(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 108776576 B
(45)授权公告日 2023.08.15

(21)申请号 201810263102.0

(22)申请日 2018.03.28

(65)同一申请的已公布的文献号
申请公布号 CN 108776576 A

(43)申请公布日 2018.11.09

(30)优先权数据
62/480,113 2017.03.31 US
62/483,913 2017.04.10 US
15/618,081 2017.06.08 US

(73)专利权人 三星电子株式会社
地址 韩国京畿道

(72)发明人 拉姆达斯·P·卡恰尔
颂蓬·保罗·奥拉里希
弗莱德·沃利

(74)专利代理机构 北京天昊联合知识产权代理有限公司 11112
专利代理师 张帆 赵南

(51)Int.Cl.
G06F 3/06(2006.01)
G06F 12/02(2006.01)

(56)对比文件
CN 101956936 A,2011.01.26
CN 104346287 A,2015.02.11
CN 105556930 A,2016.05.04
CN 105589661 A,2016.05.18
CN 106020723 A,2016.10.12
CN 106104500 A,2016.11.09
CN 106126447 A,2016.11.16
KR 20170007103 A,2017.01.18
US 2012026423 A1,2012.02.02
US 2015356020 A1,2015.12.10
Nusrat Sharmin Islam.et al.High Performance Design for HDFS with Byte-Addressability of NVM and RDMA.ICS '16: Proceedings of the 2016 International Conference on Supercomputing.2016,第1-14页.

审查员 陈娜

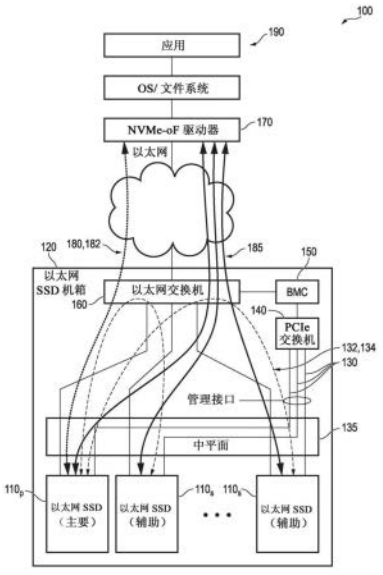
权利要求书3页 说明书12页 附图9页

(54)发明名称

用于聚合的网上NVMe装置的聚合存储方法

(57)摘要

本公开提供了一种用于NVMe-oF装置的存储聚合方法、一种在一组NVMe-oF以太网SSD中的NVMe-oF SSD容量聚合的方法以及一种聚合的以太网SSD组。所述用于NVMe-oF装置的存储聚合方法包括:将聚合组识别为包括多个NVMe-oF SSD的聚合的以太网SSD;选择聚合组的NVMe-oF SSD之一作为主要NVMe-oF SSD;选择聚合组的其它NVMe-oF SSD作为辅助NVMe-oF SSD;以及利用管理NVMe-oF SSD的处理器初始化主要NVMe-oF SSD中的映射分配表。



CN 108776576 B

1.一种用于NVMe-oF装置的存储聚合方法,包括:

将聚合组识别为包括多个NVMe-oF SSD的聚合的以太网SSD;

选择所述聚合组的NVMe-oF SSD之一作为主要NVMe-oF SSD;

选择所述聚合组的其它NVMe-oF SSD作为辅助NVMe-oF SSD;以及

利用管理NVMe-oF SSD的处理器初始化所述主要NVMe-oF SSD中的映射分配表,

其中,仅所述主要NVMe-oF SSD对于主机可见,并且

其中,所述处理器被构造为初始确定所述多个NVMe-oF SSD中的哪一个被选择为所述主要NVMe-oF SSD。

2.根据权利要求1所述的方法,其中,在连接至所述聚合组的存储管理员的指导下进行利用所述处理器初始化所述映射分配表。

3.根据权利要求1所述的方法,其中,所述映射分配表包括针对所述聚合组的NVMe-oF SSD中的每一个的NVMe-oF SSD的容量、NVMe-oF SSD的地址和NVMe-oF SSD的容量的剩余量。

4.根据权利要求3所述的方法,还包括:

将所述主要NVMe-oF SSD的地址提供至用户应用,以能够在所述聚合组与所述用户应用之间进行数据转移。

5.根据权利要求1所述的方法,还包括:

在所述主要NVMe-oF SSD处从连接至所述聚合组的所述主机接收管理命令;

确定对应于所述管理命令的数据是仅存储在所述主要NVMe-oF SSD上还是存储在一个或多个辅助NVMe-oF SSD上;

当数据仅存储在所述主要NVMe-oF SSD上时,将所述数据转移至所述主机;以及

当数据存储在所述一个或多个辅助NVMe-oF SSD上时,将所述管理命令分割为分别对应于所述一个或多个辅助NVMe-oF SSD的一个或多个管理子命令,从所述一个或多个辅助NVMe-oF SSD接收子命令完成条目,以及生成完成条目并将其从所述主要NVMe-oF SSD发送至所述主机。

6.根据权利要求5所述的方法,还包括:当数据存储在所述一个或多个辅助NVMe-oF SSD上时,

在所述一个或多个辅助NVMe-oF SSD中的对应的辅助NVMe-oF SSD处接收所述一个或多个管理子命令中的管理子命令;

根据该管理子命令确定是否将所述数据从所述对应的辅助NVMe-oF SSD转移至所述主要NVMe-oF SSD;

生成完成条目;以及

将所述完成条目发送至所述主要NVMe-oF SSD。

7.根据权利要求1所述的方法,还包括:

在所述主要NVMe-oF SSD处接收生成命名空间的命令或者删除命名空间的命令;

利用所述主要NVMe-oF SSD参照所述映射分配表;

当所述命令是生成命名空间时,在所述主要NVMe-oF SSD和/或一个或多个辅助NVMe-oF SSD中分配容量,或者当所述命令是删除命名空间时,检索所述主要NVMe-oF SSD和/或所述一个或多个辅助NVMe-oF SSD中的对应的一个;以及

更新所述映射分配表。

8.根据权利要求1所述的方法,还包括:

在所述主要NVMe-oF SSD处接收读/写命令;

利用所述主要NVMe-oF SSD查找所述映射分配表;

生成一个或多个读/写子命令;

将所述一个或多个读/写子命令分别发送至一个或多个辅助NVMe-oF SSD;

根据所述读/写命令在所述主机与所述主要NVMe-oF SSD和/或所述一个或多个辅助NVMe-oF SSD之间转移数据;以及

在转移所述数据之后将完成发送至所述主机。

9.根据权利要求8所述的方法,还包括:

在所述一个或多个辅助NVMe-oF SSD中的对应的辅助NVMe-oF SSD处接收所述一个或多个读/写子命令中的读/写子命令;

提取对应于该读/写子命令的传输信息;

将读/写请求从所述对应的辅助NVMe-oF SSD发送至所述主机;以及

在完成对应于所述读/写子命令的数据转移之后将完成条目发送至所述主要NVMe-oF SSD。

10.一种在一组NVMe-oF以太网SSD中的NVMe-oF SSD容量聚合的方法,包括:

识别聚合组的多个NVMe-oF SSD;

将所述多个NVMe-oF SSD之一指定为主要NVMe-oF SSD;以及

将其余NVMe-oF SSD指定为辅助NVMe-oF SSD,

其中,只有主机的主机驱动器可见的NVMe-oF SSD是所述主要NVMe-oF SSD,

其中,仅所述主要NVMe-oF SSD对于主机可见,并且

其中,由处理器初始确定所述多个NVMe-oF SSD中的哪一个被指定为所述主要NVMe-oF SSD。

11.根据权利要求10所述的方法,还包括:

利用所述主要NVMe-oF SSD根据所述主要NVMe-oF SSD从所述主机接收的命令维护映射分配表,

其中,所述映射分配表指示在所述聚合组的主要NVMe-oF SSD和一个或多个辅助NVMe-oF SSD之间划分的逻辑块地址空间。

12.根据权利要求11所述的方法,还包括:

利用处理器初始化映射分配表,以根据将各个NVMe-oF SSD之一指定为所述主要NVMe-oF SSD来构造所述聚合组。

13.根据权利要求10所述的方法,还包括:

利用所述主要NVMe-oF SSD根据所述主要NVMe-oF SSD从所述主机接收的命令聚合各个辅助NVMe-oF SSD的容量,以使得所述聚合组的所述多个NVMe-oF SSD针对所述主机呈现为单个聚合的逻辑容量。

14.根据权利要求10所述的方法,还包括:

利用所述主要NVMe-oF SSD将容量分配至一个或多个辅助NVMe-oF SSD;以及

利用所述主要NVMe-oF SSD在映射分配表中记录分配的容量和关联的映射逻辑块地址

范围。

15.根据权利要求10所述的方法,还包括:

利用所述主要NVMe-oF SSD为所述辅助NVMe-oF SSD和所述主要NVMe-oF SSD的聚合的容量设置预留空间。

16.根据权利要求10所述的方法,还包括:

在所述主要NVMe-oF SSD处从所述主机接收命令;

将所述命令划分为多个子命令,所述多个子命令各自对应于各个对应的辅助NVMe-oF SSD中的对应的一个;以及

将子命令从所述主要NVMe-oF SSD发送至各个对应的辅助NVMe-oF SSD。

17.根据权利要求16所述的方法,还包括:

基于各自的子命令,将数据从各个对应的辅助NVMe-oF SSD直接转移至所述主机。

18.根据权利要求16所述的方法,还包括:

在各个对应的辅助NVMe-oF SSD处从所述主要NVMe-oF SSD接收各自的子命令;

执行对应于各自的子命令的任务;以及

在完成任务时将各自的子命令完成条目从各个对应的辅助NVMe-oF SSD发送至所述主要NVMe-oF SSD。

19.根据权利要求18所述的方法,还包括:

利用所述主要NVMe-oF SSD维护子命令上下文表;

在所述主要NVMe-oF SSD处从各个辅助NVMe-oF SSD接收子命令完成条目;以及

利用所述主要NVMe-oF SSD根据接收到的所述子命令完成条目跟踪所述子命令的执行。

20.一种聚合的以太网SSD组,包括:

以太网SSD机箱;

所述以太网SSD机箱上的以太网交换机,其用于使得能够与主机的主机驱动器通信;

耦接至所述以太网交换机的处理器;

耦接至所述处理器的PCIe交换机;

多个NVMe-oF SSD,其中包括:

主要NVMe-oF SSD;以及

多个辅助NVMe-oF SSD,其经由包括所述以太网交换机和所述PCIe交换机的专用通信信道连接至所述主要NVMe-oF SSD,其中,仅所述主要NVMe-oF SSD对于主机可见,并且

其中,所述处理器被构造为初始确定所述多个NVMe-oF SSD中的哪一个被指定为所述主要NVMe-oF SSD。

用于聚合的网上NVMe装置的聚合存储方法

[0001] 相关申请的交叉引用

[0002] 本申请要求于分别于2017年3月31日和2017年4月10日在美国专利和商标局提交的临时申请No.62/480,113和No.62/483,913的优先权,所述申请的内容以引用方式并入本文中。

技术领域

[0003] 本公开的一些实施例总体涉及一种系统和方法,其用于聚合多个存储器驱动(例如,eSSD),以被主机感知为单个的大逻辑容量。

背景技术

[0004] 固态盘(SSD)迅速变为现代IT基础设施的优选存储元件,从而替代传统硬盘驱动(HDD)。SSD提供非常低的延迟、高数据读/写吞吐量以及可靠的数据存储。

[0005] 网上高速非易失性存储器(Non-Volatile Memory express over Fabrics,NVMe-oF)是一种允许几百个或几千个NVMe-oF装置(例如,高速非易失性存储器(NVMe)SSD)通过诸如IB、FC和以太网的网络结构连接的新兴技术。NVMe-oF协议允许实现远程直接附加存储(rDAS)。这样,允许大量SSD连接至远程主机。NVMe-oF协议使用远程直接存储器存取(RDMA)协议来提供NVMe命令、数据和响应的可靠通信。用于提供RDMA服务的传输协议包括iWARP、RoCE v1和RoCE v2。

[0006] NVMe-oF接口允许大量SSD连接至远程主机。按照常规,针对每个NVMe-oF SSD,在远程主机上运行驱动器实例。对于一些应用,单个SSD提供的存储容量会是不足的。

发明内容

[0007] 本公开的一些实施例提供了一种聚合多个SSD的方法,所述多个SSD被主机感知为单个大容量逻辑卷,以及一种用于实现所述方法的网络状结构。

[0008] 根据一些实施例,一种网上NVMe装置的存储聚合方法,包括:将聚合组识别为包括多个NVMe-oF SSD的聚合的以太网SSD;选择聚合组的NVMe-oF SSD之一作为主要NVMe-oF SSD;选择聚合组的其它NVMe-oF SSD作为辅助NVMe-oF SSD;以及利用管理NVMe-oF SSD的处理器初始化主要NVMe-oF SSD中的映射分配表。

[0009] 根据一些示例实施例,在连接至聚合组的存储管理员的指导下进行利用处理器初始化映射分配表。

[0010] 根据一些示例实施例,映射分配表包括针对聚合组的NVMe-oF SSD中的每一个的NVMe-oF SSD的容量、NVMe-oF SSD的地址和NVMe-oF SSD的容量的剩余量。

[0011] 根据一些示例实施例,所述方法还包括:将主要NVMe-oF SSD的地址提供至用户应用,以能够在聚合组与用户应用之间进行数据转移。

[0012] 根据一些示例实施例,所述方法还包括:在主要NVMe-oF SSD处从连接至聚合组的主机接收管理命令;确定对应于管理命令的数据是仅存储在主要NVMe-oF SSD上还是存储

在一个或多个辅助NVMe-oF SSD上;当数据存储在所述一个或多个辅助NVMe-oF SSD上时,将管理命令分割为分别对应于所述一个或多个辅助NVMe-oF SSD的一个或多个管理子命令;将数据转移至主机;从所述一个或多个辅助NVMe-oF SSD接收子命令完成条目;以及生成完成条目并将其从主要NVMe-oF SSD发送至主机。

[0013] 根据一些示例实施例,所述方法还包括:在所述一个或多个辅助NVMe-oF SSD中的对应的辅助NVMe-oF SSD处接收所述一个或多个管理子命令中的管理子命令;根据该管理子命令确定是否将数据从所述对应的辅助NVMe-oF SSD转移至主要NVMe-oF SSD;生成完成条目;以及将完成条目发送至主要NVMe-oF SSD。

[0014] 根据一些示例实施例,所述方法还包括:在主要NVMe-oF SSD处接收生成命名空间的命令或者删除命名空间的命令;利用主要NVMe-oF SSD参照映射分配表;当命令是生成命名空间时,在主要NVMe-oF SSD和/或所述一个或多个辅助NVMe-oF SSD中分配容量,或者当命令是删除命名空间时,检索主要NVMe-oF SSD和/或所述一个或多个辅助NVMe-oF SSD中的对应的一个;以及更新映射分配表。

[0015] 根据一些示例实施例,所述方法还包括:在主要NVMe-oF SSD处接收读/写命令;利用主要NVMe-oF SSD查找映射分配表;生成一个或多个读/写子命令;将所述一个或多个读/写子命令分别发送至所述一个或多个辅助NVMe-oF SSD;根据读/写命令在主机与主要NVMe-oF SSD和/或所述一个或多个辅助NVMe-oF SSD之间转移数据;以及在转移数据之后将完成条目发送至主机。

[0016] 根据一些示例实施例,所述方法还包括:在所述一个或多个辅助NVMe-oF SSD中的对应的辅助NVMe-oF SSD处接收所述一个或多个读/写子命令中的读/写子命令;提取对应于该读/写子命令的传输信息;将读/写请求从所述对应的辅助NVMe-oF SSD发送至主机;以及在完成对应于读/写子命令的数据转移之后将完成条目发送至主要NVMe-oF SSD。

[0017] 根据一些示例实施例,一种在一组NVMe-oF以太网SSD中的NVMe-oF SSD容量聚合的方法,包括:识别聚合组的多个NVMe-oF SSD;将各个NVMe-oF SSD之一指定为主要NVMe-oF SSD;以及将其余NVMe-oF SSD指定为辅助NVMe-oF SSD,其中,只有主机的主机驱动器可见的NVMe-oF SSD是主要NVMe-oF SSD。

[0018] 根据一些示例实施例,所述方法还包括:利用主要NVMe-oF SSD根据主要NVMe-oF SSD从主机接收的命令维护映射分配表,其中,映射分配表指示在聚合组的主要NVMe-oF SSD和一个或多个辅助NVMe-oF SSD之间划分的逻辑块地址(LBA)空间。

[0019] 根据一些示例实施例,所述方法还包括:利用处理器初始化映射分配表,以根据将各个NVMe-oF SSD之一指定为主要NVMe-oF SSD来构造聚合组。

[0020] 根据一些示例实施例,所述方法还包括:利用主要NVMe-oF SSD根据主要NVMe-oF SSD从主机接收的命令聚合各个辅助NVMe-oF SSD的容量,以使得聚合组的所述多个NVMe-oF SSD针对主机呈现为单个聚合的逻辑容量。

[0021] 根据一些示例实施例,所述方法还包括:利用主要NVMe-oF SSD将容量分配至一个或多个辅助NVMe-oF SSD;以及利用主要NVMe-oF SSD在映射分配表中记录分配的容量和关联的映射逻辑块地址(LBA)范围。

[0022] 根据一些示例实施例,所述方法还包括:利用主要NVMe-oF SSD为辅助NVMe-oF SSD和主要NVMe-oF SSD的聚合的容量设置预留空间(overprovisioning)。

[0023] 根据一些示例实施例,所述方法还包括:在主要NVMe-oF SSD处从主机接收命令;将所述命令划分为多个子命令,所述多个子命令各自对应于各个对应的辅助NVMe-oF SSD中的对应的一个;以及将子命令从主要NVMe-oF SSD发送至各个对应的辅助NVMe-oF SSD。

[0024] 根据一些示例实施例,所述方法还包括:基于各自的子命令,将数据从各个对应的辅助NVMe-oF SSD直接转移至主机。

[0025] 根据一些示例实施例,所述方法还包括:在各个对应的辅助NVMe-oF SSD处从主要NVMe-oF SSD接收各自的子命令;执行对应于各自的子命令的任务;以及在完成任务时将各自的子命令完成条目从各个对应的辅助NVMe-oF SSD发送至主要NVMe-oF SSD。

[0026] 根据一些示例实施例,所述方法还包括:利用主要NVMe-oF SSD维护子命令上下文表;在主要NVMe-oF SSD处从各个辅助NVMe-oF SSD接收子命令完成条目;以及利用主要NVMe-oF SSD根据接收到的子命令完成条目跟踪子命令的执行。

[0027] 根据一些示例实施例,提供了一种聚合的以太网SSD组,包括:以太网SSD机箱;以太网SSD机箱上的以太网交换机,其用于使得能够与主机驱动器通信;耦接至以太网交换机的处理器;耦接至板管理控制器的PCIe交换机;以及多个NVMe-oF SSD,其中包括:主要NVMe-oF SSD;以及多个辅助NVMe-oF SSD,其经由包括以太网交换机和PCIe交换机的专用通信信道连接至主要NVMe-oF SSD,其中,仅主要NVMe-oF SSD对于主机可见,并且其中,板管理控制器被构造为初始确定各个NVMe-oF SSD中的哪一个包括主要NVMe-oF SSD。

[0028] 因此,因为单个主要eSSD在通过辅助eSSD跟踪所有相关子命令的完成并保持辅助eSSD对于主机不可见的同时,执行了所有NVMe-oF协议处理,所以eSSD的聚合组对主机呈现为单个大逻辑容量。

附图说明

[0029] 通过以下结合附图的描述可更加详细地理解一些实施例,其中:

[0030] 图1示出了根据本公开的实施例的在单个eSSD机箱中包括多个聚合的eSSD的NVMe-oF以太网SSD(eSSD)存储中使用的系统架构的框图;

[0031] 图2示出了根据本公开的实施例的图1的实施例所示的多个eSSD机箱在以太网SSD机架中连接在一起的框图;

[0032] 图3是根据本公开的实施例的由主要eSSD维护的“映射分配表”的示例;

[0033] 图4示出了根据本公开的实施例的描绘映射分配表的初始化的流程图;

[0034] 图5示出了根据本公开的实施例的NVMe-oF以太网SSD(eSSD)存储中使用的系统架构的框图,所述NVMe-oF以太网SSD(eSSD)存储包括进出分别位于多个机架中的多个eSSD机箱中的多个聚合的eSSD的数据;

[0035] 图6是描绘示例命令上下文的表;

[0036] 图7示出了根据本公开的实施例的描绘通过主要eSSD处理管理命令的流程图;

[0037] 图8示出了根据本公开的实施例的描绘命名空间生成和删除命令的执行的流程图;

[0038] 图9示出了根据本公开的实施例的描绘在P-eSSD的控制下执行读/写命令的流程图;

[0039] 图10示出了根据本公开的实施例的描绘通过S-eSSD执行管理子命令的流程图;

[0040] 图11示出了根据本公开的实施例的描绘通过S-eSSD执行读/写子命令的流程图；以及

[0041] 图12示出了根据本公开的实施例的描绘在S-eSSD中的数据转移和子命令完成同步的流程图。

具体实施方式

[0042] 通过参照以下对实施例和附图的详细描述可更容易地理解本发明构思的特征及其实现方法。下文中，将参照附图更详细地描述实施例，其中相同的附图标记始终指代相同元件。然而，本发明可按照许多不同形式实施，并且不应理解为仅限于本文示出的实施例。相反，这些实施例作为示例提供，以使得本公开将是彻底和完整的，并且将把本发明的各方面和各特征完全传递给本领域技术人员。因此，本领域普通技术人员对于完全理解本发明的各方面和各特征所不需要的处理、元件和技术可不描述。除非另有说明，否则相同的附图标记在附图和撰写的说明书中始终指代相同元件，因此，将不重复对其的描述。在附图中，为了清楚起见，可夸大元件、层和区的相对尺寸。

[0043] 在以下描述中，出于解释的目的，阐述了许多特定细节，以提供对各个实施例的彻底理解。然而，应该清楚，可不需要这些特定细节或者通过一个或多个等同布置方式来实施各个实施例。在其它实例中，公知的结构和装置按照框图形式示出，以避免不必要地使各个实施例不清楚。

[0044] 应该理解，虽然本文中可使用术语“第一”、“第二”、“第三”等来描述各个元件、组件、区、层和/或部分，但是这些元件、组件、区、层和/或部分不应被这些术语限制。这些术语用于将一个元件、组件、区、层或部分与另一元件、组件、区、层或部分区分开。因此，下面描述的第一元件、第一组件、第一区、第一层或第一部分可被称作第二元件、第二组件、第二区、第二层或第二部分，而不脱离本发明的精神和范围。

[0045] 为了方便描述，本文中可使用诸如“在……下方”、“在……之下”、“下”、“在……下”、“在……之上”、“上”等的空间相对术语，以描述附图中所示的一个元件或特征与另一元件或特征的关系。应该理解，空间相对术语旨在涵盖使用或操作中的装置的除图中所示的取向之外的不同取向。例如，如果图中的装置颠倒，则被描述为“在其它元件或特征之下”或“在其它元件或特征下方”或“在其它元件或特征下”的元件将因此被取向为“在其它元件或特征之上”。因此，示例性术语“在……之下”和“在……下”可涵盖在……之上和在……之下这两个取向。装置可按照其它方式取向(旋转90度或位于其它取向)，并且本文所用的空间相对描述语将相应地解释。

[0046] 应该理解，当将元件、层、区或组件称作“位于”另一元件、层、区或组件“上”、“连接至”或“结合至”另一元件、层、区或组件时，其可直接位于所述另一元件、层、区或组件上、直接连接至或结合至所述另一元件、层、区或组件，或者可存在一个或多个中间元件、层、区或组件。另外，还应该理解，当一个元件或层被称作“位于”两个元件或层“之间”时，其可为所述两个元件或层之间的唯一元件或层，或者也可存在一个或多个中间元件或层。

[0047] 针对本公开的目的，“X、Y和Z中的至少一个”以及“选自X、Y和Z中的至少一个”可理解为仅X、仅Y、仅Z，或者X、Y和Z中的两个或更多个的任意组合，诸如，举例来说，XYZ、XYY、YZ和ZZ。相同的附图标记始终指代相同元件。如本文所用，术语“和/或”包括一个或多个相关

所列项的任意和所有组合。

[0048] 在下面的示例中，x轴、y轴和z轴不限于直角坐标系的三个轴，而是可由更宽的含义解释。例如，x轴、y轴和z轴可彼此垂直，或者可代表彼此不垂直的不同方向。

[0049] 本文使用的术语仅是为了描述特定实施例，并且不旨在限制本发明。如本文所用，除非上下文清楚地另有说明，否则单数形式“一种”和“一”也旨在包括复数形式。还应该进一步理解，当术语“包括”、“包含”、“包括……的”和“包含……的”用于本说明书中时，指明存在所列特征、整体、步骤、操作、元件和/或组件，但是不排除存在或添加一个或多个特征、整体、步骤、操作、元件、组件和/或它们的组。如本文所用，术语“和/或”包括一个或多个相关所列项的任何和所有组合。当诸如“中的至少一个”的表达出现于元件的列表之后时，其修饰元件的整个列表而不修饰列表中的单独的元件。

[0050] 如本文所用，术语“基本上”、“约”和类似术语用作表达接近而不是表达程度，并且旨在说明本领域普通技术人员应该能够认识到的测量或计算的值的固有偏差。此外，当描述本发明的实施例时，使用“可”是指“本发明的一个或多个实施例”。如本文所用，可认为术语“使用”与术语“利用”同义。另外，术语“示例性”旨在指代示例或示出。

[0051] 当可不同地实施特定实施例时，可与描述的次序不同地执行特定的处理次序。例如，两个连续地描述的处理可同时执行或者按照与描述的次序相反的次序执行。

[0052] 另外，本文公开和/或说明的任何数的范围旨在包括说明的范围内的相同数值精度的所有子范围。例如，范围“1.0至10.0”旨在包括在说明的最小值1.0与说明的最大值10.0之间(包括端点)的所有子范围，也就是说，最小值等于或大于1.0，最大值等于或小于10.0，诸如，例如，2.4至7.6。本文说明的数的任何上限旨在包括其中包含的所有较小数限，并且本文说明的数的任何下限旨在包括其中包含的所有较大数限。因此，申请人保留修改本说明(包括权利要求)的权利，以特别说明在本文特别说明的范围内包含的任何子范围。

[0053] 本文参照作为实施例和/或中间结构的示意图的剖视图描述各个实施例。这样，作为例如制造技术和/或公差的结果，附图中的形状的变化是可以预见的。因此，本文公开的实施例不应被理解为限于区的具体示出的形状，而是包括例如由制造工艺导致的形状的偏差。例如，示为矩形的注入区将通常具有圆形或弯曲特征和/或在其边缘具有注入浓度的梯度，而不是从注入区至非注入区二值变化。同样地，通过注入形成的掩埋区可在掩埋区与通过其发生注入的表面之间的区中导致一些注入。因此，图中示出的区实际上是示意性的，并且它们的形状不旨在示出装置的区的实际形状，并且不旨在限制。

[0054] 根据本文所述的本发明的实施例的电子装置或电装置和/或任何其它相关装置或组件可利用任何合适的硬件，固件(例如专用集成电路)、软件或者软件、固件和硬件的组合来实施。例如，这些装置的各个组件可形成在一个集成电路(IC)芯片或者分离的IC芯片上。此外，这些装置的各个组件可在柔性印刷电路膜、带载封装(TCP)、印刷电路板(PCB)上实施，或者形成在一个衬底上。此外，这些装置的各个组件可为在一个或多个计算装置中在一个或多个处理器上运行的一个处理或线程，以执行计算机程序指令和与其它系统组件交互，以执行本文所述的各种功能。所述计算机程序指令存储在可利用标准存储器装置在计算装置中实施的存储器中，诸如例如，随机存取存储器(RAM)。所述计算机程序指令也可存储在诸如例如CD-ROM、闪速驱动器等的其它非易失性计算机可读介质中。另外，本领域技术人员应该理解，本领域技术人员应该理解，各个计算装置的功能可组合或者集成在单个计

算装置中,或者特定计算装置的功能可分布在一个或多个其它计算装置上,而不脱离本发明的示例性实施例的精神和范围。

[0055] 除非另外限定,否则本文中使用的所有术语(包括技术和科学术语)具有与本发明所属领域的普通技术人员之一通常理解的含义相同的含义。还应该理解,除非本文中明确这样定义,否则诸如在通用词典中定义的那些的术语应该被解释为具有与它们在相关技术和/或本说明书的上下文中的含义一致的含义,而不应该理想化地或过于正式地解释它们。

[0056] 图1示出了根据本公开的实施例的在单个eSSD机箱120中包括多个聚合的eSSD 110的NVMe-oF以太网SSD(eSSD)存储中使用的系统架构100的框图。图2示出了根据本公开的实施例的诸如图1的实施例所示的那些的多个eSSD机箱120在以太网SSD机架中连接在一起的框图。

[0057] 如上所述,NVMe-oF接口允许大量SSD 110连接至远程主机190。按照常规,针对各个NVMe-oF SSD 110,在远程主机190上运行驱动器实例。然而,对于一些应用,通过单个SSD 110提供的存储容量不足。这种应用可受益于几百太字节(terabyte)的容量的单个逻辑卷。因此,这种应用可受益于提供了在“聚合组”中聚合在一起的大量单独的SSD 110的本公开的实施例,并且所述大量单独的SSD 110对于所述应用呈现为单个逻辑卷。

[0058] 例如,24个16太字节(16TB)的eSSD可呈现为单个逻辑384TB驱动。需要大量聚合SSD 110的一些应用示例包括大数据挖掘和分析、石油化工、天然气与能源勘探、实验粒子物理学和药物开发。这些示例会需要高性能计算(HPC),这需要大存储容量和高性能二者。

[0059] 虽然可存在聚合底层SSD 110并且提供单个逻辑、易扩展卷的系统软件层,但是这种系统软件通常是高度复杂和庞杂的。这种软件会需要在主机190上运行的大量NVMe-oF驱动器实例,从而消耗诸如存储器、CPU周期和功率的系统资源。目标侧的解决方案可潜在地使用x86服务器或者RAID-on-Chip(ROC)系统,以提供大容量作为单个逻辑卷。然而,这种解决方案通常是复杂的、昂贵的,并且具有性能和能量的负面影响。例如,根据本发明的实施例,通过CPU接收和发送数据会消耗比通过DMA引擎、ASIC等消耗的能量大好几倍的能量的量。

[0060] 因此,本公开的实施例按照有效和成本经济的方式提供了一种用于以太网NVMe-oF SSD中聚合多个eSSD 110的方法和结构。

[0061] 参照图1,(例如,奉存储管理员的命令)为eSSD 110分配两个角色之一,从而各个eSSD 110用作主要eSSD(P-eSSD)110p或者辅助eSSD(S-eSSD)110s。单个机箱120(或者给定机架中的多个机箱120,或者分布在宽范围的多个机架230中的多个机箱120)中的单个P-eSSD 110p和一组多个S-eSSD 110s作为单个逻辑驱动整体提供由远程主机190使用的必要的闪速存储器容量。eSSD机箱120包括eSSD 110以及诸如板管理控制器(BMC)装置150的处理器和用于外部连接的以太网交换机160。虽然eSSD机箱120用于在以下实施例的描述中指代一组NVMe-oF装置,但是本发明的其它实施例可简单地应用于任何其它多个NVMe-oF装置,而不管其物理外壳如何(例如,机箱、机架或者基于容器的外壳)。此外,虽然eSSD 110用于描述下面描述的实施例的NVMe-oF装置,但是其它NVMe-oF装置可等同地应用于本发明的其它实施例。

[0062] 因此,经由对应的以太网交换机160,可跨越机架230中的多个机箱120并且跨越各自包括多个机箱120的多个机架230来聚合eSSD 110。

[0063] P-eSSD 110p是远程主机NVMe-oF驱动器170可见的唯一eSSD 110,因此终止了NVMe-oF协议。P-eSSD 110p代表其本身和同一聚合组中的所有其余S-eSSD 110s为远程主机190呈现单个、大的聚合的逻辑容量。P-eSSD 110p从远程主机NVMe-oF驱动器170接收所有输入/输出(I/O)命令180,并且将命令响应(例如,完成条目)182提供至远程主机190。

[0064] P-eSSD 110p还维护映射分配表(MAT),其指示了在eSSD 110的相同的聚合组的P-eSSD 110p以及一些或全部S-eSSD 110s之间划分的逻辑块地址(LBA)空间。当通过P-eSSD 110p接收到I/O命令180时,P-eSSD 110p首先查找MAT(例如,图3的MAT 300,下面进一步描述)以确定哪一个eSSD 110(例如,P-eSSD 110p、一个或多个S-eSSD 110s或者这两者的集合)可满足I/O命令180。根据MAT,P-eSSD 110p随后将适当修改的NVMe-oF I/O子命令132发送至适当的S-eSSD 110s的集合。

[0065] 为了发送子命令132,P-eSSD 110p在通电后还在PCIe总线140和中平面135上与S-eSSD 110s中的每一个建立专用以太网RDMA连接(或专有通信信道)130。该专用队列偶(queue-pair,QP)通信信道130被P-eSSD 110p使用来将I/O命令(例如,子命令132)发送至S-eSSD 110s,并且从S-eSSD 110s接收完成条目134。专用通信信道130可为以太网,并且可允许经以太网交换机160发送数据。然而,专用通信信道130也可为基于PCIe的信道,并且可允许经PCIe交换机发送数据。也就是说,所有eSSD 110可使用两种或更多种通信模式彼此通信是可能的。例如,以太网信道可通常用于数据的传输,并且PCIe信道可用于管理,同时任一个信道可用作专用通信信道130。

[0066] S-eSSD 110s是利用NVMe-oF协议仅执行相对于远程主机190进出的数据转移的正常NVMe-oF SSD 110。利用RDMA READ和WRITE服务相对于远程主机190直接进行这些数据转移。S-eSSD 110s从P-eSSD 110p接收命令(例如,子命令132),但是不直接从远程主机NVMe-oF驱动器170接收命令。S-eSSD 110s将子命令完成条目134发送至P-eSSD 110p而不是远程主机190,以指示子命令132的完成。

[0067] P-eSSD 110p处理所有NVMe-oF协议终止,处理所有主机命令和完成列队(例如,提交队列/完成队列(SQ/CQ)),并且对于在远程主机启动器上运行的远程主机NVMe-oF驱动器170是可见的。当远程主机驱动器170发出NVMe管理命令180或I/O命令180时,命令180被发送至P-eSSD 110p,并且所有管理命令180被P-eSSD 110p执行。然而,I/O命令180可分布在多个eSSD 110之间。

[0068] P-eSSD 110p也可根据I/O命令180执行其自身的数据转移共享。在将对应于原始命令180的命令完成条目182发送至远程主机190之前,P-eSSD 110p则等待(例如,来自S-eSSD 110s的集合的)所有子命令完成条目134到达专用通信信道130。

[0069] P-eSSD 110p还针对在命令上下文表(例如,见图6)中执行的各个命令维护“命令上下文”。该命令上下文被P-eSSD 110p使用,以跟踪子命令132的执行、数据转移和任何错误状态。当所有子命令132完成后,将命令响应/完成条目182提供至远程主机190,并且将命令上下文表解除分配。

[0070] 参照图2,多个eSSD机箱120可在以太网SSD机架230中连接在一起,其中机架顶(TOR)交换机240用于提供共同机架230中的多个机箱120之间的连接。相似地,位于不同的对应的地理位置的多个机架230可通过对应的TOR交换机240彼此连接(直接彼此连接或者通过外部交换机彼此连接)。以太网机架230可位于单个数据中心建筑物中,或者可分布在

宽地理区域上。

[0071] 总而言之，本发明的实施例提供了用于聚合多个以太网NVMe-oF SSD(eSSD)110以使其呈现为单个大容量NVMe-oF SSD的机制。eSSD110可位于单个机箱120中，可位于单个机架230中的多个机箱120中，或者甚至可散布于各自具有多个机箱120的大量以太网机架230上。为eSSD 110之一指定主要eSSD(P-eSSD)110p的角色。为其它eSSD 110指定为辅助eSSD(S-eSSD)110s。S-eSSD 110s从P-eSSD 110p接收子命令132，完成子命令132，并且将用于这些子命令132的完成条目134发回P-eSSD 110p，尽管S-eSSD 110s执行与远程主机启动器的直接数据转移。因此，当前实施例允许容量聚合有效地用作单个以太网SSD，而不牺牲任何存储带宽。

[0072] 图3是根据本公开的实施例的由P-eSSD 110p维护的“映射分配表”300的示例。图4示出了根据本公开的实施例的描绘映射分配表300的初始化的流程400。

[0073] 参照图3和图4，如上所述，当前实施例利用两种eSSD 110(例如，P-eSSD 110p和S-eSSD 110s)。P-eSSD 110p和S-eSSD 110s二者使用网上NVMe(NVMe-oF)协议，以为主机190提供存储服务。P-eSSD 110p维护包含被主机190感知为单个逻辑卷的eSSD 110的聚合组中的S-eSSD 110s的细节的表(例如，映射分配表(MAT)300)。

[0074] MAT 300可由与P-eSSD 110p在相同机箱120中的BMC 150初始化。BMC 150可管理以太网机箱120和类似以太网交换机160和eSSD 110的组件。BMC 150具有用于系统管理目的的PCIe和SMBus接口。另外，BMC将(例如，在存储管理员的指导下)确定将聚合哪些eSSD 110(S410)，并且当确定eSSD 110时，BMC 150可配置以太网交换机160。

[0075] MAT 300的左侧的三列在存储管理员的指导下被BMC 150初始化。BMC 150和存储管理员对于聚合组/存储系统中具有的所有eSSD 110可见并且知晓所有eSSD 110。这种知晓包括eSSD 110中的每一个的容量311和地址位置312。存储管理员可决定需要哪些S-eSSD 110s来形成“聚合的以太网SSD”(例如，聚合组)。BMC 150和存储管理员可向用户宣告或提供P-eSSD 110p的网络地址，以使得对应于远程主机NVMe-oF驱动器170的用户应用知道在哪里找到聚合的以太网SSD。BMC 150和存储管理员还可选择或任命eSSD 110之一作为P-eSSD 110p(S420)，并且还可在初始指派之后针对多种原因中的一个或多个来改变被指明为P-eSSD 110p的eSSD 110。然后，BMC 150可为聚合组的主要模式和辅助模式编程(S430)。

[0076] P-eSSD 110p也可在BMC 150上保持MAT 300的副本，并且可定期更新与BMC 150一起存储的MAT 300的副本。在一些实施例中，仅P-eSSD 110p包含官方MAT 300，并且eSSD索引313“0”指示P-eSSD 110p，而其余的eSSD索引值对应于各个S-eSSD 110s中的对应的一个。

[0077] P-eSSD 110p终止对于主机驱动器170的NVMe-oF协议，并且执行主机驱动器170发出的所有命令180。当主机190命令180完成时，P-eSSD 110p将完成条目182按照“完成条目”182的形式发回主机驱动器170。相对于主机命令180，远程主机NVMe-oF驱动器170完全不知道S-eSSD 110s的存在。P-eSSD 110p还维护提交队列(SQ)，并且将命令完成条目提交至完成队列(CQ)。

[0078] MAT 300右边的三列通过P-eSSD 110p更新和维护。当远程主机NVMe-oF驱动器170生成“命名空间”时，将特定闪存容量分配至命名空间。命名空间LBA范围314映射至一组eSSD 110，并且记录在通过P-eSSD 110p维护的MAT 300中。下面将参照图8描述该处理的细

节。

[0079] P-eSSD 110p还可执行特定的初始化。一旦MAT 300在p-eSSD 110p中初始化(S440),P-eSSD 110p就知道哪个eSSD 110是对应的S-eSSD 110s。P-eSSD 110p随后与聚合组中的S-eSSD 110s中的每一个设置通信信道130。通信信道130可遍及经过机箱120中的以太网交换机160的以太网接口,或者替代性地可遍及经过机箱120中的PCIe交换机的PCIe接口。如果S-eSSD 110s之一位于在相同机架230中的不同机箱120中,则通过TOR交换机240建立通信信道130。给定机箱120中的通信信道130也可遍及PCIe总线140。

[0080] 图5示出了根据本公开的实施例的NVMe-oF以太网SSD(eSSD)存储中使用的系统架构500的框图,所述NVMe-oF以太网SSD(eSSD)存储包括进出分别位于多个机架230中的多个eSSD机箱120中的多个聚合的eSSD 110的数据。

[0081] 参照图5,P-eSSD 110p可通过外部网络交换机和路由器建立与位于广域网(WAN)中的S-eSSD 110s的以太网通信信道530。这种专用以太网通信信道530可为RDMA队列偶(QP),或者可为专用方法。以太网通信信道530用于交换子命令132和关联的完成条目。

[0082] 图6是示出示例命令上下文的表600。

[0083] 参照图6,子命令132中的每一个具有命令ID 640,并且携带“命令标签”610,从而当P-eSSD 110p接收完成条目134时,完成条目134可被追溯回原始命令180。一旦追溯到子命令132的完成条目134,“子命令编号”字段620递减,并且为当前状态锁存接收到的错误状态。当子命令编号字段620达到零时,对应的命令180(子命令132的父)完成,并且P-eSSD 110p可将完成条目182发回远程主机190。此时的P-eSSD 110p产生具有积累的错误状态630的完成条目,并且将其放入关联的CQ。

[0084] 图7示出了根据本公开的实施例的示出通过P-SSD 110p处理管理命令180的流程700。

[0085] 参照图1和图7,并且如上所述,各个P-eSSD 110p维护命令SQ。当存在被P-eSSD 110P接收(S710)并能够被执行的命令180时,P-eSSD 110p仲裁SQ,并且随后选择执行命令180。P-eSSD 110p执行所有NVMe命令(例如,管理命令和I/O命令)180。也就是说,虽然S-eSSD 110s可将数据直接发送至主机190,但是S-eSSD 110s不直接从主机190接收命令,并且不直接将完成条目发送至主机190。通过P-eSSD 110p执行管理命令可不需要与任何S-eSSD 110s的任何通信。

[0086] 在接收命令180之后,P-eSSD 110P确定数据在哪儿,以及是否具有所有数据(S720)。如果P-eSSD 110p具有所有数据,则P-eSSD 110p将数据转移至主机190(S770)。

[0087] 如果P-eSSD 110p确定其不具有所有数据(S720),则P-eSSD 110p随后咨询MAT 300以确定请求的数据位于何处(S730)。一旦识别一组相关的eSSD 110,P-eSSD 110p就推进命令180的执行。当所有的相关请求的数据在P-eSSD 110p本身中可用时,P-eSSD 110p执行数据转移185。然而,当请求的数据散布在P-eSSD 110p和/或S-eSSD 110s的集合上时,P-eSSD 110p将原始命令180划分为合适数量的子命令132(S740)。子命令132的数量对应于散布有请求的数据的eSSD 110的数量。各个子命令132对应于请求的数据由各个eSSD 110占据的那部分。

[0088] P-eSSD 110p将合适的开始LBA(SLBA)、块数(NLB)和远程散布/收集列表(SGL)放在子命令132中。SGL包含远程主机190上的地址、密钥和转移缓存器的大小。P-eSSD 110p随

后在命令分割处理中将这些子命令132在专用QP通信信道130上发送至对应的S-eSSD 110s(S750),并且等待从对应的S-eSSD 110s中的每一个接收完成条目134(S760)。因此,将原始命令180分割为子命令132,以使得合适的各个eSSD 110可并行地执行数据转移,从而使得数据能够转移至主机190(S770)。

[0089] P-eSSD 110p针对参照图6描述的执行的命令180生成命令上下文。命令上下文用于保持对子命令132的执行以及子命令132的任何中间错误状态(S780)的跟踪。一旦P-eSSD 110p确认管理命令完成,则P-eSSD 110p将完成条目发送至主机190(S790)。

[0090] 因此,远程主机NVMe-oF驱动器170发送的所有管理命令180由P-eSSD 110p接收(S710)并执行。P-eSSD 110p可单独地完成一些或全部管理命令180(例如,当P-eSSD 110p在(S720)中确定其已单独具有完成管理命令所需的所有信息时)。在一些情况下,P-eSSD 110p可在完成管理命令180之前从S-eSSD 110s取回特定多条信息。如果P-eSSD 110p从S-eSSD 110s中寻找特定非用户数据信息,则P-eSSD 110p可生成管理子命令132并将其发送至对应的S-eSSD 110s。S-eSSD 110s可利用它们之间的专用通信信道130将任何必要的数据和子命令完成条目134发回P-eSSD 110p。

[0091] 图8示出了根据本公开的实施例的描绘命名空间生成和删除命令的执行的流程图800。

[0092] 参照图8,P-eSSD 110p可接收并执行命名空间生成命令(S810)和/或删除命令(S811)。当P-eSSD 110p接收命名空间生成命令时(S810),P-eSSD 110p可查找MAT 300(S820),并且可从总可用池分配合适量的容量(S830)。新生成的命名空间可具有单独来自P-eSSD 110p或者单独来自特定的S-eSSD的闪存容量,或者新生成的命名空间可具有来自P-eSSD 110p和S-eSSD 110s的任何组合的闪存容量。P-eSSD 110p随后可在MAT 300中记录分配的容量和关联映射的LBA范围(S840)。

[0093] 当P-eSSD 110p接收用于删除命名空间的命名空间删除命令时(S811),P-eSSD 110p查找MAT 300(S821),检索对应的eSSD(S831),并且解除分配关联的容量(S841),然后对应地更新MAT 300。

[0094] 相对于通过S-eSSD 110s的命名空间生成/删除命令执行,S-eSSD 110s不直接接收命名空间生成/删除命令。通常,S-eSSD 110s应该包含代表整个容量的单个命名空间。当合适时,P-eSSD 110p可将命名空间生成命令或删除命令发送至S-eSSD 110s,作为子命令132。S-eSSD 110s随后分别执行这些命令,并且将对应的完成条目134发回P-eSSD 110p。这个流程可以与从P-eSSD 110p接收的任何这种管理子命令的流程是相同的。

[0095] 图9示出了根据本公开的实施例的描绘在P-eSSD 110p的控制下执行读/写命令的流程图900。

[0096] 参照图9,P-eSSD 110p可接收并执行包括读和写命令180的所有I/O命令180。当P-eSSD 110p接收读/写命令180时(S910),P-eSSD 110p首先查找MAT 300(S920)。从MAT 300中,P-eSSD 110p识别出关联的用户数据所在的一组eSSD。

[0097] 如参照图6的描述,P-eSSD 110p随后针对原始命令生成命令上下文(S930),从而P-eSSD 110p可跟踪子命令132的执行。P-eSSD 110p随后生成对应的读/写子命令132(S940),并且将合适的子命令132发送至合适的S-eSSD 110s(S950)。P-eSSD 110p还将所有必要的传输网络相关信息(例如,地址)提供至S-eSSD 110s。作为子命令132的一部分,S-

eSSD 110s接收包含远程缓存器命令、大小和安全密钥的远程主机190SGL。

[0098] 将子命令132中的数据转移字段合适地修改至右偏移。S-eSSD 110s向远程主机190缓存器执行直接数据转移(S960)，并且当完成时，S-eSSD 110s将完成条目发送至P-eSSD 110p(不同于直接发送至主机190)。如有必要，P-eSSD 110p针对远程主机190执行其自身的数据转移共享(S960)。

[0099] 此外，S-eSSD 110s中的每一个从P-eSSD 110p接收足够的信息以执行针对远程主机190的任何直接数据转移(S960)。在NVMe-oF协议中，RDMA传输服务(RDMA读和RDMA写)用于从S-eSSD 110s至远程主机190的数据转移。远程主机190可需要支持RDMA协议的共享接收队列(SRQ)特征，以允许多个S-eSSD 110s将数据转移至远程主机190(S960)。RDMA协议可在以太网/IP/TCP(iWARP)、以太网/InfiniBand(RoCE v1)或者以太网/IP/UDP(RoCE v2)上运行。相对于S-eSSD 110s与远程主机190之间的通信，S-eSSD 110s仅严格执行与远程主机190的数据转移(S960)。也就是说，仅通过S-eSSD 110s与远程主机190执行RDMA读和RDMA写操作(S960)。通过P-eSSD 110p利用RDMA发送操作或一些其它专用协议执行子命令完成条目134和任何非用户数据转移。

[0100] 当完成所有子命令(由通过P-eSSD 110p接收到所有子命令完成条目所指示)时(S970)，P-eSSD 110p针对原始命令180生成完成条目182，并且将完成条目182发送至主机190上的合适CQ(S980)。P-eSSD 110p随后解除分配命令上下文(S990)。

[0101] 图10示出了根据本公开的实施例的描绘通过S-eSSD 110s执行管理子命令的流程图1000。

[0102] 参照图10，在当前实施例中，没有S-eSSD 110s直接从主机190NVMe-oF驱动器170接收到管理命令180或者任何命令。作为替代，P-eSSD 110p仅当必要时将管理子命令132发送至S-eSSD 110s(S1010)。S-eSSD 110s随后确定是否需要任何数据转移(S1020)，并且随后在专用通信信道130上与P-eSSD 110p执行任何需要的数据转移(S1020)。S-eSSD 110s随后生成完成条目134(S1040)并且将其发送至P-eSSD 110p(S1050)。在其它实施例中，S-eSSD 110s可使用RDMA发送操作来将数据以及完成条目134发送至P-eSSD 110p。

[0103] 图11示出了根据本公开的实施例的描绘通过S-eSSD 110s执行读/写子命令的流程图1100。

[0104] 参照图11，S-eSSD 110s主要处理读或写子命令132(例如，相对于管理子命令132)。也就是说，S-eSSD 110s主要执行相对于远程主机190进出的数据运动，而不用参与协议处理的其它方面。当S-eSSD 110s接收读/写子命令时(S1110)，S-eSSD 110s使用接收到的传输网络信息(S1120)以向远程主机190发出RDMA读或RDMA写请求(S1130)。作为子命令132的一部分，S-eSSD 110s接收远程缓存器地址/偏移、大小和安全密钥的细节。当完成必要数据转移时(S1140)，S-eSSD 110s将具有合适错误状态的完成条目134发送至P-eSSD 110p(S1150)。

[0105] 图12示出了根据本公开的实施例的描绘在S-eSSD 110S中的数据转移和子命令完成同步的流程图。

[0106] 参照图12，对于给定的主机命令180，尽管P-eSSD 110p将完成条目182发送至主机190，一组S-eSSD 110s也可执行针对主机190的数据转移。在关联的数据可导致未定义行为/错误之前，随着针对命令180的完成条目182到达主机190，在针对该给定的命令180的完

成条目182被提供至主机190之前，所有数据应该被转移至远程主机190。

[0107] 如上所述，P-eSSD 110p从主机190接收读/写命令180(S1210)，并且将命令180分割为多个子命令132(S1220)。然后，从P-eSSD 110p接收对应的子命令132的每一个S-eSSD 110s向主机190发出数据转移(S1250)，并且一旦确定数据转移完成(S1260)，S-eSSD 110s就将完成条目134发送至P-eSSD 110p(S1270)。然后，一旦从每个相关的S-eSSD 110s接收所有子命令完成条目134(S1230)，P-eSSD 110p就将完成条目182发送至主机190(S1240)。

[0108] 因为聚合的以太网SSD数据转移和完成条目公告(posting)分布在一组eSSD 110上，所以应该实现数据和完成同步。通常，当单个eSSD执行命令执行的这些阶段(数据转移+完成公告)二者时，不会出现这种问题。然而，聚合的以太网SSD并不这样。因此，在公告针对命令180的完成条目182之前，P-eSSD 110p必须等待来自对应的S-eSSD 110s的所有子命令完成条目134。此外，在将子命令132的完成条目134发送至P-eSSD 110p之前，S-eSSD 110s必须确保所有它们的数据转移均完全并可靠地完成。这种两个步骤同步处理确保了在聚合的以太网SSD系统中始终实现NVMe-oF协议完整度。

[0109] 根据以上内容，因为仅单个主要eSSD对于主机可见，并且在通过辅助eSSD跟踪所有相关子命令的完成的同时执行所有NVMe-oF协议处理，所以eSSD的聚合组对主机呈现为单个大逻辑容量。

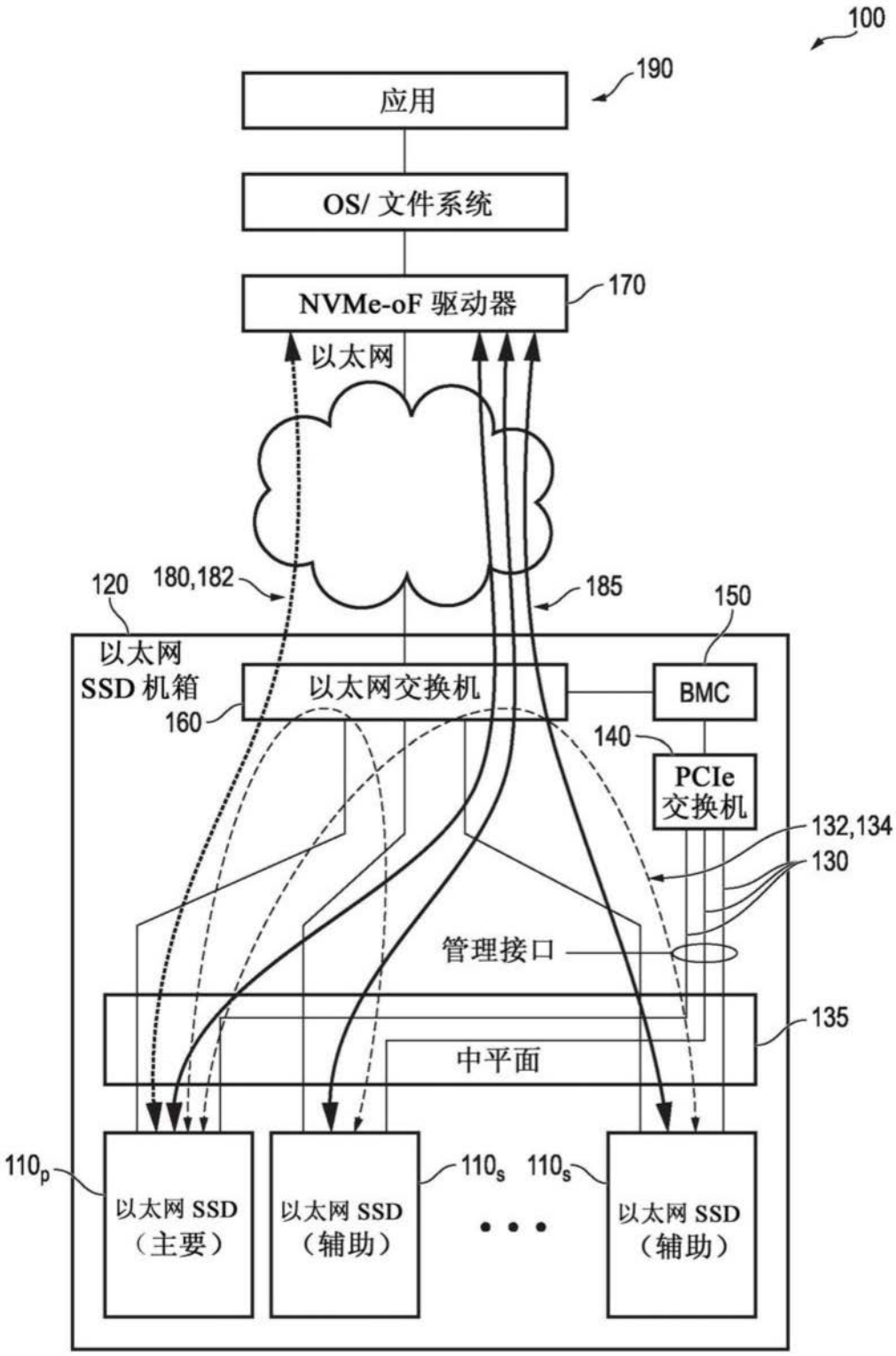

100
190
应用
OS/ 文件系统
NVMe-oF 驱动器
170
以太网
120
180,182
185
150
以太网
SSD 机箱
以太网交换机
BMC
160
140
PCIe
交换机
132,134
130
管理接口
135
中平面
110p
以太网 SSD
（主要）
以太网 SSD
（辅助）
110s
110s
以太网 SSD
（辅助）

图1

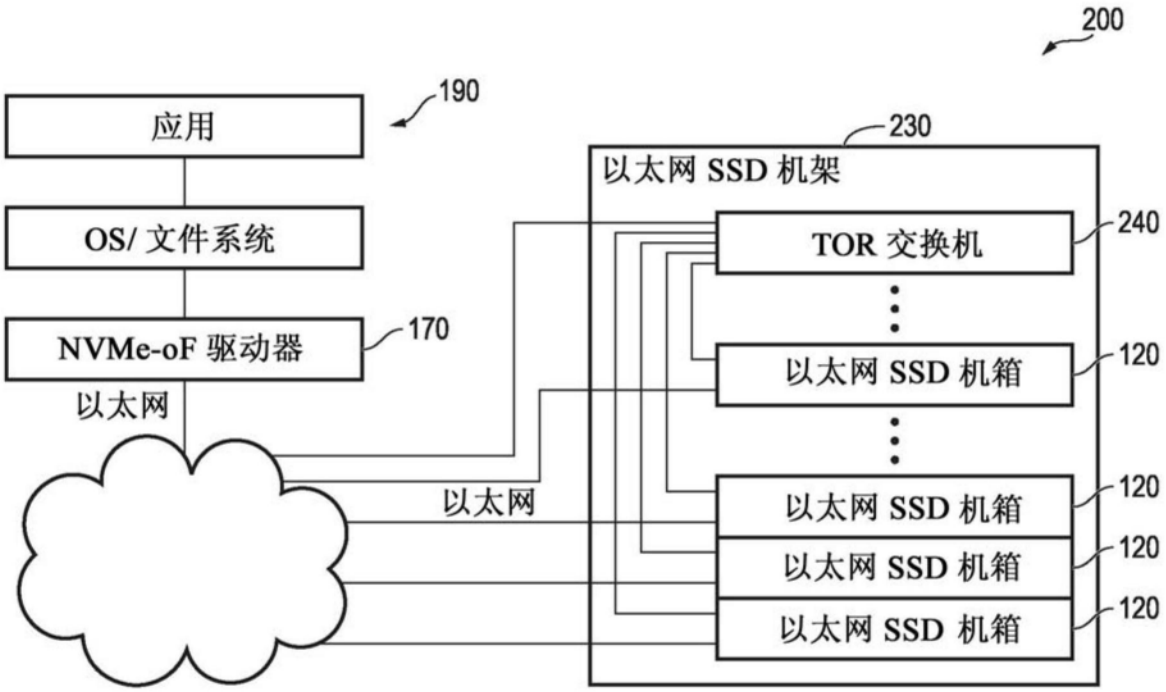


图2

300

313 eSSD 索引	311 容量（TB）	312 传输地址（MAC/IP）	314 映射地址 NS.LBA 范围	剩余容量 (TB)	其它
0	16				
1	8				
2	32				
3	16				

图3

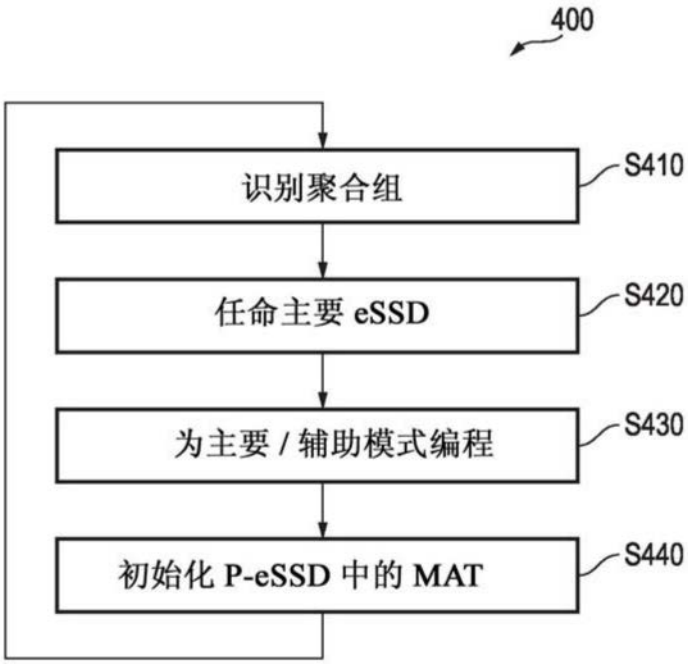

400
识别聚合组
S410
任命主要 eSSD
S420
为主要 / 辅助模式编程
S430
初始化 P-eSSD 中的 MAT
S440

图4

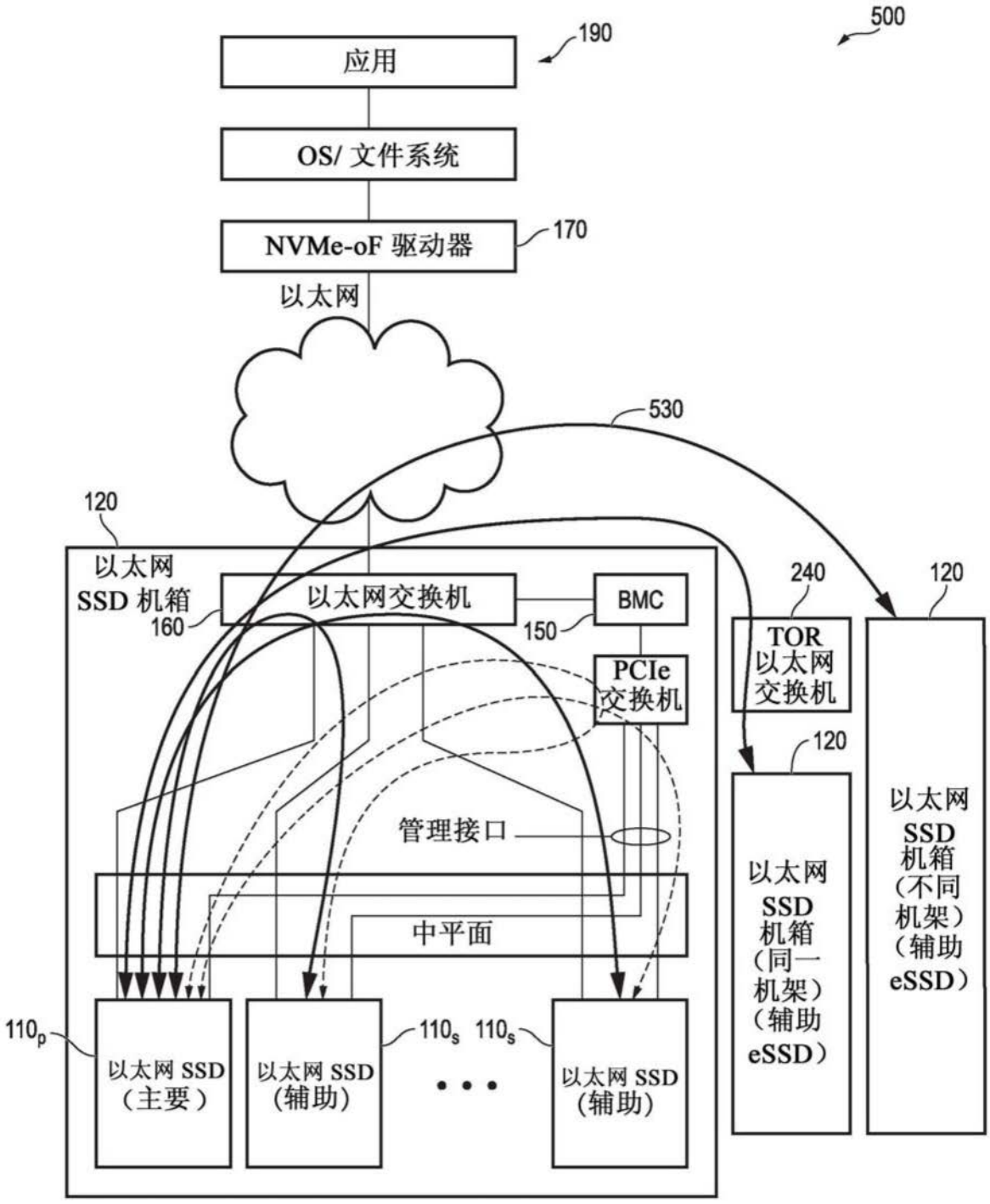

500
190
应用
OS/ 文件系统
NVMe-oF 驱动器
170
以太网
530
120
以太网
SSD 机箱
160
以太网交换机
BMC
150
PCIe
交换机
管理接口
中平面
240
TOR
以太网
交换机
120
以太网
SSD
机箱
（同一
机架）
（辅助
eSSD）
120
以太网
SSD
机箱
（不同
机架）
（辅助
eSSD）
110_p
以太网 SSD
（主要）
以太网 SSD
（辅助）
110_s
110_s
以太网 SSD
（辅助）

图5

600

640 命令 ID	610 命令标签	620 子命令编号	630 积累的错误状态
123			
15			
39			
3			

图6

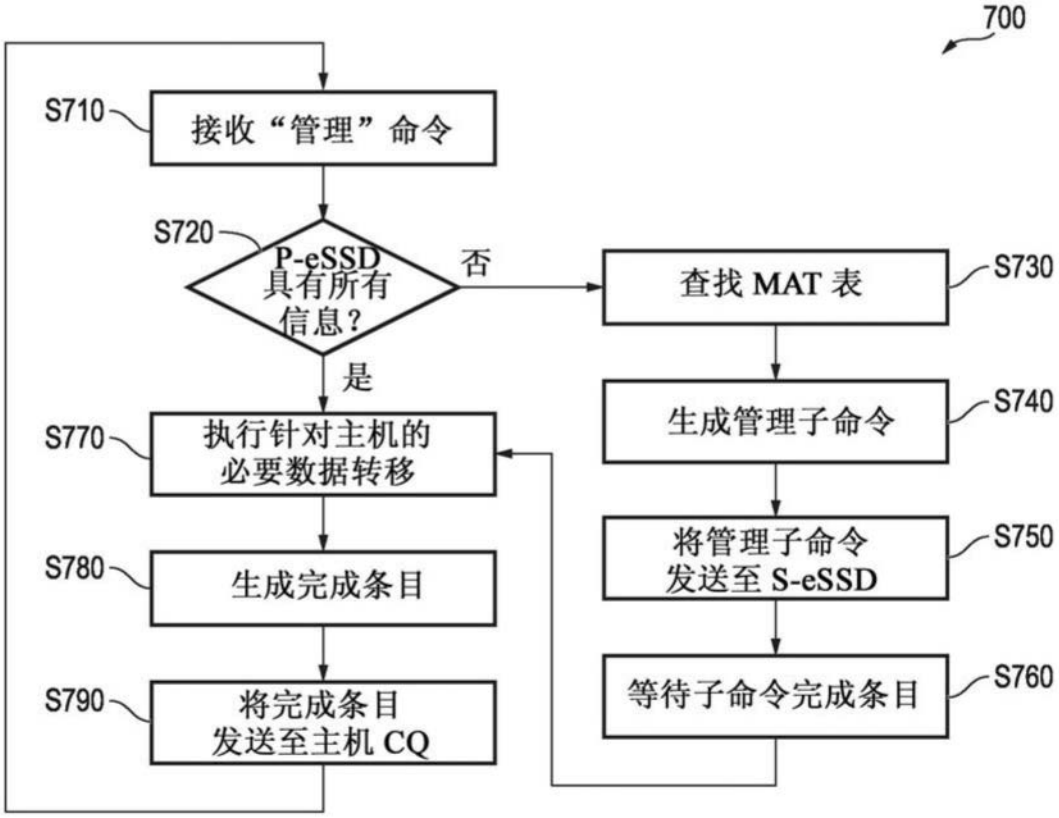


图7

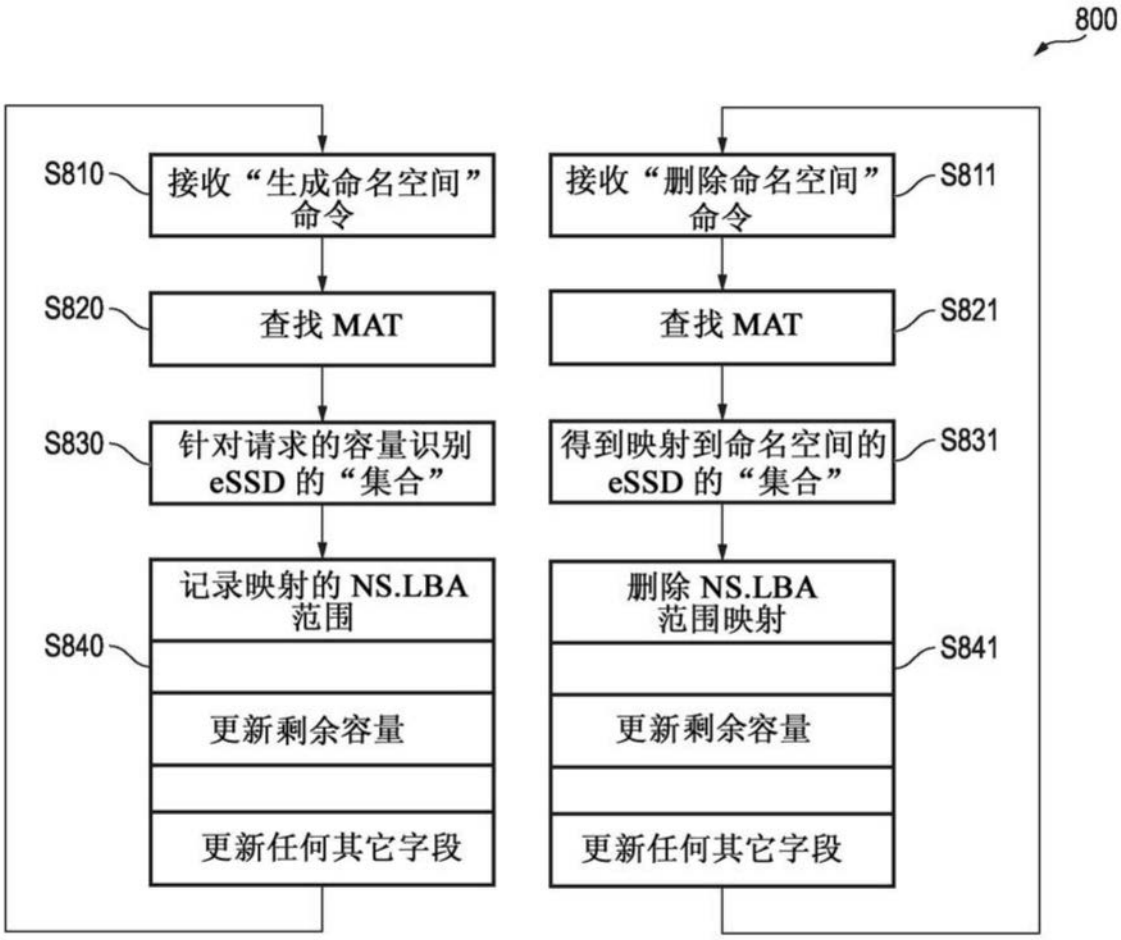

800
S810
接收“生成命名空间”命令
S820
查找 MAT
S830
针对请求的容量识别 eSSD 的“集合”
记录映射的 NS.LBA 范围
S840
更新剩余容量
更新任何其它字段
S811
接收“删除命名空间”命令
S821
查找 MAT
S831
得到映射到命名空间的 eSSD 的“集合”
删除 NS.LBA 范围映射
S841
更新剩余容量
更新任何其它字段

图8

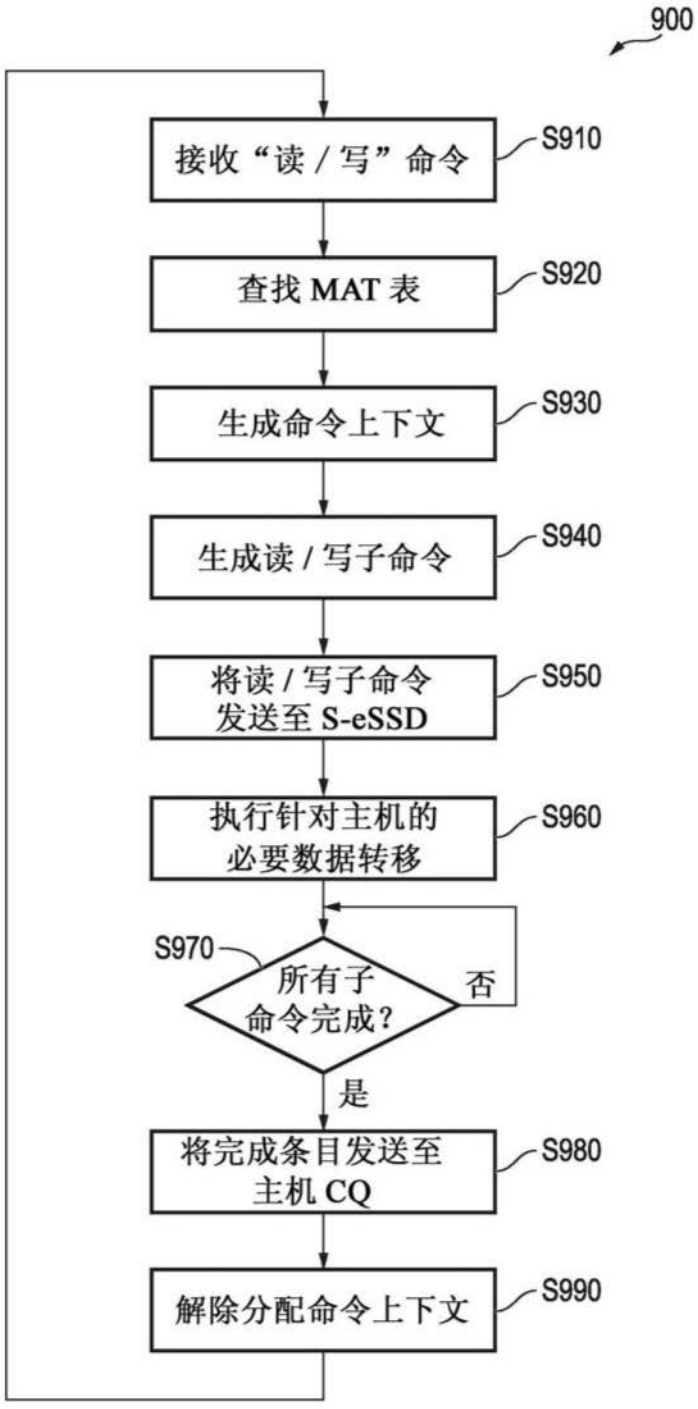

900
接收“读 / 写”命令
S910
查找 MAT 表
S920
生成命令上下文
S930
生成读 / 写子命令
S940
将读 / 写子命令
发送至 S-eSSD
S950
执行针对主机的
必要数据转移
S960
S970
所有子
命令完成?
否
是
将完成条目发送至
主机 CQ
S980
解除分配命令上下文
S990

图9

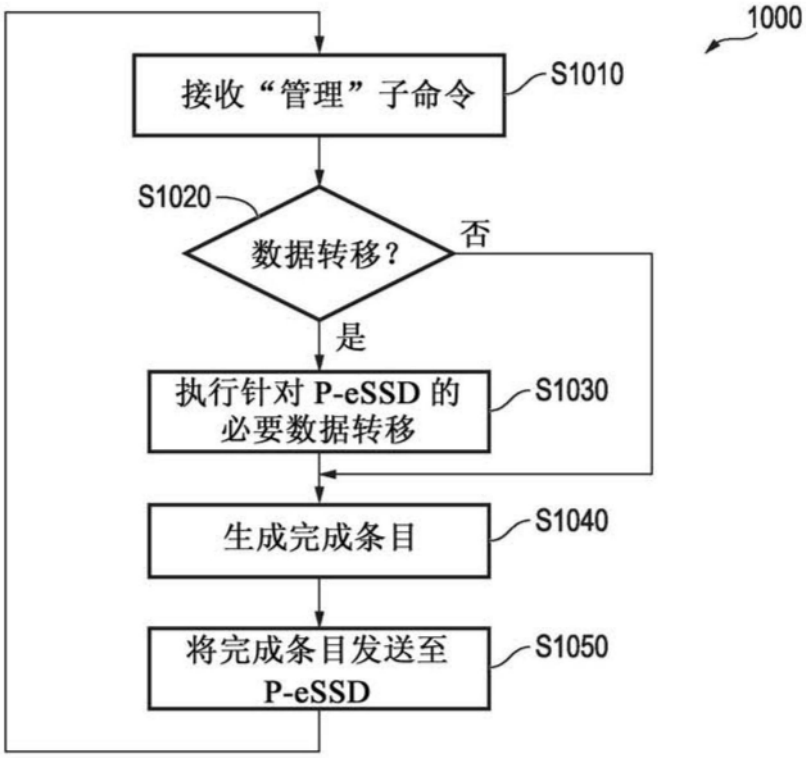

1000
接收“管理”子命令
S1010
S1020
数据转移?
否
是
执行针对 P-eSSD 的必要数据转移
S1030
生成完成条目
S1040
将完成条目发送至 P-eSSD
S1050

图10

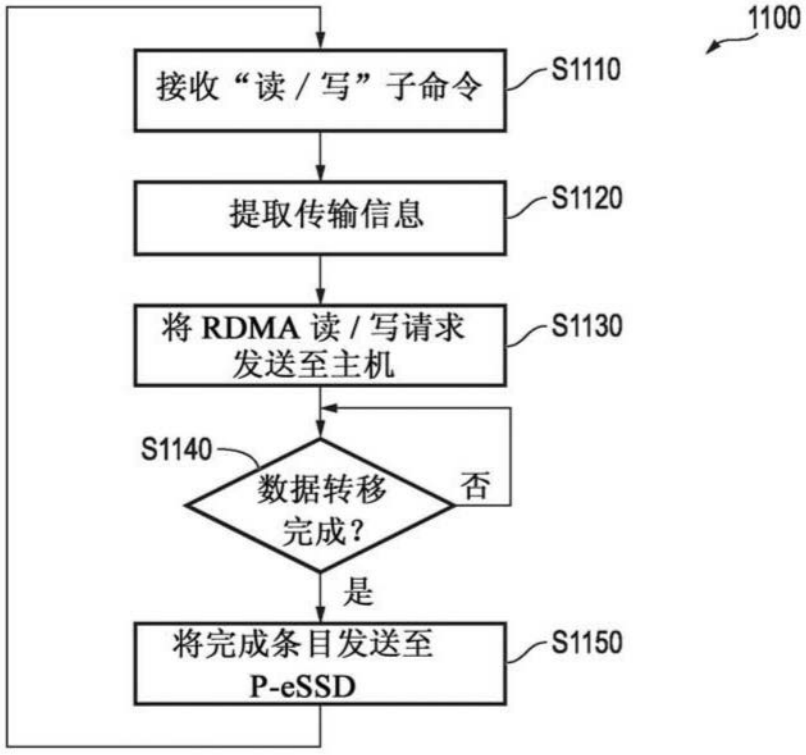

1100
接收“读/写”子命令
S1110
提取传输信息
S1120
将 RDMA 读/写请求发送至主机
S1130
S1140
数据转移完成?
否
是
将完成条目发送至 P-eSSD
S1150

图11

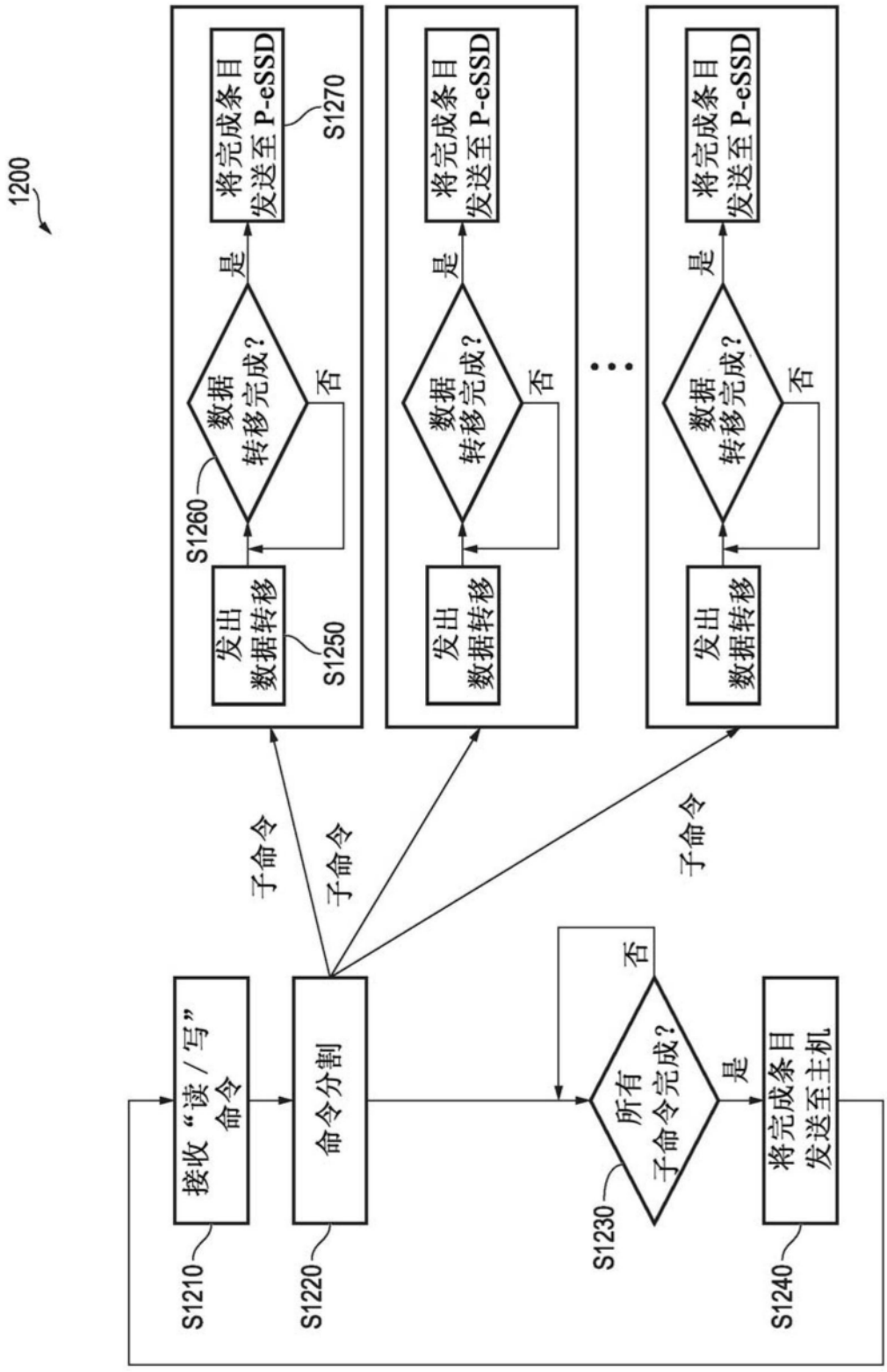

1200
接收“读/写”命令
S1210
命令分割
S1220
子命令
子命令
子命令
发出数据转移
S1250
数据转移完成?
S1260
是
否
将完成条目发送至 P-eSSD
S1270
发出数据转移
数据转移完成?
是
否
将完成条目发送至 P-eSSD
发出数据转移
数据转移完成?
是
否
将完成条目发送至 P-eSSD
所有子命令完成?
S1230
否
是
将完成条目发送至主机
S1240

图12