(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112115090 A
(43)申请公布日 2020.12.22

(21)申请号 202010229307.4

(22)申请日 2020.03.27

(30)优先权数据
62/864252 2019.06.20 US
16/827460 2020.03.23 US

(71)申请人 英特尔公司
地址 美国加利福尼亚州

(72)发明人 杨子夜 刘长鹏 曹刚 万群

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 李啸 姜冰

(51)Int.Cl.
G06F 13/42(2006.01)
G06F 13/16(2006.01)
H04L 29/06(2006.01)

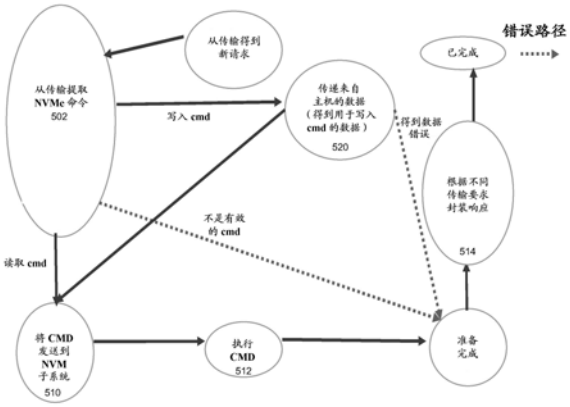
权利要求书3页 说明书16页 附图10页

(54)发明名称

用于事务的多协议支持

(57)摘要

本文中描述的示例涉及执行轮询器以轮询在多个传输层协议上来自连接的所接收到的通信,以根据多个传输层协议中的一个传输层协议识别所接收到的通信,并且根据多个传输层协议中的不同的一个传输层协议识别第二所接收到的通信。响应于多个传输层协议中的一个传输层协议的失败或者使用多个传输层协议中的一个传输层协议的慢传输速率,发生对多个传输层协议中的不同的一个传输层协议的改变。在一些示例中,轮询器在用户空间中被执行,并且所接收到的通信和第二所接收到的通信的传输层协议处理发生在内核空间中。



CN 112115090 A

1.一种非暂时性计算机可读介质,包括存储在其上的指令,所述指令如果由一个或多个处理器执行,则使得所述一个或多个处理器:

执行轮询器以:

轮询在多个传输层协议上来自连接的所接收到的通信,以根据所述多个传输层协议中的一个传输层协议识别所接收到的通信,并且根据所述多个传输层协议中的不同的一个传输层协议识别第二所接收到的通信。

2.根据权利要求1所述的计算机可读介质,其中所接收到的通信包括构造上高速非易失性存储器(NVMe-oF)兼容命令。

3.根据权利要求1所述的计算机可读介质,其中响应于所述多个传输层协议中的所述一个传输层协议的失败或者使用所述多个传输层协议中的所述一个传输层协议的慢传输速率,发生对所述多个传输层协议中的所述不同的一个传输层协议的改变。

4.根据权利要求1所述的计算机可读介质,其中所述轮询器在用户空间中被执行并且包括存储在其上的指令,所述指令如果由一个或多个处理器执行,则使得所述一个或多个处理器执行内核空间中的所接收到的通信和第二所接收到的通信的传输层协议处理。

5.根据权利要求1所述的计算机可读介质,包括存储在其上的指令,所述指令如果由一个或多个处理器执行,则使得所述一个或多个处理器:

配置由所述轮询器应用的服务质量,以轮询从所述多个传输层协议之中所接收到的通信。

6.根据权利要求1所述的计算机可读介质,包括存储在其上的指令,所述指令如果由一个或多个处理器执行,则使得所述一个或多个处理器:

执行所述轮询器以针对所执行的命令从存储装置轮询完成指示,其中所述轮询器将基于所配置的服务质量来轮询完成指示。

7.根据权利要求1所述的计算机可读介质,其中所述传输层协议包括以下中的一个或多个:远程直接存储器存取(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议(UDP)、或快速UDP因特网连接(QUIC)。

8.根据权利要求1所述的计算机可读介质,包括存储在其上的指令,所述指令如果由一个或多个处理器执行,则使得所述一个或多个处理器:

对应用于之前第一传输层协议的第二传输层协议应用限制。

9. 根据权利要求1所述的计算机可读介质,包括存储在其上的指令,所述指令如果由一个或多个处理器执行,则使得所述一个或多个处理器:

从存储装置接收响应,以及

使用所述多个传输层协议中的一个传输层协议来生成具有对启动器的所述响应的分组。

10. 一种用于使用至少两个不同传输层协议中的一个或多个来提供构造上高速非易失性存储器(NVMe-oF)兼容服务的方法,所述方法包括:

执行轮询器以从多个不同传输层协议之中轮询存储命令的接收,以及

使得协议处理使用传输层协议所接收到的并且由所述轮询器所检测到的命令。

11.根据权利要求10所述的方法,包括:

许可使用不同传输层协议来同时访问NVMe兼容存储装置。

12.根据权利要求10所述的方法,其中所述轮询器在用户空间中被执行,并且所述协议处理在内核空间中被执行。

13.根据权利要求10所述的方法,其中执行轮询器以从多个不同传输层之中轮询存储命令的接收应用每传输层类型的服务质量。

14.根据权利要求10所述的方法,包括:

将所接收到的命令的地址转译成在存储装置处可访问的逻辑块地址。

15. 根据权利要求10所述的方法,包括:

轮询对所接收到的命令的响应,以及

提供所述响应以用于使用传输层协议的传输。

16. 根据权利要求10所述的方法,包括:

从启动器接收通信,所述启动器从使用第一传输层协议改变成使用第二传输层协议,以及

对应用于所述第一传输层协议的所述第二传输层协议应用限制。

17.根据权利要求10所述的方法,其中所述传输层协议包括以下中的一个或多个:远程直接存储器存取(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议(UDP)、或快速UDP因特网连接(QUIC)。

18.一种系统,包括:

网络接口;

耦合到所述网络接口的存储器;以及

耦合到所述存储器的第一处理器,所述第一处理器被配置成:

执行第一轮询器以轮询使用第一传输协议和第二传输协议所接收到的接收到的存储命令,以及

耦合到所述存储器的第二处理器,所述第二处理器被配置成:

执行第二轮询器以轮询使用所述第一传输协议和所述第二传输协议所接收到的接收到的存储命令。

19. 根据权利要求18所述的系统,其中所述第一处理器将:

应用所述第一轮询器的服务质量以轮询使用第一传输协议和第二传输协议所接收到的接收到的存储命令,以及

其中所述第二处理器将:

应用所述第二轮询器的服务质量以轮询使用第一传输协议和第二传输协议所接收到的接收到的存储命令。

20. 根据权利要求18所述的系统,其中所述第一处理器将:

执行所述第一轮询器以轮询存储命令的完成,以及

使得使用所述第一传输协议或第二传输协议向远程装置传输响应,以及

其中所述第二处理器将:

执行所述第二轮询器以轮询存储命令的完成,以及

使得使用所述第一传输协议或第二传输协议向远程装置传输响应。

21.根据权利要求18所述的系统,其中所述传输协议包括以下中的一个或多个:远程直接存储器存取(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议

(UDP)、或快速UDP因特网连接(QUIC)。

用于事务的多协议支持

[0001] 相关申请

本申请要求保护2019年6月20日提交的序列号为62/864252的美国临时专利申请的优先权日期的权益,其全部公开通过引用并入本文中。

背景技术

[0002] 存在各种存储协议,其使得能够使用网络或构造来访问存储装置。例如,构造上NVMe(NVMe-oF)规范被设计成使得能够访问远程NVMe兼容固态驱动器(SSD)。例如,NVMe-oF至少在NVM高速基础规范修订版1.4(2019)中被描述。NVMe-oF兼容装置向可通过网络或构造访问的远程系统提供高性能NVMe存储驱动器。

附图说明

[0003] 图1描绘兼容NVMe-oF和因特网小型计算机系统接口(iSCSI)协议的系统的示例。

[0004] 图2描绘其中主机可以使用连接与一个或多个存储装置通信的系统的示例。

[0005] 图3描绘具有在NVMe兼容SSD中启用的TCP和RDMA传输的NVMe-oF服务设计的示例。

[0006] 图4A描绘用于与NVMe-oF服务经由不同传输的连接提供可靠性的示例节点。

[0007] 图4B描绘图4A的系统的示例操作。

[0008] 图5示出传输层处理的图。

[0009] 图6A描绘用于配置目标存储节点以使用一个或多个传输类型进行通信的示例过程。

[0010] 图6B描绘可以由节点用于与另一个节点或主机装置进行通信的示例过程。

[0011] 图7描绘系统。

[0012] 图8描绘环境。

[0013] 图9描绘网络接口。

具体实施方式

[0014] 高速非易失性存储器(NVMe)规范描述了用于通过高速外围组件互连(PCIe)端口访问数据存储系统的系统。NVMe被描述于例如NVM Express™基础规范修订版1.3c(2018)以及其前驱、后继和专有变体,其整体通过引用被并入。NVMe允许主机装置将存储区指定为单独名称空间(namespace)。名称空间可以是非易失性存储器中的可寻址域,所述非易失性存储器具有已被格式化用于块访问的所选数量的存储块。名称空间可以包括固态驱动器(SSD)中的介质的可寻址部分,或者跨越多个SSD或其它数据存储装置的多装置存储器空间。名称空间ID可以是相关联的名称空间的唯一标识符。主机装置可以通过指定NSID、控制器ID和一个或多个块的相关联的逻辑地址(例如,逻辑块地址(LBA))来访问特定的非易失性存储器。

[0015] 分布式块存储系统通过呈现存储在跨远程存储装置的大的池(pool)潜在地分散(scatter)的段(segment)中的逻辑块装置,向应用提供块装置功能性。为了使用这些逻辑

块装置,应用确定要访问的所有段的位置。计算平台可以使用构造或网络来访问存储装置。用于使用构造或网络访问存储的示例方案包括NVMe-oF或者构造或网络规范上的其它专有存储。

[0016] 图1描绘兼容NVMe-oF和因特网小型计算机系统接口(iSCSI)协议的系统的示例。随着NVMe兼容固态驱动器(SSD)的出现,用于允许远程主机装置访问SSD的现有iSCSI协议的再使用可以涉及协议转换100(例如,NVMe到SCSI到iSCSI以及iSCSI到SCSI到NVMe)。NVMe-oF协议可以通过为NVMe事务提供各种传输层来缓解这种问题,所述NVMe事务诸如远程直接存储器访问(RDMA)、光纤通道(FC)以及最近的传输控制协议(TCP)。

[0017] 在2018年,为NVMe-oF发布了涉及TCP传输(例如,TP-8000 NVMe-oF™ TCP传输结合)的规范。T-8000的目的是用于再使用现有的联网设施,因为传统的网络接口控制器(NIC)和交换机可能不支持远程直接存储器访问(RDMA)相关的协议。利用TCP传输支持,再使用现有的网络设施可以发生,从而缓和从iSCSI到NVMe-oF的转换。因此,存在由NVMe-oF定义并支持的若干众所周知的传输层协议(例如RDMA、FC和TCP)。开发者可以根据本说明书使用一个或多个不同传输来实现NVMe-oF目标或启动器(initiator),用于向远程主机(启动器)呈现NVMe子系统。

[0018] 通常,可以经由不同传输提供单个NVMe-oF服务以避免使NVMe-oF服务禁用的单个传输协议层失败。换句话说,通过经由不同传输提供单个NVMe-oF服务可以改进服务可靠性,因为如果一个传输不起作用,则可以通过另一个传输提供NVMe-oF服务。然而,不同构造传输提供不同的性能。例如,RDMA传输可以提供比TCP传输快得多的传输速度。而存储事务经受服务级别协议(SLA)并且将在规定的时间量内完成。

[0019] 由内核提供用于管理传输的选择的一些技术方案。例如,Linux操作系统内核可以提供Linux I/O目标(LIO)服务,以经由不同传输导出(export)NVMe-oF服务。当前,Linux内核模式支持通过NVMe-oF的RDMA和TCP。例如,如果利用nvmet_rdma和nvmet_tcp启用内核,则RDMA传输和TCP传输两者都可用于NVMe-oF服务。

[0020] 基于内核的技术方案经由多个传输提供可靠的服务可能是具有挑战性的。如果传输依赖于内核模块并且内核崩溃(crash),则即使服务被部署在虚拟机中,服务恢复中也可能存在延迟。此外,如果内核中的TCP或RDMA传输实现不稳定,则内核可能发生故障(malfunction)并且可能影响整个操作系统(OS)。

[0021] 在目标装置处,NVMe-oF命令的执行可以包括:

1. 接收从基于传输层的请求提取的NVMe-oF命令。传输层可以是开放系统互连(OSI)通信模型的第4层。传输层可用于确保消息跨网络的可靠到达,并且提供错误检查机制和数据流控制。

[0022] 2. 解析命令。如果命令是NVMe命令,则将命令发送到底层NVMe驱动程序(driver),分配共享数据缓冲器(如果命令使用缓冲器的话),并且将命令发送到提交队列(SQ)。可选地,可以增加当前队列深度(QD)以便管理所有空闲的NVMe命令槽。

[0023] 3. 完成NVMe命令(例如,在目标介质处执行写入或读取),并且在完成队列(CQ)中提供NVMe响应(例如,RSP,16字节)。系统可以将NVMe响应映射到NVMe命令。如果读取相关的数据,则将数据放入与NVMe命令相关联的分配的共享缓冲器中。

[0024] 4. NVMe响应和数据(用于读取)可以根据传输层响应定义来封装,并且发送到主

机(启动器)。

[0025] 图2描绘其中主机可以使用连接与一个或多个存储装置通信的系统的示例。主机或启动器200可以与一个或多个目标子系统204-0至204-a通信,其中a是大于或等于2的整数。例如,主机或启动器200可以执行具有操作系统和应用的虚拟机或容器。NVMe-oF主机或启动器200可以发现NVMe-oF存储事务使用一个或多个传输协议是可用的。可以至少为了连接可靠性或用于减少与远程存储或存储器装置或池的存储事务的等待时间而提供传输协议选择或切换。例如,根据其自身应用的策略,如果与NVMe子系统的连接被破坏或太慢(例如,不可接受的等待时间),则主机或启动器200将利用其自身配置的顺序经由另一个传输层与目标连接。例如,RDMA可以是默认协议,随后是FC和TCP,尽管可以使用传输层选择的另一个顺序。NVMe子系统在NVMe规范中定义。

[0026] NVMe子系统可以在不同传输上与不同传输IP地址和服务ID(例如,端口)一起注册。因此,用户(例如,应用、虚拟机、容器)可以经由不同传输访问NVMe子系统。因此,如果单个传输失败或太慢,则另一个传输可以用于NVMe子系统,因为至少两个传输被配置用于一个子系统。

[0027] 在主机200处的非RDMA兼容NIC和任何接口202-0至202-a可以用于支持一些实施例。在一些情况下,对于使用TCP传输的NVMe-oF,可以利用将TCP操作卸载到网络接口。

[0028] 目标系统204-0至204-a可以使用具有服务质量和保留系统的相应多传输层轮询器(poller)206-0至206-a,其根据服务质量方案轮询在多个传输协议上的命令的接收,并且应用本文中描述的可应用的保留方案。目标系统204-0至204-a可以至少包括NVMe兼容子系统,其具有一个或多个控制器以管理对定义所连接的存储装置(例如,固态驱动器(SSD))中的地址范围的名称空间标识符(NSID)的读取/写入请求。NVMe子系统可以使用PCIe总线与附着的存储装置通信。

[0029] 连接250可以是网络、互连、构造、PCIe接口、总线或其它有线或无线介质。

[0030] 图3描绘具有在NVMe兼容SSD中启用的TCP和RDMA传输的NVMe-oF服务设计的示例。启动器(未示出)可以使用具有目标的NVMe发现服务,以确定利用目标支持多个传输层。在该示例中,NVMe-oF目标支持RDMA和TCP传输两者,但是每个传输独立地起作用。在一些示例中,目标可以利用与支持NVMe-oF有关的存储性能开发工具包(Kit)(SPDK)的特征。在图3的系统中,对于每个传输类型,存在中央处理单元(CPU)核302-1至302-n上执行的对应轮询组(polling group)。轮询组被定期性地执行以搜索关于其传输类型的所接收到的命令。例如,如果存在两个传输,例如RDMA和TCP传输,则存在在不同的CPU核302-1至302-n上运行的独立的两个轮询组。每个传输具有其自身的资源管理(例如,存储器资源),并且可以不执行在传输类型之间的平衡存储器管理。

[0031] 每个CPU核302-1至302-n可以运行针对轮询组的专用线程,以监测传输类型的端口并维护用于处置所接收到的NVMe命令的数据缓冲池(buffer pool)。可以为每个传输协议层类型分配轮询组,以轮询所接收到的命令或完成的事务。但是该方案在传输之中不提供优先级,并且在传输类型之中不支持服务质量(QoS)。如果存在传输失败,则可能不存在用于处置传输失败的策略。例如,如果RDMA传输失败或经历高等待时间,则不提供后动作(post action)来在不同传输上恢复NVMe-oF连接,因为传输是独立处置的。

[0032] 在一些情况下,使用传输协议层所接收到的命令和数据被分配给专用缓冲池。如

果NVMe子系统通过多个传输导出，则可能不存在保留支持，以防止对存储器缓冲器同时访问的资源竞争，所述存储器缓冲器存储从主机或其它装置所接收到的、或将被提供给主机或其它装置的通信或数据。此外，因为存在访问NVMe子系统的多个传输，所以在NVMe子系统处在经由传输的请求之中可能存在竞争，使得不同的请求可能请求不一致的动作(例如，不同的写入操作)。

[0033] 将处理器的虚拟存储器空间划分成内核空间和用户空间。应用的代码在用户空间中运行。如果应用请求内核执行高特权服务，则应用对内核空间进行系统调用。如果NVMe驱动程序功能性在Linux内核中实现，那么对于在NVMe装置上读取/操作的用户空间应用而言，应用向内核发出系统调用，这引入来自内核空间和用户空间之间的上下文切换以及数据复制的延时。如果NVMe-oF目标服务在内核空间中，则可能出现各种问题。例如，(1)在多个线程之中具有大量锁定和资源竞争的情况下，在内核I/O栈中存在用于NVMe命令执行的长I/O栈，或(2)如果用户(例如，云存储提供商)使用这样的NVMe-oF服务，则他们需要升级他们的内核或者将NVMe-oF相关的新特征向后移植(backport)到内核中(例如，用于支持目标侧中的TCP传输的nvmet_tcp模块)。

[0034] 各种实施例提供至少一些支持的传输到单个用户空间服务中。各种实施例提供用户空间技术方案，以经由不同传输使用轮询组来启用NVMe-oF服务。例如，NVMe命令可以在不同传输(例如，远程直接存储器访问(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议(UDP)、快速UDP因特网连接(QUIC)等)中的一个上被提供。各种实施例可以提供可靠的服务，由此NVMe子系统服务经由不同传输被提供给主机，并且如果一个传输失败或被认为太慢，则客户端可以切换到另一个传输。各种实施例提供具有细粒度传输资源管理和调度的统一传输管理，以从不同传输提供服务质量(QoS)。

[0035] 各种实施例为所有可用或利用的传输(或可用或利用的传输的子集)提供NVMe-oF事务的单个轮询组。线程模型可用于在一个或多个核或处理器上执行单个统一轮询组。轮询组可以以循环(round robin)方式在传输之中轮询。此外，服务质量(QoS)可以被应用于根据优先级来调度传输的轮询。可以通过配置文件或使用远程过程调用(RPC)来设置用于处置不同传输的QoS策略，以在NVMe-oF目标服务的运行期间动态地改变QoS策略。与具有其自身的轮询组的传输的技术方案相比，在一个或多个核上使用单个轮询组可以更平衡来自不同传输的NVMe命令，并且通过使用共享缓冲池而不是用于不同传输的池来简化资源管理。

[0036] 图4A描绘用于为与NVMe-oF服务经由不同传输的连接提供可靠性的示例节点。节点可以用于一个或多个NVMe兼容存储装置的NVMe集群中。例如，节点可以使用接口402用于与启动器装置经由网络、互连、构造、互连或其它通信路径的通信。从接口402所接收到的分组可以由可应用的传输层处理器404-0至404-n处理。传输层处理器404-0至404-n可以执行用于RDMA、TCP、FC或包括用户数据协议(UDP)、QUIC等的其它传输层的传输层处理。传输层处理可以使用特权操作或受保护的存储器区域在内核空间中执行。内核空间可以是由特权操作系统内核、内核扩展和一些装置驱动程序保留的存储器区域。用户空间可以是在其中应用软件和一些驱动程序执行并写入或读取内容的存储器区域。

[0037] 共享存储器缓冲池410可用于在各种传输上从启动器所接收到的命令或分组。然而，各种实施例可提供用于一个或多个传输的缓冲器。多个CPU核406-0至406-1可以共享对

缓冲池410的访问。例如，在轮询组线程的开始时，可以在缓冲池410中预分配数据缓冲池(例如，64个缓冲器)。例如，如果存在从传输提取的具有读取类型的NVMe命令，则可以在池410中分配缓冲器。对于提供给NVMe驱动程序418的NVMe读取命令，将分配的缓冲器的地址封装在NVMe命令中，并且在读取操作之后，可以将从NVMe装置读取的数据复制到池410中的缓冲器。

[0038] CPU核上的线程可以使用从启动器所接收到的多个传输来执行通信的轮询组。例如，CPU核406-0上的线程可以执行轮询组408-0，以轮询由传输层处理器404-0至404-n中的任何一个传输处理器处理的分组的接收。当通过传输的连接被接受时，可以使用选择方案(例如，循环)来将连接处置分派到指定的线程/CPU核，以处置将连接添加到指定的线程/CPU核上的轮询组。线程/CPU核上的轮询组用于轮询来自轮询组中的一个或多个连接的命令。轮询组可以与一个或多个连接相关联，并且可以与一个启动器或不同启动器形成连接。其它CPU核可用于执行轮询组，所述轮询组轮询由传输层处理器404-0至404-n中的任何一个传输层处理器处理的分组的接收。在该示例中，CPU核406-1上的线程可执行轮询组408-1，以轮询由传输层处理器404-0至404-n中的任何一个传输层处理器处理的分组的接收，但可使用任何其它数量的核或线程。在一些示例中，轮询组408-0和408-1可以在用户空间中实现。

[0039] 例如，轮询组可执行寄存器或存储器空间的读取以确定新的I/O事务是否可用于处理。如果新的I/O事务可用于处理，则轮询组可以完成用于I/O事务的描述符，并且处理新的I/O事务，并且然后从VM向I/O提交器(submitter)(例如，上层应用)通知I/O事务的完成状态。轮询组可以通过执行协议处置、缓冲器管理或与硬件装置交互来处理I/O事务。执行协议处置可以包括提取NVMe命令的命令类型以及准备来自以太网封装的I/O请求。缓冲器管理可以包括准备用于NVMe SSD装置的缓冲器，并且还准备用于从以太网NIC装置的分组接收的数据缓冲器。与存储或网络装置的交互可以包括向NVMe SSD、NVMe-oF远程装置或网络装置发送I/O请求。

[0040] 轮询组可以根据可应用的服务质量(QoS)来应用跨不同传输层处理器404-0至404-n的轮询的优先级。例如，应用的QoS可以包括使用轮询组，以基于至少一些传输层的最小或最大每秒输入/输出操作(IOP)或最小或最大带宽的轮询速率，轮询来自不同传输的NVMe命令。QoS还可以牵涉根据优先级针对每个传输层类型的NVMe命令执行的最小或最大数量。可以至少基于用于传达存储命令的传输层来使用其它QoS方案。

[0041] NVM子系统412-0和412-1可以被分配供名称空间的范围上使用。可以使用其它数量的NVM子系统。在一些示例中，由NVM子系统管理的名称空间的范围是不重叠的。NVM子系统412-0和412-1可以对不同的名称空间应用不同的规则。NVM子系统412-0和412-1可以解封装NVMe命令并且应用访问控制。例如，NVM子系统412-0和412-1可以使用相应的多个传输访问控制414-0和414-1，来防止在不同传输层上发送的事务而引起的冲突。例如，访问控制可以通过不同传输避免不同的NVMe命令之中的读取/写入冲突。多个传输访问控制414-0和414-1可以使用例如NVMe保留注册命令(例如，NVMe规范修订版1.3c(2018)中的6.11节)来配置。保留注册命令用于注册、注销或替换保留密钥。保留注册命令可与名称空间相关联。多个传输访问控制414-0和414-1可以管理从一个传输协议被应用到用于相同连接或启动器的另一个传输协议的限制。启动器客户端服务器或虚拟机(VM)具有其自身的主机标识

符,如NVMe规范中所定义的那样。

[0042] 例如,如果启动器使用保留注册命令来保留存储区域,并且使用RDMA传输层来写入区域。如果TCP传输层用于向区域写入,则如果RDMA传输被中断、断开或改变,则TCP传输层继承RDMA传输层中应用的限制,并且使用TCP传输层传送的命令经受限制。限制可以例如正在防止从区域读取或正在防止对区域的写入操作。其它示例包括缓解经由不同传输的请求之中的竞争,使得不同的请求可以请求不一致的动作(例如,不同的写入操作),所述不一致的动作可以根据策略来解决。

[0043] 例如,如果启动器使用保留注册命令来保留存储区域,并且使用RDMA传输层来写入区域。特权和访问角色对于其它传输(例如,TCP传输)而言将是相同的,因为目标经由NVMe规范中定义的128位“主机标识符”来标识不同的主机。通过来自使用相同主机标识符的相同主机的不同传输的连接经受相同的特权。如果RDMA传输被中断或断开,并且相同的主机使用TCP传输层来向存储区写入,则TCP传输层继承RDMA传输层中应用的限制,并且使用TCP传输层传送的命令经受相同的限制。

[0044] 限制可以例如正在防止从来自另一个主机的定义区域(例如,子系统中的名称空间1)读取或正在防止来自其它主机对该区域的写入操作。限制可以缓解经由不同传输来自不同主机的请求之中的竞争,使得在不同的请求将导致不一致的动作(例如,不同的写入操作)的情况下,可以根据策略来解决动作,而不管由主机使用哪个传输。

[0045] 另一个示例可以是,存在两个主机A和B,并且A和B从目标访问子系统C中的相同名称空间2。在这种场景下,主机A是名称空间#2的“持有者”,而主机B是“注册者”,主机A将访问类型设置成“写入独占”,而主机B只能读取这个名称空间,直到A释放注册或改变访问类型为止。如果主机A正在使用RDMA传输,则主机B正在使用FC传输,并且主机A改变成使用TCP传输,则主机A仍然是名称空间#2的“持有者”,并且管理其它角色的访问特权,并且读取/写入策略仍然由其角色控制,直到主机A注销保留为止。无论传输B正在使用什么,主机B都不能向名称空间#2写入。

[0046] 在一些示例中,多个传输访问控制414-0和414-1可以基于QoS或预协商的服务级别协定(SLA)来控制轮询组针对通过每个传输发送的通信的轮询速率。多个传输访问控制可以挑选传输或轮询速率来使用,以满足QoS或SLA要求。

[0047] NVMe或NVMe-oF子系统中的多个传输访问控制414-0和414-1将不同的命令(例如,读取或写入)从客户端或虚拟机或容器路由到不同的名称空间。多个传输访问控制414-0和414-1可以在用户空间中操作。

[0048] NVM子系统412-0和412-1可以使得所接收到的NVMe命令被转换成块IO命令。用户空间块装置(bdev)层416可以将NVMe SSD划分成若干路径或分区。用户空间块装置可以是SPDK bdev层。在该示例中,块装置bdev1-bdevx可以表示块装置1-x,尽管可以表示更多或更少的块装置。用户空间bdev层416可以将NVMe命令转译成块命令。用户空间bdev层416可以将NVMe命令的地址(包括虚拟地址)变换成与存储装置处的地址访问兼容的块层命令。例如,可以为地址确定逻辑块层地址。例如,bdev1可以对应于逻辑块地址lba0至lba100,并且块大小可以是4KB或任何大小。代替bdev,可以使用异步IO(AIO)方式。

[0049] SPDK块装置层是旨在等同于操作系统块存储层的库。该库可以提供以下功能性:(1)可插拔模块API,用于实现与不同类型的块存储装置通过接口连接的块装置;(2)驱动程

序模块,用于NVMe、malloc(ramdisk)、Linux AIO、virtio-scsi、Ceph RBD、Pmem和Vhost-SCSI启动器以及更多;(3)应用API,用于枚举和声明SPDK块装置并且然后在这些装置上执行操作(读取、写入、解映射等);(4)用于堆叠块装置以创建复杂I/O流水线的设施,包括逻辑卷管理(lvol)和分区支持(GPT);(5)经由JSON-RPC配置块装置;(6)请求排队、超时和复位处置;或(7)多个无锁队列,用于向块装置发送I/O。

[0050] NVMe驱动程序418可以在诸如用户空间IO(UIO)或VFIO(虚拟功能I/O)之类的内核模块中或在用户空间中执行。NVMe驱动程序418可以向相关联的NVMe兼容SSD或其它存储或存储器装置发出NVMe命令。NVMe命令可以通过队列对(qpair)来执行。队列对可以包括提交队列(SQ)和完成队列(CQ)。命令被放入SQ中,并且CQ用于从SSD 420接收响应。

[0051] 尽管本文中描述的示例是关于主机计算系统的,但是各种实施例可以用于网络或构造接口中的处理器。云服务提供商或数据中心所有者或承租者可以使用本文中描述的跨一个或多个网络接口的技术来配置数据中心,以轮询经由一个或多个传输层协议所接收到的命令的接收,或者使用在处理器上执行的轮询器来轮询命令的完成。本文中描述的示例可以应用于存储器或存储池,包括持续存储器。

[0052] 图4B描绘图4A的系统的示例操作。示例处理流程可以如下。在(1),接口402接收分组。在(2),传输层处理器处理分组。在(3),基于通过轮询组408-0在不同传输层之中的轮询,根据使用RDMA来标识所接收到的分组,并且传输层处理器404-0处理所接收到的分组并提取NVMe命令。传输层处理器404-0可以在共享缓冲器410中分配缓冲器(例如,根据NVMe命令需要),以存储与从启动器所接收到的命令相关联的数据,或者缓冲器用于放置要从由该命令标识的NVMe SSD读取的内容。在(4),NVM子系统412-0接收NVMe命令,并且使用用户空间bdev层416或者使用直通(passthrough)模式将与命令相关联的地址转换成块IO命令,如果NVMe SSD可以解译该地址的话。在(5),将块IO命令提供给SSD 420的NVMe接口。在(6)和(7),在执行由NVMe命令传达的存储事务之后,将NVMe IO完成指示提供给NVM子系统412-0以完成IO并通知轮询组(例如,轮询组408-0)。在(8),使用所选传输层,结果被封装并且被发送到启动器。例如,对于读取请求,使用传输层封装读取数据,并且在(9),将数据分组提供给接口402以用于传输。在(10),接口402向启动器系统发送数据分组。

[0053] 图5示出用于传输层处理的图。当通过传输层从远程装置接收NVMe命令时,关于图5来描述处理。目标从特定传输层接收新请求。在502,目标从传输提取NVMe命令(例如,NVMe命令可以封装在TCP PDU(协议数据单元)中以用于TCP传输)。对于读取命令(CMD),在510,CMD可以被发送到NVM子系统,并被执行(512),并且是根据传输要求提供(514)的封装的响应。对于写入命令(CMD),在520,来自主机用于写入CMD的数据被访问,并且发生来自510、512和514的过程。

[0054] 图6A描绘用于配置目标存储节点以使用一个或多个传输类型进行通信的示例过程。在602,该过程配置使用一个或多个传输类型在主机和节点之间所接收到的事务的轮询组。轮询组可由单个核(或多个核)执行,以从多个传输类型之中轮询来自连接的所接收到的通信。另外,轮询组还可以轮询所接收到的完成通知,其指示由(一个或多个)存储装置提供的读取或写入操作完成。例如,目标存储节点可以使用NVMe-oF兼容通信来接收和执行NVMe命令。可以使用NVMe命令的各种传输层类型,诸如以下中的一个或多个:远程直接存储器存取(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议(UDP)、QUIC

等等。在一些示例中，单个核可以针对使用一个或多个传输类型在主机和节点之间所接收到的事务执行轮询组。

[0055] 另外，在602，可以设置对传输类型轮询的优先级，由此第一传输层类型的轮询可以以第一速率发生，并且第二传输层类型的轮询可以以第二速率发生。第一速度和第二速率可以被设置成提供针对不同传输层类型的服务质量，从而潜在地使一种类型的传输层的通信的轮询优先于另一种类型的传输层。

[0056] 在604，过程可以配置共享数据缓冲器以存储从一个或多个传输类型所接收到的命令和相关内容。可以为来自启动器或连接的一个或多个传输类型的命令分配共享数据缓冲器。如果轮询组访问数据缓冲器(例如，读取或写入)，则共享数据缓冲器可由一个或多个轮询组访问而无需锁定。共享数据缓冲器可以被分配以存储与读取类型或写入类型的NVMe命令相关联的数据。例如，对于写入NVMe命令，与NVMe命令相关联的数据可以存储在共享缓冲器中。对于读取NVMe命令，从存储介质读取的并且与NVMe读取命令相关联的数据可以在传输到启动器之前存储在共享缓冲器中。

[0057] 图6B描绘可以由节点用于与另一个节点或主机装置进行通信的示例过程。在610，过程使用轮询组以根据编程的优先级方案使用一个或多个传输类型轮询存储相关的命令通信。在610，轮询组可以轮询命令(例如，NVMe命令)的提交或命令的完成(例如，数据与读取命令有关是可用的或数据已被写入用于写入命令)。可以配置传输层的轮询速率。为轮询组分配的共享缓冲池可以用于存储与NVMe命令相关联的数据。对于写入命令，数据可以是要被写入到存储介质的数据。对于读取命令，数据可以是从存储介质读取的数据。

[0058] 在612，在限制可应用于第一传输协议的情况下，可以将限制应用于用于传达NVMe命令的第二传输协议。例如，如果区域经受对阻止第一传输协议的读取或写入操作的限制，则将限制传播到使用第二传输协议。本文中描述了限制和保留的各种示例。

[0059] 在614，响应于检测到用于传输类型的通信，过程将与命令(例如，NVMe命令)相关联的地址转译成由目标存储装置用于与该命令相关联的地址的地址或地址范围。例如，用户空间块装置可以将地址转译成物理存储装置上的地址。

[0060] 在616，存储装置执行命令并且通知轮询组命令的完成。如果命令用于数据读取，则数据被存储在共享缓冲器中并且可用于被传送。在618，使用所选传输层将存储命令的结果提供给启动器。例如，可以选择多个传输层中的一个供发送对于存储命令的响应使用。在从存储装置读取数据的情况下，可以将数据传送给请求者。在数据将被写入到存储装置的情况下，完成通知可以被发送给请求者。可以通过CQ(完成队列)取(fetch)响应。

[0061] 图7描绘系统。系统可以使用本文中描述的实施例，以利用在一个或多个处理器上执行的轮询器轮询在多个传输层协议上所接收到的分组。系统700包括处理器710，其提供用于系统700的指令的处理、操作管理和执行。处理器710可以包括任何类型的微处理器、中央处理单元(CPU)、图形处理单元(GPU)、处理核或其它处理硬件以为系统700提供处理、或处理器的组合。处理器710控制系统700的整体操作，并且可以是或包括一个或多个可编程通用或专用微处理器、数字信号处理器(DSP)、可编程控制器、专用集成电路(ASIC)、可编程逻辑装置(PLD)等，或这样的装置的组合。

[0062] 在一个示例中，系统700包括耦合到处理器710的接口712，其可以表示用于需要较高带宽连接的系统组件的较高速接口或高吞吐量接口，所述系统组件诸如存储器子系统

720或图形接口组件740或加速器742。接口712表示接口电路，其可以是独立组件或被集成到处理器管芯上。目前，图形接口740与图形组件通过接口连接，以用于向系统700的用户提供可视显示。在一个示例中，图形接口740可以驱动向用户提供输出的高清晰度(HD)显示器。高清晰度可以指具有大约100 PPI(每英寸的像素)或更大的像素密度的显示，并且可以包括诸如全HD(例如，1080p)、视网膜显示器、4K(超高清晰度或UHD)或其它之类的格式。在一个示例中，显示器可以包括触摸屏显示器。在一个示例中，图形接口740基于存储在存储器730中的数据或基于由处理器710执行的操作或两者来生成显示。在一个示例中，图形接口740基于存储在存储器730中的数据或基于由处理器710执行的操作或两者来生成显示。

[0063] 加速器742可以是固定功能或者是可以由处理器710访问或使用的可编程卸载引擎。例如，在加速器742之中的加速器可以提供压缩(DC)能力、诸如公钥加密(PKE)之类的密码服务、密码(cipher)、散列/认证能力、解密或其它能力或服务。在一些实施例中，另外或备选地，在加速器742之中的加速器提供如本文中所描述的字段选择控制器能力。在一些情况下，加速器742可以被集成到CPU插槽(例如，到包括CPU并且提供与CPU的电接口的主板或电路板的连接器)中。例如，加速器742可以包括单核或多核处理器、图形处理单元、逻辑执行单元、单或多级高速缓存、可用于独立地执行程序或线程的功能单元、专用集成电路(ASIC)、神经网络处理器(NNP)、可编程控制逻辑、以及诸如现场可编程门阵列(FPGA)或可编程逻辑装置(PLD)之类的可编程处理元件。加速器742可以提供多个神经网络、CPU、处理器核、通用图形处理单元，或者可以使图形处理单元可用于供人工智能(AI)或机器学习(ML)模型使用。例如，AI模型可以使用或包括以下中的任何一个或组合：强化学习方案、Q学习方案、深度Q学习或异步优势动作评价(A3C)、组合神经网络、循环组合神经网络或其它AI或ML模型。可以使多个神经网络、处理器核或图形处理单元可用于供AI或ML模型使用。

[0064] 存储器子系统720表示系统700的主存储器，并且为要由处理器710执行的代码或要在执行例程中使用的数据值提供存储。存储器子系统720可以包括一个或多个存储器装置730，诸如只读存储器(ROM)、闪速存储器、诸如DRAM之类的一种或多种随机存取存储器(RAM)、或其它存储器装置、或这样的装置的组合。存储器730此外存储和托管操作系统(OS)732以提供用于执行系统700中的指令的软件平台。另外，应用734可以在来自存储器730的OS 732的软件平台上执行。应用734表示具有其自身的操作逻辑以执行一个或多个功能的执行的程序。进程736表示向OS 732或一个或多个应用734或组合提供辅助功能的代理或例程。OS 732、应用734和进程736提供软件逻辑以向系统700提供功能。在一个示例中，存储器子系统720包括存储器控制器722，所述存储器控制器722是用于生成命令并将命令发出到存储器730的存储器控制器。将理解，存储器控制器722可以是处理器710的物理部分或接口712的物理部分。例如，存储器控制器722可以是集成存储器控制器，其与处理器710集成到电路上。

[0065] 虽然没有特定图示，但是将理解，系统700可以包括装置之间的一个或多个总线或总线系统，诸如存储器总线、图形总线、接口总线或其它。总线或其它信号线可以将组件通信地或电耦合在一起，或者既通信地耦合又电耦合组件。总线可以包括物理通信线、点到点连接、桥、适配器、控制器或其它电路或组合。总线可以包括例如系统总线、外围组件互连(PCI)总线、超传输或工业标准架构(ISA)总线、小型计算机系统接口(SCSI)总线、通用串行总线(USB)或电气和电子工程师协会(IEEE)标准1394总线(火线)中的一个或多个。

[0066] 在一个示例中，系统700包括接口714，其可以耦合到接口712。在一个示例中，接口714表示接口电路，其可以包括独立组件和集成电路。在一个示例中，多个用户接口组件或外围组件或两者耦合到接口714。网络接口750向系统700提供通过一个或多个网络与远程装置(例如，服务器或其它计算装置)通信的能力。网络接口750可以包括以太网适配器、无线互连组件、蜂窝网络互连组件、USB(通用串行总线)或其它有线或无线的基于标准的或专有的接口。网络接口750可以将数据传送到在相同数据中心或架(rack)中的装置或远程装置，这可以包括发送存储在存储器中的数据。网络接口750可以从远程装置接收数据，这可以包括将所接收到的数据存储到存储器中。各种实施例可以结合网络接口750、处理器710和存储器子系统720使用。网络接口750的各种实施例使用本文中描述的实施例来接收或传送定时相关信号，并且提供免受来自错误配置的端口使用的电路损坏的保护，同时提供可接受的传播延迟。

[0067] 在一个示例中，系统700包括一个或多个输入/输出(I/O)接口760。I/O接口760可以包括一个或多个接口组件，用户通过所述一个或多个接口组件与系统700交互(例如，音频、字母数字、触觉/触摸或其它接口)。外围接口770可以包括上面没有特别提到的任何硬件接口。外围设备通常是指依赖地连接到系统700的装置。依赖连接是系统700提供在其上执行操作并且用户与其交互的软件平台或硬件平台或两者的连接。

[0068] 在一个示例中，系统700包括以非易失性方式存储数据的存储子系统780。在一个示例中，在某些系统实现中，存储设备780的至少某些组件可以与存储器子系统720的组件重叠。存储子系统780包括(一个或多个)存储装置784，其可以是或包括用于以非易失性方式存储大量数据的任何常规介质，诸如一个或多个基于磁、固态或光的盘或组合。存储设备784将代码或指令和数据786保持在持续状态(即，尽管系统700的电源中断，值仍被保持)。存储设备784可以一般地被认为是“存储器”，尽管存储器730通常是执行或操作存储器以向处理器710提供指令。尽管存储设备784是非易失性的，但是存储器730可以包括易失性存储器(即，如果系统700的电源被中断，则数据的值或状态)是不确定的。在一个示例中，存储子系统780包括与存储设备784通过接口连接的控制器782。在一个示例中，控制器782是接口714或处理器710的物理部分，或者可以包括处理器710和接口714两者中的电路或逻辑。

[0069] 易失性存储器是如果装置的电源被中断则其状态(以及因此存储在其中的数据)是不确定的存储器。动态易失性存储器使用刷新存储在装置中的数据来维持状态。动态易失性存储器的一个示例包括DRAM(动态随机存取存储器)，或一些变体，诸如同步DRAM(SDRAM)。如本文中所描述的存储器子系统可与许多存储器技术兼容，所述存储器技术诸如DDR3(双倍数据速率版本3，由JEDEC(联合电子装置工程委员会)在2007年6月27日的原始发布)。DDR4(DDR版本4，由JEDEC在2012年9月出版的初始规范)、DDR4E(DDR版本4)、LPDDR3(低功率DDR版本3，JESD209-3B，由JEDEC在2013年8月)、LPDDR4(LPDDR版本4、由JEDEC在2014年8月最初出版的JESD209-4)、WIO2(宽输入/输出版本2，由JEDEC在2014年8月最初出版的JESD229-2)、HBM(高带宽存储器，由JEDEC在2013年10月最初出版的JESD325)、LPDDR5(当前由JEDEC在讨论中)、HBM2(HBM版本2)，当前由JEDEC在讨论中，或存储器技术的其它或组合，以及基于这样的规范的衍生或扩展的技术。JEDEC标准在www.jedec.org处可获取。

[0070] 非易失性存储器(NVM)装置是即使装置的电源被中断、其状态也是确定的存储器。在一个实施例中，NVM装置可以包括块可寻址存储器装置，诸如NAND技术，或者更特别地，多

阈值级NAND闪速存储器(例如,单级单元(“SLC”)、多级单元(“MLC”)、四级单元(“QLC”)、三级单元(“TLC”)或一些其它NAND)。NVM装置还可以包括字节可寻址就地写入(write-in-place)三维交叉点存储器装置,或其它字节可寻址就地写入NVM装置(也称为持续存储器),诸如单或多级相变存储器(PCM)或具有开关的相变存储器(PCMS)、Intel® Optane™、使用硫族化物相变材料(例如,硫族化物玻璃)的NVM装置、包括金属氧化物基、氧空位基和导电桥随机存取存储器(CB-RAM)的阻性存储器、纳米线存储器、铁电随机存取存储器(FeRAM、FRAM)、结合忆阻器技术的磁阻随机存取存储器(MRAM)、自旋转移矩(STT)-MRAM、基于自旋电子磁性结存储器的装置、基于磁隧道结(MTJ)的装置、基于DW(畴壁)和SOT(自旋轨道转移)的装置、基于晶闸管的存储器装置、或以上任何的组合、或其它存储器。

[0071] 电源(未描绘)向系统700的组件提供功率。更特别地,电源通常与系统700中的一个或多个电力供应通过接口连接,以向系统700的组件供电。在一个示例中,电力供应包括AC到DC(交流电到直流电)适配器以插入到壁装插座(wall outlet)中。这种AC电源可以是可再生能源(例如,太阳能)功率源。在一个示例中,电源包括DC电源,诸如外部AC到DC转换器。在一个示例中,电源或电力供应包括无线充电硬件,以经由接近充电场进行充电。在一个示例中,电源可以包括内部电池、交流电流供应、基于运动的电力供应、太阳能电力供应或燃料电池源。

[0072] 在示例中,系统700可以使用处理器、存储器、存储设备、网络接口和其它组件的互连计算板(sled)来实现。可以使用高速互连,诸如:以太网(IEEE 802.3)、远程直接存储器存取(RDMA)、InfiniBand、因特网广域RDMA协议(iWARP)、快速UDP因特网连接(QUIC)、融合以太网上的RDMA(RoCE)、高速外围组件互连(PCIe)、Intel快速通道(QuickPath)互连(QPI)、Intel超级通道互连(UPI)、Intel片上系统构造(IOSF)、OmniPath、计算高速链路(CXL)、超传输(HyperTransport)、高速构造、NVLink、高级微控制器总线架构(AMBA)互连、OpenCAPI、Gen-Z、加速器高速缓存一致互连(CCIX)、3GPP长期演进(LTE)(4G)、3GPP 5G及其变型。可以使用诸如构造上NVMe(NVMe-oF)或NVMe之类的协议将数据复制或存储到虚拟化的存储节点。

[0073] 图8描绘包括多个计算架802的环境800,每个计算架802包括架顶部(ToR)交换机804、pod管理器806和多个池化系统抽屉。通常,池化系统抽屉可以包括池化计算抽屉和池化存储抽屉。可选地,池化系统抽屉还可以包括池化存储器抽屉和池化输入/输出(I/O)抽屉。在图示的实施例中,池化系统抽屉包括Intel® Xeon®处理器池化计算机抽屉808和Intel® ATOM™处理器池化计算抽屉810、池化存储抽屉812、池化存储器抽屉814和池化I/O抽屉816。池化系统抽屉中的每个经由高速链路818(诸如40吉比特/秒(Gb/s)或100 Gb/s以太网链路或100+ Gb/s硅光子(SiPh)光链路)连接到ToR交换机804。

[0074] 多个计算架802可以经由它们的ToR交换机804互连(例如,到pod级交换机或数据中心交换机),如由到网络820的连接所示的那样。在一些实施例中,计算架802的群组经由(一个或多个)pod管理器806作为单独的pod来管理。在一个实施例中,使用单个pod管理器来管理pod中的所有架。备选地,分布式pod管理器可以用于pod管理操作。

[0075] 环境800还包括用于管理环境的各个方面的管理接口822。这包括管理架配置,其中对应的参数被存储为架配置数据824。

[0076] 在示例中,环境800可以使用处理器、存储器、存储设备、网络接口和其它组件的互

连计算板来实现。可以使用高速互连，诸如：以太网(IEEE 802.3)、远程直接存储器存取(RDMA)、InfiniBand、因特网广域RDMA协议(iWARP)、快速用户数据报协议(UDP)因特网连接(QUIC)、融合以太网上的RDMA(RoCE)、高速外围组件互连(PCIe)、Intel快速通道(QuickPath)互连(QPI)、Intel超级通道互连(UPI)、Intel片上系统构造(IOSF)、OmniPath、计算高速链路(CXL)、超传输(HyperTransport)、高速构造、NVLink、高级微控制器总线架构(AMBA)互连、OpenCAPI、Gen-Z、加速器高速缓存一致互连(CCIX)、3GPP长期演进(LTE)(4G)、3GPP 5G及其变型。可以使用诸如构造上NVMe(NVMe-oF)或NVMe之类的协议将数据复制或存储到虚拟化的存储节点。

[0077] 本文中的实施例可以在各种类型的计算和联网设备(诸如交换机、路由器、架和刀片服务器，诸如在数据中心和/或服务器场环境中所采用的那些)中实现。在数据中心和服务器场中使用的服务器包括阵列服务器配置，诸如基于架的服务器或刀片服务器。这些服务器经由各种网络供应(provision)通信互连，诸如将服务器集合划分成在LAN之间具有适当的交换和路由设施以形成专用内联网的局域网(LAN)。例如，云托管设施可以通常采用具有众多服务器的大型数据中心。刀片包括被配置成执行服务器类型功能的单独的计算平台，即“卡上的服务器”。因此，每个刀片包括为常规服务器所共用的组件，包括提供内部接线(即，总线)以用于耦合适当的集成电路(IC)和安装到板的其它组件的主印刷电路板(主板)。

[0078] 图9描绘可以使用实施例或由实施例使用的网络接口。网络接口可以使用本文中结合各种实施例描述的实施例来接收或传送分组。网络接口900可以包括收发器902、处理器904、传送队列906、接收队列908、存储器910、和总线接口912、以及DMA引擎926。收发器902可能能够依照可应用的协议(例如，如IEEE 802.3中描述的以太网)接收和传送分组，尽管可以使用其它协议。收发器902可以经由网络介质(未描绘)从网络接收分组和向网络传送分组。收发器902可以包括物理层(PHY)电路914和介质访问控制(MAC)电路916。PHY电路914可以包括编码和解码电路(未示出)，以根据可应用的物理层规范或标准来对数据分组进行编码和解码。MAC电路916可以被配置成将要传送的数据装配成分组，所述分组包括目的地和源地址以及网络控制信息和错误检测散列值。MAC电路916可以被配置成通过验证数据完整性、移除前导码和填充、以及提供供较高层处理的内容，来处理所接收到的分组的MAC报头。

[0079] 处理器904可以是以下的任何组合：处理器、核、图形处理单元(GPU)、现场可编程门阵列(FPGA)、专用集成电路(ASIC)或允许对网络接口900进行编程的其它可编程硬件装置。例如，处理器904可以提供中间队列的分配或解除分配。例如，“智能网络接口”可以使用处理器904在网络接口中提供分组处理能力。

[0080] 分组分配器924可以提供所接收到的分组的分发供使用本文中描述的时隙分配或RSS的多个CPU或核处理。当分组分配器924使用RSS时，分组分配器924可以基于所接收到的分组的内容来计算散列或作出另一个确定，以确定哪个CPU或核将处理分组。

[0081] 中断合并922可以执行中断调节，由此网络接口中断合并922在生成主机系统的中断以处理(一个或多个)所接收到的分组之前等待多个分组到达或者等待超时期满。接收段合并(RSC)可以由网络接口900执行，由此将传入分组的部分组合成分组的段。网络接口900将该合并的分组提供给应用。

[0082] 直接存储器存取(DMA)引擎926可以将分组报头、分组有效负载和/或描述符直接从主机存储器复制到网络接口，或反之亦然，而不是将分组复制到主机处的中间缓冲器并且然后使用另一个复制操作从中间缓冲器到目的地缓冲器。

[0083] 存储器910可以是任何类型的易失性或非易失性存储器装置，并且可以存储用于对网络接口900进行编程的任何队列或指令。传送队列906可以包括供网络接口传输的数据或对该数据的引用。接收队列908可以包括由网络接口从网络接收的数据或对该数据的引用。描述符队列920可以包括引用传送队列906或接收队列908中的数据或分组的描述符。总线接口912可以提供与主机装置(未描绘)的接口。例如，总线接口912可以与外围连接外围组件互连(PCI)、高速PCI、PCI-x、串行ATA(SATA)和/或通用串行总线(USB)兼容接口兼容(尽管可以使用其它互连标准)。

[0084] 在一些示例中，本文中描述的网络接口和其它实施例可以结合以下来使用：基站(例如，3G、4G、5G等)、宏基站(例如，5G网络)、微微站(例如，兼容IEEE 802.11的访问点)、纳米站(例如，用于点到多点(PtMP)应用)、内部部署(on-premises)数据中心、外部部署(off-premises)数据中心、边缘网络元件、雾网络元件和/或混合数据中心(例如，使用虚拟化、云和软件定义的联网来跨物理数据中心和分布式多云环境递送应用工作负荷的数据中心)。

[0085] 可以使用硬件元件、软件元件或两者的组合来实现各种示例。在一些示例中，硬件元件可以包括装置、组件、处理器、微处理器、电路、电路元件(例如，晶体管、电阻器、电容器、电感器等)、集成电路、ASIC、PLD、DSP、FPGA、存储器单元、逻辑门、寄存器、半导体装置、芯片、微芯片、芯片集等。在一些示例中，软件元件可以包括软件组件、程序、应用、计算机程序、应用程序、系统程序、机器程序、操作系统软件、中间件、固件、软件模块、例程、子例程、函数、方法、过程、软件接口、API、指令集、计算代码、计算机代码、代码段、计算机代码段、字、值、符号或其任何组合。确定是否使用硬件元件和/或软件元件来实现示例，可以根据任意数量的因素而变化，所述因素诸如期望的计算速率、功率水平、耐热性、处理循环预算、输入数据速率、输出数据速率、存储器资源、数据总线速度和其它设计或性能约束，如给定实现所期望的那样。处理器可以是硬件状态机、数字控制逻辑、中央处理单元或任何硬件、固件和/或软件元件中的一个或多个组合。

[0086] 一些示例可以使用或作为制品或至少一个计算机可读介质来实现。计算机可读介质可以包括用于存储逻辑的非暂时性存储介质。在一些示例中，非暂时性存储介质可以包括能够存储电子数据的一个或多个类型的计算机可读存储介质，包括易失性存储器或非易失性存储器、可移除或不可移除存储器、可擦除或不可擦除存储器、可写或可重写存储器等。在一些示例中，逻辑可以包括各种软件元件，诸如软件组件、程序、应用、计算机程序、应用程序、系统程序、机器程序、操作系统软件、中间件、固件、软件模块、例程、子例程、功能、方法、过程、软件接口、API、指令集、计算代码、计算机代码、代码段、计算机代码段、字、值、符号或其任何组合。

[0087] 根据一些示例，计算机可读介质可以包括非暂时性存储介质，用于存储或维护指令，所述指令在由机器、计算装置或系统执行时，使得机器、计算装置或系统执行根据所描述的示例的方法和/或操作。指令可以包括任何适合类型的代码，诸如源代码、编译代码、解译代码、可执行代码、静态代码、动态代码等。指令可以根据预定义的计算机语言、方式或语法来实现，以用于指示机器、计算装置或系统执行特定功能。指令可以使用任何适合的高

级、低级、面向对象、可视、编译和/或解译编程语言来实现。

[0088] 至少一个示例的一个或多个方面可以由存储在至少一个机器可读介质上的代表性指令来实现，该代表性指令表示处理器内的各种逻辑，该代表性指令当由机器、计算装置或系统读取时，使得机器、计算装置或系统制造逻辑以执行本文中描述的技术。被称为“IP核”的这样的表示可以被存储在有形的机器可读介质上，并且被提供给各种客户或制造设施以加载到实际制造逻辑或处理器的制作机器中。

[0089] 短语“一个示例”或“示例”的出现未必全部指的是相同的示例或实施例。本文中描述的任何方面可以与本文中描述的任何其它方面或类似方面组合，而不管该方面是否是关于相同的附图或元素来描述的。附图中描绘的块功能的划分、省略或包括不推断用于实现这些功能的硬件组件、电路、软件和/或元件将必须被划分、省略或包括在实施例中。

[0090] 一些示例可以使用表达“耦合的”和“连接的”以及它们的派生词来描述。这些术语未必旨在作为彼此的同义词。例如，使用术语“连接的”和/或“耦合的”的描述可以指示两个或更多个元件彼此直接物理或电接触。然而，术语“耦合的”也可以意味着两个或更多个元件彼此不直接接触，但是仍然彼此协作或交互。

[0091] 术语“第一”、“第二”等在本文中不表示任何顺序、数量或重要性，而是用于将一个元件与另一个元件进行区分。术语“一”和“一个”在本文中不表示数量的限制，而是表示存在所引用的项中的至少一个。本文中参考信号所使用的术语“断言”表示信号的状态，其中信号是活动的，并且这可以通过向信号应用逻辑0或逻辑1中的任何逻辑电平来实现。术语“跟随”或“之后”可以指紧随或跟随在一些其它事件或多个事件之后。根据备选实施例，也可以执行其它步骤序列。此外，取决于特定的应用，可以添加或移除附加的步骤。可以使用改变的任意组合，并且具有本公开的益处的本领域技术人员将理解其许多变化、修改和备选实施例。

[0092] 除非另外特定说明，否则诸如短语“X、Y或Z中的至少一个”之类的分离性语言在如通常用于呈现以下内容的上下文内以其它方式来理解：项、术语等可以是X、Y或Z或其任意组合(例如，X、Y和/或Z)。因此，这样的分离性语言通常不旨在并且不应当暗示某些实施例要求X中的至少一个、Y中的至少一个或Z中的至少一个而各自存在。另外，除非另有特定说明，否则诸如短语“X、Y和Z中的至少一个”之类的连接性语言应当也被理解成表示X、Y、Z或其任意组合，包括“X、Y和/或Z”。

[0093] 下面提供了本文中公开的装置、系统和方法的说明性示例。装置、系统和方法的实施例可以包括以下描述的示例中的任一个或多个以及任意组合。

[0094] 示例1包括一种非暂时性计算机可读介质，包括存储在其上的指令，所述指令如果由一个或多个处理器执行，则使得所述一个或多个处理器：执行轮询器以：轮询在多个传输层协议上来自连接的所接收到的通信，以根据所述多个传输层协议中的一个传输层协议识别所接收到的通信，并且根据所述多个传输层协议中的不同的一个传输层协议识别第二所接收到的通信。

[0095] 示例2可以与任何示例组合，其中所接收到的通信包括构造上高速非易失性存储器(NVMe-oF)兼容命令。

[0096] 示例3可以与任何示例组合，其中响应于所述多个传输层协议中的所述一个传输层协议的失败或者使用所述多个传输层协议中的所述一个传输层协议的慢传输速率，发生

对所述多个传输层协议中的所述不同的一个传输层协议的改变。

[0097] 示例4可以与任何示例组合，其中所述轮询器在用户空间中被执行并且包括存储在其上的指令，所述指令如果由一个或多个处理器执行，则使得所述一个或多个处理器执行内核空间中的所接收到的通信和第二所接收到的通信的传输层协议处理。

[0098] 示例5可以与任何示例组合，并且包括存储在其上的指令，所述指令如果由一个或多个处理器执行，则使得所述一个或多个处理器：配置由所述轮询器应用的服务质量，以轮询从所述多个传输层协议之中所接收到的通信。

[0099] 示例6可以与任何示例组合，并且包括存储在其上的指令，所述指令如果由一个或多个处理器执行，则使得所述一个或多个处理器：执行所述轮询器以针对所执行的命令从存储装置轮询完成指示，其中所述轮询器将基于所配置的服务质量来轮询完成指示。

[0100] 示例7可以与任何示例组合，其中所述传输层协议包括以下中的一个或多个：远程直接存储器存取(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议(UDP)、或快速UDP因特网连接(QUIC)。

[0101] 示例8可以与任何示例组合，并且包括存储在其上的指令，所述指令如果由一个或多个处理器执行，则使得所述一个或多个处理器：对应用于之前第一传输层协议的第二传输层协议应用限制。

[0102] 示例9可以与任何示例组合，并且包括存储在其上的指令，所述指令如果由一个或多个处理器执行，则使得所述一个或多个处理器：从存储装置接收响应，以及使用所述多个传输层协议中的一个传输层协议来生成具有对启动器的所述响应的分组。

[0103] 示例10可以与任何示例组合，并且包括一种用于使用至少两个不同传输层协议中的一个或多个来提供构造上高速非易失性存储器(NVMe-oF)兼容服务的方法，所述方法包括：执行轮询器以从多个不同传输层协议之中轮询存储命令的接收，以及使得协议处理使用传输层协议所接收到的并且由所述轮询器所检测到的命令。

[0104] 示例11可以与任何示例组合，并且包括：许可使用不同传输层协议来同时访问NVMe兼容存储装置。

[0105] 示例12可以与任何示例组合，其中所述轮询器在用户空间中被执行，并且所述协议处理在内核空间中被执行。

[0106] 示例13可以与任何示例组合，并且其中执行轮询器以从多个不同传输层之中轮询存储命令的接收应用每传输层类型的服务质量。

[0107] 示例14可以与任何示例组合，并且包括：将所接收到的命令的地址转译成在存储装置处可访问的逻辑块地址。

[0108] 示例15可以与任何示例组合，并且包括：轮询对所接收到的命令的响应，以及提供所述响应以用于使用传输层协议的传输。

[0109] 示例16可以与任何示例组合，并且包括：从启动器接收通信，所述启动器从使用第一传输层协议改变成使用第二传输层协议，以及对应用于所述第一传输层协议的所述第二传输层协议应用限制。

[0110] 示例17可以与任何示例组合，并且其中所述传输层协议包括以下中的一个或多个：远程直接存储器存取(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议(UDP)、或快速UDP因特网连接(QUIC)。

[0111] 示例18可以与任何示例组合，并且包括一种系统，包括：网络接口；耦合到所述网络接口的存储器；以及耦合到所述存储器的第一处理器，所述第一处理器被配置成：执行第一轮询器以轮询使用所述第一传输协议和所述第二传输协议所接收到的接收到的存储命令，以及耦合到所述存储器的第二处理器，所述第二处理器被配置成：执行第二轮询器以轮询使用所述第一传输协议和所述第二传输协议所接收到的接收到的存储命令。

[0112] 示例19可以与任何示例组合，并且其中所述第一处理器将：应用所述第一轮询器的服务质量以轮询使用第一传输协议和第二传输协议所接收到的接收到的存储命令，以及其中所述第二处理器将：应用所述第二轮询器的服务质量以轮询使用第一传输协议和第二传输协议所接收到的接收到的存储命令。

[0113] 示例20可以与任何示例组合，并且其中所述第一处理器将：执行所述第一轮询器以轮询存储命令的完成，以及使得使用所述第一传输协议或第二传输协议向远程装置的传输响应，以及其中所述第二处理器将：执行所述第二轮询器以轮询存储命令的完成，以及使得使用所述第一传输协议或第二传输协议向远程装置的传输响应。

[0114] 示例21可以与任何示例组合，并且其中所述传输协议包括以下中的一个或多个：远程直接存储器存取(RDMA)、传输控制协议(TCP)、光纤通道(FC)、InfiniBand、用户数据协议(UDP)、或快速UDP因特网连接(QUIC)。

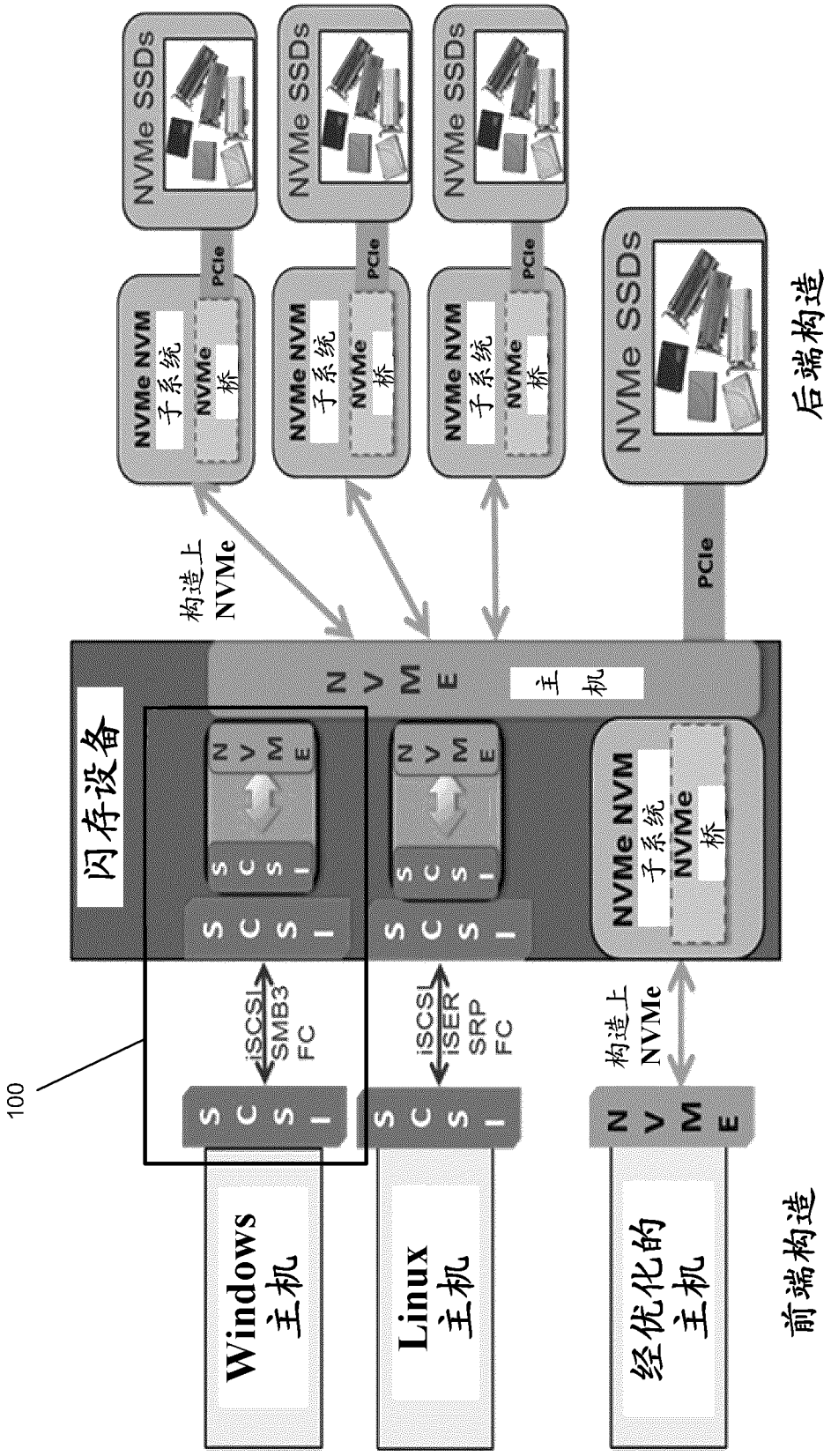

NVMe SSDs
NVMe SSDs
NVMe SSDs
NVMe SSDs
PCIe
PCIe
PCIe
PCIe
NVMe NVM
子系统
NVMe
桥
NVMe NVM
子系统
NVMe
桥
NVMe NVM
子系统
NVMe
桥
NVMe NVM
子系统
NVMe
桥
后端构造
构造上
NVMe
闪存设备
N V M E
主 机
S C S I
N V M E
S C S I
S C S I
S C S I
iSCSI
SMB3
FC
iSCSI
iSER
SRP
FC
构造上
NVMe
100
Windows
主机
Linux
主机
经优化的
主机
前端构造

图 1

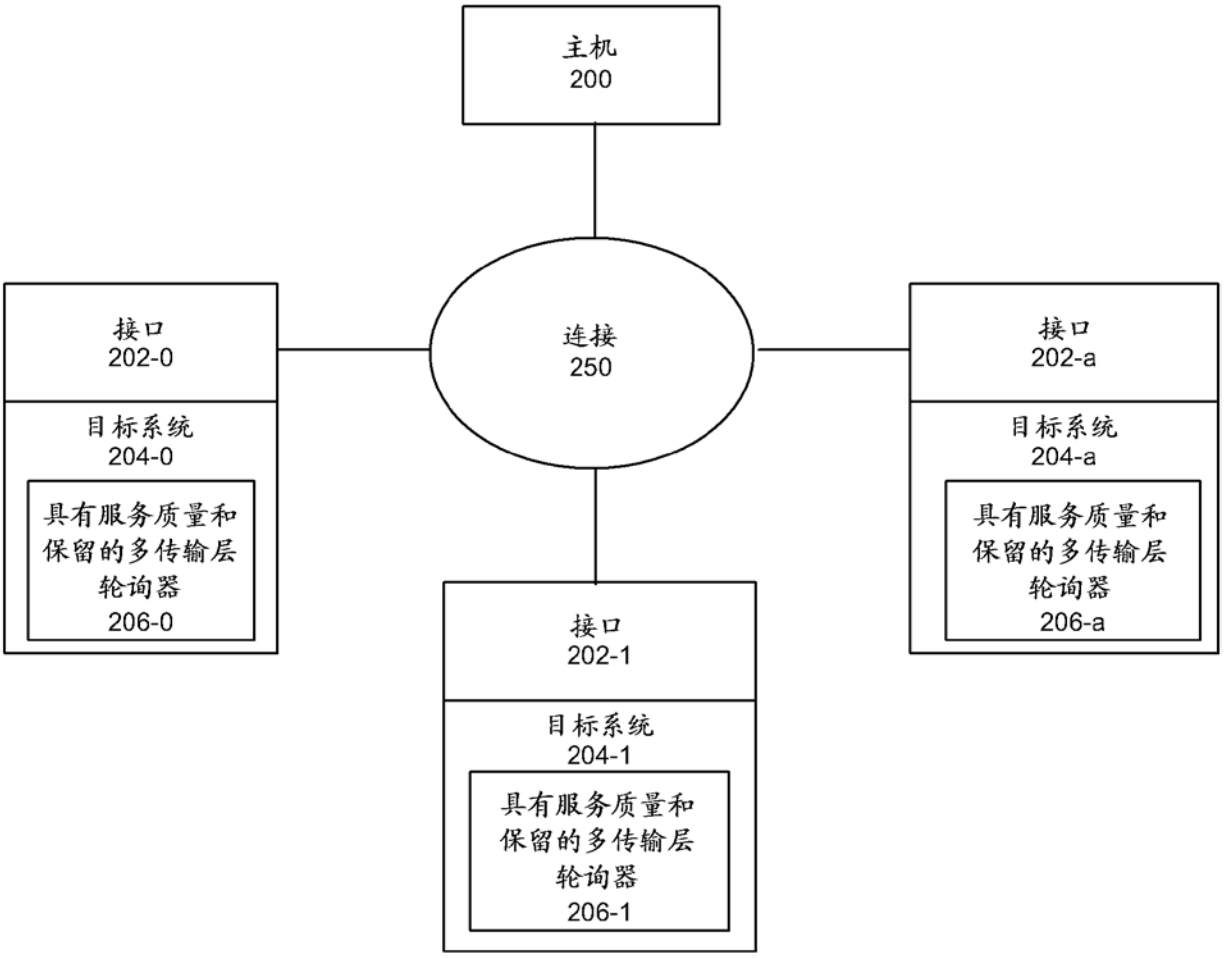

主机
200
连接
250
接口
202-0
目标系统
204-0
具有服务质量和保留的多传输层轮询器
206-0
接口
202-1
目标系统
204-1
具有服务质量和保留的多传输层轮询器
206-1
接口
202-a
目标系统
204-a
具有服务质量和保留的多传输层轮询器
206-a

图 2

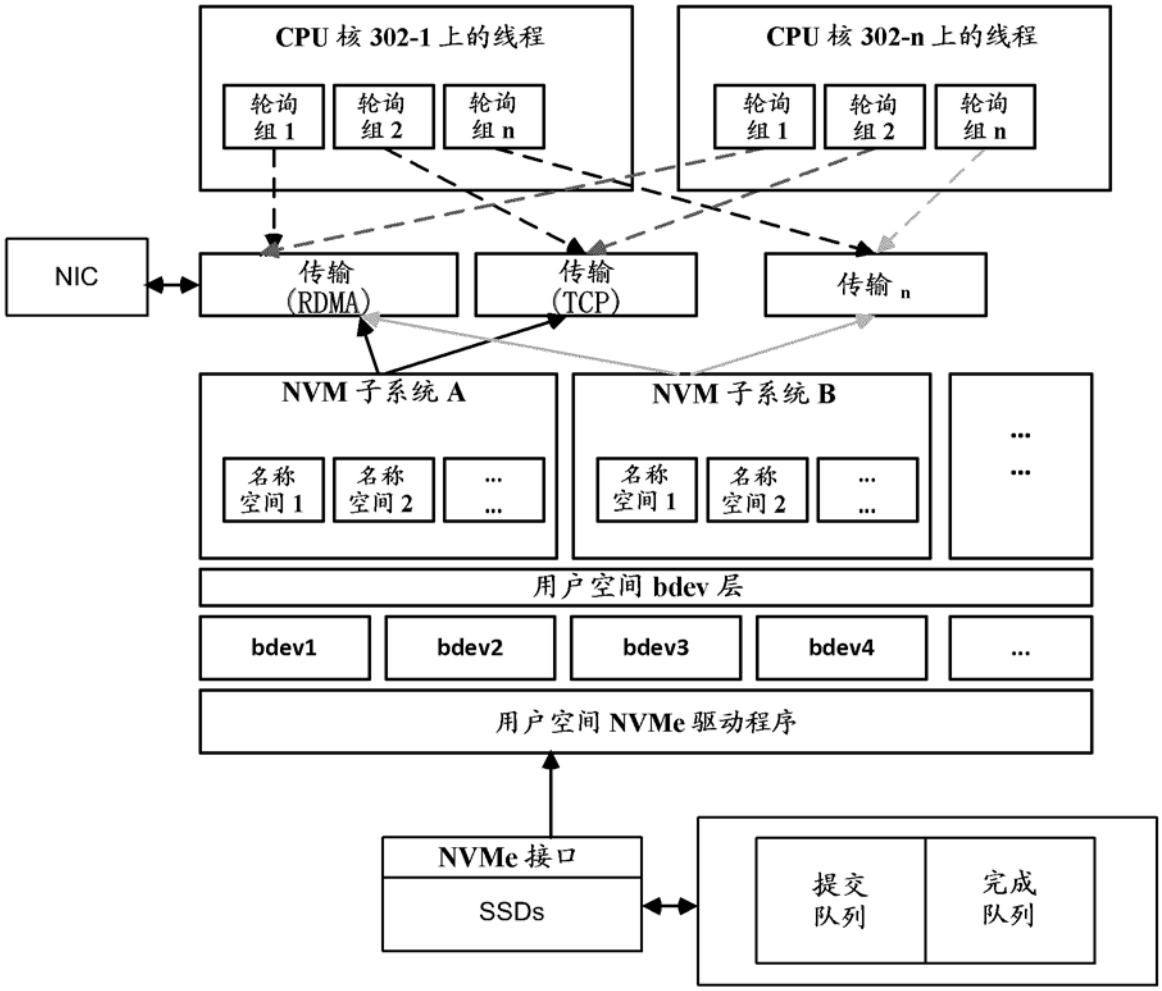

CPU 核 302-1 上的线程
轮询组 1
轮询组 2
轮询组 n
CPU 核 302-n 上的线程
轮询组 1
轮询组 2
轮询组 n
NIC
传输 (RDMA)
传输 (TCP)
传输 n
NVM 子系统 A
名称空间 1
名称空间 2
...
NVM 子系统 B
名称空间 1
名称空间 2
...
...
用户空间 bdev 层
bdev1
bdev2
bdev3
bdev4
...
用户空间 NVMe 驱动程序
NVMe 接口
SSDs
提交队列
完成队列

图 3

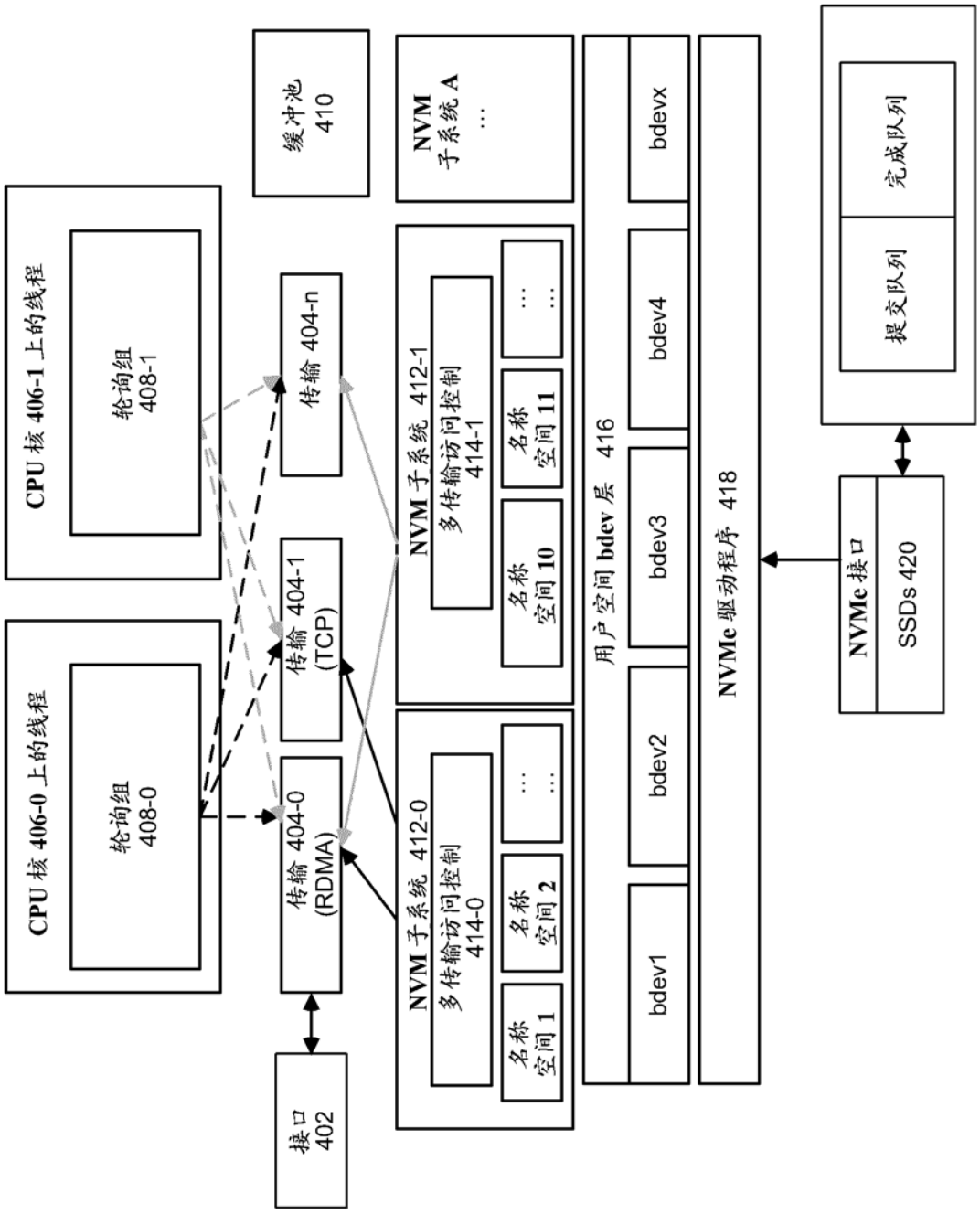

CPU 核 406-0 上的线程
轮询组 408-0
CPU 核 406-1 上的线程
轮询组 408-1
接口 402
传输 404-0 (RDMA)
传输 404-1 (TCP)
传输 404-n
缓冲池 410
NVM 子系统 412-0
多传输访问控制 414-0
名称空间 1
名称空间 2
NVM 子系统 412-1
多传输访问控制 414-1
名称空间 10
名称空间 11
NVM 子系统 A
用户空间 bdev 层 416
bdev1
bdev2
bdev3
bdev4
bdevx
NVMe 驱动程序 418
NVMe 接口
SSDs 420
提交队列
完成队列

图 4A

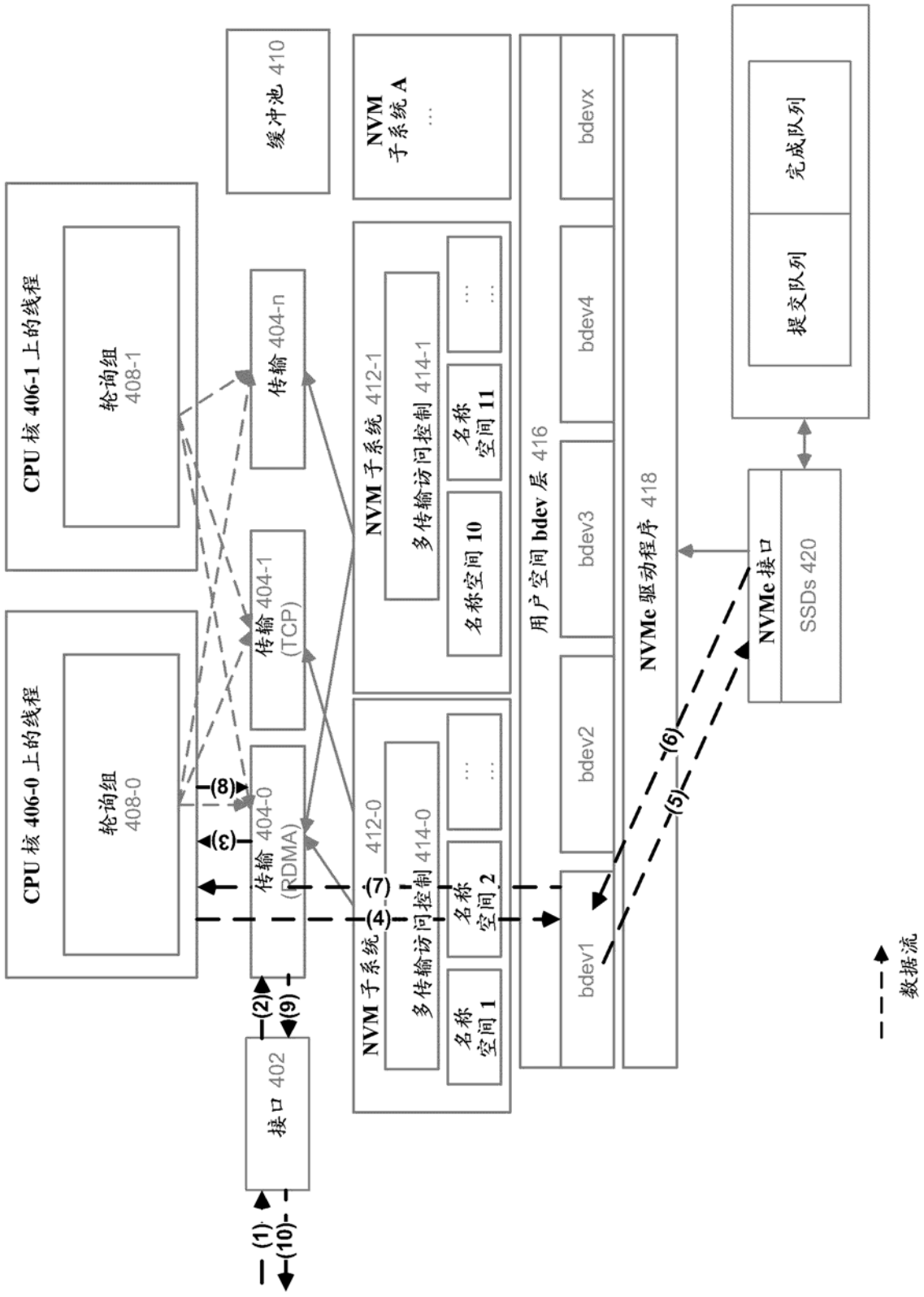

CPU 核 406-0 上的线程
轮询组 408-0
CPU 核 406-1 上的线程
轮询组 408-1
缓冲池 410
接口 402
传输 404-0 (RDMA)
传输 404-1 (TCP)
传输 404-n
NVM 子系统 412-0
多传输访问控制 414-0
名称空间 1
名称空间 2
NVM 子系统 412-1
多传输访问控制 414-1
名称空间 10
名称空间 11
NVM 子系统 A
用户空间 bdev 层 416
bdev1
bdev2
bdev3
bdev4
bdevx
NVMe 驱动程序 418
NVMe 接口
SSDs 420
提交队列
完成队列
(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
(9)
(10)
数据流

图 4B

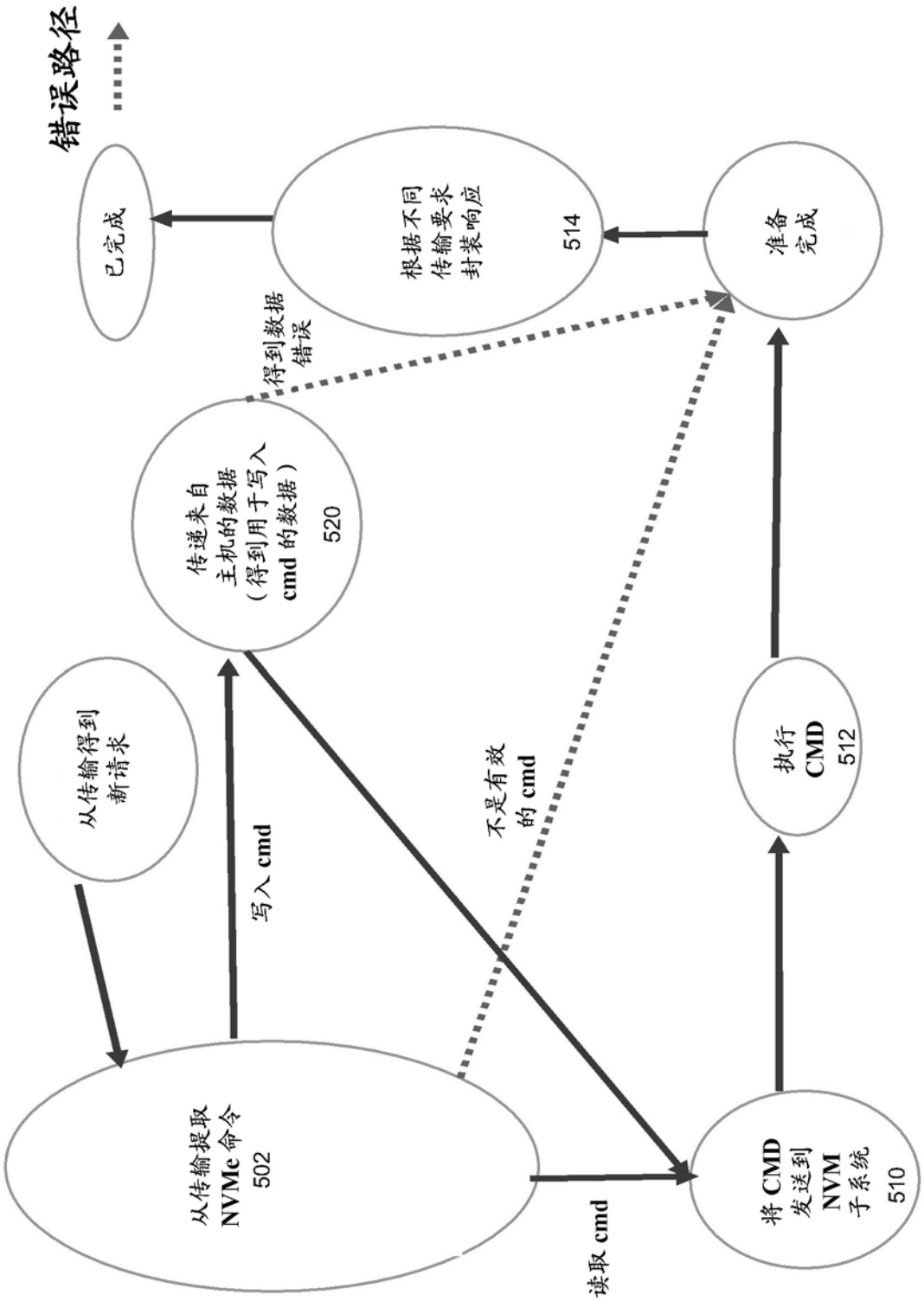

错误路径
已完成
根据不同传输要求封装响应
514
准备完成
得到数据错误
传递来自主机的数据（得到用于写入cmd的数据）
520
从传输得到新请求
写入 cmd
不是有效的 cmd
执行CMD
512
从传输提取NVMe 命令
502
读取 cmd
将 CMD发送到NVM子系统
510

图 5

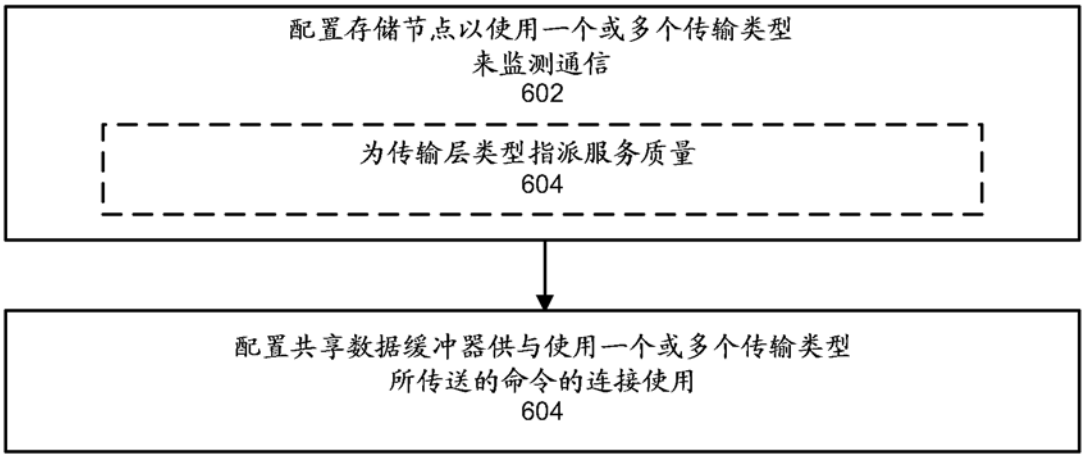

配置存储节点以使用一个或多个传输类型
来监测通信
602
为传输层类型指派服务质量
604
配置共享数据缓冲器供与使用一个或多个传输类型
所传送的命令的连接使用
604

图 6A

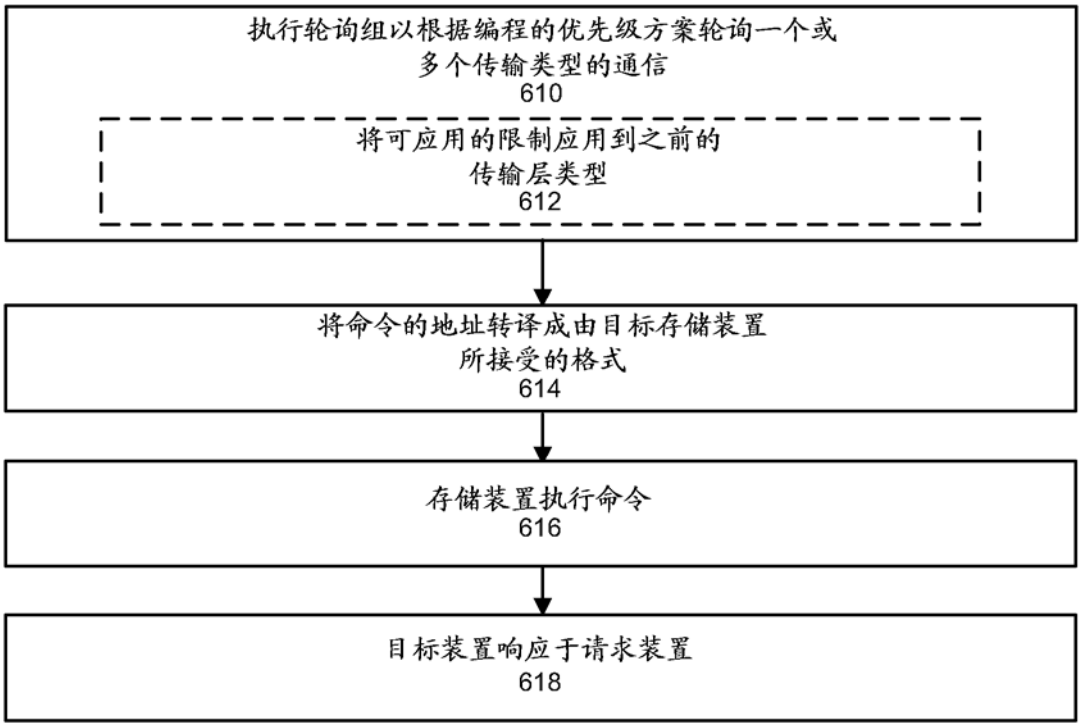

执行轮询组以根据编程的优先级方案轮询一个或
多个传输类型的通信
610
将可应用的限制应用到之前的
传输层类型
612
将命令的地址转译成由目标存储装置
所接受的格式
614
存储装置执行命令
616
目标装置响应于请求装置
618

图 6B

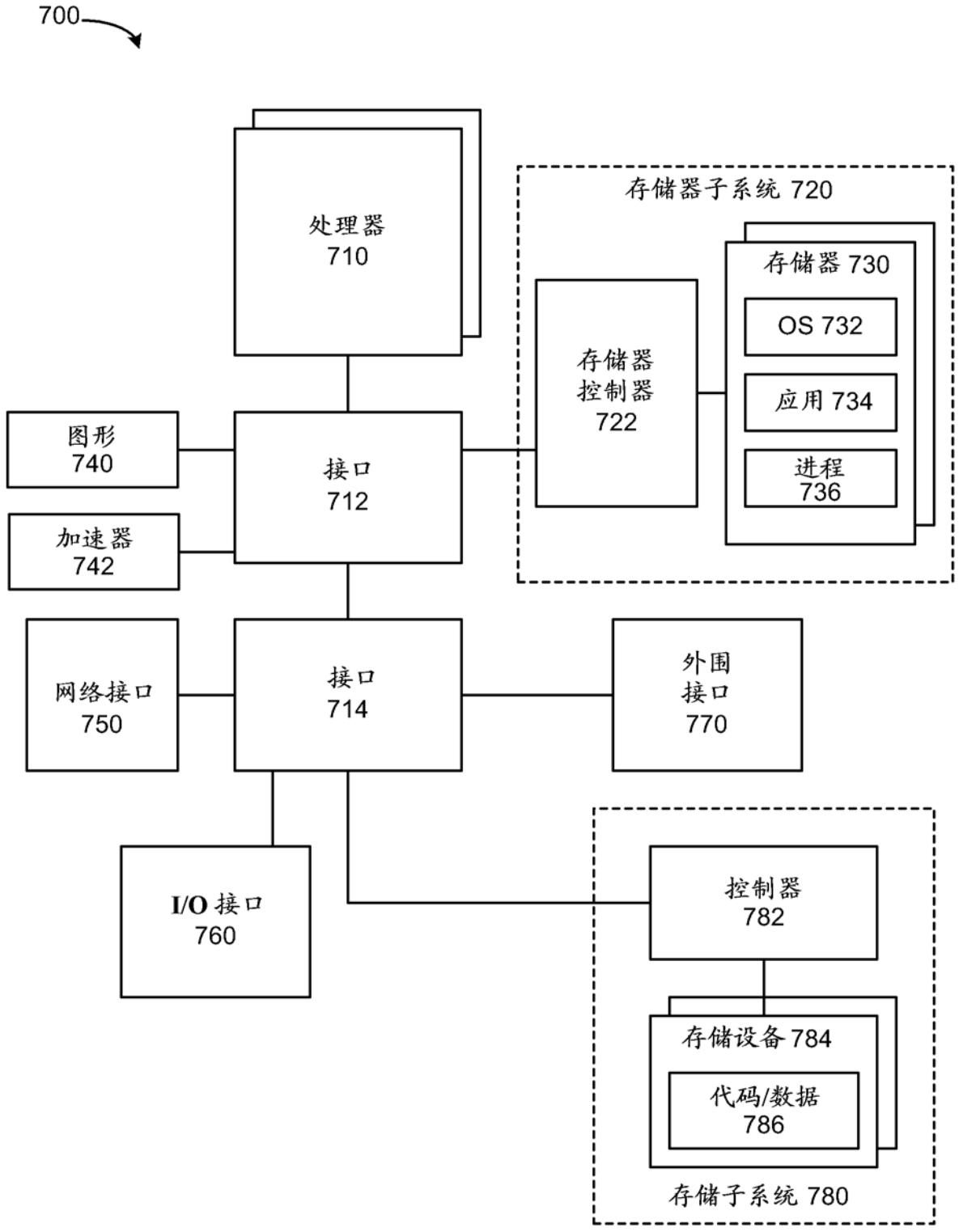

700
处理器
710
存储器子系统 720
存储器 730
OS 732
应用 734
进程
736
存储器
控制器
722
图形
740
接口
712
加速器
742
网络接口
750
接口
714
外围
接口
770
I/O 接口
760
控制器
782
存储设备 784
代码/数据
786
存储子系统 780

图 7

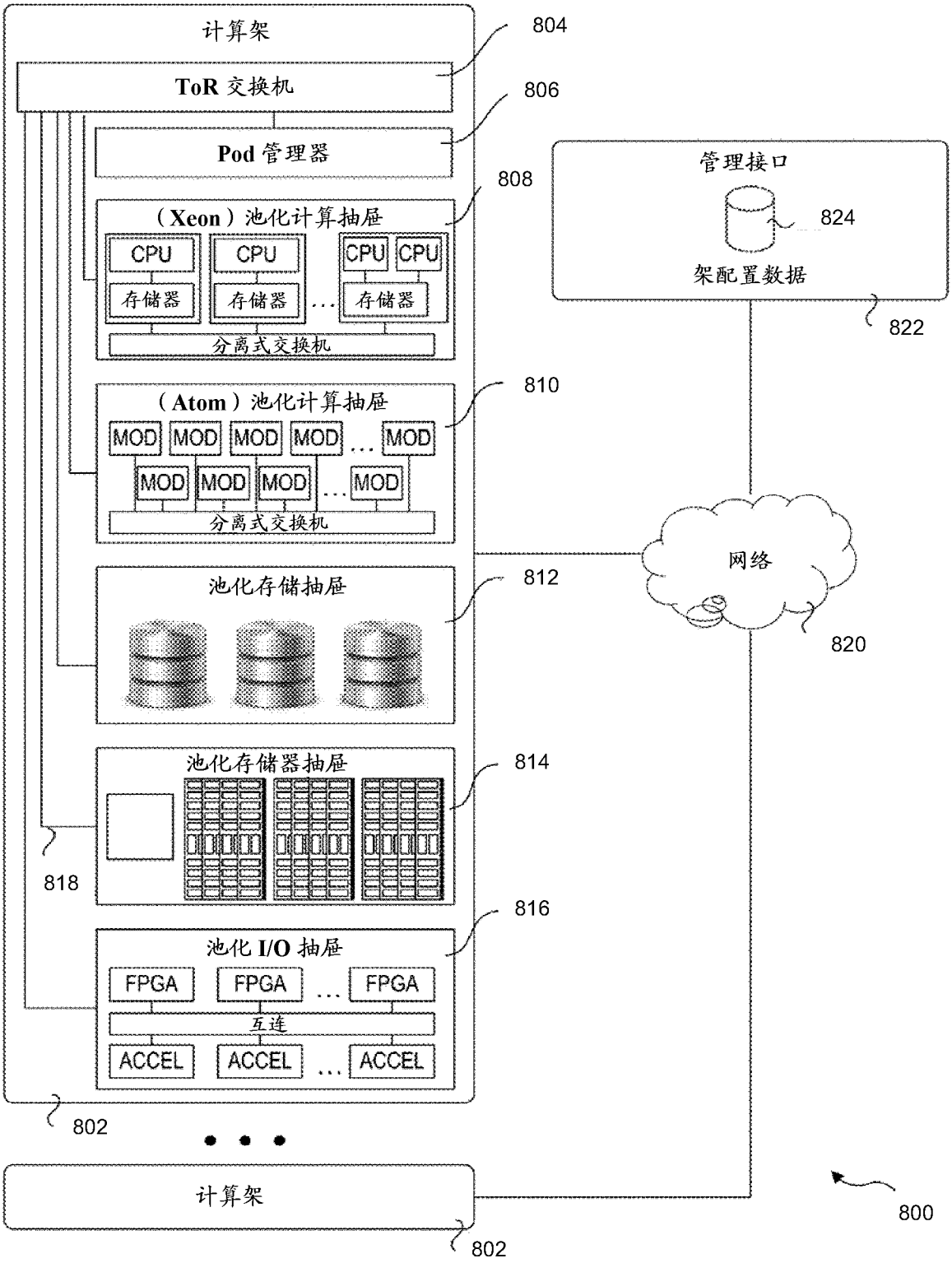

计算架
804
ToR 交换机
806
Pod 管理器
808
（Xeon）池化计算抽屉
CPU
CPU
CPU
CPU
存储器
存储器
存储器
分离式交换机
管理接口
824
架配置数据
822
（Atom）池化计算抽屉
810
MOD
MOD
MOD
MOD
MOD
MOD
MOD
MOD
MOD
分离式交换机
网络
820
池化存储抽屉
812
池化存储器抽屉
814
818
池化 I/O 抽屉
816
FPGA
FPGA
FPGA
互连
ACCEL
ACCEL
ACCEL
802
计算架
802
800

图 8

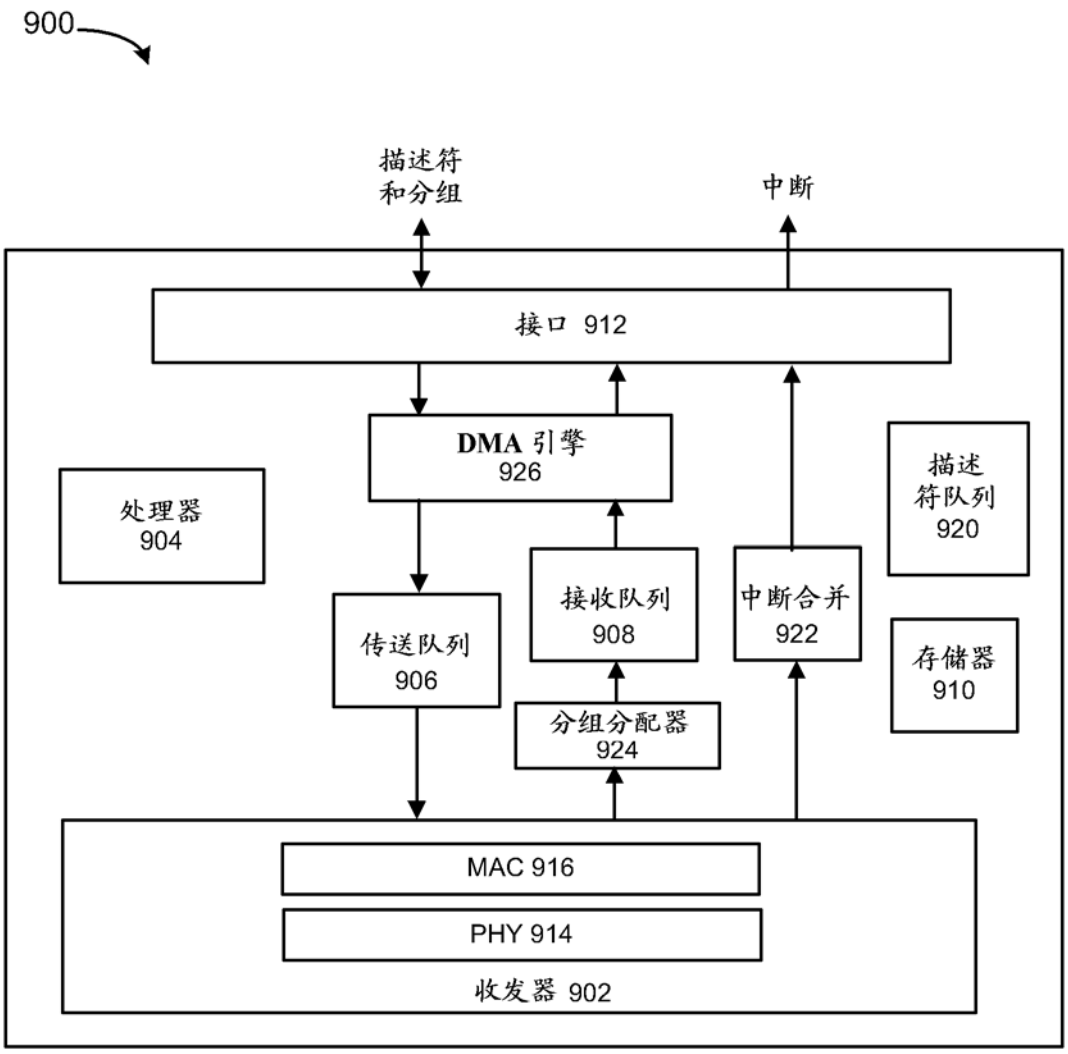

900
描述符
和分组
中断
接口 912
DMA 引擎
926
处理器
904
描述
符队列
920
传送队列
906
接收队列
908
中断合并
922
存储器
910
分组分配器
924
MAC 916
PHY 914
收发器 902

图 9