(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 111201756 A
(43)申请公布日 2020.05.26

(21)申请号 201880066224.9

(22)申请日 2018.10.12

(30)优先权数据
2982147 2017.10.12 CA

(85)PCT国际申请进入国家阶段日
2020.04.10

(86)PCT国际申请的申请数据
PCT/IB2018/057945 2018.10.12

(87)PCT国际申请的公布数据
WO2019/073452 EN 2019.04.18

(71)申请人 洛克波特网络股份有限公司
地址 加拿大安大略渥太华

(72)发明人 M.R.威廉斯

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 高巍

(51)Int.Cl.
H04L 12/66(2006.01)

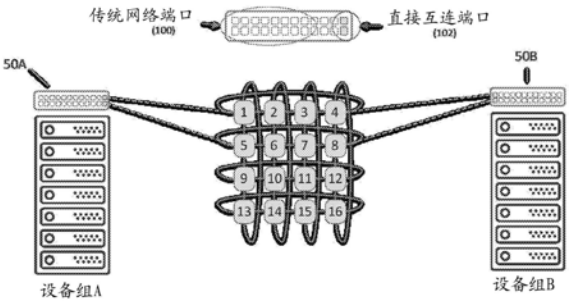
权利要求书3页 说明书9页 附图15页

(54)发明名称

直接互连网关

(57)摘要

一种专用网络网关设备，其能够在传统网络和直接互连网络之间桥接、交换或路由网络业务，该设备包括：第一组一个或多个传统网络端口，每个端口具有单个链路，通常包括一个或多个SFP+、QSFP和QSFP+连接器，这样的端口连接到形成传统网络的交换机或设备；以及，第二组一个或多个直接互连端口，每个端口具有大量链路(两个或更多)，通常包括MXC、MTP和MTO连接器中的一个或多个，这样的端口连接到直接互连网络。



CN 111201756 A

1.一种专用网络网关设备(50),其能够在传统网络和直接互连网络之间桥接、交换或路由网络业务,所述设备包括:

第一组至少一个传统网络端口(100),每个端口具有单个链路,以及

第二组至少一个直接互连端口(102),每个端口具有至少两个链路;

其中所述第一组传统网络端口(100)连接到形成传统网络的交换机和设备中的一个;且

其中第二组至少一个直接互连端口(102)连接到直接互连网络。

2.根据权利要求1所述的设备(50),其中,所述第一组传统网络端口(100)包括SFP+、QSFP和QSFP+连接器中的至少一个。

3.根据权利要求1所述的设备(50),其中,所述第二组至少一个直接互连端口(102)包括MXC、MTP和MTO连接器中的至少一个。

4.根据权利要求1所述的设备(50),其中,所述传统网络端口(100)连接到所述传统网络中的交换机端口和路由器端口中的至少一个。

5.根据权利要求1所述的设备(50),其中,所述第二组直接互连端口(102)连接到在所述直接互连网络的实现中使用的无源配线架/集线器(60)。

6.根据权利要求1所述的设备(50),其中,所述第二组直接互连端口(102)各自连接到它们自己的专用直接互连专用集成电路(ASIC)104、现场可编程门阵列(FPGA)、通用处理器、网络处理器和能够充当直接互连网络中的节点的设备中的至少一个。

7.根据权利要求1所述的设备(50),其中,所述第二组直接互连端口(102)各自连接到共享专用集成电路(ASIC)、现场可编程门阵列(FPGA)、通用处理器、网络处理器,和能够充当直接互连网络中的节点的设备中的至少一个。

8.根据权利要求1所述的设备(50),其中,所述桥接、交换和路由功能中的至少一项由网络交换机ASIC、现场可编程门阵列(FPGA)、通用处理器、网络处理器和能够执行网络业务转发的设备中的至少一个执行。

9.根据权利要求1所述的设备(50),其中,所述桥接、交换和路由功能中的至少一项由网络控制器ASIC、现场可编程门阵列(FPGA)、通用处理器、网络处理器和能够执行网络业务转发的设备中的至少一个执行。

10.根据权利要求9所述的设备(50),其中,所述ASIC、FPGA、通用处理器、网络处理器和节点中的至少一个能够充当直接互连节点,其具有在本地通过传统网络接口线发送的目的地/源业务。

11.根据权利要求1所述的设备(50),其中,所述直接互连端口代替所述直接互连网络内的设备。

12.根据权利要求1所述的设备(50),其中,所述第二组直接互连端口连接到多个直接互连网络。

13.一种专用网络网关设备(50),其能够在传统网络和直接互连网络之间进行桥接、交换和路由网络业务中的至少一个,所述设备包括:

第一端口,所述第一端口是被配置为连接到直接互连网络的直接互连端口,和

第二端口,所述第二端口是被配置为连接到形成传统网络的交换机和设备中的至少一个的标准网络端口。

14.根据权利要求13所述的设备(50),其中,所述第一端口包括MXC、MTP和MTO连接器中的至少一个。

15.根据权利要求13所述的设备(50),其中,所述第二端口包括SFP+、QSFP和QSFP+连接器中的至少一个。

16.根据权利要求13所述的设备(50),其中,所述第一端口连接到在直接互连网络的实现中使用的无源配线架/集线器。

17.根据权利要求13所述的设备(50),其中,所述第二端口连接到所述传统网络中的交换机端口和路由器端口中的至少一个。

18.一种专用网络网关设备(50),用于在传统网络与直接互连网络之间进行桥接和路由网络业务中的至少一个,所述设备包括:

第一组传统网络端口(100),每个端口具有单个链路,这样的端口连接到形成第一传统网络的末端设备;和

第二组直接互连端口(102),每个端口具有两个或更多链路,这样的端口连接到所述直接互连网络,

其中,所述直接互连网络充当骨干网,其允许网络业务从专用网络网关设备路由(50)路由到另一个专用网络网关设备(50),所述另一个专用网络网关设备(50)包括:

第一组传统网络端口(100),每个端口具有单个链路,这样的端口连接到形成第二传统网络的末端设备;和

第二组直接互连端口(102),每个端口具有两个或更多链路,这样的端口连接到直接互连网络。

19.一种专用网络网关设备(50),用于在传统网络和直接互连网络之间进行桥接、交换和路由网络业务中的至少一个,所述设备包括:

第一组至少一个传统网络端口(100),每个端口具有单个链路,这样的端口连接到形成传统网络的交换机和设备中的至少一个;

第二组至少一个直接互连端口(102),每个端口具有一个或多个链路,这样的端口连接到直接互连网络;以及

逻辑上相关联以充当单个直接互连节点的多个直接互连端口。

20.一种用于在传统网络和直接互连网络之间进行桥接、交换和路由网络业务中的至少一个的计算机实现的方法,所述方法包括以下步骤:

将权利要求1所述的专用网关设备(50)连接到传统网络以及直接互连网络,所述专用网关设备(50)充当所述直接互连网络内的一个或多个节点;并且

基于所述网络业务的标头或内容,通过传统网络与直接互连网络之间的网关设备(50)来转发所述网络业务。

21.一种用于在可由多于一个的网关设备(50)访问位于传统网络中的资源时协调哪个网关设备(50)具有提供对所述资源进行访问的能力的计算机实现的方法,该方法包括以下步骤:

在直接互连端口上接收ARP、广播、多播或选播业务,其中,所述业务正请求访问位于所述传统网络中的所述资源,并且其中,所述直接互连端口经由一个或多个跃点链接到所述多于一个的网关设备(50),每一个网关设备都能够提供对所述资源的访问;

从应该提供对所述资源的访问的所述多于一个的网关设备中计算出最佳网关设备端口；

在所述业务、所述直接互连节点和所计算出的最佳网关设备端口之间创建关联；以及

将所述关联传达给所述多于一个的网关设备中的每一个，以确保所计算出的最佳网关设备端口提供对所述资源的访问。

22.根据权利要求21所述的计算机实现的方法，其中，计算应当提供对所述资源的访问的所述最佳网关设备端口的步骤包括：确定所述多于一个的网关设备端口中的哪个最接近所述直接互连端口。

23.根据权利要求21所述的计算机实现的方法，其中，计算应当提供对所述资源的访问的所述最佳网关设备端口的步骤包括：采用共识算法来确保所述业务的一致性。

24.根据权利要求21所述的计算机实现的方法，其中，传达所述关联的步骤通过专用或共享协调总线来处理。

直接互连网关

技术领域

[0001] 本发明涉及计算机网络中的I/O(输入/输出)业务。更具体地,本发明涉及在传统网络和直接互连网络之间桥接、交换或路由数据的专用设备。

背景技术

[0002] 计算机网络允许许多节点相互路由或另外地交换数据。因此,计算机网络能够支持大量的应用和服务,例如共享使用存储服务器、访问万维网、使用电子邮件等。

[0003] 节点本身的特征经常可以基于它们执行的专门任务而分为三类:计算节点,诸如具有执行计算的CPU的服务器(但通常只有很少甚至没有本地磁盘空间);包含系统辅助存储并提供并行文件系统服务的I/O节点;以及,提供与外部数据服务器和大容量存储系统的连接性的网关节点。一些节点甚至可以提供一种以上的功能,例如处理I/O及网关功能。

[0004] 然而,用于并行和分布式系统的I/O已引起计算机系统的用户和设计者的极大关注。在这方面,虽然CPU的速度实际上每年都以指数级的速度增长,但不幸的是I/O设备的速度却在以较慢的步调增长,这经常是由于它们可能更加受限于机械组件的速度。因此I/O性能(衡量节点之间的I/O数据业务)经常是网络性能的限制因素。确实,在并行和分布式计算机系统中,CPU和I/O之间速度的不匹配加剧了,这使I/O成为严重限制可伸缩性的瓶颈。当网络牵涉涉及例如多媒体和科学建模的商业应用,每个应用都对I/O要求很大的时候,尤其会是这种情况。

[0005] 直接互连网络,例如在第WO 2015/027320A1号PCT专利申请公开中披露的网络(其描述了一种新颖的环面或更高基数的互连拓扑,用于在并行计算机系统中以网孔状方式连接网络节点),其业务通常限制于节点,节点是直接互连的一部分。尽管在第WO 2015/027320A1号PCT专利申请公开中披露的这种新颖的系统和架构对于数据中心和云数据中心中的商业部署特别有益和实用,但不幸的是当今运行的大多数数据中心仍然基于传统三层架构、胖树架构,或以DCell服务器为中心的架构等。对于基于这些架构的数据中心,不幸的是并不希望也不可能将它们加入直接互连,因此它们无法发挥这种网络拓扑的优势。一些现有技术的直接互连架构已经提供了一种系统,其中每个节点或节点的子集(即网关节点)都具有与直接互连和与传统网络二者的双重连通性,但是这样的节点在两个网络之间桥接或路由时难于管理和加载设备资源。

[0006] 因此可期望的是具有一种直接互连网关,其被设计成并且能够允许直接互连设备和非直接互连设备进行通信。而且,具有可有助于克服一些上述针对I/O业务的缺点的网关将是有益的。

发明内容

[0007] 在一方面,本发明提供了一种专用设备,即网关设备,其能够在传统网络和直接互连网络之间桥接、交换或路由。

[0008] 在另一方面,本发明提供了一种高度可管理的网关设备,其可以由网络管理系统

来管理。

[0009] 在另一方面，本发明提供了一种网关设备，该网关设备允许在多个直接互连端口之间协调MAC表和ARP、广播、多播和选播响应。

[0010] 在一个实施例中，本发明提供了一种专用网络网关设备，其能够在传统网络和直接互连网络之间桥接、交换或路由网络业务。该设备包括：第一组一个或多个传统网络端口，每个端口具有单个链路，这样的端口连接到形成传统网络的交换机或设备；以及第二组一个或多个直接互连端口，每个端口具有两个或更多链路，这样的端口连接到直接互连网络。传统网络端口可以包括SFP+、QSFP和QSFP+连接器中的一个或多个，并且可以连接到传统网络中的交换机端口或路由器端口，而直接互连端口可以包括MXC、MTP和MTO连接器中的一个或多个，并且可以连接到在直接互连网络的实现中使用的无源配线架/集线器。备选地，直接互连端口可以各自连接到它们自己的专用直接互连专用集成电路(ASIC)，或者它们可以各自连接到一个或多个共享专用集成电路(ASIC)。桥接、交换或路由功能可以由网络交换机ASIC或网络控制器ASIC执行。ASIC可以能够充当直接互连节点，具有在本地通过传统网络接口线发送的目的地/源业务。另外，在其他实施例中，直接互连端口可以代替直接互连网络内的设备，而且它们甚至可以连接到多个直接互连网络。

[0011] 在另一个实施例中，本发明提供了一种专用网络网关设备，该设备能够在传统网络和直接互连网络之间桥接、交换或路由网络业务，其中，所述设备包括两个端口，即第一端口，就是能够连接到直接互连网络的直接互连端口；以及第二端口，就是可以连接到形成传统网络的交换机或设备的标准网络端口。第一端口可以包括MXC、MTP或MTO连接器中的一个，并且可以连接到在直接互连网络的实现中使用的无源配线架/集线器。第二端口可以包括SFP+、QSFP或QSFP+连接器之一，并且可以连接到传统网络中的交换机端口或路由器端口。

[0012] 在再一个实施例中，本发明提供了一种专用网络网关设备，其能够在传统网络和直接互连网络之间桥接或路由网络业务，该设备包括：第一组传统网络端口，每个端口具有单个链路，这样的端口连接到形成第一传统网络的末端设备；第二组直接互连端口，每个端口具有两个或更多链路，这样的端口连接到直接互连网络，其中，所述直接互连网络充当骨干，其允许网络业务从专用网络网关设备路由到另一个专用网络网关设备，所述另一专用网络网关设备包括：第一组传统网络端口，每个端口具有单个链路，这样的端口连接到形成第二传统网络的末端设备；以及第二组直接互连端口，每个端口具有两个或更多链路，这样的端口连接到直接互连网络。

[0013] 在又一个实施例中，本发明提供了一种专用网络网关设备，其能够在传统网络和直接互连网络之间桥接、交换或路由网络业务，该设备包括：第一组一个或多个传统网络端口，每个端口具有单个链路，这样的端口连接到形成传统网络的交换机或设备；第二组一个或多个直接互连端口，每个端口具有一个或多个链路，这样的端口连接到直接互连网络；以及，逻辑上相关联以充当单个直接互连节点的多个直接互连端口。

[0014] 在另一个实施例中，本发明提供了在传统网络和直接互连网络之间桥接、交换或路由网络业务的计算机实现的方法，该方法包括以下步骤：将专用网关设备连接到传统网络和直接互连网络，所述专用网关设备充当直接互连网络内的一个或多个节点；以及，基于网络业务的标头或内容，通过传统网络与直接互连网络之间的网关设备转发网络业务。

[0015] 在进一步的实施例中，本发明提供了一种计算机实现的方法，该方法在多于一个的网关设备都可访问位于传统网络中的资源时协调哪个网关设备应提供对所述资源的访问，该方法包括以下步骤：(i)在直接互连端口上接收ARP、广播、多播或选播业务，其中所述业务正在请求访问位于传统网络中的资源，并且其中所述直接互连端口经由一个或多个跃点链接到多于一个网关设备，每个网关设备都能够提供对该资源的访问；(ii)计算应提供对该资源的访问的多于一个的网关设备中的最佳网关设备端口；(iii)在业务、直接互连节点和计算出的最佳网关设备端口之间创建关联；以及(iv)与多于一个的网关设备中的每一个传达该关联，以确保所计算出的最佳网关设备端口提供对该资源的访问。计算应提供对该资源的访问的最佳网关设备端口的步骤可以包括确定多于一个的网关设备端口中的哪一个最接近直接互连端口，或者可以包括采用共识算法来确保业务的一致性。传达该关联的步骤可以由专用或共享的协调总线来处理。

附图说明

[0016] 现在将参考附图通过示例的方式描述本发明的实施例，其中：

[0017] 图1是根据本发明的一个实施例的网关设备的高层总览图。

[0018] 图2是根据本发明的一个实施例的网关设备的总览图，其包括传统网络交换机ASIC，所配备的某些端口连接到传统网络连接器，而其他端口连接到它们自己的专用直接互连ASIC。

[0019] 图2b是根据本发明的一个实施例的网关设备的总览图(关于图2所示)，其中每个直接互连ASIC可以连接到多于一个的直接互连端口，并且可以连接到多于一个的传统网络交换机ASIC端口。

[0020] 图2c是根据本发明的一个实施例的网关设备的总览图，其包括单个交换机和直接互连ASIC，将传统网络交换机ASIC和一个或多个直接互连ASIC的功能相结合。

[0021] 图2d是根据本发明的一个实施例的网关设备的总览图，其包括主机接口卡，该主机接口卡包含传统的网络端口和直接互连端口。

[0022] 图2e是根据本发明的一个实施例的网关设备的总览图，其包括主机接口卡，在此直接互连ASIC和传统网络控制器ASIC的功能被组合为单个直接互连和传统ASIC。

[0023] 图2f是根据本发明的一个实施例的网关设备的总览图，其中直接互连端口每个端口包含一个或多个链路，并且由网关设备组合成组，使得每个组在逻辑上都是由网关设备关联以充当直接互连网络中的单个节点。

[0024] 图3是根据本发明的一个实施例的网关设备的总览图，其中直接互连端口连接到无源配线架/集线器。

[0025] 图4是根据本发明的一个实施例的网关设备的总览图，其中直接互连端口代替了直接互连拓扑内的服务器。

[0026] 图4b显示了如图4所示的实施例，其中直接互连端口的链路连接到不同的节点。

[0027] 图5是本发明的一个实施例的总览图，其中网关上的直接互连端口链接到不同的直接互连，以允许在多个直接互连之间进行桥接、交换或路由。

[0028] 图6是本发明的一个实施例的总览图，其中直接互连内的所有或大多数节点由网关端口组成，并且其中所有或大多数其他设备将直接连接至一个或多个更多的网关。

[0029] 图7提供了如何最小化从直接互连中的每个设备到最近的网关端口的平均距离(以跃点为单位)的示例。

[0030] 图8提供了网关如何协调哪个直接互连端口应响应在同一直接互连中多于一个的端口接收的ARP请求的示例。

[0031] 图8b提供了协调机制的实施例的示例(根据图8),其中协调功能分布在直接互连ASIC上,且每个直接互连ASIC连接到用于传达协调信息和决策的协调总线。

[0032] 图9提供了逻辑树的一个实施例,该逻辑树解释了当多于一个的网关连接到同一个环面时,网关如何协调其对环面拓扑的认知以及对ARP、广播、多播和选播业务的响应。

[0033] 图10提供了中间网关节点如何处理数据包而不是将数据包路由到另一个网关节点的示例。

[0034] 图11和12描述了直接互连ASIC的优选实施例的操作。

具体实施方式

[0035] 本发明提供了一种专用设备,即网关设备,其能够在传统网络和直接互连网络之间桥接、交换或路由。采用这样的专用设备,直接互连节点上的资源不必由直接互连和传统网络之间的桥接、切换或路由来负担,从而将对I/O性能的影响最小化。另外,与网关节点的现有技术使用相反,本网关设备是可以由网络管理系统管理的高度可管理的设备。此外,本发明的网关设备允许在多个直接互连端口之间协调MAC表和ARP、广播、多播和选播响应。

[0036] 图1示出了根据本发明的一个实施例的网关设备50的高层总览图。网关设备50包括两组端口。第一组端口(在此示例中,最左边的十二个端口)是标准的传统网络端口100,每个端口具有单个链路(例如SFP+、QSFP、QSFP+连接器),这样的端口连接到形成传统网络的现有的交换机和/或设备。在例如Clos拓扑(多级电路交换网络)中,这些端口100最有可能连接到主干(spine)或超级主干端口。第二组端口(最右边的十二个端口)是直接互连端口102,每个端口具有大量链路(两个或更多)(例如,具有MXC/MTP/MTO连接器)。例如,如果类似于第WO 2015/027320A1号PCT专利申请公开中所公开的无源配线架/集线器60被用于支持直接互连拓扑,则将直接互连端口102连接到无源配线架/集线器60(参见图3)。网关设备50可以具有1机架单元(RU)可机架安装设备的形式,或者甚至具有1/2RU(或如期望的其他)的形式,以有效地节省机架空间。图2示出了包括传统网络交换机专用集成电路(ASIC)106的实施例,其中第一组传统网络端口100各自连接到传统网络连接器,并且其中第二组直接互连端口102各自连接到它们自己的专用直接互连ASIC 104。这些专用直接互连ASIC 104优选地各自都能够充当直接互连节点,并具有在本地通过传统网络接口线(例如100Gbps以太网)发送的目的地/源业务。交换机ASIC 106和直接互连ASIC 104的端口之间的连接可以使用任何标准组件互连(例如100GBase-KR4、100GBase-KP4、40GBase-KR4、25GBase-KR、10GBase-KR或任何其他类似标准)来实现。

[0037] 传统的网络交换机ASIC 106包含标准的网络业务转发功能,包括学习经过其端口可到达的设备、经过适当的出站端口、网络过滤、业务监督,以及通常在层2、层3和层4及以上网络交换机、路由器和网桥实现的其它功能发送所接收到的业务。转发决定可以是基于一个或多个因素的,包括但不限于源和目的地层2(MAC)地址、源端口、源和目的地层3(IPv4、IPv6等)地址,源和目的地层4端口,以及层5及以上的标头和数据有效载荷。

[0038] 在直接互连ASIC 104(其具有可以经过交换机ASIC 106的标准端口到达的最终目的地)处,从直接互连接收到的网络业务将从直接互连ASIC 104发送到交换机ASIC 106,在此交换机ASIC 106标准业务转发功能将经过适当的标准端口发出该业务。

[0039] 类似地,由交换机ASIC 106(标准端口具有经过直接互连ASIC 104可以到达的最终目的地)从标准端口接收的网络业务将由交换机ASIC 106转发到直接互连ASIC 104。

[0040] 在另一个实施例(未示出)中,且适用于每个可能的实施例,将理解的是交换机ASIC 106(以及本文讨论的所有类似的ASIC)可以用现场可编程门阵列(FPGA)、通用处理器、网络处理器或任何其他能够执行网络业务转发的设备来代替。

[0041] 在另一实施例(未示出)中,且适用于每个可能的实施例,将理解的是直接互连ASIC 104(以及本文所讨论的所有类似ASIC)可以由现场可编程门阵列(FPGA)、通用处理器、网络处理器或任何其他能够充当直接互连网络中的节点的设备来代替。

[0042] 图2b示出了另一实施例,其中每一个直接互连ASIC 104可以连接到一个或多个直接互连端口102,并且可以连接到一个或多个传统网络交换机ASIC 106端口。

[0043] 图2c示出了另一实施例,其中单个交换机及直接互连ASIC 105结合了传统网络交换机ASIC 106和一个或多个直接互连ASIC 104的功能。再一次将理解的是,单个交换机及直接互连ASIC 105(以及本文讨论的所有类似的ASIC)可以由现场可编程门阵列(FPGA)、通用处理器、网络处理器或任何其他能够执行网络业务转发和/或充当直接互连网络中的节点的设备来代替。

[0044] 图2d示出了网关设备的另一实施例,该网关设备具有包含传统网络端口和直接互连端口的主机接口卡110的形式。主机接口卡110设计为经由通信总线(例如PCIe、Gen-Z或CCIX)连接到主机设备。主机设备可以是服务器、计算机、硬盘驱动器、固态驱动器、驱动器柜或任何其他能够发送和接收数据的设备。传统网络端口连接到标准网络控制器ASIC 108(或类似的现场可编程门阵列(FPGA)、通用处理器、网络处理器或任何其他能够执行网络业务转发的设备等)且直接互连端口连接到直接互连ASIC104。直接互连ASIC 104也连接到网络控制器ASIC108。以这种形式,网络控制器ASIC108将能够在传统网络和直接互连网络之间交换业务,而无需主机干预。

[0045] 图2e示出了另一个主机接口卡111的实施例,其中直接互连ASIC 104和网络控制ASIC 108的功能被组合到单个直接互连及传统ASIC 107中。再一次将理解的是,单个直接互连及传统ASIC 107(以及本文讨论的所有类似的ASIC)可以由现场可编程门阵列(FPGA)、通用处理器、网络处理器或任何其他能够执行网络业务转发和/或充当直接互连网络中的节点的设备来代替。

[0046] 图2f示出了另一个实施例,其中直接互连端口102包含每个端口一个或多个链路,并且由网关设备50组合成组55。每个组55由网关设备50逻辑关联以在直接互连网络中充当单个节点。

[0047] 如图4所示,直接互连端口102和组55将分别代替直接互连拓扑中的设备。

[0048] 在又一个实施例中,如果在直接互连中未使用无源配线架/集线器60,则每个网关端口(即,直接互连端口102)的各个链路可以连接到作为直接互连的一部分的设备。在这方面,图4b示出了一实施例,其中直接互连端口102中的一个端口的链路逐个连接到直接互连网络内的相邻节点。在图4b中,直接互连采取二维环面的形式,并且包括直接互连端口102

的4个链路被编号为1、2、3和4。为了形成二维环面，链路1连接到设备A，链路2连接到设备B，链路3连接到设备C，链路4连接到设备D。类似地，对于其他拓扑，直接互连端口102的每个链路都应连接到拓扑所需的适当的相邻设备。

[0049] 在进一步的实施例中，如图5所示，网关设备50还可以用于在多个直接互连之间进行桥接、交换或路由。在这种情况下，网关设备50上的直接互连端口102将在不同的直接互连之间划分(在图5中以A和B示出)。然后，直接互连A中的所有设备都可以从直接互连A经过网关设备50到达，且反之亦然。在所提供的示例中，目的地为直接互连B的设备的来自直接互连A的设备的业务将首先穿越直接互连A到网关设备50上显示为A4的直接互连端口102。网关设备50然后将该业务转发经过示出为B2的直接互连端口102，然后在此被转发经过直接互连B去往目的地节点。

[0050] 在本发明的又一个实施例中，网关可以用作接入交换机，并且直接互连将形成骨干网(见图6)。在这种情况下，直接互连内的所有或大多数节点将由网关端口组成，而所有或大多数其他设备将直接连接到网关设备50的传统网络端口100。在图6提供的示例中，来自设备组A目的地为设备组B的业务将首先被转发到网关设备50A。网关50A转发功能将识别出该目的地设备可经过直接互连到达，并且将经过网关设备50A的直接互连端口102中的一个端口转发该业务。然后直接互连将业务转发到网关设备50B的直接互连端口中的一个端口。网关50B转发功能将识别出该目的地设备可经过其标准网络端口中的一个端口到达，并将适当地转发该业务。

[0051] 为了最大化I/O业务效率，图7提供了在单个直接互连部署中如何选择直接互连端口102以最小化从直接互连中的每一个网关设备50到最近的网关端口的平均距离(跃点数)的示例。在具有按图7编号的节点的4x42D环面中，使用网关充当节点1、6、11和16将使到直接互连中其他节点的距离最小化。然而，本领域技术人员还将理解的是，如果节点的子集倾向于产生高于平均的I/O量，则网关节点的部署可能会偏向于更接近于这些更高的I/O节点。已知有多种算法可用于确定这些I/O节点的最佳位置(例如，参见Bae,M.,Bose,B.发表的“Resource Placement inTorus-BasedNetworks”载于Proc.IEEE国际并行处理研讨会第327-331页，IEEE计算机协会出版社，洛斯阿拉米托斯(1996年)；Dillow,DavidA.等人发表的：“I/O Congestion Avoidance via Routing and ObjectPlacement”美国:N.p.,2011年印刷的Cray用户组会议记录(CUG:2011)，美国阿拉巴马州费尔班克斯；Almohammad,B.、Bose,B.发表的“Resource Placements in 2D Tori”，载于:Proc.IPPS 1998年，第12届国际并行处理研讨会国际会议论文集第431页，IEEE计算机学会，华盛顿特区(1998)；Dillow,David A.等人发表的：“Enhancing I/O Throughput via Efficient Routing and Placement for Largescale Parallel File Systems”，载于:IEEE国际性能、计算和通信会议(IPCCC)的会议论文集，2011年IEEE第30届国际；Ezell,M.等人发表的：“I/O RouterPlacement and Fine-GrainedRouting onTitanto Support Spider II”，载于:Cray用户小组会议(CUG)的会议记录，2014年5月；以及，Babatunde,A.等人发表的：“I/O Node Placement forPerformance andReliabilityinTorus Networks”，德克萨斯农工大学计算机科学系，2006年1月。)

[0052] 在一优选实施例中，直接互连端口102将充当标准直接互连端口，并自动转发保留在直接互连中的业务(即，转发FLIT)。它们还将针对目的地为不在直连互连中的设备的业

务将FLIT重新组合成网络数据包(见第WO2015/120539A1号PCT专利申请公开:一种在分布式直接互连网络中路由数据包的最佳方法)。网关设备50还应当优选地具有向/从传统网络端口100和直接互连端口102中的每一个端口发出/接收网络数据包的能力,并且还能够基于层2、3或根据以上来解释和转发该业务。

[0053] 在一优选实施例中,标准的北向网络管理接口将显露(例如CLI、OpenFlow、SNMP、REST等)以允许网络管理系统来管理网关设备50。

[0054] 在一个实施例中,当多个网关端口连接到同一直接互连时,来自给定流的所有数据包应优选地在同一网关端口上出站,以帮助保证按顺序传送数据包。

[0055] 网关设备50应当优选地被配置为汇总连接到同一直接互连的直接互连端口102之间的MAC转发表(即,当直接互连端口学习VLAN/MAC地址/node_id元组时,该元组应当优选地与连接到同一直接互连的其他直接互连端口102共享)。

[0056] 在一优选实施例中,当在连接到同一直接互连的一个或多个直接互连端口102处接收到ARP请求时,应由网关50协调哪个直接互连端口应作出响应的决定以确保仅发出单个响应(例如,参见图8)。这可以通过以下方式完成:例如经过轮询选择方案、经过对ARP请求的某些部分(IP地址、源MAC等)进行杂凑(hash),或本领域技术人员已知的其他算法,来选择最接近ARP请求的源的直接互连端口、选择接收ARP的第一个直接互连端口。图8b示出了这一协调机制的一个实施例,其中协调功能分布在直接互连ASIC104上,并且每一个直接互连ASIC 104连接至用于传达协调信息和决定的协调总线112。在另一实施例中,该协调功能可以集中在专用协调ASIC中,并且每一个直接互连ASIC 104经由专用或共享协调总线112连接到协调ASIC。

[0057] 当多于一个的网关连接到同一环面时,网关设备50应优选地以与上述单个网关情况类似的方式协调其对环面拓扑的认知以及对ARP请求的响应(见图9关于这种协调的逻辑树)。在一个实施例中,网关可以通过广播协议发现彼此。在另一个实施例中,网关可以被配置为知道其他网关的位置。在又一个实施例中,可以使用诸如Raft的共识算法来确保网关的协调和业务-网关/端口关联的一致性(参阅例如Ongaro,D.,Ousterhout,J.发表的:“In Search ofan Understandable Consensus Algorithm(扩展版)”载于:2014年USENIX年度技术会议,2014年6月19日,宾夕法尼亚州费城;Woos,D.等人发表的:“Planning for Change in a Formal Verification of the Raft Consensus Protocol”,载于:CertifiedPrograms and Proofs(CPP),2016年1月)。

[0058] 通常,每当环面节点想要与可通过一个或多个网关设备50访问的资源进行通信时,该网关优选地应协调选择哪个网关端口以与上述ARP示例类似的方式提供对该资源的访问。这样的示例包括选播、广播和多播业务、节点和服务发现协议以及IPv6邻居发现。

[0059] 作为进一步的考虑,重要的是要注意,在许多情况下,在直接互连内使用非最小路由。由于直接互连中的网关在拓扑结构内的多个位置都具有网关端口,因此目的地为一个网关端口的业务可能首先穿越其他网关端口中的一个。因此可优选的是,通过使单个的第一个网关端口处理业务来提高效率,而不是允许业务穿越到更远的网关端口。在图10中提供了这样的一个示例,在其中节点5正在向节点1发送业务,但是由于是非最小路由,数据包将穿越节点6。在一个实施例中,为了提高效率,端口位于节点6的网关将识别出该数据包的目的地为另一个网关端口,并将处理该数据包就像其目的地为节点6一样,而不是通过直接

互连经由节点2转发。

[0060] 如上所述，直接互连ASIC 104在交换机ASIC 106和直接互连之间提供连接。为了确保本领域技术人员将能够制造并使本发明的网络网关设备起作用，图11和图12描述了直接互连ASIC 104的优选实施例的操作。对于本发明范围内的其他ASIC，包括适当情况下ASIC 105、107和108，该操作以及必要的任何修改也都是可用的。

[0061] 如本领域中众所周知的，交换机ASIC 106发送和接收以太网帧。图11描述了由交换机ASIC 106生成的以太网帧200是如何由直接互连ASIC 104处理的。MAC地址数据库201包含以太网MAC地址列表以及与每一个以太网MAC地址关联的直接互连节点。检查从交换机ASIC 106接收到的以太网帧200，并从其标头中检索源MAC地址和目的地MAC地址。将这些MAC地址用作MAC地址数据库201的索引。以太网帧301的源MAC地址与当前节点号组合以创建或更新此节点号与MAC地址数据库201中的源MAC地址之间的关联。

[0062] 如果以太网帧200的目的地MAC地址在MAC地址数据库201中，则从MAC地址数据库201中检索与该MAC地址关联的节点号202，然后将节点号202用作索引进入源路由数据库206并从源路由数据库206中检索与节点号202相关联的源路由203。如本领域中众所周知的，源路由数据库包含网络目的地的列表以及通过该网络到达每一个目的地的一个或多个路径。源路由数据库可以手动填充，也可以依据众所周知的自动拓扑发现和路由计算算法。然后将以太网帧200转换为FLIT 204。FLIT是在直接互连中使用的专用帧类型，并且可以具有固定尺寸也可以具有可变尺寸。在优选的实施例中，FLIT将具有固定尺寸。在另一个实施例中，FLIT将具有在最小和最大尺寸内的可变尺寸。在又一个实施例中，将调整FLIT的尺寸，使得以太网帧200完全适合于FLIT有效载荷。

[0063] 如果以太网帧200大于单个FLIT的有效载荷，则将创建多个FLIT 204。如果以太网帧200适合于单个FLIT的有效载荷，则将创建单个FLIT。然后将源路由203与当前节点的节点号一起插入到FLIT 204中的第一个的标头中。然后，将FLIT 204从源路由203中指定的出站端口205发出。

[0064] 如果以太网帧200的目的地MAC地址不在MAC地址数据库201中，或者如果以太网帧200的目的地MAC地址指示它是一个广播以太网数据包，则尽管将不包括源路由，以太网帧200也像在上述情况的以太网中那样转换为FLIT 204。一旦创建了FLIT 204，就在第一个FLIT的标头中设置一个标志，以指示应将这些FLIT广播到直接互连中的每个节点。在第一个FLIT的标头中还设置了“生存时间”(TTL)值。TTL确定可以通过直接互连转发广播FLIT的最大次数。在一个实施例中，选播和组播以太网帧将视为它们是如上所述的广播帧来对待。

[0065] 图12描述了在直接互连ASIC 104的优选实施例中直接互连ASIC 104如何处理多个FLIT204。直接互连ASIC 104接收来自直接互连301的FLIT204，并检查其中第一个FLIT的标头，查看是否设置了广播标志。如果未设置广播标志，则从第一个FLIT中检索源路由203，并确定源路由是否指示当前节点是FLIT 204的目的地节点。在另一个实施例中，FLIT包含目的地节点的节点号。该节点号是用于确定当前节点是否是FLIT 204的目的地节点的节点号。如果当前节点是FLIT 204的目的地节点，则将FLIT 204组合以形成以太网帧301。以太网帧301的源MAC地址与第一个FLIT的标头中的节点号相结合，以在MAC地址数据库201中创建或更新所述节点号与源MAC地址之间的关联。然后将以太网帧301发出至交换机ASIC 106。

[0066] 如果确定这不是目的地节点,则使用源路由来确定FLIT 204的出站端口302。然后将FLIT 204发出至出站端口302。

[0067] 如果在第一个FLIT标头中设置了广播标志,则将FLIT 204组合以形成以太网帧301。将以太网帧301的源MAC地址与第一个FLIT标头中的节点号组合以在MAC地址数据库301中创建或更新所述节点号和源MAC地址之间的关联。以太网帧301被发出至交换机ASIC 106。

[0068] 然后,第一个FLIT的标头中的TTL被减去一。如果TTL现在等于零,则丢弃该FLIT 204。如果TTL大于零,则将FLIT 204从除最初接收到FLIT204的入站端口以外的所有出站端口发出。

[0069] 在直接互连ASIC 104的其他实施例中,可以不使用源路由。在一个实施例中,将由每一个节点使用以太网帧200的目的地MAC地址来执行本地下一跃点路由查找。在另一个实施例中,FLIT标头中的目的地节点信息将由每一个节点用来执行本地下一跃点路由查找。

[0070] 对于本领域的技术人员将显而易见的是,直接互连104和交换机ASIC106的其他实施例可以被设计为使用以太网以外的协议工作。在一个实施例中,这些元件将被设计为使用Gen-Z工作。在这种情况下,直接互连104将期望接收到Gen-Z Core64数据包而不是以太网帧。代替以太网MAC地址的是将使用Gen-Z GCID(全局组件ID)并将其与直接互连节点号关联。

[0071] 尽管已经描述了本发明的特定实施例,但是对于本领域的技术人员显而易见的是,可在所附权利要求的范围内作出对这些实施例的变化和修改。

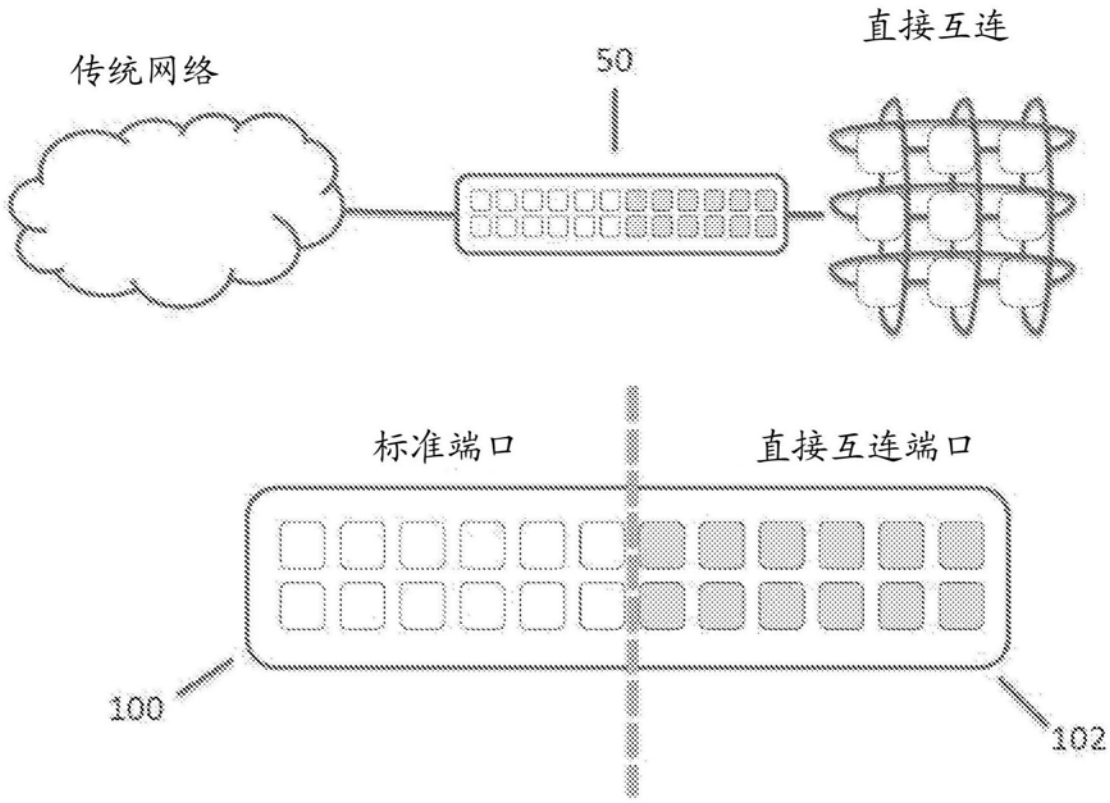


图1

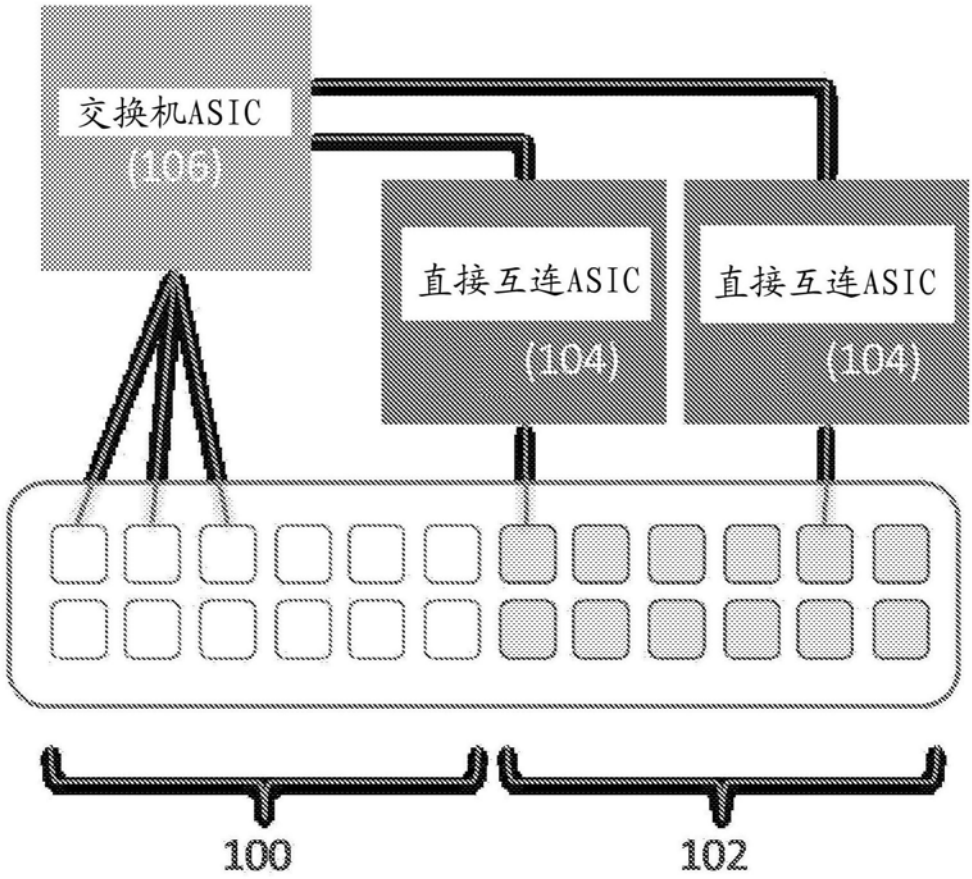


图2

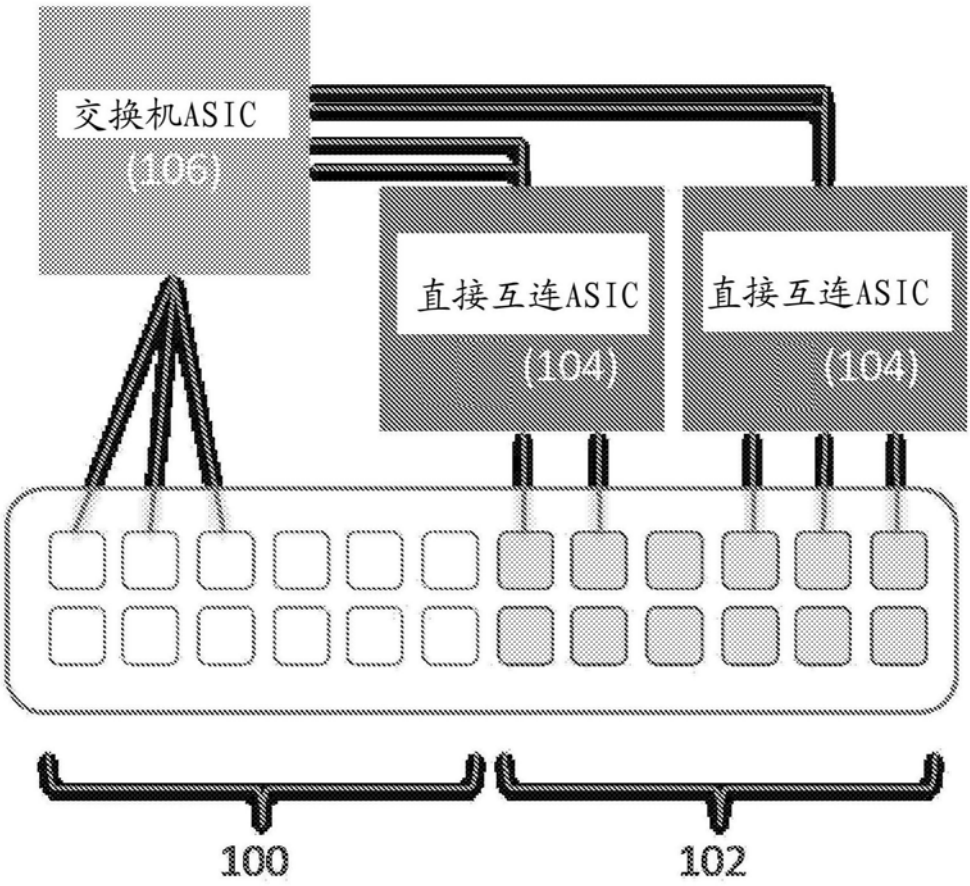


图2b

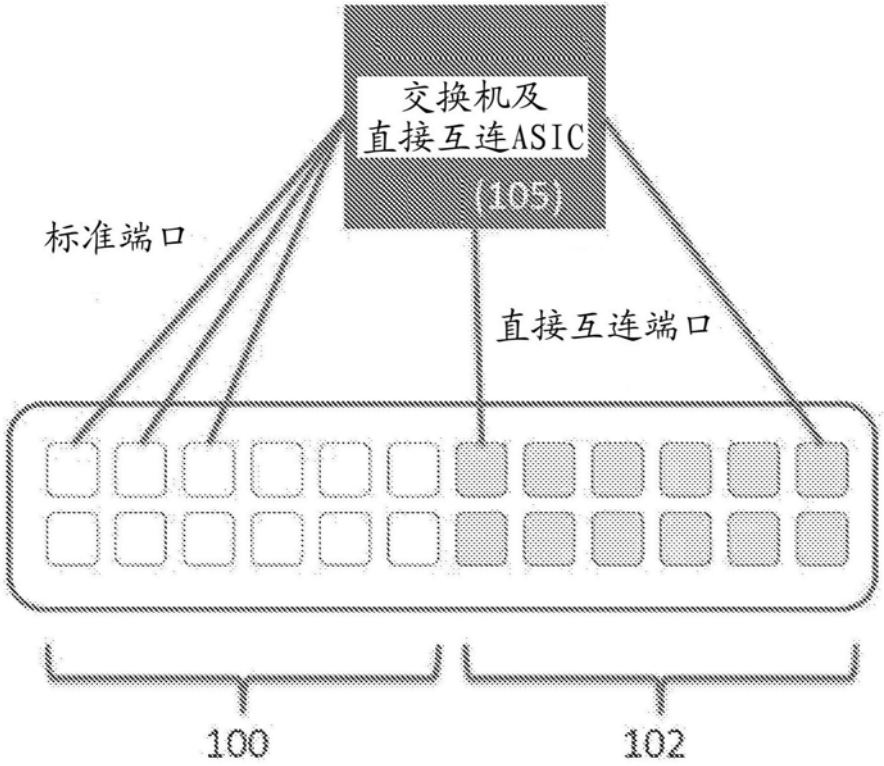


图2c

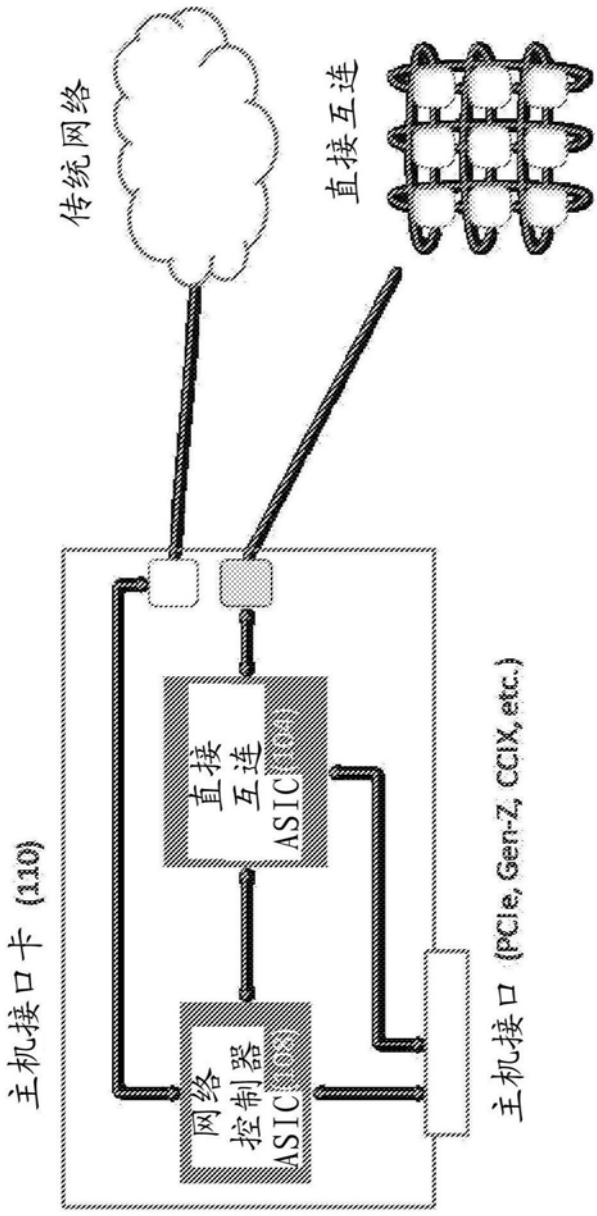

传统网络
直接互连
主机接口卡 (110)
直接互连ASIC (104)
网络控制器ASIC (103)
主机接口 (PCIe, Gen-Z, CCIX, etc.)

图2d

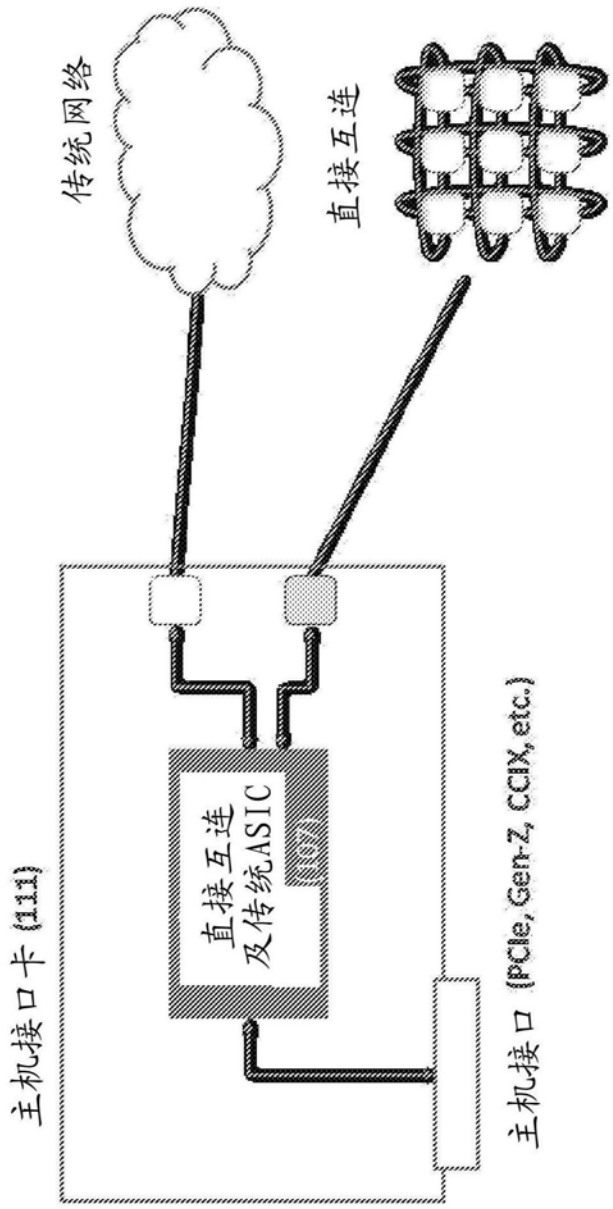


图2e

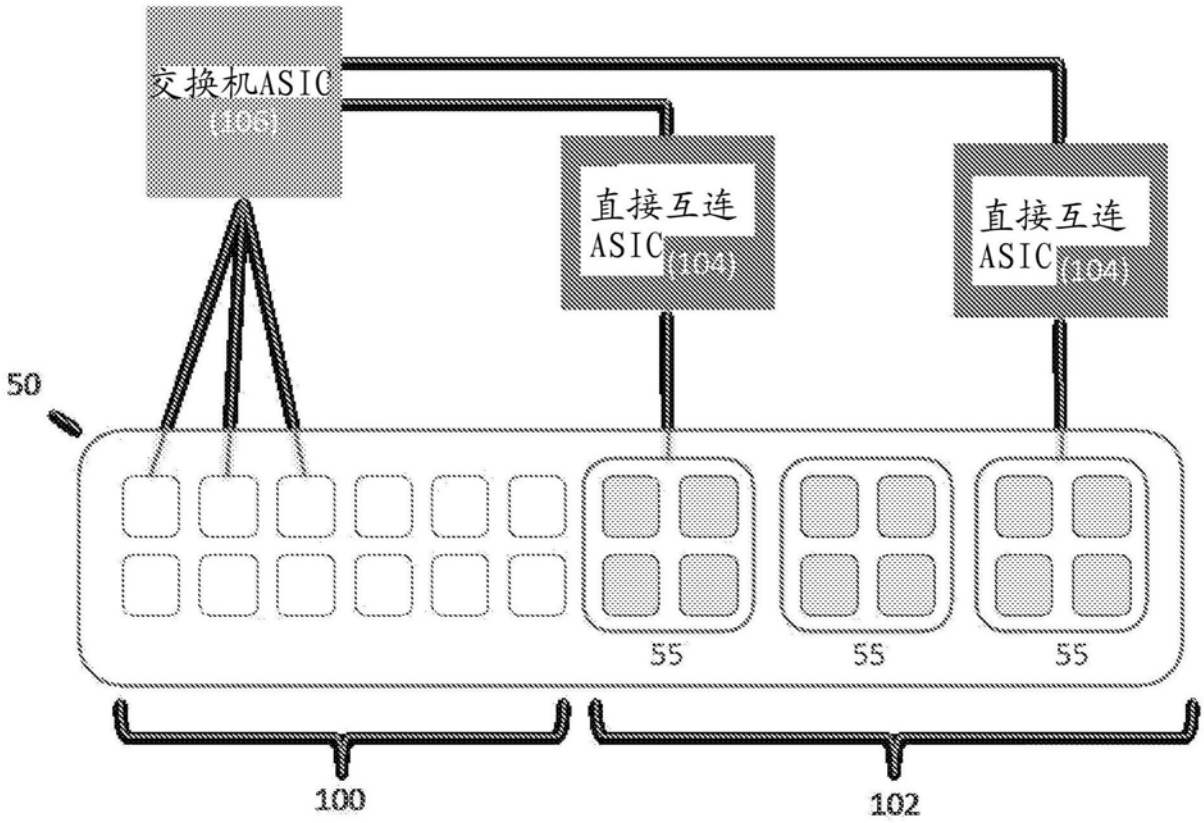

交换机ASIC
(106)
直接互连
ASIC (104)
直接互连
ASIC (104)
50
55
55
55
100
102

图2f

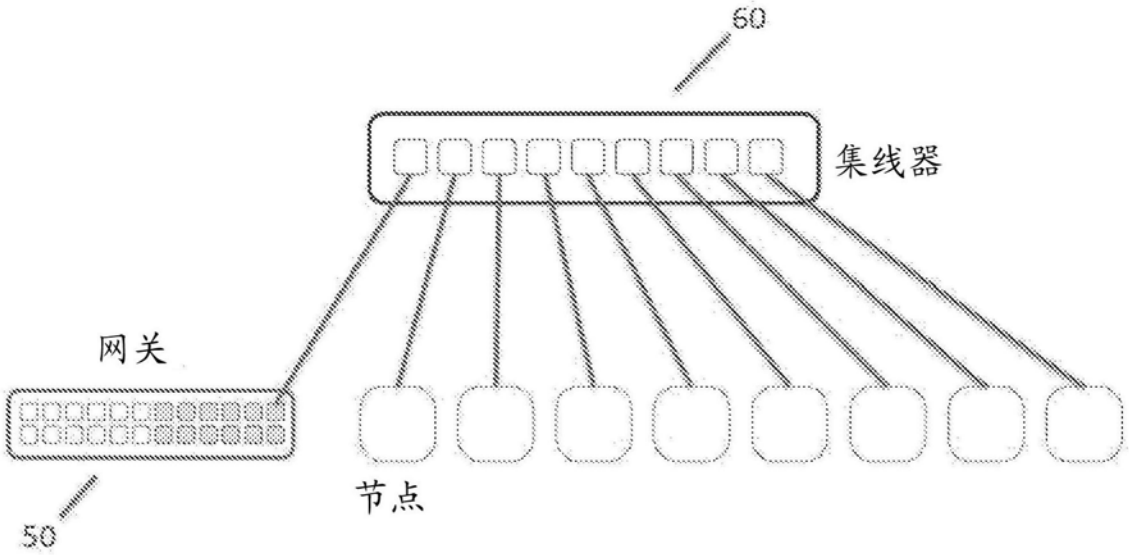

60
集线器
网关
节点
50

图3

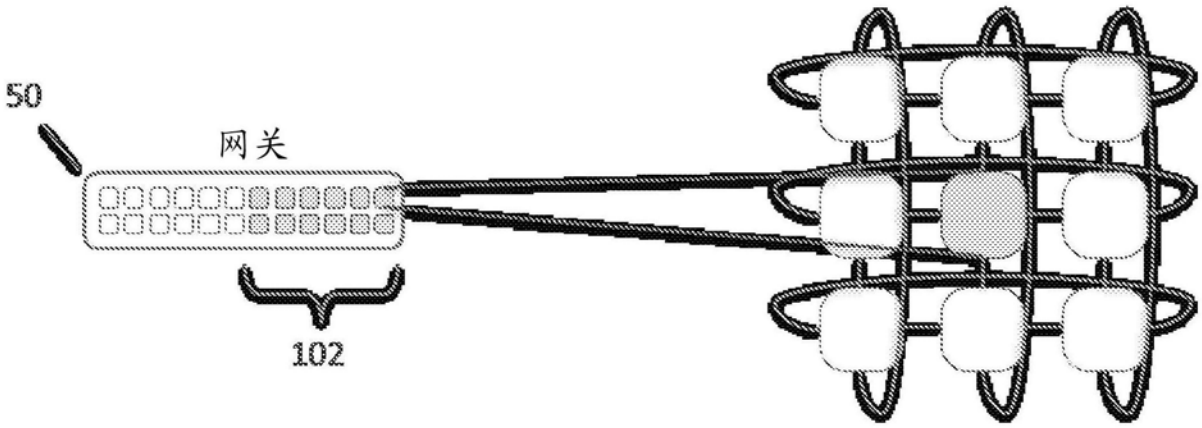


图4

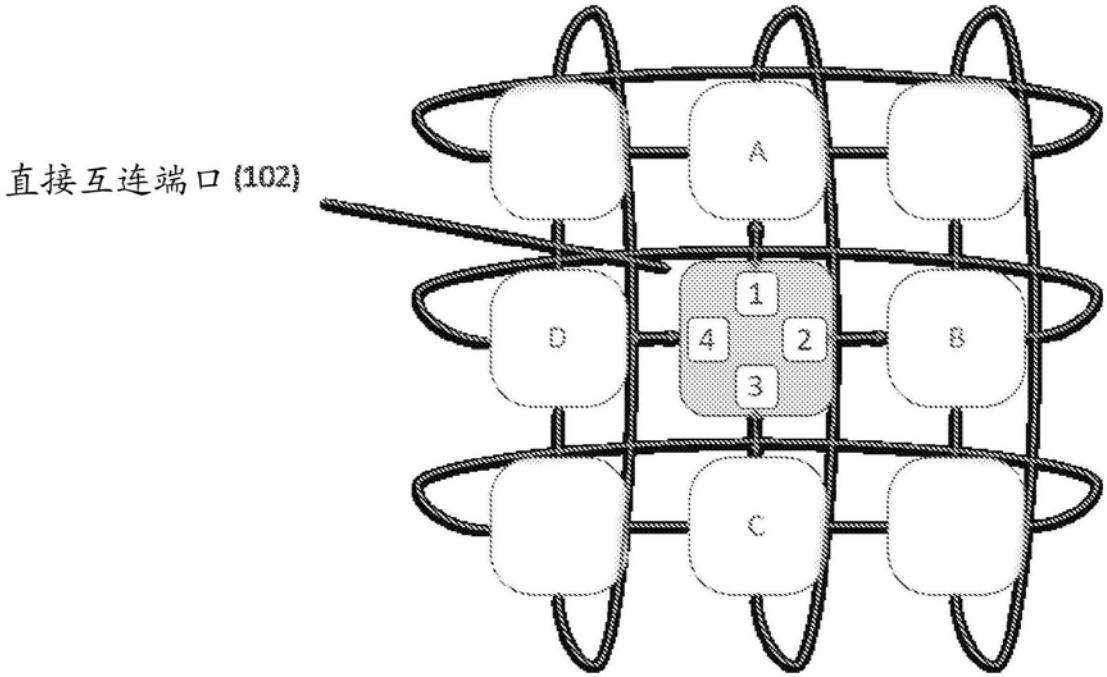


图4b

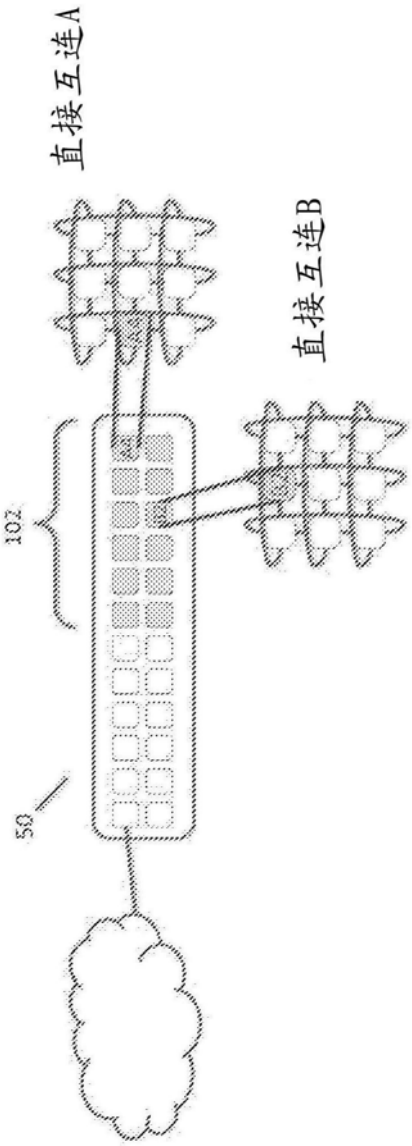


图5

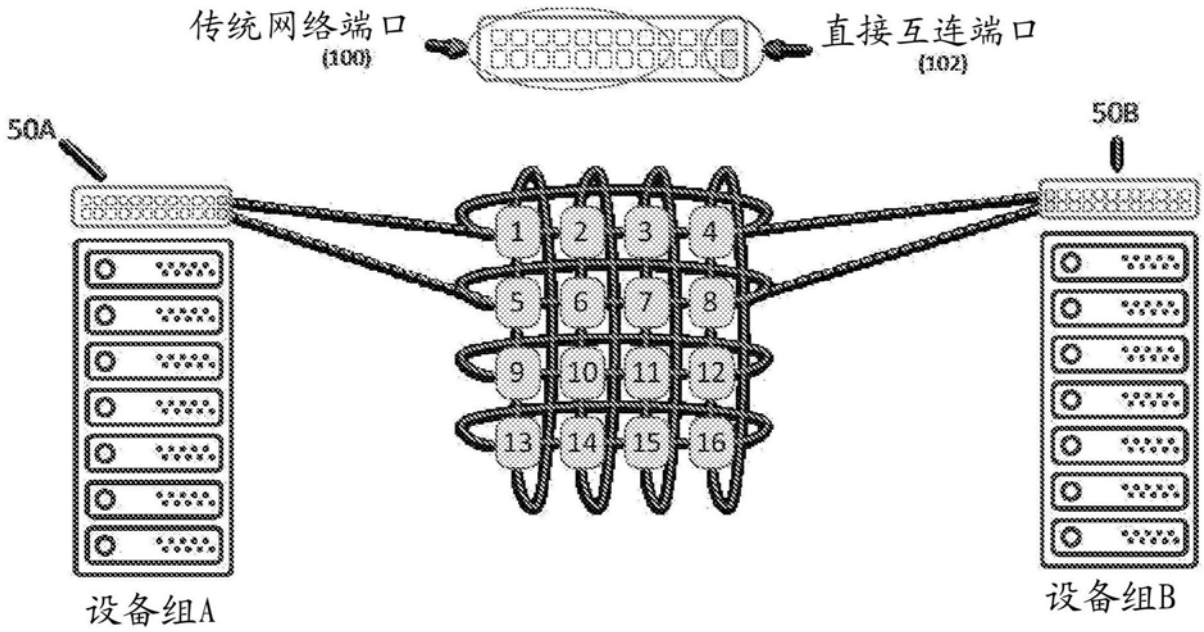

传统网络端口
(100)
直接互连端口
(102)
50A
50B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
设备组A
设备组B

图6

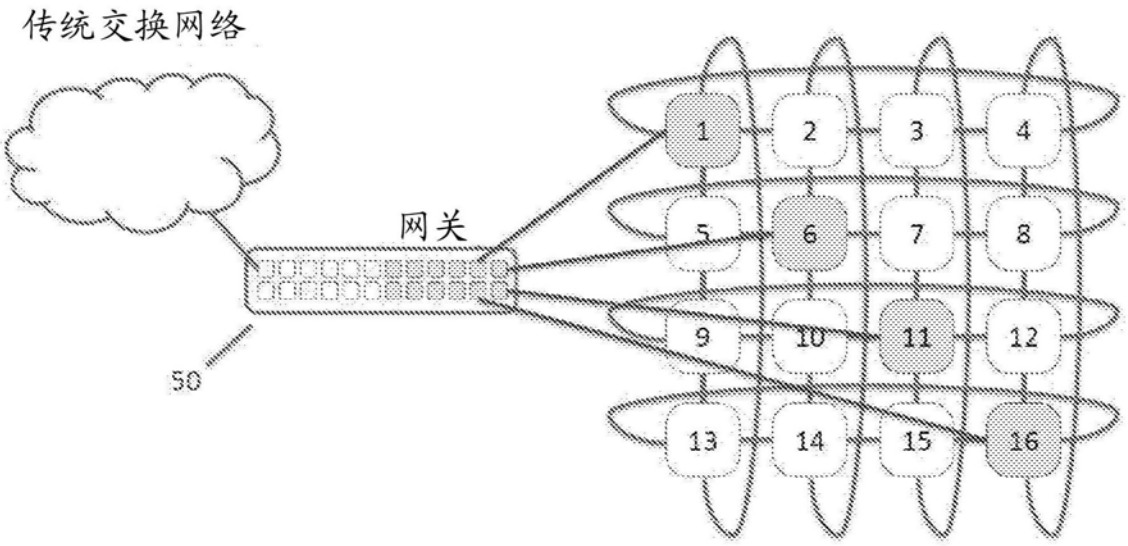

传统交换网络
网关
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

图7

- 节点1、6、11和16为同一网关设备上的端口
- 节点5发送广播数据包
- 节点1、6、11和16接收广播数据包
- 网关确定节点1和6为最接近节点5的节点
- 网关选择节点1为最佳节点
- 节点1与该业务相关联并将处理
  未来关于广播数据包的业务

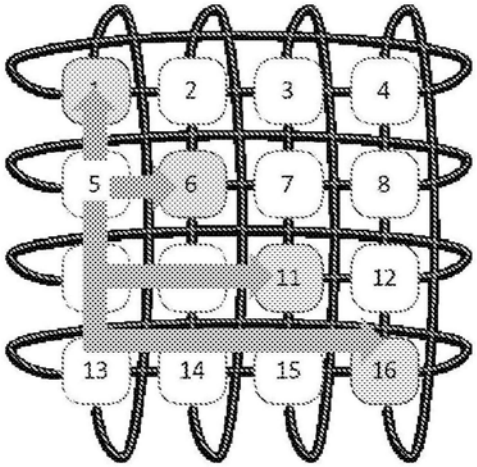


图8

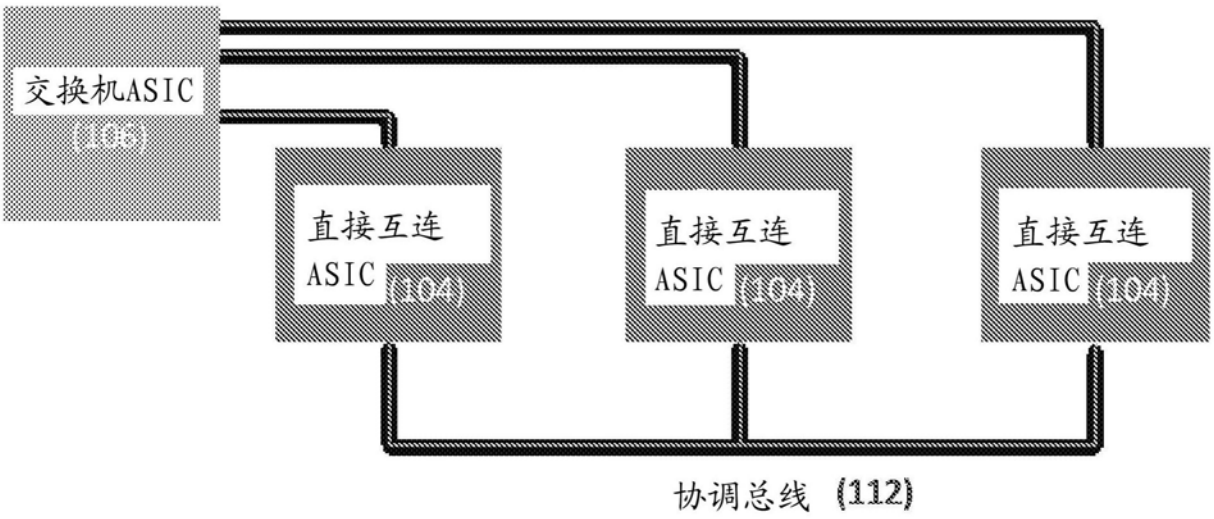


图8b

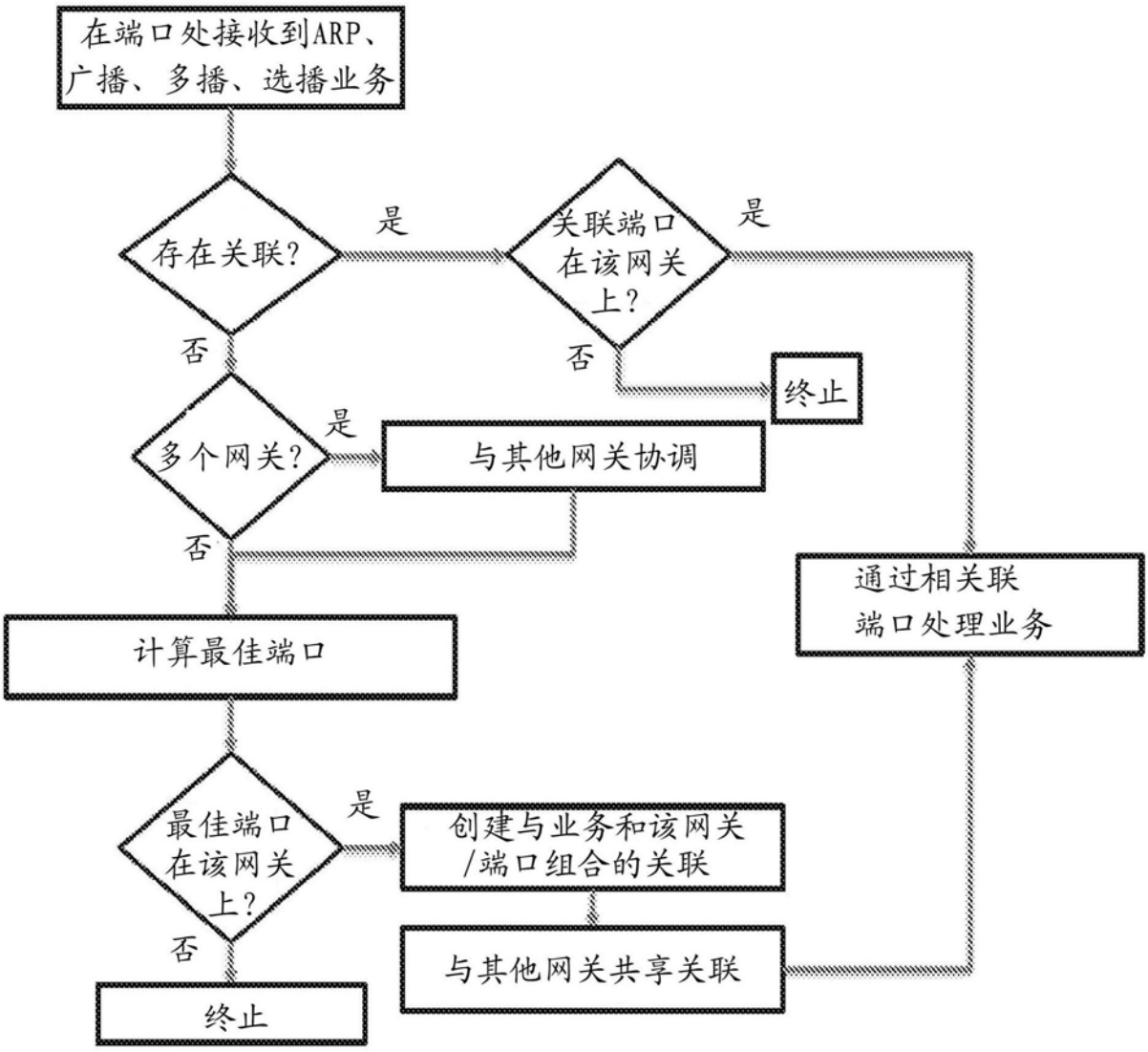

在端口处接收到ARP、广播、多播、选播业务
存在关联?
是
关联端口在该网关上?
是
否
终止
否
多个网关?
是
与其他网关协调
否
通过相关联端口处理业务
计算最佳端口
最佳端口在该网关上?
是
创建与业务和该网关/端口组合的关联
否
与其他网关共享关联
终止

图9

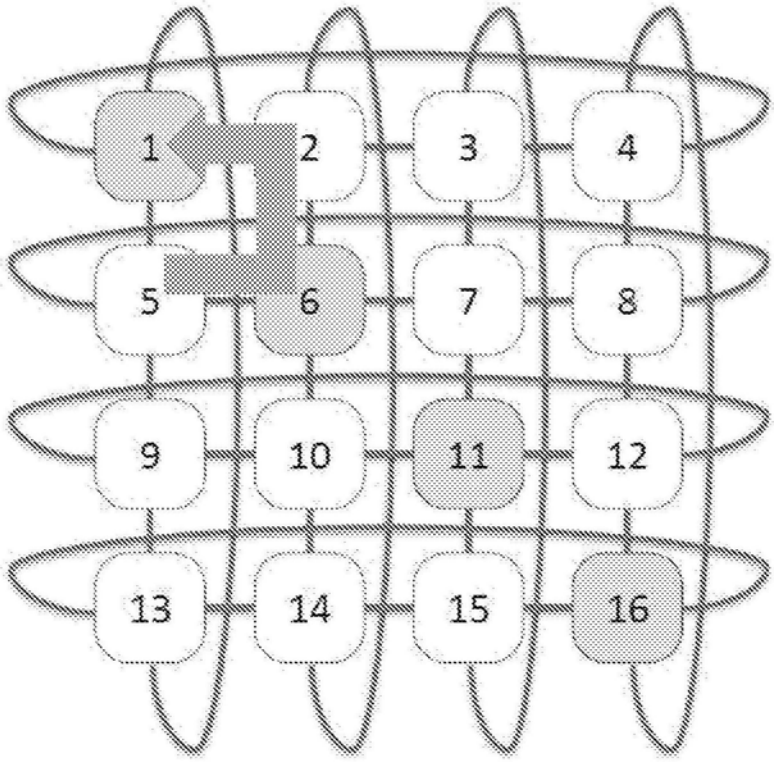


图10

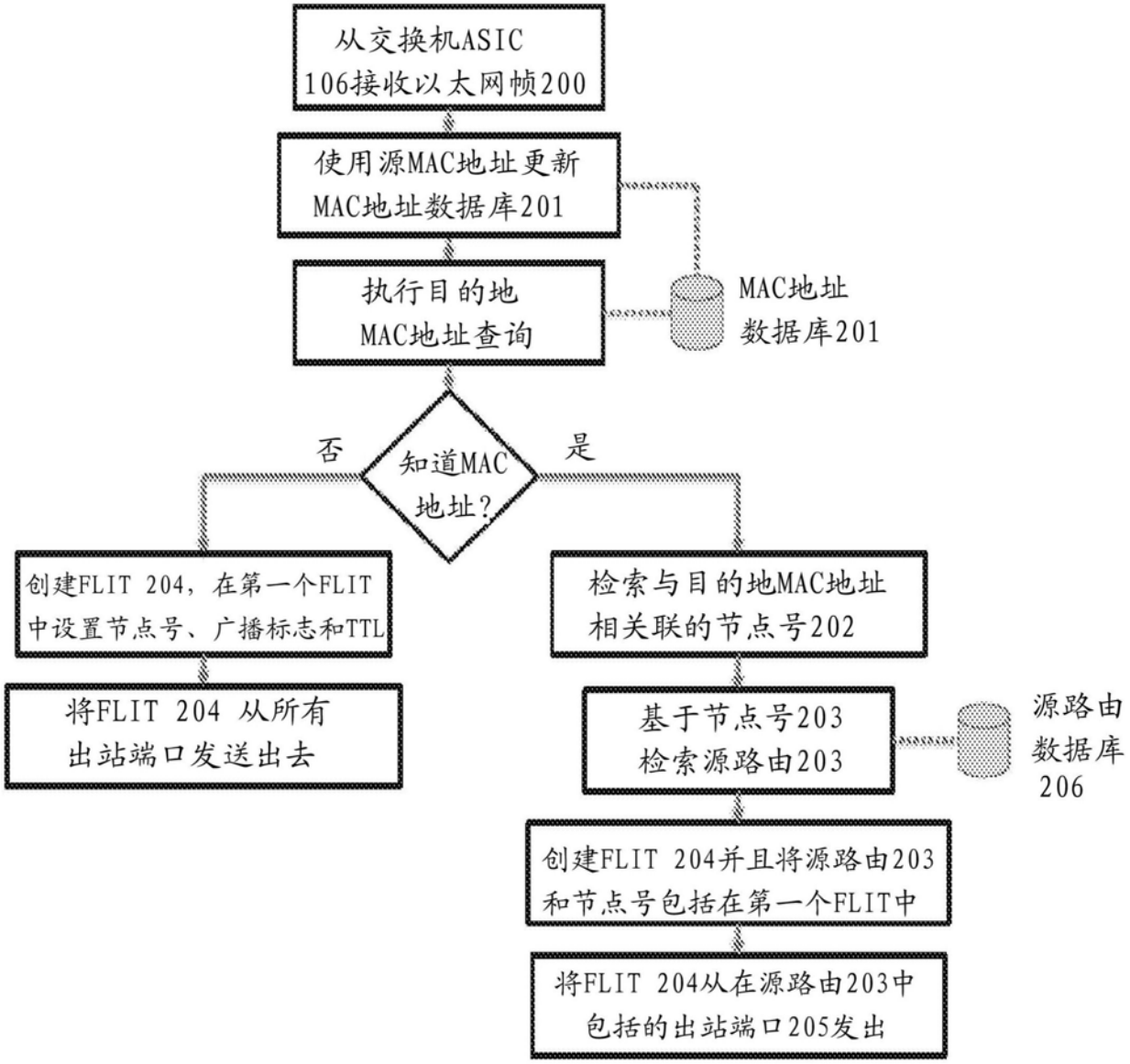

从交换机ASIC
106接收以太网帧200
使用源MAC地址更新
MAC地址数据库201
执行目的地
MAC地址查询
MAC地址
数据库201
知道MAC
地址?
否
是
创建FLIT 204，在第一个FLIT
中设置节点号、广播标志和TTL
将FLIT 204 从所有
出站端口发送出去
检索与目的地MAC地址
相关联的节点号202
基于节点号203
检索源路由203
源路由
数据库
206
创建FLIT 204并且将源路由203
和节点号包括在第一个FLIT中
将FLIT 204从在源路由203中
包括的出站端口205发出

图11

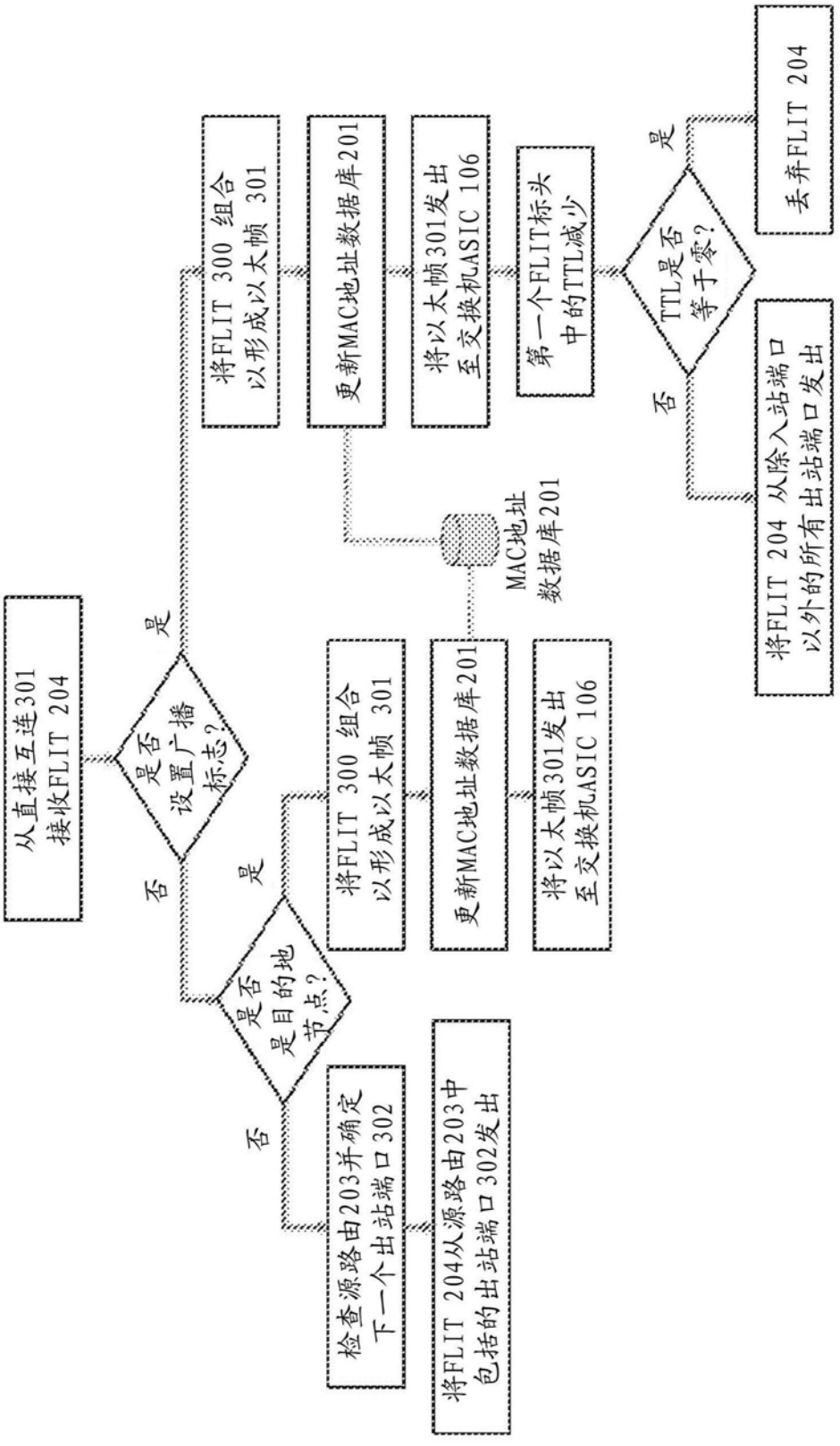

从直接互连301接收FLIT 204
是否设置广播标志?
是
否
将FLIT 300 组合以形成以太帧 301
更新MAC地址数据库201
将以太帧301发出至交换机ASIC 106
第一个FLIT标头中的TTL减少
TTL是否等于零?
是
否
丢弃FLIT 204
将FLIT 204 从除入站端口以外的所有出站端口发出
MAC地址数据库201
是否是目的地节点?
是
否
将FLIT 300 组合以形成以太帧 301
更新MAC地址数据库201
将以太帧301发出至交换机ASIC 106
检查源路由203并确定下一个出站端口 302
将FLIT 204从源路由203中包括的出站端口 302发出

图12