(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110088724 B
(45)授权公告日 2022.08.26

(21)申请号 201880005146.1

(22)申请日 2018.02.27

(65)同一申请的已公布的文献号
申请公布号 CN 110088724 A

(43)申请公布日 2019.08.02

(30)优先权数据
15/468,620 2017.03.24 US
15/602,874 2017.05.23 US

(85)PCT国际申请进入国家阶段日
2019.06.18

(86)PCT国际申请的申请数据
PCT/US2018/019909 2018.02.27

(87)PCT国际申请的公布数据
WO2018/175060 EN 2018.09.27

(73)专利权人 西部数据技术公司
地址 美国加利福尼亚州

(72)发明人 S·贝尼斯蒂

(74)专利代理机构 北京纪凯知识产权代理有限公司 11245
专利代理师 魏利娜

(51)Int.Cl.
G06F 3/06(2006.01)
G06F 13/16(2006.01)

(56)对比文件
JP 2007183959 A,2007.07.19
CN 106528461 A,2017.03.22
CN 104821887 A,2015.08.05
US 2008282031 A1,2008.11.13
US 2016124876 A1,2016.05.05
US 2015019798 A1,2015.01.15
US 2003204552 A1,2003.10.30
CN 106527967 A,2017.03.22
CN 102467968 A,2012.05.23
US 2017010992 A1,2017.01.12
曹文斌 等.应用多GPU的可压缩湍流并行计算.《国防科技大学学报》.2015,第37卷(第3期),第78-83页.
W. Choi et al.An in-depth study of next generation interface for emerging non-volatile memories.《2016 5th Non-Volatile Memory Systems and Applications Symposium (NVMSA)》.2016,第1-6页.

审查员 黄烨腾

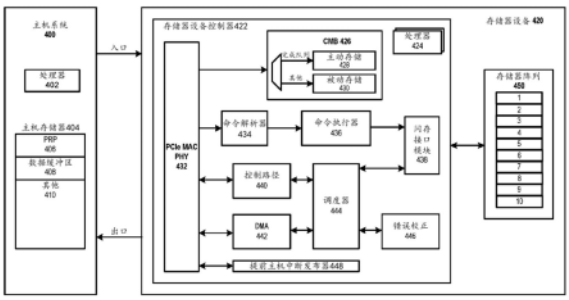
权利要求书2页 说明书17页 附图11页

(54)发明名称

使用控制器存储器缓冲区进行自适应提前完成发布的系统和方法

(57)摘要

本发明公开了用于动态和自适应中断合并的系统和方法。NVM Express(NVMe)实现了成对提交队列和完成队列机制,主机设备上的主机软件将命令放置到所述提交队列中。存储器设备通过中断向所述主机设备通知所述完成队列上的条目。响应于接收到所述中断,所述主机设备访问所述完成队列以访问由所述存储器设备放置在所述完成队列中的条目。所述主机设备可能花费一定量的时间来服务所述中断,从而导致主机时延。给定所述主机时延的了解,所述存储器设备对所述中断的发送计时,使得给定所述主机时延,所述存储器设备可以以及时的方式将所述条目发布到所述完成队列。



CN 110088724 B

1.一种用于存储器设备的方法,包括:

在存储器设备中:

从主机设备接收读取队列的请求,所述队列指示所述存储器设备对命令的执行;

响应于所述请求,确定在接收所述请求之后的预定时间段内,所述存储器设备是否将用主机命令的执行状态来更新所述队列;以及

响应于确定所述存储器设备将在所述预定时间段内更新所述队列,延迟对所述请求的响应,直到所述存储器设备用所述主机命令的所述执行状态更新所述队列之后。

2.根据权利要求1所述的方法,其中,所述队列包括指示所述存储器设备完成主机命令的完成队列。

3.根据权利要求2所述的方法,其中,用所述主机命令的所述执行状态更新所述队列包括更新所述完成队列以指示所述主机命令的执行已经完成。

4.根据权利要求3所述的方法,其中,所述存储器设备包括配置为存储所述完成队列的缓冲区。

5.根据权利要求1所述的方法,其中,所述预定时间段包括预定时间。

6.根据权利要求1所述的方法,其中,所述预定时间段包括预定数量的处理周期。

7.一种非易失性存储器设备,包括:

非易失性存储器;

通信接口,所述通信接口被配置为与主机设备通信;和

控制器,所述控制器与所述非易失性存储器和所述通信接口通信,所述控制器包括控制器存储器缓冲区,所述控制器存储器缓冲区被配置为存储完成队列并且被配置为:

完成第一命令的执行;将第一条目发布到所述完成队列,

所述第一条目指示所述第一命令的执行的完成;从所述主机设备接收读取驻留在所述控制器存储器缓冲区中的所述完成队列的请求,所述完成队列指示由所述主机设备放置在对应的提交队列上的命令的执行的完成;

响应于接收到所述请求,基于所述完成队列中的预期未来活动来确定是否延迟响应所述请求;

其中,所述预期未来活动包括在接收所述请求之后的预定时间段内用第二条目更新所述完成队列,所述第二条目指示第二命令的完成;并且

响应于确定延迟响应,延迟响应所述请求,直到所述控制器执行所述完成队列中的所述预期的未来活动之后。

8.根据权利要求7所述的存储器设备,其中,所述控制器被进一步配置为:

向所述主机设备发送中断,其中,接收所述中断导致所述主机设备访问发布到所述完成队列的所述第一条目。

9.根据权利要求7所述的存储器设备,其中,所述控制器被进一步配置为向所述主机设备发送中断,所述中断指示所述主机设备发送读取所述完成队列的所述请求。

10.根据权利要求9所述的存储器设备,其中:

所述控制器存储器缓冲区被进一步配置为存储顺序队列,所述主机设备被配置为将一个或多个命令存储在所述顺序队列上以供所述存储器设备执行;并且

所述控制器被进一步配置为:

响应于所述主机设备导致所述一个或多个命令被存储在所述顺序队列上，提取存储在所述顺序队列上的所述一个或多个命令；

开始执行所述一个或多个提取的命令；并且

在完成所述一个或多个提取的命令的执行之前，将所述中断发送到所述主机设备。

11.根据权利要求10所述的存储器设备，其中，所述控制器被配置为基于所述控制器确定所述一个或多个提取的命令已经开始执行但尚未完成执行来确定是否延迟响应所述请求。

12.一种非易失性存储器设备，包括：

非易失性存储器；

用于从主机设备接收读取完成队列的请求的装置，所述完成队列指示命令的执行的完成，所述命令中的至少一个命令用于从所述非易失性存储器读取数据或将数据写入所述非易失性存储器；

用于响应于所述请求而确定所述存储器设备是否将在接收所述请求之后的预定时间段内用主机命令的执行状态来更新所述完成队列的装置；以及

响应于确定所述存储器设备将在所述预定时间段内更新所述完成队列，用于延迟对来自所述主机设备的所述请求的响应的装置，直到所述存储器设备用所述主机命令的所述执行状态来更新所述完成队列之后。

13.根据权利要求12所述的非易失性存储器设备，其中，所述预定时间段包括预定时间或预定数量的处理周期。

14.根据权利要求12所述的非易失性存储器设备，还包括用于存储所述完成队列的装置。

使用控制器存储器缓冲区进行自适应提前完成发布的系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求2017年3月24日提交的美国申请No.15/468,620和2017年5月23日提交的美国申请No.15/602,874的优先权,这两个申请据此全文以引用方式并入本文。

技术领域

[0003] 本申请涉及存储器系统。

背景技术

[0004] NVM Express(NVMe)是访问经由PCI Express(PCIe)总线附接的非易失性存储介质的标准。NVMe可与多种非易失性存储介质诸如固态驱动器(SSD)一起使用。NVMe的一个焦点涉及主机设备(其可以访问和/或写入非易失性存储介质)和存储器设备(其包括非易失性存储介质)之间的I/O通信。在这点上,NVMe实现了成对提交队列和完成队列机制,主机设备上的主机软件将命令放置到提交队列中。完成由存储器设备控制器放置到相关联的完成队列中。

发明内容

[0005] 根据本发明的一方面,提供了一种用于存储器设备的方法,包括:在存储器设备中:从主机设备接收读取队列的请求,所述队列指示所述存储器设备对命令的执行;响应于所述请求,确定在接收所述请求之后的预定时间段内,所述存储器设备是否将用主机命令的执行状态来更新所述队列;以及响应于确定所述存储器设备将在所述预定时间段内更新所述队列,延迟对所述请求的响应,直到所述存储器设备用所述主机命令的所述执行状态更新所述队列之后。

[0006] 根据本发明的另一方面,提供了一种非易失性存储器设备,包括:非易失性存储器;通信接口,所述通信接口被配置为与主机设备通信;和控制器,所述控制器与所述非易失性存储器和所述通信接口通信,所述控制器包括控制器存储器缓冲区,所述控制器存储器缓冲区被配置为存储完成队列并且被配置为:完成第一命令的执行;将第一条目发布到所述完成队列,所述第一条目指示所述第一命令的执行的完成;从所述主机设备接收读取驻留在所述控制器存储器缓冲区中的所述完成队列的请求,所述完成队列指示由所述主机设备放置在对应的提交队列上的命令的执行的完成;响应于接收到所述请求,基于所述完成队列中的预期未来活动来确定是否延迟响应所述请求;其中,所述预期未来活动包括在接收所述请求之后的预定时间段内用第二条目更新所述完成队列,所述第二条目指示第二命令的完成;并且响应于确定延迟响应,延迟响应所述请求,直到所述控制器执行所述完成队列中的所述预期的未来活动之后。

[0007] 根据本发明的另一方面,提供了一种非易失性存储器设备,包括:非易失性存储器;用于从主机设备接收读取完成队列的请求的装置,所述完成队列指示命令的执行的完

成,所述命令中的至少一个命令用于从所述非易失性存储器读取数据或将数据写入所述非易失性存储器;用于响应于所述请求而确定所述存储器设备是否将在接收所述请求之后的预定时间段内用主机命令的执行状态来更新所述完成队列的装置;以及响应于确定所述存储器设备将在所述预定时间段内更新所述完成队列,用于延迟对来自所述主机设备的所述请求的响应的装置,直到所述存储器设备用所述主机命令的所述执行状态来更新所述完成队列之后。

附图说明

[0008] 并入本说明书中并且构成本说明书的一部分的附图示出了本发明的各个方面,并与说明书一起用于解释其原理。在方便的情况下,相同的参考号将在整个附图中用来指代相同或相似的元件。

[0009] 图1A是示例性非易失性存储器系统的框图。

[0010] 图1B是包括多个非易失性存储器系统和主机的存储模块的框图。

[0011] 图1C是分级存储系统的框图。

[0012] 图2A是图1A的非易失性存储器系统的控制器的示例性部件的框图。

[0013] 图2B是图1A的非易失性存储器系统的非易失性存储器管芯的示例性部件的框图。

[0014] 图3是主机设备和NVMe控制器的框图,示出了主机设备和存储器设备请求和处理NVMe命令的序列。

[0015] 图4是主机系统和存储器设备的其他示例性部件的框图。

[0016] 图5是确定是否延迟响应来自主机设备的读取控制器存储器缓冲区中的完成队列的请求的第一示例方法的流程图。

[0017] 图6是确定是否延迟响应来自主机设备的读取控制器存储器缓冲区中的完成队列的请求的第二示例方法的流程图。

[0018] 图7是确定在发布到完成队列之前是否向主机设备发送中断的示例方法的流程图。

[0019] 图8是示出现有技术和主机设备访问存储在控制器存储器缓冲区中的完成队列的一种实施方式之间的差异的时序图。

[0020] 图9是示出现有技术发布到完成队列和存储在控制器存储器缓冲区中的完成队列的提前中断发布的一种实施方式之间的差异的时序图。

[0021] 图10是示出现有技术发布到完成队列和存储在主机设备中的完成队列的提前中断发布的一种实施方式之间的差异的时序图。

具体实施方式

[0022] 概述

[0023] NVM Express基于成对提交队列和完成队列机制。如下面参考图3更详细讨论的,NVMe标准包括用于处理命令的步骤序列。例如,该序列可以如下:主机设备将主机命令发布到提交队列;主机设备将主机命令被发布到提交队列的通知发送到存储器设备;存储器设备从提交队列提取主机命令;存储器设备执行主机命令;存储器设备将指示主机命令已经完成执行的条目(例如,完成队列上的条目指示命令的执行状态是完成)发布到完成队列;

存储器设备将条目被发布到完成队列的通知发送到主机设备;主机设备从完成队列中检索条目;以及主机设备发送已经从完成队列中检索到条目的通知。

[0024] 在这点上,命令由主机设备的主机软件放置到提交队列中。完成由存储器设备的控制器放置到相关联的完成队列中。因此,完成队列是指示命令的执行的队列。在一个实施方式中,提交队列和完成队列被分配在主机设备上的主机存储器中。特别地,每个提交队列和完成队列可以物理上相邻地位于主机存储器中,或者不相邻地位于主机存储器中。另选地,主机设备可以将提交队列和完成队列放置在存储器设备的控制器存储器中的控制器存储器缓冲区(CMB)中。

[0025] 在主机设备将命令被发布到提交队列的通知发送到存储器设备(例如,主机设备向提交队列发出门铃写入)的时间和主机设备发送条目已经从完成队列中检索到的通知(例如,主机设备向对应的完成队列门铃写入)的时间之间,NVMe命令执行性能被测量。

[0026] 因此,在处理命令时,主机设备可以请求动作,诸如读取完成队列上的条目。如上文所讨论的,在一个实施方式中,完成队列驻留在存储器设备上的CMB中。在该实施方式中,主机设备可以通过使用用于从存储器设备中的CMB读取的传输分组层(TLP)读取请求来请求读取CMB中的完成队列。响应于接收到TLP请求,存储器设备执行动作,诸如读取CMB中的完成队列,并发送该动作的结果,诸如从CMB中的完成队列读取的条目。通常,存储器设备的控制器将调度以执行动作,诸如读取CMB,以及调度各种其他任务来执行。在这点上,何时执行动作的调度基于响应于请求执行该动作的优先级和各种其他任务的优先级。因此,存储器设备以被动方式响应于TLP请求,简单地基于存储器设备的内部资源响应于TLP请求。

[0027] 在一个实施方式中,存储器设备的控制器可以以主动方式响应于TLP请求,由此存储器设备审查请求的内容,即请求试图从CMB中的完成队列读取的内容,并且基于可能影响完成队列的其他动作或预期的未来活动来确定是否延迟响应请求(例如,存储器设备确定CMB上的完成队列将在一定数量的硬件周期中发布另一条目)。特别地,控制器可以监控主机设备和存储器设备之间的接口,诸如PCIe接口,以便识别对CMB上的完成队列的TLP读取请求。以这种方式,存储器设备的控制器可以基于影响完成队列的其他动作主动监控是否延迟响应来自主机设备的请求。

[0028] 作为一个示例,存储器设备可能先前已经将第一条目发布到完成队列,其中第一条目指示第一命令的执行的完成。此外,响应于完成第一命令的执行,存储器设备可能已经向主机设备发送中断,使得主机设备访问发布到完成队列的第一条目。响应于接收到中断,主机设备发送访问完成队列的请求。响应于接收到该请求,存储器设备可以确定是否将完成另一命令(诸如第二命令),从而导致在第二条目的预定时间段内更新完成队列,其中第二条目指示第一命令的执行的完成。在这点上,存储器设备可以延迟访问完成队列的请求的响应。作为另一示例,存储器设备可以在其中存储顺序队列,其中主机设备导致命令被放置以供存储器设备执行。存储器设备可以提取命令,并开始执行提取的命令。在完成提取的命令的执行之前,存储器设备可以发布中断,如下文进一步讨论的。响应于发布的中断,主机设备发送访问完成队列的请求。在存储器设备还没有将指示提取的命令已经完成执行的条目发布到完成队列的情况下,存储器设备可以延迟对主机设备请求的响应。

[0029] 另选地或除此之外,存储器设备的控制器可以确定向主机设备发送条目被发布到完成队列的通知,甚至在条目已经被发布之前。如下文更详细讨论的,在主机设备响应于由

存储器设备发布的中断的过程中存在时延。特别地，存储器设备可以确定主机设备为了响应于中断通常花费的时间量，诸如硬件处理周期的数量或以微秒为单位的时间。有了对主机时延的了解以及对存储器设备何时将条目发布到完成队列的了解，存储器设备可以将中断提前传输到主机设备。

[0030] 如上文所讨论的，完成队列可以驻留在存储器设备上的CMB中或主机设备中的主机存储器中。在完成队列驻留在CMB中的第一具体实施方式中，存储器设备可以将中断提前发布到主机设备。存储器设备可以基于主机时延(例如，从存储器设备发布中断的时间到主机设备请求从完成队列读取的时间的时间段)和基于存储器设备对于存储器设备将条目发布到完成队列的时间的估计来对中断的发送计时。例如，存储器设备可以确定存储器设备将完成命令处理的第一时间段，诸如处理周期的数量或以微秒为单位的时间，以及用于主机时延的第二时间段，其可以由处理周期的数量或以微秒为单位的时间表示。在一个实施方式中，当第一时间段等于第二时间段时，存储器设备可以对中断的发送进行计时。换句话说，当完成命令的执行的时间段等于主机设备响应于中断的时间段时，存储器设备可以对中断的发送进行计时。在一个具体实施方式中，主机设备响应于中断的时间段包括预定数量的处理周期。在该具体实施方式中，当完成命令的执行的时间段等于预定数量的处理周期时，存储器设备可以接着发送中断。在主机设备在存储器设备将条目发布到完成队列之前请求从完成队列读取的情况下，存储器设备可以延迟响应，直到条目被发布到完成队列之后。在完成队列驻留在主机存储器中的第二具体实施方式中，存储器设备同样可以将中断提前发布到主机设备。类似于第一具体实施方式，存储器设备可以基于主机时延和基于存储器设备对于存储器设备将条目发布到完成队列的时间的估计来对中断的发送进行计时。然而，因为完成队列驻留在主机设备中，所以存储器设备不能延迟主机设备从完成队列读取条目的请求。不过，在任一实施方式中，存储器设备可以减少NVMe命令的寿命。

[0031] 实施方案

[0032] 以下实施方案描述了用于处理命令的非易失性存储器设备和相关方法。在转向这些和其他实施方案之前，以下段落提供了可与这些实施方案一起使用的示例性非易失性存储器设备和存储模块的讨论。当然，这些仅仅是示例，并且可以使用其他合适类型的非易失性存储器设备和/或存储模块。

[0033] 图1A是示出非易失性存储器设备100的框图。非易失性存储器设备100可以包括控制器102和可以由一个或多个非易失性存储器管芯104构成的非易失性存储器。如本文所述，术语管芯指的是在单个半导体基板上形成的一组非易失性存储器单元，以及用于管理那些非易失性存储器单元的物理操作的相关联的电路。控制器102可以与主机设备或主机系统进行交互，并且将用于读取、编程和擦除操作的命令序列传输到非易失性存储器管芯104。如下文所讨论的，命令可以包括逻辑地址。

[0034] 控制器102(可以是闪存存储器控制器)可以采用以下形式：例如处理电路、微处理器或处理器，以及存储可由(微)处理器执行的计算机可读程序代码的计算机可读介质(例如，软件或固件)、逻辑门、开关、专用集成电路(ASIC)、可编程逻辑控制器和嵌入式微控制器。控制器102可以配置有硬件和/或固件，以执行下面描述并在流程图中示出的各种功能。另外，示出为控制器内部的一些部件也可以存储在控制器外部，并且可以使用其他部件。此外，短语“操作地与…通信”可能意味着直接或间接地(有线或无线)与一个或多个部件通信

或通过一个或多个部件通信,其可在本文中示出或未示出。

[0035] 如本文所用,闪存存储器控制器是管理存储在闪存存储器上的数据并与主机诸如计算机或电子设备通信的设备。除了这里描述的特定功能外,闪存存储器控制器可以具有各种功能。例如,闪存存储器控制器可以对闪存存储器进行格式化以确保存储器正确操作,标出坏的闪存存储器单元,并且分配备用单元以替代将来的故障单元。备用单元中的部分备用单元可以用来容纳固件以操作闪存存储器控制器并实现其他特征。固件的一个示例是闪存转换层。在操作中,当主机设备需要从闪存存储器读取数据或向闪存存储器写入数据时,它将与闪存存储器控制器通信。在一个实施方案中,如果主机设备提供要读取/写入数据的逻辑地址,则闪存存储器控制器可以将从主机接收的逻辑地址转换为闪存存储器中的物理地址。闪存存储器控制器还可以执行各种存储器管理功能,诸如但不限于损耗均衡(分配写入以避免损耗否则将被重复写入的特定存储器块)和垃圾收集(在块已满之后,仅将有效的数据页面移动到新块,因此可以擦除并重用完整块)。

[0036] 控制器102和非易失性存储器管芯104之间的接口可以是任何合适的闪存接口,诸如切换模式200、400或800。在一个实施方案中,存储器设备100可以是基于卡的系统,诸如安全数字(SD)卡或微型安全数字(微SD)卡。在另选的实施方案中,非易失性存储器设备100可以是嵌入式存储器设备的一部分。

[0037] 虽然在图1A所示的示例中,非易失性存储器设备100可以包括控制器102和非易失性存储器管芯104之间的单个信道,但是本文描述的主题不限于具有单个存储器信道。例如,在一些NAND存储器设备架构中,控制器和NAND存储器管芯104之间存在2、4、8个或更多个NAND信道,具体取决于控制器的能力。在本文描述的任何实施方案中,即使在附图中示出单个信道,控制器和存储器管芯104之间也可以存在多于一个信道。

[0038] 图1B示出了包括多个非易失性存储器设备100的存储模块200。因此,存储模块200可以包括存储控制器202,该存储控制器与主机220以及包括多个非易失性存储器设备100的存储系统204交互。存储控制器202和非易失性存储器设备100之间的接口可以是总线接口,诸如作为示例的串行高级技术附件(SATA)、外围组件快速互连(PCIe)接口、嵌入式多媒体卡(eMMC)接口、SD接口或通用串行总线(USB)接口。在一个实施方案中,存储系统200可以是固态驱动器(SSD),诸如在诸如膝上型计算机和平板电脑的便携式计算设备和移动电话中存在的。

[0039] 图1C是示出分级存储系统250的框图。分级存储系统250可以包括多个存储控制器202,该多个存储控制器中的每个存储控制器控制相应的存储系统204。主机系统252可以经由总线接口访问分级存储系统250内的存储器。示例总线接口可以包括作为示例的非易失性存储器express(NVMe)、以太网光纤信道(FCoE)接口、SD接口、USB接口、SATA接口、PCIe接口或eMMC接口。在一个实施方案中,图1C所示的分层存储系统250可以是机架可安装的大容量存储系统,该机架可安装的大容量存储系统可由多个主机计算机访问,诸如在数据中心中或在需要大容量存储的其他位置中可以找到的。在一个实施方案中,主机系统252可以包括主机220中描述的功能。

[0040] 图2A是更详细地示出控制器102的示例性部件的框图。控制器102包括与主机交互的前端模块108、与非易失性存储器管芯104交互的后端模块110、以及执行非易失性存储器设备100的各种功能的各种其他模块。通常,模块可以是硬件或硬件和软件的组合。例如,每

个模块可以包括专用集成电路(ASIC),现场可编程门阵列(FPGA),电路,数字逻辑电路,模拟电路,离散电路、门或任何其他类型的硬件的组合,或者其组合。除此之外或另选地,每个模块可以包括存储器硬件,该存储器硬件包括可由处理器或处理器电路执行的指令,以实现模块的一个或多个特征。当任何一个模块包括存储器中包括可由处理器执行的指令的部分时,该模块可以包括或不包括处理器。在一些示例中,每个模块可以只是存储器的一部分,该部分包括可由处理器执行的指令,以实现对应模块的特征,而该模块不包括任何其他硬件。因为每个模块包括至少一些硬件,即使当所包括的硬件包括软件时,每个模块也可以可互换地称为硬件模块。

[0041] 控制器102可以包括缓冲区管理器/总线控制模块114,其管理随机存取存储器(RAM)116中的缓冲区,并控制用于在控制器102的内部通信总线117上通信的内部总线仲裁。只读存储器(ROM)118可以存储和/或访问系统引导代码。虽然图2A中示出为与控制器102分开定位,但在其他实施方案中,RAM 116和ROM 118中的一者或两者可以位于控制器102内。在其他实施方案中,RAM 116和ROM 118的部分可以位于控制器102内和控制器102外部。此外,在一些实施方式中,控制器102、RAM 116和ROM 118可以位于单独的半导体管芯上。如下文所讨论的,在一个实施方式中,提交队列和完成队列可以存储在控制器存储器缓冲区中,控制器存储器缓冲区可以容纳在RAM 116中。

[0042] 另外,前端模块108可以包括提供与主机或下一级存储控制器的电接口的主机接口120和物理层接口(PHY)122。主机接口120类型的选择可以取决于所使用的存储器的类型。主机接口120的示例类型可以包括但不限于SATA、SATA Express、SAS、光纤信道、USB、PCIe和NVMe。主机接口120通常可以有利于传输数据、控制信号和定时信号。

[0043] 后端模块110可以包括错误校正控制器(ECC)引擎124,该ECC引擎对从主机接收的数据字节进行编码,并且对从非易失性存储器管芯104读取的数据字节进行解码和错误校正。如下文更详细讨论的,ECC引擎可以是可调的,诸如以基于模式生成不同量的ECC数据(例如,在正常编程模式下生成正常模式ECC数据,并且在突发编程模式下生成突发模式ECC数据,其中突发模式ECC数据大于正常模式ECC数据)。后端模块110还可以包括命令定序器126,该命令定序器126生成命令序列,诸如编程、读取和擦除命令序列,以传输到非易失性存储器管芯104。另外,后端模块110可以包括RAID(独立驱动器冗余阵列)模块128,其管理RAID奇偶校验的生成和失败数据的恢复。RAID奇偶校验可以用作写入到非易失性存储器设备100中的数据的附加级的完整性保护。在一些情况下,RAID模块128可以是ECC引擎124的一部分。存储器接口130向非易失性存储器管芯104提供命令序列,并且从非易失性存储器管芯104接收状态信息。与命令序列和状态信息一起,要编程到非易失性存储器管芯104中和从非易失性存储器管芯104读取的数据可以通过存储器接口130传送。在一个实施方案中,存储器接口130可以是双倍数据速率(DDR)接口,诸如切换模式200、400或800接口。闪存控制层132可以控制后端模块110的总体操作。

[0044] 因此,控制器102可以包括一个或多个管理表,用于管理存储系统100的操作。一种类型的管理表包括逻辑到物理地址映射表。逻辑到物理地址映射表的大小可能会随着存储器大小而增加。在这点上,高容量存储设备(例如,大于32G)的逻辑到物理地址映射表可能太大而不能存储在SRAM中,并且可以与用户和主机数据一起存储在非易失性存储器104中。因此,对非易失性存储器104的访问可能首先需要从非易失性存储器104读取逻辑到物理地

址映射表。

[0045] 图2A所示非易失性存储器设备100的附加模块可以包括媒体管理层138，该媒体管理层执行非易失性存储器管芯104的存储器单元的损耗均衡。非易失性存储器设备100还可以包括其他分立部件140，诸如外部电气接口、外部RAM、电阻器、电容器或可以与控制器102进行交互的其他部件。在另选实施方案中，RAID模块128、媒体管理层138和缓冲区管理/总线控制器114中的一者或多者是控制器102中可能不需要的任选部件。

[0046] 图2A所示的非易失性存储器设备100的其他模块可以包括对主机读取队列的请求的响应定时模块112和向主机发送中断的定时模块113。如下文更详细讨论的，存储器设备可以使用对主机读取队列的请求的响应定时模块112来确定对主机对读取队列(诸如完成队列)的请求的响应定时，包括基于影响队列的一个或多个动作(诸如在预定数量的硬件周期内更新完成队列)来延迟响应。存储器设备还可以使用向主机发送中断的定时模块113来确定向主机设备发送中断的定时(诸如发送中断，该中断指示甚至在存储器设备将条目放置在完成队列上之前条目就被放置在完成队列上)。

[0047] 图2B是更详细地示出非易失性存储器管芯104的示例性部件的框图。非易失性存储器管芯104可以包括非易失性存储器阵列142。非易失性存储器阵列142可以包括多个非易失性存储器元件或单元，其各自被配置成存储一个或多个数据位。非易失性存储器元件或单元可以是任何合适的非易失性存储器单元，包括采用二维和/或三维配置的NAND闪存存储器单元和/或NOR闪存存储器单元。存储器单元可以采用固态(例如，闪存)存储器单元的形式，并且可以是一次可编程、几次可编程或多次可编程的。此外，存储器元件或单元可以被配置为每个单元存储单位数据的单层单元(SLC)、每个单元存储多位数据的多层单元(MLC)或其组合。对于一些示例配置，多层单元(MLC)可以包括每个单元存储三位数据的三层单元(TLC)。

[0048] 另外，闪存存储器单元可以在阵列142中包括具有浮栅和控制栅的浮栅晶体管(FGT)。浮栅被绝缘体或绝缘材料包围，这有助于将电荷保留在浮栅中。浮栅内电荷的存在或不存在可能导致用于区分逻辑电平的FGT的阈值电压的偏移。也就是说，每个FGT的阈值电压可以指示存储在存储器单元中的数据。在下文中，FGT、存储器元件和存储器单元可以互换使用，以指代相同的物理实体。

[0049] 存储器单元可以根据存储器单元的行和列的矩阵状结构设置在存储器阵列142中。在行和列的交叉点是FGT(或存储器单元)。一列FGT可以称为一串。串或列中的FGT可以串联电连接。一行FGT可以称为一页。一页或一行中的FGT的控制栅可以电连接在一起。

[0050] 存储器阵列142还可以包括连接到FGT的字线和位线。每页FGT都耦接到字线。特别地，每个字线可以耦接到一页中的FGT的控制栅。此外，每串FGT可以耦接到位线。此外，单个串可以跨越多个字线，并且串中的FGT的数量可以等于块中的页数。

[0051] 非易失性存储器管芯104还可以包括页面缓冲区或数据高速缓存144，其高速缓存从存储器阵列142感测到的和/或将被编程到存储器阵列142的数据。非易失性存储器管芯104还可以包括行地址解码器146和列地址解码器148。当从存储器阵列142中的存储器单元读取或向存储器单元写入数据时，行地址解码器146可以解码行地址并选择存储器阵列142中的特定字线。列地址解码器148可以解码列地址，以选择存储器阵列142中要电耦接到数据高速缓存144的特定位线组。

[0052] 此外，非易失性存储器管芯104可以包括外围电路150。外围电路150可以包括提供状态信息到控制器102的状态机151。状态机151的其他功能将在下面进一步详细描述。

[0053] 图3示出了经由NVMe标准用于处理命令的步骤序列。如图所示，主机设备300包括主机存储器302，并且存储器设备包括控制器，诸如NVMe控制器310。在一个实施方式中，主机存储器302包括提交队列304和完成队列306。此外，在一个实施方式中，提交队列和完成队列可以具有1:1的相关性。另选地，提交队列和完成队列不具有1:1的相关性。

[0054] 实际上，在初始化阶段，主机设备300创建一个或多个提交队列和一个或多个对应的完成队列。特别地，主机设备300可以通过向存储器设备发送诸如每个队列的基地址的信息，来通知存储器设备提交队列和完成队列。在这点上，每个提交队列具有对应的完成队列。当提交队列和完成队列驻留在主机设备中时，主机设备向存储器设备发送信息，以便存储器设备确定提交队列和完成队列在主机设备中的位置。在一个具体实施方式中，主机设备发送指示提交队列和完成队列的创建的命令。该命令可以包括PRP1指针，其是指向主机设备上具体提交队列或具体完成队列的位置列表的指针。实际上，存储器设备使用PRP1发送TLP读取请求，以便获得PRP列表，并且将PRP列表存储在存储器设备中，以确定主机设备内的存储器位置，用于在从具体提交队列读取或写入具体完成队列的未来命令中使用。另选地，主机设备300可以指示存储器设备在驻留在存储器设备中的存储器(诸如控制器存储器缓冲区)中创建提交队列和对应的完成队列。

[0055] 提交队列304可以基于环形缓冲区，诸如图3所示，其具有头指针和尾指针。在创建提交队列并将有关创建的提交队列的情况通知给存储器设备之后，主机设备300可以向提交队列写入命令(或几个命令)。这在图3中表示为步骤1，标记为“队列命令”。特别地，图3示出了四个命令被写入提交队列。在一个实施方式中，存储器设备不知道主机设备300已经用四个命令更新了提交队列304，因为主机设备300更新了它自己的主机存储器302。在另一个实施方式中(诸如当提交队列和完成队列驻留在控制器存储器缓冲区中时)，存储器设备可以监控主机设备300和存储器设备之间的通信接口，以进行特定通信，诸如写入驻留在存储器设备上的提交队列。例如，存储器设备可以监控PCI Express总线上的传输层分组(TLP)，以确定主机设备300是否已经发送了导致对驻留在控制器存储器缓冲区中的提交队列进行更新的TLP。在这点上，存储器设备可以识别正被写入提交队列的一个或多个条目。

[0056] 在步骤2中，主机设备300写入存储器设备中的提交队列尾部门铃寄存器312。向提交队列尾部门铃寄存器312的这种写入向存储器设备表明主机设备在该具体提交队列304中排队了一个或多个命令(例如，如图3所示的4个命令)。向提交队列尾部门铃寄存器312的写入可以采取几种形式之一。一方面，主机设备300指示提交队列304的新尾部，从而指示写入提交队列304的命令数量。因此，由于存储器设备知道提交队列304的基地址，所以存储器设备只需要知道尾地址来指示写入提交队列304的新命令的数量。在命令(或一组命令)被处理之后，存储器设备随后相应地设置提交队列304的新头部。因此，尾指针可以表示与头指针的“偏移”。另一方面，主机设备300指示写入提交队列304的命令的数量。实际上，每个提交队列304在存储器设备中具有对应的提交队列尾部门铃寄存器，使得当主机设备300更新特定门铃寄存器(与特定提交队列304相关)时，存储器设备可以基于门铃寄存器确定哪个特定提交队列304已经被更新。

[0057] 在步骤2(其中存储器设备被通知提交队列304上的命令)之后和步骤3(其中存储

器设备提取命令)之前,存储器设备知道提交队列304中有挂起的命令。在一般情况下,可能有几个提交队列(在几个提交队列中可能有许多挂起的命令)。因此,在执行步骤3之前,存储器设备控制器可以在各种提交队列之间进行仲裁,以选择从中提取命令的特定提交队列。

[0058] 响应于确定从哪个特定提交队列304提取命令,在步骤3,存储器设备从特定提交队列304提取命令。实际上,存储器设备可以访问特定提交队列304的基地址加上主机设备300中实现的当前头指针上的指针。

[0059] 如上文所讨论的,提交队列或完成队列可以被分配存储器区域(诸如在主机设备中或在存储器设备中的控制器存储器缓冲区中)。提交队列和完成队列可以包括多个条目,每个条目与具体命令相关联。每个条目的大小可以是预定的大小,诸如64Kb。在这点上,通过使用提交队列的基地址,并且将基地址偏移条目数量与每个条目的大小(例如,64Kb)的乘积,提交队列中的条目可以被确定。

[0060] 如上文所讨论的,存储器设备知道已经通过步骤2通知的尾指针。因此,存储器设备可以从提交队列304获得所有新命令。在驻留在主机设备上的提交队列中,存储器设备可以发送TLP请求以从提交队列304获得命令。响应于接收到TLP请求,主机设备300发送带有提交队列304中的命令的完成TLP消息。在这点上,在步骤3结束时,存储器设备从提交队列304接收命令。

[0061] 在步骤4,存储器设备处理该命令。在一个实施方式中,存储器设备解析命令,并确定执行命令(例如,读/写/等)的步骤。例如,命令可以包括读取命令。响应于读取命令的接收,存储器设备解析读取命令,实现地址转换,并访问闪存以接收数据。在接收到数据之后,存储器设备基于命令中的信息(例如,下文讨论的PRP 1)使数据存储在主机设备上。作为另一示例,命令可以包括写入命令。响应于写入命令的接收,存储器设备解析写入命令,确定数据在经受写入的主机设备上的位置,从主机设备上的位置读取数据,并将数据写入闪存存储器。

[0062] 特别地,存储器设备可以接收带有PRP1指针的读取命令或写入命令。例如,其中主机设备请求存储器设备从闪存存储器读取的读取命令包括指向PRP列表的PRP1指针。存储器设备获得PRP列表,以便确定主机设备内的存储器位置,以写入从闪存存储器读取的数据。作为另一示例,其中主机设备请求存储器设备将数据写入闪存存储器的写入命令包括指向PRP列表的PRP1指针。存储器设备获得PRP列表,以便确定从主机设备中读取数据的存储器位置(然后将读取的数据保存到闪存存储器)。

[0063] PRP列表中的每个条目可以与主机设备存储器中的某个部分相关联,并且可以是预定大小,诸如4Kb。因此,在1Mb的传输中,PRP列表中可能有250个引用,每个引用的大小为4Kb。实际上,存储器设备可以无序地检索数据。这可能是由于要检索的数据在几个闪存管芯上,这些管芯可用于在不同时间的数据检索。例如,在检索对应于1Mb传输的0Kb-100Kb的数据之前,存储器设备可以检索对应于1Mb传输的100Kb-200Kb的数据。然而,因为存储器设备具有PRP列表(并且因此知道主机设备期望存储对应于100Kb-200Kb的数据的存储器位置),所以存储器设备可以传输对应于1Mb传输的100Kb-200Kb的数据,而不必首先检索对应于1Mb传输的0Kb-100Kb的数据。

[0064] 在NVMe中,可以有多个PCI Express TLP来将数据从存储器设备传输到主机设备

300。通常，基于命令中的指示(例如，命令包括存储所请求的数据的地址)，将传输的数据存储在主机设备300的主机存储器302中。

[0065] 在完成数据传输之后，在步骤5，存储器设备控制器向相关完成队列306发送完成消息。如上所述，在初始化阶段，主机设备300将提交队列与完成队列相关联。因此，主机设备300基于存储器设备写入哪个完成队列来知道在提交队列中完成的命令。完成消息可以包含关于命令处理的信息，诸如命令是否成功完成或者在执行命令时是否有错误。

[0066] 在步骤5之后，主机设备300不知道发布到完成队列306的存储器设备。这是由于存储器设备导致数据被写入完成队列306。在这点上，在步骤6，存储器设备通知主机设备300已经对完成队列306进行了更新。特别地，存储器设备向主机设备300发布中断(例如，在NVMe中，主机设备300可以使用MSIe中断)。如下文更详细讨论的，存储器设备可以在将条目发布到完成队列之前对中断的发送进行计时。

[0067] 响应于接收到中断，主机设备300确定在该完成队列306中有针对主机设备300挂起的一个或多个完成条目。在步骤7，主机设备300然后处理完成队列306中的条目。例如，在完成队列驻留在存储器设备中的情况下，主机设备可以发送TLP读取请求来读取驻留在存储器中的完成队列。如下文更详细讨论的，存储器设备可以延迟对主机设备的读取请求的响应。

[0068] 在主机处理来自完成队列306的条目之后，在步骤8，主机设备300将主机设备300处理的来自完成队列306的条目通知给存储器设备。这可以通过更新完成队列头部门铃寄存器314来执行，完成队列头部门铃寄存器314指示存储器设备主机设备300处理了来自完成队列306的一个或多个条目。当主机发出完成队列门铃写入时，相关中断合并向量的参数可以被更新以反映这种变化。例如，完成队列的状态可以从几乎满的状态变为几乎空的状态。结果，中断可能被刷新到主机设备。

[0069] 响应于更新完成队列头部门铃寄存器314，存储器设备更新完成队列306的头部。给定新的头部，存储器设备知道完成队列306中的哪些条目已经被主机设备300处理并且可以被重写。

[0070] 图4是主机系统400和存储器设备420的其他示例性部件的框图。主机系统400包括一个或多个处理器402和主机存储器404。主机存储器404可以包括物理区域页面(PRP)406、数据缓冲区408和其他存储器410。某些NVMe命令，诸如读取命令和写入命令，可以包括指向PRP列表的指针，该列表限定了主机设备存储器中的部分。例如，读取命令可以包括指向PRP列表的指针，其中PRP列表指示存储器中的部分，在所述部分中，存储器设备应该存储响应于读取命令而读取的数据。作为另一示例，写入命令可以包括指向PRP列表的指针，其中PRP列表指示存储器中的部分，在所述部分中，存储器设备应该读取要存储在存储器设备的闪存存储器上的数据。在处理命令时，存储器设备可以通过向主机设备发送一个或多个PRP提取请求来获得PRP列表。在这点上，存储器设备可以发送几个PRP提取请求，这些请求与不同的NVMe命令相关联。

[0071] 图4进一步示出了主机设备400和存储器设备420之间的通信接口。在第一实施方式中(图4中未示出)，主机设备和存储器设备之间的通信接口是单工的，向存储器设备的通信和来自存储器设备的通信在同一路径上。在第二实施方式中(图4中示出)，主机设备400和存储器设备420之间的通信接口是双工的，具有单独的入口路径和单独的出口路径。从存

储器设备420的角度来看，入口路径包括从主机设备400到存储器设备420的传入请求。相反，从存储器设备420的角度来看，出口路径包括从存储器设备420到主机设备400的传出请求。

[0072] 传入请求(从主机设备400到存储器设备420的请求)可以以不同的方式分割，诸如传入的读取请求和传入的写入请求。例如，主机设备400可以经由入口路径发送读取存储器设备420中的存储器的部分(诸如下面讨论的控制器存储器缓冲区(CMB)426)的读取请求或者写入存储器设备420中的存储器的部分的写入请求。同样，存储器设备420可以经由出口路径发送向主机设备400中的存储器的部分的读取请求或向主机设备400中的存储器的部分的写入请求。

[0073] 在使用NVMe的实践中，可能存在一系列读取请求(主机设备对读取驻留在存储器设备上的数据的请求，反之亦然)和一系列写入请求(主机设备对将数据写入驻留在存储器设备上的位置的请求，反之亦然)。特别地，在NVMe中，存储器设备和主机设备使用事务层分组(TLP)请求彼此通信，诸如在其他设备上执行读取的TLP读取请求，或者在其他设备上执行写入的TLP写入请求。在一个示例中(提交队列和完成队列驻留在主机设备上)，响应于主机设备对存储器设备上的门铃寄存器的TLP写入请求(经由入口路径发送)(对门铃寄存器的写入指示提交队列上有命令)，存储器设备使用TLP读取请求(经由出口路径发送)从提交队列(驻留在主机设备上)提取写入命令。因此，写入命令是对存储器设备向非易失性存储器写入数据的请求。然后，存储器设备解析写入命令以获得信息，诸如指向PRP列表的PRP指针(例如PRP1)的指示。PRP列表是诸如指针或地址的一系列信息，指示数据在主机设备中的位置。然后，存储器设备使用另一个TLP读取请求从PRP列表中的指针或地址读取数据。此后，存储器设备通过将数据存储在存储器设备上的非易失性存储器(例如闪存存储器)中来执行写入。在存储数据之后，存储器设备使用TLP写入请求将条目写入完成队列(指示写入命令已经完成)。最后，存储器设备使用TLP写入请求来生成对主机设备的中断，其中中断向主机设备发信号通知完成队列上有条目。响应于中断，主机设备读取完成队列上的条目，然后向CQ门铃写寄存器发出TLP写入请求，指示主机设备已经审查了完成队列上的条目。

[0074] 作为另一示例(提交队列和完成队列同样驻留在主机设备上)，响应于主机对存储器设备上的门铃寄存器的TLP写入请求(对门铃寄存器的写入指示提交队列上有命令)，存储器设备使用TLP读取请求从提交队列(驻留在主机设备上)提取读取命令。因此，读取命令是对存储器设备从非易失性存储器读取数据并将读取的数据发送到主机设备的请求。然后，存储器设备读取非易失性存储器(例如闪存存储器)以读取数据。存储器设备可以对数据执行一系列操作，诸如错误校正、加密/解密等，存储缓冲区散布在该系列操作中的每一个之间。然后，存储器设备可以解析读取命令以获得信息，诸如指向PRP列表的PRP指针(例如PRP1)的指示。PRP列表是诸如指针或地址的一系列信息，其指示主机设备中存储从非易失性存储器读取的数据(以及任选地错误校正、加密等的数据)的位置。存储器设备使用TLP读取请求从PRP列表中的指针或地址读取数据。此后，存储器设备使用TLP写入请求来写入从非易失性存储器读取的数据。在将数据写入主机设备之后，存储器设备使用TLP写入请求将条目写入完成队列(指示读取命令已经完成)。最后，存储器设备使用TLP写入请求来生成对主机设备的中断，其中中断向主机设备发信号通知完成队列上有条目。响应于中断，主机设备读取完成队列上的条目，然后向CQ门铃写寄存器发出TLP写入请求，指示主机设备已经

审查了完成队列上的条目。

[0075] 任选地，完成队列和提交队列可以驻留在存储器设备中，诸如在控制器存储器缓冲区(CMB)426中，其部分或全部被分配给主机设备400。在这种情况下，主机设备可以向存储器设备发送TLP读取请求(经由入口路径发送)，以从完成队列中读取。同样，存储器设备可以向主机设备发送TLP写入请求(经由出口路径发送)以生成中断。例如，图4示出了完成队列驻留在主动存储428中，而其他数据结构驻留在被动存储430中。处理器424可以监控到CMB的一些或全部通信。在一个实施方式中，处理器424可以监控与完成队列相关的通信，诸如TLP读取请求。响应于处理器424检测到与完成队列相关的通信，处理器424可以分析该通信并相应地动作。例如，处理器424可以识别通信涉及读取完成队列。响应于该确定，处理器424可以基于在预定时间段内添加到完成队列的其他条目来确定延迟响应，如下面更详细讨论的。在这点上，存储428基于对指向它的通信(诸如读取)的主动监控而是主动的。相反，存储430基于处理器424不监控指向它的通信而是被动的。

[0076] 在一个实施方式中，当主机设备400访问CMB时，存储器设备420首先检测访问是到完成队列区还是到其他区。对于其他区，实现诸如SRAM的被动存储器，并且主机设备400能够直接向该存储器发出读取/写入请求。对于完成队列区，存储器设备420实现有效逻辑，该有效逻辑解析事务并以不同的方式对每个事务进行响应，如下面更全面解释的。

[0077] 此外，在一个实施方式中，主机设备将其中的存储器分配给提交队列和完成队列，提交队列和完成队列可以物理上相邻或不相邻定位。另选地，主机设备400指示存储器设备420将存储器分配给CMB 426中的提交队列和完成队列。

[0078] 存储器设备420包括存储器设备控制器422和存储器阵列450。存储器阵列450可以以各种方式分割，诸如分割成图4所示的10个部分。存储器设备控制器422可以包括一个或多个处理器424，并入PCIe MAC和PHY接口432中的一个或全部，并且并入其他HW和FW部件。

[0079] 命令解析器434被配置为解析从提交队列提取的命令(无论提交队列是驻留在存储器设备420中还是主机设备400中)。命令执行器436被配置为仲裁和执行从提交队列中提取和解析的命令。调度器444被配置为调度一种或多种类型的数据传输。作为一个示例，读取的数据可以经由闪存接口模块438从不同的存储器阵列450并行到达。调度器444可以从不同的数据传输中进行仲裁。作为另一示例，调度器444负责控制数据传输，同时激活控制路径440以提取PRP、发布完成和中断，并激活DMA 442以在主机设备400和存储器设备420之间进行实际数据传输。

[0080] 闪存接口模块438被配置为控制和访问存储器阵列450。存储器设备控制器422还包括错误校正446，错误校正446可以对从存储器阵列450提取的数据进行错误校正，并且可以包括低密度奇偶校验(LDPC)，其是线性错误校正码。可以设想其他错误校正方法。

[0081] 提前主机中断发布器448被配置为在启用时发布主机设备中断。如下文更详细讨论的，考虑到PCIe和主机时延，提前主机中断发布器448甚至可以在完成命令之前将中断发布到主机设备400。在一个实施方式中，提前发布时间是自适应的，并且可能取决于先前的事务时延。该先前的事务时延可以存储在存储器设备中，并且可以指示主机设备响应中断的时延。实际上，就在存储器设备420用条目更新完成队列之后，主机设备400将提取相关的完成队列条目。

[0082] 图5是确定是否延迟响应来自主机设备的读取控制器存储器缓冲区中的完成队列

的请求的第一示例方法的流程图500。在502，存储器设备从主机设备接收读取CMB中的队列(诸如完成队列)的请求。如上文所讨论的，存储器设备控制器422可以监控到存储器的各个部分(诸如CMB)的通信，以确定通信是否与队列(诸如完成队列)相关。在504，存储器设备确定是否延迟对请求的响应。如上文所讨论的，存储器设备可以基于检测到的各种操作来确定延迟对读取队列的请求的响应。一个示例包括影响队列的操作，诸如预定时间段(例如，预定数量的硬件周期)内的预期未来活动(例如，向完成队列发布条目)。如果在504，没有确定延迟响应，则在506，存储器设备发送响应。如果在504，确定延迟响应，则在508，存储器设备可以确定延迟的长度(例如，条目将被发布到完成队列的估计时间)，在510，等待延迟的长度，并且在512，发送响应。另选地，存储器设备可以基于事件的检测来触发响应的发送，而不是基于经过的时间来触发响应的发送。例如，存储器设备可以确定等待，直到存储器设备检测到条目被发布到完成队列。响应于该确定，存储器设备然后可以响应于主机设备查询来读取完成队列。

[0083] 图6是确定是否延迟响应来自主机设备的读取控制器存储器缓冲区中的完成队列的请求的第二示例方法的流程图600。在602，当完成队列驻留在CMB中时，主机设备发出试图访问完成队列条目的完成队列TLP读取请求。在604，存储器设备确定相关完成队列条目是否可用并存储在存储器设备内部。如果相关完成队列可用并存储在内部，在606，存储器设备可以通过向主机设备提供所需的条目来立即完成事务。否则，在608，存储器设备检查是否存在将很快完成的与该完成队列相关联的命令(例如，在预定数量的硬件周期内，以便不引起超时错误)。如果是，在610，存储器设备推迟事务并在条目可用时完成它。否则，在612，存储器设备立即完成事务，同时提供主机设备将理解所提供的条目无效的条目。

[0084] 图7是确定在发布到完成队列之前是否向主机设备发送中断的示例方法的流程图700。在702，存储器设备提取命令并开始执行该命令。在704，存储器设备确定在发布到完成队列之前是否发送中断。另选地，存储器设备可以确定在完成命令的执行之前发送中断。例如，在写入请求中，存储器设备可以确定在完成将数据写入存储器设备上的闪存存储器之前发布中断。如果不发送，在706，存储器设备仅在命令的执行完成之后才发布到完成队列。此后，在708，存储器设备向主机设备发送中断。

[0085] 如果存储器设备确定在发布完成队列或执行完成之前发送中断，则在710，存储器设备可以响应于中断而访问主机设备的时延。存储器设备可以记录主机设备响应于先前中断的定时，以便确定主机时延(例如，主机设备响应中断需要多长时间，以及主机设备向完成队列发送TLP读取请求需要多长时间)。在712，存储器设备进一步估计在发布到完成队列之前的时间(包括完成命令的执行的时间)。基于主机时延和发布到完成队列的估计时间，在714，存储器设备确定发送中断的时间。在716，存储器设备在确定的时间发送中断。

[0086] 在718，存储器设备确定在存储器设备将条目发布到完成队列之前，读取完成队列的主机设备TLP读取请求是否已经到达。如果没有，在720，存储器设备立即发送响应(包括从完成队列读取的条目)。如果到达，存储器设备没有正确估计，并且主机设备比预期更快地发送读取请求。在这种情况下，在724，存储器设备延迟对读取请求的响应，直到存储器设备将条目发布到完成队列之后。在存储器设备估计值明显出错，使得延迟大于超时误差的情况下，存储器设备可以从完成队列发送旧条目，以向主机设备指示发送的旧条目无效。

[0087] 图8是示出现有技术时序图800和主机设备访问存储在控制器存储器缓冲区中的

完成队列的一个实现的时序图850之间的差异的时序图。在现有技术的实施方式中，存储器设备中的控制器以被动方式管理CMB，其中访问CMB以读取其中的完成队列的请求作为普通命令被处理，而不考虑与CMB(或其中的完成队列)相关的任何其他活动。在这点上，现有技术实施方式中的存储器设备发布到CMB中的完成队列，并且主机设备从CMB中提取条目。结果，在PCIe上，从完成队列(CQ)读取请求TLP到其CQ完成TLP(存储器设备响应于报告来自完成队列的条目而发回主机的TLP)的时间段相对固定，并且等于PCIe周转时间。这在图8中示出为“周转时间”805、810、815在不同的CQ读取请求TLP之间是相同的。

[0088] 相反，在一个实施方式中，存储器设备中的控制器可以主动管理CMB，并且在某些情况下，推迟事务(例如，CQ完成TLP)，从而导致自适应延迟。如上文所讨论的，存储器设备可以基于CMB中的活动来延迟响应，诸如存储器设备确定服从TLP读取请求的完成队列将在预定时间段内已经在其中发布了条目。因此，如图8所示，周转时间855不受延迟的影响，并且与周转时间805、810和815相同。然而，周转时间860和865不同于周转时间805、810和815。特别地，周转时间860和865是比周转时间805、810和815更长的时间段，并且示出了存储器设备在响应中的延迟。因此，周转时间860和865表示存储器设备完成完成队列上的活动(诸如向完成队列发布新条目)并发送CQ完成TLP的时间段。尽管周转时间860和865比周转时间805、810和815长，但是可以提高主机设备和存储器设备之间的通信的整体效率。

[0089] 图9是示出现有技术时序图900和存储在控制器存储器缓冲区中的完成队列的提前中断发布的一个实现的时序图950之间的差异的时序图。如图9所示，定时是从主机设备顺序队列(SQ)门铃写入(主机指示顺序队列上有命令)到主机设备完成队列(CQ)门铃写入(主机指示它已经从完成队列中检索到条目)来测量的。在现有技术时序图900的操作中，在接收到SQ门铃写入操作之后，存储器设备从提交队列中提取命令并启动数据传输。在完成数据传输后，存储器设备将条目写入CMB中的完成队列，并发布中断。此后，主机设备提取相关的CQ条目(使用CQ读取请求TLP)，接收响应(以CQ完成TLP的形式)，并最终发送CQ门铃写入(指示主机设备读取完成队列上的条目)。

[0090] 如图9所示，SQ门铃写入、NVMe命令提取(其中存储器设备从顺序队列中提取命令)以及数据传输的开始对于900和950是相同的定时。相反，在现有技术时序图900中，存储器设备仅在数据传输已经完成之后(例如，在读取命令中，存储器设备已经将从闪存存储器读取的所有数据写入主机设备)向主机设备发出中断，并且条目被发布到完成队列。在实现的时序图950中，中断甚至在数据传输完成之前就被发送到主机设备。响应于中断，主机设备发送CQ读取请求TLP，实际上是向存储器设备发送读取驻留在CMB中的完成队列的请求。因此，中断会提前发布，甚至在完成数据传输之前。在该实施方式中，存储器设备可以确定发布中断的精确时间。在一个具体实施方式中，存储器设备可以自适应地确定定时。例如，优选的定时是当存储器设备的内部逻辑接收到CQ读取请求TLP并考虑周转时间时(就服务于CQ读取请求TLP的存储器设备而言)，相关条目已经可用并存储到完成队列中。在这点上，存储器设备可以发出CQ完成TLP(从完成队列读取相关条目)。此后，主机设备可以发出CQ门铃写入，从而结束时序。如下所示，与现有技术相比，NVMe命令的寿命缩短了。

[0091] 因此，通过考虑处理中断时的主机设备时延和通信中的PCIe时延，存储器设备可以向主机设备发送关于完成队列条目可用性的提前通知(以中断的形式)。如图9所示，NVMe命令时间线显著缩短，直接导致性能的提高，尤其是当具有低队列深度时。在一个实施方式

中，中断的提前通知时间是自适应的，并且可以取决于一个或多个方面，诸如存储器设备队列深度、过去测量的时延和配置。关于队列深度，发布与主机设备提取相关完成队列条目同步的提前中断是相关的，尤其是在低队列深度配置中(例如，一次处理1或2个命令)。例如，在队列深度为一的情况下，主机设备仅在接收到前一个命令的完成指示之后才发送下一个命令。本方法使主机设备更早获得该完成条目，从而提高性能。作为另一示例，高队列深度中的精确定时不太重要，因为带宽可能是相同的。然而，时延得到了改善。

[0092] 关于时延，主机设备中断时延的先前测量可以改变存储器设备发送提前中断的定时。这些测量可以基于完成队列(例如，分别测量一个、一些或每个完成队列的时延)，因为不同的完成队列可以被分配给不同的主机设备CPU，每个CPU潜在地具有不同的时延。此外，时延可能取决于相关完成队列的状态。例如，完成队列可能具有空(没有完成队列条目)、几乎空、几乎满和满的状态。特别地，当相应的完成队列已满时，存储器设备可以调整定时，使得主机设备将尽可能快地获得条目，因为主机设备在此期间处于空闲状态。关于配置，固件可以基于本发明的自适应方法来分析响应。在一个实施方式中，固件可以基于分析禁用自适应方法。在另选的实施方式中，固件可以将自适应方法应用于NVMe协议的某些方面，而不将自适应方法应用于NVMe协议的其他方面。

[0093] 在存储器设备过早发送中断，从而导致主机设备在条目被发布到完成队列之前发送CQ读取请求TLP的情况下，存储器设备仍然可以处理CQ读取请求TLP。特别地，由于完成队列位于CMB中，存储器设备可以延迟对CQ读取请求TLP的响应，并且一旦完成条目可用，就发送响应。此外，在存储器设备发送中断之后的数据传输期间存在任何错误的情况下，存储器设备可以简单地使用将在完成数据传输之后提供的完成队列条目来更新主机设备。

[0094] 图10是示出现有技术时序图1000和存储在主机设备中的完成队列的提前中断发布的一个实现的时序图1050之间的差异的时序图。与图9相反，顺序队列和完成队列驻留在主机设备中。因此，不能在连接到存储器设备的PCIe总线上监控主机设备完成队列访问；然而，存储器设备可以监控所有其他事务，包括SQ门铃写入、NVMe命令提取、数据传输、CQ条目写入、中断发布和CQ门铃写入，

[0095] 如图10所示，定时是从主机设备顺序队列(SQ)门铃写入(主机指示顺序队列上有命令)到主机设备完成队列(CQ)门铃写入(主机指示它已经从完成队列中检索到条目)来测量的。在现有技术时序图1000的操作中，在接收到SQ门铃写入操作之后，存储器设备从提交队列中提取命令并启动数据传输。在完成数据传输后，存储器设备将条目写入主机设备中的完成队列，并发布中断。响应于中断，主机设备读取驻留在主机设备中的完成队列上的条目(显示为主机时延)。此后，主机设备执行CQ门铃写入(指示主机设备读取完成队列上的条目)。在这点上，由存储器设备执行的事务被一个接一个地执行，而没有并行性。

[0096] 如图10所示，SQ门铃写入、NVMe命令提取(其中存储器设备从顺序队列中提取命令)以及数据传输的开始对于1000和1050是相同的定时。相反，在现有技术时序图1000中，存储器设备仅在数据传输已经完成之后(例如，在读取命令中，存储器设备已经将从闪存存储器读取的所有数据写入主机设备)向主机设备发出中断，并且条目被发布到完成队列(CQ写入)。在实现的时序图1050中，中断甚至在数据传输完成之前和条目被发布到完成队列之前就被发送到主机设备。响应于中断，主机设备读取驻留在主机设备中的完成队列上的条目(显示为主机时延)。因为完成队列驻留在主机设备上，所以存储器设备不能延迟响应，诸

如图9中可能完成的那样。在这点上，存储器设备对中断的发送进行计时，使得在主机设备读取CQ(主机CQ读取)之前，将条目发布到完成队列(CQ写入)。此后，主机设备执行CQ门铃写入。如图所示，图1050的NVMe命令的时间寿命比图1000的NVMe命令的时间寿命短。因此，存储器设备提前发布中断，所以在存储器设备更新完成队列上的条目之后立即执行相关的主机CQ提取。

[0097] 最后，如上所述，可以使用任何合适类型的存储器。半导体存储器设备包括易失性存储器设备，诸如动态随机存取存储器(“DRAM”)或静态随机存取存储器(“SRAM”)设备，非易失性存储器设备，诸如电阻式随机存取存储器(“ReRAM”)、电可擦除可编程只读存储器(“EEPROM”)、闪存存储器(也可以被认为是EEPROM的子集)、铁电随机存取存储器(“FRAM”)和磁阻随机存取存储器(“MRAM”)，以及能够存储信息的其他半导体元件。每种类型的存储器设备可具有不同的配置。例如，闪存存储器设备可以NAND配置或NOR配置进行配置。

[0098] 该存储器设备可由无源元件和/或有源元件以任何组合形成。以非限制性示例的方式，无源半导体存储器元件包括ReRAM设备元件，其在一些实施方案中包括电阻率切换存储元件诸如反熔丝、相变材料等，以及任选地包括导引元件诸如二极管等。进一步以非限制性示例的方式，有源半导体存储器元件包括EEPROM和闪存存储器设备元件，其在一些实施方案中包括包含电荷存储区域的元件，诸如浮栅、导电纳米粒子或电荷存储介电材料。

[0099] 多个存储器元件可被配置为使得它们串联连接或者使得每个元件可被单独访问。以非限制性示例的方式，NAND配置中的闪存存储器设备(NAND存储器)通常包含串联连接的存储器元件。NAND存储器阵列可被配置为使得该阵列由存储器的多个串构成，其中串由共享单个位线并作为组被访问的多个存储器元件构成。另选地，存储器元件可被配置为使得每个元件均为单独可访问的，例如，NOR存储器阵列。NAND和NOR存储器配置是示例性的，并且存储器元件可以其他方式配置。

[0100] 位于基板内和/或上方的半导体存储器元件可被布置成两个或三个维度，诸如二维存储器结构或三维存储器结构。

[0101] 在二维存储器结构中，半导体存储器元件被布置在单个平面或单个存储器设备级中。通常，在二维存储器结构中，存储器元件被布置在平面中(例如，在x-z方向平面中)，所述平面基本上平行于支撑存储器元件的基板的主表面延伸。基板可以是存储器元件的层在其之上或之中形成的晶圆，或者其可以是在存储器元件形成后附接到其的承载基板。作为非限制性示例，基板可包括半导体，诸如硅。

[0102] 存储器元件可被布置在处于有序阵列中(诸如在多个行和/或列中)的单个存储器设备级中。然而，存储器元件可以非常规配置或非正交配置排列。存储器元件可各自具有两个或更多个电极或接触线，诸如位线和字线。

[0103] 三维存储器阵列被布置成使得存储器元件占据多个平面或多个存储器设备级，从而形成三个维度(即，在x方向、y方向和z方向上，其中y方向基本上垂直于基板的主表面，并且x方向和z方向基本上平行于基板的主表面)的结构。

[0104] 作为非限制性示例，三维存储器结构可被垂直地布置为多个二维存储器设备级的叠堆。作为另一个非限制示例，三维存储器阵列可被布置为多个垂直列(例如，基本上垂直于基板的主表面延伸的列，即，在y方向上)，其中在每一列中每一列均具有多个存储器元件。列可以以二维配置例如在x-z平面中布置，从而得到存储器元件的三维布置，其中元件

位于多个垂直堆叠的存储器平面上。三维存储器元件的其他配置也可构成三维存储器阵列。

[0105] 以非限制性示例的方式，在三维NAND存储器阵列中，存储器元件可耦接在一起以在单个水平(例如，x-z)存储器设备级内形成NAND串。另选地，存储器元件可耦接在一起以形成横贯多个水平存储器设备级的垂直NAND串。可设想到其他三维配置，其中一些NAND字符串包含在单个存储器级中的存储器元件，而其他字符串则包含跨越多个存储器级的存储器元件。三维存储器阵列也可以NOR配置以及ReRAM配置来设计。

[0106] 通常，在单片三维存储器阵列中，一个或多个存储器设备级在单个基板上方形成。任选地，单片三维存储器阵列还可具有至少部分地在单个基板内的一个或多个存储器层。作为非限制性示例，基板可包括半导体，诸如硅。在单片三维阵列中，构成阵列的每个存储器设备级的层通常形成在阵列的底层存储器设备级的层上。然而，单片三维存储器阵列的相邻存储器设备级的层可被共享或具有介于存储器设备级之间的居间层。

[0107] 然后，可单独形成二维阵列，然后封装在一起以形成具有多个存储器层的非单片存储器设备。例如，非单片的堆叠存储器可通过在单独的基板上形成存储器级并然后将存储器级堆叠在彼此之上而构造。可在堆叠前将基板减薄或从存储器设备级移除，但由于存储器设备级在单独基板上初始形成，因此所得的存储器阵列不是单片的三维存储器阵列。此外，多个二维存储器阵列或三维存储器阵列(单片或非单片)可在单独的芯片上形成，然后封装在一起以形成堆叠的芯片存储器设备。

[0108] 通常需要相关联的电路来操作存储器元件并与存储器元件通信。作为非限制性示例，存储器设备可具有用于控制并驱动存储器元件以实现诸如编程和读取的功能的电路。该相关联的电路可与存储器元件位于同一基板上和/或位于单独的基板上。例如，用于存储器读取-写入操作的控制器可位于单独的控制器芯片上和/或位于与存储器元件相同的基板上。

[0109] 预期将前面的详细描述理解为本发明可以采用的选定形式的说明，而不是作为本发明的定义。预期只有以下权利要求(包括所有等同物)限定要求保护的本发明的范围。最后，应当指出的是，本文所述的任何优选实施方案的任何方面均可单独使用或彼此组合使用。

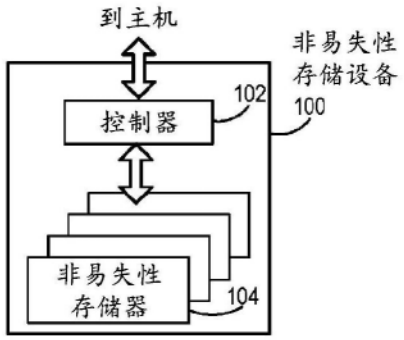

到主机
非易失性
存储设备
控制器
102
100
非易失性
存储器
104

图1A

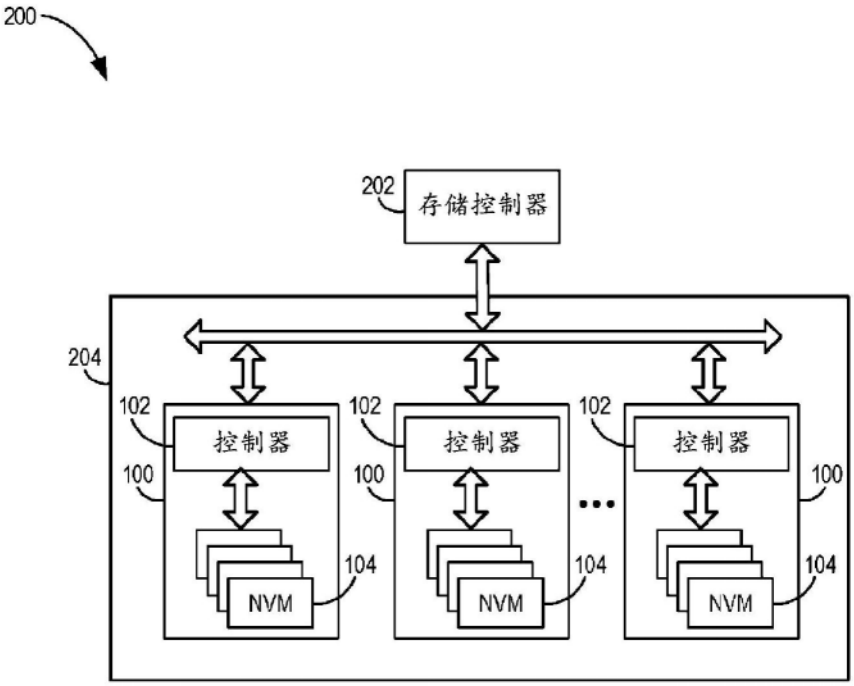

200
202
存储控制器
204
102
控制器
100
104
NVM

图1B

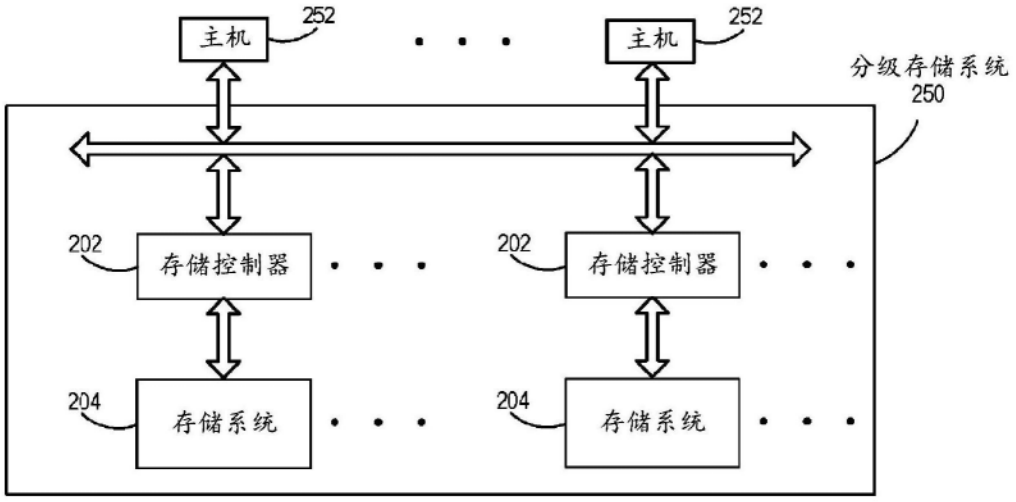

主机
252
主机
252
分级存储系统
250
202
存储控制器
202
存储控制器
204
存储系统
204
存储系统

图1C

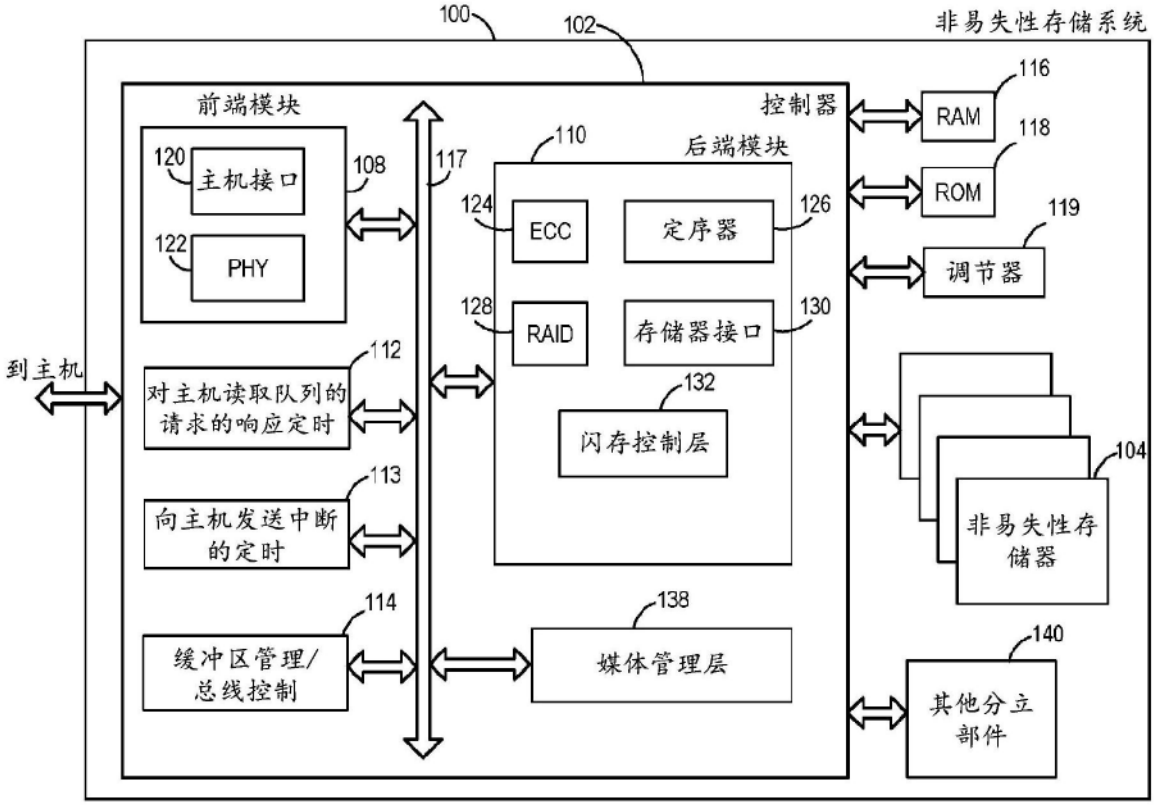

100
102
非易失性存储系统
前端模块
控制器
后端模块
110
116
RAM
118
ROM
119
调节器
120
主机接口
108
117
122
PHY
124
ECC
定序器
126
128
RAID
存储器接口
130
132
闪存控制层
到主机
112
对主机读取队列的请求的响应定时
113
向主机发送中断的定时
114
缓冲区管理/总线控制
138
媒体管理层
104
非易失性存储器
140
其他分立部件

图2A

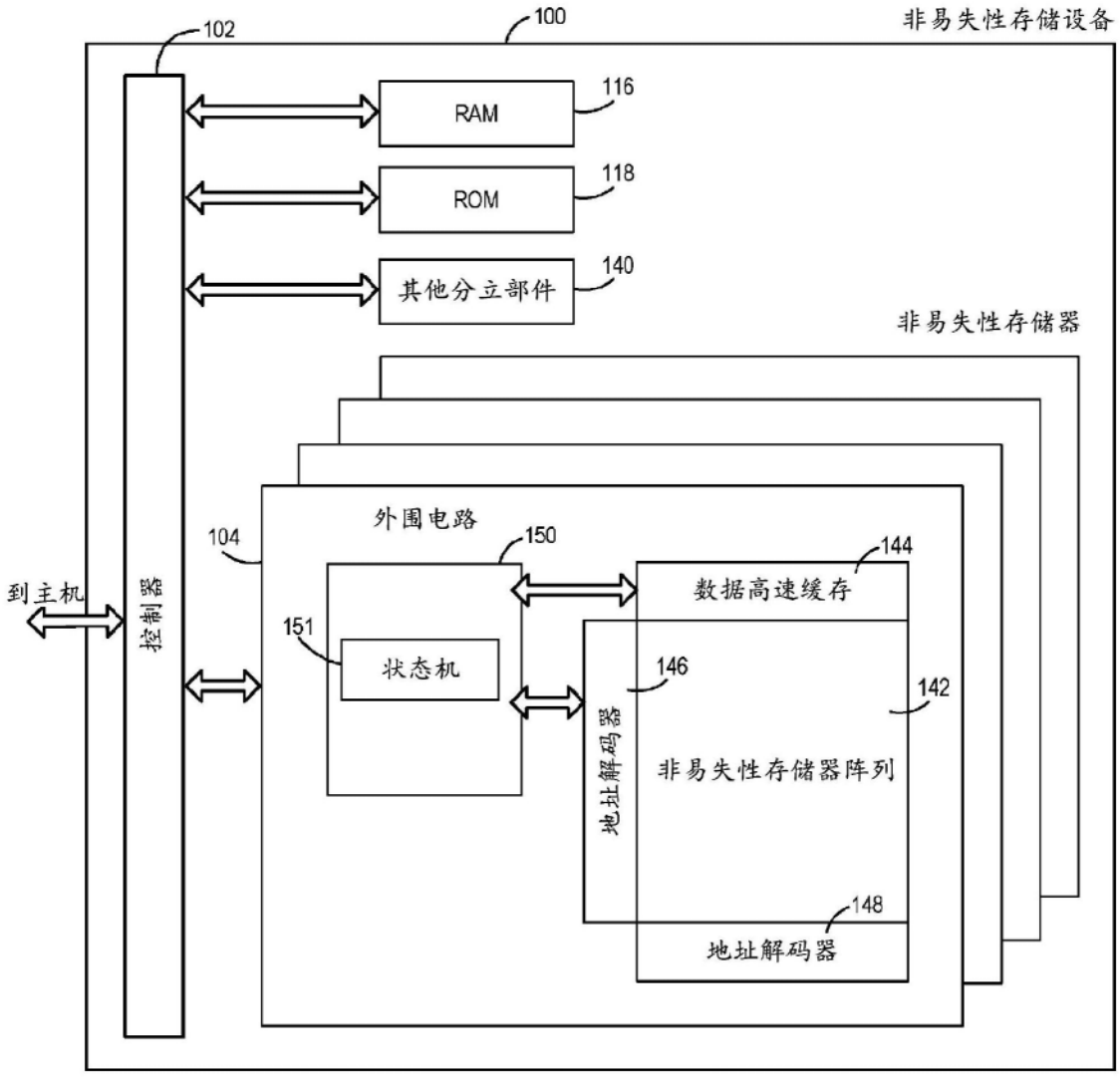

非易失性存储设备
100
102
RAM
116
ROM
118
其他分立部件
140
非易失性存储器
外围电路
104
150
144
数据高速缓存
到主机
控制器
151
状态机
146
142
地址解码器
非易失性存储器阵列
148
地址解码器

图2B

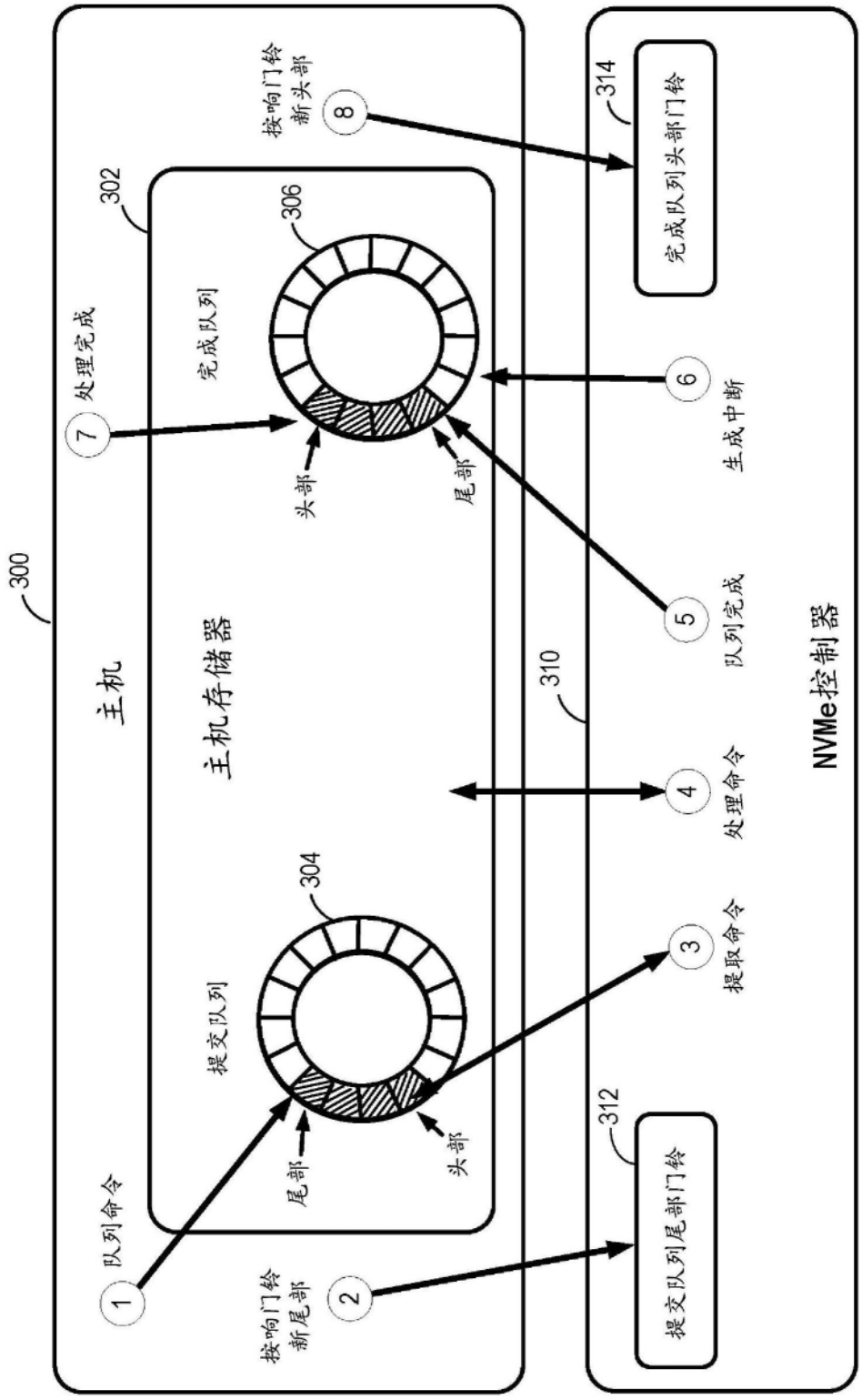

300
主机
主机存储器
302
提交队列
304
完成队列
306
310
NVMe控制器
312
提交队列尾部门铃
314
完成队列头部门铃
1 队列命令
2 按响门铃新尾部
3 提取命令
4 处理命令
5 队列完成
6 生成中断
7 处理完成
8 按响门铃新头部
尾部
头部
头部
尾部

图3

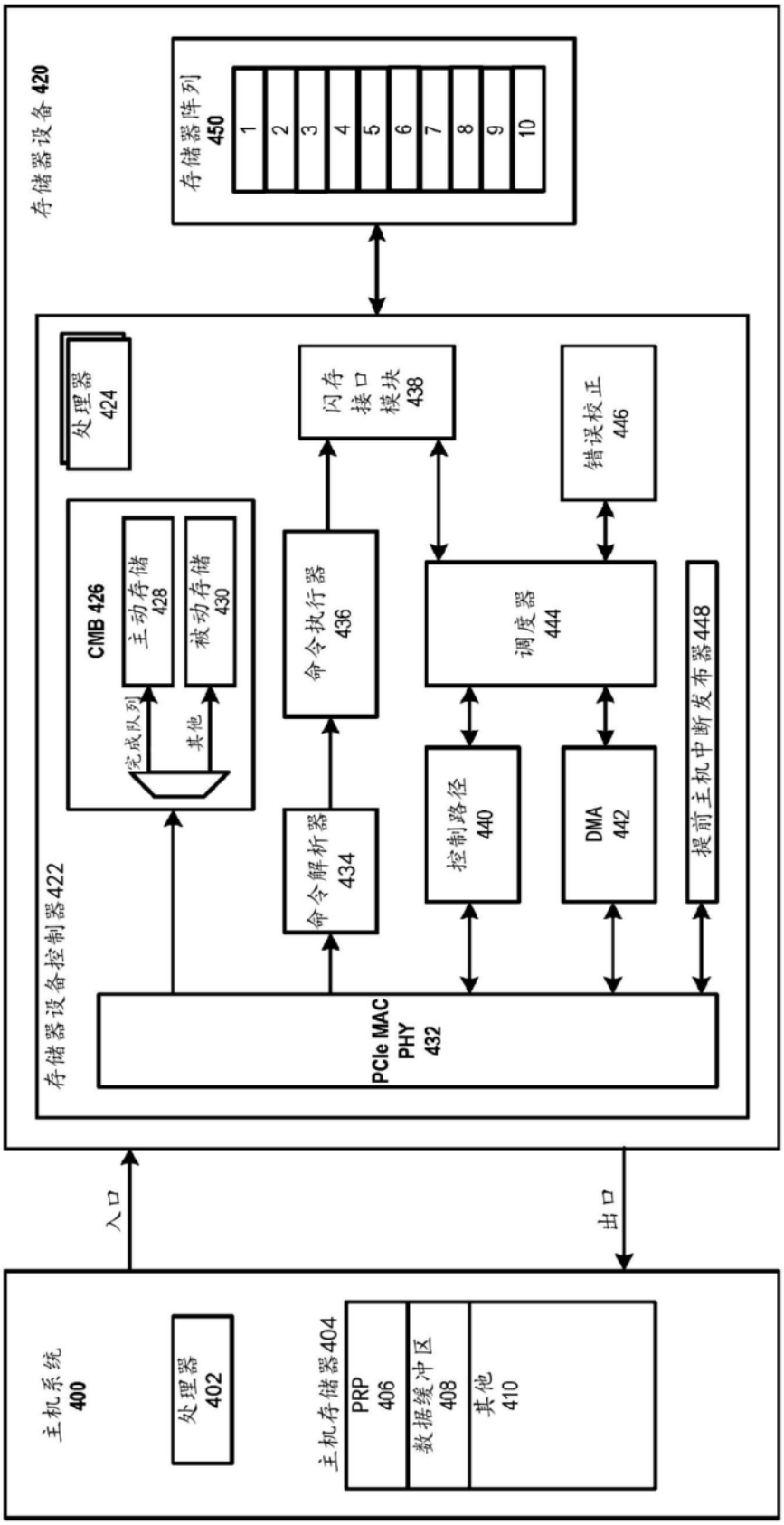

主机系统
400
处理器
402
主机存储器404
PRP
406
数据缓冲区
408
其他
410
入口
出口
存储器设备 420
存储器阵列
450
1
2
3
4
5
6
7
8
9
10
存储器设备控制器422
处理器
424
CMB 426
主动存储
428
被动存储
430
完成队列
其他
命令解析器
434
命令执行器
436
闪存接口模块
438
PCIe MAC
PHY
432
控制路径
440
DMA
442
调度器
444
错误校正
446
提前主机中断发布器448

图4

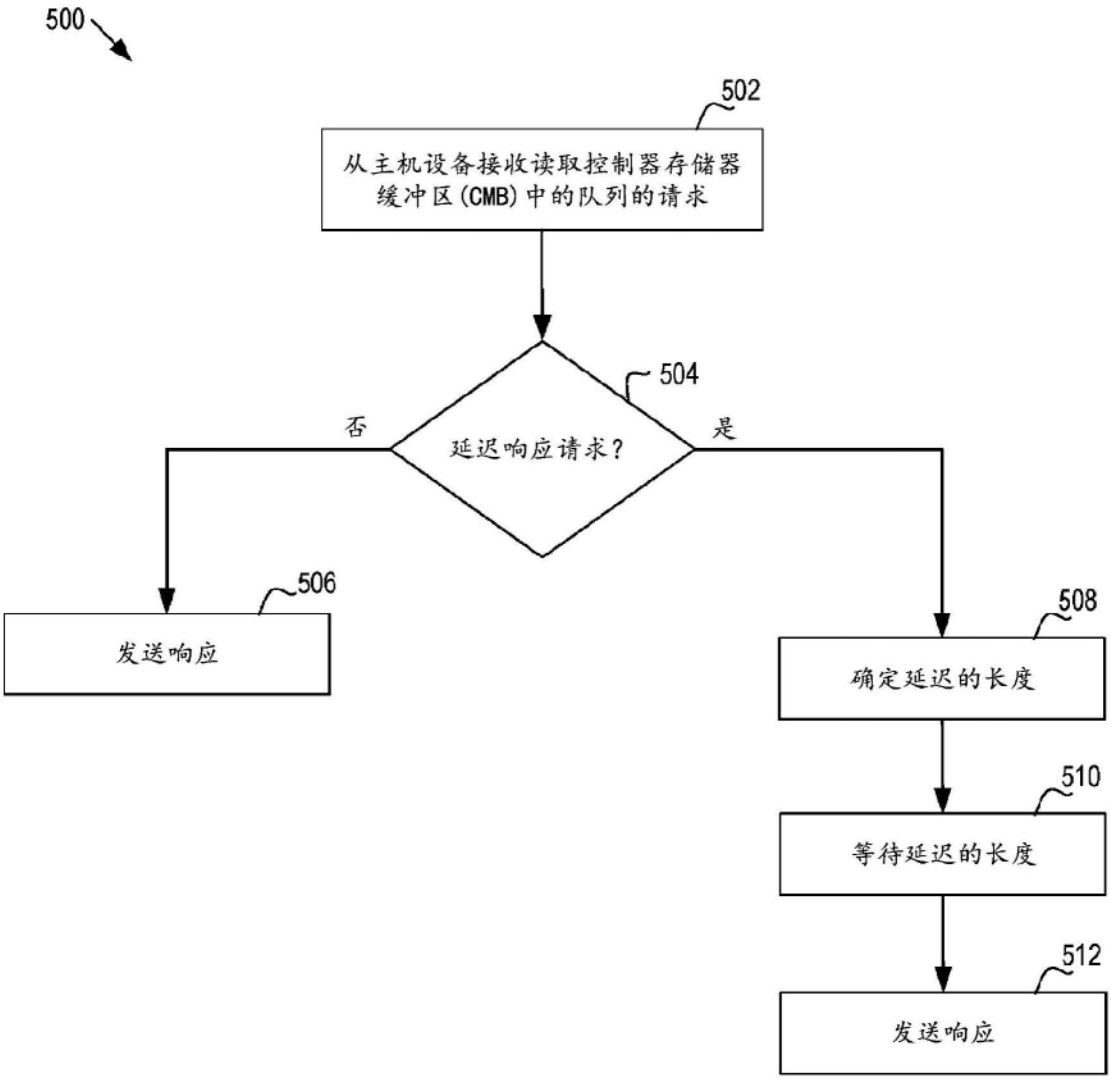

500
502
从主机设备接收读取控制器存储器缓冲区(CMB)中的队列的请求
504
延迟响应请求?
否
是
506
发送响应
508
确定延迟的长度
510
等待延迟的长度
512
发送响应

图5

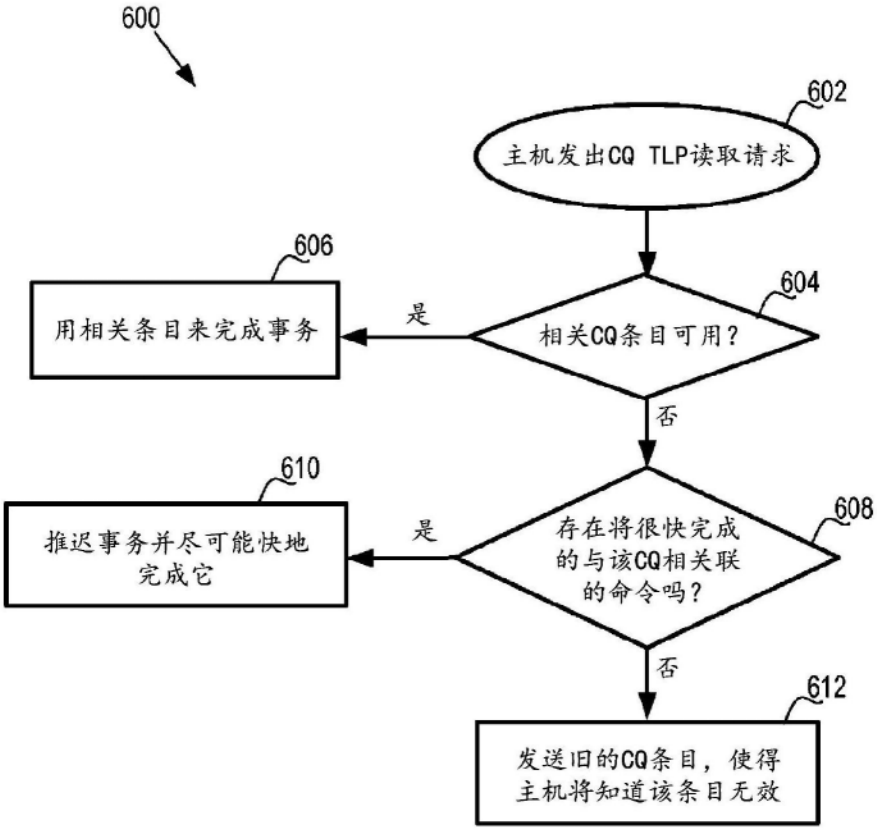

600
602
主机发出CQ TLP读取请求
604
相关CQ条目可用？
是
606
用相关条目来完成事务
否
608
存在将很快完成的与该CQ相关联的命令吗？
是
610
推迟事务并尽可能快地完成它
否
612
发送旧的CQ条目，使得主机将知道该条目无效

图6

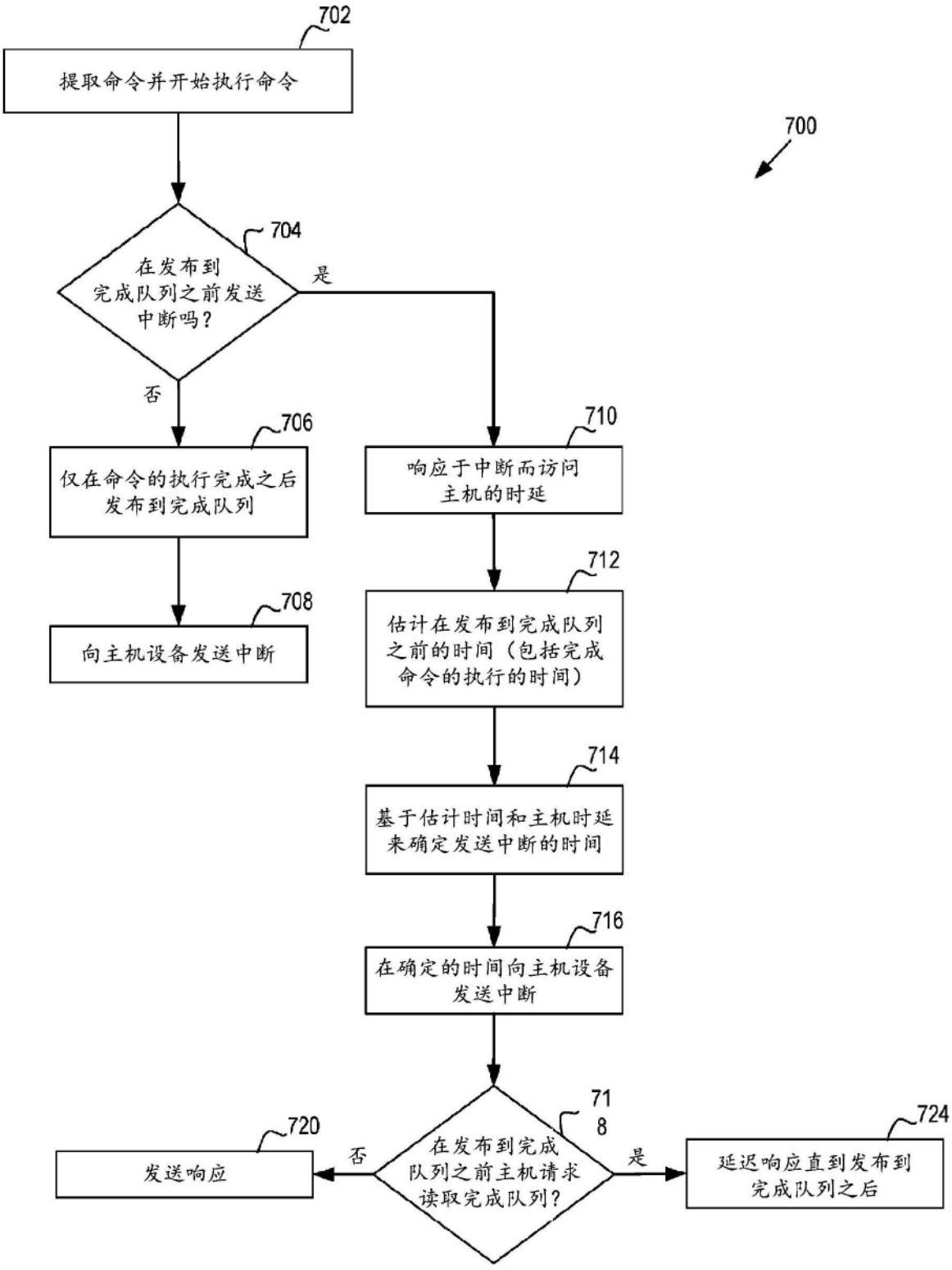

700
702
提取命令并开始执行命令
704
在发布到完成队列之前发送中断吗？
是
否
706
仅在命令的执行完成之后发布到完成队列
708
向主机设备发送中断
710
响应于中断而访问主机的时延
712
估计在发布到完成队列之前的时间（包括完成命令的执行的时间）
714
基于估计时间和主机时延来确定发送中断的时间
716
在确定的时间向主机设备发送中断
718
在发布到完成队列之前主机请求读取完成队列？
否
720
发送响应
是
724
延迟响应直到发布到完成队列之后

图7

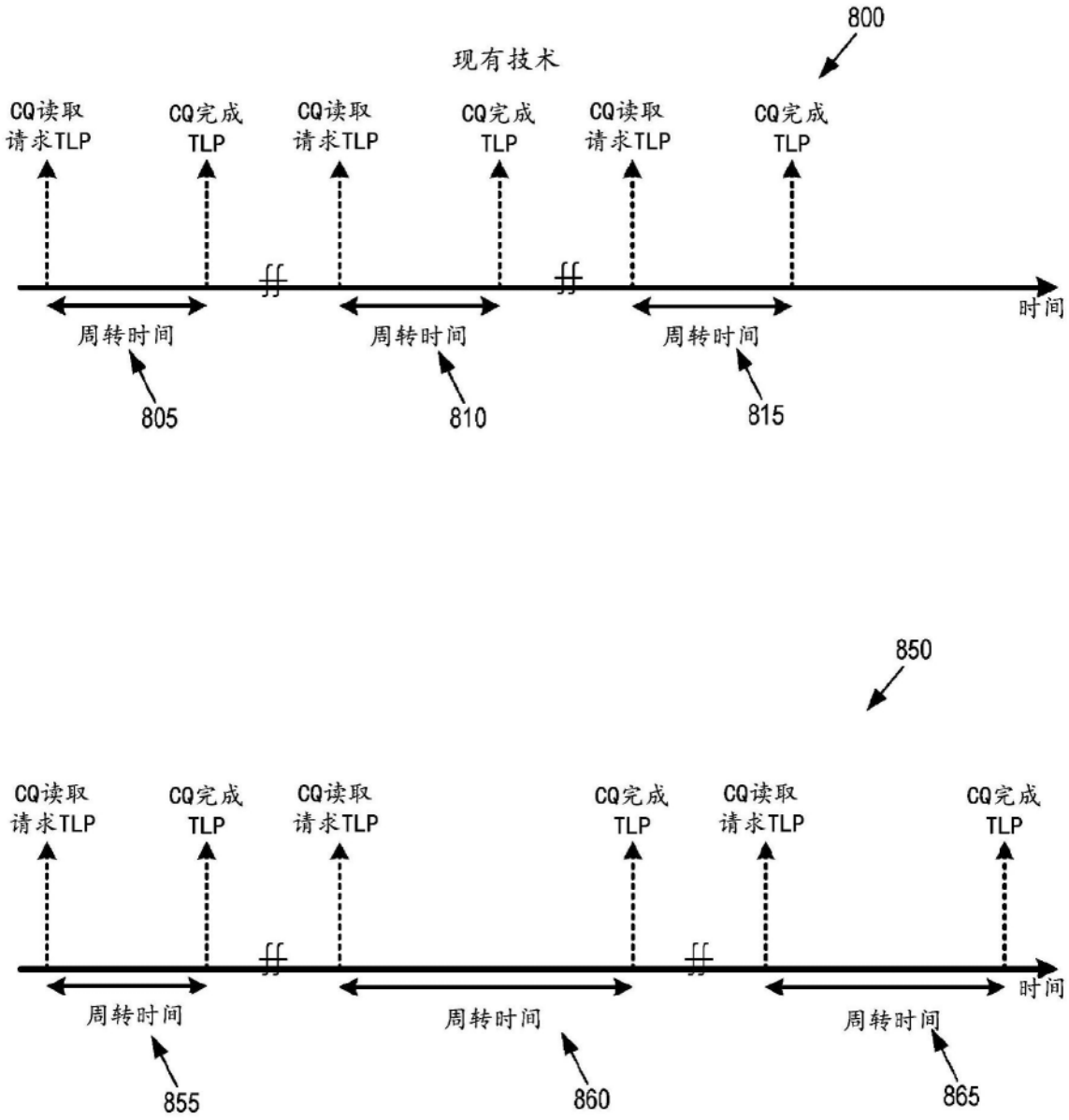

800
现有技术
CQ读取请求TLP
CQ完成TLP
CQ读取请求TLP
CQ完成TLP
CQ读取请求TLP
CQ完成TLP
时间
周转时间
周转时间
周转时间
805
810
815
850
CQ读取请求TLP
CQ完成TLP
CQ读取请求TLP
CQ完成TLP
CQ读取请求TLP
CQ完成TLP
时间
周转时间
周转时间
周转时间
855
860
865

图8

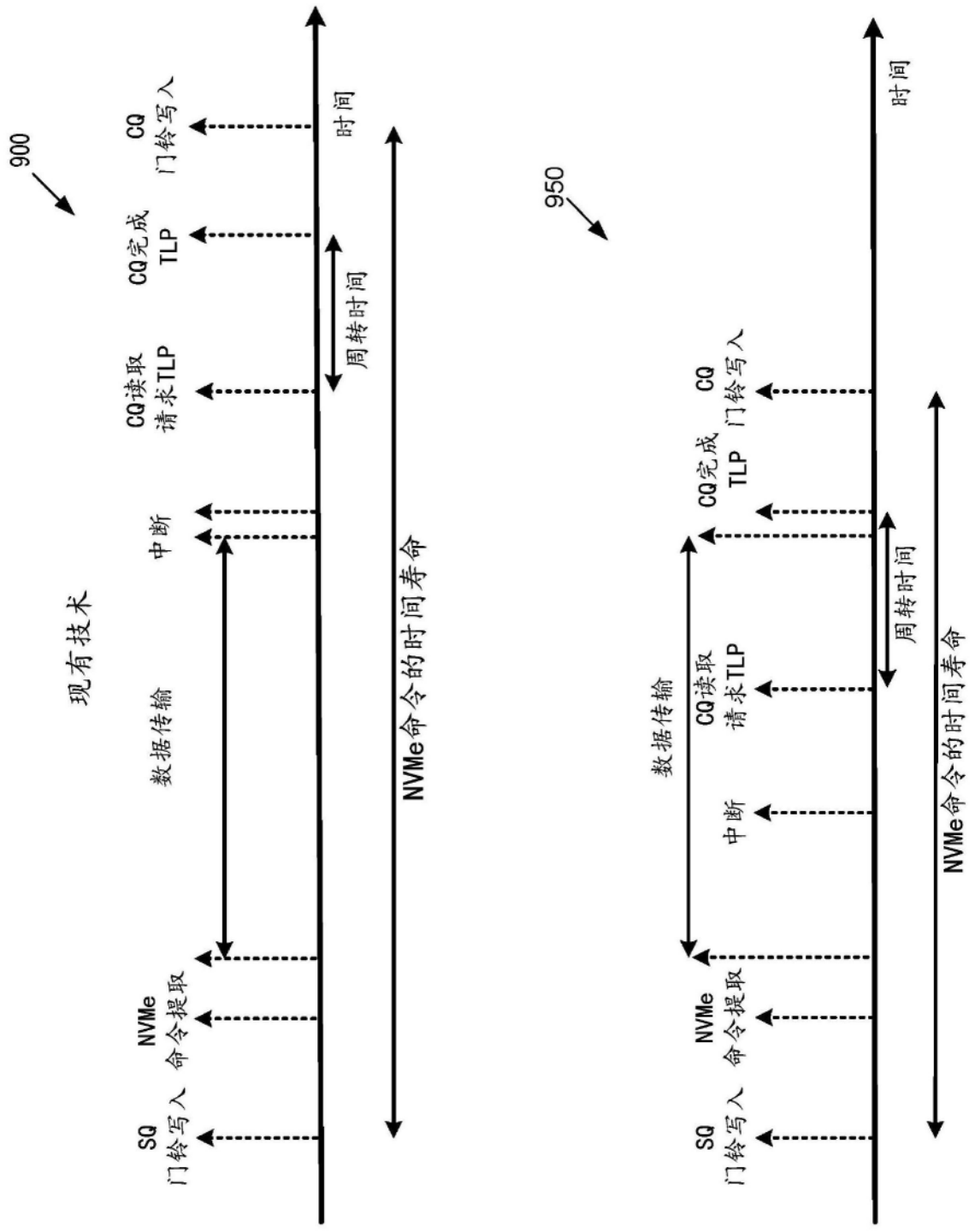

900
现有技术
SQ
门铃写入
NVMe
命令提取
数据传输
中断
CQ读取
请求TLP
CQ完成
TLP
CQ
门铃写入
时间
周转时间
NVMe命令的时间寿命
950
SQ
门铃写入
NVMe
命令提取
数据传输
中断
CQ读取
请求TLP
CQ完成
TLP
CQ
门铃写入
时间
周转时间
NVMe命令的时间寿命

图9

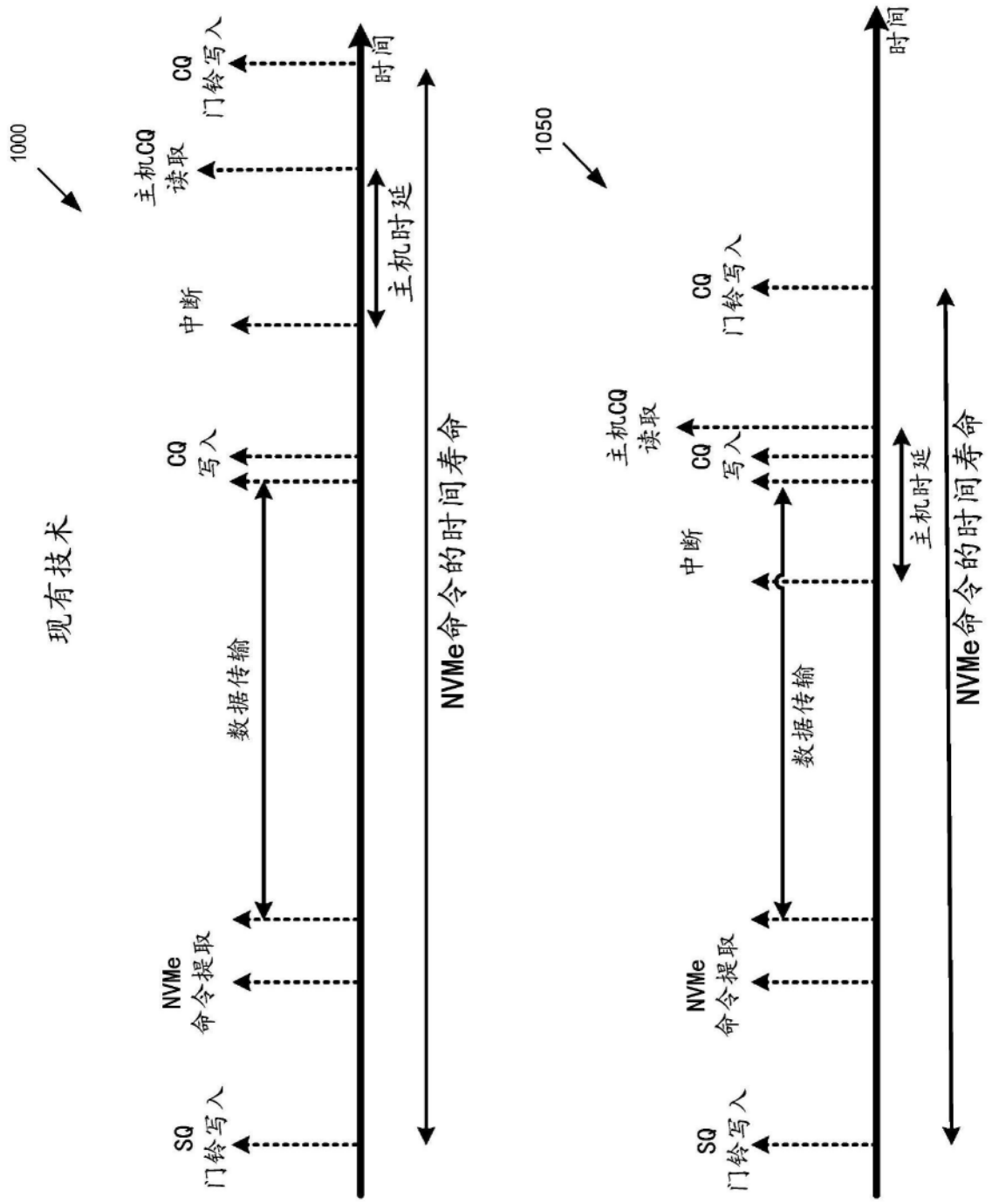

1000
现有技术
SQ
门铃写入
NVMe
命令提取
数据传输
CQ
写入
中断
主机CQ
读取
CQ
门铃写入
时间
主机时延
NVMe命令的时间寿命
1050
SQ
门铃写入
NVMe
命令提取
数据传输
中断
CQ
写入
主机CQ
读取
CQ
门铃写入
时间
主机时延
NVMe命令的时间寿命

图10