(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110869919 B
(45)授权公告日 2023.09.19

(21)申请号 201880045638.3
(22)申请日 2018.07.04
(65)同一申请的已公布的文献号
申请公布号 CN 110869919 A
(43)申请公布日 2020.03.06
(30)优先权数据
15/647,561 2017.07.12 US
15/840,600 2017.12.13 US
(85)PCT国际申请进入国家阶段日
2020.01.08
(86)PCT国际申请的申请数据
PCT/IB2018/054942 2018.07.04
(87)PCT国际申请的公布数据
WO2019/012375 EN 2019.01.17
(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 M·P·莱昂斯 P·K·斯茨韦德 R·埃里克森 S·刘
(74)专利代理机构 北京市中咨律师事务所 11247
专利代理师 刘都 于静
(51)Int.Cl.
G06F 15/16(2006.01)
(56)对比文件
CN 103828310 A,2014.05.28
CN 104158747 A,2014.11.19
CN 104541314 A,2015.04.22
US 2009157844 A1,2009.06.18
WO 2007120127 A2,2007.10.25
审查员 张琦

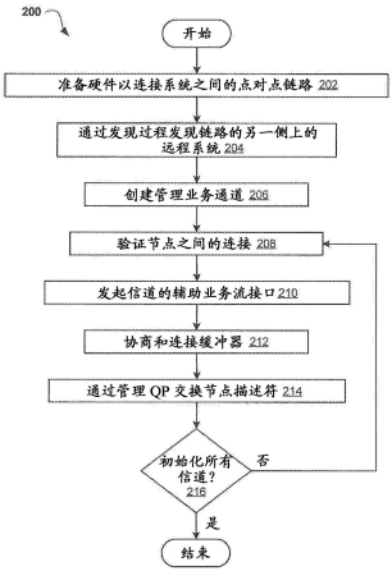
权利要求书2页 说明书11页 附图6页

(54)发明名称

用于远程节点发现、通信信道确认和连接的方法

(57)摘要

根据一个实施例,提供了一种用于计算机网络的方法、计算机系统和计算机程序产品。本发明可包括使用媒体接入控制(MAC)多播过程来发现远程节点可寻址性,以及在实现一个或多个通信信道时使用**InfiniBand®**(**InfiniBand®**和所有基于**InfiniBand ®**的商标和标志是System I/O,Incorporated和/或其附属机构的商标或注册商标)协议来连接队列对的组的方法,包括准备硬件以连接系统之间的点对点链路;通过发现过程发现链路另一侧的远程系统;创建管理业务通道;验证节点之间的连接;发起所述辅助业务流接口;协商和连接缓冲器;以及交换节点描述符。



CN 110869919 B

1.一种用于远程节点发现和链接多个队列对以用于通信信道实施的处理器实施的方法,所述方法包括:

准备多个硬件以连接多个节点之间的一个或多个点对点链路;

通过发现过程发现所述一个或多个点对点链路内的所述多个节点;

在所述多个节点之间创建多个管理业务通道,所述多个管理业务通道利用不特定于信道的多个发起协议;

验证所述多个节点之间的多个连接;

在所述多个连接上发起一个或多个辅助业务流接口;

协商和连接用于所述多个连接的多个缓冲器;以及

在所述多个节点之间交换多个节点描述符。

2.如权利要求1所述的方法,其中所述方法在设备驱动层内执行。

3.如权利要求1所述的方法,其中,准备多个硬件以连接多个节点之间的一个或多个点对点链路进一步包括:

从所述多个硬件顶端预留多个缓冲器资源。

4.如权利要求3所述的方法,其中所述从所述多个硬件顶端预留多个缓冲器资源是使用单根输入/输出虚拟化接口来进行的。

5.如权利要求1所述的方法,其中来自该多个硬件的多个缓冲器资源的预留被执行一次。

6.如权利要求1所述的方法,进一步包括:

利用信道初始化仲裁器功能以确保所述多个连接内所有的多个信道已经被初始化。

7.如权利要求1所述的方法,其中,协商和连接所述多个缓冲器包括四步队列对连接处理。

8.一种用于远程节点发现和链接多个队列对以用于通信信道实施的计算机系统,所述计算机系统包括:

一个或多个处理器、一个或多个计算机可读存储器、一个或多个计算机可读有形存储介质、以及存储在所述一个或多个有形存储介质中的至少一个上的程序指令,用于由所述一个或多个处理器中的至少一个经由所述一个或多个存储器中的至少一个执行,其中所述计算机系统能够执行一种方法,所述方法包括:

准备多个硬件以连接多个节点之间的一个或多个点对点链路;

通过发现过程发现所述一个或多个点对点链路内的所述多个节点;

在所述多个节点之间创建多个管理业务通道,所述多个管理业务通道利用不特定于信道的多个发起协议;

验证所述多个节点之间的多个连接;

在所述多个连接上发起一个或多个辅助业务流接口;

协商和连接用于所述多个连接的多个缓冲器;以及

在所述多个节点之间交换多个节点描述符。

9.如权利要求8所述的计算机系统,其中所述方法在设备驱动层内执行。

10.如权利要求8所述的计算机系统,其中准备多个硬件以连接多个节点之间的一个或多个点对点链路进一步包括:

从所述多个硬件顶端预留多个缓冲器资源。

11.如权利要求10所述的计算机系统，其中所述从所述多个硬件顶端预留多个缓冲器资源是使用单根输入/输出虚拟化接口来进行的。

12.如权利要求8所述的计算机系统，来自该多个硬件的多个缓冲器资源的预留被执行一次。

13.如权利要求8所述的计算机系统，进一步包括：

利用信道初始化仲裁器功能以确保所述多个连接内所有的多个信道已经被初始化。

14.如权利要求8所述的计算机系统，其中，协商和连接该多个缓冲器包括四步队列对连接处理。

15.一种用于远程节点发现和链接多个队列对以用于通信信道实施的计算机程序产品，所述计算机程序产品包括：

一个或多个计算机可读有形存储介质和存储在所述一个或多个有形存储介质中的至少一个之上的程序指令，所述程序指令可由处理器执行，所述程序指令指导所述处理器执行一种方法，所述方法包括：

准备多个硬件以连接多个节点之间的一个或多个点对点链路；

通过发现过程发现所述一个或多个点对点链路内的所述多个节点；

在所述多个节点之间创建多个管理业务通道，所述多个管理业务通道利用不特定于信道的多个发起协议；

验证所述多个节点之间的多个连接；

在所述多个连接上发起一个或多个辅助业务流接口；

协商和连接用于所述多个连接的多个缓冲器；以及

在所述多个节点之间交换多个节点描述符。

16.如权利要求15所述的计算机程序产品，其中所述方法在设备驱动层内执行。

17.如权利要求15所述的计算机程序产品，其中准备多个硬件以连接多个节点之间的一个或多个点对点链路进一步包括：

从所述多个硬件顶端预留多个缓冲器资源。

18.如权利要求17所述的计算机程序产品，其中所述从所述多个硬件顶端预留多个缓冲器资源是使用单根输入/输出虚拟化接口来进行的。

19.如权利要求15所述的计算机程序产品，其中来自该多个硬件的多个缓冲器资源的预留被执行一次。

20.如权利要求15所述的计算机程序产品，进一步包括：

利用信道初始化仲裁器功能以确保所述多个连接内所有的多个信道已经被初始化。

用于远程节点发现、通信信道确认和连接的方法

背景技术

[0001] 本发明总体上涉及计算领域，并且更具体地涉及计算机网络领域。

[0002] 计算机网络涉及两个或更多个计算设备与数据链路的连接以使得计算设备能够交换数据。计算机网络用作大量应用和服务的基础，例如访问互联网、云存储、打印机、传真机、电子邮件、即时消息、电话服务和数字媒体等等。随着诸如膝上型计算机和蜂窝电话之类的个人计算设备的广泛普及，计算机网络已经成为现代生活的普遍且至关重要的要素。这样，在计算机联网技术的维护和改进方面的进步具有为现代社会产生普遍益处的潜力。

发明内容

[0003] 根据一个方面，提供了一种用于远程节点发现和链接多个队列对以用于通信信道实施的处理器实施方法，该方法包括：准备多个硬件以连接多个节点之间的一个或多个点对点链路；通过发现过程发现所述一个或多个点对点链路内的所述多个节点；在所述多个节点之间创建多个管理业务通道；验证所述多个节点之间的多个连接；在所述多个连接上发起一个或多个辅助业务流接口；协商和连接用于所述多个连接的多个缓冲器；以及在所述多个节点之间交换多个节点描述符。

[0004] 根据另一个方面，提供了一种用于远程节点发现和链接多个队列对以用于通信信道实施的计算机系统，所述计算机系统包括：一个或多个处理器；一个或多个计算机可读存储器；一个或多个计算机可读有形存储介质；以及存储在所述一个或多个有形存储介质中的至少一个上的用于由所述一个或多个处理器中的至少一个经由所述一个或多个存储器中的至少一个执行的程序指令，其中所述计算机系统能够执行一种方法，所述方法包括：准备硬件以连接多个节点之间的一个或多个点对点链路；通过发现过程发现所述一个或多个点对点链路内的所述多个节点；在所述多个节点之间创建多个管理业务通道；验证所述多个节点之间的多个连接；在所述多个连接上发起一个或多个辅助业务流接口；协商和连接用于所述多个连接的多个缓冲器；以及在所述多个节点之间交换多个节点描述符。

[0005] 根据另一个方面，提供了一种用于远程节点发现和链接多个队列对以用于通信信道实施的计算机程序产品，该计算机程序产品包括：一个或多个计算机可读有形存储介质以及存储在该一个或多个有形存储介质中的至少一个上的程序指令，这些程序指令可由处理器执行，这些程序指令指导该处理器执行一种方法，该方法包括：准备多个硬件以连接多个节点之间的一个或多个点对点链路；通过发现过程发现所述一个或多个点对点链路内的所述多个节点；在所述多个节点之间创建多个管理业务通道；验证所述多个节点之间的多个连接；在所述多个连接上发起一个或多个辅助业务流接口；协商和连接用于所述多个连接的多个缓冲器；以及在所述多个节点之间交换多个节点描述符。

[0006] 根据一个实施例，提供了用于计算机网络的方法、计算机系统和计算机程序产品。本发明可包括使用媒体接入控制(MAC)多播过程来发现远程节点可寻址性，以及在实现一个或多个通信信道时使用**InfiniBand®**(**InfiniBand®**和所有基于**InfiniBand®**的商标和标志是System I/O，Incorporated和/或其附属机构的商标或注册商标)协议来连接

队列对的组的方法，包括准备硬件以连接系统之间的点对点链路；通过发现过程发现链路另一侧的远程系统；创建管理业务通道；验证节点之间的连接；发起所述辅助业务流接口；协商和连接缓冲器；以及交换节点描述符。

附图说明

[0007] 现在将参考以下附图仅通过举例来描述本发明的优选实施例：

[0008] 图1示出根据至少一个实施例的示例性联网计算机环境；

[0009] 图2是根据至少一个实施例的示出节点发现和仲裁过程的操作流程图；

[0010] 图3是根据至少一个实施例的示出链路初始化过程的功能框图；

[0011] 图4是根据至少一个实施例的图1中描述的计算机和服务器的内部和外部组件的框图；

[0012] 图5描绘了根据本发明的实施例的云计算环境；以及

[0013] 图6描绘了根据本发明的实施例的抽象模型层。

[0014] 附图的不同特征不是按比例的，因为说明是为了清晰起见，促进本领域的技术人员结合详细说明理解本发明的优选实施例。

具体实施方式

[0015] 在此披露了所要求保护的结构和方法的实施例；然而，可以理解，所公开的实施例仅仅是对可以以不同形式实施的所要求保护的结构和方法的说明。然而，本发明可以以许多不同的形式实施，并且不应被解释为限于在此阐述的示例性实施例。在说明书中，可省略众所周知的特征和技术的细节以避免不必要地模糊所呈现的实施例。

[0016] 本发明的实施例涉及计算领域，并且更具体地涉及计算机网络。以下描述的示例性实施例提供了一种系统、方法和程序产品，该系统、方法和程序产品尤其使用媒体访问控制(MAC)多播过程来发现远程节点可寻址性，并且在实现一个或多个通信信道时使用**InfiniBand®**(**InfiniBand®**和基于所有**InfiniBand®**的商标和标志是System I/O，Incorporated和/或其附属机构的商标或注册商标)协议来连接队列对。因此，本实施例具有通过提供允许交换环境的节点发现的带内方法来改进计算机网络的技术领域的能力。

[0017] 如之前所描述的，计算机网络涉及两个或更多个计算设备与数据链路的连接以使得计算设备能够交换数据。计算机网络用作大量应用和服务的基础，例如访问互联网、云存储、打印机、传真机、电子邮件、即时消息、电话服务和数字媒体等等。随着诸如膝上型计算机和蜂窝电话之类的个人计算设备的广泛普及，计算机网络已经成为现代生活的普遍且至关重要的要素。这样，在计算机联网技术的维护和改进方面的进步具有为现代社会产生普遍益处的潜力。

[0018] 许多网络连接利用构建在可商购操作系统中的标准的一组协议。这些协议是一般性的，并且是网络连接性增长的基础。然而，这种一般性是以网络性能为代价的。对诸如网络时延之类的度量敏感的系统可选择开发非通用的、但是针对特定应用和系统的操作环境定制的定制通信协议。因此，需要一种在点对点链路中使用节点发现而不需要集中式服务器，并且基于经由初始发现建立的固定信道而不是通过使用使用动态协议创建的信道来传输数据的方法。

[0019] 根据一个实施例，通过使用多播协议发现带内的远程可寻址性，然后建立直接连接的管理路径，通过该管理路径验证和初始化远程信道定义，在两个系统之间建立一组一个或多个信道。

[0020] 在这种实施例中，首先使用如Ethernet的物理层协议，如**InfiniBand®**（**InfiniBand®**以及基于所有**InfiniBand®**的商标和标志是System I/O，Incorporated和/或其附属公司的商标或注册商标）支持传输层协议的通信适配器在两个服务器系统之间建立物理连接。通过该连接，通过建立本地多播队列对(QP)来接收包含远程系统上的连接端口的地址的多播消息，并在本地系统上发送包含连接端口的地址的多播消息，来获得远程系统的可寻址性。

[0021] 当已经获得远程端口地址时，它被用作建立与远程系统的直接连接的管理QP的参数。管理QP被用作在其上发送信息的路径，诸如信道的本地和远程定义，并且诸如包括信道的QP的数量的参数被协商。

[0022] 对于具有匹配的本地和远程定义的每个信道，管理QP路径用于定义和连接辅助QP和一组构成缓冲QP。辅助QP用于信道特定的非数据消息(诸如消息中断命令)，并且缓冲QP用于发送和接收用于信道的数据消息。

[0023] 本发明可以是以任何可能的技术细节集成水平的系统、方法、和/或计算机程序产品。计算机程序产品可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。

[0024] 计算机可读存储介质可以是有形设备，该有形设备可以保留和存储供指令执行设备使用的指令。计算机可读存储介质例如可以是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、半导体存储设备或上述的任意合适的组合。计算机可读存储介质的更具体的例子的非穷尽列表包括以下：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)、便携式致密盘只读存储器(CD-ROM)、数字通用盘(DVD)、记忆棒、软盘、诸如其上记录有指令的凹槽中的穿孔卡或凸起结构的机械编码设备、和前述的任意合适的组合。如本文中所使用的，计算机可读存储介质不应被理解为是瞬态信号本身，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，通过光纤电缆的光脉冲)、或通过导线传输的电信号。

[0025] 在此描述的计算机可读程序指令可以经由网络(例如，互联网、局域网、广域网和/或无线网络)从计算机可读存储介质下载到相应的计算/处理设备或者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、光传输纤维、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器卡或网络接口从网络接收计算机可读程序指令，并且转发所述计算机可读程序指令以存储在相应的计算/处理设备内的计算机可读存储介质中。

[0026] 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器指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、用于集成电路的配置数据、或以一种或多种编程语言的任何组合写入的源代码或目标代码，编程语言包括面向对象的编程语言，诸如Smalltalk、C++等，以及过程编程语言，诸如"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计算机上执行、部分在用户计算机上执行、

作为独立软件包执行、部分在用户计算机上部分在远程计算机上执行、或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(包括局域网(LAN)或广域网(WAN))连接到用户的计算机，或者可以连接到外部计算机(例如，通过使用互联网服务提供商的互联网)。在一些实施例中，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个性化电子电路来执行计算机可读程序指令，以便执行本发明的方面。

[0027] 在此参照根据本发明的实施例的方法、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个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

[0028] 这些计算机可读程序指令可以提供给通用计算机、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诸如指令，使得通过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在流程图和/或框图的一个或多个块中指定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，其中存储有指令的计算机可读存储介质包括一个包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各方面的指令的制造品。

[0029] 计算机可读程序指令也可以加载到计算机、其他可编程数据处理装置或其他设备上，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在流程图和/或方框图的一个或多个方框中指定的功能/动作。

[0030] 附图中的流程图和框图展示了根据本发明的不同实施例的系统、方法和计算机程序产品的可能实现方式的架构、功能和操作。对此，流程图或框图中的每个方框可以表示指令的模块、段或一部分，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。在一些替代实现方式中，在这些框中指出的功能可以不按照在这些图中指出的顺序发生。例如，连续示出的两个块实际上可基本上同时执行，或者这些块有时可按相反的顺序执行，这取决于所涉及的功能。还将注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以由执行指定功能或动作或执行专用硬件与计算机指令的组合的基于专用硬件的系统来实现。

[0031] 以下描述的示例性实施例提供了一种系统、方法和程序产品，该系统、方法和程序产品通过使用多播协议发现带内的远程可寻址性、然后建立直接连接的管理路径来在两个计算系统之间建立一个或多个通信信道的组，在该路径上远程信道定义被验证和初始化。

[0032] 参见图1，描绘了根据至少一个实施例的示例性联网计算机环境100。联网计算机环境100可以包括经由通信网络112互连的网络节点102A和102B。根据至少一个实现方式，联网计算机环境100可以包括多个网络节点102A、102B，为了简洁起见，仅示出了其中的两个。

[0033] 通信网络112可以包括不同类型的通信网络，如广域网(WAN)、局域网(LAN)、电信网络、无线网络、公共交换网络和/或卫星网络。通信网络112可包括连接，诸如有线、无线通信链路或光纤电缆。可以认识到，图1仅提供一个实现方式的图示并且不暗示关于其中可以实现不同实施例的环境的任何限制。可基于设计和实施要求对所描绘的环境作出许多修

改。

[0034] 根据本发明的一个实施例，网络节点102A可以包括处理器104A和数据存储设备106A，该数据存储设备106A被启用以托管和运行软件程序108A和节点发现和仲裁程序110A，并且经由通信网络112与网络节点102B通信。网络节点102A可以是例如移动设备、电话、个人数字助理、上网本、膝上型计算机、平板计算机、台式计算机、或能够运行程序并且接入网络的任何类型的计算设备。如将参见图5讨论的，网络节点102A可分别包括内部组件502a和外部组件504a。网络节点102A还可以在云计算服务模型中操作，诸如软件即服务(SaaS)、平台即服务(PaaS)、或者基础设施即服务(IaaS)。网络节点102A还可位于云计算部署模型中，诸如私有云、社区云、公共云或混合云。

[0035] 根据本发明的一个实施例，网络节点102B可以包括处理器104B和数据存储设备106B，该数据存储设备106B能够托管和运行软件程序108B和节点发现和仲裁程序110B，并且经由通信网络112与网络节点102A通信。网络节点102B可以是例如移动设备、电话、个人数字助理、上网本、膝上型计算机、平板计算机、台式计算机、或能够运行程序并访问网络的任何类型的计算设备。如将参见图5讨论的，网络节点102B可分别包括内部组件502a和外部组件504a。网络节点102B还可在云计算服务模型中操作，诸如软件即服务(SaaS)、平台即服务(PaaS)或基础设施即服务(IaaS)。网络节点102A还可位于云计算部署模型中，诸如私有云、社区云、公共云或混合云。

[0036] 软件程序108A、B可以是利用由节点发现和仲裁程序110A、110B建立的连接的多个更高级软件程序中的任一个。

[0037] 根据本实施例，节点发现和仲裁器程序110A、110B可以是能够执行组功能以发现远程主节点的身份并且使用多播协议并且随后出于数据信道通信的目的链接队列对的组的程序。下面参见图2进一步详细解释节点发现和仲裁器方法。在一个实施例中，节点发现和仲裁器程序110A、110B在计算节点102A、B的操作系统的设备驱动层中运行。

[0038] 现在参见图2，示出了用于远程节点发现和数据信道初始化的方法的操作流程图。在202处，节点发现和仲裁器程序110A、110B准备硬件以连接系统之间的点对点链路。该适配器初始化步骤可以分为三个阶段；阶段1可能需要节点发现和仲裁器程序110A、110B触发对由适配器采用的虚拟功能的功能级别重置，以便将适配器置于已知状态。如果不存在为适配器定义的此类信道，那么节点发现和仲裁器程序110A、110B可中止初始化。接下来，节点发现和仲裁器程序110A、110B可启用用于DMA事务以及MSI-X中断的适配器。

[0039] 适配器初始化的阶段2可能需要节点发现和仲裁器程序110A、110B建立通信信道，通过该通信信道向和从适配器传递命令。适配器初始化的阶段3可涉及节点发现和仲裁器程序110A、110B向适配器查询关键硬件配置信息，所述关键硬件配置信息可稍后在对适配器的命令中使用。此外，节点发现和仲裁器程序110A、110B注册事件队列和存储器保护表条目以变成为在适配器上被通知事件并且允许远程直接存储器存取(RDMA)事务。节点发现和仲裁器程序110A、110B此时还保留分配中的QP，所述分配中的QP然后被分成组，其中数据信道由一组QP组成。节点发现和仲裁器程序110A、110B顺序地分配QP组以提供初始化的简化并拆除流接口。一旦适配器被准备好用于通知、DMA、中断和其他此类流量，节点发现和仲裁程序110A、110B可建立流接口，发现、定时和管理和数据流量可通过所述流接口流动。

[0040] 在本发明的一个实施例中，在适配器初始化过程的阶段3期间，节点发现和仲裁器

程序110A、110B可以从适配器顶端预留缓冲器资源。缓冲池的保留是昂贵的操作;在该实施例中,在适配器初始化期间,缓冲池预留操作可仅被执行一次,这使得资源可稍后用于灵活分配。在一个实施例中,节点发现和仲裁器程序110A、110B使用被称为单根输入/输出虚拟化(SRIOV)的接口来进行缓冲池预留,SRIOV是允许诸如网络适配器的设备在不同PCIe硬件功能之间单独地访问其资源的PCI Express扩展。在此接口下,一个设备驱动器可为资源的管理所有者,且网络发现和仲裁器程序110A、110B可为访客。

[0041] 接下来,在204处,节点发现和仲裁程序110A、110B通过发现过程发现链路的另一侧上的远程系统。为了执行节点/远程主机发现,节点发现和仲裁器程序110A、110B可使用多播目的地MAC地址以及专用有效载荷。当端口出现且主机固件已被配置时,节点发现和仲裁器程序110A、110B可建立QP(也称为流接口),且可周期性地以原始以太网包的形式发送出具有多播MAC地址和有效载荷的消息。除了建立流接口之外,节点发现和仲裁器程序110A、110B可配置硬件中的MAC地址,以便知道何时转发具有指定MAC地址的业务。一旦远程主机固件启动并且运行,远程主机就可以在这个指定的MAC地址上接收以太网分组并且将有效载荷放置到存储器中。有效载荷包含关于链路的远程侧的信息,并且可用于验证连接。除了链路信息之外,为了实现网络流量的发送,节点发现和仲裁程序110A、110B可发送用于仲裁链路的哪一侧将开始较高级初始化过程的信息。在一个实施例中,该仲裁信息可以由时间戳数据组成。除了仲裁信息之外,通过有效载荷发送的信息可以包括:以太网地址、全局标识、管理QP号、系统时间协议(STP)QP、远程直接内存访问(RDMA)密钥信息、以及主机固件版本。以太网地址和组标识符(GID)可以是用于在点对点和联网交换机环境中发送定向网络流量的标识信息。节点发现和仲裁器程序110A、110B可使用QP编号和RDMA密钥信息来翻译一旦分组到达特定端点适配器则分组应去往何处。最后,固件版本可设置在控制块中,并且节点发现和仲裁器程序110A、110B可使用固件版本来识别所支持的代码功能。

[0042] 然后,在206处,节点发现和仲裁器程序110A、110B创建管理业务通道。管理业务通道可以是用于网络业务的通信信道,这些通信信道为可能需要的任何特定协议提供通用路径。这可以包括信息收集协议,诸如时间同步协议,或者sysplex宽记录协议。在备选实施例中,管理通道可以利用不特定于信道的多于一个发起协议,例如STP。

[0043] 步骤208至216包括信道初始化过程。一旦所有端口级/管理流接口已在早期步骤中建立,下一步骤就是信道初始化过程。执行这个信道初始化过程的实体(被称为信道初始化仲裁器,并且可以确定哪些信道是有效的)不处于错误状态或已经被置于阈值中,并且处于正确的架构状态。对于每个定义的信道,信道初始化仲裁器启动一个定时器,并开始信道初始化过程,该信道初始化过程包括验证连接、辅助队列对连接、缓存连接和节点描述符交换的步骤。

[0044] 在208处,节点发现和仲裁器程序110A、110B验证节点之间的连接。验证连接阶段用于确保节点发现和仲裁器程序110A、110B所发现的节点是正确的。该阶段可首先需要链路的活动侧的动作,该动作使用链路初始化的节点发现部分中的仲裁信息来建立该节点发现和仲裁器程序110A、110B;首先,节点发现和仲裁程序110A、110B将本地节点(有源侧)的信道的定义发送到远程节点(无源侧)。信道定义可以由被称为信道子系统标识(CSSID)和信道路径标识符(CHPID)的两个标识符组成。在被动侧,相对于被动侧的信道信息来交叉检查主动侧的信道信息。如果信息匹配,则可给出肯定响应并且可将无源侧的信息发送至有

源侧。如果任何信息不匹配，则可将响应代码发送回活动侧以表示错误，并且可将特殊错误代码向上发送给用户。使用管理QP发送验证连接分组。

[0045] 然后，在210，节点发现和仲裁器程序110A、110B发起信道的辅助业务流接口(AUX QP)。该流接口用于信道特定的非数据信息。特定于信道的信息包括心跳和小恢复动作以维持该信道而不影响端口上的其他信道。节点发现和仲裁器程序110A、110B使用在下一步骤中描述的相同连接过程来建立AUX QP流接口。除了建立AUX QP流接口之外，节点发现和仲裁器程序110A、110B选择先前从适配器保留的QP组，并且分配相应的控制存储器区域。节点发现和仲裁器程序110A、110B还建立一个完成队列，该完成队列用于通知主机固件一些新分组已到达为信道分配的QP中的一个且必须被处理。

[0046] 接下来，在212处，节点发现和仲裁器程序110A、110B协商和连接缓冲器。在一个实施例中，节点发现和仲裁器程序110A、110B利用如下描述的四步过程：最初，有源和无源侧的主机固件枚举每种QP，每个QP需要连接。然后，利用此信息，任一侧的固件逐一连接每个QP。为了协商要使用的QP的数量，使用管理员QP将消息从主动侧发送到被动侧，该消息包含它支持的QP的最大数量。然后，被动侧检查其具有的值，并且如果那些值相同，则被动侧将值返回到主动侧。如果这些值不相同，那么这两个值中的较小者返回到活动侧。然后，活动侧检查其接收到的QP数目，并且然后一次顺序地滚动一个协商数目的QP，从而建立每侧两个阶段中的每个QP。主动侧的QP由复位状态设置并初始化，然后主动侧通过管理QP向被动侧发送连接请求，相应的QP由复位状态初始化。一旦被动侧完成，被动侧就向主动侧返回连接响应，并且使主动QP准备好接收流量，然后准备好发送并向被动侧重新发送连接确认。然后，被动侧对其QP执行相同的两个步骤，并且向主动侧返回“连接完成”消息，其中，主动侧随后初始化下一个QP连接。

[0047] 然后，在214，节点发现和仲裁器程序110A、110B通过管理QP交换节点描述符。这向另一侧给出特定的信道信息并且使得数据流量能够开始。该信息是从主动侧到被动侧和返回的简单乒乓消息。一旦这个步骤完成，信道被初始化，并且节点发现和仲裁器程序110A、110B进行测试，以确保业务能够被发送并且信道名义上和没有错误地执行。节点发现和仲裁器程序110A、110B可以在后端调用链路初始化程序，以向软件程序108A、B通知信道的可用性。

[0048] 然后，在216，节点发现和仲裁器程序110A、110B程序确定是否已使用信道初始化仲裁器功能初始化所有定义的信道。节点发现和仲裁器程序110A、110B通过遍历所定义的信道列表来作出该确定。根据一个实现方式，如果信道初始化仲裁器功能确定所有信道已经被初始化，该方法可以终止(步骤216，“是”分支)。如果信道初始化仲裁器功能确定不是所有的信道都已经被初始化(步骤216，“否”分支)，则节点发现和仲裁器程序110A、110B可以继续至步骤208以重新开始信道初始化过程并且验证节点之间的连接。

[0049] 图3是功能框图，展示了根据至少一个实施例的链路初始化过程。在图2的步骤202中进一步描述适配器初始化304的性能。在本发明的一个实施例中，在适配器初始化304期间，在图2的步骤204和206中进一步描述端口初始化306。在图2的步骤208至216中进一步讨论信道初始化308。在信道初始化308期间，信道初始化仲裁器用于确定哪些信道有效、不处于错误状态或已被置于阈值中和处于正确架构状态。在信道初始化308之后，该过程将新创建的虚拟信道切换到软件程序108。

[0050] 可以认识到，图2-3仅提供一个实现方式的说明并且不暗示关于可以如何实施不同的实施例的任何限制。可基于设计和实施要求对所描绘的环境作出许多修改。

[0051] 图4是根据本发明的实施例的在图1中描绘的网络节点102A、102B的内部和外部组件的框图400。应当理解，图4仅提供一个实现方式的图示，并不暗示关于其中可以实现不同实施例的环境的任何限制。可基于设计和实施要求对所描绘的环境作出许多修改。

[0052] 数据处理系统402、404代表能够执行机器可读程序指令的任何电子设备。数据处理系统402、404可以表示智能电话、计算机系统、PDA或其他电子设备。可以由数据处理系统402、404表示的计算系统、环境和/或配置的示例包括但不限于个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持式或膝上型设备、多处理器系统、基于微处理器的系统、网络PC、小型计算机系统和包括任何上述系统或设备的分布式云计算环境。

[0053] 网络节点102A、102B可以包括图4中展示的内部组件402a、b和外部组件404a、b的相应组。每组内部组件402包括一个或多个处理器420、一个或多个计算机可读RAM422和一个或多个总线426上的一个或多个计算机可读ROM424以及一个或多个操作系统428和一个或多个计算机可读有形存储设备430。网络节点102A、102B上的一个或多个操作系统428、软件程序108和节点发现和仲裁器程序110A、110B存储在相应的计算机可读有形存储设备530中的一个或多个上，以供相应的处理器420中的一个或多个经由相应的RAM422(其通常包括高速缓冲存储器)执行。在图4所示的实施例中，计算机可读有形存储设备430中的每一个都是内部硬盘驱动器的磁盘存储设备。可替代地，计算机可读有形存储设备430中的每一个是半导体存储设备，诸如ROM 424、EPROM、闪存或可以存储计算机程序和数字信息的任何其他计算机可读有形存储设备。

[0054] 每组内部组件402a、b还包括R/W驱动器或接口432，以便从一个或多个便携式计算机可读有形存储设备438(诸如CD-ROM、DVD、记忆棒、磁带、磁盘、光盘或半导体存储设备)读取和向其写入。软件程序(诸如节点发现和仲裁器程序110A、110B)可存储在相应的便携式计算机可读有形存储设备438中的一个或多个上，经由相应的R/W驱动器或接口432读取，并加载到相应的硬盘驱动器430中。

[0055] 每组内部组件402a、b还包括网络适配器或接口436，如TCP/IP适配器卡、无线Wi-Fi接口卡、或3G或4G无线接口卡或其他有线或无线通信链路。可经由网络(例如，互联网、局域网或其他广域网)和相应的网络适配器或接口536从外部计算机将网络节点102A、102B中的软件程序108和节点发现和仲裁器程序110A、110B下载到网络节点102A、102B。从网络适配器或接口436，将网络节点102A、102B中的软件程序108和节点发现和仲裁器程序110A加载到相应的硬盘驱动器430中。网络可以包括铜线、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。

[0056] 每组外部组件404a、b可以包括计算机显示监视器444、键盘442和计算机鼠标434。外部组件404a、b还可包括触摸屏、虚拟键盘、触摸板、定点设备和其他人机接口设备。每组内部组件402a、b还包括设备驱动器440以接口到计算机显示监视器444、键盘442和计算机鼠标434。设备驱动器440、R/W驱动器或接口432和网络适配器或接口436包括硬件和软件(存储在存储设备430和/或ROM424中)。

[0057] 首先应当理解，尽管本公开包括关于云计算的详细描述，但其中记载的技术方案的实现却不限于云计算环境，而是能够结合现在已知或以后开发的任何其它类型的计算环

境而实现。

[0058] 云计算是一种服务交付模式，用于对共享的可配置计算资源池进行方便、按需的网络访问。可配置计算资源是能够以最小的管理成本或与服务提供者进行最少的交互就能快速部署和释放的资源，例如可以是网络、网络带宽、服务器、处理、内存、存储、应用、虚拟机和服务。这种云模式可以包括至少五个特征、至少三个服务模型和至少四个部署模型。

[0059] 特征包括：

[0060] 按需自助式服务：云的消费者在无需与服务提供者进行人为交互的情况下能够单方面自动地按需部署诸如服务器时间和网络存储等的计算能力。

[0061] 广泛的网络接入：计算能力可以通过标准机制在网络上获取，这种标准机制促进了通过不同种类的瘦客户机平台或厚客户机平台(例如移动电话、膝上型电脑、个人数字助理PDA)对云的使用。

[0062] 资源池：提供者的计算资源被归入资源池并通过多租户(multi-tenant)模式服务于多重消费者，其中按需将不同的实体资源和虚拟资源动态地分配和再分配。一般情况下，消费者不能控制或甚至并不知晓所提供的资源的确切位置，但可以在较高抽象程度上指定位置(例如国家、州或数据中心)，因此具有位置无关性。

[0063] 迅速弹性：能够迅速、有弹性地(有时是自动地)部署计算能力，以实现快速扩展，并且能迅速释放来快速缩小。在消费者看来，用于部署的可用计算能力往往显得是无限的，并能在任意时候都能获取任意数量的计算能力。

[0064] 可测量的服务：云系统通过利用适于服务类型(例如存储、处理、带宽和活跃用户帐号)的某种抽象程度的计量能力，自动地控制和优化资源效用。可以监测、控制和报告资源使用情况，为服务提供者和消费者双方提供透明度。

[0065] 服务模型如下：

[0066] 软件即服务(SaaS)：向消费者提供的能力是使用提供者在云基础架构上运行的应用。可以通过诸如网络浏览器的瘦客户机接口(例如基于网络的电子邮件)从各种客户机设备访问应用。除了有限的特定于用户的应用配置设置外，消费者既不管理也不控制包括网络、服务器、操作系统、存储、乃至单个应用能力等的底层云基础架构。

[0067] 平台即服务(PaaS)：向消费者提供的能力是在云基础架构上部署消费者创建或获得的应用，这些应用利用提供者支持的程序设计语言和工具创建。消费者既不管理也不控制包括网络、服务器、操作系统或存储的底层云基础架构，但对其部署的应用具有控制权，对应用托管环境配置可能也具有控制权。

[0068] 基础架构即服务(IaaS)：向消费者提供的能力是消费者能够在其中部署并运行包括操作系统和应用的任意软件的处理、存储、网络和其他基础计算资源。消费者既不管理也不控制底层的云基础架构，但是对操作系统、存储和其部署的应用具有控制权，对选择的网络组件(例如主机防火墙)可能具有有限的控制权。

[0069] 部署模型如下：

[0070] 私有云：云基础架构单独为某个组织运行。云基础架构可以由该组织或第三方管理并且可以存在于该组织内部或外部。

[0071] 共同体云：云基础架构被若干组织共享并支持有共同利害关系(例如任务使命、安全要求、政策和合规考虑)的特定共同体。共同体云可以由共同体内的多个组织或第三方管

理并且可以存在于该共同体内部或外部。

[0072] 公共云:云基础架构向公众或大型产业群提供并由出售云服务的组织拥有。

[0073] 混合云:云基础架构由两个或更多部署模型的云(私有云、共同体云或公共云)组成,这些云依然是独特的实体,但是通过使数据和应用能够移植的标准化技术或私有技术(例如用于云之间的负载平衡的云突发流量分担技术)绑定在一起。

[0074] 云计算环境是面向服务的,特点集中在无状态性、低耦合性、模块性和语意的互操作性。云计算的核心是包含互连节点网络的基础架构。

[0075] 现在参考图5,其中显示了示例性的云计算环境60。如图所示,云计算环境50包括云计算消费者使用的本地计算设备可以与其相通信的一个或者多个云计算节点100,本地计算设备例如可以是个人数字助理(PDA)或移动电话54A,台式电脑54B、笔记本电脑54C和/或汽车计算机系统54N。云计算节点100之间可以相互通信。可以在包括但不限于如上所述的私有云、共同体云、公共云或混合云或者它们的组合的一个或者多个网络中将云计算节点10进行物理或虚拟分组(图中未显示)。这样,云的消费者无需在本地计算设备上维护资源就能请求云计算环境50提供的基础架构即服务(IaaS)、平台即服务(PaaS)和/或软件即服务(SaaS)。应当理解,图5显示的各类计算设备54A-N仅仅是示意性的,云计算节点100以及云计算环境50可以与任意类型网络上和/或网络可寻址连接的任意类型的计算设备(例如使用网络浏览器)通信。

[0076] 现在参考图6,其中显示了云计算环境50提供的一组功能抽象层600。首先应当理解,图6所示的组件、层以及功能都仅仅是示意性的,本发明的实施例不限于此。如图6所示,提供下列层和对应功能:

[0077] 硬件和软件层60包括硬件和软件组件。硬件组件的例子包括:主机61;基于RISC(精简指令集计算机)体系结构的服务器62;服务器63;刀片服务器64;存储设备65;网络和网络组件66。软件组件的例子包括:网络应用服务器软件67以及数据库软件68。

[0078] 虚拟层70提供一个抽象层,该层可以提供下列虚拟实体的例子:虚拟服务器71、虚拟存储72、虚拟网络73(包括虚拟私有网络)、虚拟应用和操作系统74,以及虚拟客户端75。

[0079] 在一个示例中,管理层80可以提供下述功能:资源供应功能81:提供用于在云计算环境中执行任务的计算资源和其它资源的动态获取;计量和定价功能82:在云计算环境内对资源的使用进行成本跟踪,并为此提供帐单和发票。在一个例子中,该资源可以包括应用软件许可。安全功能:为云的消费者和任务提供身份认证,为数据和其它资源提供保护。用户门户功能83:为消费者和系统管理员提供对云计算环境的访问。服务水平管理功能84:提供云计算资源的分配和管理,以满足必需的服务水平。服务水平协议(SLA)计划和履行功能85:为根据SLA预测的对云计算资源未来需求提供预先安排和供应。

[0080] 工作负载层90提供云计算环境可能实现的功能的示例。在该层中,可提供的工作负载或功能的示例包括:地图绘制与导航91;软件开发及生命周期管理92;虚拟教室的教学提供93;数据分析处理94;交易处理95;以及节点发现和仲裁器方法96。节点发现和仲裁器方法96可能与使用媒体访问控制(MAC)多播过程来发现远程节点可寻址性,以及**InfiniBand®**(**InfiniBand®**和所有基于**InfiniBand®**的商标和标记是System I/O,Incorporated和/或其分支机构的商标或注册商标)协议有关,以在实施一个或多个通信信道时连接队列对的组。

[0081] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

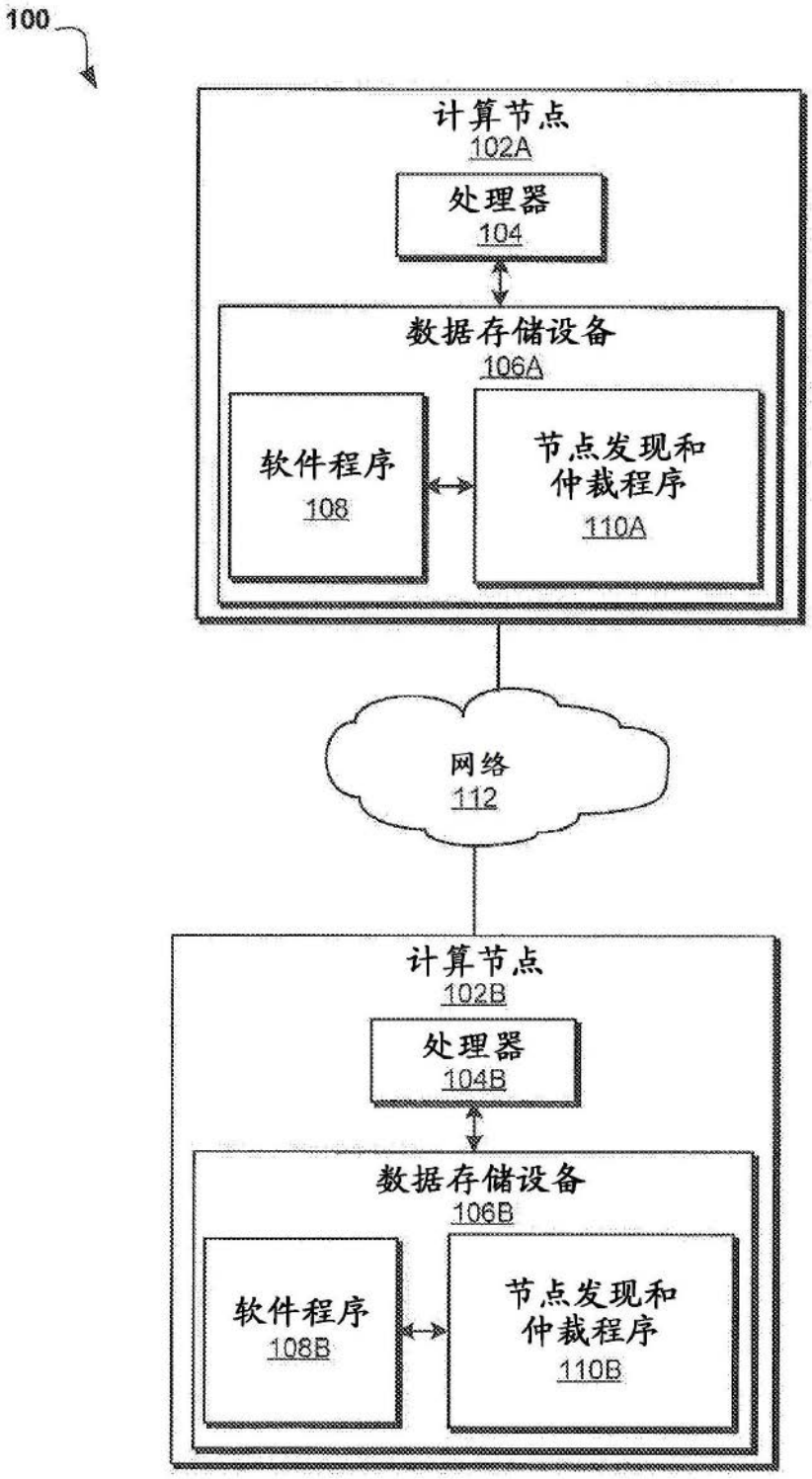

100
计算节点
102A
处理器
104
数据存储设备
106A
软件程序
108
节点发现和
仲裁程序
110A
网络
112
计算节点
102B
处理器
104B
数据存储设备
106B
软件程序
108B
节点发现和
仲裁程序
110B

图1

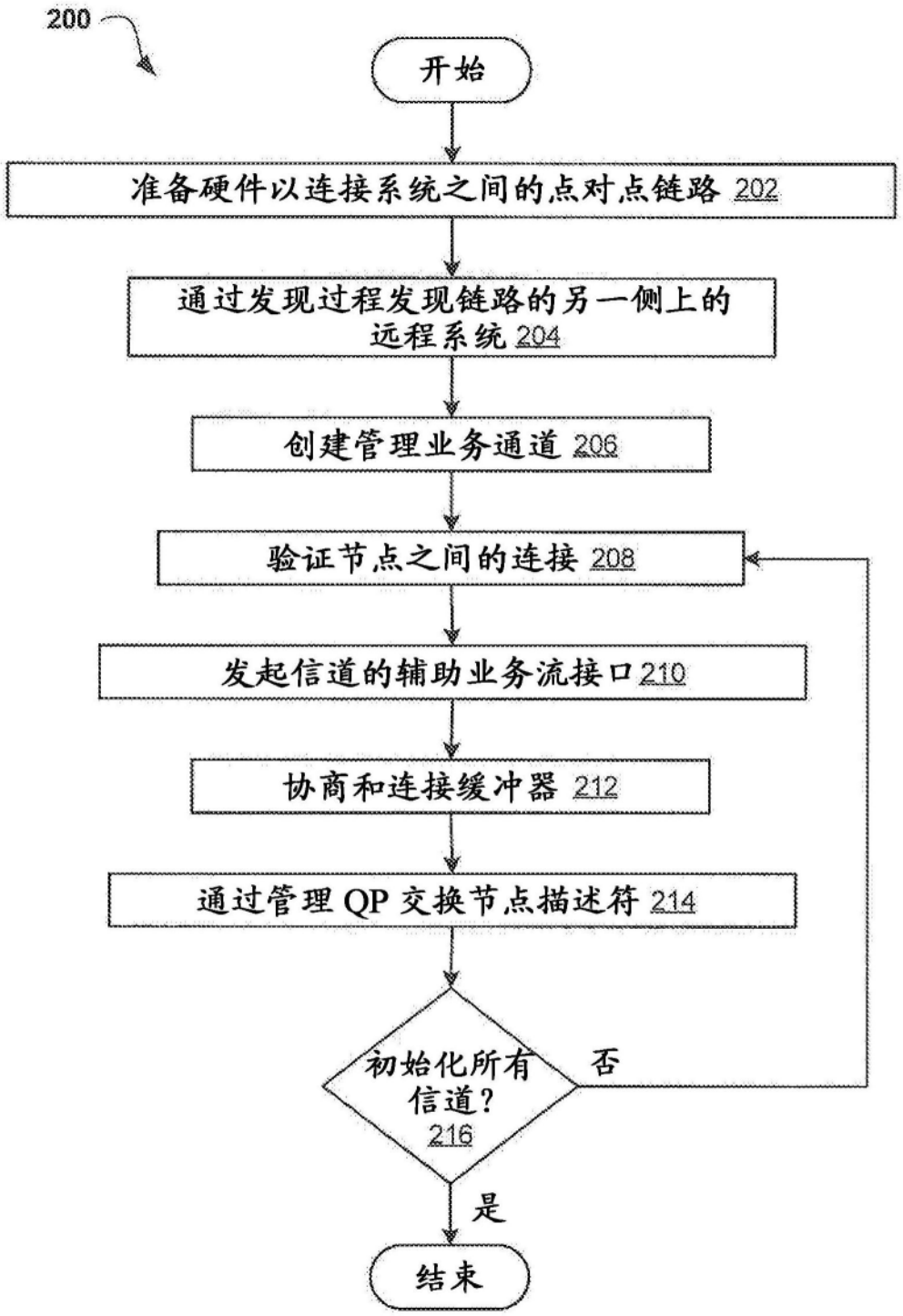

200
开始
准备硬件以连接系统之间的点对点链路 202
通过发现过程发现链路的另一侧上的远程系统 204
创建管理业务通道 206
验证节点之间的连接 208
发起信道的辅助业务流接口 210
协商和连接缓冲器 212
通过管理 QP 交换节点描述符 214
初始化所有信道? 216
否
是
结束

图2

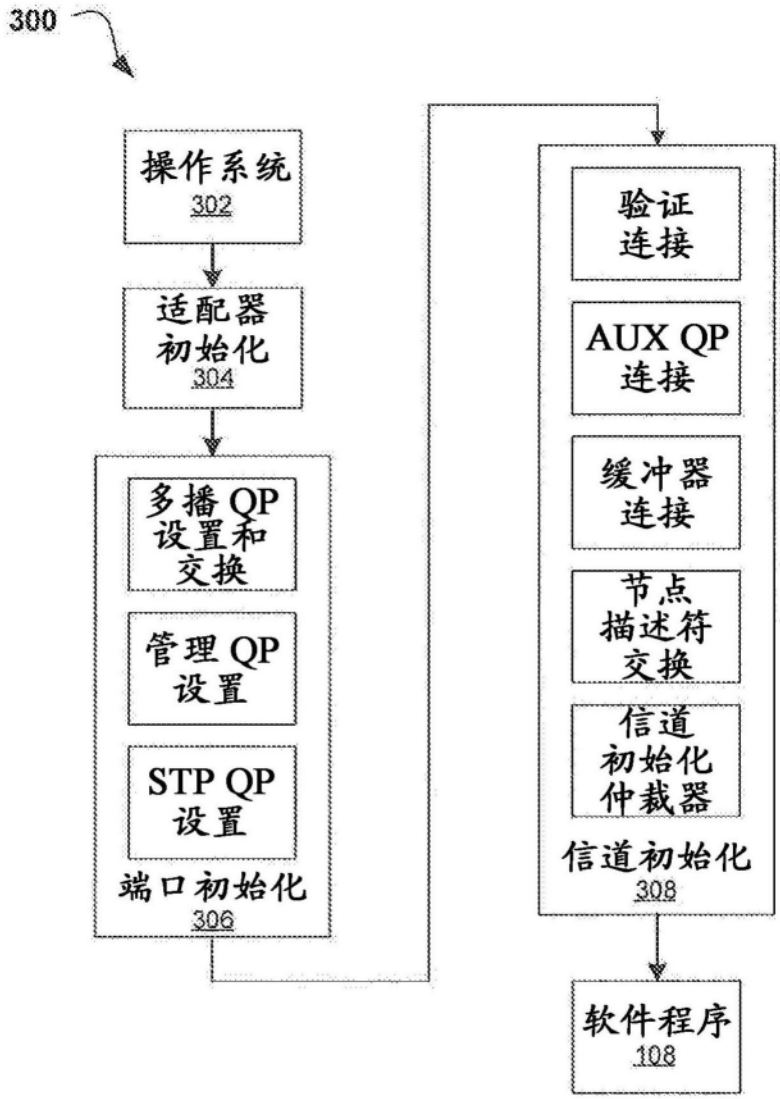

300
操作系统
302
适配器
初始化
304
多播 QP
设置和
交换
管理 QP
设置
STP QP
设置
端口初始化
306
验证
连接
AUX QP
连接
缓冲器
连接
节点
描述符
交换
信道
初始化
仲裁器
信道初始化
308
软件程序
108

图3

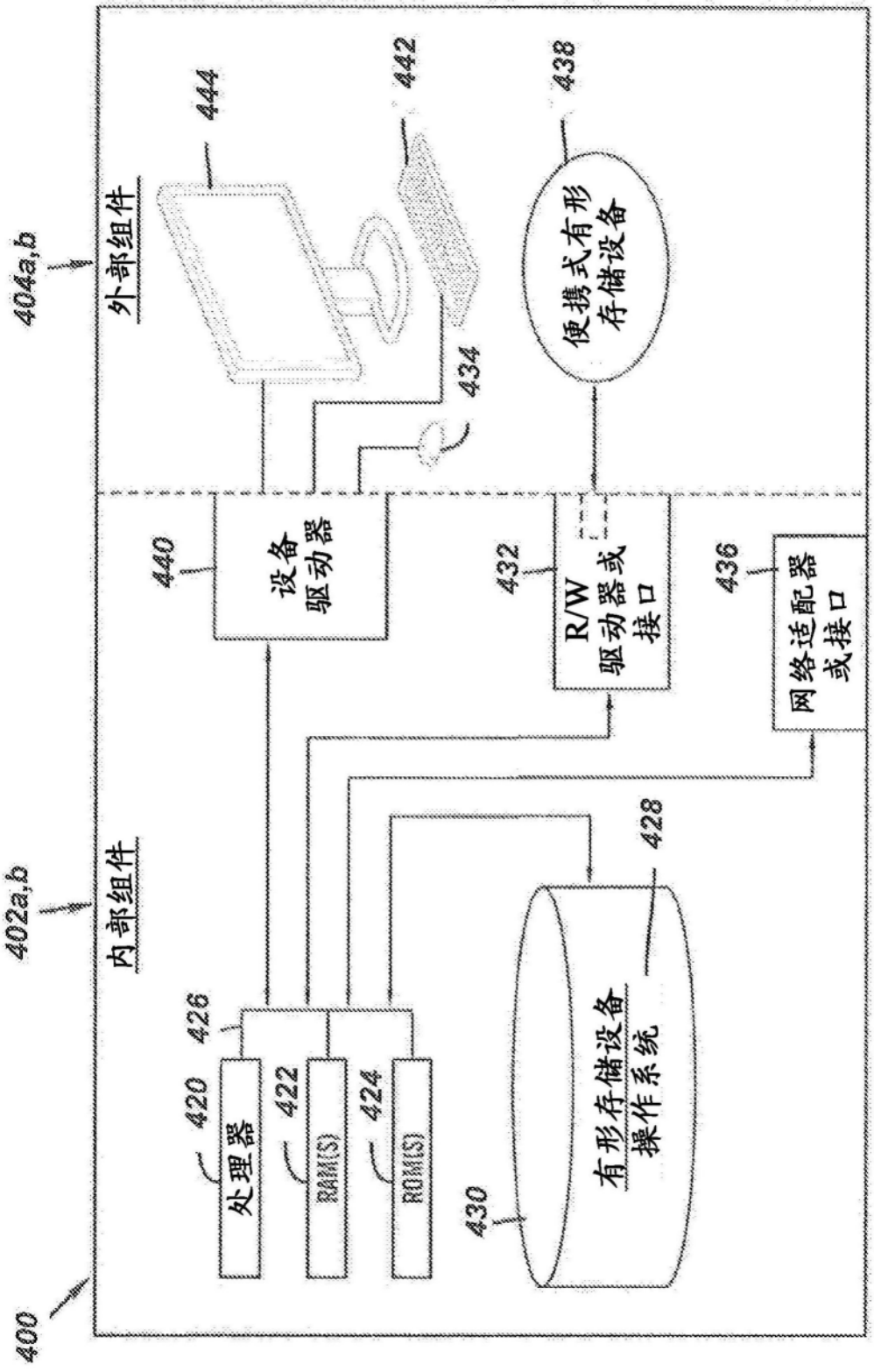

400
402a,b
内部组件
404a,b
外部组件
420
处理器
422
RAM(S)
424
ROM(S)
426
428
有形存储设备
操作系统
430
432
R/W
驱动器或
接口
434
436
网络适配器
或接口
438
便携式有形
存储设备
440
设备
驱动器
442
444

图4

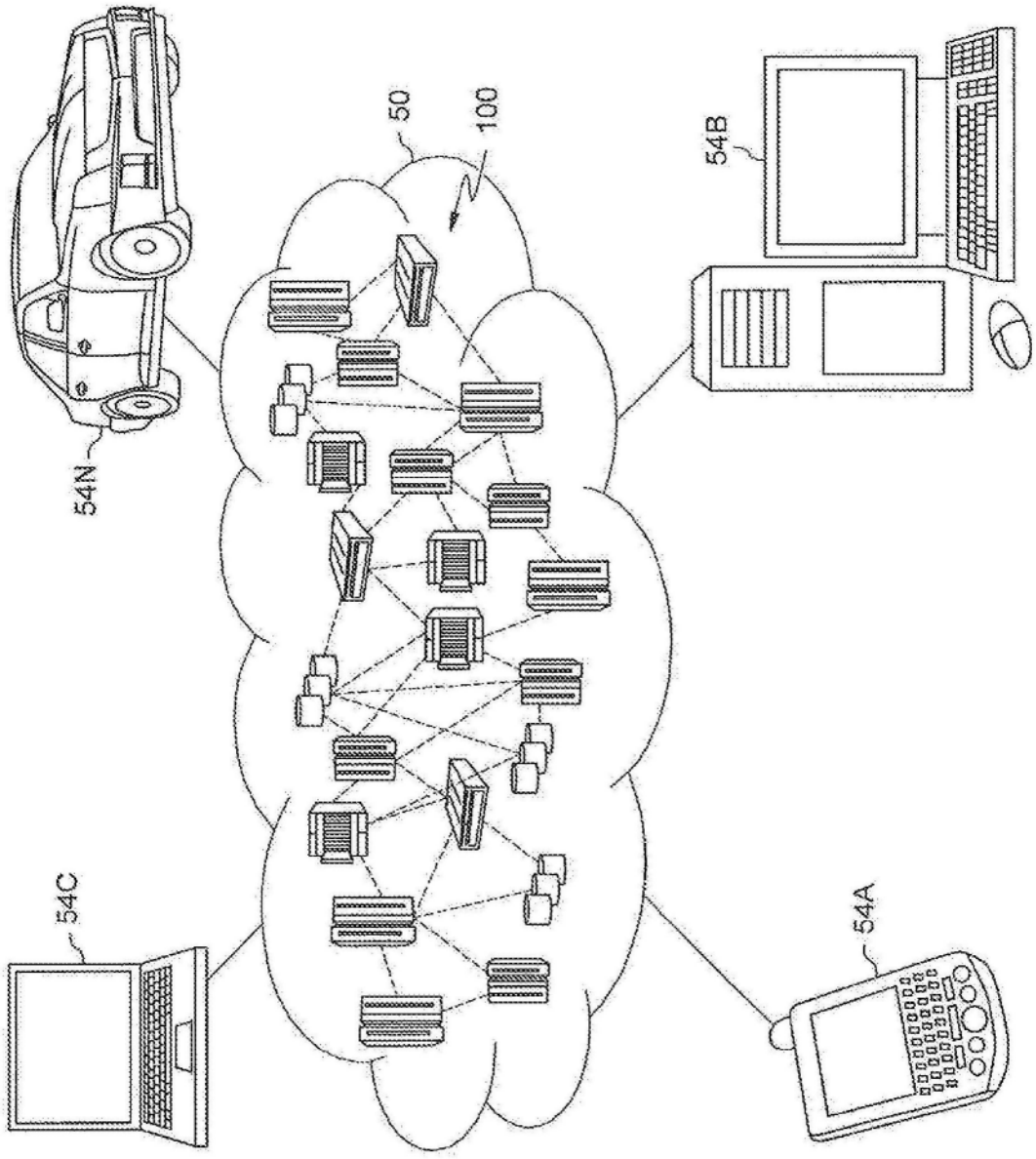


图5

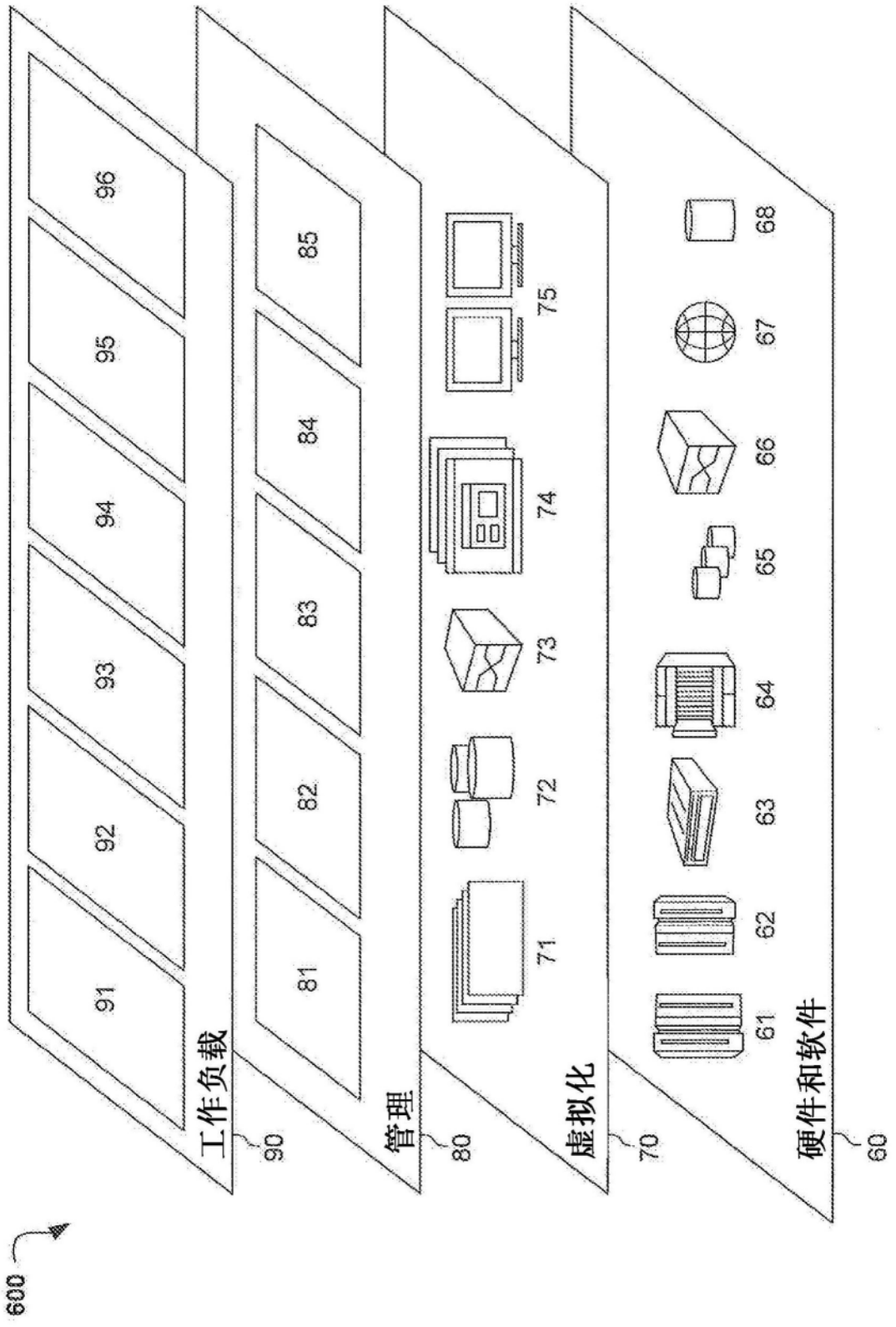

600
工作负载
90
91
92
93
94
95
96
管理
80
81
82
83
84
85
虚拟化
70
71
72
73
74
75
硬件和软件
60
61
62
63
64
65
66
67
68

图6