(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116057507 A
(43)申请公布日 2023.05.02

(21)申请号 202180056209.8
(22)申请日 2021.08.12
(30)优先权数据
17/003,344 2020.08.26 US
(85)PCT国际申请进入国家阶段日
2023.02.07
(86)PCT国际申请的申请数据
PCT/IB2021/057438 2021.08.12
(87)PCT国际申请的公布数据
WO2022/043812 EN 2022.03.03
(71)申请人 国际商业机器公司
地址 美国纽约阿芒克
(72)发明人 K·帕特尔 S·帕特尔 S·罗伊
(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 陈金林
(51)Int.Cl.
G06F 9/50(2006.01)

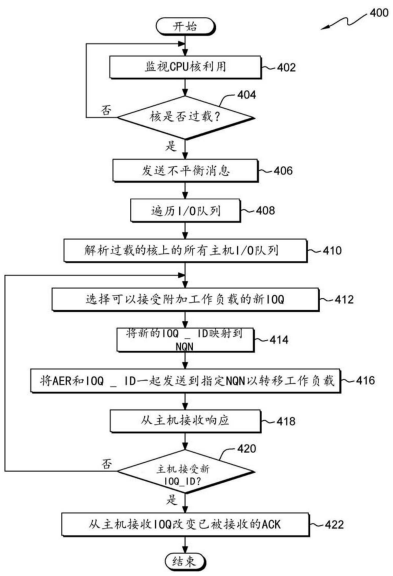
权利要求书5页 说明书13页 附图7页

(54)发明名称

存储级负载平衡

(57)摘要

在存储级负载平衡的方法中,监测存储系统的负载水平,其中负载水平是存储系统中的多个CPU核的利用率百分比。基于一个或多个CPU核的利用率百分比超过阈值来检测过载状况,其中过载状况是由来自访问单个CPU核的多个主机计算机的一个或多个I/O队列的重叠引起的。响应于检测到过载状况,在第二CPU核上选择新I/O队列,其中第二CPU核具有小于第二阈值的利用率百分比。向主机计算机发送推荐,其中该推荐是将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡存储系统的负载水平。



CN 116057507 A

1.一种用于存储级负载平衡的计算机实现的方法,包括以下步骤:

由一个或多个计算机处理器监测存储系统的负载水平,其中所述负载水平是所述存储系统中的多个CPU核的利用率百分比;

由所述一个或多个计算机处理器基于所述多个CPU核中的一个或多个CPU核的利用率百分比超过第一阈值来检测过载状况,其中所述过载状况是由来自访问所述存储系统中的所述多个CPU核中的第一CPU核的多个主机计算机中的每个主机计算机的一个或多个I/O队列的重叠引起的;

响应于检测到所述过载状况,由所述一个或多个计算机处理器在所述存储系统中的所述多个CPU核中的第二CPU核上选择新I/O队列,其中第二CPU核具有小于第二阈值的利用率百分比;以及

由所述一个或多个计算机处理器向所述多个主机计算机中的第一主机计算机发送推荐,其中所述推荐是将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

2.根据权利要求1所述的计算机实现的方法,其中向所述多个主机计算机中的第一主机计算机推荐第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平还包括:

由所述一个或多个计算机处理器从所述多个主机计算机中的一个主机计算机接收响应;以及

响应于所述响应是对所述推荐的拒绝,由所述一个或多个计算机处理器向所述多个主机计算机中的第二主机计算机推荐第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

3.根据权利要求1或权利要求2所述的计算机实现的方法,其中监测所述存储系统的负载水平还包括,使用守护程序来收集所述多个CPU核中的所述一个或多个CPU核的利用率百分比。

4.根据任一前述权利要求所述的计算机实现的方法,其中检测所述过载状况还包括:

由所述一个或多个计算机处理器收集被包含在一个或多个存储系统配置映射和一个或多个存储系统使用表中的CPU核利用率数据,其中所述一个或多个存储系统配置映射包括所述多个CPU核中的每个CPU核的I/O队列配置;以及

由所述一个或多个计算机处理器分析在所述一个或多个存储系统配置映射和所述一个或多个存储系统使用表中收集的所述CPU核利用率数据,以确定所述多个CPU核中的每个CPU核中每个I/O队列的利用率百分比。

5.根据任一前述权利要求所述的计算机实现的方法,其中响应于检测到所述过载状况,选择所述存储系统中的所述多个CPU核中的第二CPU核上的新I/O队列,其中第二CPU核具有小于第二阈值的利用率百分比,还包括,执行所述存储系统上的所述一个或多个I/O队列中的每个I/O队列上的所述工作负载的对称工作负载平衡。

6.一种用于存储级负载平衡的计算机程序产品,所述计算机程序产品包括一个或多个计算机可读存储介质和存储在所述一个或多个计算机可读存储介质上的程序指令,所述程序指令包括用于以下操作的指令:

监测存储系统的负载水平,其中所述负载水平是所述存储系统中的多个CPU核的利用率百分比;

基于所述多个CPU核中的一个或多个CPU核的利用率百分比超过第一阈值来检测过载状况,其中所述过载状况是由来自访问所述存储系统中的所述多个CPU核中的第一CPU核的多个主机计算机中的每个主机计算机的一个或多个I/O队列的重叠引起的;

响应于检测到所述过载状况,在所述存储系统中的所述多个CPU核中的第二CPU核上选择新I/O队列,其中第二CPU核具有小于第二阈值的利用率百分比;以及

向所述多个主机计算机中的第一主机计算机发送推荐,其中所述推荐将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

7.根据权利要求6所述的计算机程序产品,其中用于向所述多个主机计算机中的第一主机计算机推荐第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平的所述程序指令还包括,存储在所述一个或多个计算机可读存储介质上的用于以下操作的以下程序指令中的一个或多个:

从所述多个主机计算机中的所述一个主机计算机接收响应;以及

响应于所述响应是对所述推荐的拒绝,向所述多个主机计算机中的第二主机计算机推荐第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

8.根据权利要求6或权利要求7所述的计算机程序产品,其中用于监测所述存储系统的负载水平的所述程序指令还包括,使用守护程序来收集所述多个CPU核中的所述一个或多个CPU核的利用率百分比。

9.根据权利要求6所述的计算机程序产品,其中用于检测所述过载状况的所述程序指令还包括存储在所述一个或多个计算机可读存储介质上的用于以下操作的以下程序指令中的一个或多个:

收集被包含在一个或多个存储系统配置映射和一个或多个存储系统使用表中的CPU核利用率数据,其中所述一个或多个存储系统配置映射包括所述多个CPU核中的每个CPU核的I/O队列配置;以及

分析在所述一个或多个存储系统配置映射和所述一个或多个存储系统使用表中收集的所述CPU核利用率数据,以确定所述多个CPU核中的每个CPU核中的每个I/O队列的利用率百分比。

10.根据权利要求6所述的计算机程序产品,还包括程序指令,响应于检测到所述过载状况,选择所述存储系统中的所述多个CPU核中的第二CPU核上的新I/O队列的程序指令,其中第二CPU核具有小于第二阈值的利用率百分比,还包括,执行所述存储系统上的所述一个或多个I/O队列中的每个I/O队列上的所述工作负载的对称工作负载平衡。

11.一种用于存储级负载平衡的计算机系统,所述计算机系统包括:

一个或多个计算机处理器;

一个或多个计算机可读存储介质;以及

存储在所述一个或多个计算机可读存储介质上以供所述一个或多个计算机处理器中的至少一个计算机处理器执行的程序指令,所存储的程序指令包括用于以下操作的指令:

监测存储系统的负载水平,其中所述负载水平是所述存储系统中的多个CPU核的利用率百分比;

基于所述多个CPU核中的一个或多个CPU核的利用率百分比超过第一阈值来检测过载状况,其中所述过载状况是由来自访问所述存储系统中的所述多个CPU核中的第一CPU核的

多个主机计算机中的每个主机计算机的一个或多个I/O队列的重叠引起的；

响应于检测到所述过载状况，在所述存储系统中的所述多个CPU核中的第二CPU核上选择新I/O队列，其中第二CPU核具有小于第二阈值的利用率百分比；以及

向所述多个主机计算机中的第一主机计算机发送推荐，其中所述推荐将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

12.根据权利要求11所述的计算机系统，其中向所述多个主机计算机中的第一主机计算机发送所述推荐，其中所述推荐是将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平，还包括，存储在所述一个或多个计算机可读存储介质上的用于以下操作的以下程序指令中的一个或多个：

从所述多个主机计算机中的一个主机计算机接收响应；以及

响应于所述响应是对所述推荐的拒绝，向所述多个主机计算机中的第二主机计算机推荐第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

13.根据权利要求11所述的计算机系统，其中监测所述存储系统的负载水平，其中所述负载水平是所述存储系统中的所述多个CPU核的利用率百分比，还包括，使用守护程序来收集所述多个CPU核中的所述一个或多个CPU核的利用率百分比。

14.根据权利要求11所述的计算机系统，其中基于所述多个CPU核中的一个或多个CPU核的利用率百分比超过第一阈值来检测所述过载状况，其中所述过载状况是由来自访问所述存储系统中的所述多个CPU核中的第一CPU核的多个主机计算机中的每个主机计算机的一个或多个I/O队列的重叠引起的，还包括，存储在所述一个或多个计算机可读存储介质上的用于以下操作的以下程序指令中的一个或多个：

收集被包含在一个或多个存储系统配置映射和一个或多个存储系统使用表中的CPU核利用率数据，其中所述一个或多个存储系统配置映射包括所述多个CPU核中的每个CPU核的I/O队列配置；以及

分析在所述一个或多个存储系统配置映射和所述一个或多个存储系统使用表中收集的所述CPU核利用率数据，以确定所述多个CPU核中的每个CPU核中的每个I/O队列的利用率百分比。

15.根据权利要求11所述的计算机系统，响应于检测到所述过载状况，选择所述存储系统中的所述多个CPU核中的第二CPU核上的新I/O队列，其中第二CPU核具有小于第二阈值的利用率百分比，还包括，执行所述存储系统上的所述一个或多个I/O队列中的每个I/O队列上的工作负载的对称工作负载平衡。

16.一种用于存储级负载平衡的计算机实现的方法，包括以下步骤：

响应于从主机计算机接收到建立I/O队列对的命令，由一个或多个计算机处理器分配处理器资源和存储系统中的存储器资源，其中所述存储系统实现结构上的非易失性存储器快速(NVMe-oF)架构；

由所述一个或多个计算机处理器检测所述存储系统中的多个CPU核中的第一CPU核上的过载状况，其中所述过载状况是使用相同I/O队列对的多个主机计算机的重叠；以及

响应于检测到所述过载状况，由所述一个或多个计算机处理器向所述多个主机计算机中的第一主机计算机发送推荐，其中所述推荐将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

17.根据权利要求16所述的计算机实现的方法,其中响应于检测到所述过载状况,由所述一个或多个计算机处理器将所述推荐发送到所述多个主机计算机中的第一主机计算机,其中所述推荐将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平,还包括:

由所述一个或多个计算机处理器从所述多个主机计算机中的一个主机计算机接收响应;以及

响应于所述响应是对所述推荐的拒绝,由所述一个或多个计算机处理器向所述多个主机计算机中的第二主机计算机推荐第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

18.根据权利要求16或权利要求17所述的计算机实现的方法,其中由所述一个或多个计算机处理器检测所述存储系统中的所述多个CPU核中的第一CPU核上的所述过载状况,其中所述过载状况是使用相同I/O队列对的多个主机计算机的重叠,还包括:

由所述一个或多个计算机处理器收集被包含在一个或多个存储系统配置映射和一个或多个存储系统使用表中的CPU核利用率数据,其中所述一个或多个存储系统配置映射包括所述多个CPU核中的每个CPU核的I/O队列配置;以及

由所述一个或多个计算机处理器分析在所述一个或多个存储系统配置映射和所述一个或多个存储系统使用表中收集的所述CPU核利用率数据,以确定所述多个CPU核中的每个CPU核中的每个I/O队列的利用率百分比。

19.根据权利要求18所述的计算机实现的方法,其中所述过载状况基于所述存储系统中的所述多个CPU核中的一个或多个CPU核的利用率百分比超过第一阈值。

20.根据权利要求16所述的计算机实现的方法,其中响应于检测到所述过载状况,由所述一个或多个计算机处理器将所述推荐发送到所述多个主机计算机中的第一主机计算机,其中所述推荐是将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平,还包括,基于所述存储系统中的所述多个CPU核中的第二CPU核具有小于第二阈值的利用率百分比,选择所述存储系统中的所述多个CPU核中的第二CPU核上的新I/O队列。

21.一种用于存储级负载平衡的计算机系统,所述计算机系统包括:

一个或多个计算机处理器;

一个或多个计算机可读存储介质;以及

存储在所述一个或多个计算机可读存储介质上以供所述一个或多个计算机处理器中的至少一个计算机处理器执行的程序指令,所存储的程序指令包括用于以下操作的指令:

响应于从主机计算机接收到建立I/O队列对的命令,分配处理器资源和存储系统中的存储器资源,其中所述存储系统实现结构上的非易失性存储器快速(NVMe-oF)架构;

检测所述存储系统中的多个CPU核中的第一CPU核上的过载状况,其中所述过载状况是使用相同I/O队列对的多个主机计算机的重叠;以及

响应于检测到所述过载状况,向所述多个主机计算机中的第一主机计算机发送推荐,其中所述推荐将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

22.根据权利要求21所述的计算机系统,其中响应于检测到所述过载状况,将所述推荐

发送到所述多个主机计算机中的第一主机计算机，其中所述推荐将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平，还包括，存储在所述一个或多个计算机可读存储介质上的用于执行以下操作的以下程序指令中的一个或多个：

从所述多个主机计算机中的所述一个主机计算机接收响应；以及

响应于所述响应是对所述推荐的拒绝，向所述多个主机计算机中的第二主机计算机推荐第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平。

23.根据权利要求21所述的计算机系统，其中检测所述存储系统中的多个CPU核中的第一CPU核上的所述过载状况，其中所述过载状况是使用相同I/O队列对的多个主机计算机的重叠，还包括，存储在所述一个或多个计算机可读存储介质上的用于执行操作的以下程序指令中的一个或多个：

收集被包含在一个或多个存储系统配置映射和一个或多个存储系统使用表中的CPU核利用率数据，其中所述一个或多个存储系统配置映射包括所述多个CPU核中的每个CPU核的I/O队列配置；以及

分析在所述一个或多个存储系统配置映射和所述一个或多个存储系统使用表中收集的所述CPU核利用率数据，以确定所述多个CPU核中的每个I/O队列的利用率百分比。

24.根据权利要求23所述的计算机系统，其中所述过载状况基于所述存储系统中的所述多个CPU核中的一个或多个CPU核的利用率百分比超过第一阈值。

25.根据权利要求21所述的计算机系统，响应于检测到所述过载状况，将所述推荐发送到所述多个主机计算机中的第一主机计算机，其中所述推荐是将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡所述存储系统的负载水平，还包括，存储在所述一个或多个计算机可读存储介质上的用于以下操作的以下程序指令中的一个或多个：基于所述存储系统中的所述多个CPU核中的第二CPU核具有小于第二阈值的利用率百分比，选择所述存储系统中的所述多个CPU核中的第二CPU核上的新I/O队列。

存储级负载平衡

技术领域

[0001] 本发明一般涉及计算机存储领域，尤其涉及存储级负载平衡。

背景技术

[0002] 非易失性存储器快速(NVMe™)是优化的、高性能的可扩展主机控制器接口，其被设计为基于快速外围组件互连(PCIe®)接口解决利用固态存储的计算机存储系统的需求。从非易失性存储器技术的发展开始设计，NVMe被设计为提供对利用非易失性存储器构建的存储设备的高效访问，从当前的NAND闪存技术到未来的更高性能的永久性存储器技术。

[0003] NVMe协议利用到底层介质的并行、低延迟数据路径，类似于高性能处理器架构。与传统存储接口(例如串行附接SCSI(SAS)和串行高级技术附接(SATA)协议)相比，这提供了明显更高的性能和更低的延迟。NVMe可以支持多达65,535个输入/输出(I/O)队列，其中每个队列具有65,535个条目。传统SAS和SATA接口只能支持单个队列，其中每个SAS队列具有254个条目，并且每个SATA队列仅具有32个条目。NVMe主机软件可以根据系统配置和预期工作负载创建队列，直到NVMe控制器所允许的最大值。NVMe支持分散/聚集I/O，最小化数据传送上的CPU开销，并且甚至提供基于工作负载要求改变它们的优先级的能力。

[0004] 结构上的NVMe(NVMe-oF)是用于通过网络(aka结构)在主机和存储系统之间通信的网络协议。NVMe-oF定义了通过存储联网结构支持用于NVMe块存储协议的一系列存储联网结构的公共架构。这包括启用到存储系统中的前侧接口、扩展到大量NVMe设备、以及扩展可以访问NVMe设备和NVMe子系统的距离。

发明内容

[0005] 本发明的一个方面包括一种用于存储级负载平衡的计算机实现的方法。在第一实施例中，监测存储系统的负载水平，其中负载水平是存储系统中的多个CPU核的利用率百分比。基于一个或多个CPU核的利用率百分比超过阈值来检测过载状况，其中过载状况是由来自访问存储系统中的单个CPU核的多个主机计算机的一个或多个I/O队列的重叠引起的。响应于检测到过载状况，在存储系统中的第二CPU核上选择新I/O队列(IOQ)，其中第二CPU核具有小于第二阈值的利用率百分比。向主机计算机发送推荐，其中该推荐是将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡存储系统的负载水平。

[0006] 本发明的另一方面包括用于存储级负载平衡的计算机实现的方法。在第二实施例中，响应于从主机计算机接收到建立I/O队列对的命令，在存储系统中分配处理器和存储器资源，其中存储系统实现结构上的非易失性存储器快速(NVMe-oF)架构。在存储系统中的CPU核上检测过载状况，其中过载状况是使用相同I/O队列对的多个主机计算机的重叠。响应于检测到过载状况，向主机计算机发送推荐，其中该推荐将I/O流量从第一CPU核移动到第二CPU核上的新I/O队列以重新平衡存储系统的负载水平。

附图说明

[0007] 图1是示出根据本发明实施例的分布式数据处理环境的功能框图。

[0008] 图2是根据本发明实施例的用于存储级负载平衡的、在图1的分布式数据处理环境内的计算设备上的存储系统的NVMe CPU核到队列映射的示例。

[0009] 图3a是根据本发明的实施例的用于存储级负载平衡的、在图1的分布式数据处理环境内的计算设备上的非平衡存储系统的示例。

[0010] 图3b是根据本发明的实施例的用于存储级负载平衡的、在图1的分布式数据处理环境内的计算设备上结合本发明的存储系统的示例。

[0011] 图3c是根据本发明的实施例的用于存储级负载平衡的、在图1的分布式数据处理环境内的计算设备上结合本发明的经平衡的存储系统的示例。

[0012] 图4是根据本发明实施例的在图1的计算机系统内的队列平衡程序的步骤的流程图。

[0013] 图5示出了根据本发明实施例的在图1的分布式数据处理环境内执行队列平衡程序的计算设备的组件的框图。

具体实施方式

[0014] 随着现代数据处理系统中数据的量和使用的爆炸性增长,需要新的方法来增加现代系统中主机和存储设备之间的数据传送的吞吐量并减少延迟。在典型的系统中,在一个存储系统中存在多个共存的传输信道和协议,其可以包括NVMe远程直接存储器访问(NVMe-RDMA)、光纤信道NVMe(NVMe-FC)、光纤信道到小型计算机系统接口(FC-SCSI)、以太网光纤信道(FCoE)、因特网小型计算机系统接口(iSCSI)等。

[0015] NVMe是存储协议,其被设计用于在通常使用PCIe总线作为传输机制的服务器、存储设备和闪存控制器之间的更快的数据传送。NVMe规范提供了实现高性能I/O的寄存器接口和命令集。NVMe是用于主机与存储系统之间的数据传输的传统小型计算机系统接口(SCSI)标准(以及像SAS、SATA等的其它标准)的替代。基于NVMe的PCIe闪存相比于基于SAS和基于SATA的SSD的主要优点之一是减少了主机软件栈中的访问延迟,导致了更高的每秒输入/输出操作(IOP)和更低的CPU利用率。

[0016] NVMe支持具有多核处理器的并行I/O处理,这导致更快的I/O分派,这导致I/O延迟的减少。由于多个CPU核同时处理I/O请求,所以系统性能由于CPU资源的最优利用而提高。另外,NVMe被设计为每I/O使用更少的CPU指令。NVMe还支持单个消息队列中的64,000个命令以及最多65,535个I/O队列。

[0017] 结构上的NVMe(NVMe-oF)是本地PCIe NVMe的扩展,其允许NVMe提供的高性能和低延迟的益处,但是跨网络结构而不是本地连接。服务器和存储设备可以通过以太网或光纤信道(FC)连接,其都支持结构上的NVMe命令,并且将NVMe协议的优点扩展到互连的系统组件。NVMe-oF的设计目标是除了与访问PCIe NVMe存储设备相关联的延迟之外,增加不超过10微秒的延迟以用于NVMe主机计算机和网络连接的NVMe存储设备之间的通信。

[0018] NVMe-oF支持用于从主机到存储系统的常规I/O操作的多个I/O队列。NVMe支持最多65,535个队列,每个队列中具有多达65,535个条目。一旦建立了连接,主机驱动器的责任就是创建队列。一旦主机连接到目标系统,就创建称为管理队列的专用队列。顾名思义,管

理队列用于将控制命令从发起者传送到目标设备。一旦创建管理队列，主机就使用它来基于系统要求创建I/O队列。主机可以建立到具有相同NQN(NVMe资格名称，其用于标识远程NVMe存储目标)的单个控制器的多个I/O队列，并且具有映射到它的多个名称空间(或卷)。一旦建立了I/O队列，就将I/O命令提交给I/O提交队列(SQ)，并从完成队列(CQ)中收集I/O响应。可以使用经由用于该会话的管理队列发送的控制指令来添加或移除这些I/O队列。

[0019] 当接收到用于I/O队列创建的命令时，目标设备执行初始系统检查以寻找最大支持队列和其他相关字段、创建I/O队列、并将该I/O队列分配给存储控制器上的CPU核。接下来，目标设备经由管理完成队列发送对队列创建请求的响应。每个I/O队列被分配给存储控制器上的不同CPU核。这允许并行性并提高系统的吞吐量。在目标存储控制器处实现核分配逻辑，并且基于存储控制器处的预定义策略来执行I/O队列到CPU核的映射。

[0020] 现有技术中的问题是由于队列重叠而导致的性能下降。NVMe可以支持大约65,535个队列，其可以被分配给不同的CPU核以实现并行性。当主机发出与存储系统建立I/O队列对的命令时，存储系统将处理器资源和存储器资源分配给I/O队列对。例如，考虑两个或更多个主机建立到公共NVMe目标的连接的情况。由多个主机创建的I/O队列可能在各个CPU核上开始重叠，即，核1上的主机“A”主I/O队列可能与核1上的主机“B”主I/O队列重叠。在这种情况下，来自两个主机的I/O队列对在NVMe上发送的I/O工作负载由存储控制器处的单个核服务。这降低了存储控制器端的并行性，并影响主机应用I/O性能。在现有技术中，没有办法将CPU核分配与预期的工作负载联系起来，并且这可能导致在存储控制器节点上可用的CPU核之间的显著的I/O负载不平衡。由于每个CPU核在多个I/O队列之间共享，因此没有办法检测由于来自一个或多个主机的重叠队列而导致的工作负载不平衡或者向服务器通知工作负载不平衡。在多个主机经由NVMe队列连接到存储目标的情况下，由于不相似的主机I/O工作负载，CPU核中的一些可能过载，而一些可能负载不足。此外，不存在存储系统可以用来预测在I/O队列创建时每个队列将产生多少负载的机制。在主机多路径驱动器处，主机将使用某一I/O队列作为主队列。在多个主机具有附接到同一CPU核的主队列的情况下，该CPU核变得过载，并且访问数据的应用将经历增加的I/O延迟，并且因此将不会获得并行性的益处。

[0021] 作为I/O队列重叠的结果，IOP可能由于跨CPU核的不平衡负载而减少。在主机执行小的I/O密集型工作负载的情况下，由于重叠的队列而导致的这种开销的严重性变得更糟，并且可能导致在峰值工作负载期间的应用减速以及意外的I/O延迟问题。这也在存储控制器处产生性能问题，因为跨存储控制器系统的不平衡的CPU核在一些CPU核空闲时增加了对其它CPU核的附加负担，从而减少了并行处理并增加了总延迟和延迟。

[0022] 在各种实施例中，本发明通过检测NVMe存储控制器内的CPU核分配中的重叠I/O队列，并且将I/O队列的分配重新平衡到CPU核来解决这个问题。在一个实施例中，队列平衡程序监测在所有可用CPU核上建立的队列、工作负载和CPU核可用性。一旦遇到队列重叠的情况，队列平衡程序将确定CPU工作负载和负载不平衡。队列平衡程序识别连接到CPU核的I/O队列，并且针对具有高带宽利用率的IOP工作负载分析I/O队列。由于IOP工作负载是CPU敏感的，所以队列平衡程序收集该信息，并且映射附接到过载CPU核的每个I/O队列的CPU消耗。在一个实施例中，队列平衡程序遍历从相同主机创建的所有I/O队列，并且还分析它们的工作负载。

[0023] 在一个实施例中，队列平衡程序基于收集的工作负载信息确定哪个I/O队列工作

负载可以被增加以获得更好的性能。队列平衡程序通过执行存储系统上的I/O队列工作负载的对称工作负载平衡来实现这一点。

[0024] 在一个实施例中，一旦队列平衡程序做出新的I/O工作负载转移决定，该信息作为信号被发送到NVMe控制器的管理控制单元，并且将队列重叠情况的异步通知发送到主机。该高级错误报告(AER)消息包含存储系统期望将流量移动到的I/O队列ID(IOQ_ID)，以平衡CPU工作负载。

[0025] 一旦信号被发送到主机，主机NVMe驱动器将决定是继续当前I/O发送策略还是采用来自队列平衡程序的建议以对某个IOQ进行优先级排序。如果主机决定采用来自队列平衡程序的建议，则IOQ路由策略由主机侧的NVMe驱动器来调整。在一些情况下，如果主机能够容忍性能下降，或者主机能够容忍IOP的总体降低，或者如果主机由于任何其他原因不想改变IOQ策略，则拒绝该建议，并且向队列平衡程序发送信号，通知其拒绝。在一个实施例中，一旦队列平衡程序接收到拒绝信号，队列平衡程序就向另一主机发送AER消息以将其I/O工作负载移出过载的CPU核。这样，队列平衡程序和主机将都是决定的参与方，并且通过向第二主机发信号将适度地实现工作负载平衡。

[0026] 本发明的优点包括减少的队列重叠瓶颈、更好的性能、IOP的增加、避免IOQ的重新创建、以及跨CPU核的改进的负载平衡。

[0027] 本发明减少了队列重叠瓶颈，因为主机IOQ偏好改变了，从而减少或消除了CPU核不平衡。

[0028] 本发明导致更好的性能，因为在主机同时执行I/O时的队列重叠情况下，性能将随着核服务一次一个地每个队列而降低。但是当两个队列属于不同的主机时，本发明重新平衡I/O队列以避免整体性能下降。

[0029] 本发明导致IOP的增加，因为避免了队列重叠的情况，并且因此减少了主机I/O周转时间，这增加了总的IOP。

[0030] 本发明避免了IOQ的重新创建，因为它不将IOQ与存储系统或主机断开连接，并且仅指示主机NVMe驱动器在运行中改变目标，从而平衡存储级工作负载并且透明地创建性能增益。

[0031] 本发明导致跨CPU核的改进的负载平衡，因为对于跨存储系统中的所有CPU核的负载实现了更大的平衡，并且因此存储系统更平衡。

[0032] 图1是示出了根据本发明的至少一个实施例的适合于队列平衡程序112的操作的分布式数据处理环境(一般地表示为100)的功能框图。本文使用的术语“分布式”描述了包括多个物理上不同的设备的计算机系统，这些设备作为单个计算机系统一起操作。图1仅提供了一种实施方式的图示，并且不暗示对其中可实现不同实施例的环境的任何限制。本领域技术人员可以对所描述的环境进行许多修改，而不偏离权利要求所陈述的本发明的范围。

[0033] 在各种实施例中，分布式数据处理环境100包括多个主机计算机。在图1中描绘的实施例中，分布式数据处理环境100包括主机130、主机132和主机134，它们都连接到网络120。网络120可以是例如电信网络、局域网(LAN)、诸如因特网的广域网(WAN)、或这三者的组合，并且可以包括有线、无线、或光纤连接。网络120可以包括一个或多个有线和/或无线网络，其能够接收和发送数据、语音和/或视频信号，包括多媒体信号，该多媒体信号包括语

音、数据和视频信息。通常，网络120可以是支持主机130、主机132、主机134和分布式数据处理环境100内的其它计算设备(未示出)之间的通信的连接和协议的任何组合。

[0034] 在各实施例中，主机130、主机132和主机134各自可以是独立计算设备、管理服务器、web服务器、移动计算设备、或能够接收、发送和处理数据的任何其他电子设备或计算系统。在一个实施例中，主机130、主机132和主机134中的每一个可以是个人计算机、台式计算机、膝上型计算机、上网本计算机、平板计算机、智能电话或能够经由网络120与分布式数据处理环境100内的其他计算设备(未示出)通信的任何可编程电子设备。在另一实施例中，主机130、主机132和主机134可各自表示诸如在云计算环境中利用多个计算机作为服务器系统的服务器计算系统。在又一个实施例中，主机130、主机132和主机134各自表示利用集群的计算机和组件(例如，数据库服务器计算机、应用服务器计算机等)的计算系统，当在分布式数据处理环境100内被访问时，该集群的计算机和组件充当单个无缝资源池。

[0035] 在各种实施例中，分布式数据处理环境100还包括存储系统110，其经由结构140连接到主机130、主机132和主机134。例如，结构140可以是以太网结构、光纤信道结构、以太网光纤信道(FCoE)或InfiniBand结构。在另一实施例中，结构140可以包括任何RDMA技术，包括InfiniBand、融合以太网的RDMA(RoCE)和iWARP。在其他实施例中，结构140可以是本领域技术人员公知的能够将主机与存储系统以接口的方式连接的任何结构。

[0036] 在各种实施例中，存储系统110可以是独立的计算设备、管理服务器、web服务器、移动计算设备或能够接收、发送和处理数据的任何其他电子设备或计算系统。在一些实施例中，存储系统110可以经由结构140连接到网络120。

[0037] 在一个实施例中，存储系统110包括队列平衡程序112。在一个实施例中，队列平衡程序112是用于通过驱动器类型智能地选择跨协议的传输信道的较大程序的程序、应用或子程序。

[0038] 在一个实施例中，存储系统110包括信息储存库114。在一个实施例中，信息储存库114可由队列平衡程序112管理。在一个备选实施例中，信息储存库114可由存储系统110的操作系统单独或与队列平衡程序112一起管理。信息储存库114是可以存储、收集、比较和/或组合信息的数据储存库。在一些实施例中，信息储存库114位于存储系统110外部，并且通过诸如结构140之类的通信网络来访问。在一些实施例中，信息储存库114被存储在存储系统110上。在一些实施例中，信息储存库114可以驻留在另一计算设备(未示出)上，只要信息储存库114可由存储系统110访问。信息储存库114可包括传输信道和协议数据、协议类数据、驱动器类型和驱动器层数据、链路连接数据、传输信道表、要在主机发起方和目标存储系统之间传送的原始数据、由队列平衡程序112从一个或多个源接收的其它数据、以及由队列平衡程序112创建的数据。

[0039] 如本领域所公知的，信息储存库114可以使用用于存储信息的任何易失性或非易失性存储介质来实现。例如，信息储存库114可以用磁带库、光库、一个或多个独立硬盘驱动器、独立磁盘冗余阵列(RAID)中的多个硬盘驱动器、SATA驱动器、固态驱动器(SSD)或随机存取存储器(RAM)来实现。类似地，信息储存库114可以用本领域已知的任何合适的存储架构来实现，诸如关系数据库、面向对象的数据库、或一个或多个表。

[0040] 图2是根据本发明实施例的将I/O队列映射到基本NVMe存储系统中的CPU核的示例。在一个实施例中，存储系统200是图1中的存储系统110的队列映射的一个可能配置的示

例,在一个实施例中,存储系统200中的处理器具有控制器管理核210。在一个实施例中,一旦主机连接到目标系统,就在关联时创建专用队列,称为管理队列。管理队列用于将控制命令从发起者传送到目标设备。在一个实施例中,控制器管理核210中的管理队列包括管理提交队列,用于向I/O队列提交I/O请求,以及管理完成队列,用于从I/O队列接收完成消息。

[0041] 在典型的存储系统中,存在一个或多个CPU,每个CPU具有多个CPU核。在图2所示的示例中,存储系统200中的处理器具有n个核,本文描绘为Core_0 212、Core_1 214到Core_n-1 216。在本发明的一些实施例中,每个CPU核具有用于向I/O队列提交请求的I/O提交队列,以及用于从I/O队列接收完成消息的I/O完成队列。图2的示例存储系统还包括控制器220,其是根据本发明的一些实施例的存储系统的控制器。

[0042] 应注意,图2中描绘的示例仅示出了仅单个I/O队列被分配给每个CPU核。在本发明的典型实施例中,多个I/O队列被分配给每个CPU核。这个更典型的实施例在以下图3a-图3c中示出。

[0043] 图3a是一般地表示为300的典型存储配置的图示,并且描绘了来自上面的问题陈述的示例。在该示例中,主机A 310和主机B 312是图1的分布式数据处理环境100中的主机(130-134)的示例。结构320是允许在任意数量的主机设备和存储系统之间通信的结构。上面列出了在本发明的各种实施例中可以构成结构320的各种通信结构。存储系统330是图1中的存储系统110的一个可能实施例的示例,存储系统330包括盘子系统336、虚拟化和I/O管理堆栈337、NVMe队列管理器338和CPU 335。CPU 335包括核331、332、333和334,每个核都具有两个附加的I/O队列。在其它实施例中,核331-核334可以具有任何数量的附接的I/O队列,最多达如上所述的最大支持队列数量。连接321和322是CPU核与主机中的NVMe-oF I/O队列之间的连接的示例。

[0044] 在该示例中,主机A和主机B都连接到存储系统,并且由主机建立到所有四个CPU核的I/O队列。在该示例中,A1和B1队列比其它队列具有更多的I/O工作负载,因此变得过载。这造成了整个系统的不平衡和资源的利用不足。

[0045] 图3b描绘了来自图3a的系统的示例,但是结合了本发明的实施例。在该示例中,存储系统330向主机B发送AER消息,以通知核331过载并且核332未被充分利用,使得将流量从核331移动到核332将平衡系统并提高性能。在该示例中,示出了带内信令341和带外信令342两者。在一个实施例中,带外信令342使用带外API实例339来与主机通信。在一个实施例中,使用带内或带外信令。在另一实施例中,使用带内和带外信令两者。

[0046] 图3c描绘了来自图3a的系统的示例,但是结合了本发明的实施例。在该示例中,存储系统330已经将先前在图3b中的队列B1的核331上的流量(经由连接322)移动到先前未充分利用的核332和队列B2(经由连接323),并且由此已经重新平衡了CPU核和I/O队列的利用,从而增加了吞吐量并减少了延迟。

[0047] 图4是描绘队列平衡程序112的操作步骤以改进IOQ子系统中的工作负载管理的工作流400的流程图。在一个实施例中,队列平衡程序112使用收集关于CPU核的信息并检查所有可用CPU核的CPU核消耗的守护程序来连续地监测NVMe系统中的所有CPU核的CPU核利用率百分比。在一个实施例中,队列平衡程序112确定是否检测到一个或多个CPU核过载并且检测到另一组一个或多个CPU核未被充分利用。在一个实施例中,如果队列平衡程序112确定一个或多个CPU核被检测为过载并且另一组一个或多个CPU核被检测为未充分利用,则队

列平衡程序112将使用守护程序来向NVMe控制器发送具有不平衡消息的信号。在一个实施例中，在从监测守护程序接收到CPU_IMBALANCE消息时，队列平衡程序112将遍历连接过载CPU核的所有I/O队列，并且通过访问由存储控制器维护的数据访问图来收集I/O统计。在一个实施例中，队列平衡程序112解析作为过载CPU核的一部分的主机的所有I/O队列，并且捕获将被考虑用于过载平衡的其它IOQ信息。在一个实施例中，队列平衡程序112选择新IOQ以被推荐用于I/O平衡。在一个实施例中，队列平衡程序112使用IOQ管理器将新IOQ_ID映射到NVMe限定名。在一个实施例中，队列平衡程序112生成具有建议的新IOQ_ID的AER消息到指定的NQN，以推荐将工作负载转移到该IOQ。在一个实施例中，队列平衡程序112从主机接收具有在步骤412中选择的新IOQ的响应。在一个实施例中，队列平衡程序112确定主机是否已经接受推荐。在一个实施例中，队列平衡程序改变主机IOQ偏好设置，并且在指定了IOQ_ID的队列上发送更多的工作负载。

[0048] 在一个备选实施例中，工作流400的步骤可由任何其它程序在与队列平衡程序112一起工作时执行。应当理解，本发明的实施例至少提供了改进IOQ子系统中的工作负载管理。然而，图4仅提供了一种实施方式的图示，并且不暗示对其中可以实现不同实施例的环境的任何限制。本领域技术人员可以对所描绘的环境进行许多修改，而不背离权利要求所陈述的本发明的范围。

[0049] 队列平衡程序112监测CPU核利用率(步骤402)。在步骤402，队列平衡程序112使用监测守护程序连续监测NVMe系统中的所有CPU核的CPU核利用率百分比，监测守护程序收集包括CPU核利用率和I/O队列资源可用性以及所有可用CPU核的利用率的信息。在一个实施例中，队列平衡程序112使用与NVMe控制器和队列管理器并行运行的监测守护程序来监测在所有可用CPU核上建立的队列、工作负载和CPU核可用性。在一个实施例中，队列平衡程序112从存储系统配置映射(storage system configuration map)和存储系统使用表中收集CPU核利用率数据。

[0050] 队列平衡程序112确定CPU核是否过载(判定框404)。在一个实施例中，队列平衡程序112确定一个或多个CPU核是否处于过载状况下以及另一组一个或多个CPU核是否未被充分利用。在一个实施例中，使用预定的阈值来检测过度利用和未充分利用。在一个实施例中，一旦遇到队列重叠情况(例如，如图3a所示，其中主机A 310和主机B 312连接到同一CPU核)，队列平衡程序112将确定CPU工作负载和负载不平衡是否超过预定阈值。例如，阈值可以是，如果利用率百分比大于80％，则CPU核被过度利用。在一个实施例中，预定阈值是系统默认值。在另一实施例中，在运行时从用户接收预定阈值。

[0051] 在另一实施例中，队列平衡程序112通过测量每个核在一段时间内的平均利用率来确定一个或多个CPU核处于过载状况。在该实施例中，如果CPU核的平均利用率超过阈值达一段时间，则队列平衡程序112确定CPU核被过度利用。例如，阈值可以是，如果核超过50％利用率达超过一分钟，则CPU核过载。在一个实施例中，平均利用率百分比是系统默认值。在另一实施例中，在运行时从用户接收平均利用率。在一个实施例中，该时间段是系统默认值。在另一实施例中，在运行时从用户接收该时间段。

[0052] 在又一个实施例中，队列平衡程序112确定CPU核在核的利用率在短时间段内达到峰值时处于过载状况。在该实施例中，如果CPU核的利用率的增加在指定时间段内超过阈值增加率，则队列平衡程序112确定CPU核被过度利用。例如，阈值可以是，如果核利用率在10

秒内增加30％，则CPU核过载。在一个实施例中，阈值增加率是系统默认值。在另一实施例中，在运行时从用户接收阈值增加率。在一个实施例中，指定的时间段是系统默认值。在另一实施例中，在运行时从用户接收指定的时间段。

[0053] 在一个实施例中，如果基于累计消耗百分比来确认CPU不平衡，则队列平衡程序112识别连接到不平衡的CPU核的I/O队列，并且分析I/O队列以寻找具有高带宽利用率的IOP工作负载。在一个实施例中，高带宽利用率的阈值是系统默认值。在另一实施例中，高带宽利用率的阈值是由用户在运行时设置的值。由于IOP工作负载是CPU敏感的，所以队列平衡程序112收集该信息，并且映射附接到过载CPU核的每个I/O队列的CPU消耗。

[0054] 如果队列平衡程序112确定在过载状况下检测到一个或多个CPU核并且检测到另一组一个或多个CPU核未被充分利用(判定框312，“是”分支)，则队列平衡程序112前进到步骤406。如果队列平衡程序112确定在过载状况下没有检测到一个或多个CPU核或没有检测到另一组一个或多个CPU核未被充分利用(判定框312，“否”分支)，则队列平衡程序112返回到步骤402以继续监测CPU核利用。

[0055] 队列平衡程序112发送不平衡消息(步骤406)。在一个实施例中，监测守护程序利用不平衡消息向NVMe控制器发送信号。在一个实施例中，不平衡消息包括被检测到过载的CPU核。在另一实施例中，不平衡消息包括被检测到未充分利用的CPU核。在又一实施例中，不平衡消息包括被检测到过载的CPU核和被检测到未充分利用的CPU核。在一些实施例中，不平衡消息包括被检测到过载的核和被检测到未充分利用的核的利用率百分比。在一个实施例中，监测守护程序使用CPU控制器管理核中的管理提交队列将信号发送到NVMe控制器，所述CPU控制器管理核是诸如来自图2的控制器管理核210。

[0056] 队列平衡程序112遍历I/O队列(步骤408)。在一个实施例中，在从监测守护程序接收到CPU_IMBALANCE消息时，队列平衡程序112遍历连接到过载CPU核的所有I/O队列，并且通过访问由存储控制器维护的数据访问图(带宽和每秒输入/输出操作(IOP)操作)来收集I/O统计。

[0057] 在一个实施例中，队列平衡程序112检查存储系统中的所有其它CPU核以及哪些核具有附加带宽。在一个实施例中，队列平衡程序112确定存储系统中的所有CPU核的利用率百分比，以确定哪些核未被充分利用，并且可以潜在地使新I/O队列被分配给它们以重新平衡存储系统。

[0058] 队列平衡程序112解析过载CPU核上的所有主机I/O队列(步骤410)。在一个实施例中，队列平衡程序112解析作为过载CPU核的一部分的主机的所有I/O队列，并且捕获其它IOQ信息。在一个实施例中，队列平衡程序112使用IOQ信息来确定过载平衡的可用选项。在一个实施例中，IOQ信息包括被检测到过载的CPU核和被检测到未充分利用的CPU核，以确定用于过载平衡的可用选项。在一个实施例中，IOQ信息包括被检测到过载的核和被检测到未充分利用的核的利用率百分比，以确定用于过载平衡的可用选项。在又一实施例中，IOQ信息包括被检测到过载的CPU核和被检测到未充分利用的CPU核以及核的利用率百分比，以确定用于过载平衡的可用选项。

[0059] 队列平衡程序112选择可以接受附加工作负载的新IOQ(步骤412)。在一个实施例中，队列平衡程序112选择新IOQ以被推荐用于I/O平衡。在一个实施例中，队列平衡程序112基于在步骤410中从每个IOQ收集的工作负载信息选择新IOQ。在一个实施例中，队列平衡程

序112基于用于新IOQ的CPU核的利用率百分比小于阈值来选择新IOQ。在一个实施例中，预定阈值是系统默认值。在另一实施例中，在运行时从用户接收预定阈值。在一个实施例中，队列平衡程序112基于存储系统上的I/O队列工作负载的对称工作负载平衡来选择新IOQ。例如，假设队列A1和队列B1位于同一CPU核上，并且正在生成高工作负载。与队列A1和B1相关联的CPU核过载，因此队列平衡程序112将检查由主机A和主机B创建的所有I/O队列。在该示例中，队列平衡程序112然后按现有CPU和相关联的工作负载对这些I/O队列分类。在该示例中，队列平衡程序112确定驻留在核2上的IOQ A2和B2具有较少的队列和较低的CPU工作负载，因此队列平衡程序112将IOQ(A1或B1)工作负载之一移动到核2。

[0060] 在一个实施例中，队列平衡程序112选择多个IOQ，每个IOQ可以用于重新平衡工作负载，并且通过可用的工作负载对IOQ进行优先级排序。在一个实施例中，队列平衡程序112选择最高优先级的可用IOQ来推荐IOQ重新平衡。在一个实施例中，最高优先级的可用IOQ是附接到具有最低利用率的CPU核的IOQ。在另一实施例中，通过选择附接到CPU核的IOQ而不选择附接到该核的其它IOQ来确定最高优先级的可用IOQ。

[0061] 队列平衡程序112将新IOQ_ID映射到NQN(步骤414)。在一个实施例中，队列平衡程序112使用IOQ管理器将在步骤412中选择的新IOQ_ID映射到远程存储目标的NQN，例如图3a-图3c的存储系统330。

[0062] 队列平衡程序112将AER和IOQ_ID一起发送到指定NQN以转移工作负载(步骤416)。在一个实施例中，队列平衡程序112生成具有建议的新IOQ_ID的AER消息到指定的NQN，以推荐将工作负载转移到该IOQ。在一个实施例中，一旦队列平衡程序112做出新I/O工作负载转移决定，该信息作为信号被发送到NVMe控制器的管理控制单元。在一个实施例中，队列平衡程序112通过内部通信或通过协议级通信(经由NVMe异步事件请求命令)向主机发送队列重叠情况的异步通知。在一个实施例中，该消息包含存储系统期望将流量移动到的IOQ_ID，以平衡CPU工作负载。由于队列平衡程序112已经建立了具有新IOQ_ID的新I/O队列，所以队列平衡程序112期望主机在建议的队列上发送更多流量以获得更高的性能和更高的并行性。

[0063] 在一个实施例中，队列平衡程序112和主机通知器之间的通信可以通过带外(OOB)协议，使用OOB应用程序接口(API)，该OOB应用程序接口(API)被实现为具有在主机和存储控制器集群系统之间通信的能力。例如，在图3b中，信号342表示带外通信。在另一实施例中，队列平衡程序112和主机通知器之间的通信可以是使用NVMe标准的带内通信。在这个实施例中，队列重叠信息和致动器信号作为协议帧的一部分被程序化地传递。例如，在图3b中，信号341表示带内通信。

[0064] 队列平衡程序112从主机接收响应(步骤418)。在一个实施例中，队列平衡程序112从主机接收对在步骤412中选择的新IOQ的响应。在图3c的示例中，新IOQ的主机是主机B312。在一个实施例中，响应可以是主机接受推荐，或者主机拒绝推荐。

[0065] 队列平衡程序112确定主机是否接受新IOQ_ID(判定框420)。在一个实施例中，队列平衡程序112确定主机是否已经接受推荐。在一个实施例中，当信号被发送到主机时，主机NVMe驱动器将决定是继续当前I/O发送策略还是采用来自队列平衡程序112的建议以对特定IOQ进行优先级排序。在一个实施例中，如果主机决定采用来自队列平衡程序112的建议，则由主机侧的NVMe驱动器调整IOQ路由策略以在建议的IOQ_ID上发送更多的流量，以获得更多的性能。来自服务器/主机的所有新业务将经由新分配的IOQ_ID发送，新分配的IOQ_

ID转到新CPU核，因此主机应用体验到提高的性能。

[0066] 在另一实施例中，如果主机可以容忍性能下降，主机可以容忍IOP的总降低，或者主机由于任何其他原因不想改变IOQ策略，则拒绝该建议，并且发送信号以向队列平衡程序112通知拒绝。在一个实施例中，一旦队列平衡程序112接收到拒绝信号，队列平衡程序112就将AER消息发送到另一主机以将其I/O工作负载转移出过载的CPU核。这样，队列平衡程序和主机将都是决定的参与方，并且通过向第二主机发信号将适度地实现工作负载平衡。例如，如果队列A1和队列B1重叠，并且队列平衡程序112确定通过从队列A1或队列B1转移负载来平衡工作负载，则队列平衡程序向主机A发送信号以使用队列A2。如果主机A拒绝该建议，则队列平衡程序向主机B发送信号以将工作负载转移到B2。这个过程重复，直到主机接受改变到新IOQ的请求。该序列化被执行以防止由多个主机同时改变优选IOQ导致产生新的不平衡的情况。

[0067] 如果队列平衡程序112确定主机已经接受了推荐(判定框312，“是”分支)，则队列平衡程序112进行到步骤422。在一个实施例中，如果队列平衡程序112确定主机还没有接受推荐(判定框312，“否”分支)，则队列平衡程序112返回到步骤412以选择不同的IOQ。在另一实施例中，如果队列平衡程序112确定主机由于工作负载不是IOP敏感的而没有接受建议(判定框312，“否”分支)，则队列平衡程序112结束这个循环。

[0068] 队列平衡程序112从主机IOQ改变已被接收的主机接收ACK(步骤422)。在一个实施例中，如果队列平衡程序112确定主机已经接受了推荐，则队列平衡程序改变主机IOQ偏好设置以在具有新IOQ_ID的队列上发送更多工作负载。在一个实施例中，队列平衡程序112从目标接收带有ACCEPTANCE消息的ACK信号。这完成了重新平衡循环。

[0069] 在一个实施例中，队列平衡程序112结束该循环。

[0070] 图5是描述根据本发明至少一个实施例的适用于队列平衡程序112的存储系统110的组件的框图。图5显示了计算机500、一个或多个处理器504(包括一个或多个计算机处理器)、通信结构502、包括随机存取存储器(RAM)516和高速缓存518的存储器506、永久性存储装置508、通信单元512、I/O接口514、显示器522和外部设备520。应当理解，图5仅提供了一个实施例的图示，并不暗示对其中可以实现不同实施例的环境的任何限制。可以对所描述的环境进行许多修改。

[0071] 如所描绘的，计算机500在通信结构502上操作，该通信结构提供(一个或多个)计算机处理器504、存储器506、永久性存储装置508、通信单元512和(一个或多个)输入/输出(I/O)接口514之间的通信。通信结构502可以用适于在处理器504(例如，微处理器、通信处理器和网络处理器)、存储器506、外部设备520和系统内的任何其它硬件组件之间传递数据或控制信息的架构来实现。例如，通信结构502可以用一个或多个总线来实现。

[0072] 存储器506和永久性存储装置508是计算机可读存储介质。在所描述的实施例中，存储器506包括RAM 516和高速缓存518。通常，存储器506可以包括任何合适的易失性或非易失性计算机可读存储介质。高速缓存518是通过保存来自RAM 516的最近访问的数据和接近最近访问的数据来增强处理器504的性能的快速存储器。

[0073] 用于队列平衡程序112的程序指令可以存储在永久性存储装置508中，或更一般地，存储在任何计算机可读存储介质中，以由相应计算机处理器504中的一个或多个经由存储器506的一个或多个存储器执行。永久性存储器508可以是磁硬盘驱动器、固态盘驱动器、

半导体存储设备、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、闪存或能够存储程序指令或数字信息的任何其它计算机可读存储介质。

[0074] 永久性存储装置508所使用的介质也可以是可移动的。例如,可移动硬盘驱动器可以用于永久性存储装置508。其它示例包括光盘和磁盘、拇指驱动器和智能卡,它们被插入到驱动器中以便传送到也是永久性存储装置508的一部分的另一计算机可读存储介质上。

[0075] 在这些示例中,通信单元512提供与其他数据处理系统或设备的通信。在这些示例中,通信单元512包括一个或多个网络接口卡。通信单元512可以通过使用物理和无线通信链路中的任一个或两者来提供通信。在本发明的一些实施例的上下文中,各种输入数据的源可以在物理上远离计算机500,使得可以接收输入数据,并且类似地经由通信单元512发送输出。

[0076] (一个或多个)I/O接口514允许与可连接到计算机500的其它设备输入和输出数据。例如,(一个或多个)I/O接口514可以提供到(一个或多个)外部设备520(例如键盘、小键盘、触摸屏、麦克风、数码相机和/或一些其它合适的输入设备)的连接。(一个或多个)外部设备520还可以包括便携式计算机可读存储介质,诸如拇指驱动器、便携式光盘或磁盘、以及存储卡。用于实践本发明的实施例的软件和数据(例如队列平衡程序112)可以存储在这样的便携式计算机可读存储介质上,并且可以经由(一个或多个)I/O接口514加载到永久性存储装置508上。I/O接口514也连接到显示器522。

[0077] 显示器522提供向用户显示数据的机制,并且可以是例如计算机监测器。显示器522还可以用作触摸屏,诸如平板计算机的显示器。

[0078] 本文描述的程序是基于在本发明的特定实施例中实现它们的应用来标识的。然而,应当理解,这里的任何特定程序术语仅是为了方便而使用,因此本发明不应当限于仅在由这样的术语标识和/或暗示的任何特定应用中使用。

[0079] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括其上具有计算机可读程序指令的计算机可读存储介质(或多个介质),所述计算机可读程序指令用于使处理器执行本发明的各方面。

[0080] 计算机可读存储介质可以是能够保持和存储由指令执行设备使用的指令的任何有形设备。计算机可读存储介质可以是例如但不限于电子存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或前述的任何合适的组合。计算机可读存储介质的更具体示例的非穷举列表包括以下:便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式光盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、诸如上面记录有指令的打孔卡或凹槽中的凸起结构的机械编码装置,以及上述的任何适当组合。如本文所使用的计算机可读存储介质不应被解释为暂时性信号本身,诸如无线电波或其他自由传播的电磁波、通过波导或其他传输介质传播的电磁波(例如,通过光纤线缆的光脉冲)、或通过导线传输的电信号。

[0081] 本文描述的计算机可读程序指令可以从计算机可读存储介质下载到相应的计算/处理设备,或者经由网络,例如因特网、局域网、广域网和/或无线网络,下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光传输光纤、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或网络接口从网络

接收计算机可读程序指令，并转发计算机可读程序指令以存储在相应计算/处理设备内的计算机可读存储介质中。

[0082] 用于执行本发明的操作的计算机可读程序指令可以是汇编指令、指令集架构(ISA)指令、机器相关指令、微代码、固件指令、状态设置数据，或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言(例如Smalltalk、C++等)以及常规的过程式编程语言(例如“C”编程语言或类似的编程语言)。计算机可读程序指令可以完全在用户的计算机上执行，部分在用户的计算机上执行，作为独立的软件包执行，部分在用户的计算机上并且部分在远程计算机上执行，或者完全在远程计算机或服务器上执行。在后一种情况下，远程计算机可以通过任何类型的网络连接到用户的计算机，包括局域网(LAN)或广域网(WAN)，或者可以连接到外部计算机(例如，使用因特网服务提供商通过因特网)。在一些实施例中，为了执行本发明的各方面，包括例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)的电子电路可以通过利用计算机可读程序指令的状态信息来执行计算机可读程序指令以使电子电路个性化。

[0083] 本文参考根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述本发明的各方面。将理解，流程图和/或框图的每个框以及流程图和/或框图中的框的组合可以由计算机可读程序指令来实现。

[0084] 这些计算机可读程序指令可以被提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器以产生机器，使得经由计算机或其他可编程数据处理装置的处理器执行的指令创建用于实现流程图和/或框图的一个或多个框中指定的功能/动作的装置。这些计算机可读程序指令还可以存储在计算机可读存储介质中，其可以引导计算机、可编程数据处理装置和/或其他设备以特定方式工作，使得其中存储有指令的计算机可读存储介质包括制品，该制品包括实现流程图和/或框图的一个或多个框中指定的功能/动作的方面的指令。

[0085] 计算机可读程序指令还可以被加载到计算机、其他可编程数据处理装置或其他设备上，以使得在计算机、其他可编程装置或其他设备上执行一系列操作步骤，以产生计算机实现的过程，使得在计算机、其他可编程装置或其他设备上执行的指令实现流程图和/或框图的一个或多个框中指定的功能/动作。

[0086] 附图中的流程图和框图示出了根据本发明的各种实施例的系统、方法和计算机程序产品的可能实现的架构、功能和操作。在这点上，流程图或框图中的每个框可以表示指令的模块、段或部分，其包括用于实现指定的逻辑功能的一个或多个可执行指令。在一些替代实施方案中，框中所注明的功能可不按图中所注明的次序发生。例如，连续示出的两个框实际上可以基本上同时执行，或者这些框有时可以以相反的顺序执行，这取决于所涉及的功能。还将注意，框图和/或流程图图示的每个框以及框图和/或流程图图示中的框的组合可以由执行指定功能或动作或执行专用硬件和计算机指令的组合的专用的基于硬件的系统来实现。

[0087] 已经出于说明的目的给出了本发明的各种实施例的描述，但是其不旨在是穷尽的或限于所公开的实施例。在不背离本发明范围的情况下，许多修改和变化对于本领域普通技术人员来说是显而易见的。选择本文所使用的术语是为了最好地解释实施例的原理、实际应用或对市场上存在的技术改进，或为了使本领域的其他普通技术人员能够理解本文所

公开的实施例。

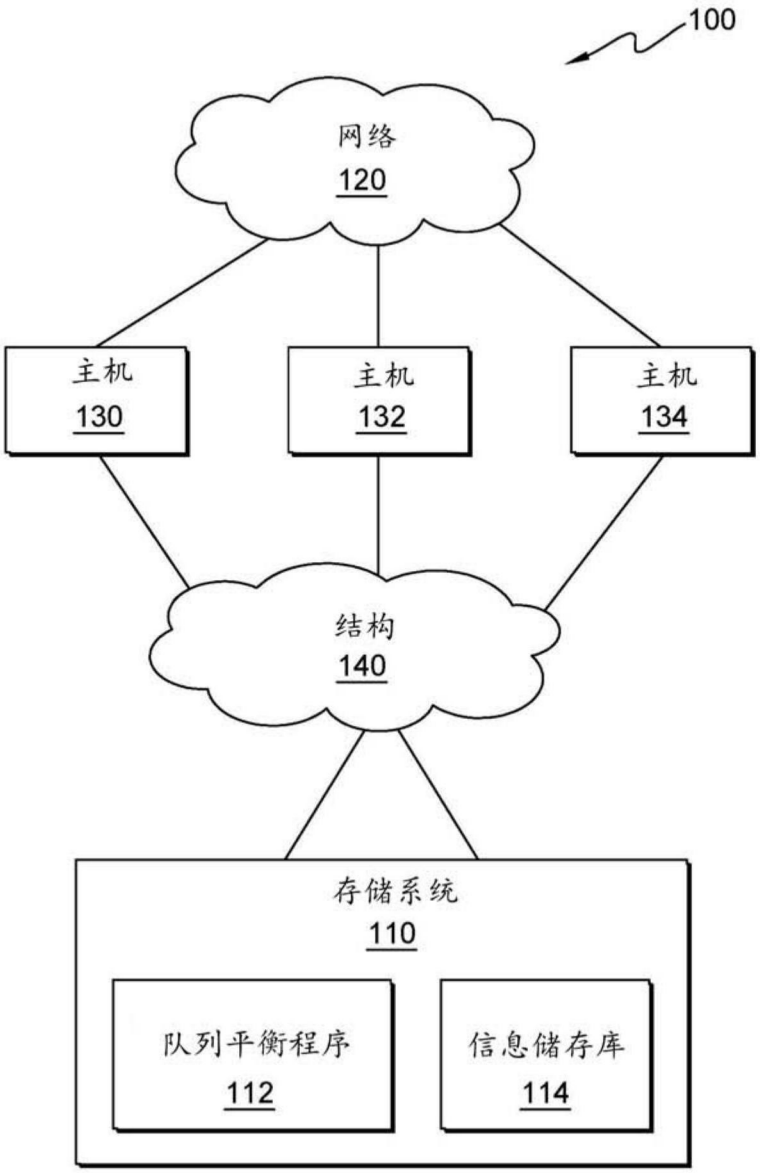

100
网络
120
主机
130
主机
132
主机
134
结构
140
存储系统
110
队列平衡程序
112
信息储存库
114

图1

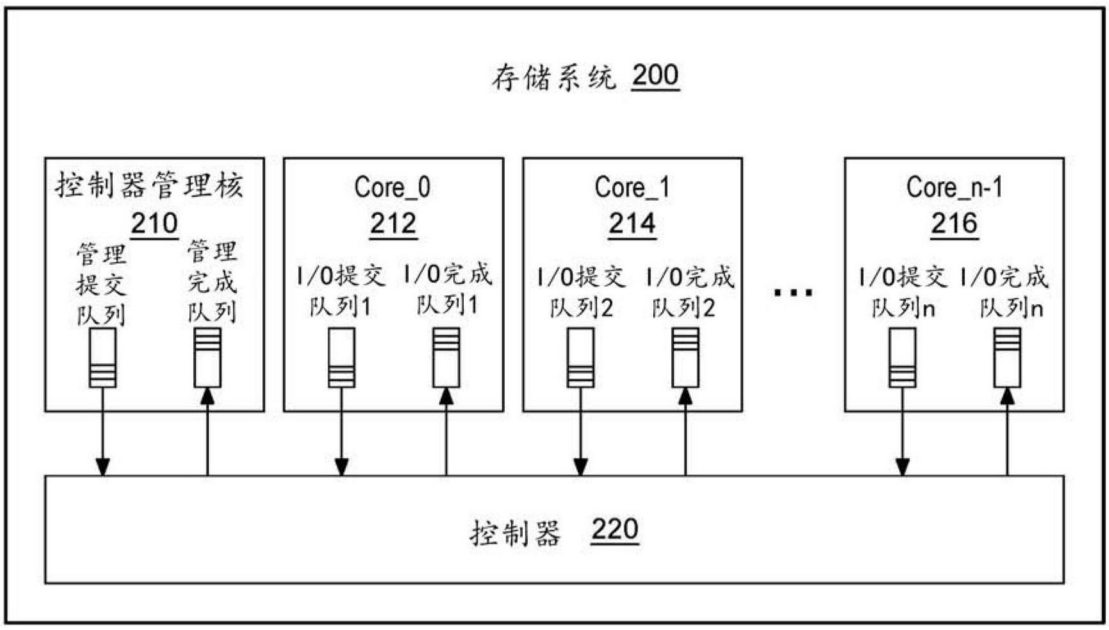

存储系统 200
控制器管理核
210
管理
提交
队列
管理
完成
队列
Core_0
212
I/O提交
队列1
I/O完成
队列1
Core_1
214
I/O提交
队列2
I/O完成
队列2
...
Core_n-1
216
I/O提交
队列n
I/O完成
队列n
控制器 220

图2

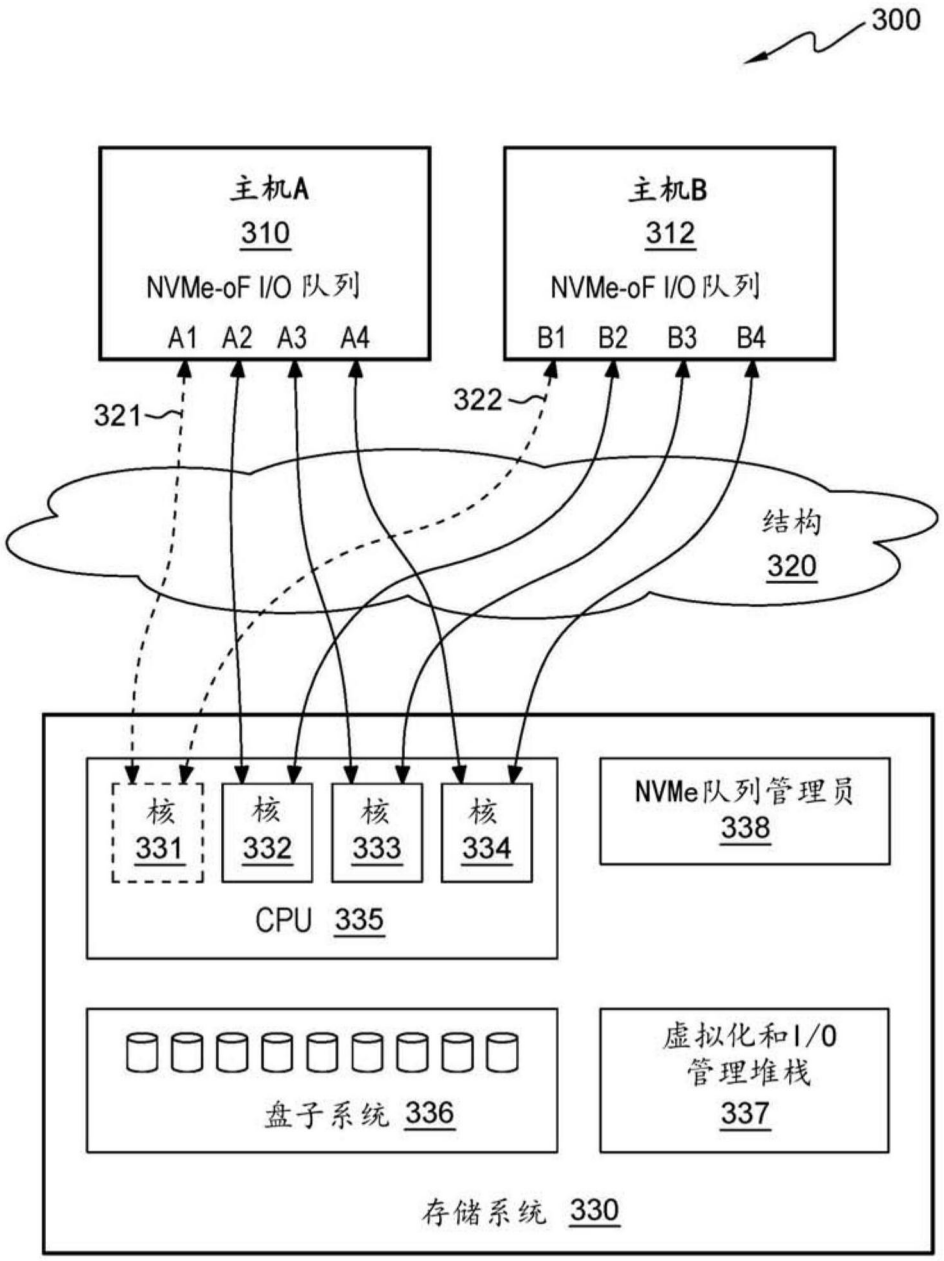

300
主机A
310
NVMe-oF I/O 队列
A1 A2 A3 A4
主机B
312
NVMe-oF I/O 队列
B1 B2 B3 B4
321
322
结构
320
核
331
核
332
核
333
核
334
CPU 335
NVMe队列管理员
338
盘子系统 336
虚拟化和I/O
管理堆栈
337
存储系统 330

图3a

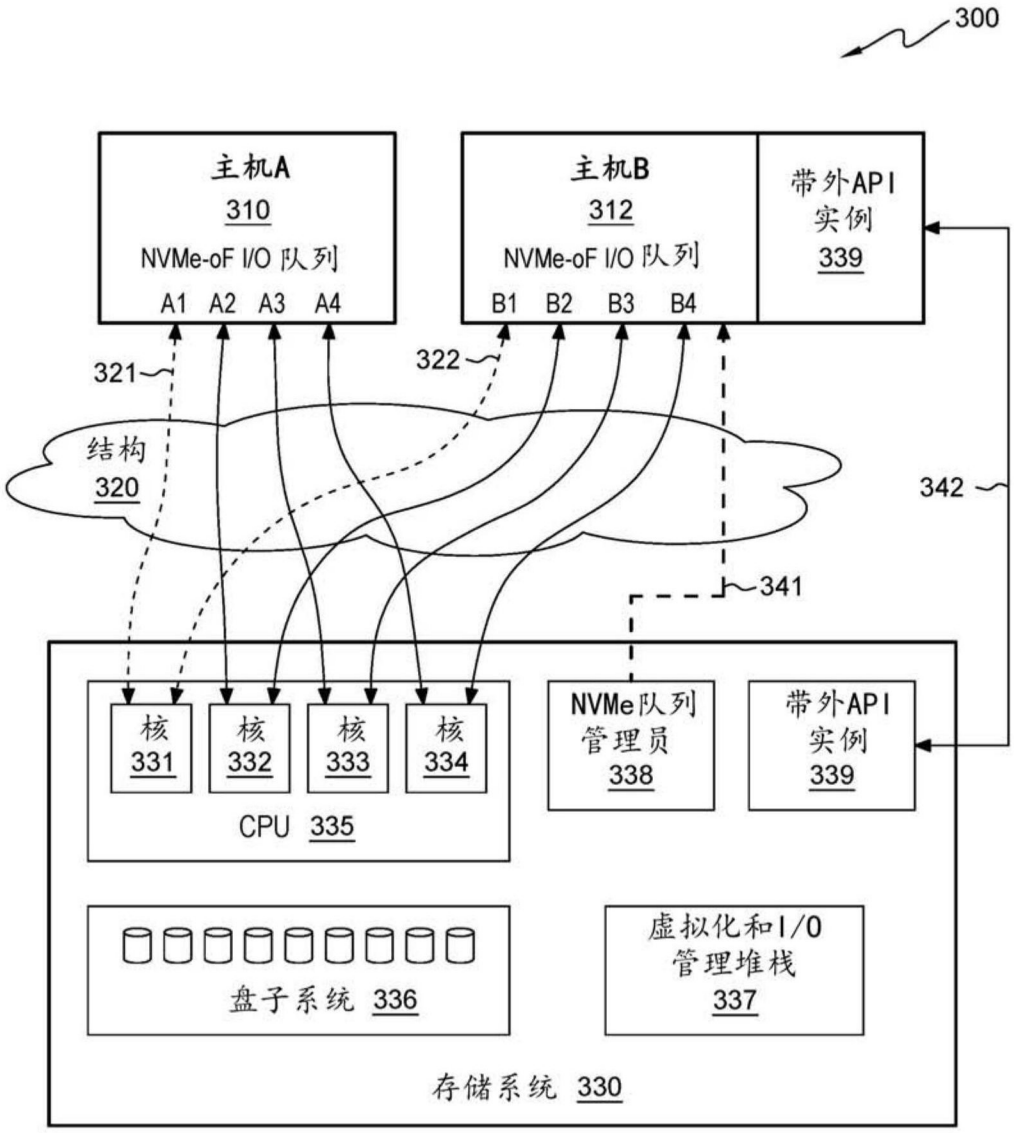

300
主机A
310
NVMe-oF I/O 队列
A1 A2 A3 A4
主机B
312
NVMe-oF I/O 队列
B1 B2 B3 B4
带外API
实例
339
321
322
结构
320
342
341
核
331
核
332
核
333
核
334
CPU 335
NVMe队列
管理员
338
带外API
实例
339
盘子系统 336
虚拟化和I/O
管理堆栈
337
存储系统 330

图3b

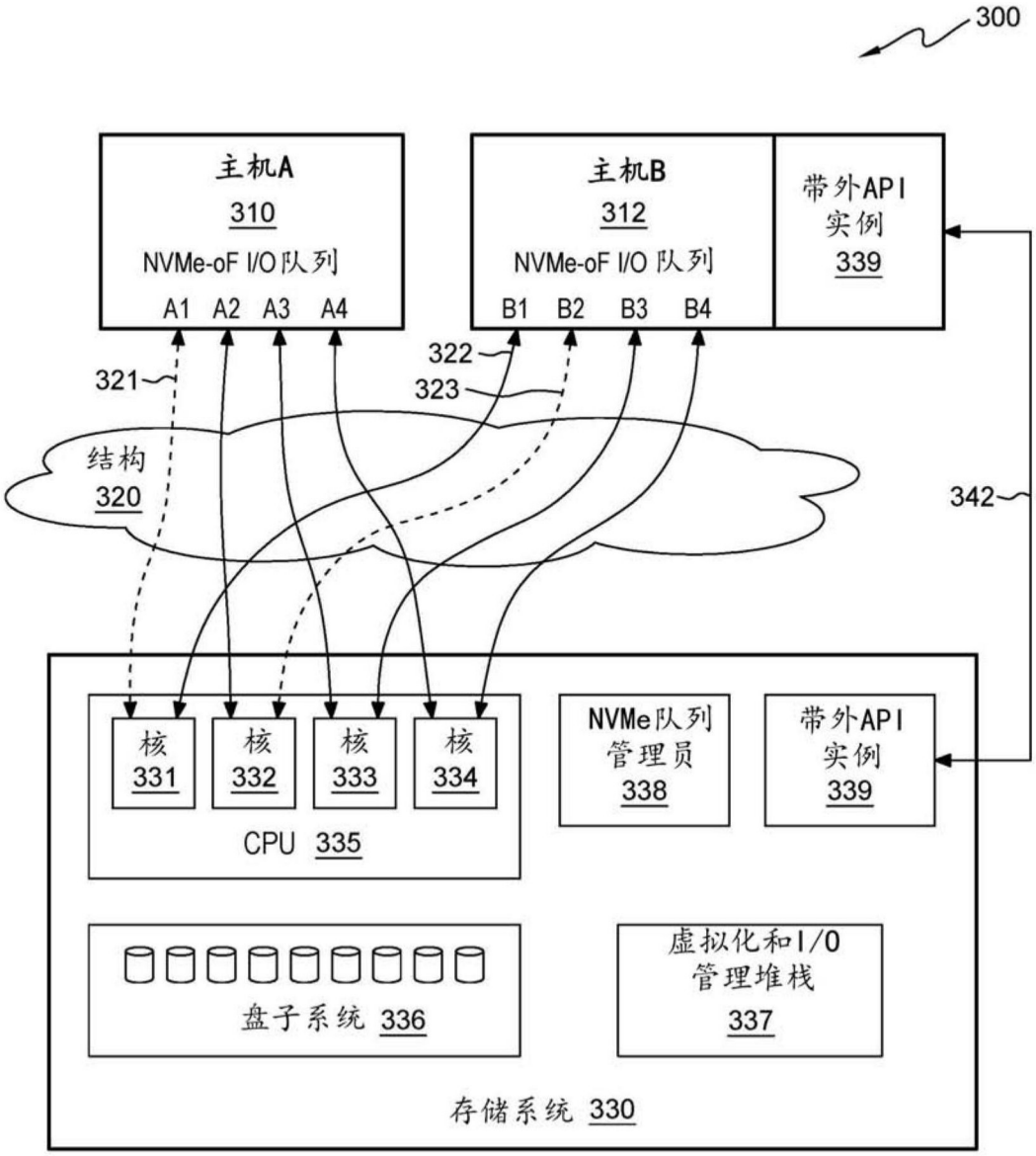

300
主机A
310
NVMe-oF I/O队列
A1 A2 A3 A4
主机B
312
NVMe-oF I/O 队列
B1 B2 B3 B4
带外API
实例
339
321
322
323
结构
320
342
核
331
核
332
核
333
核
334
CPU 335
NVMe队列
管理员
338
带外API
实例
339
盘子系统 336
虚拟化和I/O
管理堆栈
337
存储系统 330

图3c

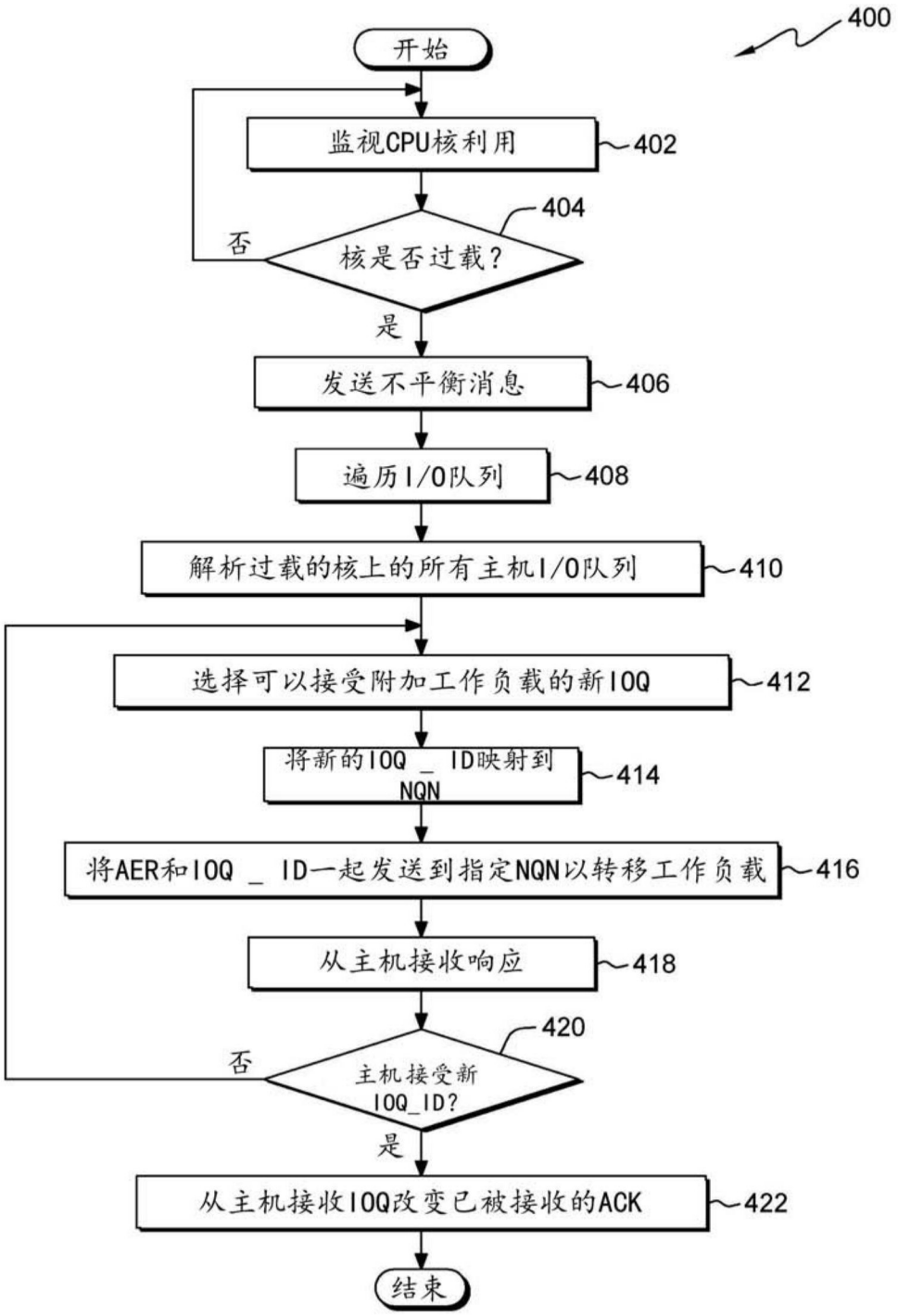

400
开始
监视CPU核利用
402
404
否
核是否过载?
是
发送不平衡消息
406
遍历I/O队列
408
解析过载的核上的所有主机I/O队列
410
选择可以接受附加工作负载的新IOQ
412
将新的IOQ _ ID映射到NQN
414
将AER和IOQ _ ID一起发送到指定NQN以转移工作负载
416
从主机接收响应
418
420
否
主机接受新IOQ_ID?
是
从主机接收IOQ改变已被接收的ACK
422
结束

图4

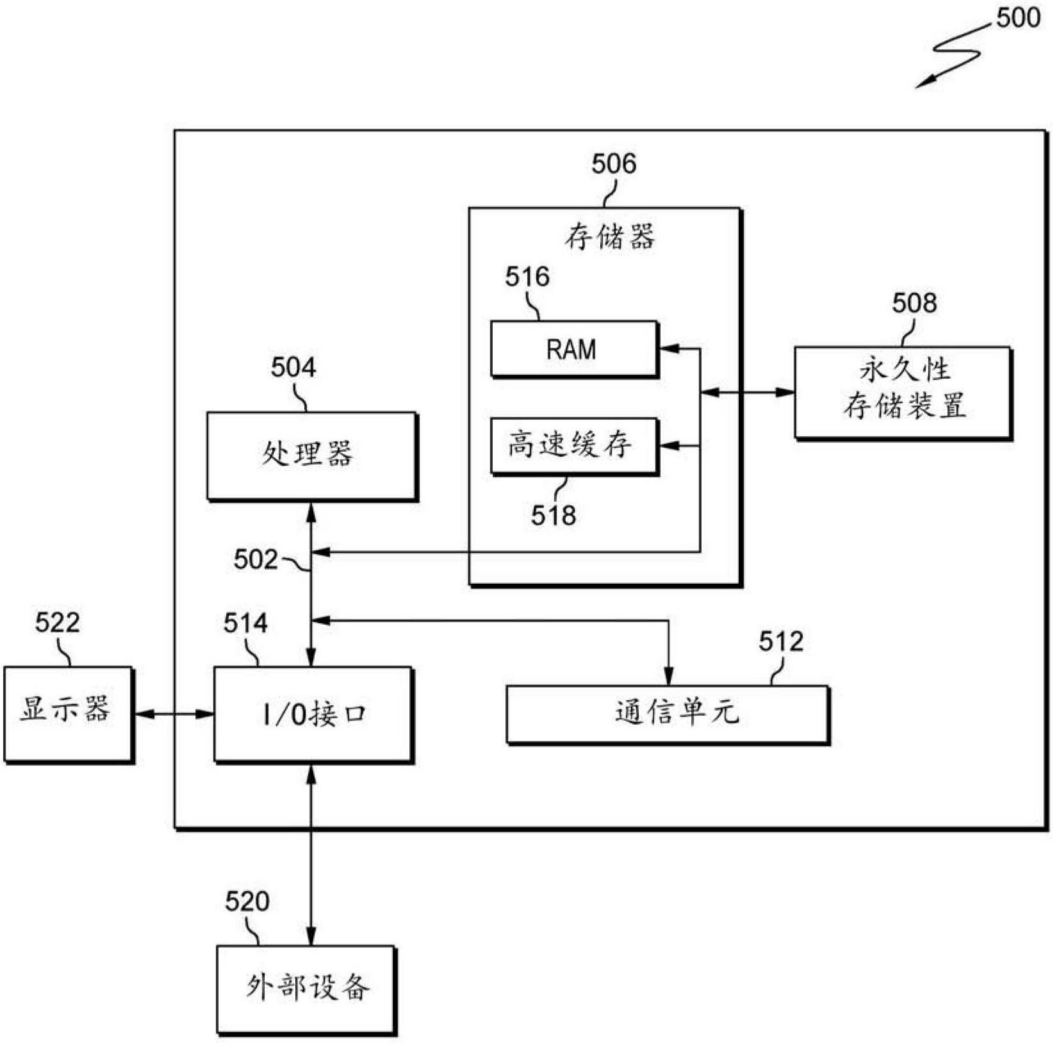

500
506
存储器
516
RAM
508
永久性
存储装置
504
处理器
高速缓存
518
502
522
514
512
显示器
I/O接口
通信单元
520
外部设备

图5