(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113692725 A
(43)申请公布日 2021.11.23

(21)申请号 202080028945.8

(22)申请日 2020.03.23

(30)优先权数据
62/852,203 2019.05.23 US
62/852,273 2019.05.23 US
62/852,289 2019.05.23 US

(85)PCT国际申请进入国家阶段日
2021.10.14

(86)PCT国际申请的申请数据
PCT/US2020/024272 2020.03.23

(87)PCT国际申请的公布数据
WO2020/236291 EN 2020.11.26

(71)申请人 慧与发展有限责任合伙企业
地址 美国德克萨斯州

(72)发明人 D·C·休森 P·昆都

(74)专利代理机构 北京市汉坤律师事务所 11602
代理人 初媛媛 吴丽丽

(51)Int.Cl.
H04L 12/863(2006.01)
H04L 12/851(2006.01)
H04L 12/931(2006.01)
H04L 12/861(2006.01)
G06F 13/10(2006.01)

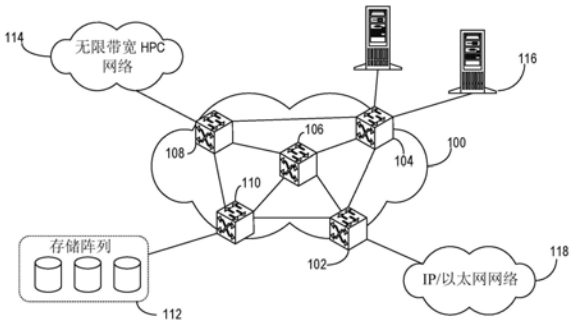
权利要求书2页 说明书8页 附图7页

(54)发明名称

促进网络接口控制器(NIC)中的高效负载均衡的系统和方法

(57)摘要

提供了一种能够在硬件引擎之间进行高效负载均衡的网络接口控制器(NIC)。所述NIC可以配备有多个排序控制单元(OCU)、队列、选择逻辑块和分配逻辑块。所述选择逻辑块可以从所述多个OCU中为来自所述队列的命令确定OCU,所述队列可以存储一个或多个命令。然后,所述分配逻辑块可以为所述OCU确定选择设置,基于所述选择设置为所述命令选择出口队列,并将所述命令发送到所述出口队列。



CN 113692725 A

1.一种网络接口控制器(NIC),包括:

多个排序控制单元(OCU);

队列,所述队列用于存储一个或多个命令;

选择逻辑块,所述选择逻辑块用于从所述多个OCU中为来自所述队列的命令确定OCU;以及

分配逻辑块,所述分配逻辑块用于:

为所述OCU确定选择设置;

基于所述选择设置为所述命令选择出口队列;以及

将所述命令发送到所述出口队列。

2.如权利要求1所述的网络接口控制器,其中,所述出口队列是流队列,并且其中,所述分配逻辑块进一步用于:

确定所述选择设置是否指示针对所述OCU的静态流队列分配;以及

响应于所述选择设置指示静态流队列分配,基于所述静态分配来选择所述流队列。

3.如权利要求2所述的网络接口控制器,其中,响应于所述选择设置指示动态流队列分配,所述分配逻辑块进一步用于从与所述OCU相关联的一组流队列中基于相应流队列上的负载来动态地选择所述流队列。

4.如权利要求3所述的网络接口控制器,其中,所述OCU在一组OCU中,并且其中,所述一组流队列与所述一组OCU相关联。

5.如权利要求2所述的网络接口控制器,其中,所述分配逻辑块进一步用于响应于确定所述流队列中的相应命令已被处理而重新分配所述流队列。

6.如权利要求1所述的网络接口控制器,其中,所述选择逻辑块进一步用于通过以下一项或多项来确定所述OCU:

基于所述命令中的指示符来确定所述OCU;以及

通过对所述命令应用散列函数来确定散列值,并基于所述散列值来识别所述OCU。

7.如权利要求1所述的网络接口控制器,所述分配逻辑块进一步用于确定所述命令是有序命令。

8.如权利要求6所述的网络接口控制器,其中,所述分配逻辑块进一步用于:

确定所述命令是否是有序命令流中的初始命令;以及

响应于所述命令不是所述初始命令,基于先前对所述有序命令流中的另一个命令的选择来选择所述出口队列。

9.如权利要求1所述的网络接口控制器,其中,响应于确定所述命令是无序命令,所述分配逻辑块进一步用于设置预定义值作为所述OCU的标识符,其中,所述预定义值指示所述命令是无序的。

10.如权利要求1所述的网络接口控制器,其中,所述分配逻辑块进一步用于维持与所述OCU相关联的命令的数量。

11.一种方法,包括:

从网络接口控制器(NIC)中的队列获得命令,其中,所述NIC包括多个排序控制单元(OCU);

从所述多个OCU中为从所述队列获得的所述命令确定OCU;

为所述OCU确定选择设置；

基于所述选择设置为所述命令选择出口队列；以及

将所述命令发送到所述出口队列。

12.如权利要求11所述的方法，其中，所述出口队列是流队列，并且其中，所述方法进一步包括：

确定所述选择设置是否指示针对所述OCU的静态流队列分配；以及

响应于所述选择设置指示静态流队列分配，基于所述静态分配来选择所述流队列。

13.如权利要求12所述的方法，进一步包括，响应于所述选择设置指示动态流队列分配，从与所述OCU相关联的一组流队列中基于相应流队列上的负载来动态地选择所述流队列。

14.如权利要求13所述的方法，其中，所述OCU在一组OCU中，并且其中，所述一组流队列与所述一组OCU相关联。

15.如权利要求12所述的方法，进一步包括响应于确定所述流队列中的相应命令已被处理而重新分配所述流队列。

16.如权利要求11所述的方法，其中，确定所述OCU进一步包括以下一项或多项：

基于所述命令中的指示符来确定所述OCU；以及

通过对所述命令应用散列函数来确定散列值，并基于所述散列值来识别所述OCU。

17.如权利要求11所述的方法，进一步包括确定所述命令是有序命令。

18.如权利要求16所述的方法，进一步包括：

确定所述命令是否是有序命令流中的初始命令；以及

响应于所述命令不是所述初始命令，基于先前对所述有序命令流中的另一个命令的选择来选择所述出口队列。

19.如权利要求11所述的方法，进一步包括，响应于确定所述命令是无序命令，设置预定义值作为所述OCU的标识符，其中，所述预定义值指示所述命令是无序的。

20.如权利要求11所述的方法，进一步包括维持与所述OCU相关联的命令的数量。

促进网络接口控制器(NIC)中的高效负载均衡的系统和方法

[0001] 发明人:David Charles Hewson(D·C·休森)和Partha Kundu(P·昆都)

[0002] 背景

技术领域

[0003] 本公开总体上涉及联网技术领域。更具体地,本公开涉及用于促进在网络接口控制器(NIC)中的硬件引擎之间进行高效负载均衡的系统和方法。

[0004] 相关技术

[0005] 随着支持网络的设备和应用变得越来越普遍,各种类型的流量以及不断增加的网络负载继续要求底层网络架构提供更高的性能。例如,诸如高性能计算(HPC)、媒体流化和物联网(IOT)等应用可以产生具有鲜明特征的不同类型的流量。因此,除了诸如带宽和延迟等传统网络性能指标外,网络架构仍继续面临诸如可扩展性、多功能性和效率等挑战。

发明内容

[0006] 提供了一种能够在硬件引擎之间进行高效负载均衡的网络接口控制器(NIC)。所述NIC可以配备有多个排序控制单元(OCU)、队列、选择逻辑块和分配逻辑块。所述选择逻辑块可以从所述多个OCU中为来自所述队列的命令确定OCU,所述队列可以存储一个或多个命令。然后,所述分配逻辑块可以为所述OCU确定选择设置,基于所述选择设置为所述命令选择流队列,并将所述命令发送到所述流队列。

附图说明

[0007] 图1示出了示例性网络。

[0008] 图2A示出了具有多个NIC的示例性NIC芯片。

[0009] 图2B示出了NIC的示例性架构。

[0010] 图3示出了NIC中的硬件引擎之间的高效负载均衡。

[0011] 图4A示出了对NIC中的无序命令进行高效负载均衡的过程的流程图。

[0012] 图4B示出了对NIC中的有序命令进行高效负载均衡的过程的流程图。

[0013] 图5示出了配备有促进在硬件引擎之间进行高效负载均衡的NIC的示例性计算机系统。

[0014] 在这些附图中,相同的附图标记指代相同的附图元素。

具体实施方式

[0015] 对所公开实施例的各种修改对于本领域技术人员来说将是显而易见的,并且在不脱离本公开的精神和范围的情况下,本文定义的一般原理可以应用于其他实施例和应用。因此,本发明不限于所示实施例。

[0016] 概述

[0017] 本公开描述了促进在网络接口控制器(NIC)中的硬件引擎之间进行高效负载均衡

的系统和方法。NIC允许主机与数据驱动的网络进行通信。

[0018] 本文描述的实施例通过以下方式解决了在NIC中为有序命令和无序命令的组合有效地分配负载的问题：(i)在一组排序控制单元(OCU)之间分配负载；以及(ii)将相应OCU动态地映射到对应的转发硬件单元。

[0019] 在操作期间，NIC可以处理可以对主机设备的存储器执行操作的命令(例如，远程直接存储器存取(RDMA)的“GET”或“PUT”命令)。NIC的主机接口可以将NIC与主机设备耦接，并促进主机设备的存储器与NIC之间的通信。在主机设备上运行的应用程序可以经由主机接口向NIC提交命令。然后，这些命令可以由NIC的多个硬件引擎同时执行。这些硬件引擎可以是用于处理直接存储器存取(DMA)命令的专用加速器。命令可以从各自的命令队列中取得，并在硬件引擎之间分配以进行处理。

[0020] 命令的子集可以是有序的。NIC应当按顺序执行这些有序命令。其余的命令可以是无序的，并且因此可以不按照顺序执行。通常，NIC可以对命令应用散列函数以确定散列值，并将命令分配到与所述散列值相对应的硬件引擎。对有序命令应用散列函数可以生成相同的散列值。因此，如果使用散列函数在一小组硬件引擎中进行选择，则NIC可能会重复选择特定的硬件引擎，而另一个硬件引擎可能仍未得到充分利用。结果是，并发性水平会受到不利影响，从而导致NIC和耦接NIC的交换机结构的性能降低。

[0021] 为了解决这一问题，NIC可以配备有可以作为散列函数的目标的一组OCU。OCU可以允许NIC促进动态负载均衡机制，使得有序命令保持有序，同时在硬件引擎之间得到均匀均衡。相应硬件引擎可以与一个或多个流队列和转发硬件相关联。转发硬件可以包括出口队列(例如，用于发送数据包的传出缓冲器)。相应硬件引擎可以从分配到该硬件引擎的一个或多个流队列中获得命令以进行处理。OCU可以在命令队列与流队列之间形成中间层。

[0022] NIC可以使用散列函数将命令分配到OCU。然而，由于OCU的数量可以明显大于硬件引擎的数量，因此命令在OCU之间的分配可以得到均衡。随后，NIC可以基于流队列的负载动态地选择流队列。例如，NIC可以选择负载低于阈值的流队列或负载量最小的流队列。通过这种方式，可以通过选择负载量最小的硬件引擎将OCU动态地映射到硬件引擎。一旦选定，同一个硬件引擎就可以用于一组有序命令。

[0023] 由于不同的应用程序可能会生成不同的有序命令流，因此NIC可以将相应的有序命令流分配到不同的流队列。一旦特定的有序命令流的执行完成，NIC就可以释放流队列以进行后续分配和重映射。在大量OCU上应用散列函数并对流队列进行动态的基于负载的选择可以促进在对应硬件引擎之间进行高效负载分配。通过这种方式，NIC可以促进在转发引擎之间进行高效负载均衡，并且可以提高NIC和耦接NIC的交换机结构的性能。

[0024] 本发明的一个实施例提供了一种NIC，所述NIC可以配备有多个排序控制单元(OCU)、队列、选择逻辑块和分配逻辑块。所述选择逻辑块可以从所述多个OCU中为来自所述队列的命令确定OCU，所述队列可以存储一个或多个命令。然后，所述分配逻辑块可以为所述OCU确定选择设置，基于所述选择设置为所述命令选择出口队列，并将所述命令发送到所述出口队列。

[0025] 在本实施例的变体中，所述出口队列可以是流队列。然后，所述分配逻辑块可以确定所述选择设置是否指示针对所述OCU的静态流队列分配。如果所述选择设置指示静态流队列分配，则所述分配逻辑块可以基于所述静态分配来选择所述流队列。

[0026] 在进一步变体中，如果所述选择设置指示动态流队列分配，则所述分配逻辑块可以从与所述OCU相关联的一组流队列中基于相应流队列上的负载动态地选择所述流队列。

[0027] 在进一步变体中，所述OCU可以在一组OCU中，并且所述一组流队列可以与所述一组OCU相关联。

[0028] 在进一步变体中，如果所述流队列中的相应命令已被处理，则所述分配逻辑块可以重新分配所述流队列。

[0029] 在本实施例的变体中，所述选择逻辑块可以通过以下一项或多项来确定所述OCU：(i)基于所述命令中的指示符来确定所述OCU；以及(ii)通过对所述命令应用散列函数来确定散列值，并基于所述散列值来识别所述OCU。

[0030] 在本实施例的变体中，所述分配逻辑块可以确定所述命令是有序命令。

[0031] 在进一步变体中，所述分配逻辑块可以确定所述命令是否是有序命令流中的初始命令。如果所述命令不是所述初始命令，则所述分配逻辑块可以基于先前对所述有序命令流中的另一个命令的选择来选择所述出口队列。

[0032] 在本实施例的变体中，如果所述命令是无序命令，则所述分配逻辑块可以设置预定义值作为所述OCU的标识符。所述预定义值可以指示所述命令是无序的。

[0033] 在本实施例的变体中，所述分配逻辑块可以维持与所述OCU相关联的命令的数量。

[0034] 在本公开中，结合图1的描述与网络架构有关，并且结合图2A及之后的描述提供了关于与支持对幂等操作进行高效管理的NIC相关联的架构和操作的更多细节。

[0035] 图1示出了示例性网络。在该示例中，交换机网络100(也可以被称为“交换机结构(switch fabric)”)可以包括交换机102、104、106、108和110。每个交换机在交换机结构100内可以具有唯一的地址或ID。各种类型的设备和网络可以耦接到交换机结构。例如，存储阵列112可以经由交换机110耦接到交换机结构100；基于无限带宽(IB)的HPC网络114可以经由交换机108耦接到交换机结构100；诸如主机116等多个终端主机可以经由交换机104耦接到交换机结构100；并且IP/以太网网络118可以经由交换机102耦接到交换机结构100。通常，交换机可以具有边缘端口和结构端口。边缘端口可以耦接到结构外部的设备。结构端口可以经由结构链路耦接到结构内的另一个交换机。通常，流量可以经由边缘交换机的入口端口注入到交换机结构100中，并经由另一个(或同一个)边缘交换机的出口端口离开交换机结构100。入口链路可以将边缘设备(例如，HPC终端主机)的NIC耦接到边缘交换机的入口边缘端口。然后，交换机结构100可以将流量传输到出口边缘交换机，所述出口边缘交换机进而可以经由另一个NIC将流量传送到目的地边缘设备。

[0036] 示例性NIC架构

[0037] 图2A示出了具有多个NIC的示例性NIC芯片。参考图1中的示例，NIC芯片200可以是为主机116设计以与交换机结构100一起工作的定制专用集成电路(ASIC)。在该示例中，芯片200可以提供两个独立的NIC 202和204。芯片200的各个NIC可以配备有主机接口(HI)(例如，用于连接到主机处理器的接口)和一个高速网络接口(HNI)，以用于与耦接到图1的交换机结构100的链路进行通信。例如，NIC 202可以包括HI 210和HNI 220，并且NIC 204可以包括HI 211和HNI 221。

[0038] 在一些实施例中，HI 210可以是外围部件互连(PCI)接口、快速外围部件互连(PCIe)接口或计算快速链路(CXL)接口。HI 210可以经由主机连接201耦接到主机，所述主

机连接可以包括N个(例如,在一些芯片中N可以是16)PCIe Gen 4通道,能够以高达每通道25Gbps的信令速率进行操作。HNI 210可以促进高速网络连接203,其可以与图1的交换机结构100中的链路进行通信。HNI 210可以使用M个(例如,在一些芯片中M可以是4)全双工串行通道以100Gbps或200Gbps的聚合速率进行操作。M个通道中的每一个都可以分别基于非归零(NRZ)调制或脉冲幅度调制4(PAM4)以25Gbps或50Gbps的速率进行操作。HNI 220可以支持电气和电子工程师协会(IEEE)802.3基于以太网的协议、以及为更高速率的小消息提供支持的增强型帧格式。

[0039] NIC 202可以支持以下一项或多项:基于消息传递接口(MPI)的点对点消息传递、远程存储器存取(RMA)操作、批量数据集体操作的卸载和进度、以及以太网包处理。当主机发出MPI消息时,NIC 202可以匹配对应的消息类型。此外,NIC 202可以针对MPI实施紧迫协议和约定协议两者,从而从主机卸载对应的操作。

[0040] 此外,NIC 202所支持的RMA操作可以包括PUT、GET和原子存储器操作(AMO)。NIC 202可以提供可靠的传输。例如,如果NIC 202是源NIC,则NIC 202可以为幂等操作提供重试机制。此外,基于连接的错误检测和重试机制可以用于可能操纵目标状态的有序操作。NIC 202的硬件可以维持重试机制所需的状态。通过这种方式,NIC 202可以消除主机(例如,软件)上的负担。决定重试机制的策略可以由主机通过驱动程序软件来指定,从而确保NIC 202的灵活性。

[0041] 此外,NIC 202可以促进触发操作、通用卸载机制以及依赖性操作序列(诸如批量数据集体)的进度。NIC 202可以支持应用编程接口(API)(例如,libfabric API),其促进由图1的交换机结构100向在主机116上运行的应用程序提供结构通信服务。NIC 202还可以支持低级别网络编程接口,诸如Portals API。另外,NIC 202可以提供高效的以太网包处理,所述以太网包处理在NIC 202为发送方时可以包括有效发送,在NIC 202为目标时可以包括流操纵,以及校验和计算。此外,NIC 202可以支持虚拟化(例如,使用容器或虚拟机)。

[0042] 图2B示出了NIC的示例性架构。在NIC 202中,HNI 220的端口宏可以促进低级别以太网操作,诸如物理编码子层(PCS)和媒体存取控制(MAC)。另外,NIC 202可以提供对链路层重试(LLR)的支持。传入数据包可以由解析器228解析并存储在缓冲器229中。缓冲器229可以是PFC缓冲器,所述PFC缓冲器被供应用于缓冲阈值量(例如,一微秒)的延迟带宽。HNI 220还可以包括分别用于管理传出数据包和传入数据包的控制发送单元224和控制接收单元226。

[0043] NIC 202可以包括命令队列(CQ)单元230。CQ单元230可以负责获取并发出主机侧命令。CQ单元230可以包括命令队列232和调度器234。命令队列232可以包括分别用于启动器命令(PUT、GET等)和目标命令(Append、Search等)的两组独立的队列。命令队列232可以被实施为在NIC 202的存储器中维持的循环缓冲器。在主机上运行的应用程序可以直接写入命令队列232。调度器234可以包括分别用于启动器命令和目标命令的两个单独的调度器。启动器命令基于散列函数被分类到流队列236中。流队列236之一可以分配到唯一的流。此外,CQ单元230可以进一步包括触发操作模块238,所述触发操作模块负责对触发命令进行排队和分派。

[0044] 出站传输引擎(OXE)240可以从流队列236中拉取命令,以对其进行处理以供分派。OXE 240可以包括地址转换请求单元(ATRU)244,所述地址转换请求单元可以将地址转换请

求发送到地址转换单元(ATU)212。ATU 212可以代表不同的引擎提供虚拟到物理地址转换，所述不同的引擎诸如OXE 240、入站传输引擎(IXE)250和事件引擎(EE)216。ATU 212可以维持大的转换高速缓存214。ATU 212既可以自己执行转换，也可以使用基于主机的地址转换服务(ATS)执行转换。OXE 240还可以包括消息切分单元(MCU)246，所述消息切分单元可以将大的消息分割为与最大发送单元(MTU)相对应的大小的数据包。MCU 246可以包括多个MCU模块。当MCU模块可用时，MCU模块可以从指派的流队列中获得下一个命令。接收到的数据可以写入数据缓冲器242中。然后，MCU模块可以将数据包报头、对应的流量类别和数据包大小发送到流量整形器248。整形器248可以确定由MCU 246提出的哪些请求可以进入网络。

[0045] 随后，可以将所选择的数据包发送到数据包和连接跟踪(PCT)270。PCT 270可以将数据包存储在队列274中。PCT 270还可以维持出站命令的状态信息，并在返回响应时更新状态信息。PCT 270还可以维持数据包状态信息(例如，允许将响应与请求匹配)、消息状态信息(例如，跟踪多数据包消息的进度)、启动器完成状态信息、以及重试状态信息(例如，维持在请求或响应丢失时对命令进行重试所需的信息)。如果在阈值时间内未返回响应，则可以将对应的命令存储在重试缓冲器272中。PCT 270可以分别基于源表276和目标表278促进用于启动器命令和目标命令的连接管理。例如，PCT 270可以更新其源表276，以跟踪可靠地传送数据包和消息完成通知所需的状态。PCT 270可以将传出数据包转发到HNI 220，所述HNI将数据包存储在出站队列222中。

[0046] NIC 202还可以包括IXE 250，所述IXE在NIC 202为目标或目的地时提供数据包处理。IXE 250可以从HNI 220获得传入数据包。解析器256可以解析传入数据包并将对应的数据包信息传递到列表处理引擎(LPE)264或消息状态表(MST)266以进行匹配。LPE 264可以将传入消息与缓冲器进行匹配。LPE 264可以确定每个消息要使用的缓冲器和起始地址。LPE 264还可以管理用于表示缓冲器以及意外消息的列表条目262的池。MST 266可以存储匹配结果和生成目标侧完成事件所需的信息。MST 266可以由不受限制的操作使用，包括多数据包PUT命令以及单数据包和多数据包GET命令。

[0047] 随后，解析器256可以将数据包存储在数据包缓冲器254中。IXE 250可以获得匹配结果以进行冲突检查。然后，DMA写入和AMO模块252可以向存储器发出由写入和AMO操作生成的更新。如果数据包包括生成目标侧存储器读操作的命令(例如，GET响应)，则可以将所述数据包传递到OXE 240。NIC 202还可以包括EE 216，所述EE可以从NIC 202中的其他模块或单元接收生成事件通知的请求。事件通知可以指定生成填充事件或计数事件。EE 216可以管理位于主机处理器存储器内的事件队列，其将完整事件写入所述主机处理器存储器。EE 216可以将计数事件转发到CQ单元230。

[0048] NIC中的高效负载均衡

[0049] 图3示出了NIC中的硬件引擎之间的高效负载均衡。在该示例中，主机设备300可以包括NIC 320。NIC 320的HI 322可以将NIC 320与设备300耦接并促进设备300与NIC 320之间的通信。设备300可以包括存储器设备302(例如，动态随机存取存储器(DRAM)模块)。应用程序308可以发出用于存储器存取的命令(例如，DMA GET或PUT)。主机设备300可以将命令存储在存储器设备302中的命令队列306中的一个命令队列中。NIC 320可以经由HI 322从命令队列306中获得命令，并在NIC 320的硬件引擎之间分配命令以进行处理。跨NIC 320的多个操作单元可以分布有相应的硬件引擎，诸如OXE 330(例如，MCU和流量整形器)。NIC

320可以使用硬件引擎同时执行这些命令。

[0050] 命令队列306中的命令的子集可以是有序的。NIC 320应当按顺序执行这些有序命令。命令队列306中的其余命令可以是无序的，并且因此可以不按照顺序执行。通常，NIC 320可以对命令应用散列函数以确定散列值，并将命令分配到与所述散列值相对应的硬件引擎。对有序命令342、344和346应用散列函数可以生成相同的散列值。因此，如果使用散列函数在一小组硬件引擎中进行选择，则NIC 320可能会重复选择特定的硬件引擎，而另一个硬件引擎可能仍未得到充分利用。结果是，并发性水平会受到不利影响，从而导致NIC 320的性能降低。

[0051] 为了解决这一问题，NIC 320可以配备有OCU模块310，所述OCU模块可以包括多个OCU 312、314和316。NIC 320还可以包括OCU映射单元(OMU)326，所述OCU映射单元可以促进散列函数从OCU模块310中选择OCU。OCU模块310可以允许NIC 320促进动态负载均衡机制，使得有序命令342、344和346保持有序，同时在硬件引擎之间得到均匀均衡。NIC 320还可以包括一组流队列328。相应硬件引擎可以与一个或多个流队列和转发硬件相关联。转发硬件可以包括出口队列。相应硬件引擎可以从分配到该硬件引擎以进行处理的一个或多个流队列中获得命令。OCU模块310可以形成命令队列306与流队列328之间的中间层。NIC 320可以从命令队列306中获得命令342、344和346，并将这些命令存储在预取队列324中。

[0052] 然后，OMU 326可以对预取队列324中的命令342应用散列函数。在获得散列值后，OMU 326可以确定所述散列值对应于哪个OCU。命令342还可以指定其应当被分配到的OCU。如果散列值对应于OCU 312，则OMU 326可以将命令342分配到OCU 312。然而，由于OCU的数量可以明显大于硬件引擎的数量，因此命令在OCU模块310中的OCU之间的分配可以得到均衡。随后，NIC 320可以基于流队列328的负载动态地为命令342选择流队列340。流队列340可以具有低于阈值或最小负载量的负载。通过这种方式，OCU 312可以动态地映射到与流队列340相关联的硬件引擎。一旦选定，同一个硬件引擎就可以用于命令344和346，因为OMU 326的散列函数可以为这些命令选择OCU 312。

[0053] 由于不同的应用程序可能会生成不同的有序命令流，因此NIC 320可以将相应的有序命令流分配到不同的流队列。例如，可以将包括命令342、344和346的流分配到流队列340，而将另一个命令流分配到另一个流队列。一旦命令342、344和346的执行完成，OMU 326就可以释放流队列340以进行后续分配和重映射。此外，当命令342和344完成并且命令346正在被处理时，OMU 326可以重新映射OCU 312。这允许OMU 326在命令流中的最后一个命令已被处理(例如，进入流队列328或由OXE 330处理)时就释放和重新分配OCU 312。在OCU模块310中的大量OCU上应用散列函数并对来自流队列328中的流队列进行动态的基于负载的选择可以促进在NIC 320中的对应硬件引擎之间进行高效负载分配，从而提高NIC 320的性能。

[0054] OCU模块310中的OCU可以划分为多个OCU集(OCUSET)。可以将与NIC 320相关联的资源(诸如命令队列306、预取队列324和流队列328)分配到OCUSET。OCUSET中的相应OCU可以共享与OCUSET相关联的资源。例如，如果OCU 312和314属于同一个OCUSET，则OCU 312和314可以共享属于OCUSET的资源，诸如流队列340。NIC 320中的OCUSET映射表可以指示哪个OCU属于哪个OCUSET。每个OCUSET可以包括一组不同的OCU。此外，流队列映射表可以基于各自的流队列标识符来确定哪些流队列被分配到OCUSET。可以将多个OCUSET映射到同一个流

队列。

[0055] 相应OCU(诸如OCU 312)可以表示有序命令342、344和346的流。来自特定命令队列的所有有序命令都应当分配到同一个OCU。应当注意，来自命令队列306中的不同命令队列的命令可以分配到同一个OCU。然而，当来自多个命令队列的命令被分配到同一个OCU时，NIC 320可以将这些命令视为同一个命令流的多个部分。因此，每个OCU都映射到流队列，并且与该OCU相关联的有序命令应当进入相同的流队列。例如，由于命令342、344和346都被映射到OCU 312，因此这些命令中的每一个命令都可以转发到流队列340。

[0056] 如果流队列340中没有针对OCU 312的未完成命令，则NIC 320可以基于负载将OCU 312映射到不同的流队列。可以将无序命令348分配到属于与所述命令相关联的OCUSET的任何流队列。命令348与其OCUSET之间的关联性可以基于命令348的命令队列来确定。换言之，命令348可以与同命令348的命令队列相关联的OCUSET相关联。对流队列的选择可以由选择设置来决定。基于选择设置，NIC可以通过基于负载动态地分配流队列或通过将OCU映射到流队列(例如，基于用户配置)静态地分配流队列来选择流。

[0057] 与OCU相关联的命令只能转发到其OCUSET内的流队列。因此，对于每个OCUSET，NIC 320可以维持OCUSET可用的流队列的数量或计数、以及指派给OCUSET的流队列中的当前最小负载值。OMU 326可以在用于将命令342指派给流队列340时计算负载指标。负载指标可以根据转发到OXE 330的命令和由OXE 330确定的命令负载来计算。NIC 320可以维持与相应类别的流量相关联的多个值，诸如报头、比例和移位。这些值可以存储在对应的寄存器(例如，控制和状态寄存器(CSR))中。NIC 320可以使用命令的长度和CSR中的值来确定负载。

[0058] 图4A示出了对NIC中的无序命令进行高效负载均衡的过程的流程图。在操作期间，NIC可以基于散列为无序命令确定OCU(操作402)，并且确定针对所述OCU的流队列分配是否是静态的(操作404)。如果针对OCU的流队列分配是静态的，则NIC可以选择分配到OCU的流队列(操作406)。另一方面，如果流队列分配不是静态的(即，动态的)，则NIC可以选择与OCUSET相关联的最小负载流队列(操作408)。

[0059] 在选择流队列(操作406或408)后，NIC可以设置OCU标识符以指示所述命令是无序命令(操作410)。应当注意，由于无序命令可以分配到与OCUSET相关联的任何流队列，因此预定义值(例如，N-1，其中，N是NIC中OCU的数量)可以设置为OCU标识符。设置预定义值可以通知流队列命令可以不按照顺序传送。然后，NIC可以使与OCU标识符相关联的计数器(例如，消息计数器)递增(操作412)，并且将命令发送到所选择的流队列并更新与流队列相关联的负载(操作414)。

[0060] 图4B示出了对NIC中的有序命令进行高效负载均衡的过程的流程图。在操作期间，NIC可以基于散列为有序命令确定OCU(操作452)，并且确定针对所述OCU的流队列分配是否是静态的(操作454)。如果针对OCU的流队列分配是静态的，则NIC可以选择分配到OCU的流队列(操作456)。另一方面，如果流队列分配不是静态的(即，动态的)，则NIC可以确定接收到的消息是否是与命令相关联的初始消息(操作458)。如果所述接收到的消息不是初始消息，则NIC可以选择已经映射的流队列(操作460)。

[0061] 另一方面，如果所述接收到的消息是初始消息，则NIC可以选择与OCUSET相关联的最小负载流队列(操作462)。然后，NIC可以利用所选择的流队列来更新流队列映射(操作464)。然后，NIC可以设置OCU标识符以指示所述命令是有序命令(操作466)。在选择静态流

队列(操作456)或映射流队列(操作460)或者设置OCU标识符(操作466)后,NIC可以使与OCU标识符相关联的计数器(例如,消息计数器)递增(操作468),并且将命令发送到所选择的流队列并更新与流队列相关联的负载(操作470)。

[0062] 示例性计算机系统

[0063] 图5示出了配备有促进高效数据包转发的NIC的示例性计算机系统。计算机系统550包括处理器552、存储器设备554和存储设备556。存储器设备554可以包括易失性存储器设备(例如,双列直插式存储器模块(DIMM))。此外,计算机系统550可以耦接到键盘562、定向设备564和显示设备566。存储设备556可以存储操作系统570。应用程序572可以在操作系统570上操作。

[0064] 计算机系统550可以配备有耦接促进高效数据请求管理的NIC 520的主机接口。NIC520可以向计算机系统550提供一个或多个HNI。NIC 520可以经由HNI之一耦接到交换机502。NIC 520可以包括OCU逻辑块530(例如,在NIC 520的CQ单元中)。OCU逻辑块530可以管理NIC 520的OCU,并且可以包括选择逻辑块532、映射逻辑块534和分配逻辑块536。

[0065] 选择逻辑块532可以基于散列函数从OCU模块540中为命令选择OCU。映射逻辑块534可以更新OCUSET映射表和流队列映射表。分配逻辑块536可以基于静态分配或动态分配将流队列分配到OCU。动态分配可以基于动态负载均衡机制。

[0066] 总之,本公开描述了一种促进在NIC中的硬件引擎之间进行高效负载均衡的NIC。所述NIC可以配备有多个排序控制单元(OCU)、队列、选择逻辑块和分配逻辑块。所述选择逻辑块可以从所述多个OCU中为来自所述队列的命令确定OCU,所述队列可以存储一个或多个命令。然后,所述分配逻辑块可以为所述OCU确定选择设置,基于所述选择设置为所述命令选择流队列,并将所述命令发送到所述流队列。

[0067] 上述方法和过程可以由硬件逻辑块、模块、逻辑块或装置来执行。硬件逻辑块、模块、逻辑块或装置可以包括但不限于专用集成电路(ASIC)芯片、场可编程门阵列(FPGA)、在特定时间执行代码块的专用或共享处理器、以及现在已知或以后开发的其他可编程逻辑设备。硬件逻辑块、模块或装置在被激活时执行其内包括的方法和过程。

[0068] 本文描述的方法和过程也可以体现为代码或数据,该代码或数据可以存储在存储设备或计算机可读存储介质中。当处理器读取并执行所存储的代码或数据时,处理器可以执行这些方法和过程。

[0069] 本发明的实施例的前述描述是仅出于说明和描述的目的而呈现的。所述描述并非旨在是穷举的或将本发明限制为所公开的形式。相应地,对于本领域普通技术人员而言,许多的修改和变化将是显而易见的。另外,以上公开内容并非旨在限制本发明。本发明的范围由所附权利要求限定。

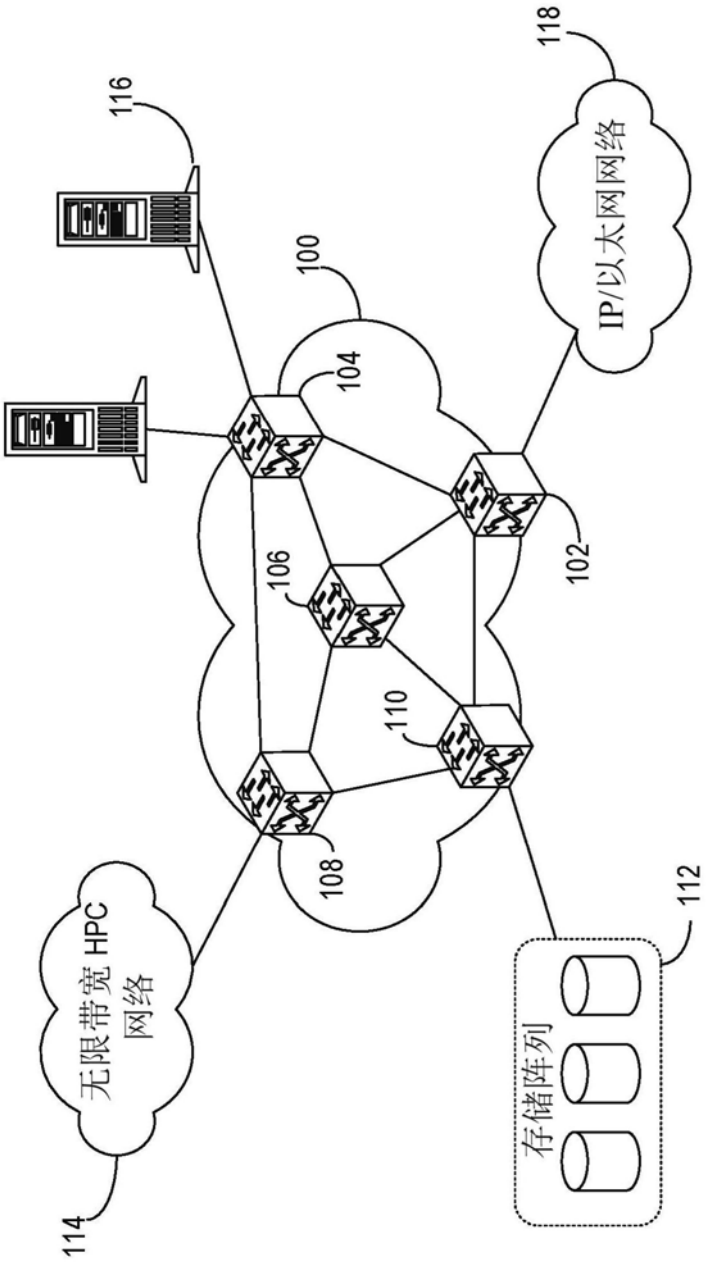

118
IP/以太网网络
116
100
104
102
106
110
108
112
存储阵列
无限带宽 HPC
网络
114

图1

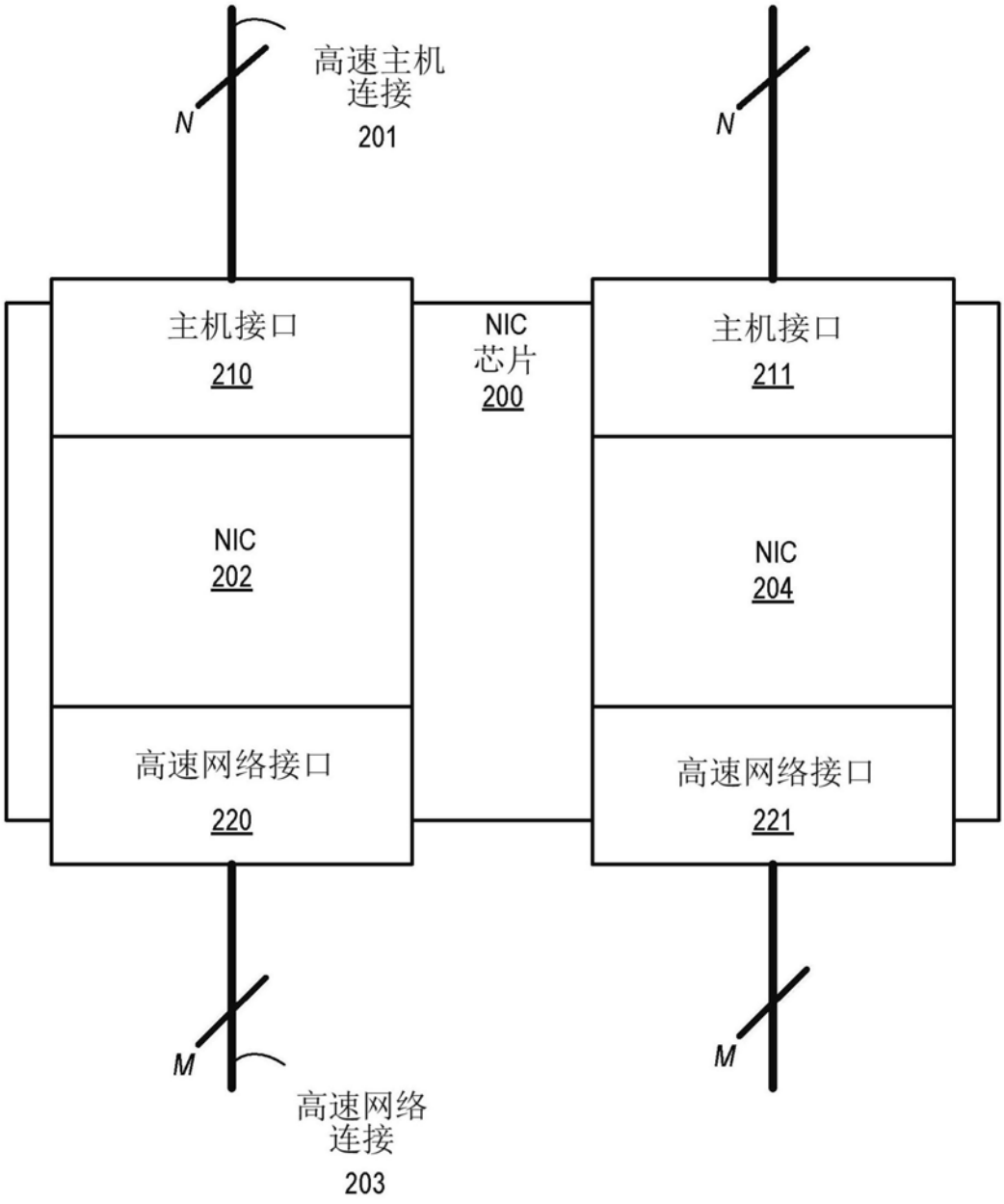

高速主机
连接
201
N
N
主机接口
210
NIC
芯片
200
主机接口
211
NIC
202
NIC
204
高速网络接口
220
高速网络接口
221
M
M
高速网络
连接
203

图2A

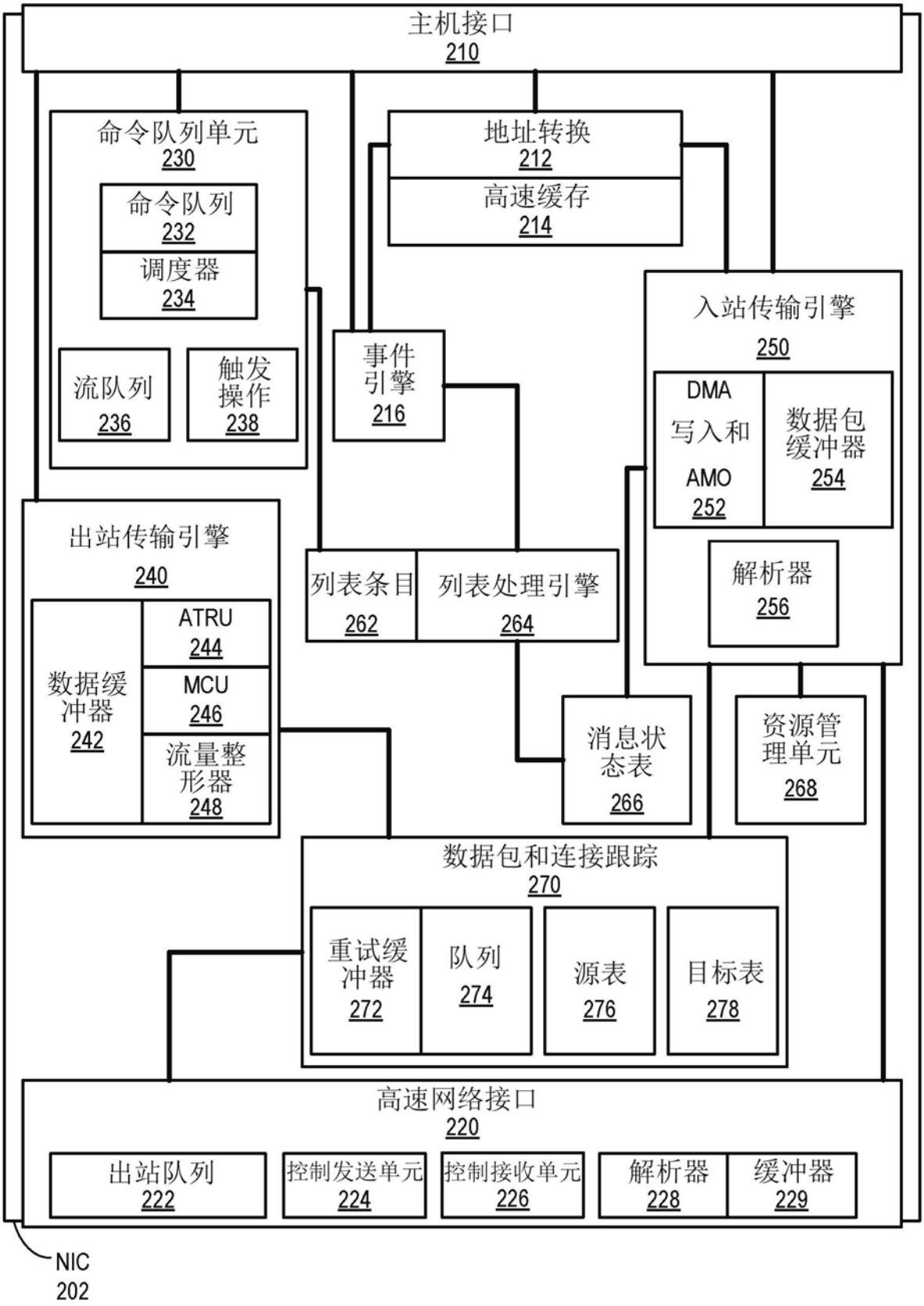

主机接口
210
命令队列单元
230
命令队列
232
调度器
234
流队列
236
触发操作
238
地址转换
212
高速缓存
214
事件引擎
216
入站传输引擎
250
DMA写入和AMO
252
数据包缓冲器
254
解析器
256
出站传输引擎
240
数据缓冲器
242
ATRU
244
MCU
246
流量整形器
248
列表条目
262
列表处理引擎
264
消息状态表
266
资源管理单元
268
数据包和连接跟踪
270
重试缓冲器
272
队列
274
源表
276
目标表
278
高速网络接口
220
出站队列
222
控制发送单元
224
控制接收单元
226
解析器
228
缓冲器
229
NIC
202

图2B

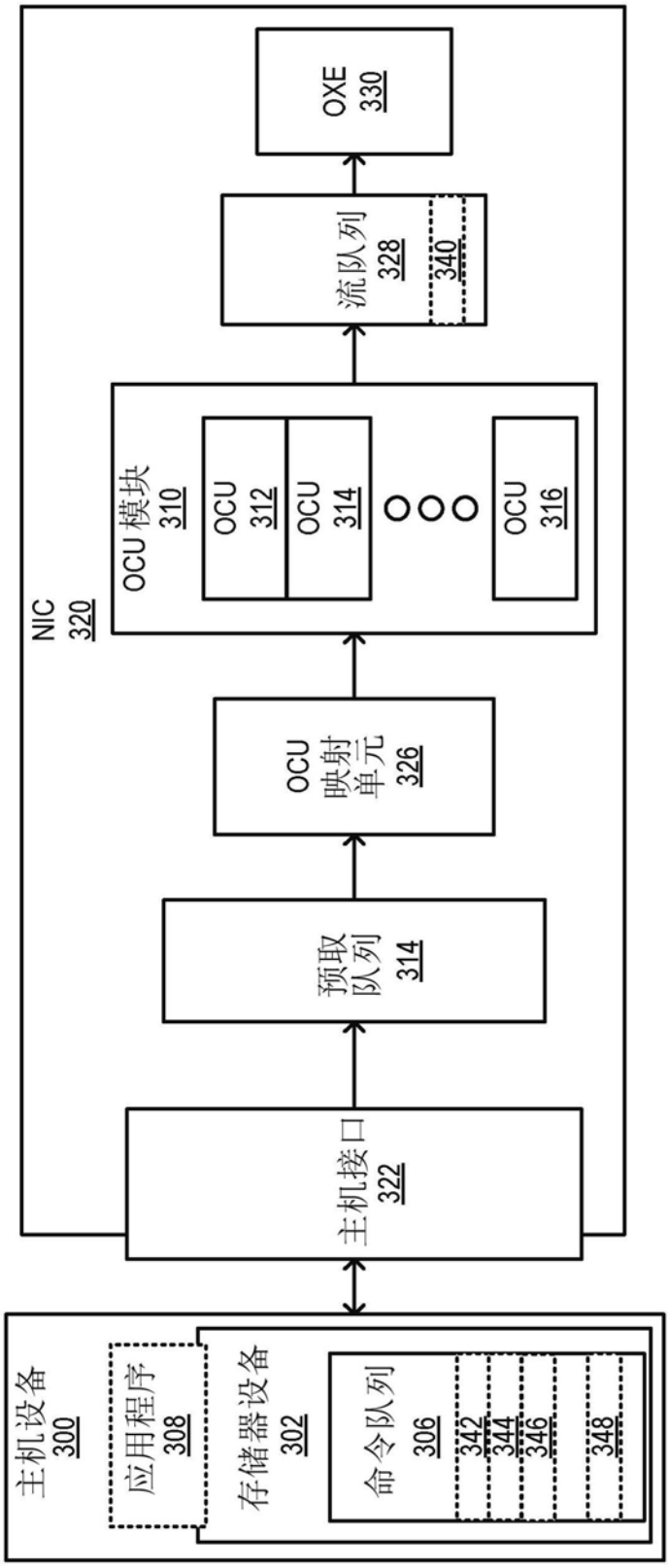

主机设备
300
应用程序
308
存储器设备
302
命令队列
306
342
344
346
348
主机接口
322
预取
队列
314
OCU
映射
单元
326
NIC
320
OCU 模块
310
OCU
312
OCU
314
OCU
316
流队列
328
340
OXE
330

图3

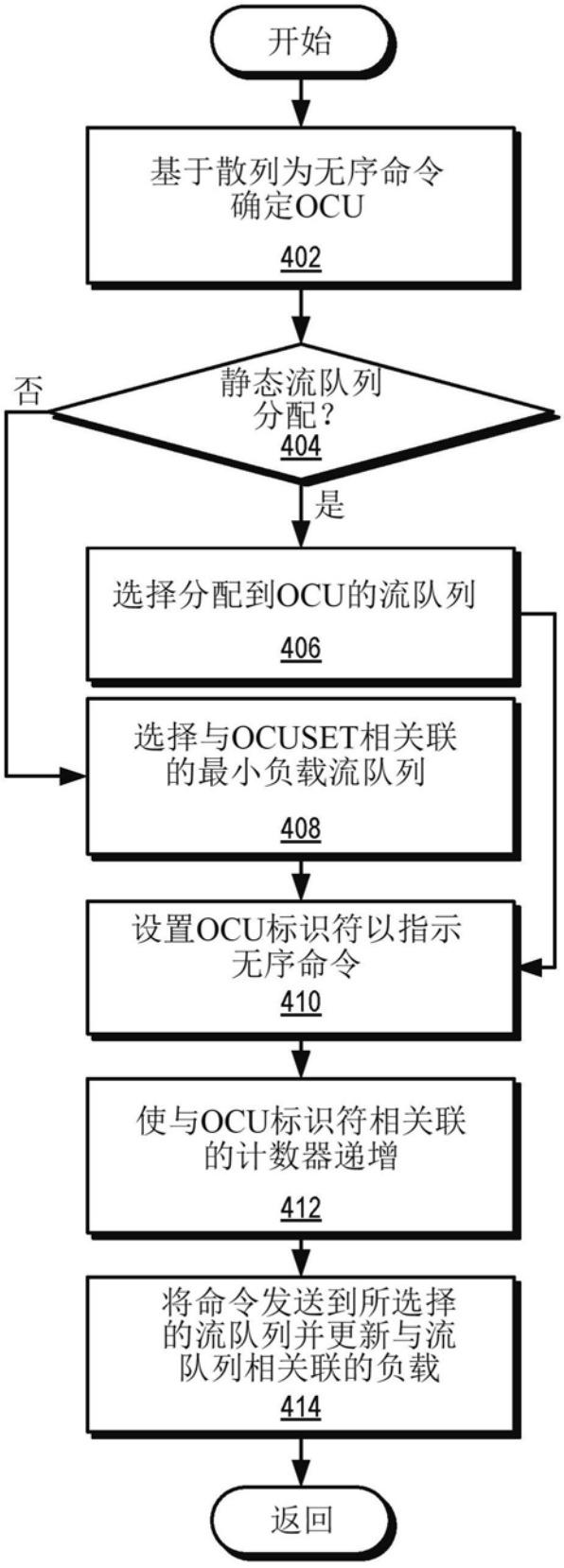

开始
基于散列为无序命令确定OCU
402
静态流队列分配？
404
否
是
选择分配到OCU的流队列
406
选择与OCUSET相关联的最小负载流队列
408
设置OCU标识符以指示无序命令
410
使与OCU标识符相关联的计数器递增
412
将命令发送到所选择的流队列并更新与流队列相关联的负载
414
返回

图4A

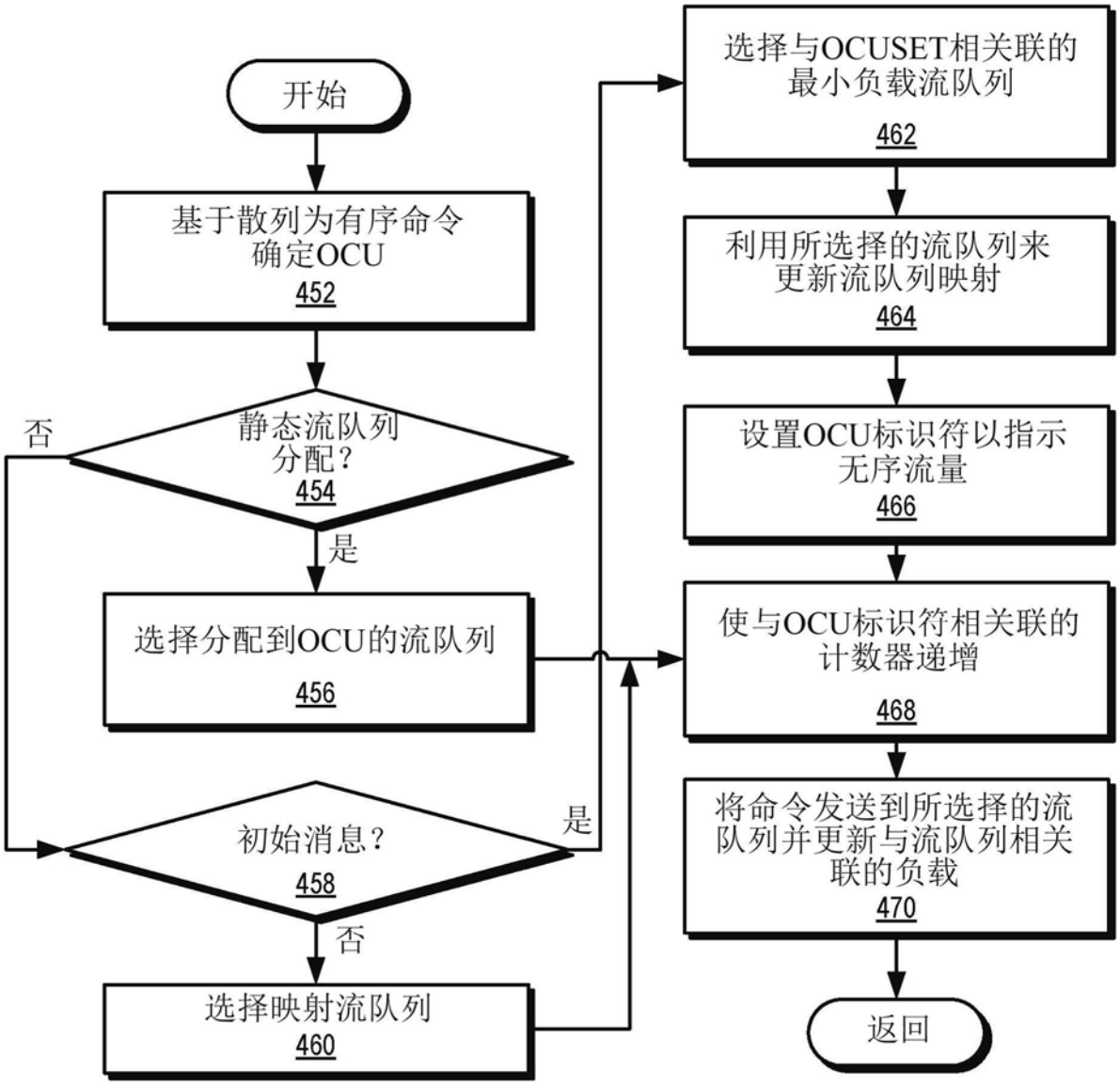

开始
基于散列为有序命令确定OCU
452
静态流队列分配？
454
否
是
选择分配到OCU的流队列
456
初始消息？
458
是
否
选择映射流队列
460
选择与OCUSET相关联的最小负载流队列
462
利用所选择的流队列来更新流队列映射
464
设置OCU标识符以指示无序流量
466
使与OCU标识符相关联的计数器递增
468
将命令发送到所选择的流队列并更新与流队列相关联的负载
470
返回

图4B

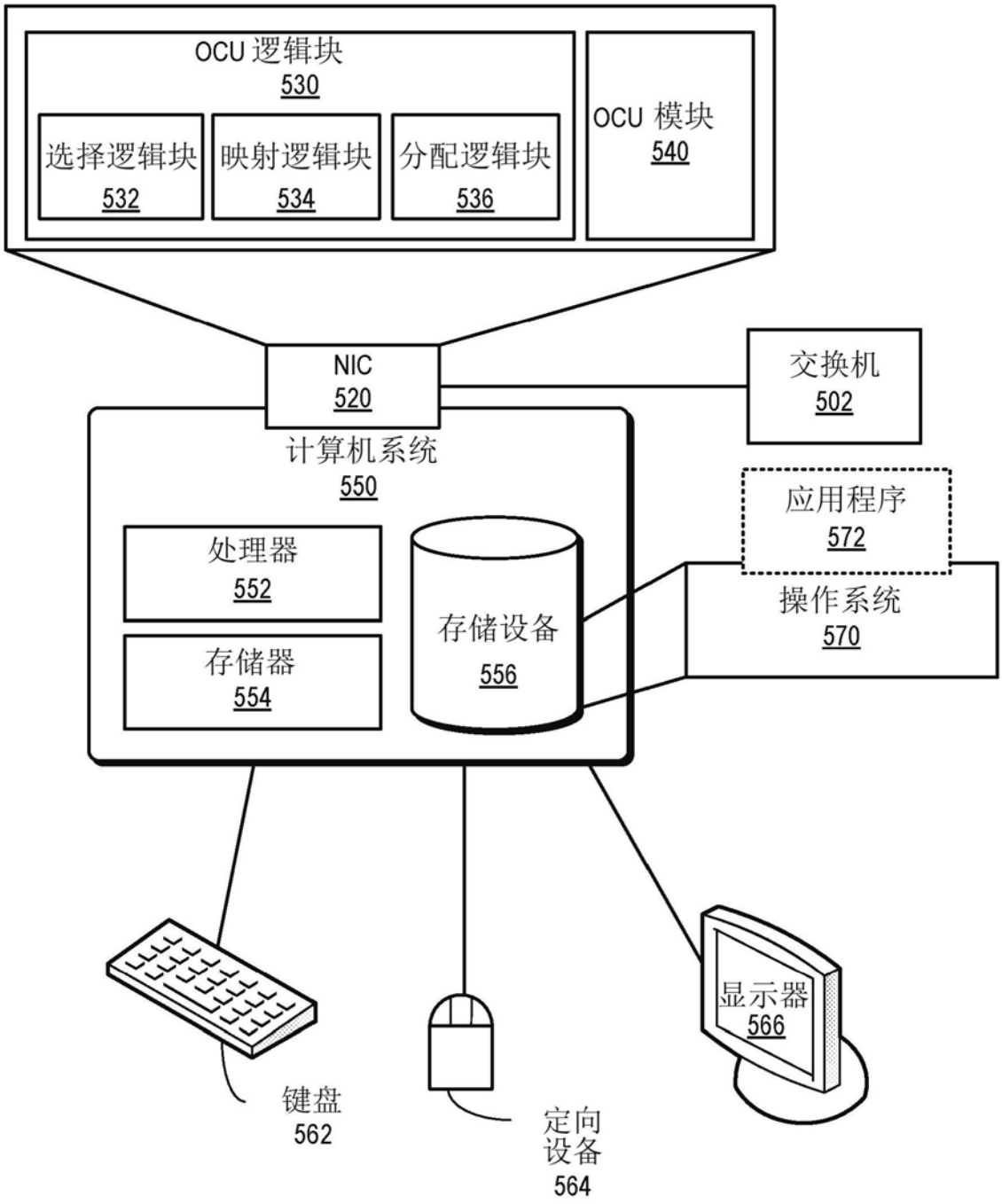

OCU 逻辑块
530
选择逻辑块
532
映射逻辑块
534
分配逻辑块
536
OCU 模块
540
NIC
520
交换机
502
计算机系统
550
处理器
552
存储器
554
存储设备
556
应用程序
572
操作系统
570
键盘
562
定向
设备
564
显示器
566

图5