备 150a 和内部客户端 114a 能够建立和管理它们自己的命令队列，并通过网络接口 156a 和独立于控制网络接口 156a 的网络代理或代理的 RDMA 将命令和数据同时双向地直接传送给外部存储设备 150b 和客户端 114、114b。在一种实施方式中，链路建立模块 168 在初始化过程（如硬件的启动或初始化）期间建立通信链路。

在另一种实施方式中，网络代理模块 170 指令至少一部分用于通过第一服务器 112 服务存储请求的命令，而至少与所述存储请求关联的数据（也可能是其他命令）在第一存储控制器和独立于第一服务器的外部存储设备之间通信。在另一种实施方式中，网络代理模块 170 代表内部存储设备 150a 和客户端 114a 转发命令或数据。

在一种实施方式中，第一服务器 11a 包括位于第一服务器 112a 内的一个或多个服务器，并包括虚拟总线模块 172，该虚拟总线模块 172 允许第一服务器 112a 内的所述一个或多个服务器通过分享的虚拟总线独立地访问一个或多个存储控制器 152a。所述虚拟总线可利用高级总线协议（如 PCIe-IOV）建立。支持 IOV 的网络接口 156a 可允许所述一个或多个服务器和所述一个或多个存储控制器独立地控制一个或多个网络接口 156a。

在不同的实施方式中，服务器内 SAN 装置允许两个或更多个存储设备 150 被配置在 RAID 中。在一种实施方式中，服务器内 SAN 装置包括将两个或更多个存储控制器配置为 RAID 的前端 RAID 模块 174。当来自客户端 114、114a 的存储请求包括存储数据的请求时，前端 RAID 模块 174 通过将所述数据以符合特定应用的 RAID 级的形式写入所述 RAID 服务所述存储请求。第二存储控制器 152 可位于第一服务器 112a 的内部或者外部。前端 RAID 模块 174 允许将存储控制器 152 配置成 RAID，从而使得存储控制器对发送所述存储请求的客户端 114、114a 可见。这种方法允许被指定为主机的存储控制器 152 或客户端 114、114a 管理条纹和校验信息。

在另一种实施方式中，服务器内 SAN 装置包括后端 RAID 模块 176，该后端 RAID 模块 176 将由存储控制器控制的两个或更多个数据存储设备 154 配置

为RAID。当来自所述客户端的存储请求包括存储数据的请求时，后端RAID模块176通过将所述数据以符合应用的RAID级的形式写入所述RAID服务所述存储请求，从而使得客户端114、114a像访问由第一存储控制器152控制的单个数据存储设备154那样访问被配置为RAID的存储设备154。这种RAID应用允许以如下方式将由存储控制器152控制的数据存储设备配置成RAID：配置成RAID的过程对任何访问数据存储设备154的客户端114、114a来说是透明的。在另一种实施方式中，前端RAID和后端RAID都具有多级RAID。本领域技术人员会认识到将存储设备152以符合此处所述的固态存储控制器104和关联的固态存储器110的形式配置为RAID的其他方法。

用于存储控制器管理的对象的装置

图2A是示意性框图，示出了根据本发明的用于存储设备内的对象管理的装置200的一种实施方式。装置200包括存储控制器152，该存储控制器152具有：对象请求接收器模块260、解析模块262、命令执行模块264、对象索引模块266、对象请求排队模块268、具有消息模块270的封包器302、及对象索引重建模块272，上述模块描述如下。

存储控制器152大体上与图1B中的系统102描述的存储控制器152类似，并且可以是图2描述的固态存储设备控制器202。装置200包括对象请求接收器模块260，该对象请求接收器模块260接收来自一个或多个请求设备155的对象请求。例如，对于存储对象数据请求，存储控制器152在数据存储设备154中以数据包的形式存储数据段，该数据存储设备154与存储控制器152相连接。所述对象请求通常由存储在或将要被存储在一个或多个对象数据包中的数据段指令存储控制器管理的对象。对象请求可请求存储控制器152创建对象，该对象随后会通过可利用本地或远程直接内存读取（“DMA”、“RDMA”）转换的稍后的对象请求来填充数据。

在一种实施方式中，对象请求为将对象的全部或一部分写入先前创建的对象的写入请求。在一个实例中，所述写入请求用于对象的数据段。可将所述对象的其他数据段写入存储设备150或者写入其他存储设备152。在另一个实例中，所述写入请求用于整个对象。在另一个实例中，所述对象请求为从由存储控制器152管理的数据段中读取数据。在又一种实施方式中，所述对象请求为删除请求，以删除数据段或对象。

有利地是，存储控制器152能接受不仅仅写新对象或为已存在的对象添加数据的写入请求。例如，由对象请求接收器模块260接收的写入请求可包括：在由存储控制器152存储的数据前添加数据的请求、在已存储的数据中插入数据的请求或者替换数据的一段的请求。由存储控制器152保持的对象索引提供了这些复杂写操作所需要的灵活性，所述写操作在其他存储控制器内不可用，但是目前仅在服务器和其他计算机文件系统内的存储控制器外可用。

装置200包括解析模块262，该解析模块262将所述对象请求解析为一条或多条命令。通常，解析模块262将所述对象请求解析为一个或多个缓存。例如，所述对象请求中的一条或多条命令可被解析为命令缓存。通常，解析模块262准备对象请求，以使得所述对象请求中的信息可以被存储控制器152理解并执行。本领域技术人员会认识到将对象请求解析为一条或多条命令的解析模块262的其他功能。

装置200包括命令执行模块264，该命令执行模块264执行从所述对象请求解析出的命令。在一种实施方式中，命令执行模块264执行一条命令。在另一种实施方式中，命令执行模块264执行多条命令。通常，命令执行模块264解释解析自所述对象请求的命令（如写入命令），然后创建、排列并且执行子命令。例如，解析自对象请求的写入命令可指令存储控制器152存储多个数据段。所述对象请求还可包括必要属性（如加密、压缩等）。命令执行模块264可命令存储控制器152压缩所述数据段、加密所述数据段、创建一个或多个数据包并为每个数据包关联包头、使用媒体加密密钥加密所述数据包、添加错误修正码并将所述数据包存储在指定位置。在指定位置存储所述数据包，并且其他子命令还可被分解为其他更高级别的子命令。本领域技术人员会认识到命令执行模块264能执行一条或多条解析自对象请求的命令的其他方法。

装置200包括对象索引模块266，该对象索引模块266在对象索引中创建对象项，以响应创建对象或存储所述对象数据段的存储控制器152。通常，存储控制器152从所述数据段中创建数据包，并且在存储所述数据段时，所述数据包存储的位置即被指定。同数据段一起接收的或作为对象请求的一部分接收的对象元数据可采用类似方法存储。

对象索引模块266在存储所述数据包和分配所述数据包的物理地址时创建进入对象索引的对象项。所述对象项包括所述对象的逻辑标识符和一个或

多个物理地址之间的映射，所述一个或多个物理地址对应于存储控制器152存储一个或多个数据包和任何对象元数据包的位置。在另一种实施方式中，在存储所述对象的数据包之前在所述对象索引中创建项。例如，如果存储控制器152较早地确定存储所述数据包的物理地址，则对象索引模块266可较早地在所述对象索引中创建项。

通常，当对象请求或对象请求组导致对象或数据段被修改时（可能在读修改写操作期间），所述对象索引模块266更新所述对象索引中的项以符合修改的对象。在一种实施方式中，所述对象索引创建新对象并在所述对象索引为所述修改的对象创建新项。通常，当仅有对象的一部分被修改时，所述对象包括修改过的数据包和一些保持不变的数据包。在这种情况下，所述新项包括到未变的数据包（与最初写入它们的位置相同）的映射和到写入新位置的修改后的对象的映射。

在另一种实施方式中，对象请求接收器模块260接收对象请求，该对象请求包括擦除数据块或其他对象元的命令，存储控制器152可至少存储一个包（如擦除包，该擦除包具有对象的引用、与对象的关系和擦除的数据块的大小的信息）。此外，这可进一步表明擦除的对象元素被填充为0。因此，擦除对象请求可用于仿真被擦除的实际的内存或存储器，并且，所述实际的内存或存储器实际上具有合适的内存/存储器的一部分，所述合适的内存/存储器实际上以0存储在所述内存/存储器的单元中。

有利地是，创建具有项（该项表明了数据段和对象元数据之间的映射）的对象索引允许存储控制器152自主的处理和管理对象。这种能力允许在存储设备150中十分灵活地存储数据。一旦创建了对象的索引项，存储控制器152可有效地处理后继关于所述对象的对象请求。

在一种实施方式中，存储控制器152包括对象请求排队模块，该对象请求排队模块在解析模块262解析之前将一个或多个由对象请求接收器模块260接收到的对象排队。对象请求排队模块268允许在接收对象请求时和在执行所述对象请时之间的灵活性。

在另一种实施方式中，存储控制器152包括封包器302，该封包器302根据一个或多个数据段创建一个或多个数据包，其中，数据包的大小适于存储在数据存储设备154内。在下文中结合图3更详细地描述封包器302。在一种实施

方式中，封包器302包括为每个包创建包头的消息模块270。所述包头包括包标识符和包长度。所述包标识符把所述包与对象（为该对象生成所述包）联系起来。

在一种实施方式中，由于包标识符包含足够的信息以确定对象和在对象内的包含在包内的对象元素之间的关系，因此每个包包括自包含的包标识符。然而，更有效的优选实施方式是在容器中存储包。

容器是一种数据结构，这种数据结构有助于更有效的存储数据包并帮助建立对象和数据包、元数据包和其他与存储在容器内的对象有关的包之间的关系。注意到存储控制器152通常以处理作为对象的一部分接收的对象元数据的类似方式处理数据段。通常，“包”可指包含数据的数据包、包含元数据的元数据包或其他包类型的其他包。对象可存储在一个或多个容器中，并且容器通常包括仅用于一个唯一的对象的包。对象可分布在多个容器之间。容器通常存储在单个逻辑擦除块内（存储部）并且通常不分散在逻辑擦除块间。

在一个实例中，容器可分散在两个或更多个逻辑/虚拟页间。通过将容器与对象关联起来的容器标签确定容器。容器可包含0个到许多个包并且容器内的这些包通常来自一个对象。包可以有许多对象元素类型（包括对象属性元、对象数据元、对象索引元和类似的元素类型）。可以创建包括不止一个对象元类型的混合包。每个包可包含0个到许多个同一类型的元。容器内的每个包通常都包含标识与对象关系的唯一标识符。

每个包与一个容器相关联。在优选实施方式中，容器被限于擦除块，以使得在每个擦除块的起始部分或在擦除块的起始部分附近能发现容器包。这有助于将数据丢失限制在具有损坏的包头的擦除块范围内。在这种实施方式中，如果对象索引不可用并且擦除块内的包头损坏，由于可能没有可靠的机制确定后继包的位置，从损坏的包头到擦除块尾的内容可能会丢失。在另一种实施方式中，更可靠的方法是采用限于页的边界的容器。这种实施方式需要更多包头开销。在另一种实施方式中，容器可流经页面和擦除块边界。这种方法需要较少的包头开销，但是，如果包头损坏，则有可能会丢失更多部分的数据。对这些实施方式来说，使用一些类型的RAID以进一步保证数据完整性是可以预期的。

在一种实施方式中，装置200包括对象索引重建模块272，该对象索引重

建模块272采用来自存储在数据存储设备154中的包头的信息重建所述对象索引中的项。在一种实施方式中，对象索引重建模块272通过读取包头（以确定每个包所属的对象）和序列信息（以确定数据或元数据在对象中所属的位置）来重建所述对象索引的项。对象索引重建模块272采用每个包的物理地址信息和时间戳或序列信息以创建包的物理地址和对象标识符和数据段序列间的映射。对象索引重建模块272使用时间戳或序列信息以再现索引变更的顺序并通常因此重建最近的状态。

在另一种实施方式中，对象索引重建模块272采用包头信息以及容器包信息放置包以识别包的物理位置、对象标识符和每个包的序列号，从而在所述对象索引中重建项。在一种实施方式中，在写入数据包时，擦除块被戳记上时间，或者赋给擦除块序列号，并且擦除块的时间戳或序列信息和来自容器头和包头的信息一起使用以重建对象索引。在另一种实施方式中，当擦除块恢复时，时间戳或序列信息被写入该擦除块。

当对象索引存储在易失性存储器中时，如果不能重建所述对象索引，错误、失电、或其他导致存储控制器152未存储所述对象索引而停工的因素可能会成为问题。对象索引重建模块272允许所述对象索引存储在具有易失性存储体优点（如快速存取）的易失性存储体中。对象索引重建模块272允许自主地快速重建所述对象索引，而并不需要依靠位于存储设备150外的设备。

在一种实施方式中，易失性存储体中的所述对象索引周期性地存储在数据存储设备154内。在具体的实例中，所述对象索引或“索引元数据”周期性地存储固态存储器110中。在另一种实施方式中，所述索引元数据存储在固态存储器110n（与固态存储器110a-110n-1存储包分离）中。独立于数据和对象元数据管理所述索引元数据，所述数据和对象元数据传送自请求设备155并且由存储控制器152/固态存储控制器202管理。管理和存储与其他来自对象的数据和元数据分离的索引元数据允许有效的数据流，同时存储控制器152/固态存储设备控制器202并不会不必要地处理对象元数据。

在一种实施方式中，其中，由对象请求接收器模块260接收到的对象请求包括写入请求，存储控制器152通过本地或远程直接存储器存取（“DMA”、“RDMA”）操作接收来自请求设备155的内存的一个或多个对象数据段。在优选实例中，存储控制器152在一次或多次DMA或RDMA操作中从请求设备155

的内存中读取数据。在另一实例中，请求设备155在一次或多次DMA或RDMA操作中将所述数据段写入存储控制器152。在另一种实施方式中，其中，所述对象请求包括读请求，存储控制器152在一次或多次DMA或RDMA操作中将对象的一个或多个数据段传送给请求设备155的内存。在优选实例中，存储控制器152在一次或多次DMA或RDMA操作中将数据写入请求设备155的内存。在另一实例中，请求设备在一次或多次DMA或RDMA操作中从存储控制器152中读取数据。在另一实施方式中，存储控制器152在一次或多次DMA或RDMA操作中从请求设备155的内存中读取对象命令请求集。在另一实例中，请求设备155在一次或多次DMA或RDMA操作中将对象命令请求集写入存储控制器152。

在一种实施方式中，存储控制器152仿真块存储，并且在请求设备155和存储控制器152之间通信的对象包括一个或多个数据块。在一种实施方式中，请求设备155包括驱动器，以使得存储设备150表现为块存储设备。例如请求设备155可与请求设备155期望数据存储的物理地址一起发送特定大小的一组数据。存储控制器152接收所述数据块，并将与所述数据块一起传送的物理块地址或者将物理块地址的转化形式作为对象标识符。然后，存储控制器152通过随意地封包所述数据块和存储数据块将所述数据块存储为对象或对象的数据段。然后，对象索引模块266利用基于物理块的对象标识符和存储控制器152存储所述数据包的实际物理位置在所述对象索引中创建项，所述数据包包括来自所述数据块的数据。

在另一种实施方式中，存储控制器152通过接收块对象仿真块存储。块对象可包括块结构中的一个或多个数据块。在一种实施方式中，存储控制器152像处理任意其他对象一样处理所述块对象。在另一种实施方式中，对象可代表整个块设备、块设备的分区或块设备的一些其他逻辑子元件或物理子元件，所述块设备包括磁道、扇区、通道及类似设备。值得特别注意的是将块设备RAID群组重映射到支持不同RAID构建（如渐进式RAID）的对象。本领域技术人员会认识到将传统的或未来的块设备映射到对象的其他方法。

固态存储设备

图2B示出了根据本发明的位于固态存储设备102内的固态存储设备控制器202的一种实施方式201的示意性框图，该固态存储设备控制器202包括写

入数据管道 106 和读取数据管道 108。固态存储设备控制器 202 可包括若干固态存储控制器 0-N, 104a-n, 每个固态存储控制器都控制固态存储器 110。在描述的实施方式中，示出了两个固态控制器：固态控制器 0 104a 和固态控制器 N 104n，并且它们中的每一个都控制固态存储器 110a-n。在描述的实施方式中，固态存储控制器 0 104a 控制数据通道，以使得附属固态存储器 110a 存储数据。固态存储控制器 N 104n 控制与存储的数据关联的索引元数据通道，以使得关联的固态存储器 110n 存储索引元数据。在替代的实施方式中，固态存储设备控制器 202 包括具有单个固态存储器 110a 的单个固态控制器 104a。在另一种实施方式中，存在大量的固态存储控制器 104a-n 和关联的固态存储器 110a-n。在一种实施方式中，一个或多个固态控制器 104a-104n-1（与它们的关联固态存储器 110a-110n-1 连接）控制数据，而至少一个固态存储控制器 104n（与其关联固态存储器 110n 连接）控制索引元数据。

在一种实施方式中，至少一个固态控制器 104 是现场可编程门阵列 (“FPGA”) 并且控制器功能被编入 FPGA。在特定的实施方式中，FPGA 是 Xilinx® 公司的 FPGA。在另一种实施方式中，固态存储控制器 104 包括专门设计为固态存储控制器 104 的组件（如专用集成电路 (“ASIC”) 或自定义逻辑解决方案）。每个固态存储控制器 104 通常包括写入数据管道 106 和读取数据管道 108，结合图 3 进一步描述了这两个管道。在另一种实施方式中，至少一个固态存储控制器 104 由 FPGA、ASIC 和自定义逻辑组件的组合组成。

固态存储器

固态存储器 110 是非易失性固态存储元件 216、218、220 的阵列，该阵列布置在内存库 214 中并且通过双向存储输入输出 (I/O) 总线 210 并行访问。在一种实施方式中，存储 I/O 总线 210 能够在任何一个时刻进行单向通信。例如，当将数据写入固态存储器 110 时，不能从固态存储器 110 中读取数据。在另一种实施方式中，数据可同时双向地流动。然而，双向（如此处针对数据总线使用的）指在同一时间数据仅在一个方向流动的数据通路，但是，当在双向数据总线上流动的数据被阻止时，数据可在所述双向总线上沿相反方向流动。

固态存储元件（如 SSS 0.0 216a）通常被配置为芯片（一个或多个小片的封装）或电路板上的小片。正如所描述的那样，固态存储元件（如 216a）独立于或半独立于其他固态存储元件（如 218a）运行，即使这些元件被一起封

装在芯片包、芯片包的堆栈或一些其他封包元件内。正如所描述的，一列固态存储元件216、218、220被指定为内存库214。正如所描述的，可以有“n”个内存库214a-n并且每个内存库可以有“m”个固态存储元件216a-m，218a-m，220a-m，从而在固态存储器110中成为固态存储元件216、218、220的n*m阵列。在一种实施方式中，固态存储器110a在每个内存库214（有8个内存库214）中包括20个固态存储元件216、218、220，并且，固态存储器110n在每个内存库214中（只有一个内存库214）包括两个固态存储元件216、218。在一种实施方式中，每个固态存储元件216、218、220由单层单元（“SLC”）设备组成。在另一种实施方式中，每个固态存储元件216、218、220由多层单元（“MLC”）设备组成。

在一种实施方式中，用于多个内存库的固态存储元件被封包在一起，所述多个内存库共享公用存储I/O总线210a行（如216b、218b、220b）。在一种实施方式中，固态存储元件216、218、220的每个芯片可具有一个或多个小片，而一个或多个芯片垂直堆叠且每个小片可被独立存取。在另一种实施方式中，固态存储元件（如SSS 0.0 216a）的每个小片可具有一个或多个虚拟小片，每个芯片可具有一个或多个小片，而一个或多个小片中的一些或全部垂直堆叠且每个虚拟小片可被独立存取。

在一种实施方式中，每组有四个堆，每堆有两个小片垂直堆叠，从而形成8个存储元件（如SSS 0.0-SSS 0.8）216a-220a，每个存储元件位于分离的内存库214a-n内。在另一种实施方式中，20个存储元件（如SSS 0.0-SSS 20.0）216形成虚拟内存库214a，因此八个虚拟内存库中的每一个都具有20个存储元件（如SSS0.0-SSS20.8）216、218、220。通过存储I/O总线210将数据发送到固态存储器110，并发送到存储元件（SSS 0.0-SSS 0.8）216a、218a、220a的特定组的所有存储元件。存储控制总线212a用于选择特定的内存库（如内存库-0 214a），从而通过连接到所有内存库214的存储I/O总线210接收到的数据仅被写入选定的内存库214a。

在优选实施方式中，存储I/O总线210由一个或多个独立I/O总线（包括210a.a-m, 210n.a-m的“IIoBa-m”）组成，其中，每一行内的固态存储元件共享独立I/O总线中的一条，所述独立I/O总线中的一条平行访问每个固态存储元件216、218、220，从而使得同时访问所有的内存库214。例如，存储I/O总线210

的一个通道可同时访问每个内存库214a-n的第一固态存储元件216a、218a、220a。存储I/O总线210的第二通道可同时访问每个内存库214a-n的第二固态存储元件216b、218b、220b。固态存储元件216、218、220的每一行都被同时访问。在一种实施方式中，其中，固态存储元件216、218、220是多层的（物理堆叠的），固态存储元件216、218、220的所有物理层被同时访问。正如此处所使用的，“同时”还包括几乎同时的访问，其中，以略有不同的时间间隔访问设备以避免切换噪声。在这种情况下，同时被用于与连续的或系列的访问相区别，其中，命令和/或数据被单独地并相继地发送。

通常，采用存储控制总线212独立地选择内存库214a-n。在一种实施方式中，采用芯片选通或芯片选择来选择内存库214。当芯片选择和芯片使能均可用时，存储控制总线212可选择多层固态存储元件216、218、220中的一层。在其他实施方式中，存储控制总线212使用其他命令来单独地选择多层固态存储元件216、218、220中的一层。还可通过控制和地址信息的结合来选择固态存储元件216、218、220，所述控制和地址信息在存储I/O总线210和存储控制总线212上传输。

在一种实施方式中，每个固态存储元件216、218、220被分割成擦除块，并且每个擦除块被分割成页。典型的页的容量为2000字节（“2kB”）。在一个实例中，固态存储元件（如SSS 0.0）包括两个寄存器并能编程为两页，从而双寄存器固态存储元件216、218、220具有4kB的容量。20个固态存储元件216、218、220的内存库214就会有80kB的页访问容量，同时同一地址流出存储I/O总线210的通道。

在固态存储元件216、218、220的内存库214中的这一组80kB大小的页可称为虚拟页。类似地，内存库214a的每个存储元件216a-m的擦除块可被分组以形成虚拟块。在优选实施方式中，当在固态存储元件216、218、220中接收到擦除命令时，擦除位于固态存储元件216、218、220内的页擦除块。然而，在固态存储元件216、218、220内的擦除块、页、平面层或其他逻辑和物理部分的大小和数量预计会随着技术的进步而变化，可以预期的是，与新配置一致的许多实施例是可能的并与本文的一般描述相一致。

通常，当将包写入固态存储元件216、218、220内的特定位置时，其中，拟将所述包写入特定页内的位置，所述特定页对应于特定内存库的特定元件

的特定擦除块的页，在发送所述包之后通过存储I/O总线210发送物理地址。所述物理地址包含足够的信息，以使得固态存储元件216、218、220将所述包导入页内的指定位置。由于存储元件行（如SSS 0.0-SSS 0.N 216a、218a、220a）上的存储元件通过存储I/O总线210a.a内的合适总线同时被访问，为了到达合适的页并将所述数据包写入在存储元件行（SSS 0.0-SSS 0.N 216a、218a、220a）中具有相似地址的页，存储控制总线212同时选择内存库214a（包括具有要将所述数据包写入其内的正确页的固态存储元件SSS 0.0 216a）。

类似地，在存储I/O总线210上传输的读命令需要同时在存储控制总线212上传输的命令，以选择单个的内存库214a和内存库214内的合适页。在优选实施方式中，读命令读取整个页，并且由于在内存库214内存在许多并行的固态存储元件216、218、220，读命令读取整个虚拟页。然而，所述读命令可分割为子命令，这将在下文中结合内存库交错进行解释。还可以在写操作中访问虚拟页。

可通过存储I/O总线210发出的擦除块擦除命令以擦除擦除块，该擦除块具有特定的擦除块地址以擦除特定的擦除块。通常，可通过存储I/O总线210的并行通路发送擦除块擦除命令以擦除虚拟擦除块，每个虚拟擦除块具有特定的擦除块地址以擦除特定的擦除块。同时，通过存储控制总线212选择特定的内存库（如内存库-0 214a）以防止擦除所有的内存库（内存库1-N 214b-n）中的具有类似地址的擦除块。还可采用存储I/O总线210和存储控制总线212的结合将其他命令发送到特定位置。本领域技术人员会认识到采用双向存储I/O总线210和存储控制总线212选择特定存储单元的其他方法。

在一种实施方式中，将包顺序地写入固态存储器110。例如，包流到存储元件216的内存库214a的存储写入缓冲器，并且当所述缓冲器饱和时，所述包被编程入指定的虚拟页。然后所述包再次填充所述存储写入缓冲器，并且当所述存储缓冲器再次饱和时，所述包被写入下一虚拟页。所述下一个虚拟页可位于同一个内存库214a内或可位于另一个内存库（如214b）内。这个过程（一个虚拟页接一个虚拟页）通常一直持续到虚拟块被填满时。在另一种实施方式中，当这个过程（一个虚拟擦除块接一个虚拟擦除块）持续时，数据流可继续越过虚拟擦除块边界。

在读、修改、写操作中，在读操作中定位并读取与所述对象关联的数据

包。已被修改的修改对象的数据段并不写入读取它们的位置。取而代之，修改的数据段再次被转化为数据包并随后被写入正在被写入的虚拟页中的下一个可用位置。各个数据包的所述对象索引项被修改为指向包含已修改的数据段的包。所述对象索引中用于与同一对象（未被修改）关联的数据包的项（或多个项）会包括指向未被修改的数据包的源位置的指针。因此，如果源对象保持不变（例如保持所述对象的先前版本不变），所述源对象将在所述对象索引中具有指向所有与最初写入的一样的数据包的指针。新对象将在所述对象索引中具有指向一些源数据包的指针和指向正在被写入的虚拟页中的修改的数据包的指针。

在复制操作中，所述对象索引包括用于源对象的项，该源对象映射到若干存储在固态存储器110中的包。当复制完拷贝时，创建了新对象并在所述对象索引中创建将所述新对象映射到源包的新项。还将所述新对象写入固态存储器110，且所述新对象的地址映射到所述对象索引中的新项。新对象包可用于确定在源对象中的包，该包被引用以防在未复制的源对象中发生改变并以防对象索引丢失或损坏。

有利地是，顺序地写入包有助于更平滑地使用固态存储器110并允许固态存储设备控制器202监测固态存储器110内的存储热点和不同虚拟页的层使用状况。相继地写入包还可有助于建立强大、高效的垃圾收集系统，这将在下文中详细描述。本领域技术人员会认识到顺序地存储数据包的其他好处。

固态存储设备控制器

在不同的实施方式中，固态存储设备控制器202还可包括数据总线204、局部总线206、缓冲控制器208、缓冲器0-N 222a-n，主控制器224、直接存储器存取（“DMA”）控制器226、存储器控制器228、动态存储器阵列230、静态随机存储器阵列232、管理控制器234、管理总线236、连接系统总线240的网桥238和杂项逻辑块242，这些将在下文中描述。在其他实施方式中，系统总线240与一个或多个网络接口卡（“NIC”）244相连接，这些网络接口卡中的一些可包括远程DMA（“RDMA”）控制器246、一个或多个中央处理器（“CPU”）248、一个或多个外部存储器控制器250和关联的外部存储器阵列252、一个或多个存储控制器254、对等控制器256和专用处理器258，这将在下文描述。连接到系统总线240的组件244-258可位于计算内112内或者可以为其他设备。

通常，固态存储控制器104通过存储I/O总线210与固态存储器110进行数据通信。在典型的实施方式中，固态存储器布置在内存库214内，且每个内存库214包括多个并行访问的存储元件216、218、220，存储I/O总线210是多条总线的阵列，每一条总线用于内存库214内的存储元件216、218、220的每一行。正如此处所使用的，术语“存储I/O总线”可指一条存储I/O总线210或多条独立的数据总线204的阵列。在优选实施方式中，访问存储元件的行(如216、218a、220a)的每条存储I/O总线210可包括在存储元件216、218a、220a的行中访问的存储部(如擦除块)的逻辑-物理映射。如果第一存储部失效、部分失效、不可访问或出现一些其他问题时，这种映射允许映射到存储部的物理地址的逻辑地址重映射到不同的存储部。相对于图3中重映射模块314进一步解释了重映射。

还可通过系统总线240、网桥238、局部总线206、缓冲器22并最终通过数据总线204将数据从请求设备155传送到固态存储控制器104。数据总线204通常连接到一个或多个由缓冲控制器208控制的缓冲器222a-n。缓冲控制器208通常控制数据从局部总线206传递到缓冲器222并通过数据总线204传递到管道输入缓冲器306和输出缓冲器330。为了解决时钟域差异、防止数据冲突等等，缓冲控制器208通常控制在缓冲器222中暂时存储来自请求设备的数据的方式，并控制此后传送给数据总线204(或相反)的方式。缓冲控制器208通常与主控制器224结合使用以协调数据流。当数据到达时，所述数据会到达系统总线240并通过网桥238传递给局部总线206。

通常，数据在主控制器224和缓冲控制器208的控制下从局部总线206传递给一个或多个数据缓冲器222。然后，所述数据通过固态控制器104从缓冲器222流向数据总线204并到达固态存储器110(如NAND闪存或其他存储媒体)。在优选实施方式中，数据与与所述数据一起到达的关联的带外元数据(“对象元数据”)采用一个或多个的数据通道被送达，所述数据通道包括一个或多个固态存储控制器104a-104n-1和关联的固态存储器110a-110n-1，而至少一个通道(固态存储控制器104n、固态存储器110n)用于带内元数据(如索引信息和其他固态存储设备102内部生成的元数据)。

局部总线206通常为双向总线或总线组，所述双向总线或总线组允许数据和命令在固态存储设备控制器202内部的设备间通信，也允许命令和数据在固

态存储设备102内部的设备和与系统总线240连接的设备244-258之间通信。网桥238有助于在局部总线206和系统总线240之间的通信。本领域技术人员会认识到其他实施方式，如总线240、206、204、210和网桥238的环结构或交换式星形配置和功能。

系统总线240通常是计算机、安装有或连接有固态存储设备102的其他设备的总线。在一种实施方式中，系统总线240可以为PCI-e总线、串行高级技术附件（“串行ATA”）总线、并行ATA或类似总线。在另一种实施方式中，系统总线240为外部总线，例如小型计算机系统接口（“SCSI”）、防火墙、光纤通道、USB、PCIe-As或类似总线。固态存储设备102可被封装为适于置于设备内部或被封装为外部连接设备。

固态存储设备控制器202包括在固态存储设备102内控制较高级别功能的主控制器224。在不同的实施方式中，主控制器224通过解释对象请求和其他请求来控制数据流，指导创建索引，该索引将与数据关联的对象标识符映射到关联的数据（或协调的DMA请求等）的物理地址。主控制器224完全地或部分地控制此处描述的许多功能。

在一种实施方式中，主控制器224采用嵌入式控制器。在另一种实施方式中，主控制器224采用局部存储器，如动态存储器阵列230（动态随机存取存储器“DRAM”）、静态存储器阵列323（静态随机存取存储器“SRAM”）等。在一种实施方式中，采用主控制器224控制局部存储器。在另一实施方式中，主控制器通过存储器控制器228访问局部存储器。在另一种实施方式中，所述主控制器运行Linux服务器并可支持各种常用服务器接口，如万维网、超文本标记语言（“HTML”）等。在另一种实施方式中，主控制器224采用纳米处理器。可采用可编程或标准逻辑或上述控制器类型的任意组合来构建主控制器224。本领域技术人员会认识到主控制器的许多实施方式。

在一种实施方式中，其中，存储设备152/固态存储设备控制器202管理多个数据存储设备/固态存储器110a-n，主控制器224在内部控制器（如固态存储控制器104a-n）之间分配工作负载。例如，主控制器224可分割将要被写入数据存储设备（如固态存储器110a-n）中的对象，使得每个附属的数据存储设备存储所述对象的一部分。这种特征是允许更快地存储和访问对象的性能增强。在一种实施方式中，主控制器224利用FPGA实施。在另一种实施方式中，位

于主控制器224内的固件可通过管理总线236、通过网络连接到NIC244的系统总线240或其他连接到系统总线240的设备更新。

在一种实施方式中，管理对象的主控制器224仿真块存储，从而使得计算机102或其他连接到存储设备152/固态存储设备102的设备将存储设备152/固态存储设备102视为块存储设备并将数据发送给存储设备152/固态存储设备120中的特定物理地址。然后，主控制器224分配块并像存储对象一样存储数据块。然后，主控制器224将块和与块一起发送的物理地址映射到由主控制器224确定的实际位置。映射存储在对象索引中。通常，对于块仿真来说在计算机112、客户端114或其他希望将存储设备152/固态存储设备102当成块存储设备来使用的设备中提供有块设备应用程序接口（“API”）。

在另一种实施方式中，主控制器224与NIC控制器244和嵌入式RDMA控制器246协同运行以提供准时的RDMA数据和命令集传输。NIC控制器244可隐藏在非透明端口后以使得能够使用自定义的驱动器。同样地，客户端114上的驱动器可通过采用标准栈API的并与NIC244结合运行的I/O存储驱动器访问计算机网络118。

在一种实施方式中，主控制器224也是独立驱动器冗余阵列（“RAID”）控制器。当数据存储设备/固态存储设备120与一个或多个其他数据存储设备/固态存储设备120联网时，主控制器224可以是用于单层RAID、多层RAID、渐进式RAID等的RAID控制器。主控制器224还允许一些对象存储在RAID阵列内而其他对象不通过RAID存储。在另一种实施方式中，主控制器224可以是分布式RAID控制器元件。在另一种实施方式中，主控制器224可包括许多RAID、分布式RAID和另行描述的其他功能。

在一种实施方式中，主控制器224与单个或多个网络管理器（如交换机）协同运行以建立路由、平衡带宽使用率、故障转移等。在另一种实施方式中，主控制器224与集成专用逻辑器件（通过局部总线206）和关联的驱动器软件协同运行。在另一种实施方式中，主控制器224与附属专用处理器258或逻辑器件（通过外部系统总线240）和关联的驱动器软件协同运行。在另一种实施方式中，主控制器224与远程专用逻辑器件（通过计算机网络118）和关联的驱动器软件协同运行。在另一种实施方式中，主控制器224与局部总线206或附属于硬盘驱动器（“HDD”）存储控制器的外部总线协同运行。

在一种实施方式中，主控制器224与一个或多个存储控制器254通信，其中存储设备/固态存储设备120可表现为通过SCSI总线、因特网SCSI(“iSCSI”)、光纤通道等连接的存储设备。同时，存储设备/固态存储设备120可自主地管理对象并可表现为对象文件系统或分布式对象文件系统。还可通过对等控制器256和/或专用处理器258访问主控制器224。

在另一种实施方式中，主控制器224与自主集成管理控制器协同运行以周期性地验证FPGA码和/或控制器软件、在运行(复位)时验证FPGA码和/或在通电(复位)期间验证控制器软件、支持外部复位请求、支持由于检查包而超时的复位请求，并支持电压、电流、功率、温度及其他环境测量和阈值中断设置。在另一种实施方式中，主控制器224管理垃圾收集以释放擦除块用于再次使用。在另一种实施方式中，主控制器224管理耗损均衡。在另一种实施方式中，主控制器224允许数据存储设备/固态存储设备102被分割成多个虚拟设备并允许基于分区的媒体加密。在又一种实施方式中，主控制器224支持具有高级的、多位的ECC修正的固态存储控制器104。本领域技术人员会认识到位于存储控制器152内(或更具体地说位于固态存储设备102内)的主控制器224的其他特征和功能。

在一种实施方式中，固态存储设备控制器202包括存储器控制器228，该存储器控制器228控制动态随机存储器阵列230和/或静态随机存储器阵列232。如上所述，存储器控制器228可独立于主控制器224使用或与主控制器224集成使用。存储器控制器228通常控制验证一些存储器类型，如DRAM(动态随机存储器阵列230)和SRAM(静态随机存储器阵列232)。在其他实例中，存储器控制器228还控制其他存储器类型，如电可擦可编程序只读存储器(“EEPROM”)等。在其他实施方式中，存储器控制器228控制两种或更多种存储器类型且存储器控制器228可包括不止一个控制器。通常，存储器控制器228在可行情况下控制尽可能多的SRAM232，并且通过DRAM230补足SRAM232。

在一种实施方式中，所述对象索引存储在存储器230、232中并周期性的被卸载到固态存储器110n或其他非易失性存储器的通道内。本领域技术人员会认识到存储器控制器228、动态存储器阵列230、静态存储器阵列232的其他运用和配置。

在一种实施方式中，固态存储设备控制器202包括DMA控制器226，该DMA控制器226控制在下列设备之间的DMA操作：存储设备/固态存储设备102、一个或多个外部存储器控制器250、关联的外部存储器阵列252和CPU248。应该注意到，外部存储器控制器250和外部存储器阵列252之所以被称为外部是因为它们位于存储设备/固态存储设备102的外部。此外，DMA控制器226还可通过NIC244和关联的RDMA控制器246控制请求设备的RDMA操作。DMA和RDMA在下文中有详细说明。

在一种实施方式中，固态存储设备控制器202包括连接到管理总线236的管理控制器234。通常管理控制器234管理存储设备/固态存储设备102的环境指标和状态。管理控制器234可通过管理总线236监测设备温度、风扇转速、电力供应设置等。管理控制器可支持电可擦可编程序只读存储器（“EEPROM”）以存储FPGA码和控制器软件。通常，管理总线236连接到存储设备/固态存储设备102内的不同组件。管理控制器234可通过局部总线206进行警报、中断等的通信或可包括单独的到系统总线240或其他总线的连接。在一种实施方式中，管理总线236为内部集成电路（“I²C”）总线。本领域技术人员会认识到通过管理总线236连接到存储设备/固态存储设备102的组件的管理控制器234的其他功能和运用。

在一种实施方式中，固态存储设备控制器202包括杂项逻辑块242，该杂项逻辑块242可被定制为专用。通常，当固态设备控制器202或主控制器224被配置为使用FPGA或其他可配置控制器时，可基于特定应用、用户需求、存储需求等而包括定制逻辑。

数据管道

图3示出了根据本发明的位于固态存储设备102内的固态存储设备控制器104的一种实施方式300的示意性框图，该固态存储设备控制器具有写入数据管道106和读取数据管道108。实施方式300包括数据总线204、局部总线206和缓冲控制器208，这些设备大体上类似于相对于图2中固态存储设备控制器202描述的设备。所述写入数据管道包括封包器302和纠错码（“ECC”）发生器304。在其他实施方式中，所述写入数据管道包括输入缓冲器306、写入同步缓冲器308、写入程序模块310、压缩模块312、加密模块314、垃圾收集器旁路316（部分位于所述读取数据管道内）、媒体加密模块318和写入缓冲器320。

读取数据管道108包括读取同步缓冲器328、ECC纠错模块322、解包器324、对齐模块326和输出缓冲器330。在另一种实施方式中，读取数据管道108可包括媒体解密模块332、垃圾收集器旁路316的一部分、解密模块334、解压缩模块336和读取程序模块338。固态存储控制器104还可包括控制与状态寄存器340和控制队列342、内存库交错控制器344、同步缓冲器346、存储总线控制器348及多路转换器（“MUX”）350。固态控制器104的组件和关联的写入数据管道106和读取数据管道108描述如下。在其他实施方式中，可采用同步固态存储器110并且可不使用同步缓冲器308、328。

写入数据管道

写入数据管道106包括封包器302，该封包器直接地或间接地通过另一写入数据管道106的级接收将要被写入固态存储器的数据或元数据段，并创建一个或多个大小适于固态存储器110的包。所述数据或元数据段通常是对象的一部分，但也可包括整个对象。在另一种实施方式中，所述数据段是数据块的一部分，但也可包括整个数据块。通常，对象接收自计算机112、客户端114或其他计算机或设备并被以流向固态存储设备102或计算机112的数据段的形式传送给固态存储设备102。数据段也可被称为另一名称（如数据包裹），本文所提及的数据段包括对象或数据块的全部或一部分。

每个对象被存为一个或多个包。每个对象可具有一个或多个容器包。每个包包含包头。所述包头可包括包头类型字段。类型字段可包括数据、对象属性、元数据、数据段定界符（多包）、对象结构、对象连接及类似物。所述包头还可包括关于包的大小的信息（如包内的数据的字节数）。所述包的长度可由包类型确定。一个实例可能是利用数据包包头的偏移值来确定对象内数据段的位置。本领域技术人员会认识到其他包含在由封包器302添加到数据上的包头内的信息和其他添加到数据包的信息。

每个包包括包头，还可能包括来自所述数据和元数据段的数据。每个包的包头包括用于将包与包所属对象联系起来的相关信息。例如，所述包头可包括对象标识符和偏移值，该偏移值表明了用于数据包形成的数据段、对象或数据块。所述包头还可包括存储总线控制器348用以存储包的逻辑地址。所述包头还可包括关于包的大小的信息（如包内字节数）。所述包头还可包括序列号，当生成数据段或对象时，该序列号识别数据段相对于对象内的其他

包所属的位置。所述包头可包括包头类型字段。类型字段可包括数据、对象属性、元数据、数据段定界符（多包）、对象结构、对象连接及类似物。本领域技术人员会认识到其他包含在由封包器302加到数据上的包头内的信息和其他添加到数据包的信息。

写入数据管道106包括ECC发生器304，该ECC发生器为一个或多个接收自封包器302的包生成一个或多个纠错码（“ECC”）。ECC发生器304通常采用纠错算法生成ECC，该ECC与包一起存储。与包一起存储的ECC通常用于探测和纠正由于传送和存储而引起的错误。在一种实施方式中，包作为长度为N的未编码块流入ECC发生器304。计算并添加长度为S的并发位，并作为长度为N+S的编码块输出。N和S的值依赖于算法的特点，该算法被选择用于实现特定的性能、效率和鲁棒性指标。在优选实施方式中，在ECC块和包之间并没有固定关系；包可包括不止一个ECC块；ECC块可包括不止一个包；且第一包可在ECC块内的任何位置终止而第二包可始于同一ECC块内的第一包终止的位置。在优选实施方式中，ECC算法不能被动态修改。在优选实施方式中，与数据包一起存储的ECC足够稳健以在两个以上的位内纠正错误。

有利地是，采用允许不止一位的修正或甚至是两位修正的稳健ECC算法允许延长固态存储器110的使用寿命。例如，如果固态存储器110内使用闪存作为存储媒体，闪存在每个擦除周期内可被写入大约100000次不出现错误。这种使用期限可通过稳健ECC算法延长。固态存储设备102板载有ECC发生器304和相应的ECC纠错模块322，固态存储设备102可在其内部纠正错误并具有比采用不甚稳健的ECC算法（如单位错误修正）更长的使用寿命。然而，在其他实施方式中，ECC发生器304可采用不甚稳健的算法并可修正单位或双位错误。在另一种实施方式中，固态存储设备110可包括不甚可靠的存储器以增加容量，所述不甚可靠的存储器例如多级单元（“MLC”）闪存，所述不甚可靠的存储器在没有稳健ECC算法的情况下可以不充分可靠。

在一种实施方式中，写入数据管道包括输入缓冲器306，该输入缓冲器接收将要被写入固态存储器110的数据段并存储输入的数据段直到写入数据管道106的下一级，例如封包器302（或其他更复杂写入数据管道106的其他级）准备处理下一个数据段。通过使用适当容量的数据缓冲器，输入缓冲器306通常允许写入数据管道106接收和处理数据段之间存在速率差异。输入缓冲器

306还允许数据总线204将数据传送给写入数据管道106的速率大于写入数据管道106能支持的速率，从而改进数据总线204运行的效率。通常，当写入数据管道106不包括输入缓冲器306时，缓冲功能在别处（如固态存储设备102）实现，但所述别处位于写入数据管道106外、位于计算机内，例如当使用远程直接存储器读取（“RMDA”）时，如在网络接口卡（“NIC”）内或其他设备上。

在另一种实施方式中，写入数据管道106还包括写入同步缓冲器308，该写入同步缓冲器308在将包写入固态存储器110之前缓冲接收自ECC发生器304的包。写入同步缓冲器308位于本地时钟域和固态存储时钟域之间的边界上，并且提供缓冲以解决时钟域差异。在其他实施方式中，可采用同步固态存储器110而移除同步缓冲器308、328。

在一种实施方式中，写入数据管道106还包括媒体加密模块318，该媒体加密模块318直接地或间接地从封包器302接收一个或多个包，并在将包发送给ECC发生器304之前利用对固态存储设备102唯一的加密密钥加密所述一个或多个包。通常，整个包（包括包头）都被加密。在另一种实施方式中，并不加密包头。在本文中，在一种实施方式中，加密密钥被理解为意味着在外部管理的秘密加密密钥，这种密钥将固态存储器110和需要加密保护的设备集成在一起。媒体加密模块318和相应的媒体解密模块332为存储在固态存储器110中数据提供安全等级。例如，当数据利用媒体加密模块加密时，如果固态存储器110连接到不同的固态存储控制器104、固态存储设备102或计算机112，通常，在不使用同一加密密钥（在将数据写入固态存储器110期间使用）时，如果不经过合理的努力，则不能读取固态存储器110的内容。

在典型的实施方式中，固态存储设备102不将所述加密密钥存储在非易失性存储器中并且不允许从外部访问所述加密密钥。在初始化期间为固态存储控制器104提供加密密钥。固态存储设备102可使用并存储非秘密性加密临时值，该非秘密性加密临时值与加密密钥结合使用。不同的临时值可与每个包一起存储。为了加强保护，加密算法可利用唯一临时值在多个包之间分割数据段。所述加密密钥可接收自客户端114、计算机112、密钥管理器或其他管理固态存储控制器104使用的加密密钥的设备。在另一种实施方式中，固态存储器110可具有两个或更多个分区，并且固态存储控制器104显得就像有两个或更多个固态存储控制器104，每一个固态存储控制器104在固态存储器110内

的单个分区上运行。在这种实施方式中，唯一的媒体加密密钥可与每个分区一起使用。

在另一种实施方式中，写入数据管道106还包括加密模块314，该加密模块314在将数据段发送给封包器302之前直接地或间接地加密接收自输入缓冲器306的数据或元数据段，利用与数据段一同接收的加密密钥来加密数据段。加密模块314与媒体加密模块318不同，这是由于：加密模块318用以加密数据的加密密钥对存储在固态存储设备102内的数据来说不是共同的并在对象基础上可能不同，并且加密密钥可不与数据段一起接收（如下所述）。例如，加密模块318用以加密数据段的加密密钥可与数据段一起被接收或可作为将对象写入数据段所属位置的命令的一部分被接收。固态存储设备102可在每个与加密密钥结合使用的对象包中使用并存储非秘密性加密临时值。不同的临时值可与每个包一起存储。为了通过加密算法加强保护，可利用唯一临时值在多个包之间分割数据段。在一种实施方式中，媒体加密模块318使用的临时值与加密模块314使用的临时值相同。

加密密钥可接收自客户端114、计算机112、密钥管理器或其他保存用于加密数据段的加密密钥的设备。在一种实施方式中，加密密钥被从固态存储设备102、计算机112、客户端114或其他外部代理中的一个传送到固态存储控制器104，所述外部代理能够执行工业标准方法以安全地传送并保护私有密钥和公共密钥。

在一种实施方式中，加密模块318利用与第一包一起接收的第一加密密钥加密第一包，并利用与第二包一起接收的第二加密密钥加密第二包。在另一种实施方式中，加密模块318利用与第一包一起接收的第一加密密钥加密第一包，而将第二数据包传递给下一级（未经加密）。有利地是，包括在固态存储设备102的写入数据管道106内的加密模块318允许对象接对象或段接段的数据加密，而不需要单独的文件系统或其他外部系统来追踪不同的用于存储相对对象或数据段的加密密钥。每个请求设备155或相关密钥管理器独立地管理加密密钥，该加密密钥仅用于加密请求设备155发送的对象或数据段。

在另一种实施方式中，写入数据管道106包括压缩模块312，该压缩模块312在将数据段发送给封包器302之前为元数据段压缩数据。压缩模块312通常利用本领域技术人员熟知的压缩程序来压缩数据或元数据段以减少段占用的

的存储空间大小。例如，如果数据段包括一串512个0位，压缩模块312可用表明512个0位的编码或令牌来替换这512个0位，其中，所述编码所占的空间比512个0位所占的空间要小得多。

在一种实施方式中，压缩模块312利用第一压缩程序压缩第一段，而输送第二段（未经压缩）。在另一种实施方式中，压缩模块312利用第一压缩程序压缩第一段并利用第二压缩程序压缩第二段。在固态存储设备102内具有这种灵活性是有利的，以便客户端或其他将数据写入固态存储设备102内的设备中每一个都可指定压缩程序或以便一个设备指定压缩程序而另一个设备指定无压缩。还可根据每个对象类型或对象类基础的默认设置来选择压缩程序。例如，特定对象的第一对象可以能够废除默认压缩程序设置，同一对象类和对象类型的第二对象可采用默认压缩程序，而同一对象类和对象类型的第三对象可不压缩。

在一种实施方式中，写入数据管道106包括垃圾收集器旁路316，该垃圾收集器旁路316接收来自读取数据管道的108（在垃圾收集系统中作为数据旁路的一部分）的数据段。垃圾收集系统通常标记不再有效的包，不再有效的原因通常是由包被标记为删除或包已被修改且修改过的数据存储在不同的位置。在某一时刻，垃圾收集系统确定存储器的某个区域可被恢复。之所以确定某个区域可被恢复可能是由于：缺乏可用的存储空间、标记为无效的数据百分比达到阈值、有效数据的合并、存储器的该区域错误检出率达到阈值或基于数据分布提高性能等。垃圾收集算法可考虑大量的因素以确定何时存储器的区域将要被恢复。

一旦存储器的区域被标记为恢复，该区域内的有效包通常必须被重新存放。垃圾收集器旁路316允许将包读入读取数据管道108，并允许然后将包直接传送给写入数据管道106而不会将包路由出固态存储控制器104。在优选实施方式中，垃圾收集器旁路316是运行在固态存储设备102内的自主垃圾收集系统的一部分。这允许固态存储设备102管理数据，从而数据系统地传播到整个固态存储器110以提升性能、数据可靠性并避免过度使用和不充分使用固态存储器110的任何一个位置或区域，并且延长了固态存储器110的使用寿命。

垃圾收集器旁路316协调将数据段插入写入数据管道106而其他数据段由客户端116或其他设备写入。在描述的实施方式中，垃圾收集器旁路316位于

写入数据管道106内的封包器302之前、读取数据管道内的解包器314之后，但也可位于写入和读取数据管道106、108内的其他位置。可在清洗写入数据管道106期间使用垃圾收集器旁路316，以填充虚拟页的剩余部分，从而提升固态存储器110内的存储效率并因此减少垃圾收集的频率。

在一种实施方式中，写入数据管道106包括写入缓冲器320，该写入缓冲器320为了高效的写操作而缓冲数据。通常，写入缓冲器320包括用于包的足够容量，以填充固态存储器110内的至少一个虚拟页。这允许写操作将数据的整个页没有中断地发送给固态存储器110。通过选择写入数据管道106的写入缓冲器320的容量并将读取数据管道108内的缓冲器的容量选为同样大小容量或比固态存储器110内存储写入缓冲器的容量大，由于单个写入命令可被设计为将数据的整个虚拟页发送给固态存储器110，从而以单条命令替代多条命令，写入和读取数据的效率更高。

当填充写入缓冲器320时，固态存储器110可用于其他读操作。这是有利的，原因是：当将数据写入存储写入缓冲器时和注入数据缓冲器的数据失速时，具有更小容量的写入缓冲器的或不具有写入缓冲器的其他固态设备可绑定固态存储器。读操作会被拦截直到整个存储写入缓冲器被填充或被编程。用于不具写入缓冲器或具有小容量的写入缓冲器的系统的另一种方法是清洗未满的存储写入缓冲器以使得能进行读操作。同样地，由于需要多写入/编程周期来填充页，因此这种方法的效率低下。

对于描述的具有容量比虚拟页容量大的写入缓冲器320的实施方式，单个的写入命令（包括大量子命令）的后续命令可以是单个程序命令，以将来自每个固态存储元件216、218、220中的存储写入缓冲器的数据页传递给每个固态存储元件216、218、220中的指定页。这种技术带来的好处是：减少了部分页编程，众所周知，这降低了数据的可靠性和稳定性并在当缓冲器填充时，为读命令和其他命令释放了目标内存库。

在一种实施方式中，写入缓冲器320为交替缓冲器，其中，所述交替缓冲器的一侧被填充，然后当所述交替缓冲器的另一侧被填充时，所述交替缓冲器的一侧被指定为在适当的时间传送数据。在另一种实施方式中，写入缓冲器320包括先进先出（“FIFO”）寄存器，该FIFO寄存器的容量比数据段虚拟页的容量大。本领域技术人员会认识到允许在将数据写入固态存储器110之前存

储数据虚拟页的其他写入缓冲器320配置。

在另一种实施方式中，写入缓冲器320的容量比虚拟页小，从而少于一页的信息可被写入固态存储器110内的存储写入缓冲器。在这种实施方式中，为了防止写入数据管道106的失速阻止读操作，采用需要从一个位置移动到另一个位置的垃圾收集系统将数据排队，这个过程是垃圾收集进程的一部分。为了防止写入数据管道106中的数据失速，可通过垃圾收集器旁路316将所述数据供应给写入缓冲器320并然后将所述数据供应给固态存储器110中的存储写入缓冲器，从而在编程所述数据之前填充虚拟页的页面。这样，写入数据管道106中的数据失速不会使读取自固态存储设备102的数据失速。

在另一种实施方式中，写入数据管道106包括写入程序模块310，该写入程序模块310具有写入数据管道106内的一个或多个用户可定义的功能。写入程序模块310允许用户自定义写入数据管道106。用户可基于特定数据请求或应用自定义写入数据管道106。当固态存储控制器104为FPGA时，用户可相对轻松地编程具有自定义命令和功能的写入数据管道106。用户还可利用写入程序模块310以使ASIC包括自定义功能，然而自定义ASIC可能比使用FPGA时更困难。写入程序模块310可包括缓冲器和旁路机制，以允许第一数据段在写入程序模块310中执行，而第二数据段通过写入数据管道106可继续传送。在另一种实施方式中，写入程序模块310可包括能通过软件编程的处理器内核。

应注意，写入程序模块310被示为位于输入缓冲器306和压缩模块312之间，然而写入程序模块310可位于写入数据管道106内的任何位置，并且可分布在不同的级302-320之间。此外，在不同的、已编程的且独立运行的级302-320之间可分布有多个写入程序模块310。此外，级302-320的顺序可以改变。本领域技术人员会认识到基于特定用户需求的级302-320的顺序的可行改变。

读取数据管道

读取数据管道108包括ECC纠错模块322，该ECC纠错模块322通过使用与请求包中的每个ECC块一起存储的ECC来确定接收自固态存储器110的请求包的ECC块中是否存在错误。然后，如果存在任何错误并且所述错误可使用ECC修正，则ECC纠错模块322修正请求包中的任何错误。例如，如果ECC能够探测6位的错误但只能修正3位的错误，那么ECC纠错模块322修正具有3位错误

的请求包ECC块。ECC纠错模块322通过把出错的位改变为正确的1或0状态来修正出错的位，从而请求数据包与其被写入固态存储器110并且为包生成ECC时一致。

如果ECC纠错模块322确定请求包包含了比ECC能修正的位数多的出错位，则ECC纠错模块322不能修正请求包毁坏的ECC块的错误并发送中断。在一种实施方式中，ECC纠错模块322发送中断以及指示请求包出错的消息。所述消息可包括指出ECC纠错模块322不能修正错误或ECC纠错模块322没有能力修正错误的信息。在另一种实施方式中，ECC纠错模块322与所述中断和/或消息一起发送请求包中毁坏的ECC块。

在优选的实施方式中，请求包中毁坏的ECC块或毁坏的ECC块的一部分（不能被ECC纠错模块322修正）由主控制器224读取，并被修正和返回给ECC纠错模块322以被读取数据管道108进一步处理。在一种实施方式中，请求包中毁坏的ECC块或毁坏的ECC块的一部分被发送给请求数据的设备。请求设备155可修正所述ECC块或用另一拷贝替换数据（如备份或镜像拷贝），然后可使用请求数据包的替换的数据或将所述替换的数据返回给读取数据管道108。请求设备155可使用出错请求包中的包头信息以识别替换毁坏请求包或替换包所属的对象所需的数据。在另一种优选实施方式中，固态存储控制器104采用一些类型的RAID存储数据并能够恢复毁坏的数据。在另一种实施方式中，ECC纠错模块322发送中断和/或消息，并且接收设备停止与请求数据包关联的读操作。本领域技术人员会认识到ECC纠错模块322确定请求包的一个或多个ECC块为毁坏的且ECC纠错模块322不能修正错误后采取的其他选择和操作。

读取数据管道108包括解包器324，该解包器324直接地或间接地接收来自ECC修正模块322的请求包ECC块，并检查和删除一个或多个包头。解包器324可通过检查包头内的包标识符、数据长度、数据位置等验证包头。在一种实施方式中，所述包头包括散列码，该散列码可用于验证传递给读取数据管道108的包为请求包。解包器324还从请求包中删除由封包器302添加的包头。解包器324可被指定为不对某些包起作用而将这些包未经修改地向前传送。一个实例可以是容器标签，当对象索引重建模块272需要包头信息时，该容器标签在重建进程期间被请求。另外的实例包括传送不同类型的包（预定在固态存储设备102内使用）。在另一种实施方式中，解包器324操作可以依赖于包的

类型。

读取数据管道326包括对齐模块326，该对齐模块326接收来自解包器324的数据并删除多余的数据。在一种实施方式中，发送给固态存储器110的读命令恢复数据包。请求数据的设备可不需要恢复的数据包内的所有数据，并且对齐模块326删除多余的数据。如果恢复页内的所有数据都是请求的数据，对齐模块326不删除任何数据。

对齐模块326在数据段传输到下一级之前以与请求数据段的设备兼容的形式按对象的数据段重新格式化数据。通常，由于数据由读取数据管道108处理，数据段或包的大小在不同级间改变。对齐模块326使用接收到的数据以将数据格式化为适于发送给请求设备155的数据段，该数据段还适于连接在一起以形成响应。例如，来自第一数据包的一部分的数据可与来自第二数据包的一部分的数据结合。如果数据段比由请求设备请求的数据大，对齐模块326可丢弃不需要的数据。

在一种实施方式中，读取数据管道108包括读取同步缓冲器328，该读取同步缓冲器328在读取数据管道108处理之前缓冲一个或多个读取自固态存储器110的请求包。读取同步缓冲器328位于固态存储时钟域和本地总线时钟域之间的边界上并提供缓冲以解决时钟域差异。

在另一种实施方式中，读取数据管道108包括输出缓冲器330，该输出缓冲器330接收来自对齐模块326的请求包并在数据包传送到所述请求设备前存储该包。输出缓冲器330解决当从读取数据管道108接收数据段时和当将数据段传送给固态存储控制器104的其他部分或传送给请求设备时之间的差异。输出缓冲器330还允许数据总线以比读取数据管道108能够支持的速率高的速率接收来自读取数据管道108的数据，以提升数据总线204运行的效率。

在一种实施方式中，读取数据管道108包括媒体解密模块332，该媒体解密模块332接收一个或多个来自ECC纠错模块322的加密过的请求包并在将一个或多个所述请求包发送给解包器324之前利用对于固态存储设备102唯一的加密密钥解密一个或多个所述请求包。通常，媒体解密模块332用以解密数据的加密密钥与媒体加密模块318使用的加密密钥一致。在另一种实施方式中，固态存储器110可具有两个或更多个分区且固态存储控制器104表现得好像有两个或更多个固态存储控制器104（每个都在固态存储器110内的单独分区内

运行)一样。在这种实施方式中，可对每个分区使用唯一的媒体加密密钥。

在另一种实施方式中，读取数据管道108包括解密模块334，该解密模块334在将数据段发送给输出缓冲器330之前解密由解包器324格式化的所述数据段。采用与读请求一起接收的加密密钥解密所述数据段，所述读请求初始化恢复由读取同步缓冲器328接收的请求包。解密模块334可利用与用于第一包的读请求一起接收的加密密钥解密第一包，然后可利用不同的加密密钥解密第二包或可将第二包未经解密地传送给读取数据管道108的下一级。通常，解密模块334使用与媒体解密模块332用以解密请求数据包的加密密钥不同的加密密钥解密数据段。当包与非秘密性加密临时值一起存储时，该临时值与加密密钥一起使用以解密数据包。加密密钥可接收自客户端114、计算机112、密钥管理器或管理固态存储控制器104使用的加密密钥的其他设备。

在另一种实施方式中，读取数据管道108包括解压缩模块336，该解压缩模块336解压缩由解包器324格式化的数据段。在优选实施方式中，解压缩模块336使用存储在包头和容器标签中的一个或两个中的压缩信息以选择补充程序，压缩模块312使用该补充程序来压缩数据。在另一种实施方式中，解压缩模块336所使用的解压缩程序由请求解压缩的数据段确定。在另一种实施方式中，解压缩模块336根据每个对象类型或对象类基础的默认设置选择解压缩程序。第一对象的第一包可以能够废除默认解压缩程序设置，具有相对的对象类和对象类型的第二对象的第二包可采用默认解压缩程序，而具有相同的对象类和对象类型的第三对象的第三包可不经过解压缩。

在另一种实施方式中，读取数据管道108包括读取程序模块338，该读取程序模块338包括一个或多个在读取数据管道108内的用户可定义功能。读取程序模块338具有与写入程序模块310类似的特点并允许用户提供自定义功能给读取数据管道108。读取程序模块338可位于图3中所示的位置、可位于读取数据管道108内的其他位置、或者可包括读取数据管道108内多个位置的多个部分。此外，在读取数据管道108内的多个不同位置可有多个独立运行的读取程序模块338。本领域技术人员会认识到读取数据管道108内的读取程序模块338的其他形式。正如写入数据管道，读取数据管道108的级可重新排序，本领域技术人员会认识到读取数据管道108内的级的其他排列顺序。

固态存储控制器104包括控制和状态寄存器340和相应的控制队列342。控

制和状态寄存器340和控制队列342有助于控制并按顺序排列与在写入和读取数据管道106、108内处理的数据相关联的命令和子命令。例如，封包器302中的数据段可具有一个或多个在与ECC发生器关联的控制队列342内的相应控制命令或指令。当数据段被封包时，可在封包器302内执行一些指令或命令中。当从数据段建立的、最新形成的数据包被传送给下一级时，其他命令或指令可通过控制和状态寄存器340直接传送给下一个控制队列342。

可同时将命令和指令加载到控制队列342上以将包转发给写入数据管道106，同时，由于每个管道级要执行各自的包，因此每个管道级读取合适的命令或指令。类似地，可同时将命令和指令加载到控制队列342上以从读取数据管道108请求包，而且，由于每个管道级要执行各自的包，因此每个管道级读取合适的命令或指示。本领域技术人员会认识到控制和状态寄存器340和控制队列342的其他特征和功能。

固态存储控制器104和/或固态存储设备102还包括内存库交错控制器344、同步缓冲器346、存储总线控制器348及多路转换器(“MUX”)350，这些设备相对于图4A和图4B描述。

内存库交错

图4A是根据本发明的位于固态存储控制器104内的内存库交错控制器344一种实施方式400的示意性框图。内存库交错控制器344连接到控制和状态寄存器340并通过MUX350、存储总线控制器348和同步缓冲器346连接到存储I/O总线210和存储控制总线212上，这在下文中有所描述。内存库交错控制器包括读取代理402、写入代理404、擦除代理406、管理代理408、读取队列410a-n、写入队列412a-n、擦除队列414a-n、用于固态存储器110中的内存库214的管理队列416a-n、内存库控制器418a-n、总线仲裁器420和状态MUX422，这些设备在下文中描述。存储总线控制器348包括具有重映射模块430的映射模块424、状态捕捉模块426和NAND总线控制器438，这些设备在下文中描述。

内存库交错控制器344将一条或多条命令送往内存库交错控制器344中的两个或更多个队列，并在固态存储器110的内存库214之间协调存储在队列中的命令的执行，以使得第一类型的命令在一个内存库214a上执行而第二类型的命令在第二内存库214b上执行。所述一条或多条命令按命令类型分别送入队列中。固态存储器110的每个内存库214在内存库交错控制器344内具有相应

的队列集，且每个队列集包括每个命令类型的队列。

内存库交错控制器344在固态存储器110的内存库214之间协调存储在队列中的命令的执行。例如，第一类型的命令在在一个内存库214a上执行而第二类型的命令在第二内存库214b上执行。通常，命令类型和队列类型包括读取和写入命令和队列410、412，但是还可包括存储媒介指定的其他命令和队列。例如，在图4A所描述的实施方式中，擦除和管理队列414、416被包括在其中且适于闪存、NRAM、MRAM、DRAM、PRAM等。

对于其他类型的固态存储器110，可包括其他类型的命令和相应的队列而不脱离本发明的范围。FPGA固态存储控制器104的灵活性允许存储媒介的灵活性。如果将闪存换成另一种固态存储类型，可改变内存库交错控制器344、存储总线控制器348和MUX350以适应媒介类型而不显著地影响数据管道106、108和其他固态存储控制器104运行。

在图4A所描述的实施方式中，对每个内存库214来说，内存库交错控制器344包括：用于从固态存储器110读取数据的读取队列410、用于将命令写入固态存储器110的写入队列412、用于擦除固态存储器中的擦除块的擦除队列414、用于管理命令的管理队列416。内存库交错控制器344还包括相应的读取、写入、擦除和管理代理402、404、406、408。在另一种实施方式中，控制和状态寄存器340和控制队列342或类似元件在没有内存库交错控制器344的情况下为了发送给固态存储器110的内存库214的数据而将命令排队。

在一种实施方式中，代理402、404、406、408将预定用于特定内存库214a的合适类型的命令送到内存库214a的修正队列。例如，读取代理402可接收用于内存库-1 214b的读命令并将所述读命令送到内存库-1读取队列410b。写入代理404可接收将数据写入固态存储器110的内存库-0 214a的写入命令并然后会将所述写入命令发送给内存库-0写入队列412a。类似地，擦除代理406可接收擦除命令以擦除内存库-1 214b中的擦除块并然后会将所述擦除命令传送给内存库-1擦除队列414b。管理代理408通常接收管理命令、状态请求及其类似消息，如复位命令或读取内存库214(如内存库-0 214a)的配置寄存器的请求。管理代理408将所述管理命令发送给内存库-0管理队列416a。

代理402、404、406、408通常还监测队列410、412、414、416的状态并当队列402、404、406、408满、接近满、丧失功能时，发送状态、中断或其

他消息。在一种实施方式中，代理402、404、406、408接收命令并生成相应的子命令。在一种实施方式中，代理402、404、406、408通过控制和状态寄存器340接收命令并生成相应的子命令，所述子命令被转发给队列410、412、414、416。本领域技术人员会认识到代理402、404、406、408的其他功能。

队列410、412、414、416通常接收命令并存储所述命令直到所述命令被要求传送给固态存储器内存库214。在典型的实施方式中，队列410、412、414、416是先进先出（“FIFO”）寄存器或以FIFO运行的类似组件。在另一种实施方式中，队列410、412、414、416按与数据、重要性或其他标准相匹配的顺序来存储命令。

内存库控制器418通常接收来自队列410、412、414、416的命令并生成合适的子命令。例如，内存库-0写入队列412a可接收将数据包的页写入内存库-0 214a的命令。内存库-0控制器418a可在合适的时间接收写入命令并可为每个存储在写入缓冲器320中的数据包生成一个或多个写入子命令（将要被写入内存库-0 214a的页中）。例如，内存库-0控制器418a可生成验证内存库-0 214a和固态存储阵列216状态的命令、选择写入一个或多个数据包的合适位置的命令、清除位于固态存储阵列216内的输入缓冲器的命令、将一个或多个数据包传送所述输入缓冲器的命令、将输入缓冲器放到选定位置中的命令、检验数据被正确编程的命令，并且如果发生程序故障，则一次或多次地中断主控制器、重试写入同一物理地址并重试写入不同的物理地址。此外，与实例中的写入命令一起，存储总线控制器348会将一条或多条命令乘以每条存储I/O总线210a-n从而翻倍，而所述命令的逻辑地址映射到用于存储I/O总线210a的第一物理地址，并映射到用于存储I/O总线210a的第二物理地址，下面将详细描述。

通常，总线仲裁器420选自内存库控制器418并从内存库控制器418的输出队列提取子命令，并且将这些子命令以最优化内存库214性能的序列形式发给存储总线控制器348。在另一种实施方式中，总线仲裁器420可响应高级中断并修改普通选择标准。在另一种实施方式中，主控制器224可通过控制和状态寄存器340控制总线仲裁器420。本领域技术人员会认识到总线控制器420可控制和交错从内存库控制器418传送到固态存储器110的命令序列。

通常，总线仲裁器420协调来自内存库控制器418适当的命令和命令类型所需的相应数据的选择，并将所述命令和数据发送给存储总线控制器348。总

线仲裁器420通常还将命令发送给存储控制总线212以选择合适的内存库214。对于闪存或其他具有异步、双向串行的存储I/O总线210的固态存储器110而言，一次只能传送一条命令（控制信息）或数据集。例如，当将写入命令或数据通过存储I/O总线210传送给固态存储器110时，读取命令、读取的数据、擦除命令、管理命令或其他状态命令不能在存储I/O总线210上传输。例如，当从存储I/O总线210读取数据时，不能向固态存储器110写入数据。

例如，在内存库-0的写操作期间，总线仲裁器420选择在其队列顶部具有写入命令或一系列写入子命令的内存库-0控制器418a，所述一系列写入子命令使得存储总线控制器348执行后继的序列。总线仲裁器420将写入命令转发给存储总线控制器348，该存储总线控制器348通过下列方式建立了写入命令：通过存储控制总线212选择内存库-0 214a、发送清除与内存库-0 214a关联的固态存储元件110的输入缓冲器的命令、发送验证与内存库-0 214a关联的固态存储元件216、218、220的状态的命令。然后，存储总线控制器348通过包含了物理地址存储I/O总线210传送写入命令，该物理地址如同映射自逻辑擦除块地址一样包括用于每个单独的物理擦除固态存储元件216a-m的逻辑擦除块地址。然后，存储总线控制器348通过多路转换器350将写入缓冲器经写入同步缓冲器多路复用到存储I/O总线210并使写入数据流向合适的页。当所述页写满时，然后，存储总线控制器348促使与内存库-0 214a关联的固态存储元件216a-m将输入缓冲器编入固态存储元件216a-m的内存单元。最终，存储总线控制器348验证状态以确保所述页被正确编程。

读操作与上文的写操作实例类似。在读操作期间，通常，总线仲裁器420或内存库交错控制器344的其他组件接收数据和相应状态信息并将数据发送给读取数据管道108，同时将状态信息发送给控制和状态寄存器340。通常，从总线仲裁器420传送给存储总线控制器348的读数据命令会促使多路转换器350将读数据通过存储I/O总线210传送给读取数据管道108并通过状态多路转换器422向控制和状态寄存器340发送状态信息。

总线仲裁器420协调不同的命令类型和数据存取模式，使得在任意给定的时间内，在总线上只有合适的命令类型或对应数据。如果总线仲裁器420已选择了写入命令，且写入子命令和对应数据正在被写入固态存储器110，总线仲裁器420不会允许在存储I/O总线210存在其他命令类型。有利地是，总线仲裁

器420使用定时信息（如预定的命令执行时间）以及接收到的关于内存库214状态的信息，以协调总线上不同命令的执行，这样做的目标是最小化或消除总线的停工时间。

通过总线仲裁器420的主控制器224通常使用存储在队列410、412、414、416中的命令的预定完成时间以及状态信息，使得在一个内存库214a上执行与命令关联的子命令时，而在其他内存库214b-n上执行其他命令的其他子命令。当内存库214a完全执行完一条命令时，总线仲裁器420将其他命令传给内存库214a。总线仲裁器420还可与协调存储在队列410、412、414、416的命令一起协调不存储在队列410、412、414、416的其他命令。

例如，可发出擦除命令以擦除固态存储器110内的一组擦除块。执行擦除命令可消耗比执行写入或读取命令多10到1000倍的时间，或消耗比执行程序命令多10到100倍的时间。对于N个内存库214，内存库交错控制器可将擦除命令分割为N条命令，每条命令擦除内存库214a的虚拟擦除块。当内存库-0 214a执行擦除命令时，总线仲裁器420可选择在其他内存库214b-n上执行的其他命令。总线仲裁器420还可与其他组件（如存储总线控制器348、主控制器224等）一起工作以在总线之间协调命令的执行。利用总线仲裁器420、内存库控制器418、队列410、412、414、416、和内存库交错控制器的代理402、404、406、408协调命令的执行可显著的提升性能（相比于其他没有内存库交错功能的固态存储系统）。

在一种实施方式中，固态控制器104包括一个内存库交错控制器344，该内存库交错控制器344为固态存储器110的所有存储元件216、218、220提供服务。在另一种实施方式中，固态控制器104内存库包括用于每个存储元件行216a-m、218a-m、220a-m的交错控制器344。例如一个内存库交错控制器344服务存储元件的一行SSS 0.0-SSS 0.N 216a、218a、220a，第二内存库交错控制器344服务存储元件的第二行SSS 1.0-SSS 1.N 216b、218b、220b，等等。

图4B示出了根据本发明的位于固态存储设备内的内存库交错控制器的一种替代实施方式401的示意性框图。图4B所示实施方式中描述的组件210、212、340、346、348、350、402-430大体上与相对于图4A描述的内存库交错装置400类似，除了下述不同点：每个内存库214包括单独的队列432a-n及用于内存库的（如内存库-0 214a）读取命令、写入命令、擦除命令、管理命令

等被传送给内存库214的单独队列432a。在一种实施方式中，队列432是FIFO。在另一种实施方式中，队列432可具有以不同于存储的顺序的顺序从队列432中提取的命令。在另一种替代实施方式（未示出）中，读取代理402、写入代理404、擦除代理406和管理代理408可结合成单个代理，所述单个代理将命令分配给合适的队列432a-n。

在另一种替代的实施方式（未示出）中，命令存储在单独的队列中，其中，可以以不同于存储的顺序的顺序从队列中提取命令，从而使得内存库交错控制器344在余下的内存库214b-n上执行。本领域技术人员会轻易地认识到其他能够在一个内存库214a上执行命令而在其他内存库214b-n上执行其他命令的队列配置和类型。

特定存储组件

固态存储控制器104包括同步缓冲器346，该同步缓冲器346从固态存储器110发送和接收的命令和状态消息。同步缓冲器346位于固态存储时钟域和本地总线时钟域之间的边界上，并提供缓冲以解决时钟域差异。同步缓冲器346、写入同步缓冲器308和读取同步缓冲器328可独立地或共同运作以缓冲数据、命令、状态消息等等。在优选实施方式中，同步缓冲器346所处的位置使得跨越时钟域的信号数量最少。本领域技术人员会认识到：时钟域间的同步可任意运行在固态存储设备102的其他位置，以优化设计实施方案的某些方面。

固态存储控制器104包括存储总线控制器348，该存储总线控制器348解释和翻译用于发送或读取自固态存储器110的数据的命令并基于固态存储器110的类型接收自固态存储器110的状态消息。例如，存储总线控制器348可针对不同的存储类型、不同性能特点、不同制造商的存储器等而具有不同的定时要求。存储总线控制器348还将控制命令发送给存储控制总线212。

在优选实施方式中，固态存储控制器104包括MUX350，该MUX350包括多路转换器350a-n的阵列，其中，每个多路转换器用于固态存储阵列110的一行。例如，多路转换器350a与固态存储元件216a、218a、220a关联。MUX350通过存储总线控制器348、同步缓冲器346和内存库交错控制器344将来自写入数据管道106的数据和来自存储总线控制器348的命令经存储I/O总线210路由至固态存储器110，并将来自固态存储器110的数据和状态消息经存储I/O总线210路由至读取数据管道108和控制和状态寄存器340。

在优选实施方式中，固态存储控制器104包括用于固态存储元件的每一行的（如SSS 0.1 216a、SSS 0.2 218a、SSS 0.N 220a）的MUX350。MUX350将来自写入数据管道106的数据和发送给固态存储器110的命令通过存储I/O总线210结合起来，并将需要由读取数据管道108处理的数据从命令中分离出来。存储在写入缓冲器320中的包通过用于固态存储元件的每一行（SSS x.0 to SSS x.N 216、218、220）的写入缓冲器308由写入缓冲器外的总线传给用于固态存储元件的每一行（SSS x.0 to SSS x.N 216、218、220）的MUX350。MUX350从存储I/O总线210接收命令和读取数据。MUX350还将状态消息传给存储总线控制器348。

存储总线控制器348包括映射模块424。映射模块424将擦除块的逻辑地址映射到擦除块的一个或多个物理地址。例如，每个内存库214a具有20个存储元件的阵列（如SSS 0.0至SSS M.0 216）的固态存储器110可具有映射到擦除块的20个物理地址的特定擦除块的逻辑地址（每个存储元件有一个物理地址）。由于平行访问存储元件，所以位于存储元件216a、218a、220a的行中的每个存储元件中的同一位置的擦除块会分享物理地址。为了选择一个擦除块（如在存储元件SSS 0.0 216a中）代替行（如在存储元件SSS 0.0、0.1、... 0.N 216a、218a、220a中）中的所有擦除块，可选择一个内存库（在这种情况下为内存库-0 214a）。

这种用于擦除块的逻辑到物理的映射是有好处的，这是由于如果一个擦除块已损坏或不可访问，所述映射可改为映射到另一擦除块。当一个元件的擦除块出错时，这种方法减少了失去整个虚拟擦除块的损失。重映射模块430将擦除块的逻辑地址的映射改为虚拟擦除块的一个或多个物理地址（遍布存储元件的阵列）。例如，虚拟擦除块1可映射到存储元件SSS 0.0 216a的擦除块1、映射到存储元件SSS 1.0 216b的擦除块1、... 和映射到存储元件M.0 216m，虚拟擦除块2可映射到存储元件SSS 0.1 218a的擦除块2、映射到存储元件SSS 1.1 218b的擦除块2、... 和映射到存储元件M.1 218m，等等。

如果存储元件SSS 0.0 216a的擦除块1损坏、由于损耗遇到错误或由于一些原因不能被使用，重映射模块可将从逻辑到物理的映射改为指向虚拟擦除块1的擦除块1的逻辑地址的映射。如果存储元件SSS 0.0 216a的空闲擦除块（将其称为擦除块221）可用且当前并未被映射，重映射模块可改变虚拟擦除块1的

映射为映射到指向存储元件SSS 0.0 216的擦除块221，而继续指向存储元件SSS 1.0 216b的擦除块1、存储元件SSS 2.0（未示出）的擦除块1、...和指向存储元件M.0 216m。映射模块424或重映射模块430可按固定顺序映射擦除块（虚拟擦除块1到存储元件的擦除块1，虚拟擦除块2到存储元件的擦除块2，等等）或可按基于其他一些标准的顺序映射存储元件216、218、220的擦除块。

在一种实施方式中，可按访问时间分组擦除块。按访问时间分组、均衡命令执行的时间（如将数据编入或写入指定擦除块的页）可平均命令补齐，从而使得在虚拟擦除块的擦除块之间执行的命令不会由于最慢的擦除块而被限制。在另一种实施方式中，可按损耗程度、运行状况来分组擦除块。本领域技术人员会认识到当映射或重映射擦除块时需要考虑的其他问题。

在一种实施方式中，存储总线控制器348包括状态捕捉模块426，该状态捕捉模块426接收来自固态存储器110的状态消息并将该状态消息发送给状态MUX422。在另一种实施方式中，当固态存储器110为闪存时，存储总线控制器348包括NAND总线控制器428。NAND总线控制器428将命令从读取和写入数据管道106、108传送给固态存储器110中的正确位置，并根据所述闪存的特点协调命令执行的时间，等等。如果固态存储器110为另一种类型的固态存储器，则将NAND总线控制器428替换为针对存储类型的总线控制器。本领域技术人员会认识到NAND总线控制器428的其他功能。

流程图

图5A是根据本发明的在固态存储设备102内采用数据管道管理数据的方法500的一种实施方式的示意性流程图。方法500始于步骤502，输入缓冲器306接收一个或多个将要被写入固态存储器110的数据段（步骤504）。通常来说，所述一个或多个数据段包括对象的至少一部分，但也可以是整个对象。封包器302可创建一个或多个对象指定包以及对象。封包器302为每个包添加包头，所述包头通常包括包的长度和对象内包的序列号。封包器302接收一个或多个存储在输入缓冲器306的数据或元数据段（步骤504），并通过创建一个或多个大小适于固态存储器110的包来封包所述一个或多个数据或元数据段（步骤506），其中，每个包包括一个包头和来自一个或个段的数据。

通常，第一包包括对象标识符，该对象标识符确定对象，为了该对象而创建包。第二包可包括具有信息的包头，该信息由固态存储设备102用于关联

第二包和第一包中确定的对象，该包头还具有在对象内定位第二包的偏移信息和数据。固态存储设备控制器202管理内存库214和包流向的物理区域。

ECC发生器304接收来自封包器302的包并为数据包生成ECC(步骤508)。通常，在包和ECC块之间没有固定关系。ECC块可包括一个或多个包。包可包括一个或多个ECC块。包可始于ECC块内的任意位置并可在ECC块内的任意位置结束。包可始于第一ECC块内的任意位置并可在相继的ECC块中的任意位置结束。

写入同步缓冲器308在将ECC块写入固态存储器110之前缓冲分布在对应ECC块中的包（步骤510），然后固态存储控制器104在考虑到时钟域差异的适当的时间写入数据（步骤512），方法500终止于步骤514。写入同步缓冲器308位于本地时钟域和固态存储器110时钟域的边界上。注意到为方便起见，方法500描述了接收一个或多个数据段并写入一个或多个数据包，但通常接收数据段流或组。通常，若干包括完整固态存储器110的虚拟页的ECC块被写入固态存储器110。通常，封包器302接收某个大小的数据段并生成另一大小的包。这必然需要数据或元数据段或数据或元数据段的部分结合起来，以形成将段的所有数据捕捉进包的数据包。

图5B是示出了根据本发明的用于服务器内SAN的方法的一种实施方式的示意性流程图。方法500开始于522，并且存储通信模块162促进第一存储控制器152a和位于第一服务器112a外部的至少一个设备之间的通信（步骤554）。第一存储控制器152a和外部设备之间的通信独立于第一服务器112a。第一存储控制器112a位于第一服务器112a内部，并且第一存储控制器152a控制至少一个存储设备154a。第一服务器112a包括与第一服务器112a和第一存储控制器152a搭配使用的网络接口156a。服务器内SAN模块164应答存储请求（步骤556）并且方法501终止于558。服务器内SAN模块164使用网络协议和/或总线协议应答存储请求（步骤556）。服务器内SAN模块164独立于第一服务器112a应答存储请求（步骤556）并且从客户端114、114a接收服务请求。

图6是根据本发明的在固态存储设备102内采用数据管道管理数据的方法600的再一种实施方式的示意性流程图。方法600始于步骤602，输入缓冲器306接收一个或多个将要被写入固态存储器110的数据或元数据段（步骤604）。封包器302为每个包添加包头，所述包头通常包括对象内包的长度。封包器302

接收一个或多个存储在输入缓冲器306中的段（步骤604），并通过创建一个或多个大小适于固态存储器110的包来封包一个或多个段（步骤606），其中每个包包括包头和来自一个或多个段的数据。

ECC发生器304接收来自封包器302的包并生成一个或多个用于包的ECC块（步骤608）。写入同步缓冲器308在将ECC块写入固态存储器110之前缓冲分布在对应ECC块中的包（步骤610），然后固态存储控制器104在考虑到时钟域差异的合适的时间写入数据（步骤612）。当从固态存储器110请求数据时，包括一个或多个数据包的ECC块被读入读取同步缓冲器328并被缓冲（步骤614）。通过存储I/O总线210接收包的ECC块。由于存储I/O总线210是双向，当读取数据时，写操作、命令操作等被停止。

ECC纠错模块322接收暂存在读取同步缓冲器328中的请求包的ECC块，并在必要时修正每个ECC块中的错误（步骤616）。如果ECC纠错模块322确定在ECC块中存在一个或多个错误并且错误可利用ECC一并修正，ECC纠错模块322修正ECC块中的错误（步骤616）。如果ECC纠错模块322确定探测到的错误不可用ECC修正，则ECC纠错模块322发送中断。

解包器324在ECC纠错模块322修正任何错误之后接收请求包（步骤618）并通过检查和删除每个包的包头解包所述包（步骤618）。对齐模块326接收经过解包的包、删除多余的数据、并采用与请求数据段的设备兼容的形式按对象的数据段重新格式化所述数据（步骤620）。输入缓冲器330接收经过解包的请求包，并在包传送给请求设备之间缓冲包（步骤622），方法600终止于步骤624。

图7是示意性流程图，示出了根据本发明的在固态存储设备102内利用内存库交错管理数据的方法700的一种实施方式。方法600始于步骤602，内存库交错控制器344将一条或多条命令传给两个或多个队列410、412、414、416（步骤604）。通常，代理402、404、406、408根据命令类型将所述命令传给队列410、412、414、416（步骤604）。队列410、412、414、416的每个集包括用于每个命令类型的队列。内存库交错控制器344在内存库214之间协调存储在队列410、412、414、416的所述命令的执行（步骤606），以使得第一类型的命令在一个内存库214a上执行，而第二类型的命令在第二内存库214b上执行，方法600结束于步骤608。

存储空间恢复

图8是示出了根据本发明的用于固态存储设备102中无用存储单元收集的装置800的一种实施方式的示意性框图。装置800包括顺序存储模块802、存储部选择模块804、数据恢复模块806、和存储部恢复模块808，这些模块描述如下。在其他实施方式中，装置800包括无用存储单元标记模块812和擦除模块810。

装置800包括顺序存储模块802，该顺序存储模块802将数据包顺序地写入存储部内的页。无论是新的包还是修改过的包，这些包都按顺序存储。在这种实施方式中，通常不将修改过的包写回其先前存储的位置。在一种实施方式中，顺序存储模块802将包写入存储部的页中的第一位置，然后写入该页中的下一个位置，并继续写入下一个位置和再下一个位置，直到该页被写满。然后，顺序存储模块802开始填充所述存储部中的下一页。这个过程一直持续到所述存储部被写满。

在优选实施方式中，顺序存储模块802开始将包写入内存库（内存库-0 214a）的存储元件（如SSS 0.0 到SSS M.0 216）中的存储写入缓冲器。当所述存储写入缓冲器写满时，固态存储控制器104使得所述存储写入缓冲器中的数据被编入内存库214a的存储元件216中的指定页。然后，另一个内存库（如内存库-1 214b）被选定，并且当一个内存库-0编程所述指定页时，顺序存储模块802开始将包写入内存库214b的存储元件218的存储写入缓冲器。当内存库214b的存储写入缓冲器写满时，该存储写入缓冲器中的内容被编入每个存储元件218中的另一指定页。这个过程是有效率的，这是因为当一个内存库214a编程页时，可填充另一个内存库214b的存储写入缓冲器。

存储部包括固态存储设备102中的固态存储器110的一部分。通常，存储部是擦除块。对于闪存来说，擦除块上的擦除操作通过充电每个单元来将一写入擦除块中的每个位。相比于以都是1的位置开始的编程操作，这是冗长的过程，并且随着数据写入，通过放电用0写入的单元将某些位改变为0。然而，在固态存储器110不是闪存或具有其中擦除周期花费与其他操作，例如读取或编程类似的时间量的闪存的情况下，可能不需要擦除存储部。

正如此处所使用的，存储部在大小上与擦除块等同，但可（或可不）被擦除。当在此处使用擦除块时，擦除块可指存储元件（如SSS 0.0 216a）内指

定大小的特定区域，并通常包括一定数量的页。当“擦除块”与闪存结合使用时，擦除块通常是在写入之前被擦除的存储部。当“擦除块”与“固态存储器”一起使用时，擦除块可（或可不）被擦除。正如此处所使用的，擦除块可包括一个擦除块或擦除块组，存储元件（如SSS 0.0到SSS M.0 216a-n）的每一行都具有该擦除块组中的一个擦除块，擦除块或擦除块组在此处还可被称为虚拟擦除块。当擦除块指与所述虚拟擦除块关联的逻辑构建时，所述擦除块在此处可被称为逻辑擦除块（“LEB”）。

通常，按照处理的顺序顺序地存储所述包。在一种实施方式中，当使用写入数据管道106时，顺序存储模块802按照包从写入数据管道106出来的顺序存储包。这种顺序可能是由于下述原因：来自请求设备155的数据段与读取自另一存储部的有效数据包（正如在下述的恢复操作期间从存储部恢复数据一样）混合。将恢复的、有效的数据包重路由到写入数据管道可包括如上文中相对于图3的固态存储控制器104描述的垃圾收集器旁路316。

装置800包括选择恢复的存储部的存储部选择模块804。选择恢复的存储部可以使顺序存储模块802将所述存储部重新用于写入数据，因此将所述恢复的存储部添加到存储池中，或者所述存储部被重新用于在确定下述条件后从所述存储部中恢复有效数据：所述存储部失效、不可靠、应该被刷新、或其他将所述存储部暂时地或永久地移出所述存储池的理由。在另一种实施方式中，存储部选择模块804通过识别具有大量无效数据的存储部或擦除块来选择恢复的存储部。

在另一种实施方式中，存储部选择模块804通过识别具有低额损耗的存储部或擦除块来选择恢复的存储部。例如，识别具有低额损耗的存储部或擦除块可包括识别无效数据少、擦除重复的次数少、位出错率低或程序计数低（缓冲器中一页数据写入所述存储部中的页的次数少；程序计数可从下列情况开始被测量：制造设备时、所述存储部最近一次被擦除时、其他任意事件发生时及这些情况的组合）的存储部。存储部选择模块804还可使用上述参数中的任意组合或其他参数以确定具有低额损耗的存储部。通过确定具有低额损耗的存储部来选择恢复的存储部可有助于发现未充分利用的存储部，还可由于损耗均衡而被恢复，等等。

在另一种实施方式中，存储部选择模块804通过识别具有高额损耗的存

储部或擦除块来选择恢复的存储部。例如，识别具有高额损耗的存储部或擦除块来选择恢复的存储部包括识别擦除重复次数多、位出错率高、具有不可恢复的ECC块或程序计数高的存储部。存储部选择模块804还可使用上述参数的任意组合或其他参数以确定具有高额损耗的存储部。通过确定具有高额损耗的存储部来选择恢复的存储部可有助于发现被过度使用的存储部，还可通过利用擦除周期刷新所述存储部而被恢复等等，或者使所述存储部像不能使用那样不提供服务。

装置800包括数据恢复模块806，该数据恢复模块806从选定为恢复的存储部中读取有效数据包、将所述有效数据包与其他将要由顺序存储模块802顺序地写入的数据包排队并更新具有由顺序存储模块802写入的有效数据的新物理地址的索引。通常，所述索引为对象索引，该对象索引将对象的数据对象标识符映射到形成包的位置的物理地址，所述数据对象存储在固态存储器110中。

在一种实施方式中，装置800包括存储部恢复模块808，该存储部恢复模块808为使用或再使用而准备所述存储部并将所述存储部标记为对顺序存储模块802可用，以在数据恢复模块806完成从所述存储部中复制有效数据之后顺序地写入数据包。在另一种实施方式中，装置800包括存储部恢复模块808，该存储部恢复模块808将选定为恢复的存储部标记为无法存储数据。通常，这是由于存储部选择模块804识别具有高额损耗的存储部或擦除块来选择恢复的存储部，从而使得所述存储部或擦除块没有条件被用于可靠的数据存储。

在一种实施方式中，装置800位于固态存储设备102的固态存储设备控制器202内。在另一种实施方式中，装置800控制固态存储设备控制器202。在另一种实施方式中，装置800的一部分位于固态存储设备控制器202内。在另一种实施方式中，由数据恢复模块806更新的对象索引也位于固态存储设备控制器202内。

在一种实施方式中，所述存储部为擦除块，并且装置800包括擦除模块810，该擦除模块810在数据恢复模块806完成从所述选定的擦除块中复制有效数据包之后并在存储部恢复模块808将所述擦除块标记为可用之前，擦除选定为恢复的擦除块。对于闪存和其他擦除操作消耗的时间比读取或写入操

作消耗的时间长得多的固态存储器来说，在使数据块可以写入新数据之前擦除所述数据块有助于高效的操作。当固态存储器 110 布置在内存库 214 内时，擦除模块 810 的擦除操作可在一个内存库上执行，而另一个内存库可执行读取、写入或其他操作。

在一种实施方式中，装置 800 包括垃圾标记模块 812，该垃圾标记模块 812 将存储部中的数据包识别为无效，以响应指示所述数据包不再有效的操作。例如，如果数据包被删除，垃圾标记模块 812 可将所述数据包识别为无效。读-修改-写操作是用于将数据包识别为无效的另一种方法。在一种实施方式中，垃圾标记模块 812 可通过更新索引将所述数据包识别为无效。在另一种实施方式中，垃圾标记模块 812 可通过存储另一数据包将所述数据包识别为无效，所述另一数据包指示无效的数据包已经被删除。这种方法是有利的，这是由于在固态存储器 110 中存储所述数据包已被删除的信息允许对象索引重建模块 262 或类似模块重建具有项的对象索引，所述项指示所述无效的数据包已经被删除。

在一种实施方式中，装置 800 可被用于在清洗命令之后填充数据的虚拟页中的剩余部分，以提升整体的性能，其中，所述清洗命令使数据停止流入写入数据管道 106，直到写入数据管道 106 为空且所有的包已被永久地写入非易失性固态存储器 110。这具有以下好处：降低了需要的垃圾收集的量、减少了用于擦除存储部的时间并减少了编程虚拟页所需的时间。例如，可仅在准备将一个小包写入固态存储器 100 的虚拟页内时，接收清洗命令。编程这个几乎为空的页可能会引起下述结果：需要立即恢复浪费的空间；导致所述存储部内的有效数据被当作垃圾不必要的收集；及擦除、恢复所述存储空间并将所述存储空间返回到可用空间池以被顺序存储模块 802 写入。

将所述数据包标记为无效而不是实际上擦除无效的数据包是有效率的，这是因为，如上所述，对于闪存和其他类似存储器来说，擦除操作消耗相当长的时间。允许垃圾收集系统（如装置 800 中所述的）在固态存储器 110 内自主地运行提供了一种将擦除操作与读取、写入或其他更快的操作分开的方法，从而使得固态存储设备 102 能比其他许多固态存储系统或数据存储设备运行得快得多。

图 9 是示意性流程图，示出了根据本发明的用于存储恢复的方法 900 的

一种实施方式。方法 900 始于步骤 902，顺序存储模块 802 将数据包顺序地写入存储部（步骤 904）。所述存储部是固态存储设备 102 中的固态存储器 110 的一部分。通常，存储部为擦除块。所述数据包源于对象，而且所述数据包按处理的顺序被顺序地存储。

存储部选择模块 804 选择恢复的存储部（步骤 906），并且数据恢复模块 806 从选定为恢复的存储部中读取有效的数据包（步骤 908）。通常，有效的数据包为未被标记为擦除、删除或其他一些无效数据标识符的数据包，所述数据包被视为有效或“好”的数据。数据恢复模块 806 将有效数据包与其他预定由顺序存储模块 802 顺序地写入的数据包排队（步骤 910）。数据恢复模块 806 更新具有由顺序存储模块 802 所写入的数据的新物理地址的索引（步骤 912）。所述索引包括从数据包的物理地址到对象标识符的映射。这些数据包存储在固态存储器 110 中，并且所述对象标识符对应于所述数据包。

在数据恢复模块 806 完成从所述存储部复制有效数据后，存储部恢复模块将选定为恢复的存储部标记为对顺序存储模块 802 可用（步骤 914），以顺序地写入数据包，方法 900 结束于步骤 916。

空白数据段指示

一般来说，当数据不再有用时就会被擦除。在许多文件系统中，擦除命令删除文件系统中的目录项，而仍将数据保持在包含该数据的存储设备中。一般来说，数据存储设备并不涉及此类擦除操作。另一种擦除数据的方法是向数据存储设备写入0、1或一些其他空数据字符，以实际上替代所擦除的文件。然而，这样做效率不高，因为在传送将被覆盖的数据时会使用宝贵的带宽。此外，用来覆盖无效数据的数据会占据存储设备中的空间。

一些存储设备（如本文所描述的固态存储设备 102）不是随机存取存储设备，因此，更新先前所存储的数据并不会覆盖现有数据。尝试在此类设备中使用一串1字符或一串0字符来覆盖数据会占据宝贵的空间，而且也无法满足所期望的覆盖现存数据的意愿。对于这些非随机存储设备（诸如固态存储设备 102）而言，客户端 114 一般来说不具备覆盖数据以擦除数据的能力。

在接收到一串重复的字符或字符串时，所接收到的数据是可高度压缩的，但通常在将所述数据发往存储设备之前先由文件系统来执行压缩。一般的存储设备无法区分已经压缩的数据和未经压缩的数据。存储设备还可接收读取

所擦除的文件的命令，从而存储设备能够向请求设备传送一连串的0、1或空字符。同样的，需要带宽来传送表示了所擦除的文件的数据。

由上述讨论可知，很明显地存在对使得存储设备接收数据将被擦除的指令的装置、系统和方法，以使得存储设备能够存储表示了空数据段、具有重复的字符或字符串的数据的数据段令牌。该装置、系统和方法还可擦除现有数据，由此产生的使用后的存储空间包括有小的数据段令牌。提出了能够克服现有技术的一些缺陷或所有缺陷的装置、系统和方法。

图10是示出了根据本发明的具有用于生成令牌指令的装置的系统1000的一种实施方式的示意性框图。该装置包括：令牌指令生成模块1002、令牌指令传输模块1004、读取接收器模块1006、读取请求传输模块1008、读取令牌指令接收器模块1010、请求客户端响应模块1012和数据段重新生成模块1014，将在下文中描述这些模块。在一种实施方式中，所述装置位于服务器112中，该服务器112与具有存储控制器152、数据存储设备154（大体与上文所描述的相类似）的存储设备150相连。

在一种实施方式中，该装置包括令牌指令生成模块1002，该令牌指令生成模块1002用于生成具有令牌指令的存储请求。令牌指令包括在存储设备150上存储数据段的请求。令牌指令旨在代替待发往存储设备150并作为数据段存储的（如果数据段令牌不在其位置发送的话）一连串的重复的、相同的字符或一连串的重复的、相同的字符串。在一种实施方式中，所述一连串的重复的、相同的字符指示所述数据段为空。例如，一连串的重复的、相同的字符可以是0也可以是1，都为0或都为1的数据段可被看作是空的。

令牌指令至少包括数据段标识符和数据段长度。数据段标识符一般来说是设法在存储设备中存储重复的、相同的字符或字符串的对象ID、文件名称或其他为文件系统、应用、服务器112所悉知的标识符等等。数据段长度一般来说是一连串的重复的、相同的字符或字符串所需的存储空间。数据段令牌和令牌指令一般来说不包括数据段的数据，如一系列重复的、相同的字符。

然而，令牌指令可以包括用于形成数据段令牌的其他相关信息，如所述重复的、相同的字符或字符串中的至少一个实例。令牌指令还可包括元数据，如数据段位置、自文件系统的地址、对应于数据段的数据存储设备中的位置等等。本领域的技术人员将会认识到可包括在令牌指令中的其他信息。在一

种实施方式中，指令生成模块1002生成令牌指令以及数据段令牌。

在一种实施方式中，令牌指令生成模块1002生成令牌指令和安全擦除命令以响应于覆盖存储设备150中的现存数据的请求。现存数据包括存储设备中的使用与令牌指令中的数据段标识符相同的数据段标识符来标识的数据。一般来说，在下列情况下发送覆盖数据的请求：仅仅将数据标记为无效的或垃圾已经不够了；删除指向数据的指针；或者其他典型的删除操作，但是，其中，需要以不可恢复所述数据的方式来覆盖所述数据。比方说，当认为数据是敏感信息，为了安全因素的考虑而必须将其删除时，就需要覆盖数据的命令。

安全擦除命令指令存储设备150覆盖现有数据，由此现有数据是不可恢复的。存储设备150随后创建数据段令牌，并对现有数据执行覆盖、恢复、擦除等操作。由此，现有数据是不可恢复的，数据段令牌存储在存储设备150中，其中，数据段令牌所占据的存储空间比现有数据小得多。

在另一种实施方式中，所述装置包括擦除确认模块1016，该擦除确认模块1016用于接收确认，即，存储设备中的现有数据已经由字符覆盖了，从而现有数据是不可恢复的。该确认可被转发至请求设备或客户端114，并可用于验证现有数据已经处于不可恢复的状况中。在其他实施方式中，安全擦除命令可指令存储设备150使用特定的字符、字符串来覆盖现有数据，或者可执行多次执行命令。本领域的技术人员将会认识到用于配置一个或多个安全擦除命令以确保现有数据不可恢复的其他方式。

可以对数据进行加密并随后将其存储在存储设备150中，其中，使用存储设备150在存储所述数据时接收到的加密密钥来完成加密过程。在另一种实施方式中，在存储现有数据之前先使用该接收到的加密密钥来加密现有数据的情况下，令牌指令生成模块1002生成令牌指令以及加密擦除命令，以响应接收了覆盖现有数据的请求。所述加密擦除命令擦除用于存储现有数据的加密密钥，由此，加密密钥不可恢复。

在一种实施方式中，擦除加密密钥包括擦除来自于请求设备的加密密钥。在另一种实施方式中，擦除加密密钥包括擦除来自服务器、密钥金库（key vault）或存储加密密钥的其他位置的加密密钥。擦除加密密钥可包括使用其他数据或使用一连串的字符来替代加密密钥，以使得用任何方式都无法恢复

该加密密钥。一般来说，在使用足够稳健以至于能够阻挠对解密现有数据的尝试的加密程序来加密现有数据的情况下，擦除加密密钥会使得存储设备150中的现有数据不可恢复。在如下情况下，覆盖现有数据的请求可以是安全擦除指令（由于安全因素的原因而覆盖数据）；覆盖数据以擦除数据的请求；设法将现有数据替换为重复的、相同的字符或字符串的请求等等。在一种实施方式中，安全擦除指令使得设备能够安全地擦除加密密钥以及能够安全地擦除现有数据。在一种实施方式中，擦除加密密钥可允许安全地擦除存储设备中的数据得以延迟，直到垃圾收集进程（存储空间恢复进程的一部分）擦除了数据为止。本领域的技术人员将会认识到擦除加密密钥的其他方法和接收覆盖现有数据的请求的其他方法。

在一种实施方式中，令牌指令包括数据段令牌，令牌指令传输模块1004发送令牌指令以及数据段令牌。在另一种实施方式中，令牌指令不包括数据段令牌，而包括使得存储设备150生成数据段令牌的命令。在该实施方式中，令牌指令传输模块1004发送命令以及令牌指令以生成数据段令牌，但并不发送数据段令牌。

所述装置包括令牌指令传输模块1004，后者用于向存储设备150发送令牌指令。一般来说，令牌指令传输模块1004发送作为存储请求的一部分的令牌指令。存储请求可以是对象请求的形式、数据请求的形式或本领域的技术人员所知的其他形式。在令牌指令生成模块1002生成了安全擦除指令的情况下，令牌指令传输模块1004将所述安全擦除指令发送给存储设备150。在令牌指令生成模块1002生成了擦除加密密钥命令的情况下，当需要时，擦除加密密钥命令被发往另一个设备来执行该命令。

在一种实施方式中，令牌指令传输模块1004发送不包括数据段令牌的令牌指令。在这种实施方式中，令牌指令包括可由存储设备150用来产生数据段令牌的指令和信息。在另一种实施方式中，令牌指令传输模块1004发送包括数据段令牌的令牌指令。在这种实施方式中，存储设备150能够识别出与令牌指令接收的数据段令牌表示了数据段，存储设备150采取适当的操作来存储数据段令牌，以使得数据段令牌表示了数据段，而不仅仅是将数据段令牌作为普通数据来存储。

在特定的实施方式中，所述装置包括：读取接收器模块1006，用于接收

来自存储设备150的读取数据段的存储请求；读取请求传输模块1008，用于向存储设备150发送存储请求。一般来说，存储请求是从请求客户端114（如外部客户端114）、服务器112内部的客户端114（如在服务器112上运行的应用或文件服务器等等）接收的。本领域的技术人员将会认识到可作为读取接收器模块1006能从其接收存储请求的请求客户端114的其他设备以及软件。

存储请求包括：读取对应于数据段令牌的数据段的请求，其中，数据段令牌被请求存储在由令牌指令传输模块1004发往存储设备150的令牌指令中。在一种实施方式中，请求客户端114不知道已经以数据段令牌的形式存储了数据段。在另一种实施方式中，请求设备知道已经以数据段令牌的形式存储了数据段，但并不清楚存储于数据段令牌中的信息。

在一种特定的实施方式中，装置还可包括读取令牌指令接收器模块1010，该令牌指令接收器模块1010用于从存储设备接收对应于所请求的数据段令牌的消息，其中，所述消息至少包括数据段标识符和数据段长度。一般来说，所述消息并不包括数据段中的数据。所述消息还包括存储在数据段令牌中的其他信息，如数据段位置或重复的、相同的字符或字符串。在这种特定的实施方式中，装置包括请求客户端响应模块1012，该客户端响应模块1012用于向请求客户端113发送根据从存储设备150接收的消息而形成的响应。

在一种实施方式中，读取令牌指令接收器模块1010还接收消息中的有关现有数据已经被字符覆盖从而现有数据不可恢复的确认，其中，所述现有数据预先存储在存储设备中并且使用来自在消息中接收的数据段令牌的相同的数据段标识符来标记。确认还可以独立于任何读取数据段的存储请求从存储设备150接收到。

在另一种实施方式中，其中，请求客户端114需要数据段，所述装置包括数据段重新生成模块1014，该数据段重新生成模块1014用于使用包含在消息中的信息来重构数据段中的数据。在这种情况下，发往请求客户端的响应包括经重构的数据段。在另一种实施方式中，发往请求客户端的响应包括包含在从存储设备150接收到的消息中的信息。请求客户端114随后重构数据段或以一些其他的方式来使用该信息。在另一种实施方式中，所述消息包括数据段令牌。数据段重新生成模块1014使用该数据段令牌以在将数据段令牌向请求客户端114转发之前重构数据段，或者，请求客户端响应模块1012可以简单

地转发该数据段令牌。

在一种实施方式中，具有令牌指令的存储请求还包括在存储设备150预留存储空间的请求，其中，所请求的预留存储空间的存储空间大小与数据段长度大致相同。在另一种实施方式，所请求的预留存储空间的存储空间大小不同于数据段长度。例如，如果存储设备150是固态存储设备102，固态存储设备102可以连接到硬驱动器或其他的长期、廉价存储器，而固态存储器110则被配置为长期存储器的缓存。预留存储空间的请求使得固态存储设备102将一部分缓存清洗到长期存储器以准备向固态存储设备102写入数据。本领域的技术人员将会认识到期望请求预留存储空间的其他情况。

在一种实施方式中，装置可具有读取接收器模块1006、读取请求传输模块1008、读取令牌指令接收器模块1010、请求客户端响应模块1012，这些模块大体类似于上文所描述的那些模块。在这种实施方式中，模块1006-1012独立于包括有令牌生成模块1002或令牌指令传输模块1004的装置。在一种实施方式中，所述装置包括大体类似于上文所描述的数据段重新生成模块1014的数据段重新生成模块1014。

图11是示意性框图，示出了根据本发明的用于生成和发送令牌指令的方法1100的实施方式。方法1100始于步骤1102，令牌指令生成模块1002生成包括令牌指令的存储请求（步骤1104），其中，令牌指令包括在存储设备150中存储数据段令牌的请求。令牌指令传输模块1004向存储设备150发送令牌指令（步骤1106），方法1100在1108结束。在一种实施方式中，存储请求包括令牌指令以存储数据段令牌，其中，存储请求大体上与数据段中的数据无关。在另一种实施方式中，存储请求包括来自于数据段的数据。在优选实施方式中，软件应用程序使用令牌指令创建存储请求，从而避免了创建数据段。在另一种实施方式中，软件应用程序请求生成令牌指令。

图12是示意性流程图，示出了根据本发明的用于读取数据段令牌的方法1200的实施方式。方法1200始于步骤1202，读取接收器模块1006从请求客户端114接收从存储设备150读取数据段的存储请求（步骤1204）。读取请求传输模块1008将存储请求发往存储设备150（步骤1206）。

读取令牌指令接收器模块1008从存储设备150接收对应于所请求的数据段令牌的消息（步骤1208），其中，所述消息至少包括数据段标示符和数据

段长度。所述消息大体上与数据段中的数据无关。请求客户端响应模块1012向请求客户端发送响应（步骤1210），其中，该响应是根据从存储设备150接收到的消息而形成的，方法1200结束于步骤1212。

图13是示意性框图，示出了根据本发明的包括用于管理数据段令牌的装置的系统1300的实施方式。系统1300包括具有写入请求接收器模块1302和数据段令牌存储模块1304的装置，在多个实施方式中，所述系统还包括具有令牌指令生成模块1306、读取请求接收器模块1308、读取数据段令牌模块1310、读取请求响应模块1312（具有发送数据段令牌模块1314和发送数据段模块1316）、重构数据段模块1318、安全擦除模块1320（具有擦除确认模块1322）和存储空间预留模块1324，下文将描述这些模块。系统1300包括具有存储控制器152和数据存储设备154的存储设备150（与上文所描述的设备大体类似）。系统1300包括与存储设备150进行通信的请求设备1326（下文将描述）。

在所描述的实施方式中，模块1302-1324被包括在存储设备150或存储控制器152中。在另一种实施方式中，模块1302-1324中的一个或多个模块的至少一部分位于存储设备150之外。在又一种实施方式中，请求设备1326以驱动器、软件或模块1302-1324中的一个或多个模块的其他功能形式包括模块1302-1324的一部分。例如，在请求设备1326中示出了令牌生成模块1306和重构数据段模块1318。本领域的技术人员将会认识到用以分布和实现模块1302-1324的功能的其他方式。

所述装置包括写入请求接收器模块1302，该写入请求接收器模块1302用于接收来自请求设备1326的存储请求，其中，所述存储请求包括将数据段存储到存储设备150的请求。数据段包括一连串的重复的、相同的字符或字符串。一般来说，所述一连串的重复的、相同的字符表明数据段为空。当一连串的重复的、相同的字符为1或0的时候，尤为如此。所述装置包括用于在存储设备150中存储数据段令牌的数据段令牌存储模块1304。数据段令牌至少包括数据段标识符和数据段长度。数据段令牌大体上与数据段中的实际数据无关。

可以以多种方式存储数据段令牌。在一种实施方式中，数据段令牌包括索引中的项，其中，所述索引对应于存储在存储设备150中的信息和数据。比方说，索引可以是上文结合图2所描绘的装置200而描述的对象索引。索引还可以是文件系统索引、块存储索引或本领域技术人员所知的其他索引。在另

一种实施方式中，数据段令牌包括存储在存储设备150中的元数据，或是采用了存储在存储设备150中的元数据的形式。在另一种实施方式中，数据段令牌作为元数据存储在存储设备中，并且数据段令牌包括索引中的项。本领域技术人员将会认识到存储数据段令牌的其他方式。

在一种实施方式中，存储请求包括用以存储数据段令牌的令牌指令，其中，存储请求本质上与数据段中的数据无关。令牌指令包括数据段令牌或用以生成数据段令牌的命令。其中，令牌指令不包括数据段令牌，数据段令牌存储模块1304根据令牌指令中的信息生成数据段令牌。如果令牌指令包括数据段令牌，那么数据段令牌存储模块1304执行如下操作：将数据段令牌辨识为表示了令牌指令中的数据段标识符所标识的数据段的数据结构；适当地存储数据段令牌。

一般来说，在数据段令牌存储模块1304辨识出了数据段令牌的情况下，该数据段令牌在某些方面不同于存储在存储设备150中的其他数据。例如，请求设备1326可以仅仅压缩数据并发送经压缩的对象、文件或数据段，从而存储设备150不将经压缩的数据段与通过其他存储请求而接收到的其他数据相区分。

在数据段令牌存储模块1304辨识出了接收到的数据段令牌是数据段令牌的情况下，数据段令牌存储模块1304以如下方式存储数据段令牌：使得当读取时，该数据段令牌表现为数据段而非数据段令牌。本领域的技术人员将会认识到数据段令牌存储模块1304在辨识出所接收到的数据段令牌是数据段令牌而非数据段之后可存储数据段令牌的其他方式。

在另一种实施方式中，存储请求包括来自数据段的数据。在该实施方式中，所述装置包括用于根据数据段生成数据段令牌的令牌生成模块1306，其中，为响应存储数据段的存储请求而创建所述数据段令牌。在又一种实施方式中，令牌生成模块1306（可能以驱动器的形式）位于请求设备1326中。

在一种实施方式中，装置包括安全擦除模块1320，该安全擦除模块1320用于使用字符覆盖现有数据，以使得现有数据不可恢复，其中，所述现有数据包括先前存储在存储设备中的数据段中的数据，所述数据段是使用与标识存储请求中的数据段时所使用的数据段标识符相同的数据段标识符来标识的。在该实施方式中，数据段令牌与数据段标识符一起存储，并且通过覆盖

现有数据擦除了数据段长度和由存储在数据段令牌中的相同的数据段标识符来标识的现有数据。一般来说，现有的字符由0、1或一些其他的字符串来覆盖，从而使得现有数据被破坏且不可恢复。

在又一种实施方式中，安全擦除模块还包括擦除确认模块1322，用于发送指示了现有数据已被覆盖的消息。一般来说，消息是发往请求设备1326的。擦除确认消息在安全擦除模块1320覆盖了现有数据之后发送。所述消息可以与存储请求在相同的交易过程中发送，也可以在与存储请求不同的交易过程中发送。

在另一种实施方式中，安全擦除模块1320在存储空间恢复操作期间覆盖现有数据。例如，如上文所述，如果存储设备150是固态存储设备102，那么存储空间恢复操作与结合图8中描绘的装置800而描述的垃圾收集相关。然而，通常会加快涉及覆盖现有数据的请求的存储空间恢复操作，以便在擦除确认模块1322发送任何确认消息之前先必要地恢复存储现有数据的存储位置。在一种实施方式中，标记或是标识现有数据以指示已经请求了安全擦除。一般来说，直到标记为要擦除的现有数据已经被覆盖并已不可恢复时才发送确认消息。在另一种实施方式中，安全擦除模块1320仅仅将现有数据标记为无效，以便随后的存储空间恢复。在另一种实施方式中，安全擦除操作更新索引，以指示现有数据无效且在随后的存储空间恢复期间防止在数据被覆盖之前访问该数据。

在一种实施方式中，安全擦除模块1320在每次存储数据段令牌时都覆盖现有数据。在另一种实施方式中，存储请求具体包括覆盖现有数据的请求，安全擦除模块1320覆盖现有数据以响应于覆盖现有数据的请求。在另一种实施方式中，安全擦除模块1320存储与确认现有数据已经被擦除相关的元数据信息，从而随后的读取能够指示该擦除。

在其他实施方式中，当未接收到安全擦除时，则删除现有数据。在一种实施方式中，删除数据包括删除索引项和地址等等。在优选实施方式中，在存储了数据段令牌时，相应的现有数据被标记为无效或已可进行存储恢复。所述数据可随后在存储恢复操作、垃圾收集操作等操作中恢复。

在特定的实施方式中，所述装置包括：读取请求接收器模块1308，用于接收读取数据段的存储请求；读取数据段令牌模块1310，用于读取对应于存

储请求所请求的数据段的数据段令牌；读取请求响应模块1312，用于向请求设备1326发送响应。所述响应是使用对应于所请求的数据段的数据段令牌生成的。

在一种实施方式中，读取数据段的请求与存储请求相关联并用于确认存储请求已经成功。在另一种实施方式中，读取数据段的请求独立于存储请求，所述请求可由生成所述存储请求的请求设备1326发起，也可由另外的不同的请求设备1326发起。

在一种实施方式中，在请求设备能够接收来自数据段令牌的信息而不是实际的数据的情况下，读取请求响应模块1312包括发送数据段令牌模块1314，该数据段令牌模块1314用于向请求设备1326发送响应中的消息。所述消息至少包括数据段标识符和数据段长度，但还可包括：数据段位置；重复的、相同的字符或字符串的至少一个例子；或其他相关信息。一般来说，所述消息并不包括数据段中的实际数据，而是包括数据段令牌所包括的其他信息。

在另一种实施方式中，在请求设备1326期望接收数据段的情况下，所述装置包括重构数据段模块1318，该重构数据段模块1318用于使用数据段令牌重构数据段中的数据。读取请求响应模块1312还包括用于向请求设备1326发送经重构的请求的数据段的发送数据段模块1316。在另一种实施方式中，重构数据段模块1318（可能以驱动器的形式）位于请求设备1326中，发送数据段令牌模块1314向请求设备1326发送包括数据段令牌信息的消息。请求设备1326的重构数据段模块1318根据消息重构所请求的数据段。

在一种实施方式中，系统1300包括一个独立的装置，该装置包括：读取请求接收器模块1308、读取数据段令牌模块1310、读取请求响应模块1312，这些模块大体上类似于上文所描述的那些模块。所述装置独立于包括写入请求接收器模块1302和数据段令牌存储模块1304的装置。在一种实施方式中，读取请求响应模块1312包括发送数据段令牌模块1314和/或发送数据段模块1316，所述装置包括重构数据段模块1318，其中，模块1314、1316和1318大体上类似于上文所描述的那些模块。

图14是示意性流程图，示出了根据本发明的用于存储数据段令牌的方法1400的实施方式。方法1400始于步骤1402，写入请求接收器模块1302从请求设备1326接收存储请求（步骤1404），其中，所述存储请求包括将数据段存

储到存储设备150的请求。数据段包括一连串重复的、相同的字符或字符串。数据段令牌存储模块1304在存储设备150处存储数据段令牌（步骤1406），方法1400结束于步骤1408。数据段令牌至少包括数据段标识符和数据段长度，在大部分情况下，数据段令牌不包括数据段中的数据。

图15是示意性流程图，示出了根据本发明的用于读取数据段令牌的方法1500的实施方式。方法1500始于步骤1502，读取请求接收器模块1308接收从存储设备150中读取数据段的存储请求（步骤1504）。所述数据段以数据段令牌的形式存在于存储设备中，所述数据段包括一连串重复的、相同的字符或字符串。所述数据段令牌至少包括数据段标识符和数据段长度，数据段令牌不包括数据段中的数据。读取数据段令牌模块1310读取对应于存储请求所请求的数据段的数据段令牌（步骤1506），读取请求响应模块1312向请求设备150发送响应（步骤1508），方法1500在1510结束。所述响应是使用对应于所请求的数据段的数据段令牌生成的。

渐进式RAID

独立驱动器冗余阵列（“RAID”）能够以许多方式构造，以实现各种目标。如本文所述，驱动器是数据的海量存储设备。驱动器或存储设备可以是固态存储器110、硬盘驱动器（“HDD”）、磁带驱动器、光学驱动器、或者本领域技术人员已知的任何其他海量存储设备。在一种实施方式中，驱动器包括作为虚拟卷访问的海量存储设备的一部分。在另一种实施方式中，驱动器包括可作为虚拟卷一起访问的并且在存储区域网络（“SAN”）中配置的两个或更多数据存储设备，所述数据存储设备还可作为RAID、简单磁盘捆绑（“JBOD”）等等。通常，驱动器作为单个单元或虚拟卷通过存储控制器152访问。在优选实施方式中，存储控制器152包括固态存储控制器104。本领域技术人员将会认识到采用可在RAID中配置的海量存储设备的形式的驱动器的其他形式。在本文描述的实施方式中，可交替地使用驱动器和存储设备150。

传统上，不同的RAID配置称为RAID级别。一个基本的RAID配置是RAID级别0，其创建存储设备150的镜像拷贝。RAID 0的优点是一个或多个存储设备150上的数据的完整拷贝在该一个或多个存储设备150的镜像拷贝上也是可用的，以使得读取主驱动器或镜像驱动器上的数据是相对快速的。RAID 0在主存储设备150故障的情况下还提供数据的备份拷贝。RAID 0的缺点是写入相

对较慢，因为所写入的数据必须被写入两次。

另一种传统RAID配置是RAID级别1。在RAID 1中，写入RAID的数据被分割成对应于存储设备150集中的N个存储设备150的N个数据段。N个数据段形成“条带”。通过在多个存储设备150之间条带化数据，增强了性能，这是因为存储设备150可并行地工作来存储N个数据段，相比于单个存储设备150保存包括N个数据段的数据更快速。然而，读取数据相对较慢，因为数据分布在多个存储设备150上，并且多个存储设备150的访问时间通常相比于从包含所有期望数据的一个存储设备150读取数据更慢。此外，RAID 1不提供故障保护。

流行的RAID配置是RAID级别5，其包括N个存储设备150条带化N个数据段，并且将奇偶校验数据段存储在第N+1个存储设备150上。RAID 5提供故障容错，因为RAID可容忍存储设备150的单个故障。例如，如果存储设备150出现故障，可使用其他可用数据段和为条带特定计算的奇偶校验数据段来创建丢失的条带的数据段。RAID 5相比于RAID 0还通常使用更少的存储空间，这是因为被RAID的存储设备150的集中的每个存储设备150不需要存储数据的完整拷贝，而仅需存储条带的数据段或奇偶校验数据段。像RAID 1一样，RAID 5对于写入数据相对较快，但是对于读取数据相对较慢。然而，对于典型的传统RAID 5写入数据相比于对于RAID 1写入数据更慢，因为必须根据条带的N个数据段为每个条带计算奇偶校验数据段。

另一种流行的RAID配置是RAID级别6，其包括双重分布奇偶校验。在RAID 6中，分配两个存储设备150作为校验镜像设备（例如1602a、1602b）。单独地计算用于条带的每个奇偶校验数据段，以使得在丢失了存储设备集中的任何两个存储设备150时，使用剩余的可用数据段和/或奇偶校验数据段可将其恢复。RAID 6具有与RAID 5类似的性能优点和缺点。

嵌套的RAID还可用于增加需要高可靠性情况下的容错性。例如，两个存储设备集（每个都被配置为RAID 5）可在RAID 0配置中被镜像。作为结果的配置可称为RAID 50。如果RAID 6用于每个镜像的集，配置可称为RAID 60。嵌套的RAID配置通常具有与底层RAID群组类似的性能结果。

如上所述，显而易见，存在对于用于渐进式RAID的装置、系统、和方法的需要，所述渐进式RAID提供相比于传统容错RAID级别（例如RAID 0、RAID

5、RAID 6等来说)容错、更快速的数据写入的好处,同时也提供相比于传统分割的RAID级别(例如RAID 1、RAID 5、RAID 6等)来说,更快速的数据读取速度。有利地是,这种装置、系统、和方法将N个数据段写入奇偶校验镜像存储设备1602,提供RAID 0系统的优点,直到需要计算奇偶校验数据段(例如在存储合并操作之前或部分存储合并操作时)。

图 10 是示意性框图,示出了根据本发明的用于渐进式 RAID 和前端分布式 RAID 的系统 1600 的一种实施方式。系统 1600 包括 N 个存储设备 150 和 M 个奇偶校验-镜像存储设备 1602,一个或多个客户端可通过计算机网络 116 访问存储设备 150 和奇偶校验-镜像存储设备 1602。N 个存储设备 150 和奇偶校验-镜像存储设备 1602 可位于一个或多个服务器 112 内。存储设备 150、服务器 112、计算机网络 116 和客户端 114 大体上与上文描述的类似。奇偶校验-镜像存储设备 1602 通常与 N 个存储设备 150 类似或相同,并且通常被指定为用于条带的奇偶校验-镜像存储设备 1602。

在一种实施方式中,N 个存储设备 150 和 M 个奇偶校验-镜像存储设备 1602 被包括在一个服务器 112 内或者可通过一个服务器 112 被访问,并可以 通过系统总线、SAN 或其他计算机网络 116 联网在一起。在另一种实施方式中,N 个存储设备 150 和 M 个奇偶校验-镜像存储设备 1602 位于多台服务器 112a-n+m 内或者可通过多台服务器 112a-n+m 被访问。例如,存储设备 150 和奇偶校验-镜像存储设备 1602 可以是上文中相对于图 1C 的系统 103 和图 5B 的方法 105 描述的服务器内 SAN 的一部分。

在一种实施方式中,奇偶校验-镜像存储设备 1602 存储存储在渐进式 RAID 中的条带的所有奇偶校验数据段。在另一种优选实施方式中,分配给渐进式 RAID 的存储设备集 1604 中的存储设备 150 被分配为用于特定条带的奇偶校验-镜像存储设备 1602,这种分配是轮换的,从而所述奇偶校验数据段在 N+M 个存储设备 150 之间为每个条带轮换。这种实施方式通过将单个存储设备 150 分配为用于每个条带的奇偶校验-镜像存储设备 1602 提供了性能上的优势。通过轮换奇偶校验-镜像存储设备 1602,与计算和存储奇偶校验数据段有关的开销可以是分散的。

在一种实施方式中,存储设备 150 为固态存储设备 102,每个存储设备 150 都具有关联的固态存储器 110 和固态存储控制器 104。在另一种实施方式

中，每个存储设备 150 包括固态存储控制器 104，并且关联的固态存储器 110 作为用于其他花费少、性能低的存储器（如磁带存储器或硬盘驱动器）的缓存。在另一种实施方式中，服务器 112 中的一个或多个包括将存储请求发送给渐进式 RAID 的一个或多个客户端 114。本领域技术人员会认识到可为渐进式 RAID 配置的具有 N 个存储设备 150 和一个或多个奇偶校验-镜像存储设备 1602 的其他系统配置。

图 17 是示意性框图，示出了根据本发明的用于渐进式 RAID 的装置 1700 的一种实施方式。在不同的实施方式中，装置 1700 包括存储请求接收器模块 1702、条带化模块 1704、奇偶校验-镜像模块 1706、奇偶校验渐进模块 1708、奇偶校验更替模块 1710、镜像集模块 1712、更新模块 1714、具有直接客户端响应模块 1718 的镜像恢复模块 1716、预整合模块 1720、后整合模块 1722、数据重建模块 1724 和奇偶校验重建模块 1726，这些模块将在下文中描述。模块 1702-1726 被描述位于服务器 112 内，但是模块 1702-1726 的一些功能或全部功能还可分布在多个服务器 112、存储控制器 152、存储设备 150、客户端 114 等设备之内。

装置 1700 包括接收存储数据的请求的存储请求接收器模块 1702，其中，所述数据是文件的数据或对象的数据。在一种实施方式中，所述存储请求是对象请求。在另一种实施方式中，所述存储请求是块存储请求。在一种实施方式中，所述存储请求不包括数据，但包括命令，存储设备 150 和奇偶校验-镜像存储设备 1602 可使用该命令以从客户端或从其他源 DAM 或 RDMA 数据。在另一种实施方式中，所述存储请求包括由于所述存储请求而将要被存储的数据。在另一种实施方式中，所述存储请求包括一条能够将数据存储在所述存储设备集 1604 中的命令。在另一种实施方式中，所述存储请求包括多条命令。本领域技术人员会认识到适于渐进式 RAID 存储数据的其他存储请求。

所述数据存储在装置 1700 可访问的位置。在一种实施方式中，所述数据在随机存取存储器（“RAM”）中可用，所述随机存取存储器如客户端 114 或服务器使用的 RAM。在另一种实施方式中，所述数据存储在硬盘驱动器、磁带存储器或其他大容量存储器中。在一种实施方式中，所述数据被配置为对象或文件。在另一种实施方式中，所述数据被配置为作为对象或文件的一部

分的数据块。本领域技术人员会认识到作为所述存储请求的目标的所述数据的其他形式和位置。

装置 1700 包括为数据计算条带模式的条带化模块 1704。所述条带模式包括一个或多个条带，其中，每个条带包括 N 个数据段的集。通过，条带中数据段的数量取决于分配给所述 RAID 群组的存储设备 150 的数量。例如，如果采用 RAID5，一个存储设备被指定为奇偶校验-镜像存储设备 1602a 以为特定的条带存储奇偶校验数据。如果四个存储设备 150a、150b、150c、150d 被分配给所述 RAID 群组，条带在除所述奇偶校验数据段外还会具有四个数据段。条带化模块 1704 将条带的 N 个数据段写入 N 个存储设备 150a-n，从而使得所述 N 个数据段中的每一个被写入分配给所述条带的存储设备 150 的集 1604 中的不同的存储设备 150a、150b、……150n。本领域技术人员会领体会到可被分配给用于特定 RAID 级别的 RAID 群组的存储设备 150 的不同组合，并会领体会到创建条带模式和将数据分割成每条带 N 个数据段的方法。

装置 1700 包括奇偶校验-镜像模块 1706，该奇偶校验-镜像模块 1706 将所述条带的 N 个数据段的集写入存储设备集 1604 中的一个或多个奇偶校验-镜像存储设备 1602，其中，奇偶校验-镜像存储设备 1602 是除 N 个存储设备 150 之外的设备。然后，所述 N 个数据段可用于后续计算奇偶校验数据段。奇偶校验-镜像模块 1706 将 N 个数据段的集复制到奇偶校验-镜像存储设备 1602 (这通常比存储所述 N 个数据段需要更少的时间)，而不是立即计算所述奇偶校验数据段。一旦所述 N 个数据段存储在奇偶校验-镜像存储设备 1602，如果 N 个存储设备 150 中的一个不可用，所述 N 个数据段就可被读取或可被用于恢复数据。读取数据还具有 RAID 0 配置的优点，这是由于全部所述 N 个数据段一起从一个存储设备 (如 1602a) 中获取。对于不止一个的奇偶校验-镜像存储设备 (如 1602a、1602b)，奇偶校验-镜像模块 1706 将所述 N 个数据段复制到每个奇偶校验-镜像存储设备 1602a、1602b。

装置 1700 包括奇偶校验渐进模块 1708，该奇偶校验渐进模块 1708 为所述条带计算一个或多个奇偶校验数据段，以响应存储整合操作。由所述 N 个数据段计算出来的所述一个或多个奇偶校验数据段存储在奇偶校验-镜像存储设备 1602 上。奇偶校验渐进模块 1708 在一个或多个奇偶校验-镜像存储设备 1602 中的每一个上存储一个奇偶校验数据段。所述存储整合操作旨在在一个

或多个奇偶校验-镜像存储设备 1602 中的至少一个上至少恢复存储空间和/或数据。例如，存储整合操作可以是上文中相对于图 8 和图 9 的装置 800 和方法 900 描述的在固态存储设备 102 上的数据垃圾收集。所述存储整合操作还可包括用于硬盘驱动器的碎片整理操作或其他整理数据以增加存储空间的类似操作。正如此处所使用的，所述存储整合操作还可包括恢复数据的操作（例如，如果存储设备 150 不可用，从错误中恢复数据，或由于其他读取数据的原因而从奇偶校验-镜像存储设备 1602 中恢复数据）。在另一种实施方式中，当奇偶校验-镜像存储设备 1602 不那么繁忙时，奇偶校验生成模块 1708 容易地计算所述奇偶校验数据段。

有利地是，通过延迟计算和存储条带的所述奇偶校验数据段，奇偶校验-镜像存储设备 1602 上的所述 N 个数据段可用于读取所述数据段、恢复数据、重建存储设备 150 上的数据，直到奇偶校验-镜像存储设备 1602 上需要更多的存储空间或其他需要存储整合操作的原因。然后，奇偶校验渐进模块 1708 可独立于存储请求接收器模块 1702、条带化模块 1704 或奇偶校验-镜像模块 1706 像后台操作一样运行。本领域技术人员会轻易地认识到延迟计算奇偶校验数据段的其他理由，其中，延迟计算所述奇偶校验数据段作为渐进式 RAID 操作的一部分。

在一种实施方式中，模块 1702-1708 的功能中（接收存储数据的请求、计算条带模式并将 N 个数据段写入 N 个存储设备、将 N 个数据段的集写入奇偶校验-镜像存储设备、计算奇偶校验数据段）的一些或全部在下述设备中实现：存储设备集 1604 的存储设备 150、客户端 114 和第三方 RAID 管理设备。所述第三方 RAID 管理设备可以是服务器 112 或其他计算机。

在一种实施方式中，装置 1700 包括奇偶校验更替模块 1710，该奇偶校验更替模块 1710 为每个条带变更被分配为一个或多个用于所述条带的奇偶校验-镜像存储设备 1602 的存储设备集 1604 中的存储设备 150。如上文中相对于图 10 的系统 1600 所描述的，通过轮换用于奇偶校验-镜像存储设备（用于条带）的存储设备 150，不同奇偶校验数据段的计算工作分布在存储设备集 1604 的存储设备 150 之间。

在另一种实施方式中，存储设备集 1604 是第一存储设备集，并且装置 1700 包括镜像集模块 1712，该镜像集模块 1712 创建除第一存储集 1604 之外的一

个或多个附加存储设备集，从而使得所述一个或多个附加存储设备集中的每一个至少包括关联的条带化模块 1704，该条带化模块 1704 将所述 N 个数据段写入所述一个或多个附加存储集中的每一个集中的 N 个存储设备。在一种相关的实施方式中，所述一个或多个附加存储设备集中的每一个集包括关联的用于存储所述 N 个数据段的集的奇偶校验-镜像模块 1706 和用于计算一个或多个奇偶校验数据段的奇偶校验渐进模块 1708。在镜像集模块 1712 创建一个或多个镜像存储设备集的情况下，RAID 可以是嵌套的 RAID(如 RAID 50)。在这种实施方式中，RAID 级可从 RAID 10 (其中，数据被条带化和镜像化) 渐进到 RAID50 或 RAID60 (其中，为每个存储设备集 1604 计算和存储奇偶校验数据段)。

在一种实施方式中，装置 1700 包括更新模块 1714。更新模块 1714 通常在奇偶校验-镜像存储设备 1602 上的 N 个数据段还未渐进为奇偶校验数据段的情况下使用。更新模块 1714 接收已更新的数据段，其中，所述已更新的数据段对应于 N 个存储设备 150 上存储的 N 个数据段中的现有数据段。更新模块 1714 将所述已更新的数据段复制到存储所述现有数据段的所述条带的存储设备 150，并将所述已更新的数据段复制到所述条带的一个或多个奇偶校验-镜像存储设备 1602。更新模块 1714 用所述已更新的数据段替换存储在 N 个存储设备 150a-n 中的存储设备 150 上的所述现有数据段，并用所述已更新的数据段替换存储在一个或多个奇偶校验-镜像存储设备 1602 上的相应的所述现有数据段。

在一种实施方式中，替换数据段包括将所述数据段写入存储设备 150 并然后将相应的数据段标记为对后续垃圾收集无效。这种实施方式的一个实例被描述用于固态存储器 110 和上文中相对于图 8 和图 9 描述的垃圾收集装置。在另一种实施方式中，替换数据段包括用已更新的数据段覆盖现有数据段。

在一种实施方式中，存储设备集 1604 是第一存储设备集，装置 1700 包括镜像恢复模块 1716，当第一存储集 1604 中的存储设备 150 不可用时，该镜像恢复模块 1716 恢复存储在第一存储集 1604 中的存储设备 150 上的数据段。所述数据段是从包含所述数据段的拷贝的镜像存储设备中恢复的。所述镜像存储设备包括一个或多个存储设备 150 的集的存储所述 N 个数据段的拷贝的一个存储设备。

在另一种实施方式中，镜像恢复模块 1716 为响应来自客户端 114 的读取所述数据段的读取请求而恢复所述数据段。在另一种相关的实施方式中，镜像恢复模块 1716 还包括直接客户端响应模块 1718，该直接客户端响应模块 1718 将请求的数据段从所述镜像存储设备发送给客户端 114。在这种实施方式中，所述请求的数据段被复制到客户端 114，从而客户端 114 不需要等到所述数据段被恢复就将所述数据段传送到客户端 114。

在一种实施方式中，装置 1700 包括预整合恢复模块 1720，该预整合恢复模块 1720 为响应读取数据段的请求而恢复存储在存储集 1604 的存储设备 150 上的所述数据段。在这种实施方式中，存储设备 150 不可用，并且所述数据段是先于奇偶校验渐进模块 1708 在一个或多个奇偶校验-镜像存储设备 1602 上生成所述一个或多个奇偶校验数据段，从奇偶校验-镜像存储设备 1602 恢复的。

在另一种实施方式中，装置 1700 包括后整合恢复模块 1724，该后整合恢复模块 1724 恢复存储在存储集的存储设备 150 上的数据段。在这种实施方式中，存储设备 150 不可用，并且所述数据段是在奇偶校验渐进模块 1708 生成所述一个或多个奇偶校验数据段之后，利用存储在一个或多个奇偶校验-镜像存储设备 150 上的一个或多个奇偶校验数据段恢复的。例如，后整合恢复模块 1724 利用奇偶校验数据段和可用的 N 个存储设备 150 上的可用的数据段重新创建丢失的数据段。

在一种实施方式中，装置 1700 包括数据重建模块 1724，该数据重建模块 1724 在重建操作期间将恢复的数据段存储在替代存储设备上，其中，所述恢复的数据段与存储在不可用的存储设备 150 上的不可用数据段相匹配。不可用的存储设备 150 是存储设备集 1602 中的 N 个存储设备 150 中的一个。通常，所述重建操作发生在存储所述不可用数据段的存储设备 150 出现故障以后。所述重建操作是将数据段恢复到所述替代存储设备上，以匹配先前存储在不可用存储设备 150 上的数据段。

可为所述重建操作而从数个来源恢复所述数据段。例如，如果匹配的数据段驻留在奇偶校验-镜像存储设备 1602 上，所述数据段可在渐进之前从奇偶校验-镜像存储设备 1602 恢复。在另一个实例中，所述数据段可从包含所述不可用数据段的拷贝的镜像存储设备中恢复。通常，如果所述恢复的数据段不

驻留在一个或多个奇偶校验-镜像存储设备 1602 上，所述数据段是从所述镜像存储设备恢复的，但是，即使匹配的数据段在镜像存储设备上可用，所述数据也可从所述镜像存储设备恢复。

在另一个实例中，如果所述恢复的数据段不驻留在奇偶校验-镜像存储设备 1604 或所述镜像存储设备中，由一个或多个奇偶校验数据段和所述 N 个数据段中的可用数据段再次生成再生数据段。通常，丢失的数据段仅在其不以某种形式存在于另一个存储设备 150 上时才会再生。

在另一种实施方式中，装置 1700 包括奇偶校验重建模块 1726，该奇偶校验重建模块 1726 在奇偶校验重建操作中在替代存储设备上重建恢复的奇偶校验数据段，其中，所述恢复的奇偶校验数据段与存储在不可用的奇偶校验-镜像存储设备上的不可用奇偶校验数据段相匹配。所述不可用的奇偶校验-镜像存储设备是一个或多个奇偶校验-镜像存储设备 1602 中的一个。所述奇偶校验重建操作将奇偶校验数据段恢复到替代存储设备以匹配先前存储在不可用奇偶校验-镜像存储设备上的奇偶校验数据段。

为了在所述重建操作中再生所述恢复的奇偶校验数据段，用于重建的数据可以来自不同的源。在一个实例中，利用存储在第二存储设备 150 集的奇偶校验-镜像存储设备 1602（存储所述条带的镜像拷贝）上的奇偶校验数据段恢复所述恢复的奇偶校验数据段。当镜像拷贝可用时，利用镜像奇偶校验数据段是可取的，这是由于不需要重新计算所述恢复的奇偶校验数据段。在另一个实例中，如果所述 N 个数据段在 N 个存储设备上可用，则由存储在 N 个存储设备 150 中的一个上的所述 N 个数据段再次生成所述恢复的奇偶校验数据段。通常，当单一故障发生在正在被重建的奇偶校验-镜像存储设备 1602 上时，所述 N 个数据段在 N 个存储设备 150 上可用。

在另一个实例中，如果 N 个数据段中的一个或多个在第一存储设备集 1604 的 N 个存储设备 150 上不可用并且匹配的奇偶校验数据段在第二存储设备 150 集上不可用，则由第二存储设备 150 集的一个或多个存储设备 150（存储所述 N 个数据段的拷贝）再次生成所述恢复的奇偶校验数据段。在又一种实施方式中，由可用数据段和不匹配的奇偶校验数据段再次生成所述恢复的奇偶校验数据段，而不考虑这些数据段在一个或多个存储设备 150 集中的位置。

在奇偶校验-镜像存储设备在存储设备集 1604 中的存储设备 150 之间更替的情况下，通常，数据重建模块 1724 和奇偶校验重建模块 1726 结合在一起工作以在重建的存储设备 150 上重建数据段和奇偶校验数据段。当第二奇偶校验-镜像存储设备 1602b 可用时，数据重建模块 1724 和奇偶校验重建模块 1726 能够在存储设备集 1604 的两个存储设备 150、1602 出现故障后重建两个存储设备。在奇偶校验-镜像存储设备 1602 还未渐进到创建奇偶校验-镜像数据段的情况下，数据段或存储设备 150 的恢复速度比下列事件之后的数据段或存储设备 150 的恢复速度快：奇偶校验-镜像存储设备 1602 已经渐进、计算并存储了用于条带的奇偶校验数据段和用于计算所述奇偶校验数据段的奇偶校验-镜像存储设备 1602 上的 N 个数据段已经被删除。

图 18 是示意性框图，示出了根据本发明的利用渐进式 RAID 更新数据段的装置 1800 的一种实施方式。通常，装置 1800 涉及 RAID 群组，其中，一个或多个奇偶校验-镜像存储设备 1602 已经渐进并且包括奇偶校验数据段(不包括用以创建所述奇偶校验数据段的所述 N 个数据段)。装置 1800 包括更新接收器模块 1802、更新复制模块 1804、奇偶校验更新模块 1806，这些模块在下文中描述。装置 1800 的模块 1802-1806 被描述位于服务器 112 内，但也可位于存储设备 150、客户端内部或位于设备的任意组合的内部，或者分布在数个设备之间。

条带、数据段、存储设备 150、存储设备集 160、奇偶校验数据段、和一个或多个奇偶校验-镜像存储设备 1602 大体上类似于上文中相对于图 11 的装置 1700 描述的条带。装置 1800 包括更新接收器模块 1802，该更新接收器模块 1802 接收已更新的数据段，其中，所述已更新的数据段对应于现有条带的现有数据段。在另一种实施方式中，更新接收器模块 1802 还可接收多个更新信息并可一起或分别处理所述更新信息。

装置 1800 包括更新复制模块 1804，该更新复制模块 1804 将已更新的数据段复制到存储相应的现有数据段的存储设备 150，并将所述已更新的数据段复制到一个或多个对应于所述现有条带的奇偶校验-镜像存储设备 1602。在另一种实施方式中，更新复制模块 1804 将所述已更新的数据段复制到奇偶校验-镜像存储设备 1602 或存储所述现有数据段的存储设备 150，并然后验证所述已更新的数据段的拷贝被转发给其他设备 1602、150。

装置 1800 包括奇偶校验更新模块 1806，该奇偶校验更新模块 1806 为响应存储整合操作而为所述现有条带的一个或多个奇偶校验-镜像存储设备计算一个或多个已更新的奇偶校验数据段。所述存储整合操作类似于上文中相对于图 11 的装置 1700 描述的存储整合操作。所述存储整合操作旨在利用一个或多个已更新的奇偶校验数据段在一个或多个奇偶校验-镜像存储设备 1602 上至少恢复存储空间和/或数据。通过等待更新一个或多个奇偶校验数据段，更新可以被推迟到更合适的时候或等到需要整合存储空间的时候。

在一种实施方式中，由所述现有奇偶校验数据段、所述更新的数据段和所述现有数据段计算所述已更新的奇偶校验数据段。在一种实施方式中，所述现有数据段在为生成所述更新的奇偶校验数据段而读取所述现有数据段之前被保持在一个位置。这种实施方式的一个好处是：可将与复制所述现有数据段到奇偶校验-镜像存储设备 1602 或其他生成所述更新的奇偶校验数据段的位置有关的开销推迟到必要的时候。这种实施方式的一个不足是：如果保持所述现有数据段的存储设备 150 出现故障，在生成所述已更新的奇偶校验数据段之前必须恢复所述现有数据段。

在另一种实施方式中，当 N 个存储设备 150a-n 中的存储所述现有数据段的存储设备 150 接收所述更新的数据段的拷贝时，所述现有数据段被复制到数据-镜像存储设备 1602。然后，存储所述现有数据段，直到所述存储整合操作。在另一种实施方式中，如果所述存储整合操作发生在触发计算所述已更新的奇偶校验数据段的存储整合操作之前，则所述现有数据段被复制到数据-镜像存储设备 1602，以响应 N 个存储设备 150a-n 中的存储所述现有数据段的存储设备 150 上的存储整合操作。后一种实施方式是有利的，这是因为直到存储所述现有数据段的存储设备 150 上的或奇偶校验-镜像存储设备 1602 上的存储整合操作需要才复制所述现有数据段。

在一种实施方式中，由所述现有奇偶校验数据段、所述已更新的数据段和增量数据段计算出所述更新的奇偶校验数据段，其中，所述增量数据段按所述更新的数据段和所述现有数据段之间的差异生成。通常，生成增量数据段是更新所述奇偶校验数据段中的部分解决方案或中间步骤。生成增量数据段是有利的，这是因为所述增量数据段可以被高度压缩并可以在传送之前被压缩。

在一种实施方式中，在为了生成所述已更新的奇偶校验数据段而读取所述增量数据段之前，所述增量数据段存储在存储所述现有数据段的存储设备上。在另一种实施方式中，当存储所述现有数据段的存储设备 150 接收所述已更新的数据段的拷贝时，所述增量数据段被复制到数据-镜像存储设备 1602。在另一种实施方式中，所述增量数据段被复制到数据-镜像存储设备 1602，以响应存储所述现有数据段的存储设备 150 上的存储整合操作。正如复制所述现有数据段一样，后一种实施方式是有利的，这是因为直到所述现有数据段上的存储整合操作之前或触发计算所述已更新的奇偶校验数据段的另一个存储整合操作之前，才移动增量数据文件。

在不同的实施方式中，模块 1802、1804、1806 的操作的一部分或全部（即接收已更新的数据段、复制所述已更新的数据段和计算所述已更新的奇偶校验数据段）发生在下述设备上：存储设备集 1604 的存储设备 150、客户端 114 或第三方 RAID 管理设备。在另一种实施方式中，独立于所述更新接收器模块 1802 的操作和更新复制模块 1804 的操作进行所述存储整合操作。

图 19 是示意性流程图，示出了根据本发明的利用渐进式 RAID 管理数据的方法 1900 的一种实施方式。方法 1900 始于步骤 1902，存储请求接收器模块 1702 接收存储数据的请求（步骤 1904），其中，所述数据是文件的数据或对象的数据。条带化模块 1704 为所述数据计算条带模式并将所述 N 个数据段写入 N 个存储设备 150（步骤 1906）。所述条带模式包括一个或多个条带。每个条带包括 N 个数据段的集，其中，所述 N 个数据段中的每一个被写入分配给所述条带的存储设备集 1604 中不同的存储设备 150。

奇偶校验-镜像模块 1706 将所述条带的 N 个数据段的集写入存储设备集 1604 中的一个或多个奇偶校验-镜像存储设备 1602（步骤 1908）。一个或多个奇偶校验-镜像存储设备是除 N 个存储设备 150 之外的设备。奇偶校验生成模块 1708 确定是否有等待中的存储整合操作（步骤 1910）。如果奇偶校验生成模块 1708 确定没有等待中的存储整合操作，方法 1900 返回并再次确定是否有等待中的存储整合操作。在其他实施方式中，存储请求接收器模块 1702、条带化模块 1704 和奇偶校验-镜像模块 1706 继续接收存储请求、计算条带模式和存储数据段。

如果奇偶校验生成模块 1708 确定没有等待中的存储整合操作（步骤

1910)，奇偶校验生成模块 1708 为所述条带计算奇偶校验数据段(步骤 1914)。由存储在奇偶校验-镜像存储设备 1602 上的 N 个数据段计算所述奇偶校验数据段。奇偶校验生成模块 1708 将所述奇偶校验数据段存储在奇偶校验-镜像存储设备 1602 上(步骤 1912)，方法 1900 结束于步骤 1916。所述存储整合操作的执行独立于接收存储 N 个数据段的请求(步骤 1904)、将 N 个数据段写入 N 个存储设备(步骤 1906)或将 N 个数据段写入一个或多个奇偶校验-镜像存储设备(步骤 1908)。所述存储整合操作旨在至少恢复奇偶校验-镜像存储设备 1602 上的存储空间或数据。

图 20 是示意性流程图，示出了根据本发明的利用渐进式 RAID 更新数据段的方法 2000 的一种实施方式。方法 2000 始于步骤 2002，更新接收器模块 1802 接收已更新的数据段(步骤 2004)，其中，所述已更新的数据段对应于现有条带的现有数据段。更新复制模块 1804 将所述已更新的数据段复制到存储相应的现有数据段的存储设备 150 和对应于所述现有条带的一个或多个奇偶校验-镜像存储设备 1602(步骤 2006)中。

奇偶校验更新模块 1806 确定是否存储整合操作在等待中(步骤 2008)。如果奇偶校验更新模块 1806 确定没有等待中的存储整合操作(步骤 2008)，奇偶校验更新模块 1806 等待存储整合操作。在一种实施方式中，方法 2000 返回并接收其他已更新的数据段(步骤 2004)，并复制所述已更新的数据段(步骤 2006)。如果奇偶校验更新模块 1806 确定没有等待中的存储整合操作(步骤 2008)，奇偶校验更新模块 1806 为所述现有条带的一个或多个奇偶校验-镜像存储设备计算一个或多个已更新的奇偶校验数据段(步骤 2010)，方法 2000 结束于步骤 2012。

前端分布式RAID

传统的 RAID 系统被配置为与 RAID 控制器一起使用，所述 RAID 控制器具有如下功能：接收数据、为所述数据计算条带模式、将所述数据分割为数据段，计算奇偶校验条带、将所述数据存储在存储设备上、更新所述数据段等等。当一些 RAID 控制器允许一些功能成为分布式的功能时，由 RAID 控制器管理的存储设备不直接与客户端通信以存储在 RAID 中条带化的数据。用于 RAID 过程的替代存储请求和数据通过所述存储控制器。

要求所述 RAID 控制器接触所有将要被存储在 RAID 中的数据是没有效率

的，这是因为这种方法产生了数据流瓶颈。这个问题在读-修改-写处理期间尤为突出，其中，RAID 群组中的全部驱动器的带宽和性能被消耗，而实际上仅更新了子集。此外，被指定用于由所述 RAID 控制器管理的数据的存储设备中的区域通常用于 RAID 群组并且不能独立地被访问。通过客户端访问存储设备 150 通常通过分区存储设备 150 来实现。当使用分区时，支持普通存储访问的分区不用于 RAID，而分配给 RAID 群组的分区不支持普通数据存储访问。为全域性地优化效用而超额预定分区的方案不仅复杂而且更加难以管理。此外，分配给一个 RAID 群组的存储空间不能通过多于一个的 RAID 控制器访问，除非一个被指定为主控制器，而其他 RAID 控制器作为从机，除非主 RAID 控制器未被激活、丧失功能等等。

典型的 RAID 控制器还在 RAID 群组的存储设备 150 之外生成奇偶校验数据段。这可能是无效率的，这是因为奇偶校验数据段通常在生成之后被发送给存储设备 150 以便于存储，这需要 RAID 控制器的计算能力。追踪奇偶校验数据段的位置和更新信息还必须在 RAID 控制器内完成而不是在自主地在存储设备 150 上完成。

如果独立的 RAID 控制器断开连接，当有必要确保所述数据保持在可用状态时，RAID 控制器通常互相交叉连接并交叉连接至驱动器，和/或像成套设备一样镜像化，但这样使数据可用性管理的花费昂贵且难以管理，还显著地降低了存储子系统的可靠性。

需要一种用于前端分布式 RAID 系统、装置和方法，所述前端分布式 RAID 允许在每个数据段、每个对象、每个文件或类似基础上使用 RAID，所述前端分布式 RAID 无需 RAID 控制器和位于客户端和存储设备之间的 RAID 控制器对。在这种系统、装置和方法中，RAID 群组可被创建用于一个数据段、对象或文件，该 RAID 群组还可在一个存储设备群组中由一个 RAID 控制器管理，而第二 RAID 控制器可被创建用于包含第一 RAID 群组的一些相同的存储设备的另一个数据段、对象或文件。RAID 控制功能可分布在客户端 114、第三方 RAID 管理设备之间，或分布在多个存储设备 150 之间，前端分布式 RAID 系统、装置和方法还可将命令发送给 RAID 群组的存储设备 150 并可允许存储设备 150 通过直接存储器存取（“DMA”）或远程 DMA（“RDMA”）直接访问和复制数据。

图 10 是示意性框图，示出了根据本发明的可被前端分布式 RAID 访问的系统 1600 的一种实施方式。上文中对图 16 中相对于渐进式 RAID 描述的组件进行的说明也可应用到前端分布式 RAID。对于前端分布式 RAID，存储设备集 1604 形成 RAID 群组并包括自主运行且能够独立地通过网络 116 或一个或多个冗余网络 116 接收和服务来自客户端 114 的存储请求的存储设备 150。

在存储设备集 1604 中的存储设备 150 之中，一个或多个存储设备 150 被指定为用于条带的奇偶校验-镜像存储设备 1602。通常，一个或多个奇偶校验-镜像存储设备 1602 的功能大体上类似于其他存储设备 150。在典型的配置中，指定的奇偶校验-镜像存储设备 1602 在存储设备集 1604 的存储设备 150 之间变更，奇偶校验-镜像存储设备 1602 实质上具有与其他存储设备 150 一样的特点，这是由于奇偶校验-镜像存储设备 1602 也必须像非奇偶校验-镜像存储设备一样运行。类似的特点是关于上述的 RAID 群组内的操作和用于客户端 114 独立通信的自主操作。在不同的实施方式中，存储设备集 1604 的存储设备 150 可在其他方面（不涉及所述 RAID 环境下的功能）不同。

存储设备集 1604 的存储设备 150 可以是独立的、可以是在一个或多个服务内成组的、可每一个驻留在一个服务器 112 内、可通过一个或多个服务器 112 被访问，等等。一个或多个客户端 114 可驻留在包括一个或多个存储设备 150 的服务器 112 内、可驻留在独立的服务器 112 内、可驻留在通过一个或多个计算网络 116 访问存储设备 150 的计算机、工作站、笔记本电脑等设备内，或位于类似设备内。

在一种实施方式中，网络 116 包括系统总线，并且存储设备集 1604 中的一个或多个存储设备 150、1602 通过所述系统总线通信。例如，系统总线可以是 PCI-e 总线、串行高级技术附件（“串行 ATA”）总线、并行 ATA 或类似总线。在另一种实施方式中，所述系统总线是外部总线，如小型计算机系统接口（“SCSI”）、火线、光纤通道、USB、PCIe-AS、无限带宽或类似总线。本领域技术人员会意识到具有存储设备 150 的其他系统 1600 配置，其中，存储设备 150 不仅自主运行，还能够独立地通过一个或多个网络 116 接收和服务来自客户端 114 存储请求。

图 11 是示意性框图，示出了根据本发明的用于前端分布式 RAID 的装置 2100 的一种实施方式。在不同的实施方式中，装置 2100 包括存储请求接收器

模块 2102、条带化关联模块 2104、奇偶校验-镜像关联模块 2106、存储请求发送器模块 2108、前端奇偶校验生成模块 2110、奇偶校验更替模块 2112、数据段恢复模块 2114、数据重建模块 2116、奇偶校验重建模块 2118 和对等通信模块 2120，这些模块在下文中描述。在不同的实施方式中，装置 2100 可被包括在下列设备中：存储设备 150（如固态存储设备 102）、存储设备控制器 152（如固态存储控制器 104）、服务器 112、第三方 RAID 管理设备等等，或者装置 2100 分布在不止一个的组件之间。

装置 2100 包括存储请求接收器模块 2102，该存储请求接收器模块 2102 接收将数据存储在存储设备集 1604 中的存储请求。所述数据可以是文件或对象的一部分，或者可以是整个文件或对象。文件可包括任意信息块或用于存储信息的源，其中，这些块或源可用于计算机程序。文件可包括由处理器访问的任意数据结构。文件可包括数据库、文本串、计算机编码等等。对象通常是用于面向对象的编程的数据结构并且可包括具有（或不具有）数据的结构。在一种实施方式中，对象是文件的子集。在另一种实施方式中，对象独立于文件。在任何情况下，对象和文件在此处被定义为包括数据、数据结构、计算机编码和其他存储在存储设备上的信息的全部集。

存储设备集 1604 包括形成 RAID 群组的自主存储设备 150，存储设备 150 自主地通过一个或多个网络 116 接收来自客户端 114 的存储请求。存储设备集 1604 的自主存储设备 150 中一个或多个被指定为用于条带的奇偶校验-镜像存储设备 1602。来自其他客户端的其他存储请求可存储在第二存储设备集上，其中，所述第二存储设备集可像第一存储设备集 1604 一样包括一个或多个相同的存储设备 150（和奇偶校验-镜像存储设备 1602）。为两个存储设备集 1604 所共用的存储设备 150 可在其内具有分配为存储空间的重叠部分。

装置 2100 包括为所述数据计算条带模式的条带化关联模块 2104。所述条带模式包括一个或多个条带。每个条带包括 N 个数据段的集。条带的 N 个数据段还可包括一个或多个空数据段。条带化关联模块 2104 将 N 个数据段中的一个数据段与被分配给所述条带的存储设备集 1604 中的 N 个存储设备 150a-n 中的一个关联。在一种实施方式中，条带化关联模块 2104 利用将要被发送给存储设备 150 的存储请求将数据段与存储设备 150 关联，该存储请求指令存储设备获取对应于来自发送所述存储请求的客户端 114 的数据段的数据。

在另一种实施方式中，所述存储请求大体上与所述数据段的数据无关。大体上与数据无关意味着所述存储请求一般来说不包括作为所述存储请求的主题的数据，但可包括可能是数据的一部分的字符、字符串等等。例如，如果所述数据包括一串重复的、相同的字符（如一串 0 字符），所述存储请求可包括所述数据包括一串 0 字符的指示而并不包括所述数据中的所有零字符。本领域技术人员会认识到发送存储请求而不发送数据的主体但同时仍然允许少量的或单个实例的某些字符或字符串存在于所述存储请求中的其他方法。所述存储请求包括命令，该命令允许 N 个存储设备 150a-n 利用 DMA 或 RDMA 操作或类似操作检索所述数据。

在另一种实施方式中，条带化关联模块 2104 通过在将要被发送给存储设备 150 的存储请求中识别数据段的数据将所述数据段与存储设备 150 关联。识别所述数据段的数据可包括数据段标识符、数据段位置或地址、数据段长度或其他允许存储设备 150 判定哪个数据包括所述数据段的信息。

在一种实施方式中，条带化关联模块 2104 在存储请求中将数据段与存储设备 150 关联，以使得客户端 114 能够在广播中发送包括所述数据段的数据，从而每个存储设备 150 能够存储关联的数据段并丢弃对应于未被分配给存储设备 150 的数据段的数据。在另一种实施方式中，条带化关联模块 2104 在存储请求中可能是通过为每个数据段分配地址将数据段与存储设备 150 关联，以使得客户端 114 可在组播中发送包括所述数据段的数据，从而每个存储设备 150 能够存储关联的数据段并丢弃对应于未被分配给存储设备 150 的数据段的数据。本领域技术人员会认识到用于条带化关联模块 2104 将数据段与存储设备 150 关联，从而将一个或多个数据段通过下述方式传给一个或多个存储设备的其他方法：广播、组播、单播、任意播等。

在一种相关的实施方式中，条带化关联模块 2104 在存储请求中将数据段与存储设备 150 关联，以使得客户端 114 能够广播、组播、单播（等）所述存储请求，并且每个存储设备 150 能够接收来自客户端 114 的涉及与存储设备 150 关联的所述数据段的存储请求的一部分，还能够丢弃不涉及与存储设备 150 关联的一个或多个数据段的存储请求的那部分。

在另一种实施方式中，由存储请求接收器模块 2102 接收的所述存储请求包括作为所述存储请求主题的数据，并且条带化关联模块 2104 通过准备用于

包括数据段的存储设备 150 的存储请求将数据段与存储设备 150 关联。条带化关联模块 2104 可运行在下列设备内：客户端 114、第三方 RAID 管理设备、存储设备 150、1602，等等。

装置 2100 包括奇偶校验-镜像关联模块 2106，该奇偶校验-镜像关联模块 2106 将 N 个数据段的集与存储设备集 1604 中的一个或多个奇偶校验-镜像存储设备 1602 关联。一个或多个奇偶校验-镜像存储设备 1602 是除 N 个存储设备 150a-n 之外的设备。在一种实施方式中，奇偶校验-镜像关联模块 2106 将 N 个数据段的集与每个奇偶校验-镜像存储设备 1602 关联，从而每个奇偶校验-镜像存储设备 1602 能够为了生成奇偶校验数据段而接收并存储条带的 N 个数据段。在另一种实施方式中，奇偶校验-镜像关联模块 2106 将条带的数据段与每个奇偶校验-镜像存储设备 1602 关联，从而奇偶校验-镜像存储设备 1602a-m 充当存储在 N 个存储设备 150a-n 中的 N 个数据段的镜像。

在不同的实施方式中，奇偶校验-镜像关联模块 2106 利用单个存储请求、多个存储请求或上文中相对于条带化关联模块 2104 描述的其他关联技术（如为了 DMA、RDMA、广播、组播而设立奇偶校验-镜像存储设备 1602 的存储请求，或将 N 个数据段包括在存储请求中）将 N 个数据段的集与一个或多个奇偶校验-镜像存储设备 1602 关联。奇偶校验-镜像关联模块 2106 可运行在下列设备内：客户端 114、第三方 RAID 管理设备、存储设备 150、1602，等等。

装置 2100 包括存储请求发送器模块 2108，该存储请求发送器模块 2108 将一个或多个存储请求发送给存储设备集 1604 中的每个存储设备，每个存储请求能够将与接收所述存储请求的存储设备 150、1602 关联的一个或多个数据段存储在存储设备 150、1602 上。在一种实施方式中，每个存储请求不包括作为所述存储请求的主题的数据。在另一种实施方式中，每个存储请求使得存储设备集 1604 的 N 个存储设备 150 和奇偶校验-镜像存储设备 1602 能够利用 DMA 或 RDMA 下载关联的数据段的数据。在另一种实施方式中，存储请求包含足够的信息以从来自客户端 114 的广播中挑选用于关联的数据段的相关存储请求或相关数据。在另一种实施方式中，存储请求包括关联的数据段的数据。

在一种实施方式中，每个存储请求识别作为条带的存储设备集 1604 的一部分的存储设备 150、1602。通过包括识别存储设备集 1604 的存储设备 150、

1604 的步骤，如果充当主机的存储设备 150 出现故障，另一个存储设备 150 可接管主机以管理 RAID 数据。在另一种实施方式中，当存储设备断开连接时，识别存储设备集 1604 使得自主存储设备 150、1602 能够恢复数据，并当替代存储设备被附加到存储设备集 1604 内时，使得自主存储设备 150、1602 能够独立于客户端重建数据。在另一种实施方式中，识别存储设备集 1604 的存储设备 150、1602 代表了用于传送数据段或存储请求的组播组。识别信息可与存储在存储设备集 1604 的存储设备 150、1602 上的、用于对象或文件的元数据一起存储。

在一种实施方式中，装置 2100 包括前端奇偶校验生成模块 2110，当奇偶校验-镜像关联模块 2106 将 N 个数据段的集与一个或多个奇偶校验-镜像存储设备 1602 中的每一个关联时，该前端奇偶校验生成模块 2110 独立于客户端 114 为所述条带计算奇偶校验数据段，并将所述奇偶校验数据段存储在奇偶校验-镜像存储设备 1602 上。由提供给奇偶校验-镜像存储设备 1602 的 N 个数据数据段的集计算所述奇偶校验数据段。当存储设备集 1604 包括了不止一个奇偶校验-镜像存储设备 1602 时，前端奇偶校验生成模块 2110 通常生成不同的奇偶校验数据段，从而存储设备集 1604 中的两个或更多个存储设备 150、1602 可出现故障，并且奇偶校验数据段信息允许恢复不可用的数据段或奇偶校验数据段。

在另一种实施方式中，当运行在存储设备集 1604 的存储设备 150 中和/或第三方 RAID 管理设备中时，前端奇偶校验生成模块 2110 计算所述奇偶校验数据段。例如，独立于客户端 114 的、发送所述存储请求的服务器 112 可计算所述奇偶校验数据段。在另一种实施方式中，前端奇偶校验生成模块 2110 运行在奇偶校验-镜像存储设备内以计算所述奇偶校验数据段。例如，奇偶校验-镜像存储设备 1602 中的存储控制器 152 可充当用于由存储设备集 1604 形成的 RAID 群组的主存储控制器。

在另一种实施方式中，前端奇偶校验生成模块 2110 计算所述奇偶校验数据段并将计算出的奇偶校验数据段发送给开成镜像的第二存储设备集中的一一个或多个附加奇偶校验-镜像存储设备 1604。这种实施方式是有利的，这是因为与计算奇偶校验数据段有关的开销只需要一次，而不需要为每个存储设备集 1604 执行开销，这样做的额外好处是减少了网络 116 的数据流量。

在一种实施方式中，装置 2100 还可包括数据段恢复模块 2112，如果存储设备 150 不可用并且接收到读取不可用数据段或包括不可用数据段的数据的请求，数据段恢复模块 2112 恢复存储在存储设备集 1604 的存储设备 150 上的数据段。利用存储设备集 1604 的可用存储设备 150 上的数据段、奇偶校验数据段和存储设备集 1604 的可用存储设备 150、1602 上的数据段的结合恢复所述数据段，或者从包括所述数据段的拷贝的镜像存储设备中恢复所述数据段。通常，镜像存储设备是存储 N 个数据段的拷贝的存储设备集的一个存储设备 150。数据段恢复模块 2112 可运行并恢复来自下述设备中的不可用数据段：存储设备 150、奇偶校验-镜像存储设备 1602、第三方 RAID 管理设备、镜像存储设备等等。

在另一种实施方式中，装置 2100 包括数据重建模块 2114，该数据重建模块 2114 在重建操作中将恢复的数据段存储在替代存储设备 150 上。例如，如果存储设备 150 由于出现故障、失去同步性等原因而变得不可用，数据重建模块 2114 可重建存储设备 150 以替换不可用的存储设备 150。在一种实施方式中，重建的存储设备 150 是已经可用的源存储设备 150。

所述恢复的数据段与存储在存储设备集 1604 的不可用存储设备 150 上的不可用数据段匹配。所述重建操作通常将一个或多个数据段和奇偶校验数据段恢复到替代存储设备 150 上，从而使其与先前存储在不可用存储设备 150 上的数据段和奇偶校验数据段相匹配。

在一种实施方式中，所述恢复的数据段利用存储设备集 1604 的可用存储设备 150 上的可用数据段被恢复用于重建操作。在另一种实施方式中，所述恢复的数据段利用来自一个或多个奇偶校验-镜像存储设备 1602 的奇偶校验数据段和存储设备集 1604 的可用存储设备 150 上的可用数据段的结合被恢复用于重建操作。在另一种实施方式中，所述恢复的数据段利用读取自奇偶校验-镜像存储设备 1602 的匹配数据段被恢复用于重建操作。在又一种实施方式中，所述恢复的数据段利用来自镜像存储设备的匹配数据段被恢复用于重建操作。数据重建模块 2114 能够运行并存储接收自下述设备的数据段：客户端 114、第三方 RAID 管理设备、存储设备 150、1602、镜像存储设备等等。

在另一种实施方式中，装置 2100 包括奇偶校验重建模块 2116，该奇偶校验重建模块 2116 在重建操作中在替代存储设备 1602 上重建恢复的奇偶校验

数据段。重建操作大体上与上文中相对于数据重建模块 2114 描述的重建操作类似。奇偶校验重建模块 2116 类似于数据重建模块 2114 运行，除了奇偶校验重建模块 2116 重建奇偶校验数据段。恢复的奇偶校验数据段与存储在分配给条带的不可用奇偶校验-镜像存储设备 1602 上的不可用奇偶校验数据段相匹配。

在不同的实施方式中，通过下述方法恢复所述奇偶校验数据段：复制存储在镜像存储设备集的奇偶校验-镜像存储设备 1602 上的所述奇偶校验数据段、从存储设备集 1604 的奇偶校验-镜像存储设备 1602 复制所述奇偶校验数据段（如果与不可用奇偶校验数据段一致）、利用存储在存储设备集 1604 的可用存储设备 150、1602 和包含数据段的拷贝的镜像存储设备上的 N 个数据段中的一个或多个和奇偶校验数据段生成所述奇偶校验数据段，等等。数据重建模块 2116 可驻留在下列设备上并运行和存储恢复的数据段：客户端 114、第三方 RAID 管理设备、存储设备 150、镜像存储设备等等。

有利地是，装置 2100 并不限于在存储设备 150、1602 中将数据存储至用于此处描述的前端分布式 RAID 操作的分区。作为替代的是，自主存储设备（如 150a）可独立地接收来自客户端 114 的将经 RAID 或未经 RAID 的数据存储在存储设备 150a 的一个或多个区域中，存储设备 150a 依然可被条带化关联模块 2104、奇偶校验-镜像关联模块 2106 和前端奇偶校验生成模块 2110 用于存储数据。

在一种实施方式中，由存储请求接收器模块 2102 接收的或由存储请求发送器模块 2108 发送的一个或多个存储请求识别包括条带的存储设备集 1604 的存储设备 150。有利地是，如果主控制器丧失功能，在存储请求中识别存储设备集 1604 的存储设备 150 有助于备份 RAID 控制器运行。例如，如果存储设备集 1604 的存储设备 150 在存储请求中被识别并且位于奇偶校验-镜像存储设备 1602 中的所述主控制器不可用，另一个奇偶校验-镜像存储设备 1602 或 N 个存储设备 150a-n 中的另一个可以成为所述主控制器。

在一种实施方式中，装置 2100 包括奇偶校验更替模块 2118，该奇偶校验更替模块 2118 为每个条带变更被分配为用于所述条带的奇偶校验-镜像存储设备 1602 的存储设备集 1604 中的存储设备 150。奇偶校验更替模块 2118 的优点已在上文中描述。在另一种实施方式中，存储设备集 1604 的存储设备 150

形成对等群组，并且装置 2100 包括对等通信模块 2120，该对等通信模块 2120 在存储设备集 1604 的存储设备 150、1602 内发送并接收存储请求。对等通信模块 2120 还可在存储设备集 1604 外部的对等设备中发送并接收存储请求。

在一种优选实施方式中，所述存储请求是通过利用装置 2100 的模块 2102-2120 在存储设备 1604 的存储设备 150、1602 之间条带化对象的数据来存储对象的对象请求。在另一种实施方式中，存储设备集 1604 的自主存储设备 150、1602 中的一个或多个被分配到第一 RAID 群组中用于第一对象或文件的至少一部分，并被分配到第二 RAID 群组中用于第二对象或文件的至少一部分。例如，一个存储设备 150a 可以是用于一个或多个条带的存储设备集 1604 的主 RAID 控制器，而第二存储设备 150b 可以是用于包括了存储设备集的一些或全部存储设备 150 的 RAID 群组的主 RAID 控制器。有利地是，装置 2100 允许灵活的分组存储设备 150、1602 以形成不同客户端 114 的 RAID 群组。

图 12 是示意性流程图，示出了根据本发明的用于前端分布式 RAID 的方法 2200 的一种实施方式。方法 2200 始于步骤 2202，存储请求接收器模块 2102 接收将数据存储在存储设备集 1604 的存储设备 150a-n 中的存储请求（步骤 2204）。条带化关联模块 2104 计算用于所述数据的条带模式（步骤 2206）并将 N 个数据段中的每一个数据段与 N 个存储设备 150a-n 中的一个关联（步骤 2208）。

奇偶校验-镜像关联模块 2106 将 N 个数据段的集与一个或多个偶校验-镜像存储设备 1602 关联（步骤 2210）。存储请求发送器模块 2108 将一个或多个存储请求发送给存储设备集 1604 中的每一个存储设备 150、1602（步骤 2212）。每个存储请求足以将与接收所述存储请求的存储设备 150 关联的一个或多个数据段存储在存储设备 150 上。然后，所述数据的数据段像被所述存储请求指令一样利用 DMA、RDMA、广播、组播等技术被传送给存储设备集 1604 的存储设备 150、1602。可选择地，前端奇偶校验生成模块 2110 为条带计算奇偶校验数据段（步骤 2214），方法 2200 结束于步骤 2216。

共享的、前端、分布式 RAID

传统的RAID利用磁盘或其他存储设备的阵列，其中，所述存储设备中的每一个的至少一部分被用于RAID并形成RAID群组。RAID控制器管理传送到

所述RAID群组的存储语法。对于冗余系统来说，RAID控制器具有备用RAID控制器，如果主RAID控制器出现故障或不可用，该备用RAID控制器准备好接管主RAID控制器。来自试图访问存储在RAID中相同数据的多个客户端的存储请求按到达的顺序被顺序地执行。

前端、分布式的RAID系统包括自主存储设备，所述自主存储设备中的每一个都包括起分布式RAID控制器作用的存储控制器，并且所述存储设备均能被配置在多个、重叠的、服务多个客户端的RAID群组中。必要的时候，两个客户端可试图访问相同的数据。如果一个存储请求先到达并执行，通常就不会出现数据的不一致。另一方面，如果用于同一数据的两个或更多个存储请求同时到达或几乎同时到达，数据可能会被损坏。

例如，如果数据存储在RAID群组中的四个存储设备中，其中，所述存储设备中的一个被指定为奇偶校验-镜像存储设备，第一客户端将存储请求发送给作为RAID控制器的第一存储控制器，第二客户端将第二存储请求发送给作为第二RAID控制器的第二存储设备，并且这两个存储请求访问相同的数据，所述第一存储设备可开始在所述第一存储设备上执行所述存储请求，然后，在RAID群组中的另一个存储设备上执行所述存储请求。同时，第二存储设备上的所述第二RAID控制器可开始在另一个存储设备上执行所述第二存储请求，然后，在RAID群组中余下的存储设备中执行所述第二存储请求。这种执行上的不匹配可能是由于下述原因：存储设备之间的物理距离、执行时间不一致，等等。这种方法可能会损坏数据。

需要一种用于处理访问相同数据的并发存储请求的共享的、前端、分布式RAID的系统、装置和方法。有利地是，这种系统、装置和方法可控制访问数据，使得执行完一个存储请求后再执行第二存储请求。

图 10 是示意性框图，示出了根据本发明的充当用于共享的、前端分布式 RAID（除了渐进式 RAID 和前端分布式 RAID 之外）的系统 1600 的一种实施方式。上文中对图 16 中相对于渐进式 RAID 和前端分布式 RAID 描述的组件的说明同样适用于共享的前端分布式 RAID。正如前端分布式 RAID 一样，存储设备集 1604 形成 RAID 群组并包括存储设备 150，存储设备自主运行并能够独立地通过网络 116 接收并服务来自客户端 114 的存储请求。

对于共享的、前端分布式 RAID，系统 1600 包括两个或更多个客户端 114，

从而两个或更多个客户端 114 中的每一个发送涉及相同数据的存储请求。所述存储请求是并发的，这是由于所述存储请求的到达使得在另一个存储请求到达之前，一个存储请求还没有完成。存储设备集 1604 的存储设备 150 之中的一个或多个被指定为用于条带的奇偶校验-镜像存储设备 1602。通常，一个或多个奇偶校验-镜像存储设备 1602 的功能大体上类似于其他存储设备 150。

在典型的配置中，指定的奇偶校验-镜像存储设备 1602 在存储设备集 1604 的存储设备 150 之间轮换，奇偶校验-镜像存储设备 1602 实质上具有与其他存储设备 150 相同的特点，这是因为奇偶校验-镜像存储设备 1602 也必须像非奇偶校验-镜像存储设备一样运行。类似的特点是相对于上述的 RAID 群组内的操作和用于客户端 114 独立通信的自主操作。在不同的实施方式中，存储设备集 1604 的存储设备 150 可在其他方面（不涉及所述 RAID 环境下的功能）不同。

存储设备集 1604 的存储设备 150 可以是独立的、可以是在一个或多个服务内成组的、可每一个驻留在一个服务器 112 内、可通过一个或多个服务器 111 被访问，等等。一个或多个客户端 114 可驻留在包括一个或多个存储设备 150 的服务器 112 内、可驻留在独立的服务器 112 内、可驻留在通过一个或多个计算网络 116 访问存储设备 150 的计算机、工作站、笔记本电脑等设备内，或位于类似设备内。

在一种实施方式中，网络 116 包括系统总线，并且存储设备集 1604 中的一个或多个存储设备 150、1602 通过所述系统总线通信。例如，系统总线可以是 PCI-e 总线、串行高级技术附件（“串行 ATA”）总线、并行 ATA 或类似总线。在另一种实施方式中，所述系统总线是外部总线，如小型计算机系统接口（“SCSI”）、火线、光纤通道、USB、PCIe-AS、无限带宽或类似总线。本领域技术人员会意识到具有存储设备 150 的其他系统 1600 配置，其中，存储设备 150 自主运行，并能够独立地通过网络 116 接收和服务来自客户端 114 存储请求。

图 13 是示意性框图，示出了根据本发明的用于共享的、前端分布式 RAID 的装置 2300 的一种实施方式。在不同的实施方式中，装置 2300 包括多存储请求接收器模块 2302、条带化模块 2304、奇偶校验-镜像模块 2306、定序器模块 2308、主验证模块 2310、主确定模块 2312、主错误模块 2314、奇偶校

验生成模块 2316 和奇偶校验更替模块 2318，这些模块在下文中描述。

装置 2300 包括多存储请求接收器模块 2302，该多存储请求接收器模块 2302 接收来自至少两个客户端 114 的至少两个存储请求，以将数据存储在存储设备集 1604 的存储设备 150 中。所述数据包括文件的数据或对象的数据。与装置有关的存储请求中的每一个至少具有一部分共有数据，此外，所述存储请求是并发存储请求，这是由于所述存储请求的到达使得在另一个存储请求到达之前，一个存储请求还没有完成。这些并发的存储请求具有损坏前端分布式 RAID 系统 1600 的共有数据的风险。在一种实施方式中，所述并发的存储请求可来自一个客户端 114。在另一种实施方式中，所述并发的存储请求来自两个或更多个客户端 114。

多存储请求可更新存储在存储设备集 1602 的存储设备 150 上的一个或多个数据段，其中，条带化模块 2304 将预先存储的数据条带化为存储在存储设备集 1604 的存储设备 150 上的数据段。在一种实施方式中，存储请求将所述数据第一次写入 RAID 群组。在这种情况下，所述数据通常会存在于其他位置，并可通过一个或多个服务器 114 访问，然后，一个存储请求将所述数据复制到 RAID 群组，而另一个存储请求同时访问所述数据。

多存储请求可包括一个更新存储在存储设备集 1604 的存储设备 150 上的一个或多个数据段的请求，还可包括目标为至少一部分共有数据的一个或多个读取请求。如果更新请求没有完成，则存储设备集 1604 的存储设备 150 返回的读取请求可由预先存在和损坏所述数据的已更新数据的结合组成。

装置 2300 包括条带化模块 2304，该条带化模块 2304 计算（为每个并发存储请求）用于所述数据的条带模式，并将 N 个数据段写入存储设备集 1604 的 N 个存储设备 150a-n。所述条带模式包括一个或多个条带，每个条带包括 N 个数据段的集。N 个数据段中的每一个被写入存储设备集 1604 中的不同的存储设备 150 并被分配给所述条带。装置 2300 包括奇偶校验-镜像模块 2306，该奇偶校验-镜像模块 2306（为每个并发存储请求）将所述条带的 N 个数据段的集写入被指定为奇偶校验-镜像存储设备 1602 的存储设备集 1604 中的存储设备 150。奇偶校验-镜像存储设备 1602 是除 N 个存储设备 150a-n 之外的设备。

条带化模块 2304 还用于计算一个或多个存储设备 150a-n 的一致性，其中，

一个或多个作为文件或对象的一部分的数据段读取自一个或多个存储设备 150a-n。

装置 2300 包括定序器模块 2308，该定序器模块 2308 确保来自第一客户端 114 的第一存储请求完成之后才执行来自第二客户端的第二存储请求，其中，至少两个并发存储请求包括第一和第二存储请求。在其他实施方式中，定序器模块 2308 确保所述第一存储请求完成之后才执行两个或多个其他并发存储请求。有利地是，定序器模块 2308 有助于并发存储请求的顺序执行，从而避免损坏数据。在一种实施方式中，定序器模块 2308 通过下述方法协调并发存储请求的执行：利用所述的存储请求必须访问所述数据的主控制器、利用锁止系统、两阶段提交或本领域技术人员熟知的其他方法。下文描述了定序器模块 2308 使用的一些方法。

在一种实施方式中，定序器模块 2308 通过下述方法确保所述第一存储请求完成之后才执行并发存储请求：接收来自存储设备集 1602 的存储设备 150 中的每一个的应答，其中，存储设备 150 在执行第二存储请求之前与一起接收第一存储请求和存储请求。通常，应答确认存储请求已完成。在一种实施方式中，存储设备 150 中受所述存储影响的每一个设备被写入，且在定序器模块 2308 开始执行第二存储请求之前从每个存储设备 150 接收应答。

在一种实施方式中，完成存储请求可包括执行单个存储设备（如 150a）上的等待中的第二存储请求的一部分之前，完成指令单个存储设备（如 150a）的第一存储请求的一部分。定序器模块 2308 可独立地验证存储设备 150 上的存储请求的一部分是否完成。在这种实施方式中，写入涉及第一存储请求的数据段直到所述第一存储请求的所有数据段被完成后才需要被延迟。定序器模块 2308 可协调发生在存储设备集 1604 的存储设备 150 上的不同请求的执行以确保所述数据不被损坏。

在一种实施方式中，在条带化模块 2304 和奇偶校验-镜像模块 2306 都将与存储请求有关的所述数据段写入存储设备集 1604 的存储设备 150 后，接收存储请求完成的应答。在另一种实施方式中，在条带化模块 2304 和奇偶校验-镜像模块 2306 都将与存储请求有关的所述数据段写入存储设备集 1604 的存储设备 150 且存储设备 150、1602 中的每一个都确认所述数据段已经被写入后，接收存储请求完成的应答。

在一种实施方式中，定序器模块 2308 通过在首先到达的并发请求之间选择存储请求选择用于执行的第一存储请求。在另一种实施方式中，定序器模块 2308 通过选择具有较早的时间戳的存储请求选择用于执行的第一存储请求。在另一种实施方式中，定序器模块 2308 通过使用一些选择标准选择存储请求选择用于执行的第一存储请求。例如，定序器模块 2308 可选择以某种方式被请求客户端 114 标记为高优先级的存储请求、可选择来自优先客户端 114 的存储请求，等等。本领域技术人员会认识到定序器模块 2308 可利用一些选择标准选择第一存储请求的其他方法。

在一种实施方式中，存储请求接收器模块 2302、条带化模块 2304、奇偶校验-镜像模块 2306 和定序器模块 2308 是主控制器（未示出）的部分，该主控制器控制并服务所述并发存储请求。所述主控制器的全部或部分可驻留并运行在下述设备内：客户端 114、第三方 RAID 管理设备、存储设备集 1604 的存储设备 150 或存储设备 150 的存储控制器 152。通过使用主控制器用于为所述数据执行服务请求，定序器模块 2308 可获悉指令所述数据的存储请求并可随后识别并发存储请求，然后还可将所述并发存储请求按下述要求排序：存储在存储设备集的存储设备 150 上的数据不会被损坏。本领域技术人员会认识到控制服务指令所述数据的存储请求的主控制器的其他实现方式。

在另一种实施方式中，所述主控制器是两个或更多个能够服务来自一个或多个客户端 114 的所述并发存储请求的主控制器的群组的一部分，其中，所述存储请求是指令存储在存储设备集 1604 的存储设备 150 上的所述数据。例如，主控制器可为第一客户端 114 服务存储请求，而第二主控制器可为第二客户端 114 服务存储请求。第一和第二客户端 114 都可访问存储在存储设备集 1604 的存储设备 150 上的数据，因此允许并发存储请求。一个主控制器可以是存储设备 150a 的一部分，而其他主控制器可以是第二存储设备 150b 的一部分。在另一种实施方式中，第一主控制器可以是第一存储设备集 1604a 一部分，而第二主控制器可以是镜像存储设备集 1604b 的一部分。

在主控制器是访问存储设备集 1604 的存储设备 150 的主控制器群组的一部分的情况下，装置 2300 可包括主验证模块 2310，该主验证模块 2310 在执行接收到的存储请求之前确认服务所述接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的所述存储请求的执行。在这种实施

方式中，其他主控制器接收所述并发存储请求，并且服务请求至少有一部分数据与其他主控制器接收到的并发存储请求相同。

例如，主控制器可接收存储请求，然后，主验证模块 2310 可在所述存储请求执行前轮询其他主控制器以验证所述主控制器仍然是用于所述存储请求的数据的主控制器。验证的一部分包括验证所述主控制器之间能够通信，从而指定的主控制器在所述存储请求执行之前被验证。这种方法可在前端 RAID 控制器被指定为主控制器而另一个前端 RAID 控制器被指定为备用控制器的情况下是有利的。在另一个实例中，主控制器可接收从文件或对象读取数据段的存储请求，然后主验证模块 2310 可轮询其他主控制器，从而验证没有进行中的文件或对象的更新。在另一个实例中，主控制器可使用主验证模块以获取控制用于所述存储请求的数据。

一种验证所述主控制器仍然是用于执行所述存储请求的主控制器的方法是：使用三路轮询方案，其中两个设备/控制器必须能够用于投票选出用于存储请求的主控制器，以便于继续轮询。这种方案使用对竞争成为主控制器的控制来说是第三方的设备（未示出），并且保留哪一个控制器被分配为主控制器的记录。这个主验证设备可以是另一个控制器、服务器上的客户端 114 等等，并能够与群组中可充当主控制器的控制器通信。然后，主验证模块 2310 的一部分可驻留在所述主验证设备内，而主验证模块 2310 的一部分位于每个控制器内。

在一个实例中，系统 1600 包括第一前端分布式 RAID 控制器（“第一控制器”）、第二前端分布式 RAID 控制器（“第二控制器”），其中的每一个控制器都可以是主控制器和分享的主验证设备。第一和第二控制器和主验证设备之间都可相互通信。主验证模块 2310 可将第一控制器指定为主控制器而把第二控制器指定为用于存储在存储设备集 1604 的存储设备 150 上的数据的备用控制器，并且主验证模块 2310 可将主控制器的信息存储在控制器和主验证设备上。只要保持第一控制器、第二控制器和主验证设备之间的通信，主验证模块 2310 就能够确认第一控制器为主控制器。

如果第一（主）控制器接收存储请求，第二（备用）控制器变得不可用或与第一控制器和主验证设备的通信丢失，主验证模块 2310 能够通过主验证设备和第一（主）控制器之间的通信验证所述第一控制器仍然是主控制器，

并且由于第一控制器和主验证设备都确认所述第控制器确实是主控制器，主验证模块 2310 可允许存储请求继续进行。由第二（备份）控制器接收的存储请求不会继续进行，这是由于第二控制器通过主验证模块 2310 识别到其不是主控制器。

另一方面，如果第一（主）控制器不可用或不能与第二（备份）控制器和主验证设备通信，并且第二（备份）控制器接收存储请求，主验证模块 2310 能够识别到第二控制器和主验证模块不能与第一控制器通信，并且主验证模块 2310 能够指定第二（备份）控制器为主控制器，存储请求也得以继续进行。然后，主控制器指定的改变被记录在第二控制器上。

如果第一控制器是操作性的，并与第二控制器和所述主验证设备完全断开通信，用于第一控制器接收的数据的任何存储请求将不会被执行。如果通信恢复，第一控制仍然不会执行存储请求，这是由于所述第二控制器和所述主验证模块都将第二控制器识别为主控制器。当然，这种主控制器指定可以被重置。本领域技术人员会认识到分配和重新分配主控制器指定给主控制器中的一个。

如果主验证设备不可用且第一存储控制器接收存储请求，主验证模块 2310 运行在第一和第二控制器上的部分能够验证所述第一控制器是主控制器，并且存储请求可继续进行。如果所述第二控制器接收存储请求，主验证模块 2310 运行在第一和第二控制器上的部分能够验证所述第一控制器是主控制器，并且存储请求不会再继续进行。在其他实施方式中，不止两个的控制器是轮询方案的一部分。本领域技术人员会认识到主验证模块 2310 能够在执行存储请求之前验证控制器是主控制器的其他方法。

在另一种实施方式中，装置 2300 包括主确定模块 2312。在发送存储请求之前，主确定模块 2312 将主确定请求发送给主控制器群组。然后，主控制器群组识别哪一个控制器被指定为用于存储请求的主控制器，并将标识主控制器的响应发回主确定模块 2312。主确定模块 2312 为所述存储请求接收主控制器的标识符并指令所述请求设备将存储请求发送给指定的主控制器。在一种实施方式中，主确定模块 2312 位于并运行在客户端 114 内。在另一种实施方式中，主确定模块 2312 位于第三方 RAID 管理设备内并在其内执行请求。在另一种实施方式中，主确定模块 2312 位于存储设备 150 内。在另一种实施方

式中，主确定模块 2312 分布在两个或更多个存储设备 150 之间。

在又一种实施方式中，装置 2300 包括返回错误指示的主错误模块 2314。在一种实施方式中，如果由主控制器控制的多存储请求接收器模块 2302 接收到不由主控制器控制的存储请求，主错误模块 2314 返回错误指示。

在另一种实施方式中，如果主确定模块 2312 或主验证模块 2310 在所述存储请求执行完成时确定主控制器不再是确定的主控制器，主错误模块 2314 返回错误指示。这种实施方式通常发生在当主控制器开始执行存储请求并与群组中的其他主控制器的通信丢失时，或者发生在轮询方案中的与其他主控制器和主验证设备的通信丢失时。在另一种实施方式中，如果由主控制器控制的多存储请求接收器模块 2302 接收不由所述主控制器控制的存储请求，主错误模块 2314 返回错误指示。

在另一种实施方式中，主控制器控制传送给一个或多个次级主控制器的存储请求。每个所述次级主控制器控制用于存储在存储设备集 1604 的存储设备 150、1602 上的数据的存储请求。在另一种实施方式中，控制所述次级主控制器的主控制器也是用于指令存储在存储设备集 1604 的存储设备 150、1602 上的数据的存储请求的次级主控制器。

在另一种实施方式中，主控制器控制传送给一个或多个次级主控制器的存储请求，并且每个所述次级主控制器控制用于存储在对所述次级主控制器来说唯一的存储设备集的存储设备 150 上的数据的存储请求。装置 2300 是灵活的，从而任何主控制器都能够成为相对于其他作为次级主控制器的控制器的主控制器。一些次级主控制器能够存储设备集 1604，而其他次级主控制器能够控制不同的存储设备集。在另一种实施方式中，主控制器可以是奇偶校验-镜像存储设备 1602 或 N 个存储设备 150a-n 中的一个。

在另一种实施方式中，当所述主控制器离线或不能确定其是指定的主控制器时，次级主控制器可以成为主控制器。本领域技术人员会认识到用于在一个或多个次级主控制器之间分配或重新分配主控制器指定的各种静态和动态的方法。

在一种优选实施方式中，装置 2300 包括奇偶校验生成模块 2316，该奇偶校验生成模块 2316 为所述条带计算奇偶校验数据段并将所述奇偶校验数据段存储在奇偶校验-镜像存储设备 1602 上。由奇偶校验-镜像存储设备 1602 上的

N 个数据段的集计算奇偶校验条带。这种实施方式通常通过 RAID5、RAID6 或其他 RAID 级别（但通常不包括 RAID0、RAID1、RAID10 等等）实现。

在另一种优选实施方式中，装置 2300 包括奇偶校验更替模块 2318，该奇偶校验更替模块 2318 为每个条带变更被分配为一个或多个用于所述条带的奇偶校验-镜像存储设备 1602 的存储设备集 1604 中的存储设备 150。轮换每个条带的奇偶校验数据段提升了性能。奇偶校验更替模块 2318 可与条带化模块 2304 一起使用，以计算一个或多个存储设备 150a-n 之间的一致性，作为文件或对象的一部分的一个数据段从一个或多个存储设备 150a-n 中读取、写入或更新。

不同的模块 2302-23018 的功能可一起在单个主控制器中实现，或者可分布在下述设备之间：一个或多个客户端 114、第三方 RAID 管理设备和一个或多个存储设备 150、1602。本领域技术人员会认识到此处描述的功能是分布式的不同的实施方式。

图 14 是示意性流程图，示出了根据本发明的用于共享的、前端分布式 RAID 的方法 2400 的一种实施方式。方法 2400 始于步骤 2402，多存储请求接收器模块 2302 接收来自至少两个客户端 114 的至少两个存储请求（步骤 2404），以读取数据或将数据存储在存储设备集 1604 中的存储设备 150。所述数据来自文件，或者是对象的数据，并且每个所述存储请求具有至少一部分共有的数据，并且，所述存储请求是并发存储请求，这是由于所述存储请求的到达使得在一个存储请求到达之前，两个存储请求中的另一个存储请求还没有完成。条带化模块 2304 为所述数据计算条带模式（步骤 2406），其中，所述条带模式包括一个或多个条带并且每个条带包括 N 个数据段的集。条带化模块 2304 还读取条带的 N 个数据段，或将条带的 N 个数据段写入存储设备集 1604 中的 N 个存储设备 150a-n（步骤 2408），其中，N 个数据段中的每一个都被写入或读取自独立的存储设备 150。

当所述存储请求是写入操作时，奇偶校验-镜像模块 2306 将所述条带的 N 个数据段的集写入存储设备集 1604 中的一个或多个奇偶校验-镜像存储设备 1602（步骤 2306），其中，奇偶校验-镜像存储设备 1602 是除 N 个存储设备 150a-n 之外的设备。奇偶校验-镜像模块 2306 还读取存储在奇偶校验-镜像存储设备 1602 中的数据段或奇偶校验数据段（2410）。定序器模块 2308 确保来

自第一客户端 114 的第一存储请求完成后才执行来自第二客户端 114 的存储请求。方法 2400 结束于步骤 2416。第一和第二存储请求是并发存储请求。

本发明可采用其他指定形式实施而不脱离本发明的宗旨或本质特点。描述的实施方式在各个方面被视为仅仅是示例性而不是限制性的。因此，本发明的范围由附属的权利要求确定，而不是由上述说明书确定。在本发明的权利要求的含义和等价范围内的所有改变被包含在本发明的保护范围内。

作为大容量、非易失性存储的高速缓存的固态存储器

通常，高速缓存是有利的，因为经常存取的或者作为应用程序或操作系统的一部分载入的数据可存储在高速缓存中，相比于必须通过大容量、非易失性（“HCNV”）存储设备访问数据的情况，后续的存取操作更快速，所述大容量、非易失性存储设备例如硬盘驱动器（“HDD”）、光盘驱动器、磁带存储器等。高速缓存通常包括在计算机内。

某些存储设备和系统在HCNV存储设备中包括高速缓存。某些存储设备包含非易失性固态高速缓存；这些提供了减少访问时间的好处，但是仅仅可提供与HCNV存储设备接口的通常受限能力一致的性能。存在通常位于主板上的某些非易失性固态高速缓存存储设备；这些设备不能用于多客户端环境中，因为没有提供高速缓存一致性。某些HCNV设备的控制器也包括高速缓存。在多个客户端共享冗余HCNV高速缓存控制器的情况下，需要复杂的高速缓存一致性算法来确保不破坏数据。

通常，在DRMA中实现高速缓存，得到额外的高速缓存能力，并且需要相对高的性能功率比。如果支持易失性高速缓存的功率失去，高速缓存中存储的数据丢失。通常，某些后备电池用于避免供电故障情况下的数据丢失，在后备电池故障之前，有足够的能力将高速缓存清洗到非易失性存储器。另外，后备电池系统消耗功率，需要冗余，消极地影响可靠性并且占据空间。电池也必须基于规则来服务并且后备电池相对昂贵。

如上所述，显而易见，存在使用作为高速缓存的固态存储器管理数据的装置、系统和方法的需求。有利地是，这种装置、系统和方法提供了消耗很少功率、提供显著更大的能力并且不需要后备电池来保持高速缓存中存储的数据的非易失性高速缓存。

图15是示出了根据本发明的具有固态存储器110的系统3400的一种实施

方式的示意性框图，该固态存储器110作为大容量、非易失性存储设备的高速缓存。系统3400包括固态存储设备102，该固态存储设备102具有包括固态存储控制器104和HCLV控制器3402的存储控制器152、固态存储器110、和网络接口156。系统3400包括通过计算机网络116连接到固态存储设备102的请求设备155和一个或多个HCNV存储设备3404a-n。本领域技术人员将会认识到图15中描述的系统3400仅仅是一种实施方式，并且允许固态存储器110作为存储设备的高速缓存的许多其他配置是可能的。

系统3400包括具有网络接口156和存储控制器152的固态存储设备102。在另一种实施方式中，网络接口156在固态存储设备102外部。例如，网络接口156可以在可包括或可不包括固态存储设备102的服务器112中。

在所述实施方式中，固态存储设备102包括存储控制器152，该存储控制器152包括固态存储控制器104和大容量、非易失性（“HCNV”）存储控制器3402。在另一种实施方式中，固态存储设备102包括不在存储控制器152中的固态存储控制器104和HCNV存储控制器3402。在其他实施方式中，固态存储设备102包括固态存储控制器104，该固态存储控制器104包括HCNV存储控制器3402，或反之亦然。

在所述实施方式中，系统3400包括具有集成的固态存储器110和外部HCNV存储设备3404a-n的固态存储设备102。在另一种实施方式中，存储控制器152、104、3402可与固态存储器110分离。在另一种实施方式中，控制器152、104、3402和固态存储器110包括在HCNV存储设备3404中。HCNV存储设备3404还可包括网络接口156。本领域技术人员将会认识到，许多其他配置是可能的。固态存储设备102、固态存储控制器104、固态存储器110、存储I/O总线210、网络接口156、计算机网络116、和请求设备155大体上类似于上述的设备和总线的其他实施方式。

在一种实施方式中，请求设备155通过系统总线连接到固态存储设备102、存储控制器152、固态存储控制器104等。请求设备155和固态存储器110之间的数据传送可在系统总线上发生。

HCNV存储设备3404通常是提供非易失性存储的大容量存储设备，并且相比于固态存储器110来说通常写入和读取数据更慢。HCNV存储设备3404相比于固态存储器110，每单位存储容量还可以更廉价。HCNV存储设备3404可以

是硬盘驱动器（“HDD”）、光盘驱动器、磁带驱动器、和类似的驱动器。提供固态存储器110以作为用于HCNV存储设备3404的高速缓存通常增加了数据的访问速度和存储速度。本领域技术人员将会认识到作为HCNV存储设备3404的高速缓存的固态存储器110的其他好处。

在一种实施方式中，HCNV存储设备3404通过存储区域网络（“SAN”）连接到存储控制器152。在一种实施方式中，分离的SAN控制器将HCNV存储设备3404连接到存储控制器152。在另一种实施方式中，HCNV存储控制器3403或存储控制器152充当SAN控制器。本领域技术人员将会认识到HCNV存储设备3404可在SAN中连接的其他方式。

图16是示出了根据本发明的具有固态存储器的装置3500的一种实施方式的示意性框图，该固态存储器作为大容量、非易失性存储设备的高速缓存。装置3500包括高速缓存前端模块3502、高速缓存后端模块3504、对象存储控制器3506、HCNV模块3508、和标准设备模拟模块3510，这些模块在下面描述。装置3500的模块3502-3510在具有固态存储控制器104和HCNV存储控制器3402的存储控制器152中描述，但是每个模块3502-3510的某些或全部可包括在固态存储控制器104、HCNV存储控制器3402、服务器112、HCNV存储设备3404或其他位置中。

装置3500包括管理关联于存储请求的数据传送的高速缓存前端模块3502，其中数据传送在请求设备155和作为一个或多个HCNV存储设备3404a-n的高速缓存的固态存储器110之间进行。装置3500还包括管理固态存储器110和HCNV存储设备3404a-n之间的数据传送的高速缓存后端模块3504。数据传送可包括数据、元数据和/或元数据索引。如上所述，固态存储器110是通常布置在内存库214中的非易失性、固态数据存储元件216、218、220的阵列。在不同的实施方式中，固态存储器110可以是闪存、nano随机访问存储器（“nano RAM”或“NRAM”）、磁电阻RAM（“MRAM”）、动态RAM（“DRAM”）、相变RAM（“PRAM”），或类似RAM。

通常，高速缓存前端模块3502、高速缓存后端模块3504、和固态存储控制器104独立于请求设备运行。例如，请求设备可将具有固态存储控制器104和HCNV存储控制器3404的存储控制器152，连同关联的存储器110、3404a-n视为单个存储设备。在另一个实例中，请求设备155可将HCNV存储设备

3404a-n和固态存储器110视为透明的。

在一种实施方式中，固态存储控制器104包括对象存储控制器模块3506，该对象存储控制器模块3506应答来自一个或多个请求设备155的对象请求并在固态存储器110中管理所述对象请求的对象。在这种实施方式中，固态控制器104连同对象存储控制模块3506一起，如上所述（尤其是如上相对于图2A中描述的装置200所述）地管理对象请求。

在一种实施方式中，装置3500包括HCNV RAID模块3508，该HCNV RAID模块3508将固态存储器110中缓存的数据存储在与RAID级别一致的独立驱动器冗余阵列（“RAID”）中的两个或更多HCNV存储设备3404a-n中。在这种实施方式中，数据对请求设备155作为整体呈现，使得RAID过程对请求设备155隐藏。例如，高速缓存前端模块3502可将来自请求设备155的数据缓存到固态存储器110中，并且高速缓存后端模块3504可与HCNV RAID模块3508合作来条带化数据并且将数据段和奇偶校验数据段存储在与RAID级别一致的HCNV存储设备3404a-n中。本领域技术人员将会认识到可在HCNV存储设备3404a-n中将来自请求设备155的数据RAID的其他方式。

在另一种实施方式中，固态存储器110和HCNV存储设备包括配置为RAID群组的混合存储设备集内的混合存储设备。例如，混合存储设备集可以是前端分布式RAID存储设备集1604，并且混合存储设备可以是如上所述分别相对于图10、11、和12中描述的系统1600、装置2100、和方法2200的存储设备集中的存储设备150、1602。在这种实施方式中，固态存储器110中缓存的和随后存储在HCNV设备3404上的数据段是条带的N个数据段之一或所述条带的奇偶校验数据段。正如在前端RAID中一样，混合存储设备接收独立于RAID条带的数据段的、来自一个或多个客户端114的存储请求。

在另一种实施方式中，混合存储设备是从两个或更多客户端114接收两个或更多同时的存储请求的共享前端分布式RAID群组1604的存储设备150、1602，如上相对于共享前端RAID所述，如上分别相对于图10、13、和14中描述的系统1600、装置2300、和方法2400所述。有利地是，该实施方式确保共享、冗余高速缓存保持一致性，而不需要另外的、复杂的一致性算法和协议。

在另一种实施方式中，固态存储器110和HCNV存储设备3404a-n包括混合存储设备，并且装置3500包括标准设备模拟模块3510，该标准设备模拟模块

3510在请求设备155加载混合存储设备的操作专用代码之前，通过模拟附属于一个或多个请求设备155的标准设备提供对混合存储设备的访问。在这种实施方式中，标准设备由工业标准的BIOS支持。该自引导操作允许具有有限功能的请求设备155识别和访问混合设备，直到固态存储控制器104、HCNV存储控制器3404、以及可能的装置3500的其他模块3502-3510的专用的驱动可加载到请求设备155上。

在一种实施方式中，固态存储设备110可分为两个或更多区域，其中一个或多个分区被作为独立于作为HCNV存储设备3404的高速缓存的固态存储器的固态存储器。例如，客户端114可访问固态存储器110的某些分区以用于一般数据存储，而一个或多个分区作为HCNV存储设备3404的高速缓存。

在一种实施方式中，不止一个的客户端114（或请求设备155）能够向高速缓存前端模块3502和高速缓存后端模块3504发送高速缓存控制消息，以管理固态存储设备110和一个或多个HCNV存储设备3404内存储的一个或多个文件或对象的状态。有利地是，客户端114/请求设备155基于每文件、每对象、或每数据段管理高速缓存的能力为共享固态存储设备110提供了高度的灵活性。

大量的高速缓存控制消息是允许的和可能的。例如，高速缓存控制消息可包括使得高速缓存后端模块3504扣牢固态存储器110中的对象或文件的一部分的控制消息。另一种高速缓存控制消息可包括使得高速缓存后端模块3504释放固态存储器110中的对象或文件的一部分的控制消息。另一种高速缓存控制消息可包括使得高速缓存后端模块3404将来自固态存储器110的对象或文件的一部分清洗到一个或多个HCNV存储设备3404的控制消息。另一种高速缓存控制消息可包括使得高速缓存后端模块3404从一个或多个HCNV存储设备3404向固态存储器110预加载对象或文件的一部分的控制消息。另一种高速缓存控制消息可包括使得高速缓存后端模块3504将来自固态存储器的一个或多个对象或文件的一部分或多个部分卸载到HCNV存储设备3404，以便释放固态存储器110中的预定量存储空间的控制消息。本领域技术人员将会认识到其他可能的高速缓存控制消息。

在一种实施方式中，高速缓存控制消息通过对象或文件的元数据（“高速缓存控制元数据”）传送。在一种实施方式中，高速缓存控制元数据是持

久化的。在另一种实施方式中，高速缓存控制元数据在创建文件或对象时通过属性集来建立。在这种实施方式中，属性可通过与特定对象类、专用文件类型的默认特点等的关系继承。在另一种实施方式中，高速缓存控制元数据从文件或对象管理系统获得。本领域技术人员将会认识到通过元数据传送高速缓存控制消息的其他方式。

在一种实施方式中，系统3400包括非易失性高速缓存存储元件。例如，除了固态存储器110，系统3400还可包括易失性的某种类型的随机访问存储器（“RAM”）。在该实施方式中，高速缓存前端模块3502和高速缓存后端模块3504在易失性高速缓存存储元件中存储某些数据并且管理固态存储器110和易失性高速缓存存储元件中存储的数据，并且后端存储模块3504还管理易失性高速缓存存储元件、固态存储器和HCNV存储设备之间的数据传送。例如，不重要的或者可容易地从另一个源恢复的数据可存储在易失性高速缓存中，而其他数据可存储在作为高速缓存的固态存储器110中。

在另一种实施方式中，用于HCNV存储设备3404中存储的对象和文件的元数据和/或索引元数据保持在固态存储设备110内和易失性高速缓存存储元件中。如上相对于图2A中描述的装置200所述，某些元数据可存储在易失性高速缓存存储元件中，并且如果易失性高速缓存存储元件中的数据丢失，所述元数据可用于重建索引。在一种实施方式中，元数据和索引元数据存储在固态存储器110中，其中不包括易失性高速缓存存储元件。本领域技术人员将会认识到使用易失性高速缓存存储元件连同作为高速缓存的固态存储器110的其他好处和方式。

图17示出了根据本发明的具有固态存储器的方法3600的一种实施方式的示意性流程图，该固态存储器作为大容量、非易失性存储设备的高速缓存。方法3600开始于步骤3602，并且高速缓存前端模块3502管理与存储请求关联的数据传送（步骤3604），其中数据传送在请求设备155和作为一个或多个HCNV存储设备3404a-n的高速缓存的固态存储器110之间进行。高速缓存后端模块3504管理固态存储器110和一个或多个HCNV存储设备110之间的数据传送（步骤3606），方法3600结束于3608。方法3600的运行大体上类似于上文中相对于图10的装置3500所述。

本发明可采用其他指定形式实施而不脱离本发明的宗旨或本质特点。描

述的实施方式在各个方面被视为仅仅是示例性而不是限制性的。因此，本发明的范围由附属的权利要求确定，而不是由上述说明书确定。在本发明的权利要求的含义和等价范围内的所有改变被包含在本发明的保护范围内。

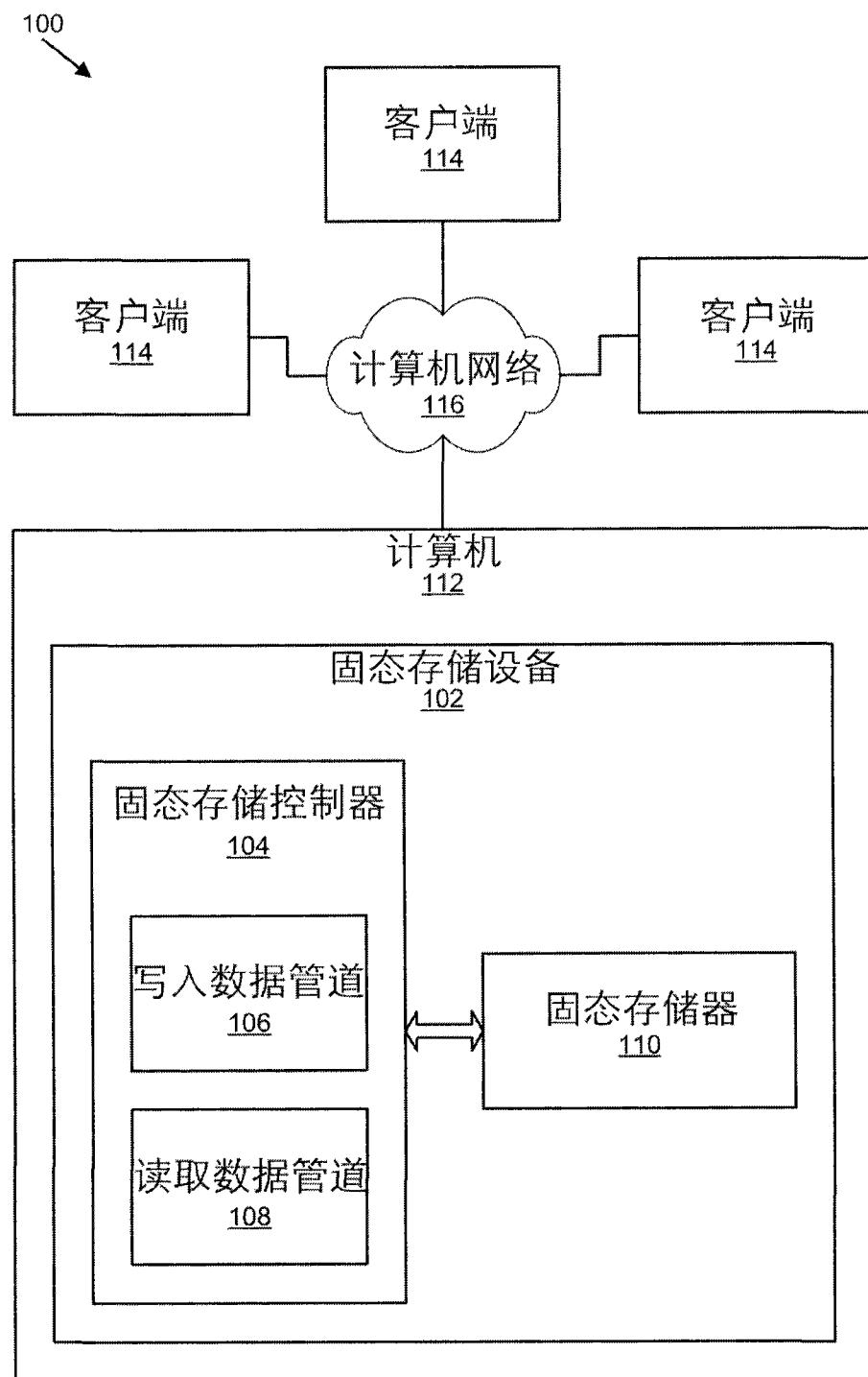


图 1A

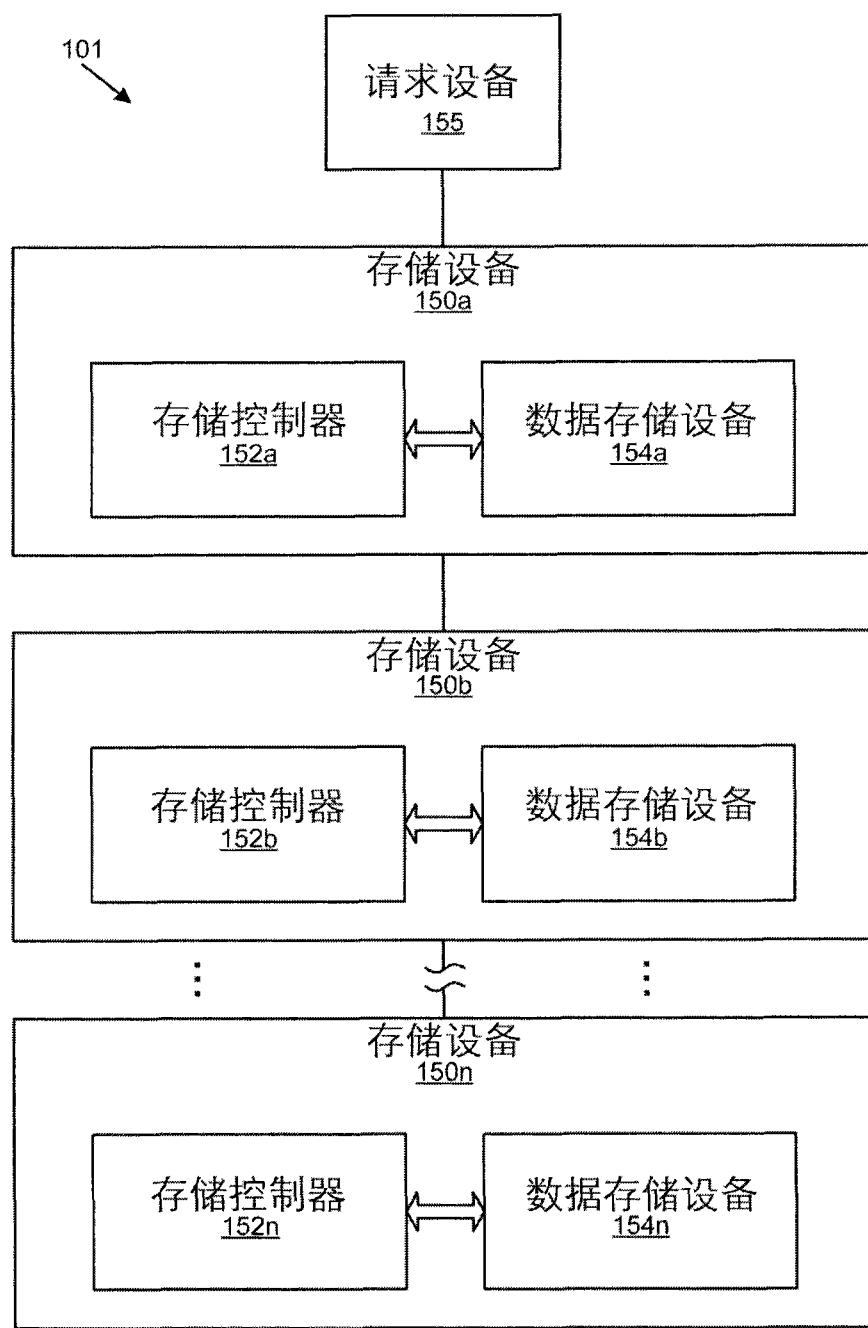


图 1B

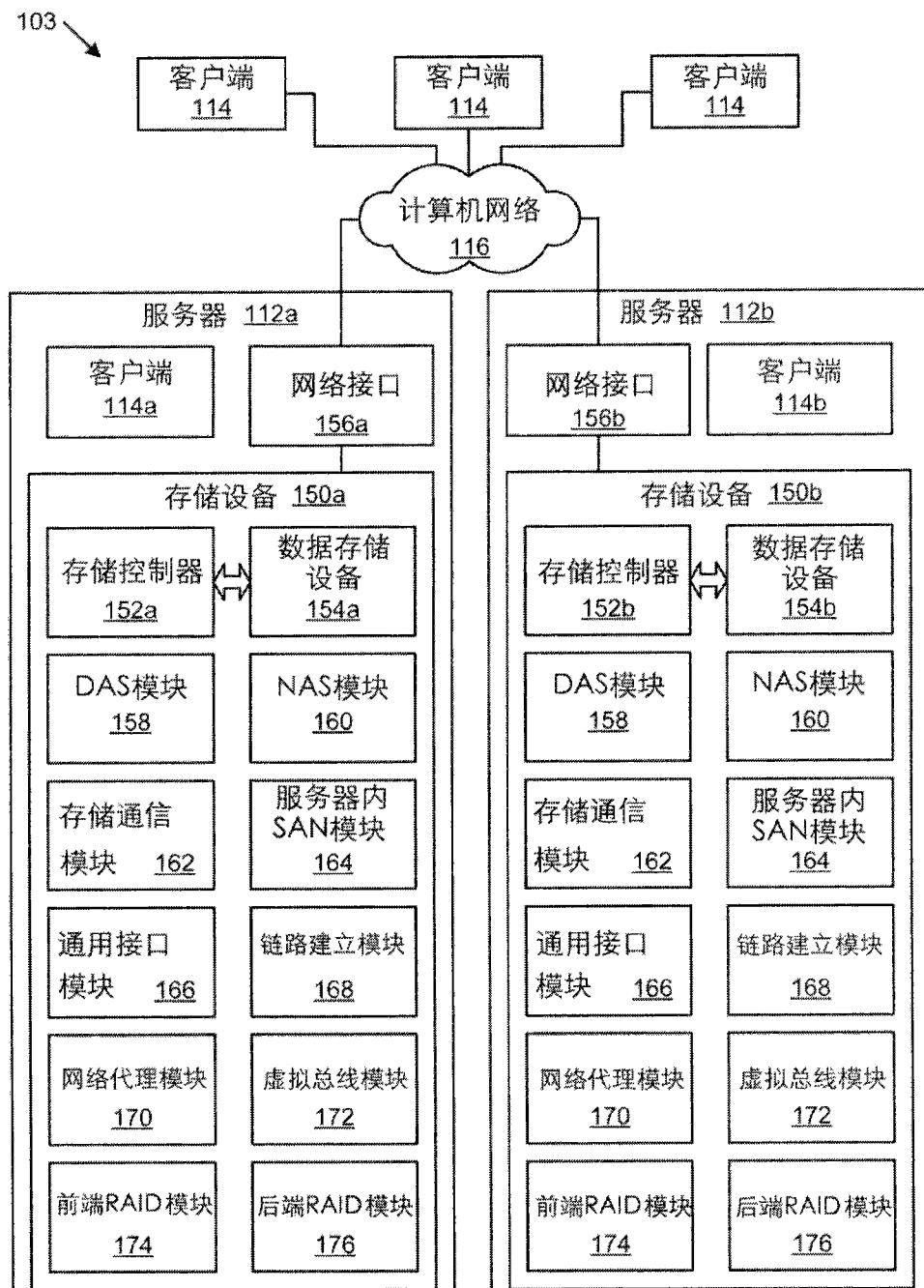


图 1C

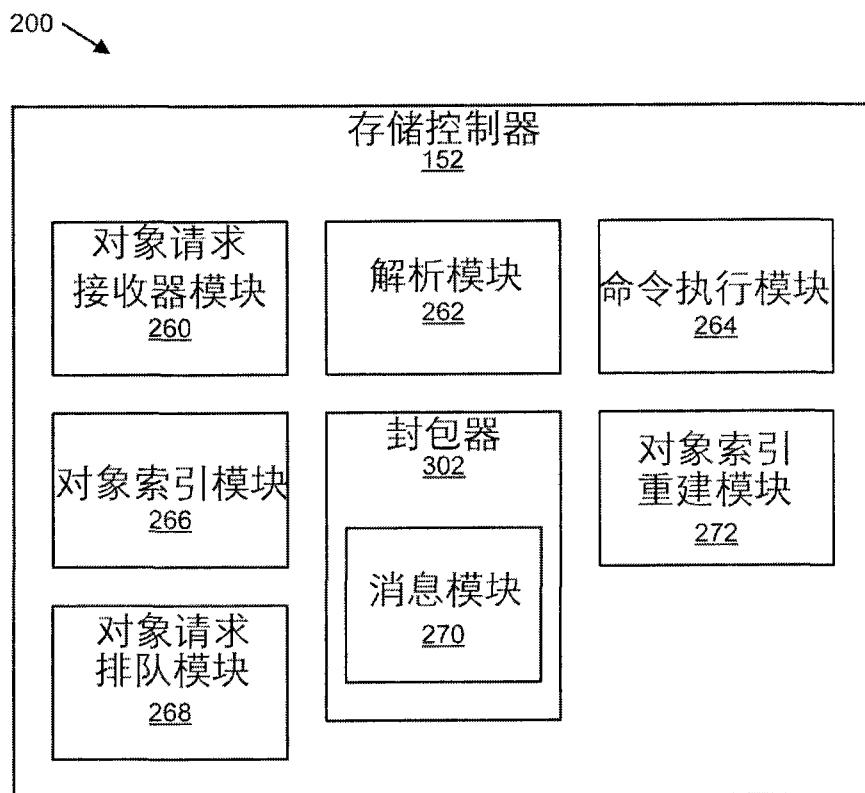


图 2A

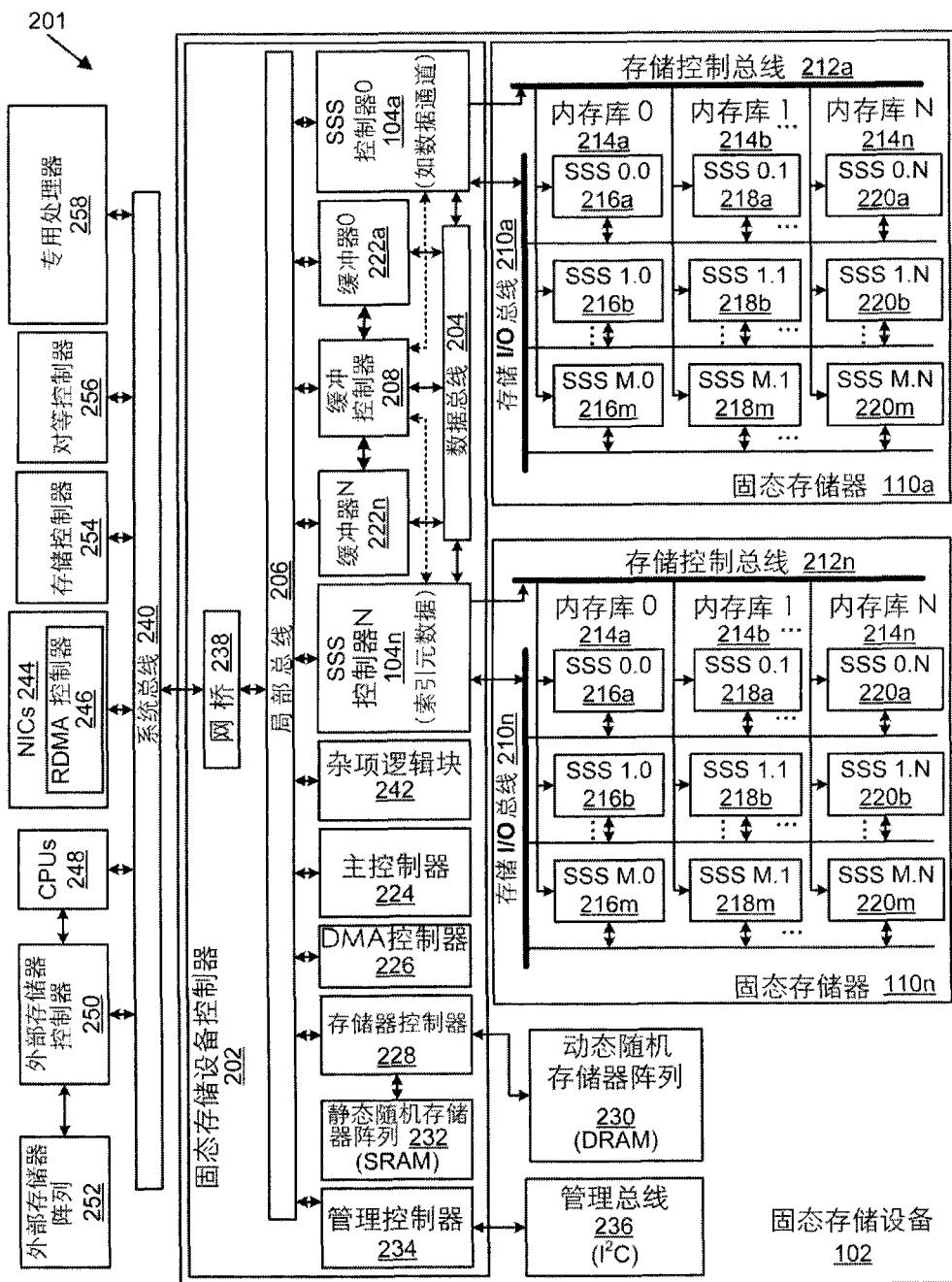


图 2B

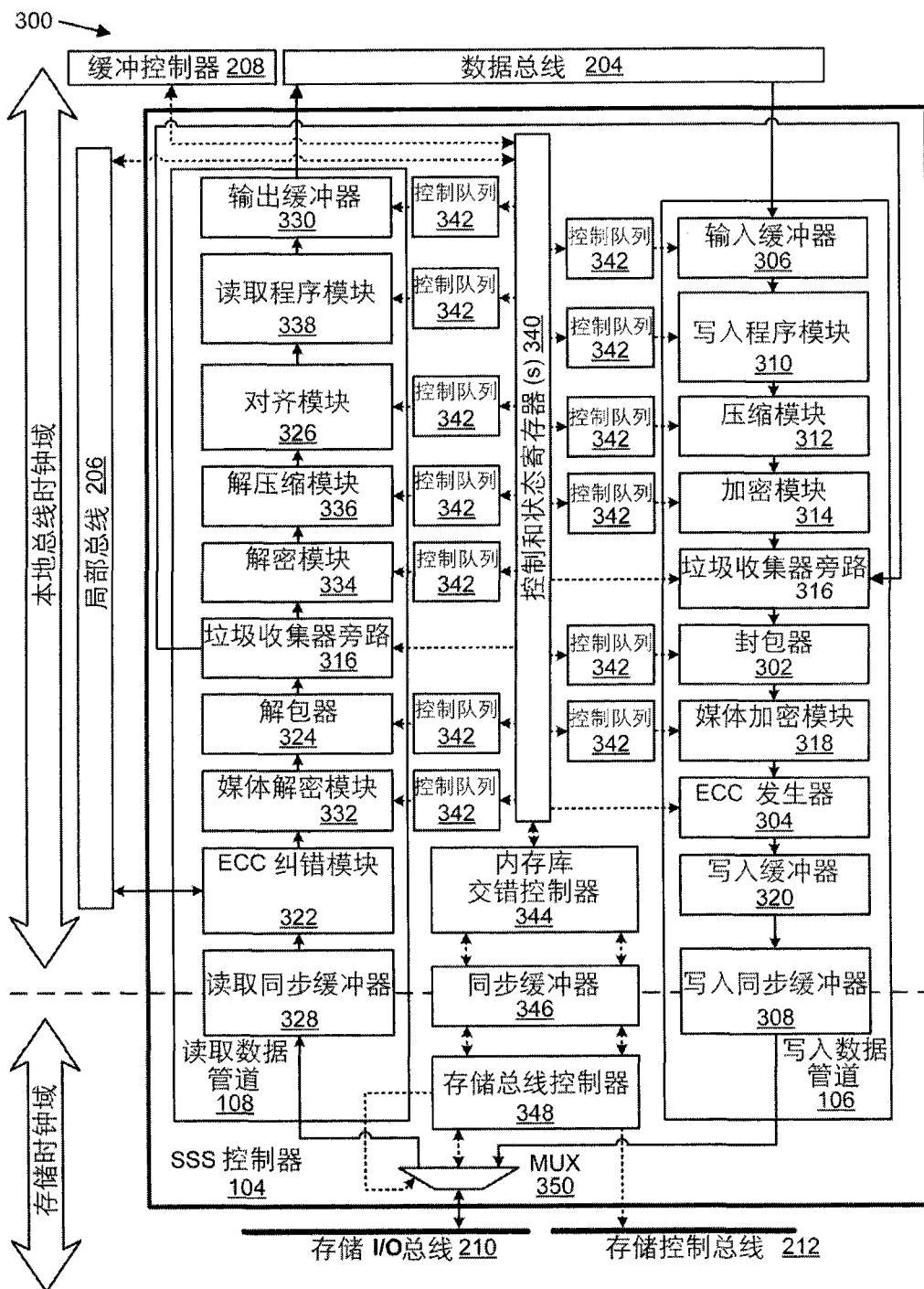


图 3

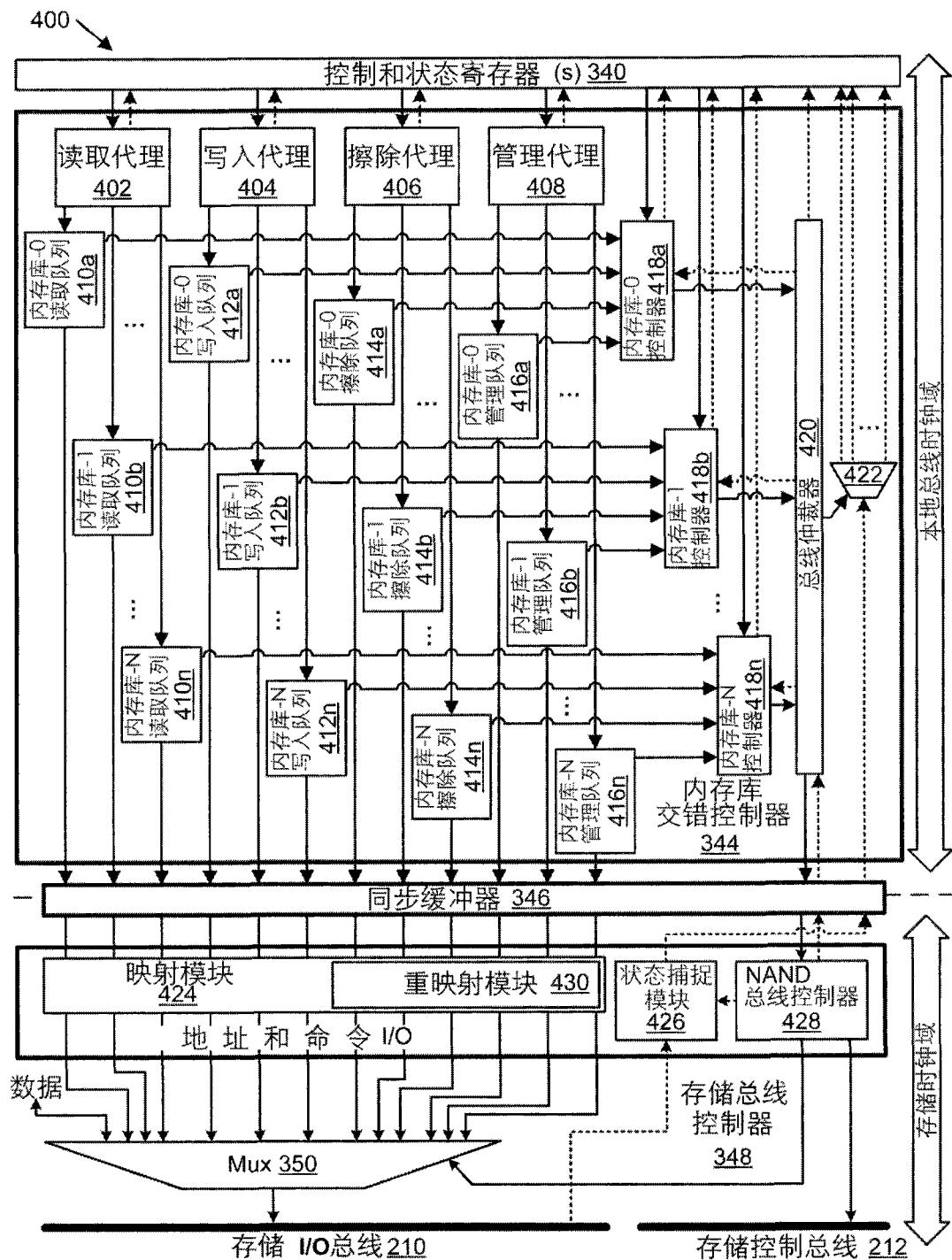


图 4A

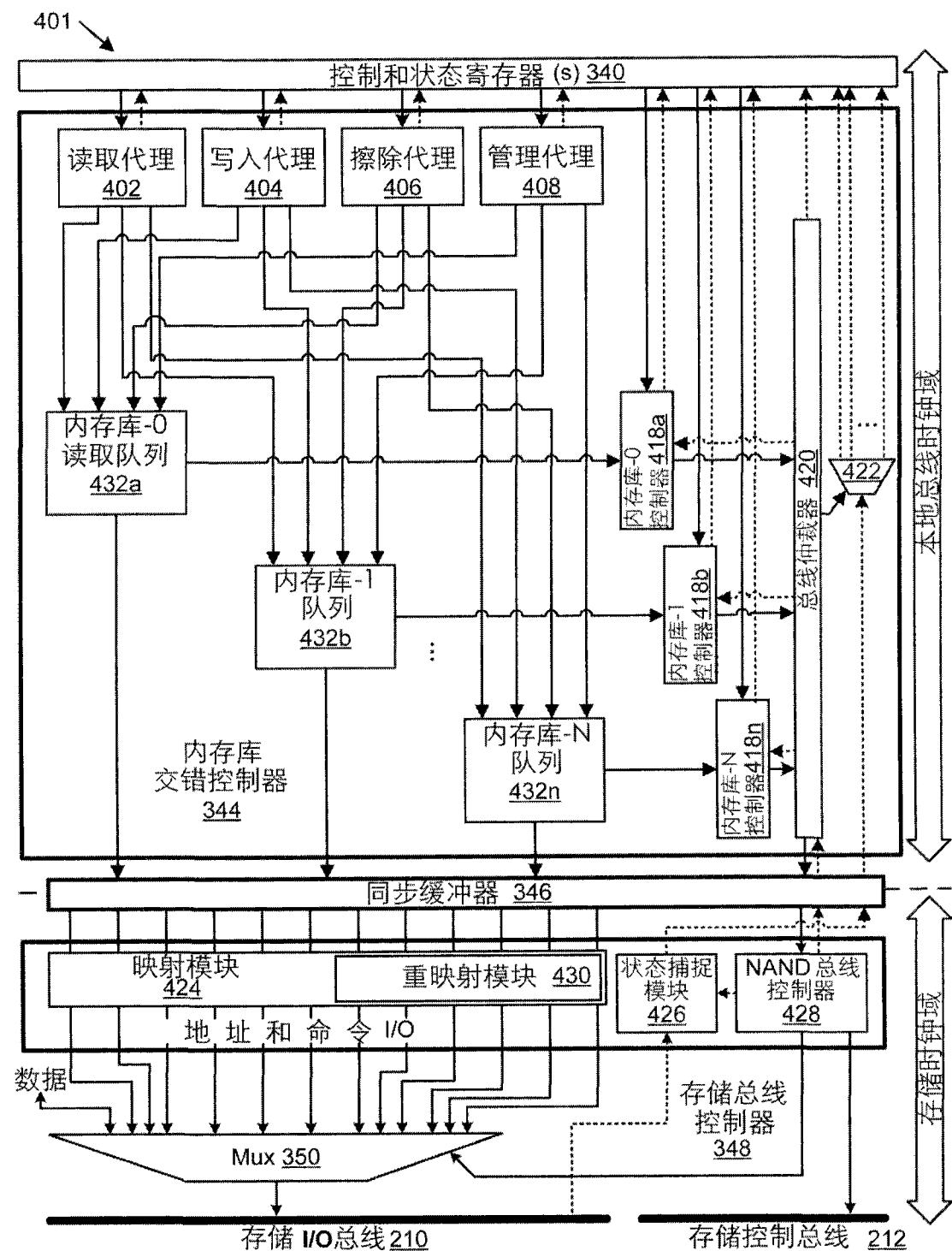


图 4B

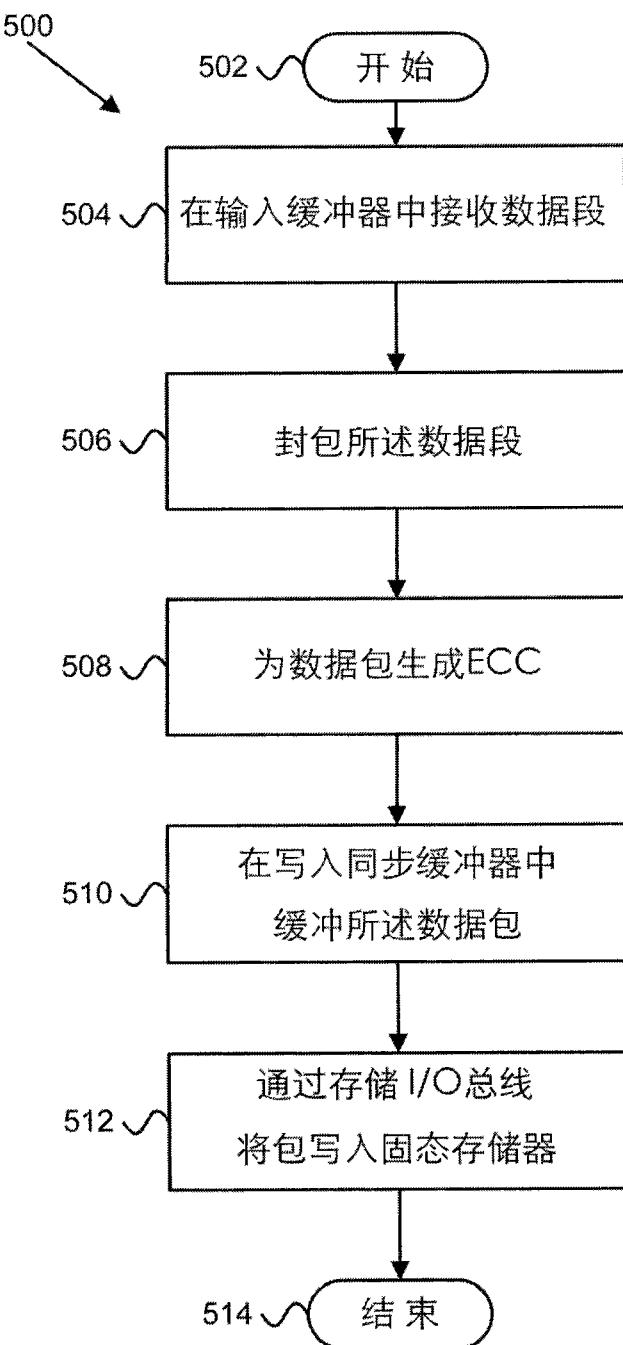


图 5A

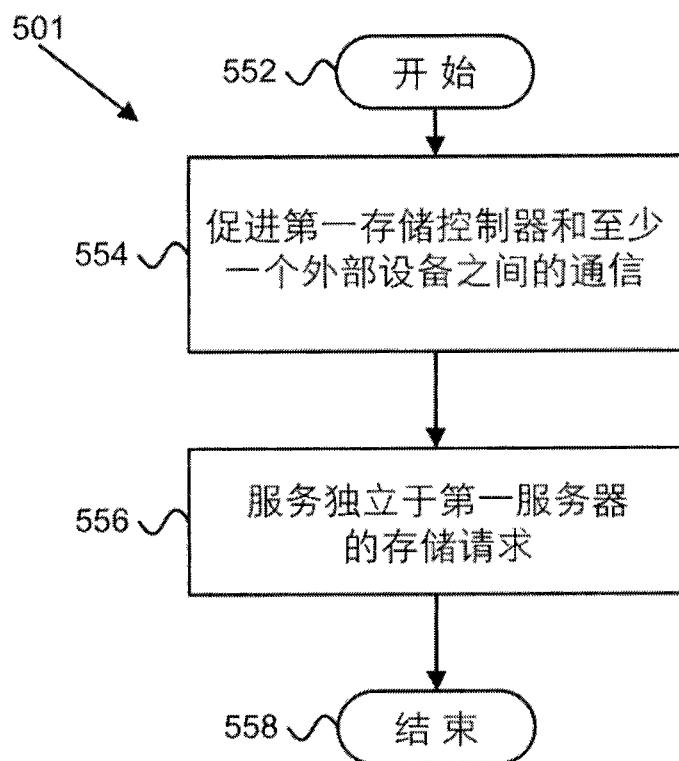


图 5B

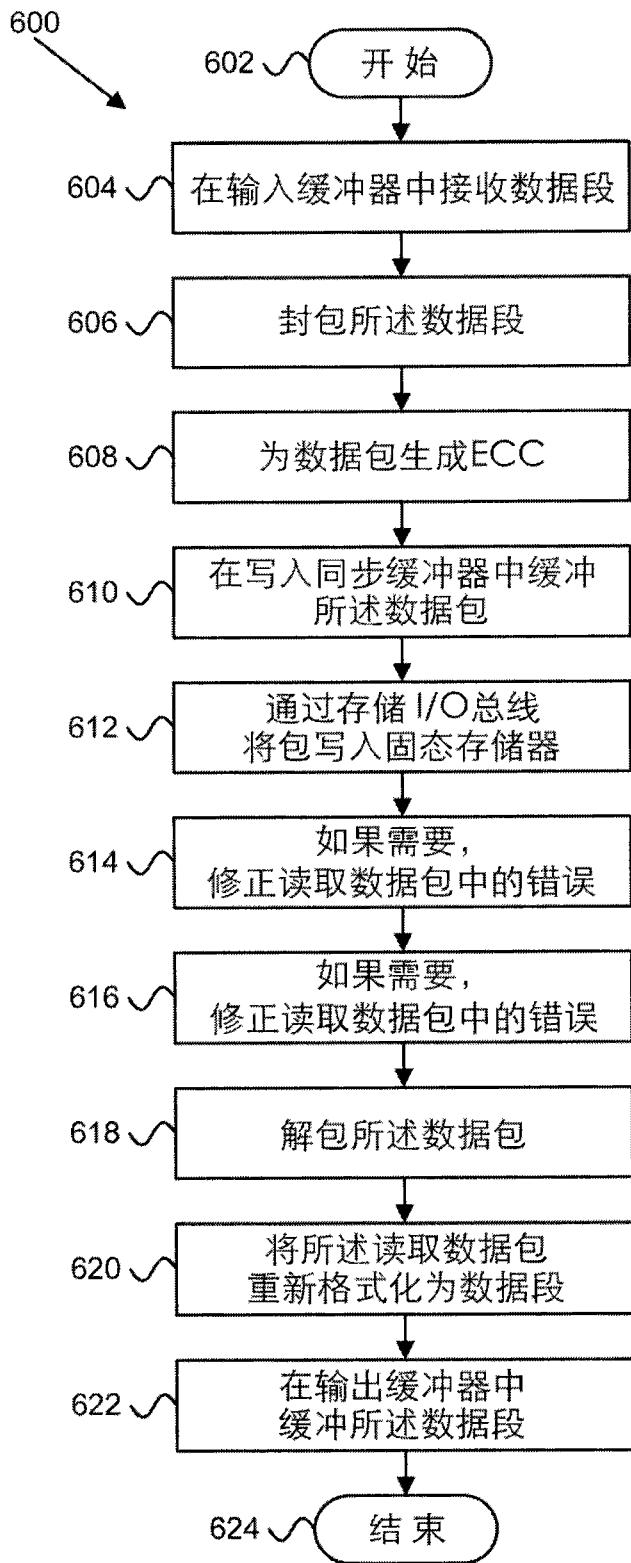


图 6

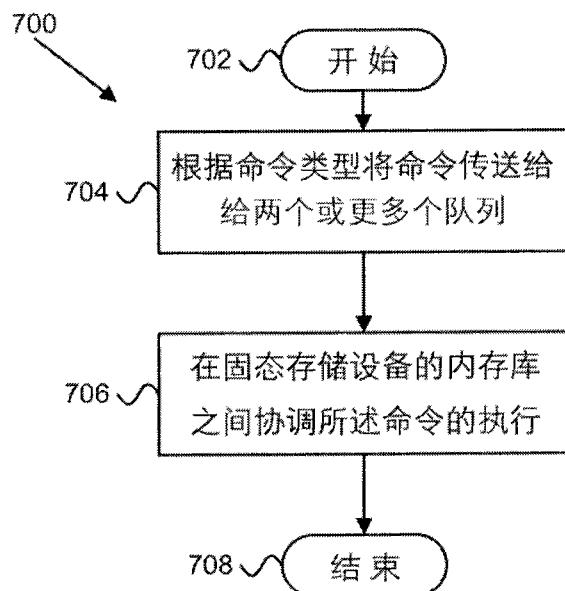


图 7

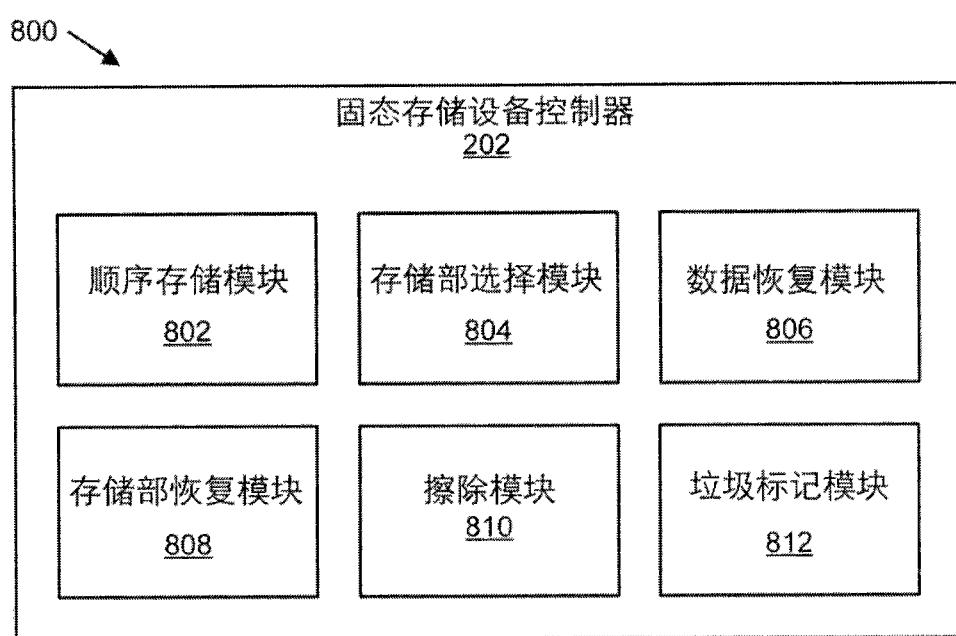


图 8

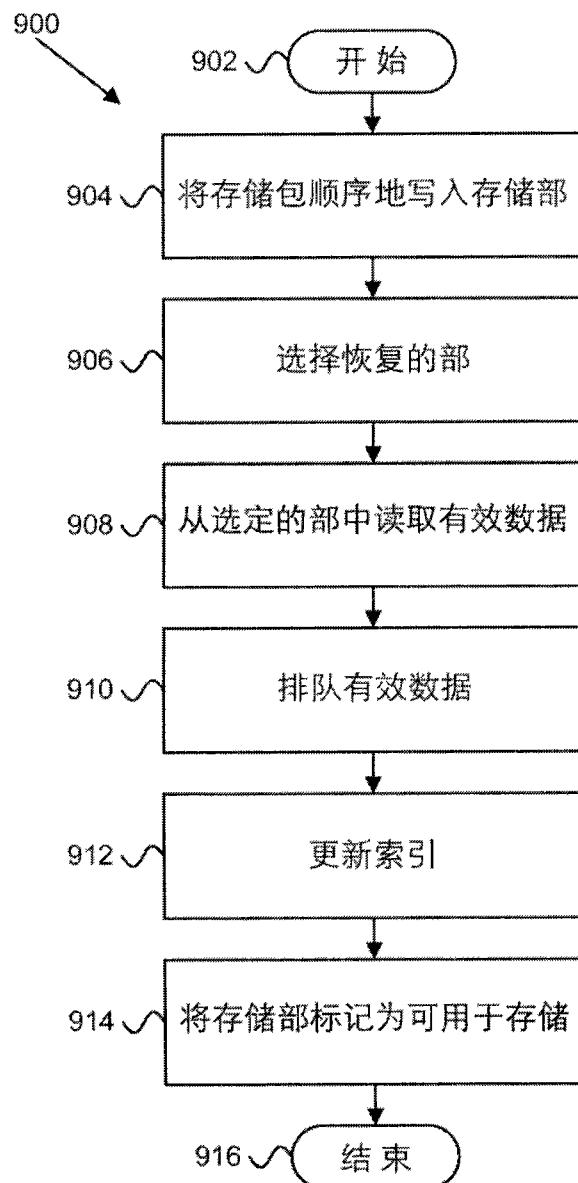


图 9

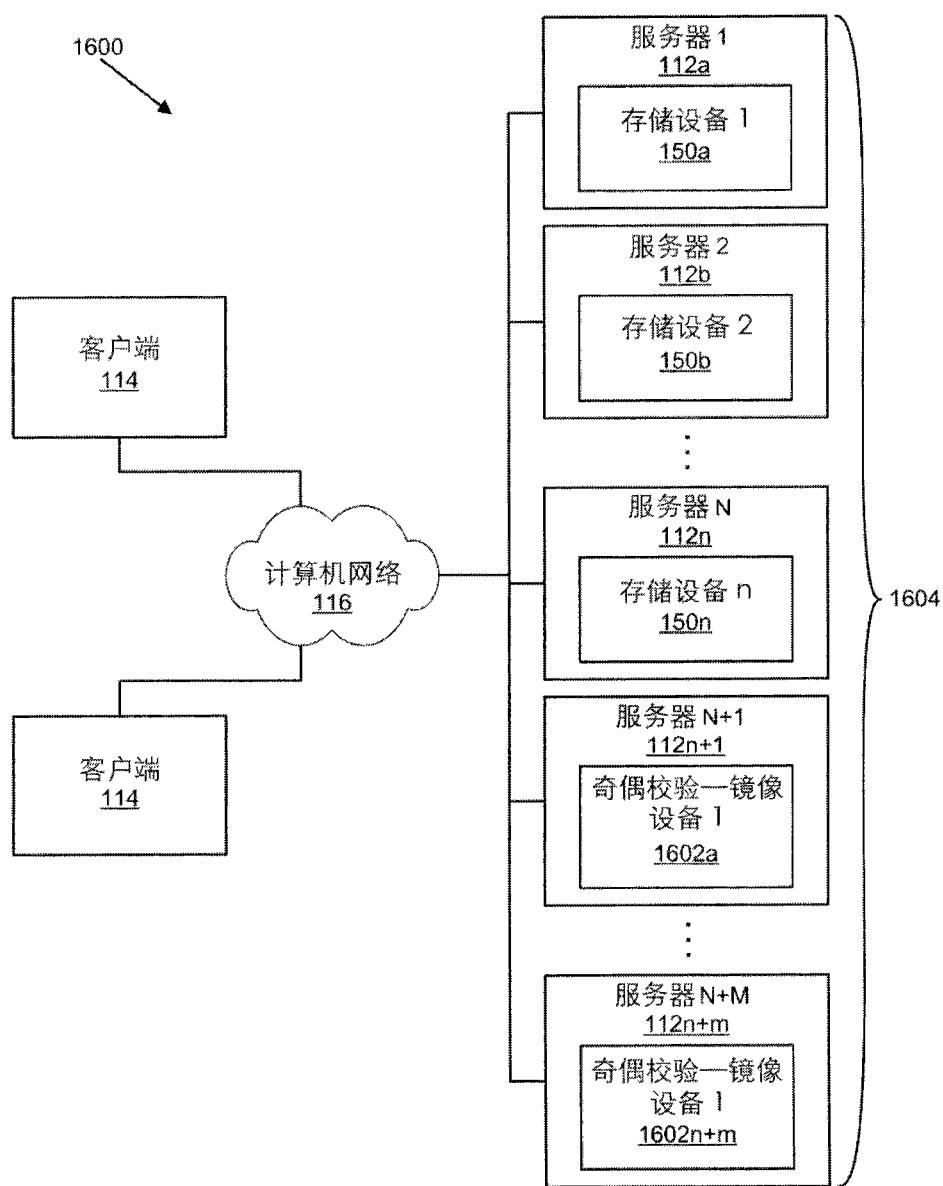


图 10

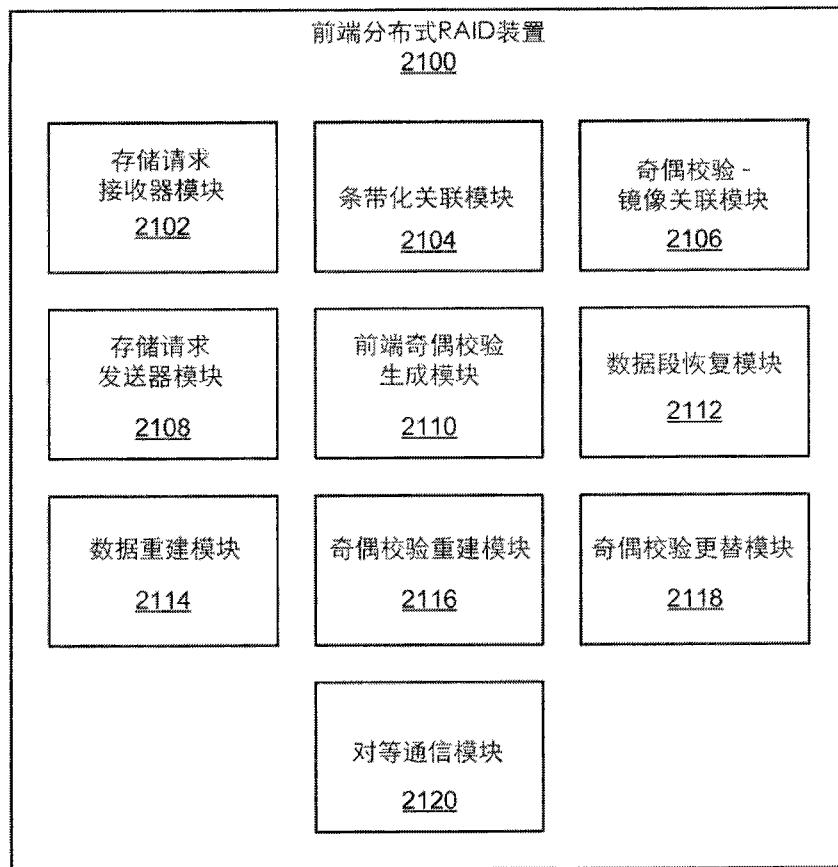


图 11

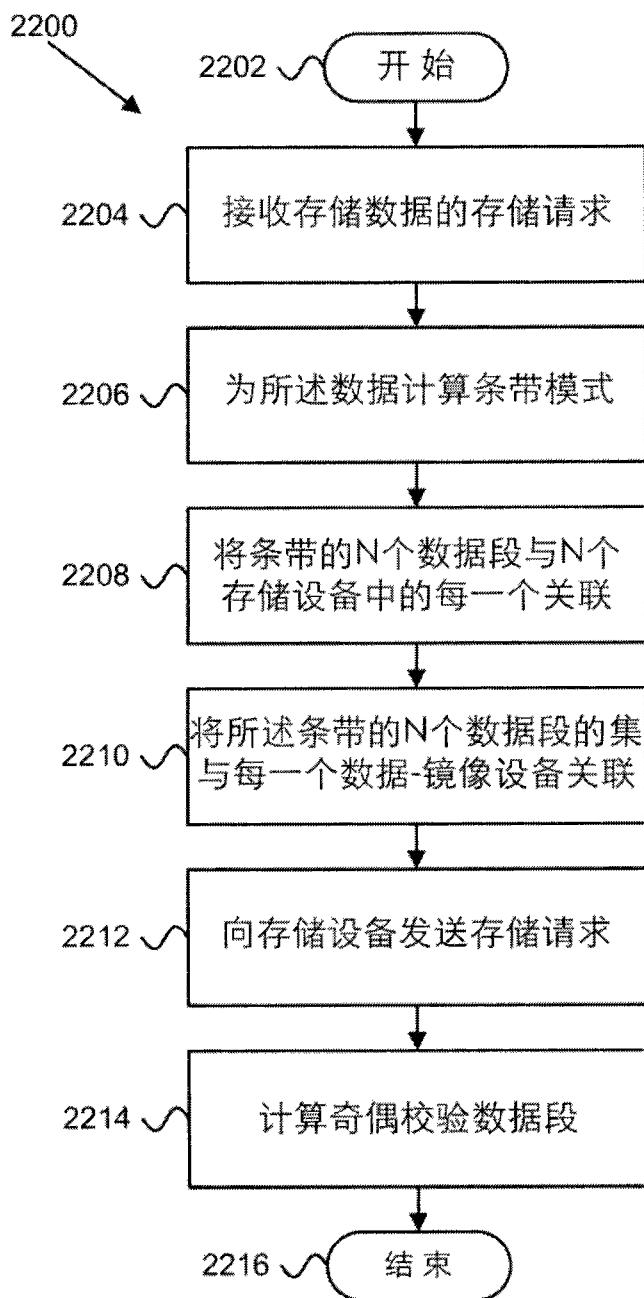


图 12

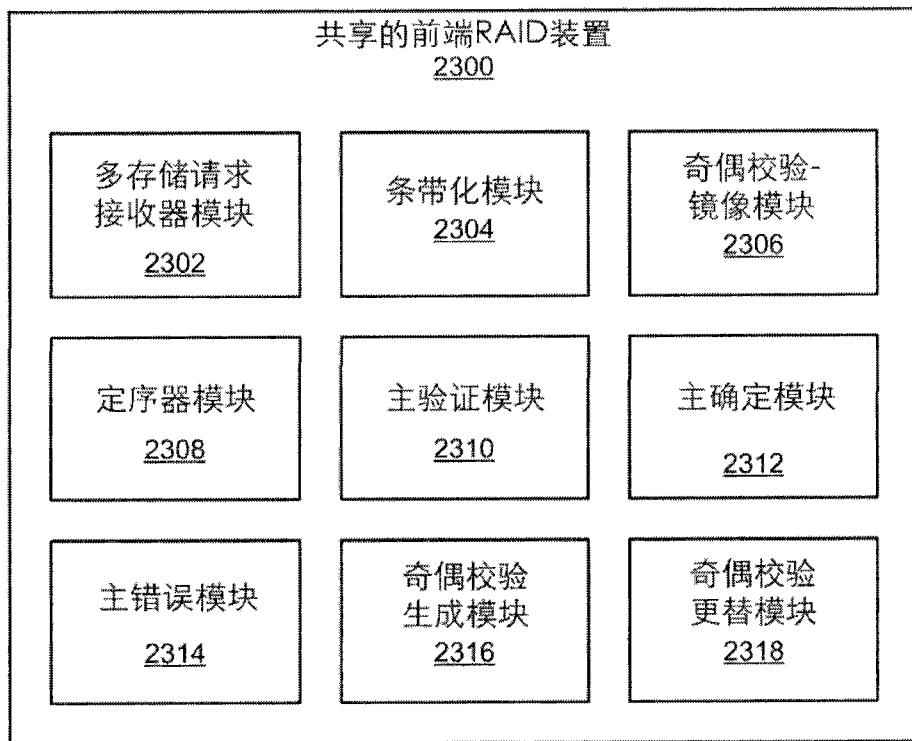


图 13

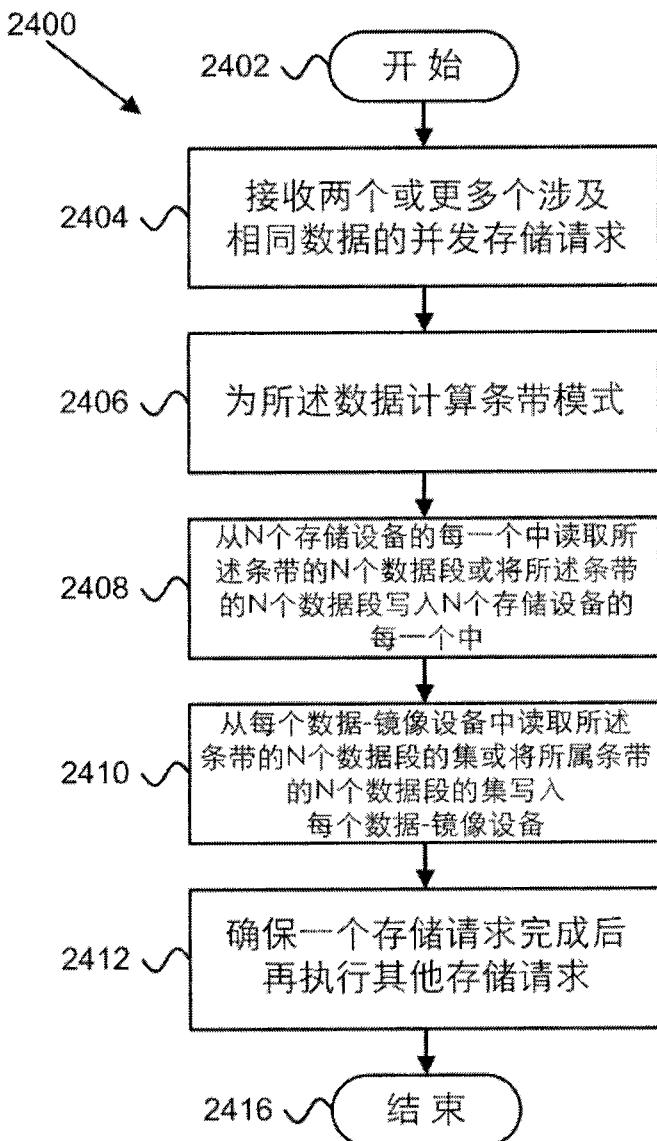


图 14

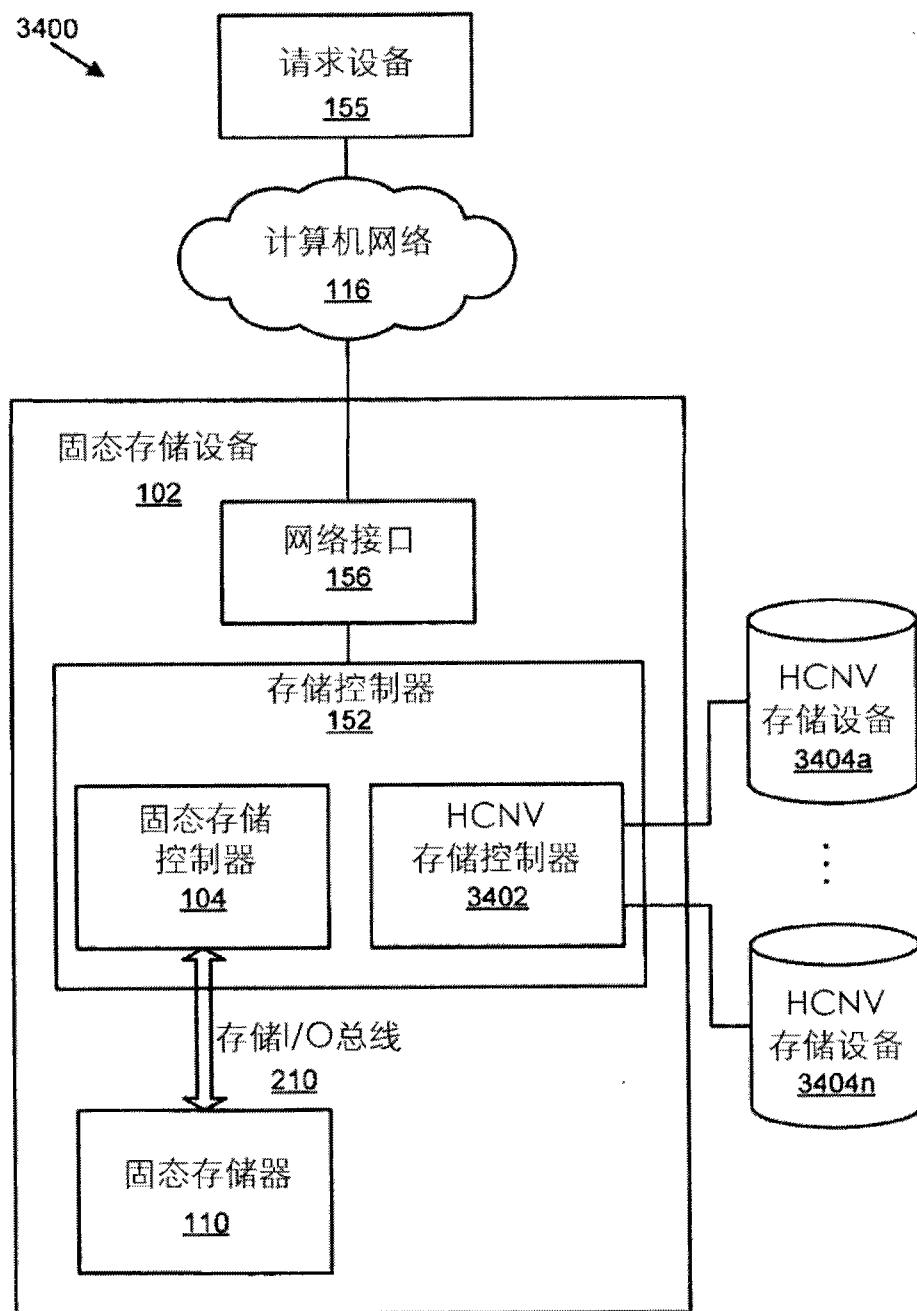


图 15

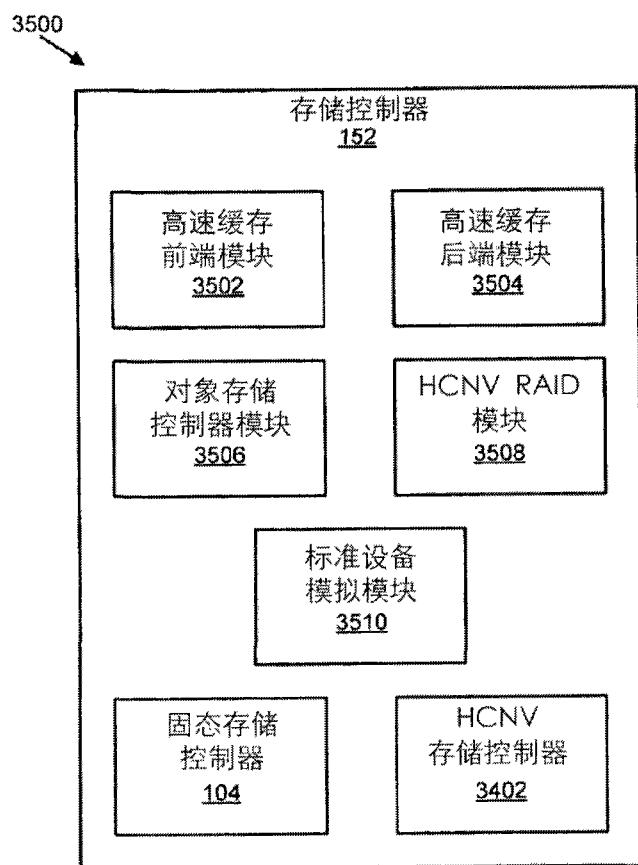


图 16

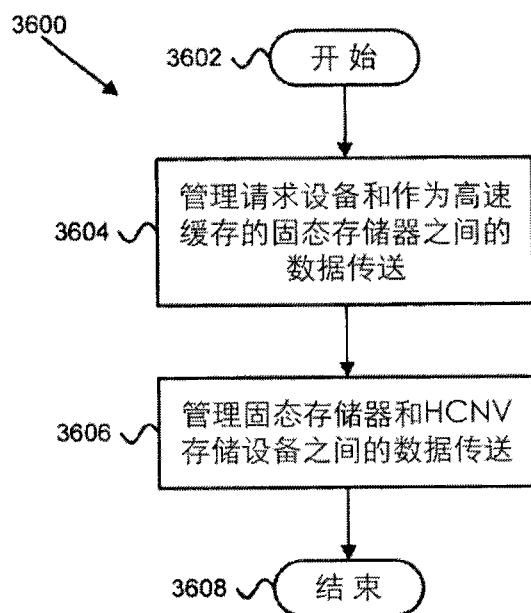


图 17