

数据中心网络负载均衡方法、装置、系统、通信设备及存储介质

相关申请的交叉引用

本申请要求于 2023 年 10 月 08 日提交中国国家知识产权局、申请号为 202311293576.7、发明名称为“一种数据中心网络负载均衡方法及装置”的中国专利申请 5 的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及网络技术领域，尤其涉及一种数据中心网络负载均衡方法、装置、系统、通信设备及存储介质。

10

背景技术

目前，数据中心网络通常采用以太网传输链路承载远程直接数据存取（Remote Direct Memory Access, RDMA）进行通信，即允许通过以太网使用 RDMA（RDMA over Converged Ethernet, RoCE）技术，该技术需要以太网支持无损传输。传统的负载均衡技术是基于流的五元组进行哈希选路，从链路的多条等价路径中随机选择一条进行转发，俗称静态等价多路径路由（Equal-Cost Multipath Routing, ECMP）。15

人工智能（Artificial Intelligence, AI）集群训练需要进行大量的分布式并行计算，显示芯片（Graphics Processing Unit, GPU）服务器需要通过交换网络进行跨主机网络通信，AI 服务器的模型训练依赖 RDMA 进行网络通信。20

发明内容

第一方面，本申请提供一种数据中心网络负载均衡方法，应用于数据中心网络中的交换机，包括以下步骤：25

获取多个计算节点的信息；

根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；

30 当接收到数据流时，根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

在一种可能的设计中，每个计算节点的信息包括计算节点的编号；

35 根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，包括：

计算每个计算节点的编号与多条等价转发路径数量的商的余数；

根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

采用上述方法，由于计算节点的编号不同，计算节点的编号与等价转发

路径数量的商的余数就不同，进而确定的转发路径不同，实现计算节点的编号与多条等价转发路径的均匀绑定。

在一种可能的设计中，每个计算节点的编号通过边界网关协议（Border Gateway Protocol, BGP）路由的扩展属性表达。

5 在一种可能的设计中，还包括：

根据所连接的计算节点的数量，依次对所连接的计算节点进行编号。

在一种可能的设计中，每个计算节点的信息还包括计算节点的 IP 地址；

在根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，还包括：

10 生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。

在一种可能的设计中，在根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，还包括：

15 根据所述目标计算节点的 IP 地址查询所述转发表项确定所述目标计算节点对应的转发路径。

在一种可能的设计中，所述计算节点为以服务器为个体的单个计算单元，或是以服务器上多个显示芯片 GPU 为个体的多个计算单元。

在一种可能的设计中，所述数据中心网络为脊交换机和叶子交换机组网架构，或者超脊交换机、脊交换机和叶子交换机组网架构。

20 第二方面，本申请还提供一种数据中心网络负载均衡系统，应用于数据中心网络中的交换机，所述系统包括：

所述数据中心网络中的叶子交换机对所连接的计算节点进行编号；

所述数据中心网络中的各个邻接的交换机之间互相同步多个计算节点信息；

25 第一交换机根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；所述第一交换机为所述数据中心网络中的各个邻接的交换机中的一个；

30 当所述第一交换机接收到数据流时，根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

在一种可能的设计中，第一交换机根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，包括：

计算每个计算节点的编号与所述多条等价转发路径数量的商的余数；

根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

35 第三方面，本申请还提供一种数据中心网络负载均衡装置，该装置包括：

接收单元，用于获取多个计算节点的信息；

处理单元，用于根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；

40 发送单元，用于当接收到数据流时，根据所述数据流的目标计算节点对

应的转发路径转发所述数据流。

在一种可能的设计中，每个计算节点的信息包括计算节点的编号；

所述处理单元，用于在根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径时，计算每个计算节点的编号与多条等价转发路径数量的商的余数；根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

第四方面，本申请还提供一种通信装置，所述装置包括：一个或多个处理器和一个或多个存储器；其中，所述一个或多个存储器存储有一个或多个程序，当所述程序被所述一个或多个处理器执行时，使得所述装置执行上述第一方面或第二方面中任一方面所述的方法。

第五方面，本申请还提供一种数据中心网络负载均衡方法，其中，应用于数据中心网络中的通信设备，包括以下步骤：

获取多个计算节点的信息；

根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；以及

当接收到数据流时，根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

在一种可能的设计中，每个计算节点的信息包括计算节点的编号；

所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，包括：

计算每个计算节点的编号与多条等价转发路径数量的商的余数；以及

根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

在一种可能的设计中，每个计算节点的编号通过边界网关协议 BGP 路由的扩展属性表达。

在一种可能的设计中，所述方法还包括：

根据所连接的多个计算节点的数量，依次对所述所连接的多个计算节点进行编号。

在一种可能的设计中，每个计算节点的信息包括计算节点的网际互连协议 IP 地址或者介质访问控制 MAC 地址；

在所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，所述方法还包括：

生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与所述计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。

在一种可能的设计中，在所述根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，所述方法还包括：

根据所述目标计算节点的 IP 地址或 MAC 地址查询所述转发表项确定所述目标计算节点对应的转发路径。

在一种可能的设计中，所述计算节点为以服务器为个体的单个计算单元，或是以服务器上多个显示芯片 GPU 为个体的多个计算单元。

在一种可能的设计中，所述数据中心网络为脊交换机和叶子交换机组网架构，或者超脊交换机、脊交换机和叶子交换机组网架构。

第六方面，本申请还提供一种数据中心网络负载均衡系统，包括：

数据中心网络中的叶子交换机，用于对所连接的计算节点进行编号；

5 数据中心网络中的各个邻接的交换机，用于互相同步多个计算节点信息；

第一交换机，用于根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；所述第一交换机为所述数据中心网络中的各个邻接的交换机中的一个；

10 当所述第一交换机接收到数据流时，所述第一交换机，还用于根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

在一种可能的设计中，所述第一交换机，用于计算每个计算节点的编号与所述多条等价转发路径数量的商的余数；以及

根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

15 在一种可能的设计中，每个计算节点的信息还包括计算节点的 IP 地址或 MAC 地址；

在所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，还包括：

20 生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址或 MAC 地址的对应关系。

在一种可能的设计中，

在所述根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，还包括：

25 根据所述目标计算节点的 IP 地址或 MAC 地址查询所述转发表项确定所述目标计算节点对应的转发路径。

第七方面，本申请还提供一种数据中心网络负载均衡装置，包括：

接收单元，用于获取多个计算节点的信息；

30 处理单元，用于根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；

发送单元，用于当接收到数据流时，根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

35 在一种可能的设计中，

每个计算节点的信息包括计算节点的编号；

40 所述处理单元，用于在根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径时，计算每个计算节点的编号与多条等价转发路径数量的商的余数；根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

在一种可能的设计中，

每个计算节点的信息还包括计算节点的 IP 地址或 MAC 地址；

在所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，还包括：

5 生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。

在一种可能的设计中，

10 在所述根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，还包括：

根据所述目标计算节点的 IP 地址查询所述转发表项确定所述目标计算节点对应的转发路径。

15 第八方面，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机执行指令，所述计算机执行指令被处理器执行时用于实现上述第五方面或第六方面中任一方面所述的方法。

第九方面，本申请还提供一种计算机程序产品，包括计算机程序，该计算机程序被处理器执行时实现上述第五方面或第六方面中任一方面所述的方法。

20

附图说明

25 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简要介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 为本申请实施例提供的数据中心网络架构示意图。

图 2 为本申请实施例提供的数据中心网络负载均衡方法中计算节点和叶子交换机连接规则示意图。

30 图 3 为本申请实施例提供的数据中心网络负载均衡方法中叶子交换机和脊交换机连接规则示意图。

图 4 为本申请实施例提供的数据中心网络负载均衡方法中脊交换机和超脊交换机连接规则示意图。

图 5 为本申请实施例提供的数据中心网络负载均衡方法的流程示意图。

35 图 6 为本申请实施例提供的数据中心网络负载均衡方法同号 GPU 通信场景路径转发示意图。

图 7 为本申请实施例提供的数据中心网络负载均衡方法异号 GPU 通信场景路径转发示意图。

图 8 为本申请实施例提供的数据中心网络负载均衡系统的流程示意图。

图 9 为本申请实施例提供的数据中心网络负载均衡装置的结构示意图一。

40 图 10 为本申请实施例提供的数据中心网络负载均衡装置的结构示意图二。

具体实施方式

为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

本申请实施例描述的应用场景是为了更加清楚的说明本申请实施例的技术方案，并不构成对于本申请实施例提供的技术方案的限定，本领域普通技术人员可知，随着新应用场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。其中，在本申请的描述中，除非另有说明，“多个”的含义是两个或两个以上。

在本申请中，使用诸如“第一”、“第二”等的序数术语来修饰元件并不表示一个元件相对于另一个元件的任何优先级，位次或顺序，或者执行方法中的动作的时间顺序。除非另外特别说明，否则此类序数词仅用作标签以将具有特定名称的一个元件与具有(除序数词外)相同名称的另一元件区分开。

在使用本申请中描述的“包括”、“具有”、和“包含”的情况下，除非使用了明确的限定用语，例如“仅”、“由……组成”等，否则还可以添加另一部件。除非相反地提及，否则单数形式的术语可以包括复数形式，并不能理解为其数量仅为一个。

传统的负载均衡技术主要有基于网络的均衡包交换技术、基于网卡(Network Interface Card, NIC)的均衡包交换技术以及静态 ECMP。

基于网络的均衡包交换技术原理为：交换机将大象流中的报文打散形成较小的数据包或者网络切片，所谓大象流即单流带宽大的数据流，并将较小的数据包或者网络切片均匀地转发给交换机的多个等价路径出口，在交换机出口对较小的数据包或者网络切片进行重组。基于网络的均衡包交换技术能够让数据流最大程度地均匀分布，从而避免链路拥塞，但是可能会出现报文乱序的情况，不能保证送达目的地址的数据流顺序到达。

基于 NIC 的均衡包交换技术与基于网络的均衡包交换技术原理类似，不同的是基于 NIC 的均衡包交换技术是在服务器网卡进行报文重组，而不是在交换机出口进行报文重组，但是同样可能会出现报文乱序的情况。

静态 ECMP 是基于流的五元组进行哈希选路，从多条等价路径中随机选择一条进行转发，对于大小流密集且随机分布的通用计算场景，静态 ECMP 的负载均衡效果还是较为理想的。但是在 AI 集群训练场景中，RDMA 通信的流量大多为大象流，当交换机入口流量大于出口流量的带宽时，会造成交换机发生网络拥塞，交换机缓存可处理临时拥塞，但当拥塞过长时，交换机缓存会过载，交换机下一个收到的新数据包将被丢弃。

显示拥塞通告算法(Explicit Congestion Notification, ECN)和基于优先级的流量控制(Priority-based Flow Control, PFC)都是对上述交换机网络过载的被动响应。ECN 通过交换机对拥塞的判断，显示拥塞标记，发送数据流的主机通过网络中返回的带拥塞标记的包发现拥塞，则减少向带有拥塞标记的交换机发送数据流或发送单流带宽较小的数据流，可以减少丢包情况的出

现。PFC 的主要原理是如果交换机检测到拥塞，则主动发送消息给交换机的上游设备，要求上游设备暂停数据流的发送，等待一定时间后再继续发送数据流，可以达到无丢包的要求。

本申请实施例应用的数据中心网络采用 CLOS 架构互联，该架构由贝尔实验室一位叫 Charles Clos 的人提出，所以命名为 CLOS 架构。CLOS 架构依赖 RDMA 进行网络通信，可以提供无阻塞的网络，常见的有两层 CLOS 架构和三层 CLOS 架构，两层 CLOS 架构包括叶子（Leaf）交换机和脊（Spine）交换机，三层 CLOS 架构如图 1 所示，包括叶子交换机、脊交换机和超脊（Super Spine）交换机，其中，叶子交换机可简称为 LE，脊交换机可简称为 SP，超脊交换机可简称为 SSP。

示例性地，本申请的数据中心网络为脊交换机和叶子交换机组网架构，或者超脊交换机、脊交换机和叶子交换机组网架构。随着网络技术的发展，本申请的数据中心网络架构可能扩展为四层 CLOS 架构、五层 CLOS 架构.....，本申请对此不做赘述。

示例性地，一个叶子交换机具有 128 个接口，同样地，一个脊交换机和一个超脊交换机均有 128 个接口，其中，叶子交换机有 64 个接口用于下行链路和计算节点连接，剩余 64 个接口用于上行链路和脊交换机连接；脊交换机有 64 个接口用于下行链路和叶子交换机连接，剩余 64 个接口用于上行链路和超脊交换机连接；超脊交换机的 128 个接口全部用于下行链路和脊交换机连接。

计算节点、叶子交换机、脊交换机以及超脊交换机之间的连接规则如下：

（1）计算节点和叶子交换机的连接规则为：计算节点组 i 中的第 j 个 GPU 口能和叶子交换机组 i 中的第 j 个叶子交换机连接，其中，i 和 j 均为正整数。如图 2 所示，计算节点有两组，每组各有 64 个计算节点，编号为 1~64，每个计算节点有 8 个 GPU 口，编号为 1~8。计算节点组 1 中的计算节点 1~计算节点 64 中每个计算节点的 1 号 GPU 口能和叶子交换机组 1 中的 LE1 连接，计算节点组 1 中的计算节点 1~计算节点 64 中每个计算节点的 2 号 GPU 口能和叶子交换机组 1 中的 LE2 连接，.....，计算节点组 1 中的计算节点 1~计算节点 64 中每个计算节点的 8 号 GPU 口能和叶子交换机组 1 中的 LE8 连接。类似地，计算节点组 2 中的计算节点 1~计算节点 64 中每个计算节点的 1 号 GPU 口能和叶子交换机组 2 中的 LE1 连接，计算节点组 2 中的计算节点 1~计算节点 64 中每个计算节点的 2 号 GPU 口能和叶子交换机组 2 中的 LE2 连接，.....，计算节点组 2 中的计算节点 1~计算节点 64 中每个计算节点的 8 号 GPU 口能和叶子交换机组 2 中的 LE8 连接。

（2）叶子交换机和脊交换机的连接规则为：叶子交换机组中第 p 个交换机能和脊交换机组 p 中的任意一个脊交换机连接，其中，p 为正整数。如图 3 所示，叶子交换机有两组，每组各有 8 个叶子交换机，编号为 1~8，脊交换机有 8 组，每组各有 64 个脊交换机，编号为 1~64。叶子交换机组 1 中的 LE1 能和脊交换机组 1 的 SP1~SP64 中的任意一个脊交换机连接，叶子交换机组 1 中的 LE2 能和脊交换机组 2 的 SP1~SP64 中的任意一个脊交换机连接，.....，

叶子交换机组 1 的 LE8 能和脊交换机组 8 的 SP1~SP64 中的任意一个脊交换机连接。类似地，叶子交换机组 2 中的 LE1 能和脊交换机组 1 的 SP1~SP64 中的任意一个脊交换机连接，叶子交换机组 2 中的 LE2 能和脊交换机组 2 的 SP1~SP64 中的任意一个脊交换机连接，.....，叶子交换机组 2 中的 LE8 能和脊交换机组 8 的 SP1~SP64 中的任意一个脊交换机连接。

(3) 脊交换机和超脊交换机之间的连接规则为：脊交换机组中的第 t 个脊交换机能和超脊交换机组 t 中的任意一个超脊交换机连接，其中， t 为正整数。如图 4 所示，脊交换机有 2 组，每组各有 64 个脊交换机，编号为 1~64；超脊交换机有 64 组，每组有 4 个超脊交换机，编号为 1~4。脊交换机组 1 中的 SP1 能和超脊交换机组 1 的 SSP1~SSP4 中的任意一个超脊交换机连接，脊交换机组 1 中的 SP2 能和超脊交换机组 2 的 SSP1~SSP4 中的任意一个超脊交换机连接，.....，脊交换机组 1 中的 SP64 能和超脊交换机组 64 的 SSP1~SSP4 中的任意一个超脊交换机连接。类似地，脊交换机组 2 中的 SP1 能和超脊交换机组 1 的 SSP1~SSP4 中的任意一个超脊交换机连接，脊交换机组 2 中的 SP2 能和超脊交换机组 2 的 SSP1~SSP4 中的任意一个超脊交换机连接，.....，脊交换机组 2 中的 SP64 能和超脊交换机组 64 的 SSP1~SSP4 中的任意一个超脊交换机连接。

基于上述静态 ECMP 在 AI 集群训练场景中失效的问题以及数据中心网络，本申请提出了一种数据中心网络负载均衡方法，用以实现数据流均匀地分布在全路径上。可以理解的是，下述方法的执行主体可以为数据中心网络中的任意一个交换机，例如，如图 1 中的叶子交换机、或脊交换机、或超脊交换机。

如图 5 所示，该方法包括如下步骤：

步骤 500：获取多个计算节点的信息。

示例性地，每个计算节点的信息包括计算节点的编号，数据中心网络中的每个叶子交换机根据该叶子交换机所连接的计算节点的数量，依次对所连接的计算节点进行编号，其中，计算节点为以服务器为个体的单个计算单元，或是以服务器上多个 GPU 为个体的多个计算单元，如上文所述，一个叶子交换机和计算节点连接的接口最多有 64 个，所以一个叶子交换机最多可连接 64 个计算节点。

例如，如图 6 所示的叶子节点组 1 中的 LE1 对所连接的计算节点组 1 中的 64 个计算节点的编号为：计算节点组 1 中计算节点 1 的编号为 1，计算节点组 1 中计算节点 2 的编号为 2，.....，计算节点组 1 中计算节点 64 的编号为 64；或者，计算节点组 1 中计算节点 1 的编号为 0，计算节点组 1 中计算节点 2 的编号为 1，.....，计算节点组 1 中计算节点 64 的编号为 63，本申请对此不做限定。

同样地，叶子节点组 2 中的 LE1 对所连接的计算节点组 2 中的 64 个计算节点的编号为：计算节点组 2 中计算节点 1 的编号为 1，计算节点组 2 中计算节点 2 的编号为 2，.....，计算节点组 2 中计算节点 64 的编号为 64。

示例性地，数据中心网络中的各个邻接的交换机之间通过 BGP 的路由通

告获取多个计算节点的信息，每个计算节点的编号通过 BGP 路由的扩展属性表达，BGP 路由的扩展属性可用标签（tag）表示，tag 值为上述计算节点对应的编号。

例如，叶子交换机 1 具有多个计算节点的信息，分别是计算节点①、计算节点②以及计算节点③，叶子交换机 1 连接的脊交换机 1 具有多个计算节点的信息，分别是计算节点④、计算节点⑤以及计算节点⑥。各个邻接的交换机之间通过 BGP 的路由通告获取多个计算节点的信息，即叶子交换机 1 会将计算节点信息①、②、③通过 BGP 的路由通告同步给脊交换机 1，脊交换机 1 也会将计算节点信息④、⑤、⑥通过 BGP 的路由通告同步给叶子交换机 1。同样地，邻接的脊交换机和超脊交换机也会通过 BGP 的路由通告互相同步多个计算节点的信息，由上述脊交换机和超脊交换机连接规则可知：一个脊交换机最多和 4 个超脊交换机连接，那么该脊交换机会通过 BGP 的路由通告将多个计算节点的信息同步给连接的 4 个超脊交换机，各超脊交换机也会将多个计算节点的信息同步给该脊交换机。

示例性地，每个计算节点的信息还包括计算节点的 IP 地址，网络中每个节点都有一个唯一的 IP 地址，IP 地址由 32 位二进制数组成，通常被表示为 4 个十进制数，每个十进制数的范围为 0~255，每个十进制数之间使用“.”进行分割，例如，128.1.1.4 和 100.2.1.3 都是一个 IP 地址。

示例性地，IP 地址由两部分组成：网段地址和主机地址，128.1.1.4/24 表示前 24 位是网段地址，后 8 位是主机地址；128.1.1.4/8 表示前 8 是网段地址，后 24 位是主机地址，“/”后面的数字可称为子网掩码，本申请中子网掩码均采用 24。

步骤 510: 根据多个计算节点的信息确定多个计算节点分别对应的转发路径，多个计算节点分别对应的转发路径为等价转发路径，且多个计算节点分别对应的转发路径不同。

示例性地，等价转发路径为等价多路径路由 ECMP，等价多路径路由 ECMP 是从交换机到交换机的下一跳交换机开销相同的路径。

下面以当前的交换机称为第一交换机，该交换机的下一跳交换机称为第二交换机为例进行说明。

如图 6 所示，假设第一交换机为叶子交换机组 1 中的 LE1，第二交换机所属层级为 Spine 层，一个脊交换机可与叶子交换机连接的接口数 M 为 64；叶子交换机共有 64 组，每组都有一个 LE1，Leaf 层共有 64 个 LE1，设 N 为 64；由上述叶子交换机和脊交换机的连接规则可知，叶子交换机组 1 中的 LE1 可以与脊交换机组 1 的 SP1~SP64 中的任意一个脊交换机连接，即第二交换机所属层级中可与第一交换机连接的交换机个数 K 为 64，因此，第一交换机 LE1 确定的等价转发路径数量为 $M/N \times K = 64/64 \times 64 = 64$ ，分别是路径 1：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP1、路径 2：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP2、路径 3：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP3、.....、路径 64：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP64，LE1 选择上述 64 条等价转发路径中的任意一条到达第二交换机的开销都是相

同的。

在一种可能的实现方式中，在根据多个计算节点的信息确定多个计算节点分别对应的转发路径时，计算每个计算节点的编号与等价转发路径数量的商的余数，根据余数从多条等价转发路径中确定与该计算节点对应的转发路径。因为计算节点的信息所包含的计算节点的编号不同，所以计算节点的编号与等价转发路径数量的商的余数不同，进而根据余数确定的转发路径不同，实现计算节点的编号与转发路径的均匀绑定。

示例一，第一交换机直接根据余数从等价转发路径中确定与该计算节点对应的转发路径。例如，假设等价转发路径数量为 10，根据每个计算节点的编号与等价转发路径数量的商的余数进行转发路径的确定，若余数为 0，则表示确定等价转发路径中的第一条路径为转发路径；若余数为 1，则表示确定等价转发路径中的第二条路径为转发路径；.....；若余数为 9，表示确定等价转发路径中的第十条路径为转发路径。

示例二，第一交换机根据余数与预设数值之和从多条等价路径中确定与该计算节点对应的转发路径。

例如，假设等价转发路径数量为 10，预设数值为 2，若余数为 0，则余数与预设数值之和为 2，表示确定多条等价转发路径中的第一条路径为转发路径；若余数为 1，则余数与预设数值之和为 3，表示确定多条等价转发路径中的第二条路径为转发路径；.....；若余数为 9，则余数与预设数值之和为 11，表示确定多条等价转发路径中的第十条路径为转发路径。

示例三，第一交换机根据多条等价转发路径的总数与余数的差值从多条等价路径中确定与该计算节点对应的转发路径。

例如，假设等价转发路径数量为 10，若余数为 0，则等价转发路径的总数与余数的差值为 10，表示确定多条等价转发路径中的第十条路径为转发路径；若余数为 1，则等价转发路径的总数与余数的差值为 9，表示确定多条等价转发路径中的第九条路径为转发路径；.....；若余数为 9，则等价转发路径的总数与余数的差值为 1，表示确定等价转发路径中的第一条路径为转发路径。

可以理解的是，上述示例一至示例三为举例，本申请对此不做限定。

进一步示例性地，在根据多个计算节点的信息确定多个计算节点分别对应的转发路径之后，生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。

示例一，假设第一交换机为图 6 中叶子交换机组 1 中的 LE1，由上文可知：叶子交换机组 1 中的 LE1 确定的等价转发路径数量为 64，分别为：路径 1：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP1；路径 2：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP2；路径 3：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP3；.....；路径 64：叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP64。叶子交换机组 1 中的 LE1 根据计算节点信息①中的 tag 值为 1，与等价转发路径数量 64 的商的余数为 1，确定其转发路径为路径 2，即叶子交换机组 1 中的 LE1->脊交换机组 1 的 SP2，生成一条转发表项，包括：

100.1.1.0/24 tag=1 脊交换机组 1 的 SP2; 同样地, 根据计算节点信息②中的 tag 值为 2 确定其转发路径为路径 3, 即叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP3, 生成一条转发表项, 包括: 100.1.2.0/24 tag=2 脊交换机组 1 中的 SP3; 根据计算节点信息③中的 tag 值为 3 确定其转发路径为路径 4, 即
5 叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP4, 生成一条转发表项, 包括: 100.1.3.0/24 tag=3 脊交换机组 1 中的 SP4.....以此类推, 叶子交换机组 1 中的 LE1 可以生成每个计算节点对应的转发表项。

示例二, 假设第一交换机为图 6 中叶子交换机组 1 中的 LE2, 由上文可知: 叶子交换机组 1 中的 LE2 确定的等价转发路径数量为 64, 分别为: 路径
10 1: 叶子交换机组 1 中的 LE2->脊交换机组 1 中的 SP1; 路径 2: 叶子交换机组 1 中的 LE2->脊交换机组 1 中的 SP2; 路径 3: 叶子交换机组 1 中的 LE2->脊交换机组 1 中的 SP3;; 路径 64: 叶子交换机组 1 中的 LE2->脊交换机组 1 中的 SP64。叶子交换机组 1 中的 LE2 根据计算节点信息①生成一条转发表项, 包括: 100.1.1.0/24 tag=1 脊交换机组 1 中的 SP2; 同样地, 根据计算
15 节点信息②生成一条转发表项, 包括: 100.1.2.0/24 tag=2 脊交换机组 1 中的 SP3; 根据计算节点信息③生成一条转发表项, 包括: 100.1.3.0/24 tag=3 脊交换机组 1 中的 SP4.....以此类推, 叶子交换机组 1 中的 LE2 可以生成每个计算节点对应的转发表项。

示例三, 假设第一交换机为图 7 中脊交换机组 1 中的 SP4, 由上文可知:
20 脊交换机组 1 中的 SP4 确定的等价转发路数量为 64, 分别为: 路径 1: 脊交换机组 1 中的 SP4->超脊交换机组 4 中的 SSP1 的接口 1; 路径 2: 脊交换机组 1 中的 SP4->超脊交换机组 4 中的 SSP1 的接口 2;; 路径 17: 脊交换机组 1 中的 SP4->超脊交换机组 4 中的 SSP2 的接口 1; 路径 18: 脊交换机组 1 中的 SP4->超脊交换机组 4 中的 SSP2 的接口 2;; 路径 64: 脊交换机组
25 1 中的 SP4->超脊交换机组 4 中的 SSP4 的接口 16。脊交换机组 1 中的 SP4 根据计算节点信息①生成一条转发表项, 包括: 100.1.1.0/24 tag=1 超脊交换机组 4 中的 SSP1; 根据计算节点信息②生成一条转发表项, 包括: 100.1.2.0/24 tag=2 超脊交换机组 4 中的 SSP1; 根据计算节点信息③生成一条转发表项, 包括: 100.1.3.0/24 tag=3 超脊交换机组 4 中的 SSP1;; 根据计算节点信
30 息⑦生成一条转发表项, 包括: 100.1.17.0/24 tag=1 超脊交换机组 4 中的 SSP2.....以此类推, 脊交换机组 1 中的 SP4 可以生成每个计算节点对应的转发表项。

步骤 520: 当接收到数据流时, 根据数据流的目标计算节点对应的转发路径转发数据流。

35 示例性地, 在根据数据流的目标计算节点对应的转发路径转发数据流之前, 根据目标计算节点的 IP 地址查询转发表项确定目标计算节点对应的转发路径, 按照与目标计算节点的 IP 地址对应的转发路径转发数据流。

示例一, 以图 6 的实线所表示的数据流 1 为例, 计算节点组 1 中的计算节点 1 的 2 号 GPU 口向计算节点组 2 中的计算节点 1 的 2 号 GPU 口传输数据, 则起始计算节点为计算节点组 1 中的计算节点 1, 目标计算节点为计算节
40

点组 2 中的计算节点 1, 其中, 计算节点组 2 中的计算节点 1 的 IP 地址为 100.2.1.2。

根据上述计算节点和叶子交换机的连接规则可知, 计算节点组 1 中的计算节点 1 的 2 号 GPU 口向叶子交换机组 1 中的 LE2 转发数据流, 当叶子交换机组 1 中的 LE2 接收到来自计算节点组 1 中的计算节点 1 的 2 号 GPU 口发送的数据流时, 叶子交换机组 1 中的 LE2 根据目标计算节点的 IP 地址 100.2.1.2 查询转发表项, 匹配到的转发表项为: 100.2.1.0/24 tag=1 脊交换机组 1 中的 SP2, 从而确定的转发路径为: 叶子交换机组 1 中的 LE2->脊交换机组 1 中的 SP2, 即叶子交换机组 1 中的 LE2 向脊交换机组 1 中的 SP2 转发数据流。根据上述计算节点和叶子交换机的规则可知: 计算节点组 2 中的计算节点 1 的 2 号 GPU 口能和叶子交换组 2 中的 LE2 连接, 所以脊交换机组 1 中的 SP2 将数据流 1 转发给叶子交换组 2 中的 LE2, 叶子交换组 2 中的 LE2 再将数据流转发给目的计算节点。因此, 图 6 的实线所表示的数据流 1 完整的转发路径为: 计算节点组 1 中的计算节点 1 的 2 号 GPU 口->叶子交换机组 1 中的 LE2->脊交换机组 1 中的 SP2->叶子交换组 2 中的 LE2->计算节点组 2 中的计算节点 1 的 2 号 GPU 口。

示例二, 以图 6 的虚线所表示的数据流 2 为例, 计算节点组 1 中的计算节点 2 的 1 号 GPU 口向计算节点组 2 中的计算节点 3 的 1 号 GPU 口传输数据, 则起始计算节点为计算节点组 1 中的计算节点 2, 目标计算节点为计算节点组 2 中的计算节点 3, 其中, 计算节点组 2 中的计算节点 3 的 IP 地址为 100.2.3.2。

根据上述计算节点和叶子交换机的连接规则可知, 计算节点组 1 中的计算节点 2 的 1 号 GPU 口向叶子交换机组 1 中的 LE1 转发数据流, 当叶子交换机组 1 中的 LE1 接收到来自计算节点组 1 中的计算节点 2 的 1 号 GPU 口发送的数据流时, 叶子交换机组 1 中的 LE1 根据目标计算节点的 IP 地址 100.2.3.2 查询转发表项, 匹配到的转发表项为: 100.2.3.2/24 tag=3 脊交换机组 1 中的 SP4, 从而确定的转发路径为: 叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP4, 即叶子交换机组 1 中的 LE1 向脊交换机组 1 中的 SP4 转发数据流。根据上述计算节点和叶子交换机的规则可知: 计算节点组 2 中的计算节点 3 的 1 号 GPU 口能和叶子交换组 2 中的 LE1 连接, 所以脊交换机组 1 中的 SP4 将数据流 2 转发给叶子交换组 2 中的 LE1, 叶子交换组 2 中的 LE1 再将数据流 2 转发给目的计算节点。因此, 图 6 的虚线所表示的数据流 2 完整的转发路径为: 计算节点组 1 中的计算节点 2 的 1 号 GPU 口->叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP4->叶子交换组 2 中的 LE1->计算节点组 2 中的计算节点 3 的 1 号 GPU 口。

示例三, 以图 7 所表示的数据流为例, 计算节点组 1 中的计算节点 2 的 1 号 GPU 口向计算节点组 2 中的计算节点 3 的 2 号 GPU 口传输数据, 则起始计算节点为计算节点组 1 中的计算节点 2, 目标计算节点为计算节点组 2 中的计算节点 3, 其中, 计算节点组 2 中的计算节点 3 的 IP 地址为 100.2.3.2。

叶子交换机组 1 中的 LE1 向脊交换机组 1 中的 SP4 转发数据流后, 在脊

交换机组 1 中的 SP4 根据目标计算节点的 IP 地址 100.2.3.2 查询转发表项, 匹配到的转发表项为: 100.2.3.2/24 tag=3 超脊交换机组 4 中的 SSP1, 从而确定的转发路径为: 脊交换机组 1 中的 SP4->超脊交换机组 4 中的 SSP1, 即脊交换机组 1 中的 SP4 向超脊交换机组 4 中的 SSP1 转发数据流。根据上述计算节点和叶子交换机的连接规则可知: 计算节点组 2 中的计算节点 3 的 2 号 GPU 口能和叶子交换机组 2 中的 LE2 通信, 所以超脊交换机组 4 中的 SSP1 的下一个转发节点属于脊交换机组 2, 又因为超脊交换机组 4 中的 SSP1 与 SP 层的 SP4 通信, 进而确定下一个转发节点为脊交换机组 2 中的 SP4, 所以超脊交换机组 4 中的 SSP1 将数据流转发给脊交换机组 2 中的 SP4。根据上述计算节点和叶子交换机的连接规则可知: 计算节点组 2 中的计算节点 3 的 2 号 GPU 口能和叶子交换组 2 中的 LE2 连接, 所以脊交换机组 2 中的 SP4 将数据流转发给叶子交换组 2 中的 LE2, 最后, 叶子交换组 2 中的 LE2 将数据流转发给目标计算节点。

因此, 图 7 所表示的数据流完整的转发路径为: 计算节点组 1 中的计算节点 2 的 1 号 GPU 口->叶子交换机组 1 中的 LE1->脊交换机组 1 中的 SP4->超脊交换机组 4 中的 SSP1->脊交换机组 2 中的 SP4->叶子交换组 2 中的 LE2->计算节点组 2 中的计算节点 3 的 2 号 GPU 口。

可以理解的是, 在根据数据流的目标计算节点确定对应的转发路径之后, 若确定该转发路径存在链路故障, 则可以从多条等价转发路径中随机重新确定一条其他转发路径, 根据重新确定的转发路径转发该数据流。其中, 重新确定的转发路径与目标计算节点对应的转发路径不同, 且重新确定的转发路径不存在链路故障。示例性地, 可以采用静态 ECMP 方法从多条等价转发路径中随机重新确定一条其他转发路径。

示例性地, 可以通过双向转发检测协议(Bidirectional Forwarding Detection, BFD) 检测是否发生链路故障。其中, 一个转发节点向另一个转发节点发送 BFD 检测报文, 若在预设时间内没有收到对方发送过来的 BFD 报文, 则确定两个转发节点之间发生了链路故障。

示例性地, 链路故障包括上行链路故障、下行链路故障、以及叶子交换机下行链路故障等。此外, 若确定出现叶子交换机下行链路故障, 还可以将数据流丢弃。

示例一, 以图 6 的实线所表示的数据流 1 为例, 计算节点组 1 中的计算节点 1 的 2 号 GPU 口向计算节点组 2 中的计算节点 1 的 2 号 GPU 口传输数据, 由上述步骤 520 可知, 叶子交换机组 1 中的 LE1 的等价转发路径有 64 条, 确定的转发路径为路径 2, 若路径 2 发生故障, 称为上行链路故障, 则叶子交换机组 1 中的 LE1 采用静态 ECMP 从剩余的 63 条等价转发路径中随机确定一条转发路径。

示例二, 以图 7 所表示的数据流为例, 计算节点组 1 中的计算节点 2 的 1 号 GPU 口向计算节点组 2 中的计算节点 3 的 2 号 GPU 口传输数据, 由上述步骤 520 可知, 超脊交换机组 4 中的 SSP1 向 SP 层转发数据流时确定的转发路径为: 超脊交换机组 4 中的 SSP1->脊交换机组 2 中的 SP4, 若该路径发生

故障，称为下行链路故障，和上行链路故障类似地，则超脊交换机组 4 中的 SSP1 采用静态 ECMP 从剩余的多条等价转发路径中随机确定一条转发路径。

示例三，仍以图 6 的实线所表示的数据流 1 为例，计算节点组 1 中的计算节点 1 的 2 号 GPU 口向计算节点组 2 中的计算节点 1 的 2 号 GPU 口传输数据，叶子交换组 2 中的 LE2 向计算节点组 2 中的计算节点 1 的 2 号 GPU 口转发数据流时，有一条等价转发路径，则确定该条等价转发路径为转发路径，若该条转发路径出现故障，称为叶子交换机下行链路故障，由于此时仅有一条转发路径，没有其他等价转发路径进行选择，所以叶子交换组 2 中的 LE2 可以将数据流丢弃。

本申请提供的数据中心网络负载均衡方法，数据中心网络中的叶子交换机按照计算节点的连接顺序对计算节点进行编号，不需要人工进行逐一配置，数据中心网络中的各个邻接的交换机之间互相同步多个计算节点信息，实现计算节点信息的自主传递，可适应网络架构的变换，根据多个计算节点的信息确定多个计算节点分别对应的转发路径，多个计算节点分别对应的转发路径为等价转发路径，且多个计算节点分别对应的转发路径不同，保证数据流在数据中心网络中的均匀分布，且当接收到数据流时，根据数据流的目标计算节点对应的转发路径转发所述数据流，不会发生报文乱序，使传输效率接近最优。

本申请提供的数据中心网络负载均衡方法，采用预先的全局路径规划进行负载均衡，非常适用于流量具有显著规律和特征的场景，例如，AI 集群训练场景，其流量大多具有单流带宽大的特征且训练任务高度并发协同。本申请提供的方法借助 BGP 的路由通告，以及扩展路由属性来实现路由自主传递，可适应网络拓扑变换，并且不需要逐一进行人工配置；此外，本申请提供的方法中所有流量都按序传输，不需要网卡或交换机对报文进行重排，因此不会出现报文乱序的问题；最后，本申请提供的方法不需要实时采集链路负载、不需要全局拓扑感知，也不需要实时计算并重路由选路，即可实现流量均匀地分布在全路径上，同时具有很好的稳定性和健壮性。

本申请还提出了一种数据中心网络负载均衡系统，用以实现数据流均匀地分布在全路径上。可以理解的是，该系统包含了图 1 中的叶子交换机、或脊交换机、或超脊交换机，其流程示意图如图 8 所示：

步骤 800：数据中心网络中的叶子交换机对所连接的计算节点进行编号。

示例性地，数据中心网络中的每个叶子交换机根据该叶子交换机所连接的计算节点的数量，依次对所连接的计算节点进行编号，其中，计算节点为以服务器为个体的单个计算单元，或是以服务器上多个 GPU 为个体的多个计算单元，如上文所述，一个叶子交换机和计算节点连接的接口最多有 64 个，所以一个叶子交换机最多可连接 64 个计算节点。

步骤 810：数据中心网络中的各个邻接的交换机之间互相同步多个计算节点信息。

示例性地，数据中心网络中的各个邻接的交换机之间通过 BGP 的路由通告互相同步多个计算节点信息，每个计算节点的信息包括计算节点的编号，

还包括计算节点的 IP 地址，其中，每个计算节点的编号通过 BGP 路由的扩展属性表达，BGP 路由的扩展属性可用标签 (tag) 表示，tag 值为步骤 800 中的计算节点对应的编号。

5 步骤 820: 第一交换机根据多个计算节点的信息确定多个计算节点分别对应的转发路径，多个计算节点分别对应的转发路径为等价转发路径，且多个计算节点分别对应的转发路径不同；第一交换机为数据中心网络中的各个邻接的交换机中的一个。

10 示例性地，等价转发路径为 ECMP，ECMP 是从交换机到交换机的下一跳交换机开销相同的路径。第一交换机根据多个计算节点的信息确定多个计算节点分别对应的转发路径时，计算每个计算节点的编号与等价转发路径数量的商的余数，根据余数从多条等价转发路径中确定与该计算节点对应的转发路径。因为计算节点的信息所包含的计算节点的编号不同，所以计算节点的编号与等价转发路径数量的商的余数不同，进而根据余数确定的转发路径不同，实现计算节点的编号与转发路径的均匀绑定。

15 示例性地，第一交换机在根据多个计算节点的信息确定多个计算节点分别对应的转发路径后，生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。

具体可以参考上述步骤 510。

20 步骤 830: 当第一交换机接收到数据流时，根据数据流的目标计算节点对应的转发路径转发数据流。

示例性地，第一交换机在根据数据流的目标计算节点对应的转发路径转发数据流之前，根据目标计算节点的 IP 地址查询转发表项确定目标计算节点对应的转发路径，按照与目标计算节点的 IP 地址对应的转发路径转发数据流。

25 具体可以参考上述步骤 520。

有关上述步骤 800、步骤 810、步骤 820 和步骤 830 更详细的描述或举例可以参考图 5 所示的方法实施例中相关描述直接得到，这里不加赘述。

可以理解的是，为了实现上述实施例中功能，主机包括了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到，结合本
30 申请中所公开的实施例描述的各示例的单元及方法步骤，本申请能够以硬件或硬件和计算机软件相结合的形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用场景和设计约束条件。

图 9 和图 10 为本申请的实施例提供的可能的数据中心网络负载均衡装置
35 的结构示意图。这些数据中心网络负载均衡装置可以用于实现上述方法实施例中交换机的功能，因此也能实现上述方法实施例所具备的有益效果。

如图 9 所示，数据中心网络负载均衡装置 900 包括接收单元 910、处理单元 920 和发送单元 930。数据中心网络负载均衡装置 900 用于实现上述图 5 所示的方法实施例中交换机的功能。

40 数据中心网络负载均衡装置 900 用于实现上述图 5 所示的方法实施例中

交换机的功能时：

所述接收单元 910，用于获取多个计算节点的信息；

所述处理单元 920，用于根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；

所述发送单元 930，用于当接收到数据流时，根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

在一种可能的设计中，所述接收单元 910，用于获取多个计算节点的信息时，通过 BGP 的路由通告获取所述多个计算节点的信息；

在一种可能的设计中，每个计算节点的信息包括计算节点的编号；

所述处理单元 920，用于在根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径时，计算每个计算节点的编号与多条等价转发路径数量的商的余数；根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

在一种可能的设计中，所述多条等价转发路径为 ECMP，所述 ECMP 是从所述交换机到所述交换机的下一跳交换机开销相同的路径。

在一种可能的设计中，所述处理单元 920，用于根据所连接的计算节点的数量，依次对所连接的计算节点进行编号。

在一种可能的设计中，每个计算节点的信息还包括计算节点的 IP 地址；

所述处理单元 920，用于在根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。

在一种可能的设计中，所述处理单元 920，用于在根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，根据所述目标计算节点的 IP 地址查询所述转发表项确定所述目标计算节点对应的转发路径。

有关上述接收单元 910、处理单元 920 和发送单元 930 更详细的描述可以直接参考图 5 所示的方法实施例相关描述直接得到，这里不加赘述。

如图 10 所示，数据中心网络负载均衡装置 1000 包括处理器 1010 和接口电路 1020，处理器 1010 和接口电路 1020 之间相互耦合。可以理解的是，接口电路 1020 可以为收发器或输入输出接口。可选的，数据中心网络负载均衡装置 1000 还可以包括存储器 1030，用于存储处理器 1010 执行的指令或存储处理器 1010 运行指令所需要的输入数据或存储处理器 1010 运行指令后产生的数据。

当数据中心网络负载均衡装置 1000 用于实现图 5 所示的方法时，处理器 1010 用于实现上述处理单元 920 的功能，接口电路 1020 用于实现上述接收单元 910 和发送单元 930 的功能。

本申请实施例中对单元的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，另外，在本申请各个实施例中的各功能单元可以集成在一个处理器中，也可以是单独物理存在，也可以两个或两个

以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

- 5 尽管已描述了本申请的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

权 利 要 求

- 1、一种数据中心网络负载均衡方法，其中，应用于数据中心网络中的通信设备，包括以下步骤：
获取多个计算节点的信息；
5 根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；以及
当接收到数据流时，根据所述数据流的目标计算节点对应的转发路径转发所述数据流。
- 10 2、如权利要求 1 所述的方法，其中，每个计算节点的信息包括计算节点的编号；
所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，包括：
计算每个计算节点的编号与多条等价转发路径数量的商的余数；以及
15 根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。
- 3、如权利要求 2 所述的方法，其中，每个计算节点的编号通过边界网关协议 BGP 路由的扩展属性表达。
- 4、如权利要求 2 所述的方法，其中，所述方法还包括：
根据所连接的多个计算节点的数量，依次对所述所连接的多个计算节点
20 进行编号。
- 5、如权利要求 1 所述的方法，其中，每个计算节点的信息包括计算节点的网际互连协议 IP 地址或者介质访问控制 MAC 地址；
在所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，所述方法还包括：
25 生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与所述计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。
- 6、如权利要求 5 所述的方法，其中，在所述根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，所述方法还包括：
根据所述目标计算节点的 IP 地址或 MAC 地址查询所述转发表项确定所
30 述目标计算节点对应的转发路径。
- 7、如权利要求 1 所述的方法，其中，所述计算节点为以服务器为个体的单个计算单元，或是以服务器上多个显示芯片 GPU 为个体的多个计算单元。
- 8、如权利要求 1-7 任一项所述的方法，其中，所述数据中心网络为脊交换机和叶子交换机组网架构，或者超脊交换机、脊交换机和叶子交换机组网
35 架构。
- 9、一种数据中心网络负载均衡系统，包括：
数据中心网络中的叶子交换机，用于对所连接的计算节点进行编号；
数据中心网络中的各个邻接的交换机，用于互相同步多个计算节点信息；
第一交换机，用于根据所述多个计算节点的信息确定所述多个计算节点

分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；所述第一交换机为所述数据中心网络中的各个邻接的交换机中的一个；

当所述第一交换机接收到数据流时，所述第一交换机，还用于根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

10、如权利要求 9 所述的系统，其中，所述第一交换机，用于计算每个计算节点的编号与所述多条等价转发路径数量的商的余数；以及

根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

11、如权利要求 9 所述的系统，其中，每个计算节点的信息还包括计算节点的 IP 地址或 MAC 地址；

在所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，还包括：

生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址或 MAC 地址的对应关系。

12、如权利要求 11 所述的系统，其中，

在所述根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，还包括：

根据所述目标计算节点的 IP 地址或 MAC 地址查询所述转发表项确定所述目标计算节点对应的转发路径。

13、一种数据中心网络负载均衡装置，包括：

接收单元，用于获取多个计算节点的信息；

处理单元，用于根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径，所述多个计算节点分别对应的转发路径为等价转发路径，且所述多个计算节点分别对应的转发路径不同；

发送单元，用于当接收到数据流时，根据所述数据流的目标计算节点对应的转发路径转发所述数据流。

14、如权利要求 13 所述的装置，其中，每个计算节点的信息包括计算节点的编号；

所述处理单元，用于在根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径时，计算每个计算节点的编号与多条等价转发路径数量的商的余数；根据所述余数从所述多条等价路径中确定与该计算节点对应的转发路径。

15、如权利要求 13 所述的装置，其中，

每个计算节点的信息还包括计算节点的 IP 地址或 MAC 地址；

在所述根据所述多个计算节点的信息确定所述多个计算节点分别对应的转发路径之后，还包括：

生成每个计算节点对应的转发表项，每个计算节点对应的转发表项指示与该计算节点对应的转发路径和该计算节点的 IP 地址的对应关系。

16、如权利要求 15 所述的装置， 其中

在所述根据所述数据流的目标计算节点对应的转发路径转发所述数据流之前，还包括：

5 根据所述目标计算节点的 IP 地址查询所述转发表项确定所述目标计算节点对应的转发路径。

17、一种通信设备，包括：一个或多个处理器和一个或多个存储器；其中，所述一个或多个存储器存储有一个或多个程序，当所述程序被所述一个或多个处理器执行时，使得所述装置执行如权利要求 1-8 中任一项所述的方法。

10 18、一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机执行指令，所述计算机执行指令被处理器执行时用于实现如权利要求 1-8 中任一项所述的方法。