



(10) 授权公告号 CN 113687770 B

(45) 授权公告日 2023. 03. 24

(21) 申请号 202110405599.7

(22) 申请日 2021.04.15

(65) 同一申请的已公布的文献号
申请公布号 CN 113687770 A

(43) 申请公布日 2021.11.23

(30) 优先权数据
16/878,444 2020.05.19 US

(73) 专利权人 慧与发展有限责任合伙企业
地址 美国得克萨斯州

(72) 发明人 B·P·勒库耶 C·图弗里
S·拉赫尔

(74) 专利代理机构 北京市金杜律师事务所
11256

专利代理师 傅远

(51) Int.Cl.

G06F 3/06 (2006.01)

G06F 9/50 (2006.01)

G06F 9/54 (2006.01)

(56) 对比文件

CN 110365604 A, 2019.10.22

CN 110825670 A, 2020.02.21

CN 107003943 A, 2017.08.01

US 2018173462 A1, 2018.06.21

US 2017249104 A1, 2017.08.31

US 2002059439 A1, 2002.05.16

审查员 姜玲玲

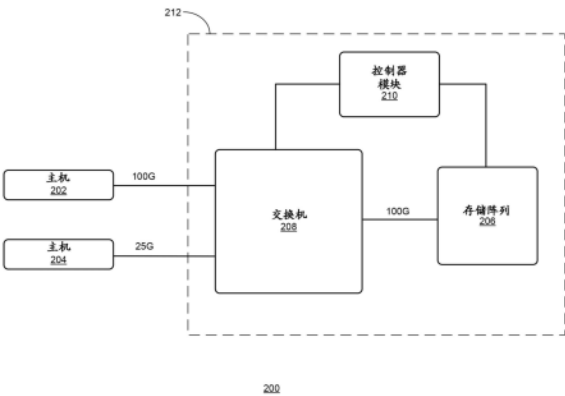
权利要求书3页 说明书10页 附图7页

(54) 发明名称

调节跨速率失配网络的NVMe-oF命令请求和数据流的系统和方法

(57) 摘要

本公开的实施例涉及用于调节跨速率失配网络的NVMe-oF命令请求和数据流的系统和方法。一个实施例可以提供一种用于实现流控制的方法和系统。在操作期间,交换机标识来自主机的命令以访问耦合到该交换机的存储设备。交换机使命令在与主机相对应的命令队列中排队。响应于确定待从存储设备传输到主机的数据量低于预先确定的阈值,交换机从命令队列中移除命令并且将所移除的命令转发到存储设备。



1. 一种方法,包括:

通过将多个主机耦合到存储设备的交换机,从所述主机接收命令以访问所述存储设备;

在所述交换机处将来自所述多个主机中不同主机的所述命令排入不同的主机专用命令队列中;

基于由所述存储设备的控制器维护的对应主机专用工作请求队列中的条目,确定从所述存储设备到所述不同主机中的相应主机的待传输的数据的总量,其中所述对应主机专用工作请求队列对应于所述相应主机,其中所述对应主机专用工作请求队列中的所述条目中的相应条目指示待传输数据帧的长度,其中所述待传输的数据的所述总量是基于所述条目中的每个条目所指示的所述长度而被确定的;

响应于确定所述待传输的数据的所述总量低于预先确定的阈值,从对应于所述相应主机的主机专用命令队列中移除命令;

将所移除的命令转发给所述存储设备的所述控制器;

通过所述控制器在所述对应主机专用工作请求队列中创建一个或多个附加条目,所述一个或多个附加条目与被转发给所述控制器的所移除的命令相关联;以及

响应于确定所述待传输的数据的所述总量超过所述预先确定的阈值,在所述交换机处停止将来自所述主机专用命令队列的所述命令转发到所述控制器。

2. 根据权利要求1所述的方法,其中主机专用命令队列的总数对应于耦合到所述交换机的所述多个主机的总数。

3. 根据权利要求1所述的方法,其中所述预先确定的阈值基于通过所述存储设备分配给所述相应主机的存储器空间的大小来确定。

4. 根据权利要求1所述的方法,包括:

将从所述相应主机接收的命令划分为多个较小命令;以及

将所述多个较小命令放置到对应于所述相应主机的所述主机专用命令队列中。

5. 根据权利要求1所述的方法,其中所述对应主机专用工作请求队列中的所述相应条目指示所述待传输数据帧的位置。

6. 根据权利要求1所述的方法,包括:

协商所述相应主机与所述存储设备之间的链接速率,其中所述相应主机经由第一通信链路耦合到所述交换机,其中所述存储设备经由第二通信链路耦合到所述交换机,并且其中所述第一通信链路和所述第二通信链路具有不同的链接速率;以及

基于所确定的链接速率和所述待传输数据帧的所述长度来设置倒数计时器的起始值。

7. 根据权利要求6所述的方法,其中所述相应主机和所述存储设备使用结构上的非易失性存储器快速NVMe-oF技术彼此通信,并且其中协商所述相应主机与所述存储设备之间的所述链接速率包括:在执行NVMe-oF连接建立的同时,交换与所述不同的链接速率相关联的信息。

8. 根据权利要求6所述的方法,其中协商所述链接速率包括:

通过耦合到所述交换机的控制器单元确定所述第一通信链路的链接速率;

通过所述控制器单元确定所述第二通信链路的链接速率;以及

基于所确定的所述第一通信链路和所述第二通信链路的链接速率来确定所述相应主

机与所述存储设备之间的所述链接速率。

9. 根据权利要求6所述的方法, 其中所述倒数计时器的所述起始值等于用于使用所协商的所述链接速率传输所述待传输数据帧所需的持续时间。

10. 根据权利要求6所述的方法, 包括:

在将所述待传输数据帧传输到所述相应主机的同时, 运行所述倒数计时器; 以及

在传输与所述对应主机专用工作请求队列中的下一条目相对应的后续数据帧之前, 等待所述倒数计时器到期。

11. 一种计算机系统, 包括:

处理器; 以及

存储器, 存储指令, 所述指令当由所述处理器执行时使得所述处理器执行方法, 所述方法包括:

通过将多个主机耦合到存储设备的交换机, 从所述主机接收命令以访问所述存储设备;

在所述交换机处将来自所述多个主机中不同主机的所述命令排入不同的主机专用命令队列中;

基于由所述存储设备的控制器维护的对应主机专用工作请求队列中的条目, 确定从所述存储设备到所述不同主机中的相应主机的待传输的数据的总量, 其中所述对应主机专用工作请求队列对应于所述相应主机, 其中所述对应主机专用工作请求队列中的所述条目中的相应条目指示待传输数据帧的长度, 其中所述待传输的数据的所述总量是基于所述条目中的每个条目所指示的所述长度而被确定的;

响应于确定所述待传输的数据的所述总量低于预先确定的阈值, 从对应于所述相应主机的主机专用命令队列中移除命令;

将所移除的命令转发给所述存储设备的所述控制器;

通过所述控制器在所述对应主机专用工作请求队列中创建一个或多个附加条目, 所述一个或多个附加条目与被转发给所述控制器的所移除的命令相关联; 以及

响应于确定所述待传输的数据的所述总量超过所述预先确定的阈值, 在所述交换机处停止将来自所述主机专用命令队列的所述命令转发到所述控制器。

12. 根据权利要求11所述的计算机系统, 其中主机专用命令队列的总数对应于耦合到所述交换机的所述多个主机的总数。

13. 根据权利要求11所述的计算机系统, 其中所述预先确定的阈值基于通过所述存储设备分配给所述相应主机的存储器空间的大小来确定。

14. 根据权利要求11所述的计算机系统, 其中所述方法还包括:

将从所述相应主机接收的命令划分为多个较小命令; 以及

将所述多个较小命令放置到对应于所述相应主机的所述主机专用命令队列中。

15. 根据权利要求11所述的计算机系统, 其中所述对应主机专用工作请求队列中的所述相应条目指示所述待传输数据帧的位置。

16. 根据权利要求11所述的计算机系统, 其中所述方法还包括:

协商所述存储设备与所述相应主机之间的链接速率, 其中所述相应主机经由第一通信链路耦合到所述交换机, 其中所述存储设备经由第二通信链路耦合到所述交换机, 并且其

中所述第一通信链路和所述第二通信链路具有不同的链接速率;以及

基于所确定的链接速率和所述待传输数据帧的所述长度来设置倒数计时器的起始值。

17. 根据权利要求16所述的计算机系统,其中所述相应主机和所述存储设备使用结构上的非易失性存储器快速NVMe-oF技术彼此通信,并且其中协商所述相应主机与所述存储设备之间的所述链接速率包括:在执行NVMe-oF连接建立的同时,交换与所述不同的链接速率相关联的信息。

18. 根据权利要求16所述的计算机系统,其中所述方法还包括:

通过耦合到所述交换机的控制器单元确定所述第一通信链路的链接速率;

通过所述控制器单元确定所述第二通信链路的链接速率;以及

基于所确定的所述第一通信链路和所述第二通信链路的链接速率来确定所述存储设备与所述相应主机之间的所述链接速率。

19. 根据权利要求16所述的计算机系统,其中所述倒数计时器的所述起始值等于用于使用所协商的所述链接速率传输所述待传输数据帧所需的持续时间。

20. 根据权利要求16所述的计算机系统,其中所述方法还包括:

在将所述待传输数据帧传输到所述相应主机的同时,运行所述倒数计时器;以及

在传输与所述对应主机专用工作请求队列中的下一条目相对应的后续数据帧之前,等待所述倒数计时器到期。

调节跨速率失配网络的NVMe-oF命令请求和数据流的系统和 方法

背景技术

[0001] 本公开一般涉及存储系统。更具体地,本公开涉及一种调节主机设备与存储阵列之间的流率的系统和方法。

[0002] 结构上的非易失性存储器快速 (NVMe-oF) 允许计算机服务器 (或主机) 通过网络与存储系统进行通信。例如,主机可以通过交换结构读取/写入位于单独硬件上的固态驱动器 (SSD)。主机上网络端口的速度可能与存储阵列的链接速度不同。主机结构链路和存储阵列结构链路之间的链接速率失配可能会导致整个链路带宽的显著未充分利用。

附图说明

[0003] 图1给出了根据现有技术的示例性存储系统。

[0004] 图2图示了根据一个实施例的示例性存储系统。

[0005] 图3给出了图示了根据一个实施例的存储阵列的示例性帧传输过程的流程图。

[0006] 图4给出了图示了根据一个实施例的用于调节命令请求的机制的图。

[0007] 图5给出了图示了根据一个实施例的调节命令请求的递送的示例性过程的流程图。

[0008] 图6图示了根据一个实施例的有助于包括NVMe-oF连接的设备的系统中的流控制的示例性装置。

[0009] 图7图示了根据一个实施例的有助于网络中的流控制的示例性计算机系统。

[0010] 在附图中,相似附图标记是指相同的附图元件。

具体实施方式

[0011] 给出以下描述以使得本领域的任何技术人员能够制造并使用实施例,并且在特定应用及其要求的上下文中提供以下描述。对于本领域技术人员而言,对所公开的实施例的各种修改是显而易见的,并且在没有背离本公开的精神和范围的情况下,本文中所定义的一般原理可以应用于其他实施例和应用。因此,本公开的范围不限于所示的实施例,而是要符合与本文中所公开的原理和特征一致的最宽范围。

[0012] 本文中所描述的实施例提供了一种用于实现NVMe-oF技术的联网设备中的端到端流控制的解决方案。流控制机制预先确定存储阵列与主机设备之间的传输流率,以防止丢帧并实现接近最大链路吞吐量。另外,为了防止存储阵列的存储资源在访问存储阵列的主机之间分布不均,其还可能造成存储阵列的总可用带宽的未充分利用,交换机可以实现用于在将从主机接收的命令请求递送到存储阵列之前对这些命令请求进行排队的命令队列。更具体地,该交换机维护命令队列集合,每个主机一个命令队列,如果待传输到该主机的数据超过预先确定的阈值,则其可以延迟新接收的命令请求从特定主机到存储阵列的提交。

[0013] 一个实施例可以提供一种用于实现流控制的方法和系统。在操作期间,交换机标识来自主机的命令以访问耦合到该交换机的存储设备。交换机将命令在与主机相对应的命

令队列中进行排队。响应于确定待从存储设备传输到主机的数据量低于预先确定的阈值，交换机从命令队列中移除命令并且将所移除的命令转发到存储设备。

[0014] 在该实施例的变型中，交换机维护多个命令队列。多个命令队列中的每个命令队列与耦合到交换机的不同主机相对应。

[0015] 在该实施例的变型中，预先确定的阈值基于通过存储设备分配给主机的存储器空间的大小来确定。

[0016] 在该实施例的变型中，交换机将命令划分为多个较小命令，并且将多个较小命令放置到命令队列中。

[0017] 在该实施例的变型中，存储设备的控制器基于该命令在与主机相对应的工作请求队列中生成条目。该条目指定与命令相关联的数据帧的大小。

[0018] 在又一变型中，主机经由第一通信链路耦合到交换机，存储设备经由第二通信链路耦合到交换机，并且第一通信链路和第二通信链路具有不同的链接速率。存储设备确定与主机的链接速率，并且基于所确定的链接速率和数据帧的大小来设置倒数计时器的起始值。

[0019] 在又一变型中，主机和存储设备使用结构上的非易失性存储器快速 (NVMe-oF) 技术彼此通信，并且通过在执行NVMe-oF连接建立的同时交换与不同链接速率相关联的信息来确定链接速率。

[0020] 在又一变型中，耦合到交换机的控制器模块被配置为确定第一通信链路的链接速率、确定第二通信链路的链接速率，并且基于所确定的第一通信链接速率和第二通信链接速率来确定链接速率。

[0021] 在又一变型中，倒数计时器的起始值等于用于使用所确定的链接速率传输数据帧所需的持续时间。

[0022] 在又一变型中，存储设备在将数据帧传输到主机的同时，运行倒数计时器；并且在传输与工作请求队列中的下一条目相对应的后续数据帧之前，等待倒数计时器到期。

[0023] NVMe-oF的流控制

[0024] 图1给出了根据现有技术的示例性存储系统。存储系统100可以包括多个主机(例如，主机102和主机104)，存储阵列(例如，SSD存储阵列)106、以及交换结构108，该交换结构108包括一个或多个互连交换机。主机和存储阵列106可以支持NVMe-oF，从而允许主机经由交换结构108访问(例如，读取或写入)存储阵列106。

[0025] 在图1中，存储阵列106经由高速链路(例如，100Gbps链路)连接到交换结构108，并且主机102也经由高速链路连接到交换结构108。然而，主机104经由低速(例如，25Gbps链路)连接到交换结构108。主机104与存储阵列106之间的链接速率之差可能导致存储阵列106的高速链路的未充分利用。更具体地，当主机104从存储阵列106读取数据时，数据以100Gbps的速率提供，这很容易使主机104的25Gbps链路不堪重负(例如，在25Gbps链路上造成拥塞)。

[0026] 处置主机的较慢链路与存储阵列的较快链路之间的速率差异的一种现有途径是在交换结构内缓冲通过较快链路发送的帧，同时等待较慢链路的输出队列(在该示例中，指派给主机104的输出队列)中的可用空间。然而，交换结构可以提供的缓冲量受限，并且在没有端到端控制来减慢发送方(在该示例中，存储阵列106)的情况下，缓冲器空间最终会被耗

尽,从而产生丢失的帧。

[0027] 为了实现端到端流控制,一种现有途径检测到由于交换机缓冲器被耗尽而导致的丢帧的发生,并且请求发送方(例如,存储阵列或主机)退避传输(例如,抑制提供所请求的数据或发送附加请求)。另一现有途径依赖于当交换机缓冲器被占用到一定阈值时,交换机向发送方发送显式拥塞通知(ECN)分组。两种途径均包括以较低速率开始传输(例如,发送帧),然后逐渐增加传输速率,直至检测到丢失的帧或生成ECN分组。继检测到拥塞之后(例如,通过接收ECN分组),发送方降低其传输速率。例如,发送方可以将其传输速率降低到一半,然后逐渐提高传输速率,直至检测到另一拥塞事件。这可能会产生锯齿形性能曲线,其中可用链路的使用率介于50%与100%之间。

[0028] 减轻拥塞的附加方案使用暂停帧来阻止发送方使交换机缓冲器溢出。然而,这种暂停机制会导致对线路头部阻塞,并且随着交换机跳数的增加,系统性能也会显著降低。

[0029] 为了在没有前述问题的情况下为存储系统提供端到端流控制,本文中所讨论的一些实施例提供了用于预先确定存储阵列与主机设备之间的流率应为多少的机制。更具体地,在主机与存储阵列的控制器之间建立通信的初始阶段期间,可以交换设备的有效链接速率。例如,在初始握手期间,主机和存储阵列可以交换其链接速率。如果主机的链接速率小于存储阵列的链接速率,则存储阵列上的网络接口可以使用主机的较低链接速率来控制帧发送回到主机的速度。可替代地,交换机的管理模块(其可以是硬件模块或可由处理器执行的软件程序)可以用于获得网络中的设备(例如,主机和存储阵列)的物理链路接口速度。

[0030] 图2图示了根据一个实施例的示例性存储系统。存储系统200可以包括多个主机(例如,主机202和204)、一个或多个存储阵列(例如,存储阵列206)、交换机208、以及控制器模块210。主机202和204可以经由各种类型的有线连接或无线连接耦合到交换机208。同样,存储阵列206还可以经由各种类型的有线连接或无线连接耦合到交换机208。主机202和204可以经由交换机208访问存储阵列206。控制器模块210可以与交换机208以及存储阵列206进行通信。在一些实施例中,存储阵列(例如,存储阵列206)、交换机208和控制器模块210可以放置在物理封壳212内部。物理封壳212可以符合预先确定的形状因子(例如,能够适配到标准机架上的插槽中)。可替代地,存储阵列、交换机208和控制器模块210中的每个都可以是独立硬件模块。不同的硬件模块可以位于不同的物理位置处,并且可以经由有线耦合或无线耦合彼此耦合。

[0031] 在一些实施例中,交换机208可以包括与以太网协议兼容的交换机。依据应用,该交换机还可以与其他类型的通信协议(例如,光纤通道或InfiniBand®协议)兼容。在一些实施例中,存储阵列206可以包括多个物理驱动器,诸如SSD和硬盘驱动器(HDD)。在一些实施例中,存储阵列206的网络接口(例如,网络接口控制器(NIC))的硬件可以维护工作请求(例如,数据传输请求)队列集合,每个主机一个队列,其中每个工作请求描述包含要通过链接发送到对应主机的数据的存储器区域。这些队列可以被描述为工作请求队列。而且,每个队列都可以维护链接速率和倒数计时器,该倒数计时器跟踪发送下一帧的时间。更具体地,当存储阵列206建立与对应主机的通信时,可以在初始握手过程期间确定链接速率。可替代地,代替使存储阵列206与主机协商链接速率,系统可以依赖控制器模块210来确定链接速率。控制器模块210与存储阵列206和交换机208通信,并且可以获知交换机208任一侧上的链接速率。因此,它可以确定存储阵列206与特定主机通信所应当使用的速率。然后,控

制器模块210可以将这样的速率信息发送到存储阵列206。

[0032] 在图2所示的示例中,存储阵列206的链接速率为100Gbps,主机202的链接速率为100Gbps,并且主机204的链接速率为25Gbps。在初始握手期间,存储阵列206和主机202可以协商其通信链接速率为100Gbps,这意味着在将数据提供给主机202时,存储阵列206可以100Gbps的速率传输数据。同样,在初始握手期间,存储阵列206和主机204可以协商其通信链接速率为25Gbps。存储阵列206所保存的用于主机202的工作请求队列可以包括用于将数据传输到主机202的未决请求,并且可以维护指示所协商的链接速率为100Gbps的记录,而用于主机204的工作请求队列可以包括用于将数据传输到主机204的未决请求并且可以维护指示所协商的链接速率为25Gbps的记录。每个工作请求可以指示待传输帧的存储器位置及其长度。

[0033] 主机202的工作请求队列也可以维护倒数计时器,该倒数计时器可以用于确定用于传输下一帧的时刻。定时器的起始值可以基于即将待传输帧的长度(即,立即待传输帧)和所协商的链接速率来确定。在图2所示的示例中,存储阵列206的网络接口硬件可以基于主机204的工作请求队列中保存的记录来确定主机204的所协商的链接速率为25Gbps。如果存储阵列206即将向主机204发送长度为2500个字节的帧,则可以确定传输该帧所需的时间量为 $2500 \text{ 字节} \times 8 \text{ 位/字节} / 25 \text{ Gbps} = 800 \text{ ns}$ 。因此,主机202的工作请求队列可以将其定时器起始值设置为800ns。存储阵列206的网络接口被配置为使得直至倒数计时器达到零之前都不会传输帧。换句话说,继开始传输帧并设置倒数计时器的起始值之后,存储阵列206的网络接口在开始传输下一帧之前等待倒数计时器达到零。应当指出,因为存储阵列206的链接速率为100Gbps,所以存储阵列206完成2500字节帧的传输所需的时间量为 $2500 \text{ 字节} \times 8 \text{ 位/字节} / 100 \text{ Gbps} = 200 \text{ ns}$ 。这意味着,在完成2500字节帧的传输之后,存储阵列206需要等待附加600ns,然后才能开始将其输出队列中的下一帧传输到主机202。另一方面,因为存储阵列206和主机202的链接速率匹配,所以当存储阵列206正在传输到主机202时,帧之间无需附加等待时间。

[0034] 图3给出了图示了根据一个实施例的存储阵列的示例性帧传输过程的流程图。在用于建立通信的初始阶段期间,存储阵列可以与访问该存储阵列的主机协商用于彼此进行通信的速率(操作302)。如果存储阵列经由链接速率低于交换结构与主机之间的链路的链接速率的链路耦合到交换结构,则所协商的链接速率将为主机的较低链接速率。一般而言,如果主机和存储阵列的链接速率失配,则所协商的通信速率可以是较低的链接速率。在一些实施例中,速率协商过程可以是NVMe-oF协议所定义的现有连接建立过程的扩展。更具体地,NVMe-oF连接建立握手可被扩展以包括端点的链接速率的交换。

[0035] 继协商链接速率之后,存储阵列可以为主机经由交换结构所请求的数据传输到主机做准备。起初,存储阵列准备要传输到主机的多个帧(操作304)。每个帧的大小和帧的数目可以通过诸如主机所请求的数据量和交换结构所实现的通信协议之类的多个因子确定。每个帧的位置和大小可以存储在存储阵列的网络接口所维护的工作请求队列中。存储阵列基于所协商的链接速率和立即待传输帧的大小来设置倒数计时器的初始值(操作306)。在传输开始时,如工作请求队列的头部条目所示,立即待传输帧可以是待传输到主机的一系列帧中的初始帧。如果已经传输了多个帧,则立即待传输帧可以是线路中的下一帧。倒数计时器的初始值被设置为等于以所协商的速率完成帧传输所需的时间量的时间值。例如,如

果所协商的速率为25Gbps并且第一帧的大小为2500字节,则倒数计时器的初始值可以被设置为以25Gbps的速度完成2500字节的传输所需的时间量,其为800ns。

[0036] 存储阵列开始传输帧,并且还启动倒数计时器(操作308)。可以基本同时启动帧的传输和倒数计时器。应当指出,在一些示例中,存储阵列可以以存储阵列的链接速率将帧从存储阵列传输出去,该链接速率可以高于所协商的速率。在这种场景中,该帧可以在被发送到主机之前,在交换结构中进行缓冲。在等待倒数计时器到期的同时,存储阵列可将数据传输到其他主机。

[0037] 继传输帧之后,存储阵列可以确定它是否是最后一个帧(操作310)。如果是,则传输结束。如果否,则存储阵列确定倒数计时器是否到期(操作312)。如果否,则存储阵列将等待倒数计时器到期。如果倒数计时器到期,则存储阵列基于下一帧来重置倒数计时器(操作306),随后启动帧传输和倒数计时器(操作308)。

[0038] 在图3所示的示例中,为每个立即待传输帧设置倒数计时器。该特定实现用于描述目的。在一些实施例中,一旦预先确定了链接速率(例如,通过链接速率协商),就可以以聚合水平而非逐帧完成对流率(例如,将帧从存储阵列发送到交换结构的速率)的计算。例如,流率可以通过除以在可编程时间间隔内发送的字节总数来计算,该可编程时间间隔可以使用倒数计时器进行编程。然后,存储阵列可以以以下方式配置:存储阵列仅在最新近编程的时间间隔内发送(多个)附加帧未超过所协商的流率的情况下,才发送(多个)附加帧(例如,在倒数计时器到期之后,发送附加帧)。换句话说,代替图3所示的逐帧传输示例,只要整体传输速率不超过所协商的链接速率,系统就可以控制任意大小的数据块的传输。在这些示例中,单个倒数计时器可以用于控制多个帧而不仅仅是单个帧的传输。

[0039] 通过针对每个帧传输在存储阵列处实现倒数计时器,所公开的实施例确保了对存储阵列的传输速率进行调节,使得主机侧的较慢链路不会被通过存储阵列进行的高速传输的重负压倒。然而,仅仅对存储阵列的传输速率的调节并不能解决所有拥塞。更具体地,具有较慢链路的主机可能会向存储阵列发出过多数目的命令请求(例如,读取请求),从而耗尽存储阵列上的存储器并且使得存储阵列延迟来自其他主机的命令请求的服务。这还可能造成存储阵列的链接速率的未充分利用。为了防止较慢主机耗尽存储阵列上的存储器资源,在一些实施例中,系统还可以调节来自不同主机的命令请求到达存储阵列的控制器的时间,以在具有各种链接速率的主机之间公平分布存储阵列的用于服务于命令请求的容量。

[0040] 在一些实施例中,为了调节来自主机的命令请求,交换结构可以维护命令队列集合,每个主机一个命令队列,其中每个命令队列存储从特定主机到达交换结构的命令请求。换句话说,来自特定主机的命令请求在被转发到存储控制器之前,在对应命令队列中进行排队。每个命令队列都维护指示对应主机正在待传输的数据量(例如,以字节为单位)的记录。当存储阵列控制器在存储阵列的存储器中为数据分配空间时,数据正在等待传输。

[0041] 图4给出了图示了根据一个实施例的用于调节命令请求的机制的图。在图4所示的示例中,多个主机(图4中未示出)可以经由交换机404将命令请求发送到存储阵列402。存储阵列402可以包括存储阵列控制器410和多个物理驱动器(例如,驱动器412、414和416)。交换机/存储管理模块406可以耦合到交换机404和存储阵列控制器410。

[0042] 交换机404可以维护多个命令队列(例如,队列422、424和426),这些命令队列可以

用于暂时保持从主机接收的命令请求。在一些实施例中,这些命令队列由位于交换机404上的硬件(例如,专用集成电路(ASIC)模块)维护。命令请求可以是读取请求或写入请求。命令队列的数目与耦合到交换机404的主机的数目相对应,其中给每个主机指派一个队列。依据发出到达交换机404的请求命令的主机的身份,交换机404将请求命令放置到对应队列中。交换机404所维护的这些命令队列有时可以称为“影子”队列,因为主机未获知这些队列。

[0043] 交换机404上的每个命令队列都可以维护正在待传输到交换机的数据量的记录。这种记录可以由交换机/存储管理模块406提供,该交换机/存储管理模块406不断检查在对应工作请求队列中待传输的数据量。工作请求队列中的条目指示待传输数据帧的存储器位置和长度。工作请求队列(例如,工作请求队列432、434和436)可以由存储阵列控制器410维护,每个主机一个队列。例如,工作请求队列432可以指派给特定主机,这意味着工作请求队列432中的每个条目都标识待传输帧的位置和长度。

[0044] 当来自主机的用于访问存储阵列402中存储的数据的命令请求到达交换机404时,代替将命令请求直接转发到存储阵列402,交换机404将命令请求存储在与发送命令请求的主机相对应的命令队列(例如,命令队列422)中。命令队列422维护指示待从存储阵列402传输到该特定主机的数据量的记录。应当指出,数据量可以与工作请求队列432中的所有条目所标识的帧的总大小相对应,该工作请求队列432与同一特定主机相对应。该数据量还反映了存储阵列控制器410已经为该特定主机供应的存储器资源的总量。在一些实施例中,交换机/存储管理模块406监测每个工作请求队列的状态,并且提供有关每个工作队列中的待传输到对应命令队列的总数据量的信息。例如,交换机/存储管理模块406可以监测工作请求队列432的状态,并且将有关工作请求队列432中的条目的信息提供给命令队列422。这种信息可以包括通过存储阵列402待传输到主机的总数据量,对应于从工作请求队列432到命令队列422。

[0045] 然后,命令队列422可以确定主机待传输的总数据量(即,与工作请求队列432中的条目相对应的数据量)是否超过预先确定的阈值。如果用于特定主机的待决数据超过该阈值,则可以认为存储阵列控制器410已经向该特定主机供应公平共享存储器资源。结果,交换机404可以暂时停止将命令请求从该特定主机转发到存储阵列控制器410,以防止存储阵列控制器410向该特定主机供应附加存储器资源。另一方面,如果特定主机的待决数据低于阈值,则交换机404可以开始将一个或多个接收的命令请求转发到存储阵列控制器410,以允许存储阵列控制器410在存储阵列402上分配存储器资源以缓冲数据。例如,如果命令请求是读取请求,则存储阵列控制器410可以在存储阵列402的驱动器中标识待读取数据,并且在其存储器中复制这样的数据。存储阵列控制器410可以在工作请求队列中为待读取数据创建一个或多个条目,其中工作请求队列中的条目描述了包含待读取数据的存储器块的位置和大小。

[0046] 在某些场景中,读取命令所请求的数据量可能会超过可用于供应给特定主机的存储器空间。例如,主机可以向交换机404发送读取请求,该读取请求请求存储阵列402上存储的大量(例如,32兆字节)数据。然而,可以供应给该主机的存储器空间的数量可以小于所请求的数量(例如,预先确定的阈值与当前供应给主机的存储器空间之间的差小于32兆字节)。在这种场景中,交换机404可以将单个大读取命令划分为多个(例如,至少两个)较小读取命令,并且在存储器空间可用时,可以将较小读取命令一一发送到存储阵列402。这样,代

替等待为该主机释放大块存储器空间以便递送大命令请求,交换机404可以在存储器空间的较小块变得可用时,开始将大命令请求的部分作为较小命令请求转发,这可能会更有效率。在一个示例中,响应于命令请求的大小超过预先确定的阈值,交换机404可以将命令请求划分为多个(例如,至少两个)较小命令请求,而与可用存储器空间的大小无关。在这种场景中,大命令请求可以在被排队到对应命令队列中之前,被划分为多个(例如,至少两个)较小命令请求。换句话说,交换机404可以将大命令划分为多个较小命令,然后将较小命令放置到对应命令队列中。

[0047] 当命令请求正在命令队列中进行调节或排队时,存储阵列402经由交换机404将工作请求队列中的条目所描述的数据传输到目的地主机。标准仲裁机制(例如,轮询(round robin))可以被用于存储阵列控制器410以从多个工作请求队列中选择用于传输的工作请求队列。每次传输数据帧时,对应主机的总待传输数据量就会减小。

[0048] 如先前所讨论的,图4所示的各种模块(诸如存储阵列402、交换机404和交换机/存储控制器406)可以放置在单个物理封壳内部以形成集成模块。在这种场景中,各种类型的连接(例如,光缆或电缆、印刷电路板(PCB)上的传输线等)可以用于形成存储阵列402与交换机404之间的高速链路。可选地,存储器阵列402、存储阵列402和交换机404可以各自是独立模块。交换机/存储控制器406可以是独立模块,也可以集成在存储阵列402或交换机404内部。在存储阵列402和交换机404是两个单独模块的实施例中,一个或多个电缆或光缆可以连接每个模块上的适当端口。

[0049] 图5给出了图示了根据一个实施例的调节命令请求的递送的示例性过程的流程图。在操作期间,耦合到存储模块的交换机从主机接收用于访问存储模块中的数据的数据访问命令(操作502)。交换机可以是允许主机设备取得对存储模块的访问的任何类型的交换机。除了单个交换机之外,该交换机还可以包括多个互连交换机。该存储模块可以包括任何类型的存储介质(例如,SSD或HDD)。交换机确定主机的身份(操作504),并且基于主机的身份来将所接收的数据访问命令放置到命令队列中(操作506)。应当指出,交换机维护每个主机的这种命令队列。在一些实施例中,命令队列可以是先进先出(FIFO)队列,其中较早条目在较晚条目之前从队列中移除。

[0050] 然后,交换机可以确定待从存储模块传输到所标识的主机的总数据量是否超过预先确定的阈值(操作508)。在一些实施例中,命令队列可以维护待传输到对应主机的总数据量的记录。这样的信息可以由与交换机和存储模块两者都通信的控制器模块提供。在一些实施例中,命令队列可以包括比较器,该比较器比较待传输数据总量和阈值。该阈值可以基于多个因子(诸如交换机模块的存储器的大小、主机的数目、每个主机的链接速率等)来确定。例如,如果存储模块的存储器的大小为100GB,则所有主机可以共享的数据缓冲量可能会略大于100GB(例如,110GB)。在这些示例中,由于瞬态数据,所以允许过量预订存储器资源。可以基于每个主机的链接速度来在多个主机之间划分这些有限的存储器资源。在一些实施例中,与具有较慢链路的主机相比,可以为具有较快链路的主机提供有较大存储器供应。例如,主机之间的存储器供应可能与它们的链接速度成比例。可替代地,存储器供应与链接速率之间的关系可以是非线性的。

[0051] 如果待传输数据总量超过阈值,则交换机等待存储阵列释放存储器空间(操作508)。当数据访问命令保持在交换机上的命令队列中时,存储模块使用与图3所示的过程相

似的过程来传输与工作请求队列中的条目相对应的待决数据。换句话说，存储模块的网络接口通过使用倒数计时器来实现流控制。继传输数据帧之后，对应存储器空间可以用于缓冲来自同一主机或来自不同主机的不同数据。

[0052] 如果待传输数据总量不超过阈值，则交换机可以确定可以供应给主机的剩余存储器空间是否足以服务于命令队列中的头部条目（操作510）。例如，交换机逻辑可以比较所请求的头部条目的存储器大小与可用存储器空间。在待传输数据总量不超过阈值的情形下，阈值与待传输数据总量之差表示可以供应给该特定主机的剩余存储器空间。如果可以供应给主机的剩余存储器空间足以服务于头部条目，则交换机可以将命令队列的头部条目转发到存储模块，以允许存储模块准备数据传输，其可以包括供应所需数据的存储器空间并且在工作请求队列中生成工作请求条目（操作512）。如果可以供应给主机的剩余存储器空间不足以服务于头部条目，则交换机可以将命令请求划分为多个较小命令请求，使得剩余存储器空间可以满足至少一个较小命令请求的存储器需求（操作514）。如果大命令请求的大小（例如，所请求的数据的大小）超过预先确定的阈值，则交换机也可以将大命令请求划分为多个较小命令请求，而与可以供应给主机的剩余存储器空间是否足以满足大命令请求无关。随后，交换机可以将至少一个命令转发到存储模块（操作512）。

[0053] 交换机可以确定该主机的命令队列中是否存在附加条目（操作516）。如果是，则交换机可以确定是否存在可用于该主机的存储器空间（操作508）。如果否，则操作结束。

[0054] 图6图示了根据一个实施例的示例性装置，该示例性装置有助于包括NVMe-oF连接设备的系统中的流控制。装置600可以包括多个单元或装置，其可以经由有线、无线、量子光或电通信信道彼此通信。装置600可以使用一个或多个集成电路来实现，并且可以包括比图6所示的单元或装置更少或更多的单元或装置。进一步地，装置600可以集成在计算机系统中，或被实现为能够与其他计算机系统和/或设备通信的（多个）单独设备。

[0055] 装置600可以包括用于与多个主机接口的主机接口602。主机可以经由主机接口602发送命令以访问存储设备上存储的数据。而且，主机可以通过经由主机接口602与存储设备的控制器（例如，存储控制器610）交换消息来建立与存储模块的通信。在一些实施例中，消息可以包括有关主机和存储设备的链接速率的信息。

[0056] 装置600可以包括多个命令队列604，其用于对经由主机接口602从主机接收的命令进行排队。命令队列的数目与耦合到主机接口602的主机的数目相匹配，其中每个命令队列被配置为存储从特定主机接收的命令请求。

[0057] 装置600可以包括多个比较器模块606。每个比较器模块可以耦合到命令队列。更具体地，每个比较器模块可以被配置为比较待通过存储模块传输到对应主机的数据量与预先确定的阈值。比较器模块606的输出可以用于控制对应命令队列604，从而允许对应命令队列将对命令队列的头部处的命令请求转发到存储模块。

[0058] 装置600可以包括交换机模块608，该交换机模块608能够在主机与存储模块之间路由命令和数据。装置600还可以包括存储控制器610，该存储控制器610用于控制耦合到交换机模块608的存储设备的操作（例如，数据缓冲和传输）。装置600可以包括多个工作请求队列612，该多个工作请求队列612用于对待处理工作请求进行排队。工作请求队列中的每个条目都描述了存储器中的位置以及待传输数据帧的大小。装置600可以包括多个倒数计时器614，其中每个倒数计时器耦合到对应工作请求队列。在传输了与工作请求队列中的条

目相对应的数据之后,工作请求队列的后续数据传输不会开始,直至倒数计时器返回至零。倒数计时器的初始值基于存储模块与对应于工作请求队列的主机之间的所协商的链接速率以及当前正在传输的帧的大小来设置。这可以确保存储模块不会将后续帧传输到主机,直至交换机模块608已经完成了对先前数据帧的传输。

[0059] 装置600可以包括工作请求处理模块616,该工作请求处理模块616被配置为处理工作请求队列612中的工作请求。工作请求处理模块616可以实现标准仲裁机制以从工作请求队列612中选择用于处理工作请求的工作请求队列。当处理工作请求时,工作请求处理模块616可以访问工作请求所指定的存储器位置以获得用于传输到对应主机的数据。装置600还可以包括数据传输模块618,该数据传输模块618被配置为将从存储模块获得的数据传输到主机。

[0060] 装置600可以包括交换机/存储控制器模块620,该交换机/存储控制器模块620能够控制交换机模块608和存储控制器610两者的操作。更具体地,交换机/存储控制器模块620可以促进耦合到交换机模块608的主机与存储控制器610所控制的存储模块之间的速率协商。交换机/存储控制器模块620可以促进将命令请求从命令队列604转发到存储控制器610。

[0061] 图7图示了根据一个实施例的有助于网络中的流控制的示例性计算机系统。计算机系统700包括处理器702、存储器704和存储设备706。计算机系统700可以耦合到显示设备710、键盘712、以及指向设备714,并且还可以经由一个或多个网络接口708耦合到网络708。存储设备706可以存储操作系统718、流控制系统720和数据742。

[0062] 流控制系统720可以包括指令,这些指令当由计算机系统700执行时,可以使得计算机系统700执行本公开中所描述的方法和/或过程。流控制系统720可以包括用于与主机接口的指令(主机接口模块722)、用于维护多个命令队列的指令(命令队列模块724)、用于维护多个比较器的指令(比较器模块726)、用于执行流量交换功能的指令(交换模块728)、用于控制存储模块的操作的指令(存储控制器模块730)、用于维护多个工作请求队列的指令(工作请求队列模块732)、用于维护多个倒数计时器的指令(倒数计时器模块734)、用于处理工作请求的指令(工作请求处理模块736)、用于将数据传输到主机的指令(数据传输模块738)、以及用于接口和控制交换机和存储控制器的操作的指令(交换机/存储控制器模块740)。

[0063] 一般而言,所公开的系统在包括链接速率失配的联网设备的系统中提供一种端到端流控制机制。更具体地说,经由不同速率的链路耦合到交换结构的两个终端设备可以在初始通信建立过程期间协商速率。当一个终端设备正在向另一终端设备传输帧时,传输终端设备的控制器可以设置倒数计时器,其中该倒数计时器的初始值被设置为使用所协商的速率传输帧所需的时间。可以同时启动帧的传输和倒数计时器。传输终端设备被配置为延迟后续帧的传输,直至倒数计时器到期。这可以防止传输终端设备以比所协商的速率更快的速率传输数据。除了调节帧传输速率之外,所公开的实施例还实现了一种机制,其用于延迟从一个终端设备到另一终端设备的数据访问命令的递送以防止请求终端设备耗尽响应终端设备的存储器资源。更具体地,当请求终端设备经由交换结构向响应终端设备发送数据请求命令时,交换结构将数据请求命令暂时保持在为请求终端设备分配的队列中。在队列的头部被移除并且被转发到响应终端设备之前,交换结构可以确定待从响应设备传输到

请求设备的数据量是否超过预先确定的阈值。阈值可以是响应终端设备上的最大存储器空间,可以分配该最大存储器空间来缓冲要传输到请求终端设备的数据。仅当待决数据量小于阈值时,队列的头部才可以被转发到响应设备。这可以确保响应终端设备的全部带宽可以由不同的请求终端设备使用,而无需一个请求终端设备耗尽存储器资源来阻止其他终端设备访问响应终端设备上的数据。这种新颖的流控制机制可以增加对联网设备的链接速率的使用,同时防止帧丢失并在访问单个终端设备上的数据的多个终端设备之间提供公平性。

[0064] 在图1至图7所示的各种示例中,终端设备可以是NVMe-oF启用的,并且可以包括主机和存储设备。实际上,还可以实现其他联网协议,并且交换结构可以包括能够进行流量交换的任何类型的交换器。而且,终端设备可以包括能够通过交换请求和响应而彼此通信的其他类型的终端设备。

[0065] 具体实施方式章节中所描述的方法和过程可以被体现为代码和/或数据,其可以存储在如上文所描述的计算机可读存储介质中。当计算机系统读取并执行计算机可读存储介质上存储的代码和/或数据时,计算机系统执行被体现为数据结构和代码并存储在计算机可读存储介质内的方法和过程。

[0066] 更进一步地,上文所描述的方法和过程可以包括在硬件模块或装置中。硬件模块或装置可以包括但不限于专用集成电路(ASIC)芯片、现场可编程门阵列(FPGA)、专用处理器或共享处理器、以及现在已知或未来开发的其他可编程逻辑设备,这些专用处理器或共享处理器在特定时间执行特定软件模块或一条代码。当激活硬件模块或装置时,它们执行其内包含的方法和过程。

[0067] 仅为了说明和描述起见已经给出了实施例的前述描述。它们不旨在是详尽的或将本公开的范围限制为所公开的形式。因而,对于本领域技术人员而言,许多修改和变化是显而易见的。

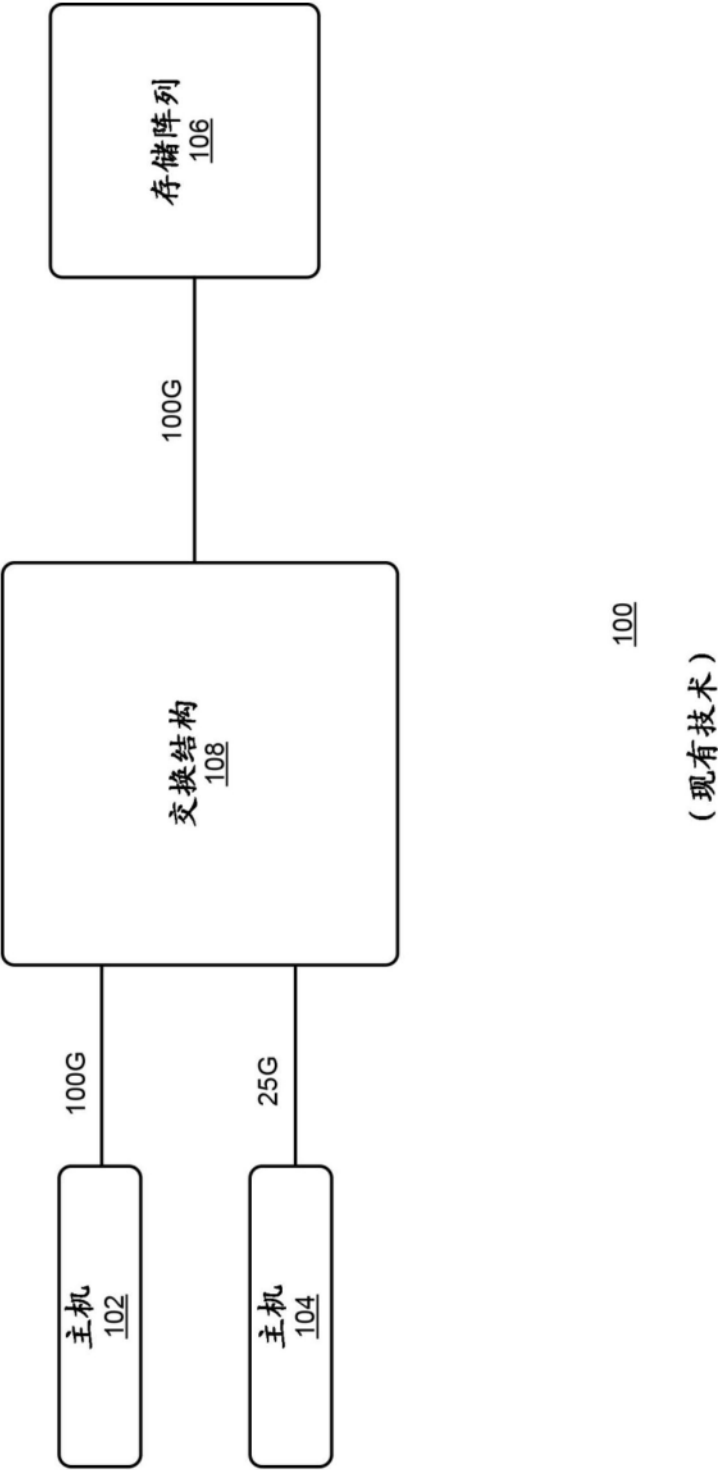


图1

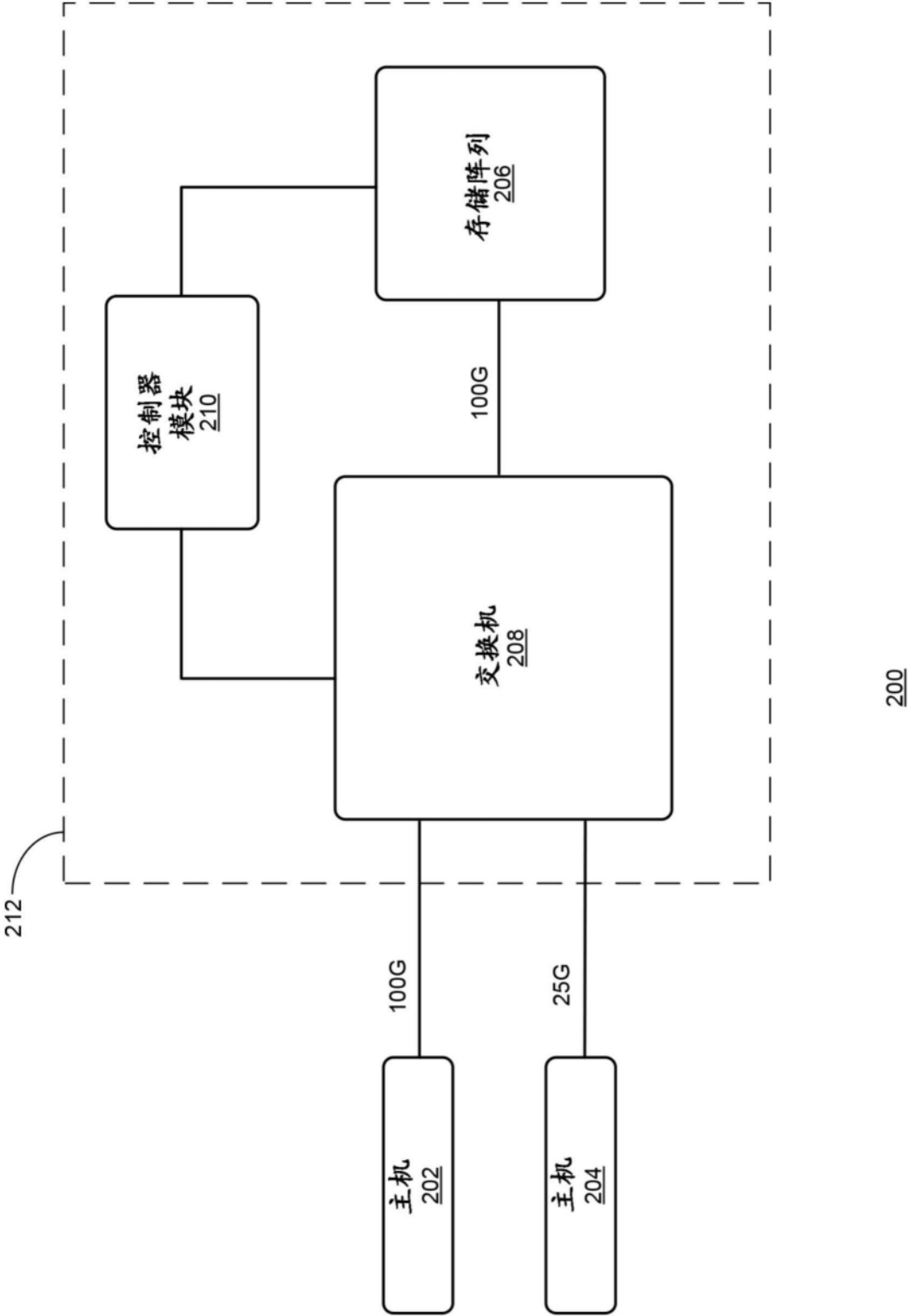


图2

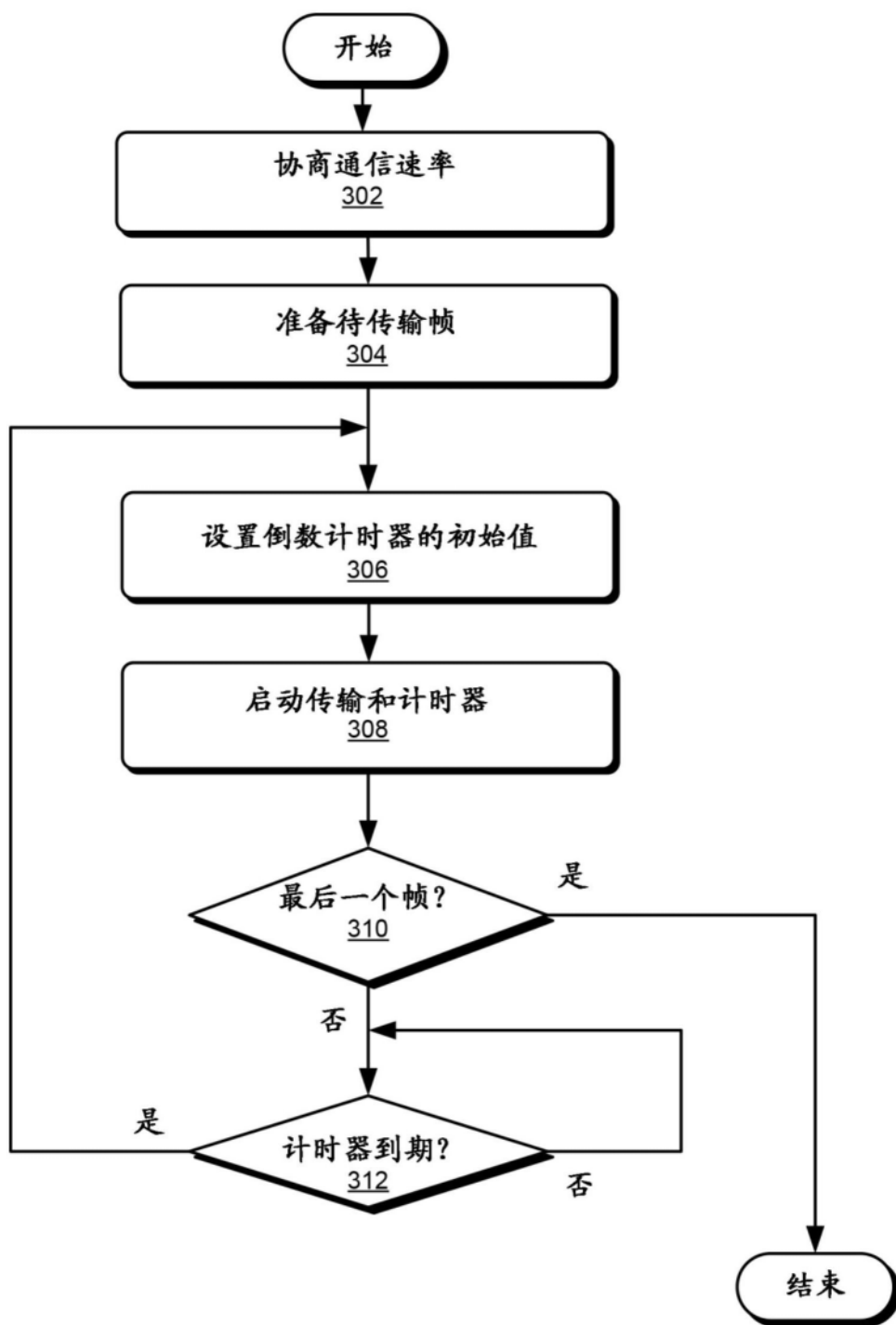


图3

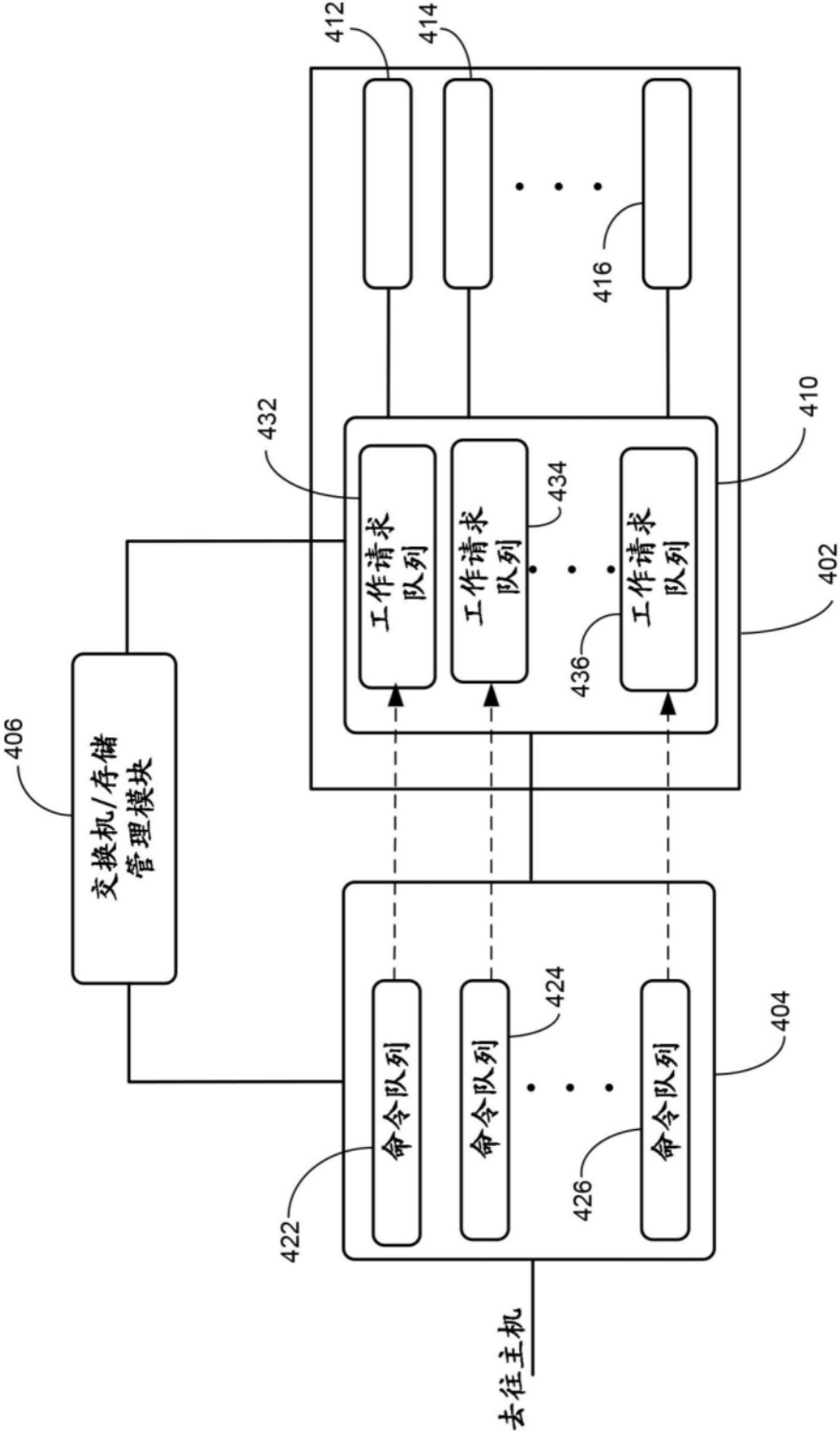


图4

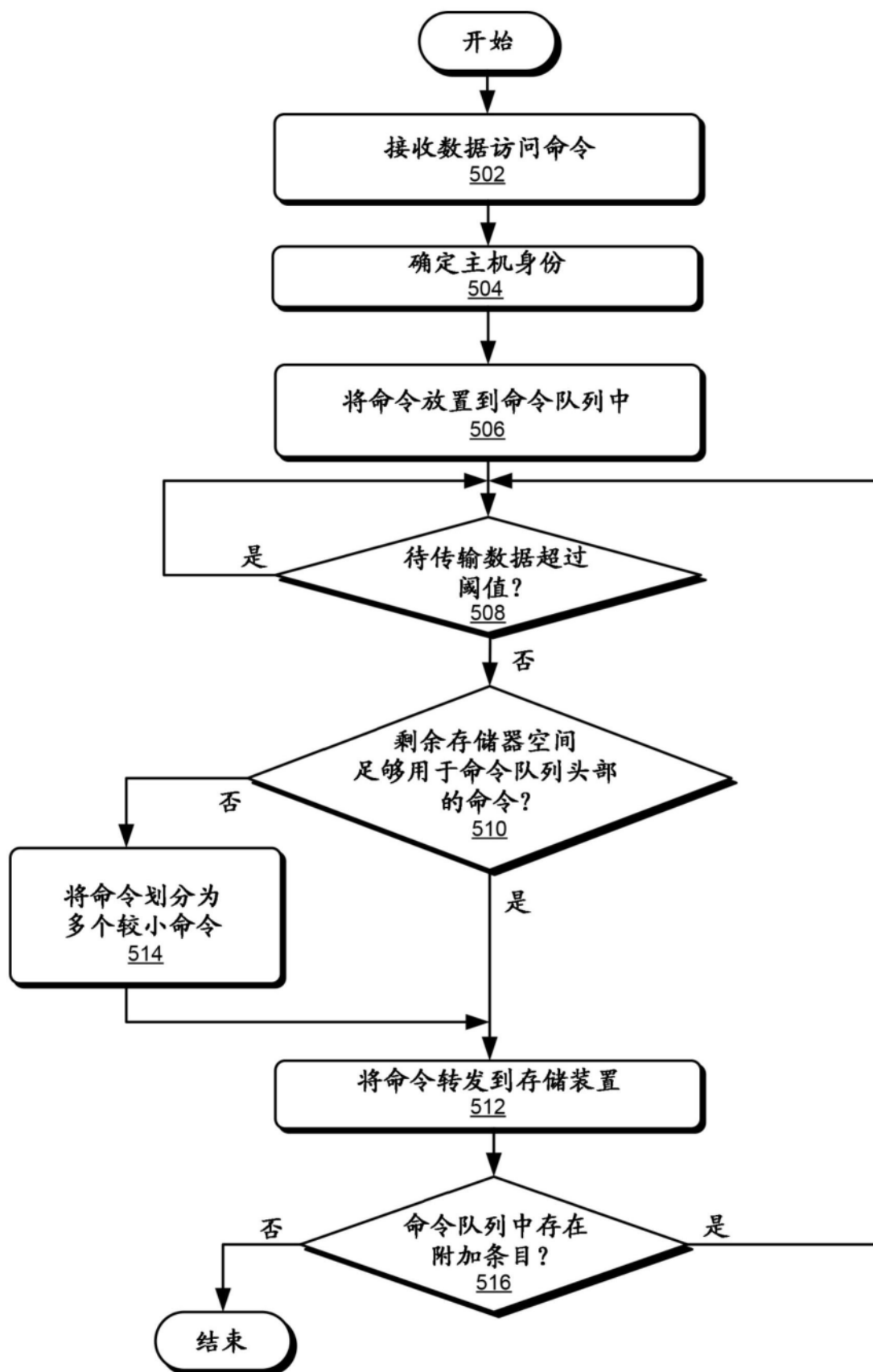


图5

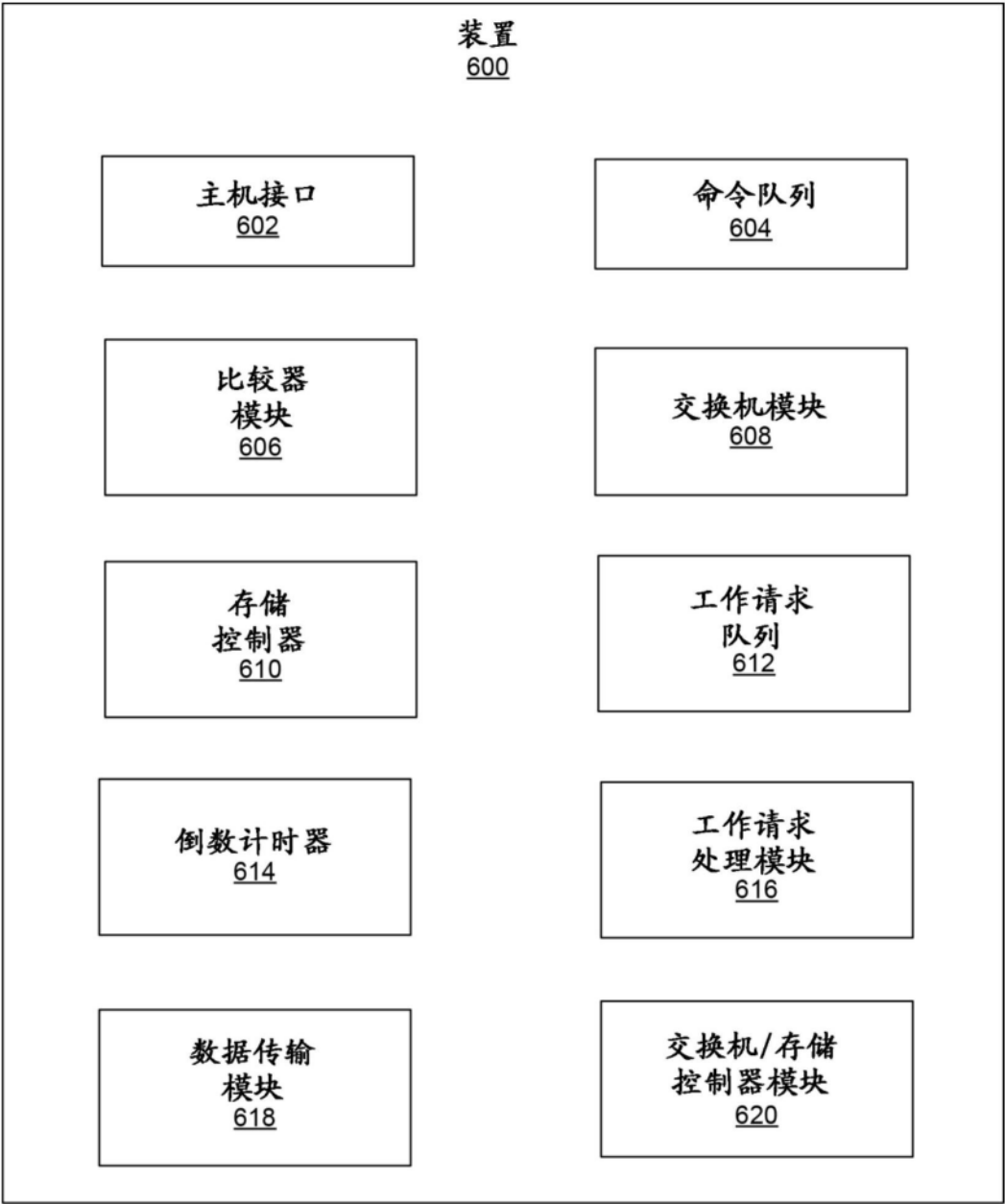


图6

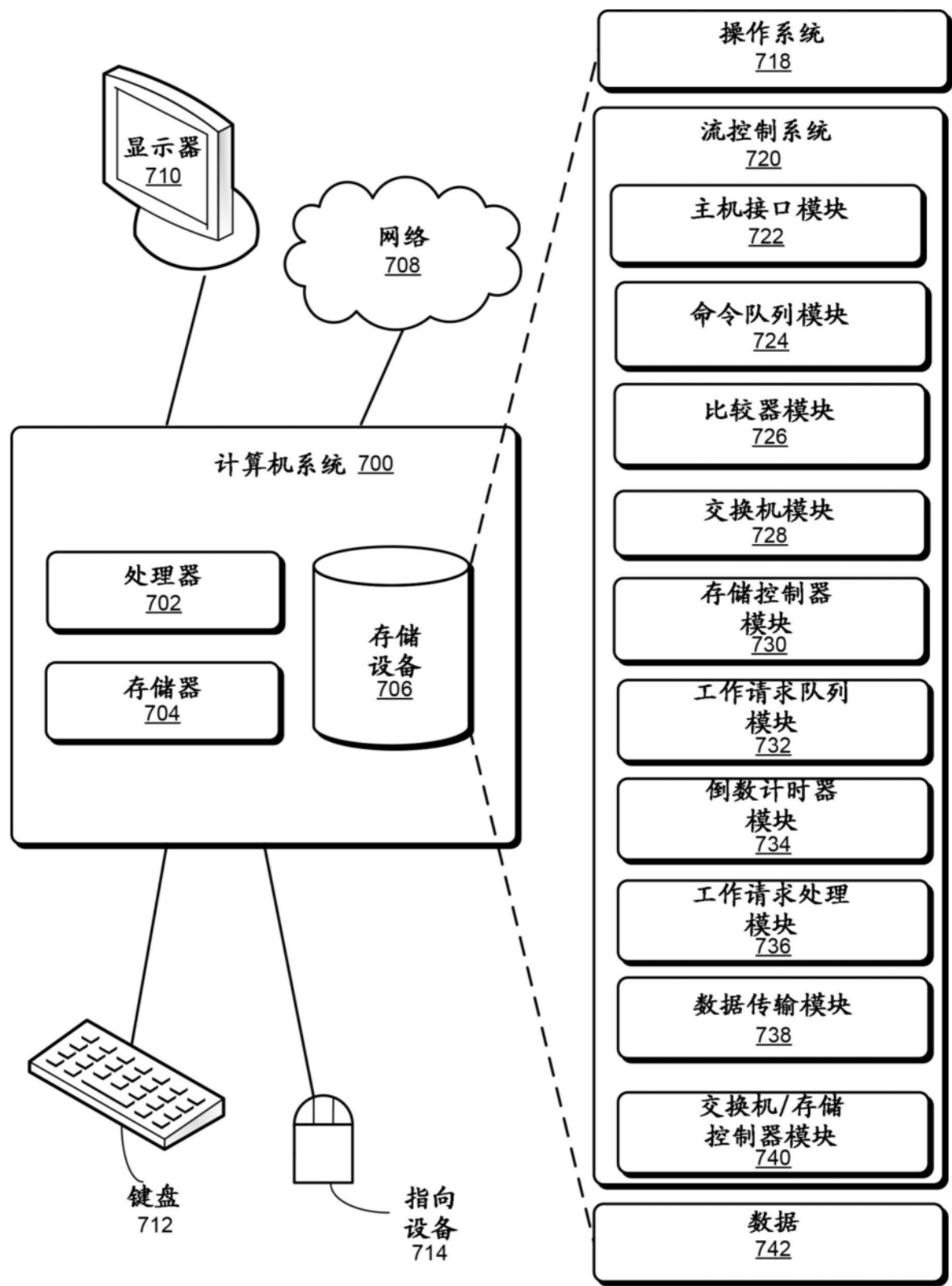


图7