(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110770710 B
(45)授权公告日 2023.09.05

(21)申请号 201880038486.4
(22)申请日 2018.05.02
(65)同一申请的已公布的文献号
申请公布号 CN 110770710 A
(43)申请公布日 2020.02.07
(30)优先权数据
62/500,794 2017.05.03 US
(85)PCT国际申请进入国家阶段日
2019.12.10
(86)PCT国际申请的申请数据
PCT/CA2018/050520 2018.05.02
(87)PCT国际申请的公布数据
WO2018/201249 EN 2018.11.08
(73)专利权人 艾德蒂克通信公司
地址 加拿大亚伯达
(72)发明人 肖恩·吉布 罗杰·伯特舒曼
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 周靖 杨明钊
(51)Int.Cl.
G06F 15/00(2006.01)
G06F 13/00(2006.01)
G06F 9/06(2006.01)
(56)对比文件
US 2015317091 A1,2015.11.05
US 2014281040 A1,2014.09.18
WO 2016135875 A1,2016.09.01
WO 2008027091 A1,2008.03.06
US 2015317088 A1,2015.11.05
US 2013198312 A1,2013.08.01
审查员 戴琦琦

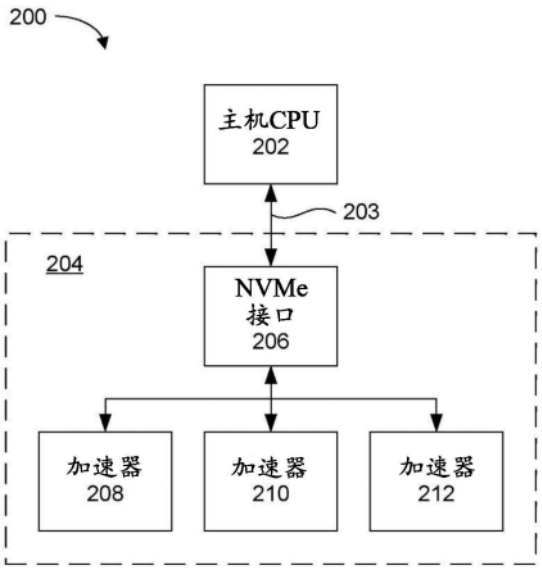
权利要求书2页 说明书10页 附图6页

(54)发明名称

用于控制数据加速的装置和方法

(57)摘要

提供了通过利用NVMe命令来控制硬件加速器从而促进执行硬件加速过程而无需使用软件和硬件特定的专用驱动器的系统和方法。所述NVMe命令可以是基于NVMe规范中提供的标准化NVMe命令,或者可以是由所述NVMe规范支持的供应商特定命令。所述命令通过主机CPU被发送到所述NVMe加速器,在一些实施例中,所述主机CPU可以定位在所述NVMe加速器远端。所述NVMe加速器可以包括CMB,主机CPU可以在其上设置NVMe队列,以减少将所述CPU与所述NVMe加速器连接的PCIe总线上的PCIe业务量。所述CMB还可以由主机CPU使用以传送用于加速的数据,以减小DMA控制器的带宽或清除主机分级缓存和存储器副本。



CN 110770710 B

1.一种用于控制硬件加速器的方法,所述方法包括:

在与所述硬件加速器通信、而与固态驱动器不相关联的NVMe接口处,从主机接收第一NVMe命令,所述第一NVMe命令具有磁盘读取或写入命令的格式但与磁盘读取或写入功能无关,并且所述第一NVMe命令包括多个命名空间中的第一命名空间,每个命名空间具有与固态驱动器(SSD)相关联的命名空间的格式,但与多个加速功能中的对应的一个相关联并且不与SSD相关联;

基于所述第一命名空间,通过所述NVMe接口确定与所接收到的第一NVMe命令相关联的加速功能,所述加速功能是除了磁盘读取或写入功能之外的功能;

在所述硬件加速器处执行加速功能以生成结果数据。

2.如权利要求1所述的方法,进一步包括:

在所述NVMe接口处从所述主机接收第二NVMe命令,所述第二NVMe命令与针对通过执行所述加速功能而生成的所述结果数据的请求相关联并且具有磁盘读取或写入功能的格式但与磁盘读取或写入功能无关,以及

响应于接收到所述第二NVMe命令,传输所述结果数据。

3.如权利要求2所述的方法,其中,从所述主机接收的所述第一NVMe命令被格式化为NVMe写入命令,并且所述第二NVMe命令被格式化为NVMe读取命令。

4.如权利要求2所述的方法,其中,所述第二NVMe命令包括所述第一命名空间,其中,所述第一NVMe命令和所述第二NVMe命令中的一个被格式化为针对所述第一命名空间的写入命令,并且所述第一NVMe命令和所述第二NVMe命令中的另一个被格式化为针对所述第一命名空间的读取命令。

5.如权利要求1所述的方法,进一步包括:

在所述NVMe接口处确定所述硬件加速器已完成执行所述加速功能;以及

将指示所述加速功能已经被执行的NVMe完成消息从所述NVME接口发送到所述主机。

6.如权利要求2所述的方法,其中,所述第一NVMe命令和所述第二NVMe命令是供应商特定命令。

7.如权利要求6所述的方法,其中,所述第一NVMe命令包括从其读取输入数据的存储设备的第一存储器地址,并且其中,除了对所述输入数据执行所述加速功能以生成所述结果数据,所述方法还包括从所述存储设备读取存储在所述第一NVMe命令中所包括的所述第一存储器地址处的所述输入数据。

8.如权利要求6所述的方法,其中,所述第二NVMe命令包括所述结果数据将被传输到的存储设备的第二存储器地址,并且其中,响应于接收到所述第二NVMe命令而传输所述结果数据包括将所述结果数据传输到所述存储设备以将所述结果数据存储在所述第二存储器地址处。

9.如权利要求1所述的方法,其中,接收所述第一NVMe命令包括经由将所述NVMe接口与所述主机连接的网络来接收所述第一NVMe命令。

10.如权利要求1所述的方法,其中,在所述NVMe接口处接收所述第一NVMe命令包括在所述NVMe接口的控制器存储缓冲器处接收所述第一NVMe命令。

11.一种用于执行加速过程的加速器,所述加速器包括:

NMVe接口和至少一个硬件加速器,所述至少一个硬件加速器与所述NVMe接口通信并且

被配置用于执行加速功能,其中,所述NVMe接口被配置用于:

从主机接收第一NVMe命令,所述第一NVMe命令具有磁盘读取或写入命令的格式但与磁盘读取或写入功能无关,并且所述第一NVMe命令包括多个命名空间中的第一命名空间,每个命名空间具有与固态驱动器(SSD)相关联的命名空间的格式,但与多个加速功能中的对应的一个相关联并且不与SSD相关联;

基于所述第一命名空间,确定与所接收到的第一NVMe命令相关联的加速功能,所述加速功能是除了磁盘读取或写入功能之外的功能;

用信号通知所述硬件加速器执行加速功能。

12.如权利要求11所述的加速器,其中,所述NVMe接口进一步被配置用于:

从所述主机接收第二NVMe命令,所述第二NVMe命令与针对通过执行所述加速功能而产生的结果数据的请求相关联并且具有磁盘读取或写入功能的格式但与磁盘读取或写入功能无关,并且

响应于接收到所述第二NVMe命令,用信号通知所述硬件加速器传输所述结果数据。

13.如权利要求12所述的加速器,其中,从所述主机接收的所述第一NVMe命令被格式化为NVMe写入命令,并且所述第二NVMe命令被格式化为NVMe读取命令。

14.如权利要求12所述的加速器,其中,所述第二NVMe命令包括所述第一命名空间,并且其中,所述第一NVMe命令和所述第二NVMe命令中的一个被格式化为针对所述第一命名空间的写入命令,并且所述第一NVMe命令和所述第二NVMe命令中的另一个被格式化为针对所述第一命名空间的读取命令。

15.如权利要求11所述的加速器,其中,所述NVMe接口进一步被配置用于:

确定所述硬件加速器已完成执行所述加速功能;并且

将指示所述加速功能已经被执行的NVMe完成消息发送到所述主机。

16.如权利要求12所述的加速器,其中,所述第一NVMe命令和所述第二NVMe命令是供应商特定命令。

17.如权利要求16所述的加速器,其中,所述第一NVMe命令包括从其读取输入数据的存储设备的第一存储器地址,并且其中,除了用信号通知所述硬件加速器对所述输入数据执行所述加速功能以生成所述结果数据,所述NVMe接口还被配置用于用信号通知所述硬件加速器从所述存储设备读取存储在所述第一NVMe命令中所包括的所述第一存储器地址处的所述输入数据。

18.如权利要求16所述的加速器,其中,所述第二NVMe命令包括所述结果数据将被传输到的第二存储器地址,并且其中,响应于接收到所述第二NVMe命令而传输所述结果数据包括将所述结果数据写入到所述第二存储器地址。

19.如权利要求11所述的加速器,其中,接收所述第一NVMe命令包括经由将所述NVMe接口与所述主机连接的网络来接收所述第一NVMe命令。

20.如权利要求11所述的加速器,进一步包括:控制器存储缓冲器(CMB),其中,在所述NVMe接口处接收所述第一NVMe命令包括在所述CMB处接收所述第一NVMe命令。

用于控制数据加速的装置和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求于2017年5月3日提交的美国临时专利申请号62/500,794的优选权权益,所述美国临时专利申请通过引用并入本文。

技术领域

[0003] 本公开涉及控制数据加速,包括但不限于算法和数据分析加速。

背景技术

[0004] 随着摩尔定律的预测终结,数据加速(包括算法和数据分析加速)已成为用于继续改进计算性能的主要研究课题。最初,通用图形处理单元(GPGPU)或视频卡是用于执行算法加速的主要硬件。最近,现场可编程门阵列(FPGA)对于执行加速而言越来越受欢迎。

[0005] 典型地,FPGA经由快速外围组件互连(PCIe)总线连接到计算机处理单元(CPU),其中,FPGA经由用于加速的特定软件和硬件平台专用的驱动器与CPU接口连接。在数据中心,已经开发出高速缓存一致性接口(包括一致性加速器处理器接口(CAPI)和高速缓存一致性互连(CCIX)),以便通过允许开发者应对与专有接口和驱动器相关联的固有困难来解决部署加速平台时的困难并且更快速地加速数据。

[0006] 非易失性存储器(NVM)(诸如闪速存储器)正越来越多地用于存储设备中。相较于旧的自旋磁盘介质,NVM固态驱动器(SSD)允许更快速的数据存储和检索。由于数据存储集中并且NVM SSD存储变得越来越普遍,所以期望这样的平台——实现更快地执行数据加速并且使用比目前已知的平台更少的功率。

[0007] 因此,期望对控制硬件加速的改进。

附图说明

[0008] 现在将参考附图仅通过举例来描述本公开的实施例。

[0009] 图1是根据现有技术的数据存储和加速系统的示意图。

[0010] 图2是根据本公开的利用NVMe接口的加速器系统架构的示意图;

[0011] 图3是根据本公开的利用NVMe接口的数据存储和加速系统的示意图;

[0012] 图4是根据本公开的用于利用NVMe接口执行加速的加速器系统的示意图;

[0013] 图5是根据本公开的用于利用NVMe接口通过网络执行加速的加速系统的示意图;并且

[0014] 图6是展示了根据本公开的用于控制硬件加速器的方法的流程图。

具体实施方式

[0015] 本公开提供了通过利用NVMe命令来控制硬件加速器从而促进执行硬件加速过程而无需使用软件和硬件特定的专用驱动器的系统和方法。所述NVMe命令可以是基于NVMe规范中提供的标准化NVMe命令,或者可以是由所述NVMe规范支持的供应商特定命令。所述命

令通过主机CPU被发送到所述NVMe加速器，在一些实施例中，所述主机CPU可以定位在所述NVMe加速器远端。所述NVMe加速器可以包括CMB，主机CPU可以在其上设置NVMe队列，以减少将所述CPU与所述NVMe加速器连接的PCIe总线上的PCIe业务量。

[0016] 本公开的实施例涉及利用非易失性存储器快速(NVMe)规范来控制硬件加速。

[0017] 在实施例中，本公开提供一种用于控制硬件加速器的方法，所述方法包括：在与所述硬件加速器相关联、而与固态驱动器不相关联的NVMe接口处，从主机接收第一NVMe命令，所述第一NVMe命令具有磁盘读取或写入功能的格式但与磁盘读取或写入功能无关；通过所述NVMe接口确定与所述接收到的第一NVMe命令相关联的加速过程；在所述硬件加速器处执行加速功能以生成结果数据。

[0018] 在示例实施例中，所述方法进一步包括：在所述NVMe接口处从所述主机接收第二NVMe命令，所述第二NVMe命令与针对通过执行所述加速功能而生成的所述结果数据的请求相关联并且具有磁盘读取或写入功能的格式但与磁盘读取或写入功能无关；以及响应于接收到所述第二NVMe命令，传输所述结果数据。

[0019] 在示例实施例中，从所述主机接收的所述第一NVMe命令是写入命令，并且所述第二NVMe命令是读取命令。

[0020] 在示例实施例中，所述第一命令和所述第二命令中的一个是针对多个命名空间中通常与SSD相关联的一个命名空间的写入命令，并且所述第一磁盘访问命令和所述第二磁盘访问命令中的另一个是针对所述多个命名空间中的所述一个命名空间的读取命令，其中，所述命名空间中的每一个都与对应的加速功能相关联。

[0021] 在示例实施例中，所述方法进一步包括：在所述NVMe接口处确定所述硬件加速器已完成执行所述加速功能；以及将指示所述加速功能已经被执行的NVMe完成消息从所述NVME接口发送到所述主机。

[0022] 在示例实施例中，所述第一NVMe命令和所述第二NVMe命令是供应商特定命令。

[0023] 在示例实施例中，所述第一NVMe命令包括所述结果数据将被写入到的第一存储器地址，并且其中，执行所述加速包括将所述结果数据写入到所述第一NVMe命令中包括的所述第一存储器地址。

[0024] 在示例实施例中，所述第二NVMe命令包括所述结果数据将被传输到的第二存储器地址，并且其中，响应于接收到所述第二NVMe命令而传输所述结果数据包括将所述结果数据写入到所述第二存储器地址。

[0025] 在示例实施例中，接收所述第一NVMe命令包括经由将所述NVMe接口与所述主机连接的网络来接收所述第一NVMe命令。

[0026] 在示例实施例中，在所述NVMe接口处接收所述第一NVMe命令包括在所述NVMe接口的控制器存储缓冲器处接收所述第一NVMe命令。

[0027] 在另一实施例中，本公开提供一种用于执行加速过程的加速器，所述加速器包括NMVe接口和至少一个硬件加速器，所述至少一个硬件加速器与所述NVMe接口通信并且被配置用于执行所述加速过程，其中，所述NVMe接口被配置用于：从主机接收第一NVMe命令，所述第一NVMe命令具有磁盘读取或写入功能的格式但与磁盘读取或写入功能无关；确定与所述接收到的第一NVMe命令相关联的加速过程；用信号通知所述硬件加速器执行加速功能。

[0028] 在示例实施例中，所述NVMe接口进一步被配置用于：从所述主机接收第二NVMe命

令,所述第二NVMe命令与针对通过执行所述加速功能而生成的所述结果数据的请求相关联并且具有磁盘读取或写入功能的格式但与磁盘读取或写入功能无关;并且响应于接收到所述第二NVMe命令,传输所述结果数据。

[0029] 在示例实施例中,从所述主机接收的所述第一NVMe命令是写入命令,并且所述第二NVMe命令是读取命令。

[0030] 在示例实施例中,所述第一命令和所述第二命令中的一个是针对多个命名空间中通常与固态驱动器(SSD)相关联的一个命名空间的写入命令,并且所述第一磁盘访问命令和所述第二磁盘访问命令中的另一个是针对所述多个命名空间中的所述一个命名空间的读取命令,其中,所述命名空间中的每一个都与对应的加速功能相关联。

[0031] 在示例实施例中,所述NVMe接口进一步被配置用于:确定所述硬件加速器已完成执行所述加速功能,并且将指示所述加速功能已经被执行的NVMe完成消息发送到所述主机。

[0032] 在示例实施例中,所述第一NVMe命令和所述第二NVMe命令是供应商特定命令。

[0033] 在示例实施例中,所述第一NVMe命令包括所述结果数据将被写入到的第一存储器地址,并且其中,执行所述加速包括将所述结果数据写入到所述第一NVMe命令中包括的所述第一存储器地址。

[0034] 在示例实施例中,所述第二NVMe命令包括所述结果数据将被传输到的第二存储器地址,并且其中,响应于接收到所述第二NVMe命令而传输所述结果数据包括将所述结果数据写入到所述第二存储器地址。

[0035] 在示例实施例中,接收所述第一NVMe命令包括经由将所述NVMe接口与所述主机连接的网络来接收所述第一NVMe命令。

[0036] 在示例实施例中,所述加速器包括命令存储缓冲器(CMB),其中,在所述NVMe接口处接收所述第一NVMe命令包括在所述CMB处接收所述第一NVMe命令。

[0037] 为了说明的简化和清楚,附图标记可以在附图当中重复以指示对应或相似的元件。阐述了许多细节,以提供对本文所描述的实施例的理解。在没有这些细节的情况下,也可以实践所述实施例。在其他实例中,并未详细描述众所周知的方法、程序和部件,以避免模糊所描述的实施例。

[0038] NVMe规范是响应于对计算机处理单元(CPU)与固态驱动器(SSD)之间的更快接口的需求而开发的协议。NVMe是用于访问经由快速外围组件互连(PCIe)总线连接到CPU的存储设备的逻辑设备接口规范,所述总线提供与旧接口相比效率更高的用于访问存储设备的接口并且被设计为具有所想到的非易失性存储器的特性。NVMe被设计为仅用于且传统上已经仅用于在存储设备上存储和检索数据,而并非用于控制硬件加速。

[0039] 在NVMe规范中,NVMe磁盘访问命令(诸如读取命令/写入命令)是使用命令队列从主机CPU被发送到存储设备的控制器的。通过管理队列处理控制器管理和配置,而输入/输出(I/O)队列处理数据管理。每个NVMe命令队列可以包括一个或多个提交队列和一个完成队列。经由提交队列将命令从主机CPU提供到存储设备的控制器,并且经由完成队列将响应返回到主机CPU。

[0040] 发送到管理队列和I/O队列的命令遵循相同的基本步骤来发出和完成命令。主机CPU在适当的提交队列中创建待执行的读取命令或写入命令,并且然后对与所述队列相关

联的tail doorbell寄存器进行写操作以用信号通知控制器已准备好执行提交条目。控制器在命令驻留在主机存储器中的情况下通过使用例如直接存储器访问(DMA)来获取读取命令或写入命令,或者在命令驻留在控制器存储器中的情况下直接获取读取命令或写入命令,并且执行所述读取命令或写入命令。

[0041] 一旦完成了对读取命令或写入命令的执行,控制器就将完成条目写入到相关联的完成队列。控制器可选地生成对主机CPU的中断,以用于指示存在待处理的完成条目。主机CPU提取并处理所述完成队列条目,并且然后针对所述完成队列对doorbell head寄存器进行写操作以指示完成条目已被处理。

[0042] 在NVMe规范中,提交队列中的读取命令或写入命令可以不按顺序完成。用于将队列和数据传送到控制器并从控制器传送队列和数据的存储器典型地驻留在主机CPU的存储器空间中;然而,NVMe规范允许使用控制器存储缓冲器(CMB)将具有队列和数据块的存储器分配在控制器的存储器空间中。NVMe标准具有供应商特定寄存器和命令空间,所述供应商特定寄存器和命令空间可以用于利用定制的配置和命令来配置NVMe存储设备。

[0043] 传统上,控制硬件加速是利用PCIe规范来执行的。然而,PCIe规范的使用需要依赖于软件(诸如例如,由主机使用的操作系统)以及目标硬件的专用驱动器。相比之下,NVMe规范利用可以与任何软件和硬件平台一起使用的标准驱动器。因此,利用NVMe规范的命令来控制硬件加速可以减少对专用驱动器的需求,并且因此与使用例如PCIe规范控制的传统硬件加速系统相比简化了硬件加速。

[0044] 传统上已经利用硬件加速的一种上下文是在数据存储中,例如在数据中心处。为了防止数据中心中存储的数据丢失,可以存储数据的多于一个副本以便提供冗余。以此方式,如果数据的一个副本因例如存储有数据的存储设备变为损坏而丢失,则此存储设备可以通过将冗余副本之一复制到新的存储设备而重新生成。

[0045] 然而,由于为数据的每个副本提供单独的存储设备的硬件费用可能极高,因此可以利用纠错(EC)过程(类似于通信中使用的纠错)来降低与冗余相关联的成本。EC过程通常基于里德-所罗门(RS)擦除编码块,其中,数据中心的多个存储设备被分配用于存储奇偶性数据,所述奇偶性数据与存储在用于数据存储而分配的其他存储设备处的数据相关联。通过使用奇偶性数据来提供冗余,硬件设备的数量与具有多个存储设备(各自存储数据的冗余副本)相比可能减少。

[0046] 当在数据丢失并且必须在存储设备上恢复时使用的计算资源增加时,硬件费用的减少被抵消。当数据块丢失、或者要重建存储设备时,通过从多个未损坏数据和奇偶性存储设备读取数据来执行缺失数据的重建,这些存储设备用于计算可以写入到替换存储设备的缺失数据块。计算从所存储的数据和奇偶性中缺失的数据块是计算密集型的,并且如果由例如数据中心的主机CPU执行,则可能导致CPU过载。在计算缺失数据块时,诸如在使用EC过程时执行的计算,可以使用硬件加速器来执行计算以减少主机CPU上的计算载荷。

[0047] 图1示出了示例已知数据存储和加速器系统100的示意图,所述系统适合使用用于数据存储的EC过程。数据存储加速器系统100包括主机CPU 102、被分配用于存储数据的数据存储设备106-1、……、106-n、被分配用于存储奇偶性信息的奇偶性存储设备108-1、……、108-m、以及用于执行例如EC过程的PCIe加速器110。主机CPU 102、数据存储设备106-1、……、106-n、奇偶性存储设备108-1、……、108-m、以及PCIe加速器110经由PCIe总线

104连接在一起。

[0048] 所示的示例系统100包括n个数据存储设备106-1至106-n以及被分配用于存储奇偶性信息的m个奇偶性存储设备108-1至108-m,其中,n和m可以是正整数并且可以基于用于生成奇偶性信息而使用的特定EC过程来确定。例如,在RS(12,4)过程的情况下,针对所包括的每十二个数据存储设备106,包括四个奇偶性存储设备108。

[0049] PCIe加速器110包括PCIe接口(未示出)以及可以是例如现场可编程门阵列(FPGA)的一或多个硬件加速器(未示出)。例如,如前文所述,恢复丢失的数据可以通过主机CPU 102通过PCIe总线向PCIe加速器110发送专有命令来发起,所述专有命令由专有加速器接口接收。响应于从主机CPU 102接收到命令,专有加速器接口用信号通知硬件加速器从非损坏数据存储设备106读取数据并从奇偶性存储设备108读取奇偶性信息,并且计算所述数据。如上所述,PCIe加速器存在需要定制驱动器的固有问题,所述定制驱动器需要跨多个OS的支持。

[0050] 本公开的实施例提供一种加速器,所述加速器利用NVMe规范的特征来减少PCIe加速器固有的至少一些上述问题。NVMe加速器可以利用NVMe命令来执行加速过程,而不是如由NVMe规范所预期的磁盘访问功能。以此方式,主机CPU可以以类似于NVMe控制器的方式处理NVMe加速器,以便利用已内置于操作系统中以支持NVMe标准的标准驱动器来执行加速过程。利用已就位的标准驱动器促进加速减少了实施硬件加速所需的软件工程。使用NVMe规范来控制硬件加速超出了NVMe规范的范围和预期,并且因此可能需要对NVMe规范进行一些修改以利用NVMe规范来控制硬件加速,如下文更详细描述的。

[0051] 参考图2,示出了示例加速系统200,其中,主机CPU 202将NVMe命令而非PCIe命令发送到NVMe加速器204。主机CPU 202可以经由PCIe总线203连接到NVMe加速器。

[0052] NVMe加速器204包括一或多个硬件加速器208、210、212,所述硬件加速器中的每一个可以例如被配置用于执行不同的加速功能。图2中示出的示例NVMe加速器204包括三个硬件加速器208、210、212。然而,其他示例NVMe加速器可以包括多于或少于三个硬件加速器,或者单个硬件加速器可以被配置用于执行多个不同的加速过程。图2中示出的示例NVMe加速器204包括NVMe接口206,所述NVMe接口从主机CPU 202接收命令并且基于所述命令用信号通知硬件加速器208、210、212中的一或多个执行适当加速。NVMe接口206包括在NVMe加速器204本身内,并且因此加速器对于主机CPU 202呈现为NVMe存储设备,但是所述加速器可能不具有接口所控制的相关联永久存储装置(诸如SSD)。使用加速器的NVMe接口206既不约束主机CPU 202具有其他NVMe设备(诸如NVMe SSD),也不限制主机CPU 202具有其他NVMe设备。

[0053] 从主机CPU 202发送到NVMe加速器204的命令可以是例如包括在NVMe规范中的标准NVMe磁盘访问命令,但是所述标准NVMe磁盘访问命令是被用作加速命令而不是磁盘访问命令。可替代地,从主机CPU 202发送的命令可以是定制命令,所述定制命令受到NVMe规范内所包括的供应商特定寄存器和命令空间的支持,如下文更详细描述的。

[0054] 现在参考图3,示出了包括NVMe加速器310的示例数据存储和加速系统300。系统300还包括经由PCIe总线304连接的主机CPU 302、n个数据存储设备306-1至306-n、和m个奇偶性存储设备308-1至308-m,这可能实质上类似于上文参考图1所描述的主机CPU 102、数据存储设备106、奇偶性存储设备108、和PCIe总线104,并且因此为避免重复不再做进一步

描述。

[0055] NVMe加速器310可能实质上类似于关于图2所描述的NVMe加速器204，使得主机CPU 302向NVMe加速器310发出NVMe命令以执行加速过程。除了包括NVMe加速器310(而不是如图1的系统100中示出的PCIe加速器)之外，图3中示出的示例系统300包括分别在数据存储设备306-1和NVMe加速器310处的CMB 312和CMB 314。尽管图3中示出的示例包括两个CMB，CMB 312、CMB 314，但是在其他示例中，系统300中可以包括多于或少于两个CMB。CMB 312、CMB 314使得主机CPU 302能够在NVMe设备上而不是在与主机CPU 302相关联的随机存取存储器(诸如例如，双倍数据速率存储器(DDR)303)中建立NVMe队列。在NVMe设备的CMB 312、CMB 314上建立NVMe队列可以用于通过减少与DMA传送相关联的PCIe业务量来减小由系统300的PCIe总线使用的PCIe带宽。

[0056] 尽管系统300包括NVMe加速器310，但是连接到同一PCIe总线304的数据存储设备306和奇偶性存储设备308(在其他示例中为数据存储设备306、奇偶性存储设备308中的一些或全部)可以定位在远端，使得数据通过网络从远程主机进行传送。

[0057] 参考图4，示出了示例加速系统400，其中，可以对例如来自通过网络424可访问的远程数据存储设备(未示出)的数据执行加速。系统400包括具有相关联DDR存储器404的主机CPU 402、和NVMe加速器410。NVMe加速器410经由PCIe开关406连接到主机CPU 402，所述开关经由PCIe总线405连接到主机CPU 402。

[0058] PCIe开关406使得NVMe加速器410与主机CPU 402断开连接并且连接到其他设备。例如，PCIe开关可以用于将NVMe加速器连接到存储设备或其他CPU。进一步地，如下文参考图5详细描述的，PCIe开关406可以用于将NVMe加速器410连接到网络。

[0059] NVMe加速器410包括现场可编程门阵列(FPGA)411以及可选地其上可以设置有控制器CMB 422的板上存储器420。例如，板上存储器420可以是双倍数据速率存储器(DDR)或任何其他合适类型的存储器。如上所述，CMB 422促进主机CPU 402在NVMe加速器410本身上设置NVMe队列，从而减少通过PCIe总线405的业务量。

[0060] FPGA 411包括控制器412(其包括DMA引擎)、NVMe接口414、一个或多个硬件加速器416、以及DDR控制器418。

[0061] 类似于上文关于图2中示出的NVMe加速器204的描述，NVMe加速器410可以通过标准NVMe命令(诸如标准NVMe读取和写入命令)来控制，或者可能通过例如如下文所述的供应商特定命令来控制。控制器412的DMA引擎可以用于传送提交命令和完成命令并且在不使用CMB的情况下将数据传送到硬件加速器416并从所述硬件加速器传送数据。

[0062] 在利用标准NVMe命令的示例中，主机CPU 402可以通过向NVMe加速器410发送标准NVMe磁盘访问命令(诸如磁盘写入命令)来发起加速过程。加速过程的结果可以由主机CPU 402通过向NVMe加速器410发送另一标准NVMe磁盘访问命令(诸如读取命令)来进行检索。在此，标准NVMe磁盘访问命令用于加速控制，而不是用于如由NVMe规范所预期的磁盘访问功能。

[0063] 在NVMe加速器410包括多个硬件加速器416的示例中，每个硬件加速器416可以与对应的NVMe命名空间相关联。例如，NVMe命名空间可以例如为原本以其他方式与SSD相关联的逻辑块地址。在实施例中，关于NVMe命名空间来发送磁盘访问命令，所述命名空间原本以其他方式与SSD相关联、但相反用于实现硬件加速并且在一些情况下实现特定类型的硬件

加速。

[0064] 在示例实施例中,NVMe加速器410被配置用于执行两种不同的加速过程:1)生成固定256位散列(SHA-256)的安全散列算法;以及2)EC。在这个示例中:SHA-256可以与命名空间1相关联;EC编码可以与命名空间2相关联;并且EC解码可以与命名空间3相关联。在这个示例中,主机CPU 402可以通过对命名空间2执行NVMe写入命令来发送待由NVMe加速器410进行EC编码的数据,并且可以通过对命名空间2执行NVMe读取命令来检索所得的经EC编码的数据。

[0065] 在利用供应商特定命令的示例中,主机CPU 402可以将供应商特定命令发送到NVMe加速器410的提交队列。所述提交队列可以驻留在主机CPU 402的DDR 404中或者NVMe加速器410的CMB 422中。供应商特定命令可以由操作码指示,并且促进提交命令向加速器416提供定制的控制和命令信息并促进完成命令向主机CPU 402提供来自加速器416的控制器412的定制反馈信息。在NVMe加速器410包括多个加速器416(每个加速器416被配置用于执行不同的加速过程)的情况下,可以向不同的加速过程分配不同的操作码。

[0066] 在示例实施例中,经由控制器412的DMA引擎,使用提交命令并通过从主机CPU 402发送的供应商特定命令中所提供的存储器地址中进行提取,向加速器416提供数据。加速器416对数据执行由供应商特定命令的操作码指定的加速过程,例如,EC解码加速。在加速器416完成对输入数据的加速过程之后,控制器412将指示加速完成的完成命令提供回到主机CPU 402。如果加速器输出数据相对较小,那么输出数据可以包括在完成命令中。例如,用于SHA-256加密散列函数的输出数据是256位(32字节),这小到足以可以包括在完成命令中。

[0067] 对于生成大量输出数据的加速过程,发起加速过程的供应商特定提交命令可以包括主机CPU 402希望将输出数据被写入到的存储设备的64位地址。在这种情况下,输出数据可以直接写入到64位存储器映射地址。64位存储器地址可以与例如包括主机CPU和NVMe加速器410的计算机的存储器相关联,或者处于另一本地或远程PCIe附接设备(诸如例如,经由PCIe开关406连接到NVMe加速器410的支持CMB的NVMe驱动器)上。在供应商特定提交命令包括64位地址的情况下,完成命令将仅在到所请求位置的数据传送完成之后被发送到主机CPU 402。

[0068] 在示例中,NVMe加速器410可以被配置为使得CMB 422映射到NVMe加速器410的板上存储器420,所述板上存储器典型地为DDR、使用DDR控制器418连接到FPGA 411。在这个示例中,可以由主机CPU 402通过发送标准NVMe命令或供应商特定命令并使用如上所述的DMA控制器412提取输入数据或直接将输入数据写入到CMB 422来提供所述输入数据和加速命令。由硬件加速器416处理输入数据而生成的输出数据可以直接写入到CMB 422,或者可以使用如上所述的完成命令来提供。在完成加速过程后,NVMe加速器410可以向主机CPU 402提供供应商特定完成消息,所述供应商特定完成消息包含指向板上存储器420中的CMB 422中的结果的存储器映射地址,使得主机CPU 402可以检索输出数据。通过在主机CPU 402与NVMe加速器410上的板上存储器420之间提供直接连接,主机CPU 402有能力从板上存储器420检索输出数据并将所述数据传输到任何其他设备,所述设备包括例如经由PCIe开关406连接到NVMe加速器的设备。

[0069] 使用用于数据传送的CMB 422降低了控制器412的DMA引擎上的带宽并且可以避免控制器412中的潜在瓶颈。使用用于数据传送的CMB 422还消除了主机CPU 402提供分级缓

存并在数据源(诸如硬盘驱动器)与加速器416之间执行存储器复制的需要,因为数据源可以直接向加速器416提供数据。使用CMB 422以接收来自一个提交命令的数据并未迫使其他提交命令将CMB 422用于其各自的数据,并且后续的命令可以使用控制器412的DMA引擎以便从主机存储器DDR 404提取数据。DDR控制器418和控制器412的DMA引擎中的瓶颈可以通过使用两种数据传送机制来缓解。

[0070] 如上所讨论的,PCIe开关406可以促进NVMe加速器410通过网络与其他设备(诸如例如,远端位置处的存储设备或CPU)连接。

[0071] 图5示出了系统500的示例,其中,主机CPU 526不具有本地连接的硬件加速器但能够通过网络524访问远程NVMe加速器510以执行加速过程,而无需在远程NVMe加速器510的位置处加载远程CPU 502。

[0072] 在图5中,远程CPU 502、DDR 504、PCIe开关506、NVMe加速器510、FPGA 511、控制器512、NVMe接口514、硬件加速器516、DDR控制器518、具有CMB 522的可选存储器520实质上类似于上文参考图4所描述的主机CPU 402、DDR 404、PCIe开关406、NVMe加速器410、FPGA 411、控制器412、NVMe引擎414、硬件加速器416、DDR控制器418、具有CMB 422的可选存储器420,并且因此为避免重复本文不再做进一步描述。远程CPU 502通过PCIe总线505连接到NVMe加速器。进一步地,PCIe开关506连接到促进将NVMe加速器510连接到网络524的远程直接访问存储器网络接口卡(RDMA NIC)508。

[0073] 主机CPU 526具有相关联DDR 528。主机CPU 526通过PCIe总线529连接到PCIe开关530。PCIe开关530连接到促进通过网络524将主机CPU 526连接到NVMe加速器510的RDMA NIC 532。网络524可以是促进在设备之间传输数据的任何合适的网络,包括有线网络、无线网络、或有线网络与无线网络的组合。

[0074] 在系统500中,主机CPU 526能够与远程NVMe加速器510直接连接以将数据从例如DDR 528直接推送到远程NVMe加速器510,而无需加载远程CPU 502并且无需远程CPU 502一定知晓主机CPU 526与远程NVMe加速器510之间的交易已经发生。类似地,可以在远程CPU 502未干涉或知晓的情况下由主机CPU 526从远程NVMe加速器510提取数据。如上所述,远程CPU 502也可以访问NVMe加速器510的加速功能。因此,图5中示出的系统500可以促进NVMe加速器510的分布式网络,所述分布式网络可以在专用NVMe加速器无保证的情况下在多个CPU之间共享以降低部署成本。

[0075] 实际上,任何数量的主机CPU 526都可以通过网络524与NVMe加速器510连接。另外,NVMe加速器510可以通过网络524连接到任何数量的存储设备。

[0076] 与系统500中的分布式加速器相关联的挑战是:鉴于NVMe加速器510远端的CPU 526将数据推送到NVMe加速器510、而其他CPU并不知晓NVMe加速器载荷的情况下管理加速过程的服务质量。此挑战可以通过实施供应商特定命令来解决,这些命令允许CPU向NVMe加速器510询问加速器载荷数据,诸如例如当前加速载荷和平均加速载荷。此询问可以促进CPU找到具有期望带宽的网络内NVMe加速器510来处理待执行的加速。可替代地,NVMe加速器510的加速载荷统计可以驻留在CMB 522中,从而允许CPU 502、CPU 526直接从NVMe加速器510的存储器520读取载荷。

[0077] 现在参考图6,示出了展示一种用于使用NVMe规范来控制加速器的方法的流程图。所述方法可以在上述示例NVMe加速器中的任一个中实施。所述方法可以由例如NVMe加速器

的处理器执行，所述处理器执行NVMe加速器的存储器中所存储的指令。

[0078] 在602处，在NVMe加速器的NVMe接口处从主机CPU接收与加速器过程相关联的第一NVMe命令。如上所公开的，第一NVMe命令的格式可以是标准NVMe命令的格式(诸如根据NVMe规范的标准磁盘访问命令(诸如例如，读取命令或写入命令))，或者可以是供应商特定命令。例如，第一NVMe命令可以是标准NVMe读取命令/写入命令，所述读取命令/写入命令可以包括原本与SSD相关联的命名空间，其中，所包括的命名空间相反与加速过程相关联。供应商特定命令可以包括由加速过程生成的结果数据将被写入到的地址。进一步地，第一NVMe命令可以从本地的主机CPU或从远端的(使得第一NVMe命令通过网络接收)主机CPU接收。

[0079] 在604处，确定与接收到的第一NVMe命令相关联的加速功能。例如，如上所述，如果第一NVMe命令是标准NVMe命令的格式，那么604处的确定可以包括确定与命名空间相关联的加速功能，所述命名空间原本以其他方式与SSD相关联、但现在与加速功能相关联，即包括在第一NVMe命令内。604处的确定也可以包括确定多个硬件加速器之一，这些硬件加速器被配置用于执行与第一NVMe命令相关联的加速过程。

[0080] 在606处，由硬件加速器执行加速过程。在606处执行加速过程可以包括将待处理的输入数据发送到硬件加速器、或者用信号通知硬件加速器检索输入数据。在606处执行加速过程也可以包括用信号通知加速硬件将所生成的结果数据写入到具体地址。

[0081] 可选地，在608处，在硬件加速器已完成执行加速过程时向主机CPU发送完成消息。所述完成消息可以是标准NVMe完成消息，或者可以是供应商特定完成消息。例如，如果结果数据小到足以包括在完成消息中，那么供应商特定完成消息可以包括结果数据。如果结果数据由硬件加速器写入到第一NVMe命令中由主机CPU指定的具体存储器地址，那么一旦已将结果数据完全写入到指定地址就可以发送完成消息。供应商特定NVMe完成消息可以包括结果数据已写入的地址。

[0082] 可选地，在610处，可以从主机CPU接收第二NVMe命令以检索结果数据，并且响应于接收到所述第二NVMe命令，可以发送结果数据。例如，第二NVMe命令可以是标准NVMe磁盘访问命令(诸如根据NVMe规范的标准读取命令或写入命令)，或者可以是供应商特定命令。标准读取命令/写入命令可以包括命名空间，其中，所包括的命名空间与加速过程相关联，使得来自与命名空间相关联的加速过程的结果数据是发送到主机CPU的数据。供应商特定命令可以包括结果数据要发送到的地址。

[0083] 本公开的实施例通过利用NVMe命令来控制硬件加速器从而促进执行硬件加速过程而无需使用软件和硬件特定的专用驱动器。所述NVMe命令可以是基于NVMe规范中提供的标准化NVMe命令，或者可以是由所述NVMe规范支持的供应商特定命令。所述命令通过主机CPU被发送到所述NVMe加速器，在一些实施例中，所述主机CPU可以定位在所述NVMe加速器远端。所述NVMe加速器可以包括CMB，主机CPU可以在其上设置NVMe队列，以减少将所述CPU与所述NVMe加速器连接的PCIe总线上的PCIe业务量。主机CPU还可以使用所述CMB来传送用于加速算法的数据，以清除主机分级缓存、减小DMA控制器的带宽或清除主机存储器副本。

[0084] 在前面的描述中，为了进行解释，阐述了很多细节以提供对实施例的透彻理解。然而，对于本领域技术人员而言将显而易见的是，这些具体细节不是必需的。在其他实例中，为了不模糊理解，采用框图的形式示出了众所周知的电气结构和电路。例如，没有提供关于本文描述的实施例是被实施为软件例程、硬件电路、固件还是其组合的具体细节。

[0085] 本公开的实施例可以被表示为存储在机器可读介质(也被称为计算机可读介质、处理器可读介质或具有在其中具体化的计算机可读程序代码的计算机可用介质)中的计算机程序产品。机器可读介质可以是任何适当的有形的非暂态介质(包括磁性的、光学的或电气的存储介质(包括磁盘、光盘只读存储器(CD-ROM)、存储器设备(易失的或非易失的)或类似的存储机制))。机器可读介质可以包含各种指令集、代码序列、配置信息或其他数据,其在执行时使处理器执行根据本公开实施例的方法中的步骤。本领域的普通技术人员将理解,实施所描述的实施方式所必需的其他指令和操作也可以存储在机器可读介质上。在机器可读介质上存储的指令可以由处理器或其他适当的处理设备来执行,并且可以与电路系统接口连接以执行所描述的任务。

[0086] 上述实施例仅旨在作为示例。在不脱离仅由所附权利要求书界定的范围的情况下,本领域的普通技术人员可以实现对具体实施例的更改、修改和变型。

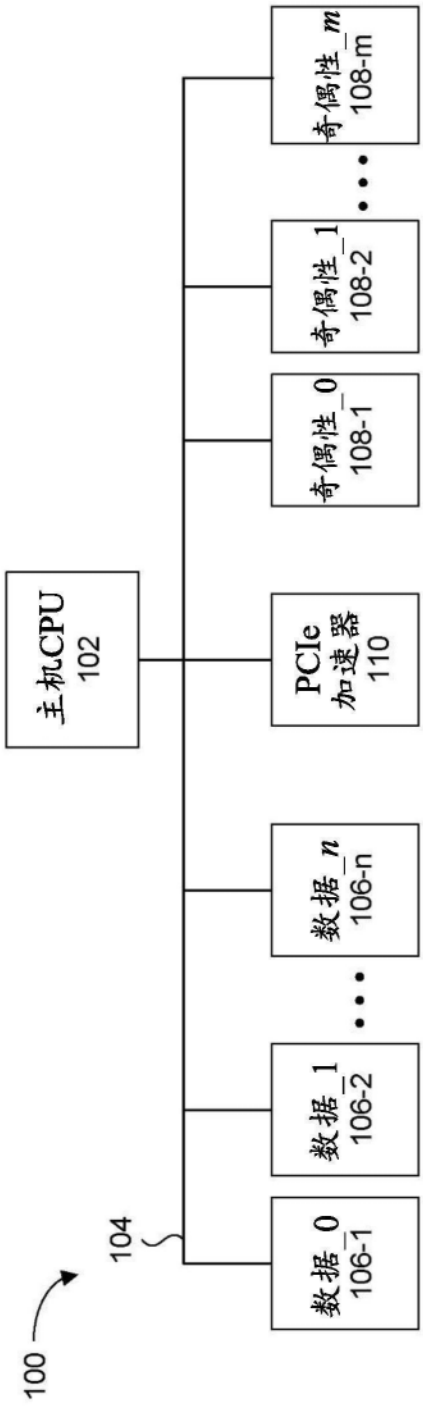

主机CPU
102
PCIe
加速器
110
奇偶性_0
108-1
奇偶性_1
108-2
奇偶性_m
108-m
数据_0
106-1
数据_1
106-2
数据_n
106-n
104
100

图1现有技术

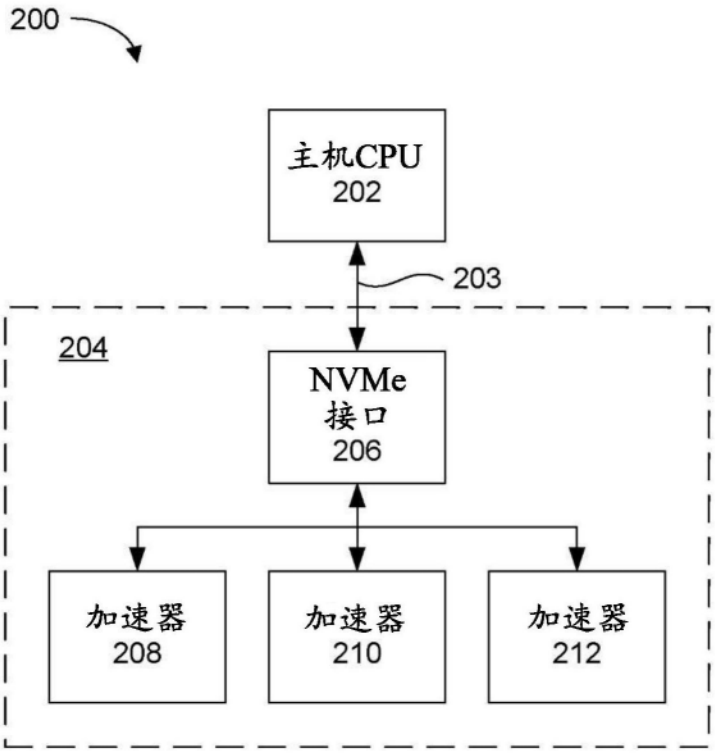


图2

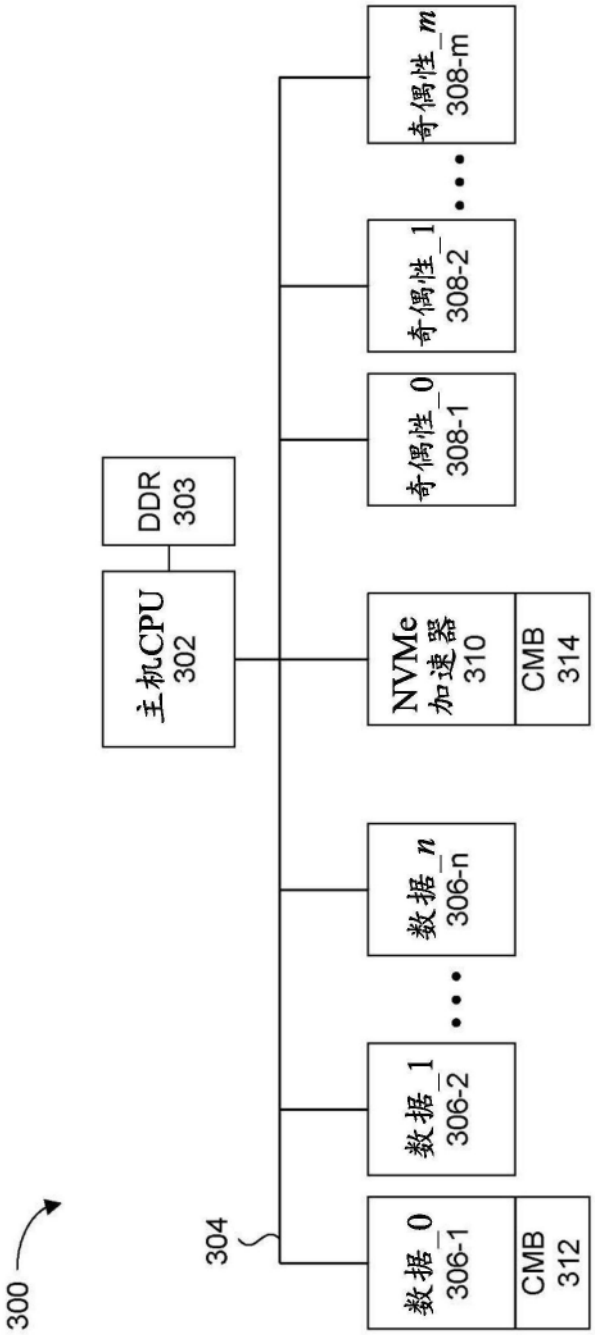

300
304
数据_0
306-1
CMB
312
数据_1
306-2
数据_n
306-n
NVMe
加速器
310
CMB
314
主机CPU
302
DDR
303
奇偶性_0
308-1
奇偶性_1
308-2
奇偶性_m
308-m

图3

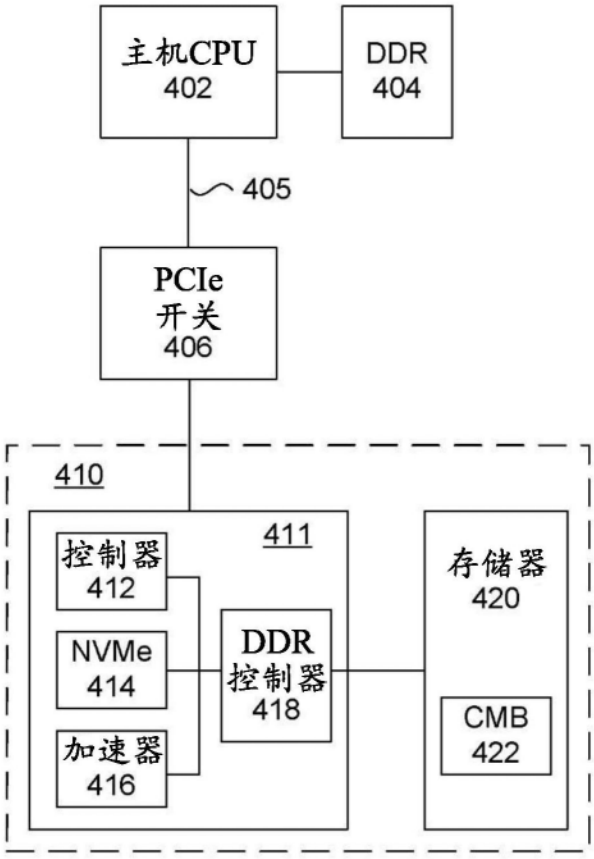

主机CPU
402
DDR
404
405
PCIe
开关
406
410
411
控制器
412
NVMe
414
加速器
416
DDR
控制器
418
存储器
420
CMB
422

图4

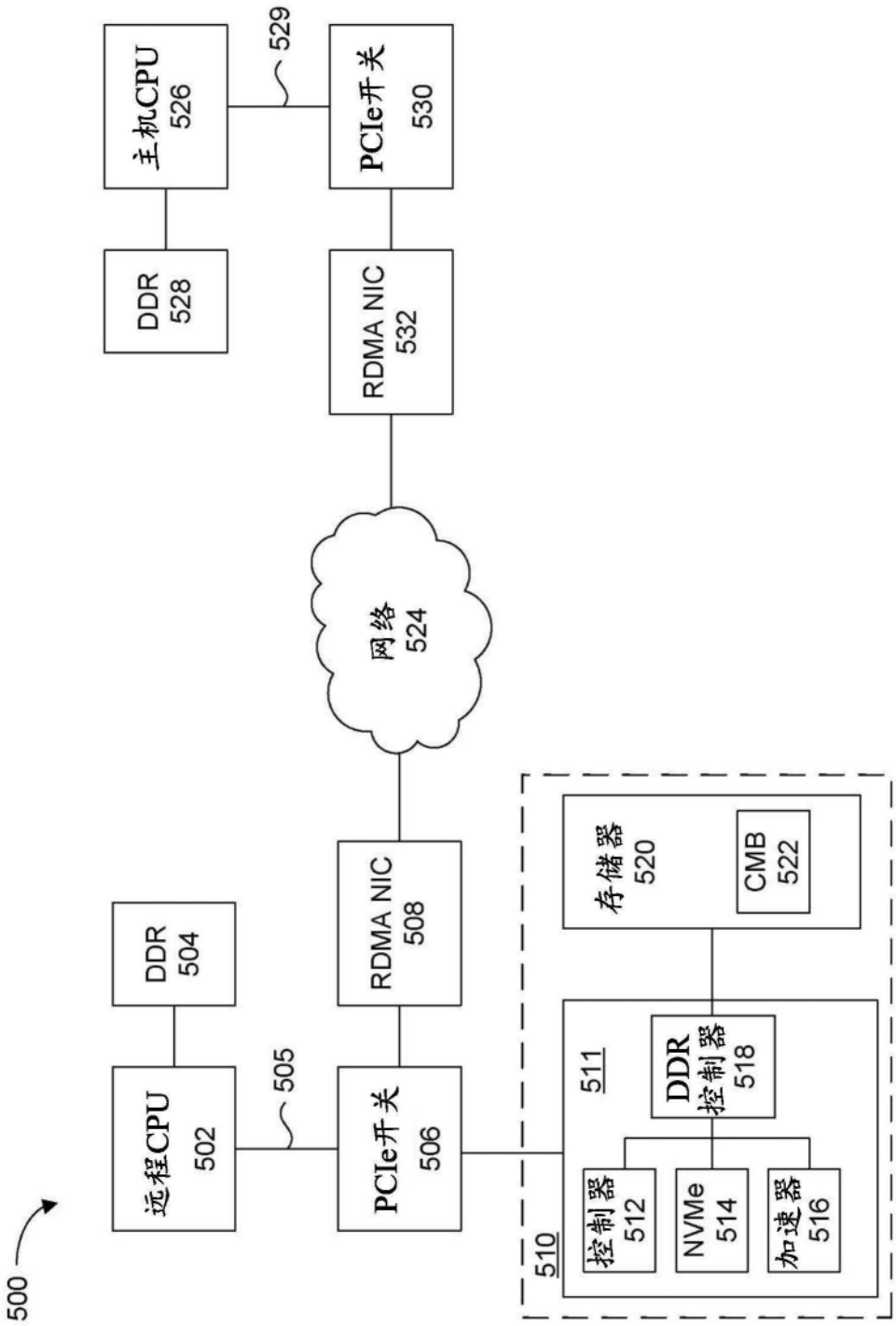

500
远程CPU
502
DDR
504
505
PCIe开关
506
RDMA NIC
508
网络
524
RDMA NIC
532
PCIe开关
530
529
主机CPU
526
DDR
528
510
511
控制器
512
NVMe
514
加速器
516
DDR
控制器
518
存储器
520
CMB
522

图5

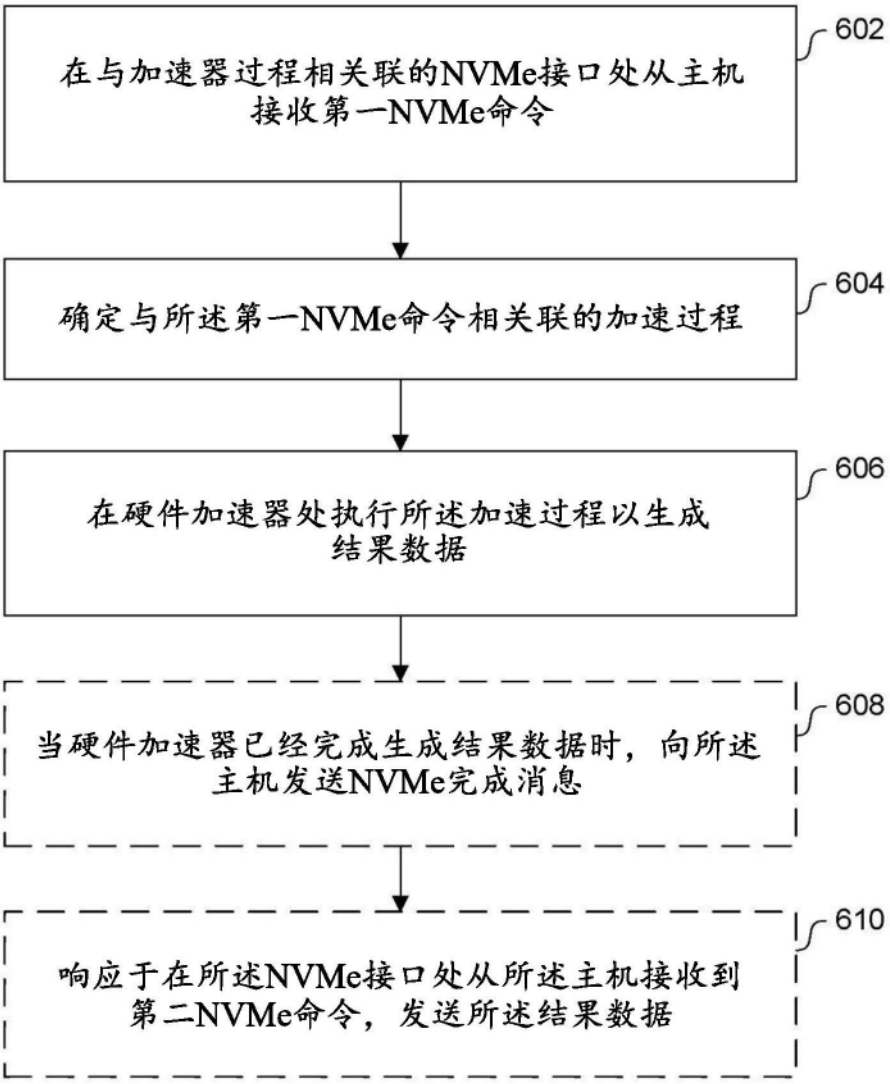

602
在与加速器过程相关联的NVMe接口处从主机接收第一NVMe命令
604
确定与所述第一NVMe命令相关联的加速过程
606
在硬件加速器处执行所述加速过程以生成结果数据
608
当硬件加速器已经完成生成结果数据时，向所述主机发送NVMe完成消息
610
响应于在所述NVMe接口处从所述主机接收到第二NVMe命令，发送所述结果数据

图6