



(12) 发明专利

(10) 授权公告号 CN 109861925 B

(45) 授权公告日 2021.12.21

(21) 申请号 201711245559.0

(22) 申请日 2017.11.30

(65) 同一申请的已公布的文献号
申请公布号 CN 109861925 A

(43) 申请公布日 2019.06.07

(73) 专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72) 发明人 刘芳 黄永成 吴炜捷

(74) 专利代理机构 广州三环专利商标代理有限公司 44202
代理人 郝传鑫 熊永强

(51) Int.Cl.
H04L 12/803 (2013.01)

(56) 对比文件

CN 101399770 A, 2009.04.01
US 2015127797 A1, 2015.05.07

审查员 邹海芳

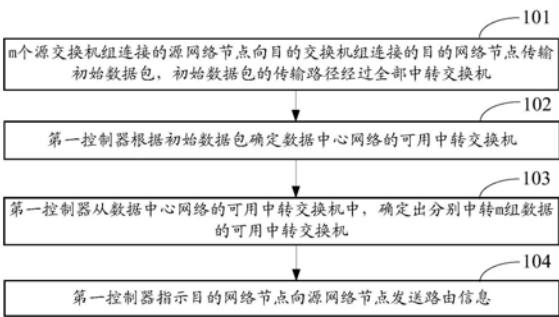
权利要求书3页 说明书17页 附图4页

(54) 发明名称

数据传输方法、相关装置及网络

(57) 摘要

本申请公开了一种数据传输方法、相关装置及网络。该方法可包括：第一控制器从数据中心网络的可用中转交换机中，确定出分别中转 m 组数据的可用中转交换机；其中，一个可用中转交换机用于中转至少一组所述数据，且任意两个可用中转交换机中转所述数据的组数的差值不超过第二阈值；所述第一控制器指示所述目的网络节点向所述源网络节点发送路由信息，所述路由信息包括用于中转所述源网络节点向所述目的网络节点传输的数据组的可用中转交换机的标识。上述方案能够实现数据中心网络中各个中转交换机的负载均衡，提高带宽资源利用率，减小传输时延。



1. 一种数据传输方法,其特征在于,所述方法应用于数据中心网络中的第一控制器,所述数据中心网络还包括多个中转交换机、 m 个源交换机组,目的交换机组、源网络节点、目的网络节点;一个所述源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输的数据称为一组数据; m 为正整数;所述方法包括:

所述第一控制器根据packing序列组,确定分别中转 m 组数据的可用中转交换机,所述可用中转交换机为所述数据中心网络中负载不超过第一阈值的中转交换机,所述packing序列组根据 v 、 k 、 m 确定, v 为所述可用中转交换机的数量, k 为所述第一控制器确定的中转一组所述数据的可用中转交换机的数量;

其中,所述packing序列组包括 m 个packing序列;所述packing序列包括 v 个元素,且,当 $v > k$ 时,所述 v 个元素中有 k 个元素取第三值,当 $v \leq k$ 时,所述 v 个元素取第三值;在一个packing序列组中,任意一个元素最多取 $\lceil mk/v \rceil$ 次第三值;

一个源交换机组对应一个packing序列,所述packing序列中的所述 v 个元素分别对应所述数据中心网络中的 v 个可用中转交换机;当一个元素的取值为第三值时,所述元素对应的可用中转交换机用于中转所述packing序列对应的源交换机组连接的源网络节点向所述目的交换机组连接的目的网络节点传输的数据;

所述第一控制器指示所述目的网络节点向所述源网络节点发送路由信息,所述路由信息包括用于中转所述源网络节点向所述目的网络节点传输的一组数据的可用中转交换机的标识。

2. 如权利要求1所述的方法,其特征在于,所述第一控制器根据packing序列组确定分别中转 m 组数据的可用中转交换机之前,还包括:

所述第一控制器获取至少一个数据包;

在所述数据包中的拥塞显示指示域的取值为第一值的情况下,确定发送所述数据包的中转交换机为所述可用中转交换机;

在所述数据包中的拥塞显示指示域的取值为第二值的情况下,确定发送所述数据包的中转交换机在预设时长后为所述可用中转交换机。

3. 如权利要求2所述的方法,其特征在于,所述至少一个数据包来自所述多个中转交换机,或者,所述至少一个数据包来自上一次数据传输中的可用中转交换机。

4. 如权利要求1所述的方法,其特征在于,所述路由信息承载在应答信号中。

5. 如权利要求1所述的方法,其特征在于,

所述数据中心网络为 n 元胖树网络;所述多个中转交换机的数量为 $n^2/4$,每 $n/2$ 个中转交换机组成一个中转交换机组;

所述 m 个源交换机组和所述目的交换机组中,任意一个交换机组包括: $n/2$ 个汇聚交换机和 $n/2$ 个边缘交换机;

其中,所述 $n/2$ 个汇聚交换机分别和所述 $n/2$ 个边缘交换机连接;

所述 $n/2$ 个汇聚交换机分别和不同的中转交换机组中的 $n/2$ 个中转交换机连接;

所述 $n/2$ 个边缘交换机分别和不同的网络节点连接。

6. 如权利要求1所述的方法,其特征在于,

所述数据中心网络为叶脊网络,所述 m 个源交换机组和所述目的交换机组中,任意一个交换机组包括:一个边缘交换机;

其中,所述边缘交换机和所述多个中转交换机连接,不同的边缘交换机分别和不同的网络节点连接。

7. 一种控制器,其特征在于,所述控制器位于数据中心网络中,所述数据中心网络还包括多个中转交换机、 m 个源交换机组,目的交换机组、源网络节点、目的网络节点;一个所述源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输的数据称为一组数据; m 为正整数;所述控制器包括:确定单元,指示单元,其中:

所述确定单元,用于根据packing序列组确定分别中转 m 组数据的可用中转交换机,所述可用中转交换机为所述数据中心网络中负载不超过第一阈值的中转交换机,所述packing序列组根据 v 、 k 、 m 确定, v 为所述可用中转交换机的数量, k 为所述控制器确定的中转一组所述数据的可用中转交换机的数量;

其中,所述packing序列组包括 m 个packing序列;所述packing序列包括 v 个元素,且,当 $v > k$ 时,所述 v 个元素中有 k 个元素取第三值,当 $v \leq k$ 时,所述 v 个元素取第三值;在一个packing序列组中,任意一个元素最多取 $\lceil mk/v \rceil$ 次第三值;

一个源交换机组对应一个packing序列,所述packing序列中的所述 v 个元素分别对应所述数据中心网络中的 v 个可用中转交换机;当一个元素的取值为第三值时,所述元素对应的可用中转交换机用于中转所述packing序列对应的源交换机组连接的源网络节点向所述目的交换机组连接的目的网络节点传输的数据;

所述指示单元,用于指示所述目的网络节点向所述源网络节点发送路由信息,所述路由信息包括用于中转所述源网络节点向所述目的网络节点传输的一组数据的可用中转交换机的标识。

8. 如权利要求7所述的控制器,其特征在于,所述控制器还包括获取单元,所述获取单元用于获取至少一个数据包;

所述确定单元还用于,在所述数据包中的拥塞显示指示域的取值为第一值的情况下,确定发送所述数据包的中转交换机为所述可用中转交换机;

在所述数据包中的拥塞显示指示域的取值为第二值的情况下,确定发送所述数据包的中转交换机在预设时长后为所述可用中转交换机。

9. 如权利要求8所述的控制器,其特征在于,所述至少一个数据包来自所述多个中转交换机,或者,所述至少一个数据包来自上一次数据传输中的可用中转交换机。

10. 如权利要求7所述的控制器,其特征在于,所述路由信息承载在应答信号中。

11. 如权利要求7所述的控制器,其特征在于,

所述数据中心网络为 n 元胖树网络;所述多个中转交换机的数量为 $n^2/4$,每 $n/2$ 个中转交换机组成一个中转交换机组;

所述 m 个源交换机组和所述目的交换机组中,任意一个交换机组包括: $n/2$ 个汇聚交换机和 $n/2$ 个边缘交换机;

其中,所述 $n/2$ 个汇聚交换机分别和所述 $n/2$ 个边缘交换机连接;

所述 $n/2$ 个汇聚交换机分别和不同的中转交换机组中的 $n/2$ 个中转交换机连接;

所述 $n/2$ 个边缘交换机分别和不同的网络节点连接。

12. 如权利要求7所述的控制器,其特征在于,

所述数据中心网络为叶脊网络,所述 m 个源交换机组和所述目的交换机组中,任意一个

交换机组包括：一个边缘交换机；

其中，所述边缘交换机和所述多个中转交换机连接，不同的边缘交换机分别和不同的网络节点连接。

数据传输方法、相关装置及网络

技术领域

[0001] 本申请涉及数据中心(data center,DC)及流量调度技术领域,特别涉及数据传输方法、相关装置及网络。

背景技术

[0002] 数据中心是用于在互联网基础设施上传递、加速、展示、计算、存储数据信息的网络,可包括计算机系统和其它与之配套的设备(例如通信和存储系统)、数据通信连接设备、环境控制设备、监控设备以及各种安全装置。数据中心广泛应用于分布式存储、大数据分析等。

[0003] 当前,大部分数据中心的网络拓扑都实现为多级交换网络,例如胖树网络(fat tree)、叶脊网络(leaf and spine architecture)等。在数据中心网络中,承担中转数据作用的中转设备(例如胖树网络中的核心交换机、叶脊网络中的脊交换机等)利用率非常高,各个中转设备的负载,即各个中转设备中转的数据量影响到整个数据中心网络的传输时延、带宽资源利用率等。如何平衡各个中转设备的负载,以降低数据包的传输时延、提高带宽资源利用率是亟需解决的问题。

发明内容

[0004] 本申请提供了一种数据传输方法、相关装置及网络,能够实现数据中心网络中各个中转交换机的负载均衡,提高带宽资源利用率,减小传输时延。

[0005] 第一方面,本申请提供了一种数据传输方法,应用于第一控制器侧,该方法可包括:第一控制器从数据中心网络的可用中转交换机中,确定出分别中转 m 组数据的可用中转交换机;所述 m 组数据分别为 m 个源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输的数据;所述数据中心网络包括多个中转交换机、所述 m 个源交换机组,所述目的交换机组、所述源网络节点、所述目的网络节点;其中,所述可用中转交换机为所述多个中转交换机中负载不超过第一阈值的中转交换机; m 为正整数;其中,一个可用中转交换机用于中转至少一组所述数据,且任意两个可用中转交换机中转所述数据的组数的差值不超过第二阈值;所述第一控制器指示所述目的网络节点向所述源网络节点发送路由信息,所述路由信息包括用于中转所述源网络节点向所述目的网络节点传输的数据组的可用中转交换机的标识。

[0006] 具体的,本申请中的数据中心网络包括多个中转交换机、交换机组、控制器和网络节点。其中,每个交换机组都和所有的中转交换机连接,控制器控制交换机组以及和交换机组连接的网络节点。

[0007] 在数据传输过程中,数据的传输路径为:网络节点-交换机组-中转交换机-另一个交换机组-另一个网络节点。本申请中,将发送数据的网络节点称为源网络节点,将连接源网络节点的交换机组称为源交换机组,将控制源交换机组和其连接的源网络节点的控制器称为第一控制器,将接收数据的网络节点称为目的网络节点,将连接目的网络节点的交换

机组称为目的交换机组。

[0008] 可选的,本申请中的数据中心网络为胖树网络,中转交换机实现为核心交换机,交换机组实现为基本物理交换单元(point of delivery,POD),控制器控制POD以及POD连接的网络节点。

[0009] 可选的,本申请中的数据中心网络为叶脊网络,中转交换机实现为脊交换机,交换机组实现为叶交换机,控制器控制叶交换机以及叶交换机连接的网络节点。

[0010] 实施第一方面描述的方法,能够实现各个中转交换机中转的负载均衡,提高带宽资源利用率,减小传输时延。

[0011] 结合第一方面,在一些实施例中,第一控制器从数据中心网络的可用中转交换机中确定出分别中转m组数据的可用中转交换机之前,还需要确定数据中心网络的所有中转交换机中当前可用的中转交换机。

[0012] 本申请中,数据分多次进行传输,第一控制器在每一次数据传输前确定当前的可用中转交换机。

[0013] 下面说明本申请中第一控制器确定可用中转交换机的方法。本申请中,中转交换机都具有显示拥塞通知(explicit congestion notification,ECN)功能,在数据包到达中转交换机时,若当前中转交换机的负载超过第一阈值,中转交换机将数据包的ECN域的取值修改为第二值;若负载不超过第一阈值,中转交换机不对数据包的ECN域作修改,或对数据包的ECN域作修改,使得数据包的ECN域的取值为第一值。可选的,第一值和第二值可以预先定义。可选的,第一阈值可以由第一控制器根据中转交换机的处理能力预存储或预配置。

[0014] (1)在初始化阶段,即源网络节点向目的网络节点传输初始数据包时,各个初始数据包的传输路径覆盖了全部的中转交换机。第一控制器获取来自全部中转交换机的初始数据包,在初始数据包的ECN域的取值为第一值时,确定发送该数据包的中转交换机为可用中转交换机;当初始数据包的ECN域的取值为第二值时,确定发送该数据包的中转交换机为不可用中转交换机。

[0015] (2)在初始化后的数据传输阶段。初始化后,源网络节点按照第一控制器确定的中转交换机确定传输数据的路径。以第2次数据传输为例,第2次数据传输时,数据中心网络的中转交换机中,部分中转交换机被使用,其余中转交换机未被使用。

[0016] 其中,被使用到的部分中转交换机为初始化阶段中,第一控制器确定用于在第2次数据传输中发送数据的中转交换机。第一控制器可根据第2次数据传输中数据包的ECN域,确定该被使用到的部分中转交换机中的可用中转交换机。

[0017] 其中,未被使用到的其余中转交换机为初始化阶段中,第一控制器确定的不可用中转交换机。在初始化阶段,从第一控制器确定不可用中转交换机时开始,经过预设时长后,第一控制器认为这些中转交换机为可用中转交换机。

[0018] 结合第一方面,在一些实施例中,为了实现任意两个可用中转交换机中转数据的组数的差值不超过第二阈值,第一控制器在确定分别中转m组数据的可用中转交换机时,确定的任意一个可用中转交换机中转数据的组数不超过 $\lceil mk/v \rceil$;其中,k为第一控制器确定的中转一组数据的可用中转交换机的数量。

[0019] 结合第一方面,在一些实施例中,为了实现一个可用中转交换机用于中转至少一组数据,且任意两个可用中转交换机中转数据的组数的差值不超过第二阈值,第一控制器

可基于packing序列确定分别中转m组数据的可用中转交换机,下面详细说明。

[0020] 首先,介绍packing序列。在数据中心网络中,多个控制器都可根据V、K、M三个参数预先存储相同的多个packing序列组。其中, $1 \leq V \leq A$, $1 \leq K \leq A$, $1 \leq M \leq B$, A为数据中心网络的中转交换机的数量, B为交换机组的数量。

[0021] 具体的,根据确定的V、K、M构造的1个packing序列组满足以下3个条件:

[0022] 1. 1个packing序列组包括M个packing序列。

[0023] 2. M个packing序列中的任意1个packing序列都包括V个元素。在V个元素中,当 $V > K$ 时,有K个元素取第三值;当 $V \leq K$ 时,V个元素都取第三值。

[0024] 3. 在M个packing序列中的每一个packing序列中,任意一个元素最多能取 $\lceil M/V \rceil$ 次第三值。其中,第三值可以预先定义,本申请不做限制。例如,第三值可以为1。

[0025] 可知,1个packing序列组具有以下特征:在packing序列组的M个packing序列中,任意一个元素至少会取一次第三值,且各个元素取第三值的次数差不多。

[0026] 其次,介绍第一控制器基于packing序列确定分别中转m组数据的可用中转交换机的过程。

[0027] 具体的,第一控制器根据三个参数 $V=v$ 、 $K=k$ 、 $M=m$ 在预先存储的多个packing序列组中找到对应的1个packing序列组。其中,v为可用中转交换机的数量,k为中转一组数据的可用中转交换机的数量,m为源交换机组的数量。

[0028] 在该对应的1个packing序列组中,一个源交换机组对应一个packing序列,所述packing序列包括v个元素,所述v个元素分别对应所述数据网络中的v个可用中转交换机;当一个元素的取值为第三值时,所述元素对应的可用中转交换机为,中转所述源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输的数据的中转交换机。

[0029] 本申请中,第一控制器基于packing序列确定分别中转m组数据的可用中转交换机,能够实现数据中心网络中各个可用中转交换机的负载均衡,提高带宽资源利用率,减小传输时延。

[0030] 结合第一方面,在一些实施例中,路由信息承载在应答信号中。

[0031] 第二方面,本申请提供一种控制器,该控制器可包括多个功能模块,用于相应的执行第一方面所提供的方法,或者第一方面可能的实施方式中的任意一种所提供的方法。

[0032] 第三方面,本申请提供一种控制器,用于执行第一方面描述的数据传输方法。所述控制器可包括:存储器以及与所述存储器耦合的处理器,其中:所述存储器用于存储第一方面描述的数据传输方法的实现代码,所述处理器用于执行所述存储器中存储的程序代码,即执行第一方面所提供的方法,或者第一方面可能的实施方式中的任意一种所提供的方法。

[0033] 第四方面,本申请提供了一种网络,所述网络包括:控制器、中转交换机、交换机组和网络节点。其中,所述控制器可以是上述第二方面或第三方面描述的控制器。

[0034] 第五方面,提供了一种计算机可读存储介质,所述可读存储介质上存储有指令,当其在计算机上运行时,使得计算机执行上述第一方面描述的数据传输方法。

[0035] 第六方面,提供了一种包含指令的计算机程序产品,当其在计算机上运行时,使得计算机执行上述第一方面描述的数据传输方法。

[0036] 实施本申请,第一控制器从数据中心网络的可用中转交换机中,确定出分别中转m

组数据的可用中转交换机,并且,一个可用中转交换机用于中转至少一组数据,任意两个可用中转交换机中转数据的组数的差值不超过第二预置。本申请能够实现数据中心网络中各个中转交换机的负载均衡,提高带宽资源利用率,减小传输时延。

附图说明

- [0037] 图1A为现有技术中胖树网络的结构示意图;
- [0038] 图1B为现有技术中叶脊网络的结构示意图;
- [0039] 图2为本申请提供的数据中心网络的结构示意图;
- [0040] 图3A为本申请提供的胖树网络的结构示意图;
- [0041] 图3B为本申请提供的叶脊网络的结构示意图;
- [0042] 图4为本申请提供的数据传输方法中初始化阶段的流程示意图;
- [0043] 图5为本申请提供的数据传输方法中初始化后的流程示意图;
- [0044] 图6为本申请提供的控制器的结构示意图;
- [0045] 图7为本申请提供的控制器的功能框图。

具体实施方式

[0046] 本申请的实施方式部分使用的术语仅用于对本申请的具体实施例进行解释,而非旨在限定本申请。

[0047] 当前,数据中心网络通常采用层级拓扑结构。下面简单介绍两种常见基于层级拓扑结构的数据中心网络:胖树网络和叶脊网络。

[0048] (一) 胖树网络

[0049] 可参见图1A,图1A为基于clos的4元胖树网络结构的示意图。

[0050] 在一个 n 元的胖树网络结构中,包括 $n^2/4$ 个核心交换机、 $n^2/2$ 个汇聚交换机、 $n^2/2$ 个边缘交换机和 $n^3/4$ 个网络节点。

[0051] 其中,每 $n/2$ 个汇聚交换机和 $n/2$ 个边缘交换机组成一个数据中心基本物理交换单元(point of delivery,POD)。在任意一个POD中,每个边缘交换机分别连接不同的 $n/2$ 个网络节点,且各个边缘交换机都连接上层的 $n/2$ 个汇聚交换机。

[0052] 其中,每 $n/2$ 个核心交换机组成一个核心交换机组。第 t 个核心交换机组中的全部核心交换机与每个POD中的第 t 个汇聚交换机相连, $0 \leq t \leq n/2$ 。

[0053] 从图1A中可以看出,在一个 n 元的胖树网络拓扑结构中,每个POD都连接到所有的核心交换机,一个确定的POD中和某个核心交换机连接的汇聚交换机是唯一的。

[0054] (二) 叶脊网络

[0055] 参见图1B,图1B为叶脊网络拓扑结构的示意图。叶脊网络拓扑结构中包括脊交换机和叶交换机,每一个叶交换机连接到所有的脊交换机。如图1B所示,每个叶交换机连接的网络节点的数量和叶交换机的下行链路端口数相关。

[0056] 可理解的,上述图1A及图1B中提到的各个有连接关系的设备之间,都是通过端口进行连接的。

[0057] 上述可知,在基于层级拓扑结构的数据中心网络中,任意两个网络节点之间可能存在多条经由不同中转设备(如胖树网络中的核心交换机、叶脊网络中的脊交换机)的路

径。

[0058] 目前,在数据中心网络中,通常使用以下两种方法来调度数据的传输路径:分布式调度和集中式调度。

[0059] (一) 分布式调度

[0060] 基于哈希(hash)算法的等价多路径(equal-cost multi-path,ECMP)转发技术是典型的分布式调度技术中的一种。在ECMP中,可参见图1A所示的胖树网络,两个POD之间的数据传输路径有多条,即当一个POD连接的网络节点向另一个POD连接的网络节点传输数据时,存在多条等价的候选路径,这些数据根据哈希算法使用多条候选路径进行传输。

[0061] 因为哈希算法的本质是随机映射,有可能出现2个或多个大象流(持续时间长,带宽要求高的数据流)同时选择同一条路径的情况,导致数据流发生碰撞,带宽资源分配不均、利用率低。

[0062] (二) 集中式调度

[0063] 在集中式调度方法中,通过集中控制器收集数据中心网络中的全局信息,包括全部的数据包信息、目的网络节点的数据请求信息、各个交换机的状态等,对各个数据包的路由做出最优的分配,从而实现各个中转设备的负载均衡。

[0064] 集中式调度的方法虽然可以得到各个数据包最优的路径选择方案,但这种调度方法需要大量的信息交互,集中控制器收集信息的过程会占用带宽资源,给数据中心网络造成额外的开销;并且,这种调度方法计算复杂度高,反应时间长。

[0065] 上述可以看出,在数据中心网络中,实现负载均衡的关键是平衡承担中转数据作用的中转设备(例如胖树网络中的核心交换机、叶脊网络中的脊交换机等)的工作量。

[0066] 本申请提出了一种数据中心网络,可通过控制器来平衡经过各个中转设备的工作量,以实现负载均衡。

[0067] 参见图2,本申请的数据中心网络包括:多个中转交换机、多个交换机组、多个控制器、和多个交换机组分别连接的网络节点。其中,任意一个交换机组都和全部中转交换机连接,多个交换机组分别连接不同的网络节点,一个交换机组连接一个控制器。

[0068] 可理解的,上述图2中提到的各个有连接关系的设备之间,都是通过端口进行连接的。

[0069] 可理解的,不限于上述提及的设备,具体实现中,本申请的数据中心网络还可包括其他的设备,本申请不做限制。

[0070] 下面分别介绍数据中心网络中的各个网元。

[0071] (一) 中转交换机

[0072] 中转交换机是指数据中心网络中最上层的、用于中转数据的交换机,可以是图1A所示胖树网络中的核心交换机,也可以是图1B所示的叶脊网络中的脊交换机,还可以是其他基于层级拓扑结构的数据中心网络中用于中转数据的交换机。

[0073] 本申请中,中转交换机具有显示拥塞通知(explicit congestion notification, ECN)功能。下面简单说明中转交换机实现ECN功能的流程。

[0074] 1. 数据包到达中转交换机时,中转交换机查看数据包的目的地址。

[0075] 2. 中转交换机根据数据包的目的地址,确定用于转发该数据包的出端口。

[0076] 可理解的,中转交换机有多个出端口。在确定数据包的目的地址后,中转交换机能

唯一地确定用于转发该数据包的出端口。

[0077] 3. 中转交换机查看该出端口处的负载。

[0078] 其中, 出端口处的负载可由该出端口处等待被转发的数据包的个数、该出端口转发数据包的速度等因素确定。等待被转发的数据包的个数越多, 或者, 转发速度越慢, 表示当前该出端口的负载越大。

[0079] 4. 如果当前该出端口的负载超过了第一阈值, 中转交换机将新到达的该数据包中的ECN域设置为第二值, 表示当前该出端口需要转发的数据包过多, 正在经历拥塞; 如果当前该出端口的负载不超过第一阈值, 则中转交换机不会对该数据包中的ECN域作任何改变, 数据包中的ECN域可以取第一值, 表示当前该出端口还可以转发新的数据包。

[0080] 其中, 第一阈值可预先设置。可选的, 第一阈值可根据中转交换机的处理能力进行设置。第一值和第二值可预先定义。

[0081] 举例说明, 如图2所示, 假设网络节点0向网络节点3发送数据包, 该数据包中的目的地址为网络节点3的地址, 该数据包的传输路径经过中转交换机0。假设中转交换机0和交换机组2通过出端口0连接。中转交换机0查看出端口0处待转发的数据包个数, 若大于5时, 将新到达的数据包的ECN域设置为11; 若小于或等于5时, 中转交换机不对新到达的数据包的ECN域做任何改变, 该数据包的ECN域可能为01或10。

[0082] 由于本申请中的中转交换机具有ECN功能, 接收到来自该中转交换机的数据包的设备可以通过数据包中的ECN域获知该中转交换机的状态: 当前是否正在经历拥塞。

[0083] (二) 交换机组

[0084] 本申请中, 交换机组是连接网络节点和中转交换机的一组设备, 可以是图1A所示胖树网络中的POD, 也可以是图1B所示的叶脊网络中的叶交换机, 还可以是其他的分层传输网络中的用于连接网络节点和中转交换机的一组设备。

[0085] (三) 控制器

[0086] 本申请中, 每个交换机组都配置有控制器, 控制器连接并控制交换机组和交换机组对应的网络节点。

[0087] 本申请中, 控制器可用于获取来自各个中转交换机的数据包、确定中转交换机的状态、进行路由调度等。

[0088] (四) 网络节点

[0089] 本申请中, 网络节点可以是工作站、服务器、终端设备等拥有唯一网络地址的设备。在数据中心网络中, 连接不同的交换机组的任意两个网络节点之间进行通信时, 需经过各自连接的交换机组和中转交换机。

[0090] 举例说明, 如图2所示, 网络节点0、网络节点2分别连接交换机组0、交换机组1, 网络节点0向网络节点2发送数据时, 该数据的传输路径为: 网络节点0-交换机组0-数据中心网络中的任意一个中转交换机-交换机组1-网络节点2。

[0091] (五) 源网络节点和目的网络节点

[0092] 本申请中, 发送数据的网络节点称为源网络节点, 接收数据的网络节点称为目的网络节点。当多个网络节点向同一个网络节点发送数据时, 有多个源网络节点, 一个目的网络节点。

[0093] 相应地, 和源网络节点连接的交换机组称为源交换机组, 和目的网络节点连接的

交换机组称为目的交换机组。本申请中,将目的交换机组的对应的控制器称为第一控制器,可理解的,第一控制器连接并控制目的交换机组和目的网络节点。

[0094] 可理解的,当有多个网络节点向同一个网络节点发送数据时,对应有多个源交换机组和一个目的交换机组。例如,参见图2,当网络节点0、1、3、4都向网络节点2发送数据时,源交换机组包括:交换机组0、交换机组2、交换机组3,目的交换机组包括交换机组1。

[0095] 可理解的,一个源网络节点向目的网络节点发送数据时,该数据的传输路径为:源网络节点-该源网络节点连接的源交换机组-中转交换机-目的网络节点连接的目的交换机组-目的网络节点。

[0096] 下面结合胖树网络及叶脊网络具体描述本申请提供的数据中心网络。

[0097] 参见图3A,图3A为本申请提供的胖树网络的结构示意图,图3A以6元胖树网络为例。和图1A相比,图3A所示的胖树网络中,每个POD都配置有POD控制器。本申请中,POD控制器可用于获取来自各个核心交换机的数据包、读取各个数据包的ECN域的取值并确定核心交换机的状态、进行路由调度等。

[0098] 参见图3B,图3B为本申请提供的叶脊网络的结构示意图。和图1B相比,图3B所示的叶脊网络中,每个叶交换机都配置有叶控制器。本申请中,叶控制器可用于获取来自各个脊交换机的数据包、读取各个数据包的ECN域的取值并确定脊交换机的状态、进行路由调度等。

[0099] 基于上述图2、图3A及图3B中的数据中心网络,为了实现各个中转交换机的负载均衡,降低数据的传输时延、提高带宽资源的利用率,本申请提供了一种数据传输方法。

[0100] 本申请的主要发明原理可包括:第一控制器获知数据中心网络中各个中转交换机的状态,在可用中转交换机中,确定出分别用于中转来自不同源网络节点的数据的可用中转交换机。其中,各个可用中转交换机中转的数据量是均衡的。

[0101] 下面对本申请涉及的几个概念进行描述:

[0102] (一) 可用中转交换机、不可用中转交换机

[0103] 本申请中,对于目的网络节点来说,数据中心网络中的各个中转交换机可分为两种:可用中转交换机和不可用中转交换机。

[0104] 可理解的,在数据中心网络的多个中转交换机中,每个中转交换机分别有一个出端口用于转发数据到该目的网络节点。

[0105] 可用中转交换机是指:该出端口处的负载不超过第一阈值的中转交换机。

[0106] 不可用中转交换机是指:该出端口处的负载超过第一阈值的中转交换机。

[0107] (二) 一组数据

[0108] 本申请中,将传输时经过同一个源交换机组的数据称为一组数据,即,将从同一个源交换机组连接的源网络节点发送给目的交换机组连接的目的网络节点的数据称为一组数据。可理解的,该一组数据可以是源交换机组连接的一个或多个源网络节点生成的。

[0109] 以图3A所示的胖树网络为例说明,假设有5个源网络节点:网络节点0、网络节点1、网络节点2、网络节点4、网络节点5、网络节点6,对应有5个源POD:POD0、POD1、POD3、POD4、POD5;有1个目的网络节点3,对应1个目的POD:POD2。则,在图3A所示的胖树网络中,共有5组数据:分别是网络节点0、网络节点1、网络节点2、网络节点4、网络节点5、网络节点6发送给网络节点3的数据。

[0110] 可理解的,一组数据的传输过程中,可通过数据包的形式传输,并且可分多次传输,本申请不做任何限制。

[0111] (三) packing序列组

[0112] 本申请中,1个packing序列组根据3个参数V、K、M确定。在V、K、M的取值确定后,根据以下3个条件构造对应的packing序列组:

[0113] 1.1个packing序列组包括M个packing序列。

[0114] 2.M个packing序列中的任意1个packing序列都包括V个元素。在V个元素中,当 $V > K$ 时,有K个元素取第三值;当 $V \leq K$ 时,V个元素都取第三值。

[0115] 3.在M个packing序列中的每一个packing序列中,任意一个元素最多能取 $[MK/V]$ 次第三值。

[0116] 其中,第三值可以预先定义,本申请不做限制。例如,第三值可以为1。

[0117] 从上述3个条件可知,1个packing序列组具有以下特征:在packing序列组的M个packing序列中,任意一个元素至少会取一次第三值,且各个元素取第三值的次数差不多。

[0118] 参见表1,表示出了1个packing序列组的可能的实现形式。在该packing序列组中, $V=6, K=3, M=5$,第三值为1。

[0119]	Packing序列1	110100
	Packing序列2	011010
	Packing序列3	001101
	Packing序列4	100110
	Packing序列5	010011

[0120] 表1 packing序列组

[0121] 如表1所示,该packing序列组中包括5个packing序列,每个packing序列都包括6个元素。

[0122] 在该packing序列组中,第1个元素取了2次“1”,第2个元素取了3次“1”,第3个元素取了2次“1”,第4个元素取了3次“1”,第5个元素取了3次“1”,第5个元素取了2次“1”。可见,任意一个元素取“1”的次数都是均衡的。

[0123] 本申请中,V、K、M的取值确定后,能够构造出对应的packing序列组。下面简单介绍根据V、K、M构造对应的packing序列组的方法。

[0124] 1.根据公式1,确定 λ_1 , λ_1 为满足公式1时 λ 的最小值。

[0125]
$$\lfloor \frac{V}{K} \lfloor \frac{\lambda_1 (V-1)}{K-1} \rfloor \rfloor \geq M \quad \text{公式 1}$$

[0126] 2.构造 (V, K, λ_1) -packing。

[0127] (V, K, λ_1) -packing由多个区组组成,每个区组为一个集合,每个区组包括的元素从V个给定的元素中选取。

[0128] (V, K, λ_1) -packing满足条件:每个区组包括K个元素,任意两个不同的元素最多同时出现在 λ_1 个区组中。

[0129] 这里,经过第(1)个步骤,构造出的 (V, K, λ_1) -packing中至少包括M个区组。

[0130] 3.从 (V, K, λ_1) -packing中的至少M个区组中,选择M个区组。其中,V个给定的元素中任意一个元素最多在 $[MK/V]$ 个区组中出现。

[0131] 4. 当 $V > K$ 时, 根据选择的 M 个区组生成对应的packing序列组。具体的, 第 i 个区组对应packing序列组中的第 i 个packing序列。其中, 具体的生成规则可参照后续相关描述

[0132] 当 $V \leq K$ 时, 直接生成packing序列组, 其中的 M 个packing序列中, 每个元素都取第三值。举例说明, 假设 $V=6, K=3, M=5$ 。给定的 V 个元素为: 1, 2, 3, 4, 5, 6。

[0133] 首先, 根据公式1确定 $\lambda_1=2$ 。

[0134] 其次, 构造 $(2, 3, 5)$ -packing, 其中包括10个区组。 $(2, 3, 5)$ -packing可以为: $\{\{1, 2, 4\}, \{2, 3, 5\}, \{3, 4, 6\}, \{1, 4, 5\}, \{2, 5, 6\}, \{1, 3, 6\}, \{2, 3, 4\}, \{4, 5, 6\}, \{1, 2, 6\}, \{1, 3, 5\}\}$ 。

[0135] 然后, 在该10个区组中选择5个区组, 使得每个元素最多在 $[5 \times 3 + 6]$ 个区组中出现。该5个区组可以为: $\{\{1, 2, 4\}, \{2, 3, 5\}, \{3, 4, 6\}, \{1, 4, 5\}, \{2, 5, 6\}\}$ 。

[0136] 最后, $V=6, K=3, V > K$, 根据选择的5个区组, 生成对应的packing序列组。参见表1, 该5个区组对应的packing序列组可以如表1所示。举例说明, 第1个区组为 $\{1, 2, 4\}$, 元素1、2、4出现, 对应第1个packing序列中的第1、2、4个元素取第三值(1), 即, 第1个packing序列为: 110100。

[0137] 显然地, 根据上述方法构造出的packing序列组中, 每一个元素取第三值的次数是均衡的。

[0138] (四) 控制器存储多个packing序列组

[0139] 本申请中, 在一个确定的数据中心网络中, 中转交换机的数量、交换机组的数量、控制器的数量是确定的。数据中心网络中的多个控制器, 都根据中转交换机的数量、交换机组的数量预先存储多个packing序列组。可理解的, 数据中心网络中多个控制器中存储的packing序列是相同的, 适用于任意一个控制器为目的控制器的场景。

[0140] 下面以数据中心网络中包括 A 个中转交换机, B 个交换机组为例, 说明控制器中存储的多个packing序列组。

[0141] 控制器根据3个参数 V, K, M 构造packing序列组, 其中, $1 \leq V \leq A, 1 \leq K \leq A, 1 \leq M \leq B$ 。因此, 控制器中可构造并存储 $A \times A \times B$ 个packing序列组。

[0142] 举例说明, 参见图3A所示的胖树网络, 其中包括9个核心交换机(中转交换机)、6个POD(交换机组)、6个控制器。图3A所示的6个控制器可根据 $1 \leq V \leq 9, 1 \leq K \leq 9, 1 \leq M \leq 6$ 来构造并存储 $9 \times 9 \times 6 = 486$ 个packing序列组。

[0143] 例如, 控制器中可存储当 $1 \leq V \leq 9, K=3, M=5$ 时的6个packing序列组, 该9个packing序列组可表2所示。其中, 1列为1个packing序列组, 例如, 当 $V=6$ 时对应的5个packing序列为1个packing序列组。

[0144]

	1	2	3	4	5	6	7	8	9
1	1	11	111	1110	11010	110100	1010100	10010010	100100100
2	1	11	111	0111	01101	011010	1101000	01001001	010010010
3	1	11	111	1011	10110	001101	0011010	10100100	001001001
4	1	11	111	1101	01011	100110	0001101	01010100	100010001
5	1	11	111	1110	10101	010011	0100011	00101010	010001100

[0145] 表2 $1 \leq V \leq 9, K=3, M=5$ 时的9个packing序列组

[0146] 和上述表2类似, 本申请中控制器还可构造其余的 $1 \leq V \leq 9, 1 \leq K \leq 9, 1 \leq M \leq 6$ 时的

多个packing序列组并进行存储。

[0147] 下面介绍本申请的数据传输方法。本申请中,数据在数据中心网络中传输时,根据数据量以数据包的形式分多次进行传输。来自同一个源交换机组连接的源网络节点的数据在每一次传输中,可能经过不同的可用中转交换机。

[0148] 本申请中,数据传输可分为两个阶段:初始化阶段(第1次数据包传输阶段),以及,初始化后的数据传输阶段。

[0149] 参见图4,图4所示实施例以数据中心网络包括多个中转交换机、 m 个源交换机组、1个目的交换机组、多个源网络节点、1个目的网络节点,且一共传输 m 组数据为例进行介绍。

[0150] 图4为本申请提供的数据传输方法中初始化阶段的流程示意图。其中,初始化阶段能让第一控制器获知数据中心网络中全部中转交换机的状态,确定其中的可用中转交换机。

[0151] 该方法可包括如下步骤:

[0152] S101、 m 个源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输初始数据包,并且,初始数据包的传输路径经过数据中心网络的全部中转交换机。

[0153] 在可选实施例中,目的网络节点在调用数据中心网络中多个源网络节点的数据时,首先向存储数据的多个源网络节点发送数据传输请求。

[0154] 具体的,多个源网络节点在接收到目的网络节点发送的数据传输请求后,首先传输初始数据包。其中,初始数据包可以为正常的一次数据传输中,多个源网络节点向目的网络节点发送的全部数据包;也可以是多个源网络节点向目的网络节点传输的全部数据中的很少一部分数据包,初始数据包较少时可以加快初始化的效率。

[0155] 可理解的,步骤S101中初始数据包的传输路径为:源网络节点-源交换机组-中转交换机-目的交换机组-目的网络节点。

[0156] 在本申请中,初始数据包的传输路径经过数据中心网络的全部中转交换机。为了使得初始数据包的传输路径经过数据中心网络的全部中转交换机,可以有以下两种策略:

[0157] 第一种策略,每一个源交换机组在接收到其连接的源网络节点的初始数据包时,都将接收到的初始数据包转发至所有的中转交换机。通过第一种传输策略,每一个源交换机组在转发初始数据包时都简单地遍历全部的中转交换机,简单易行,且能够保证初始数据包的传输路径经过全部的中转交换机。

[0158] 第二种策略,多个源交换机组在接收到其连接的源网络节点的初始数据包时,各自将初始数据包转发至不同的部分中转交换机。其中,各个源交换机组转发初始数据包的中转交换机可以预先规定,也可以由第一控制器规定并且携带在目的网络节点向多个源网络节点发送的数据传输请求中。

[0159] 举例说明,参见图2,假设共有4个中转交换机,3个源交换机组:交换机组0、1、2,1个目的交换机组3。在第二种策略中,源交换机组0可将接收到的初始数据包分别转发至中转交换机0、1,源交换机组1可将接收到的初始数据包分别转发至中转交换机1、2,源交换机组2可将接收到的初始数据包分别转发至中转交换机2、3。

[0160] 通过第二种策略,每一个源交换机组仅需转发初始数据包到部分的中转交换机,就能使得初始数据包的传输路径覆盖了全部的中转交换机,节约了带宽资源。

[0161] 可理解的,本申请的中转交换机具有ECN功能,中转交换机在接收到数据包时,会

根据当前转发数据至目的网络节点的出端口处的负载对数据包的ECN域进行改写,具体可参考图4所示数据中心网络中关于中转交换机的详细描述。

[0162] S102、第一控制器根据初始数据包确定数据中心网络的可用中转交换机。

[0163] 具体的,第一控制器连接并控制目的交换机组和目的网络节点,在初始数据包到达目的交换机组或者目的网络节点时,第一控制器能够获取初始数据包。

[0164] 由于步骤S101中,初始数据包遍历了全部的中转交换机,因此,第一控制器能够获取到每一个中转交换机转发的初始数据包,并根据初始数据包中的ECN域获知数据中心网络中全部中转交换机的状态,确定可用中转交换机和不可用中转交换机。其中,根据初始数据包的ECN域确定可用中转交换机的方法可参照图4所示数据中心网络中关于中转交换机的详细描述,在此不赘述。

[0165] 在可选实施例中,第一控制器确定当前的不可用中转交换机后,设定不可用中转交换机在预设时长后的状态为可用,这样设定的作用可参照后续图5实施例的详细描述,在此不赘述。

[0166] S103、第一控制器从数据中心网络的可用中转交换机中,确定出分别中转m组数据的可用中转交换机。其中,一个可用中转交换机用于中转至少一组所述数据,且任意两个可用中转交换机中转所述数据的组数的差值不超过第二阈值。

[0167] 具体的,第一控制器在确定数据中心网络的可用中转交换机后,可以确定在初始化后的第2次数据传输中,分别中转m组数据的可用中转交换机。并且,使得一个可用中转交换机中转至少一组数据,且任意两个可用中转交换机中转数据的组数的差值不超过第二阈值。其中,第二阈值可以预先定义,例如第二阈值可以为1或2。

[0168] 这样,在第2次的数据传输中,数据中心网络中的每一个可用中转交换机都被使用到,且,每一个可用中转交换机承担的数据量的差值在一定范围内,保证了可用中转交换机的负载均衡。

[0169] 在可选实施例中,为了保证任意两个可用中转交换机中转数据的组数的差值不超过第二阈值,第一控制器在确定分别中转m组数据的可用中转交换机时,可以使得任意一个可用中转交换机中转所述数据的组数不超过 $\lceil mk/v \rceil$ 。其中,k为所述第一控制器确定的中转一组所述数据的可用中转交换机的数量。k的具体取值由第一控制器根据当前触发该m组数据从多个源网络节点向目的网络节点传输的应用所要求的传输速率决定,例如,当该应用要求的传输速率越高时,k的取值越大。

[0170] 在另一可选实施例中,为了保证一个可用中转交换机用于中转至少一组所述数据,且任意两个可用中转交换机中转所述数据的组数的差值不超过第二阈值,第一控制器可从预先存储的多个packing序列组中选取对应的packing序列组,并根据packing序列组中的packing序列确定分别中转m组数据的可用中转交换机,下面详细说明:

[0171] 1.首先,第一控制器根据可用中转交换机的数量、中转一组所述数据的可用中转交换机的数量、源交换机组的数量确定V,K,M的取值,根据V,K,M从预先存储的多个packing序列组中选取对应的1个packing序列组。

[0172] 具体的,V的取值为可用中转交换机的数量v,K的取值为中转一组所述数据的可用中转交换机的数量k,M的取值为源交换机组的数量m。

[0173] 其中,第一控制器中预先存储多个packing序列组可参照前文本申请涉及的概念

(四) 中控制器存储多个packing序列组的相关描述。

[0174] 2. 根据选取的1个packing序列组, 确定分别中转m组数据的可用中转交换机。

[0175] 具体的, 在选取的1个packing序列组中, 包括m个packing序列。其中, 一个源交换机组对应一个packing序列。任意一个packing序列都包括v个元素, 所述v个元素分别对应所述数据网络中的v个可用中转交换机; 当一个元素的取值为第三值时, 所述元素对应的可用中转交换机为, 中转所述源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输的数据的中转交换机。

[0176] 举例说明, 假设如图2所示的数据中心网络中, 包括6个源网络节点: 网络节点0、1、2、4、5、6, 目的网络节点为网络节点3, 即数据中心网络包括5个源交换机组: 交换机组0、交换机组1、交换机组3、交换机组4、交换机组5, 目的交换机组为交换机组2。第一控制器为控制器2。假设第一控制器确定的可用中转交换机有6个: 中转交换机0、1、2、6、7、8。

[0177] 假设中转一组所述数据的可用中转交换机的数量为3, 第三值为1。即 $v=6, k=3, m=5$ 。第一控制器可根据 $V=6, K=3, M=5$, 在预先存储的多个packing序列组中找到对应的1个packing序列组。

[0178] 可选的, 该1个packing序列组可以如表1所示, 其中, packing序列1-5分别对应5个源交换机组。第1个源交换机组对应于packing序列1 (110100), 表示网络节点0、1向网络节点3发送的数据, 由第1、2、4个可用中转交换机 (中转交换机0、1、6) 中转。其他的packing序列的意义可以此类推, 在此不赘述。

[0179] S104、第一控制器指示目的网络节点向源网络节点发送路由信息, 路由信息包括用于中转源网络节点向目的网络节点传输的数据组的可用中转交换机的标识。

[0180] 具体的, 第一控制器在确定分别中转m组数据的可用中转交换机后, 指示目的网络节点将中转每一组数据的可用中转交换机的标识分别发送给对应的源网络节点。例如, 以上述步骤S103中的例子说明, 第一控制器可以指示目的网络节点将中转交换机0、1、6的标识发送给网络节点0、1。其中, 可用中转交换机的标识可以是MAC地址、交换机虚拟接口 (switch virtual interface, SVI) 地址等可唯一确定数据中心网络中的中转交换机的标识

[0181] 在可选实施例中, 路由信息可以承载在目的网络节点接收到初始化数据包后发送给各个源网络节点的应答信号中 (acknowledgement, ACK)。

[0182] 经过图4所示方法实施例, 完成了第1次的数据包传输, 即初始化阶段。参见图5, 图5为初始化后的数据传输阶段的流程示意图。图5以源网络节点接收到目的网络节点发送的路由信息, 并根据路由信息进行第2次数据包传输为例进行说明, 可包括如下步骤:

[0183] S201、源网络节点根据初始化阶段中目的网络节点发送的路由信息, 进行第2次数据包传输。

[0184] 具体的, 数据中心的多个源网络节点都接收到目的网络节点发送的路由信息, 并根据路由信息中的可用中转交换机的标识唯一地确定本次数据包传输的路径。

[0185] 以图4实施例步骤S103中的例子为例, 网络节点0、1接收到网络节点3发送的路由信息, 其中包括中转交换机0、1、6的标识。网络节点0可确定本次数据包传输的3条路径, 每条路径都包括上下行路径, 这3条路径分别为:

[0186] 路径1, 上行路径: 网络节点0-交换机组0-中转交换机0; 下行路径: 中转交换机0-

交换机组2-网络节点3;

[0187] 路径2,上行路径:网络节点0-交换机组0-中转交换机1;下行路径:中转交换机1-交换机组2-网络节点3;

[0188] 路径3,上行路径:网络节点0-交换机组0-中转交换机6;下行路径:中转交换机6-交换机组2-网络节点3。

[0189] 类似的,第2次数据包传输中,其他源网络节点的路径也可以此类推。

[0190] 下面分别结合胖树网络和叶脊网络说明第2次数据传输的具体路径。

[0191] 参见图3A,当数据中心网络为胖树网络时,网络节点0可确定本次数据包传输的3条路径,每条路径都包括上下行路径,这3条路径分别为:

[0192] 路径1,上行路径:网络节点0-边缘交换机0-汇聚交换机0-核心交换机0;下行路径:核心交换机0-汇聚交换机6-边缘交换机7-网络节点3;

[0193] 路径2,上行路径:网络节点0-边缘交换机0-汇聚交换机0-核心交换机1;下行路径:核心交换机1-汇聚交换机6-边缘交换机7-网络节点3;

[0194] 路径3,上行路径:网络节点0-边缘交换机0-汇聚交换机2-核心交换机6;下行路径:核心交换机6-汇聚交换机8-边缘交换机7-网络节点3。

[0195] 参见图3B,当数据中心网络为叶脊网络时,网络节点0可确定本次数据包传输的3条路径,每条路径都包括上下行路径,这3条路径分别为:

[0196] 路径1,上行路径:网络节点0-叶交换机0-脊交换机0;下行路径:脊交换机0-叶交换机2-网络节点3;

[0197] 路径2,上行路径:网络节点0-叶交换机0-脊交换机1;下行路径:脊交换机1-叶交换机2-网络节点3;

[0198] 路径3,上行路径:网络节点0-叶交换机0-脊交换机6;下行路径:脊交换机6-叶交换机2-网络节点3。

[0199] S202、第一控制器确定数据中心网络的可用中转交换机。

[0200] 具体的,在第2次数据包传输中,多个源网络节点经由初始化阶段中第一控制器选择的可用中转交换机传输数据。可理解的,第一控制器能够获取到第2次传输的数据包,并根据第2次传输的数据包确定初始化阶段中第一控制器选择的可用中转交换机的状态。

[0201] 以图4实施例步骤S103中的例子说明,参见图2,假设初始化阶段中第一控制器确定的可用中转交换机有6个:中转交换机0、1、2、6、7、8。那么,在步骤S202中,第一控制器能够获知这6个中转交换机的状态,重新确认在进行第2次数据包传输时,这6个中转交换机中的可用中转交换机。

[0202] 显然地,在第2次数据包传输中,第一控制器无法通过数据包来获知第2次数据包传输中,没有中转数据的中转交换机(中转交换机3、4、5)的状态。

[0203] 本申请中,通过设定不可用中转交换机的有效时长,避免以下情况:部分中转交换机被确认为不可用之后,即使在后续的数据传输中实际状态转变为可用,由于第一控制器无法获知其状态转变而无法利用其中转数据。通过这种方法能够避免带宽资源的浪费。其中,预设时长可以根据中转交换机转发数据的速度确定。可选的,中转交换机转发数据的速度越快,预设时长越短。

[0204] 以上述例子为例说明,在初始化阶段,第一控制器获知中转交换机3、4、5为不可用

中转交换机,因此,在初始化阶段后的第2次数据传输中并没有使用到中转交换机3、4、5。在进行第2次数据传输时,第一控制器无法通过数据包获知中转交换机3、4、5的状态。

[0205] 中转交换机3、4、5在初始化阶段时为不可用状态,在第2次数据传输时可能转变为可用。因此,为了避免第一控制器一直认为中转交换机3、4、5的状态为不可用,本申请中,在上一次数据传输中(即初始化阶段中)第一控制器获取到中转交换机3、4、5的状态为不可用时,第一控制器设定一个预设时长,在预设时长后,第一控制器认为中转交换机3、4、5的状态为可用。

[0206] 若第一控制器从初始化阶段中获取到中转交换机3、4、5的状态为不可用,到步骤S202中第一控制器确定数据中心网络的可用中转交换机时,已经超过了预设时长,那么第一控制器认为此时中转交换机3、4、5的状态为可用。

[0207] 通过步骤S202可知,在第2次数据传输中,第一控制器确定的可用中转交换机中,包括通过数据包的ECN域确定的可用中转交换机,还包括通过预设时长确定的可用中转交换机。

[0208] S203、第一控制器从数据中心网络的可用中转交换机中,确定出分别中转m组数据的可用中转交换机。

[0209] S204、第一控制器指示目的网络节点向源网络节点发送路由信息,路由信息包括用于中转源网络节点向目的网络节点传输的数据组的可用中转交换机的标识。

[0210] 其中,步骤S204中的路由信息用于各个源网络节点确定第3次数据传输的路径。

[0211] 可理解的,步骤S203-S204的实现和图4实施例步骤S103-S104的实现类似,可参照相关描述,在此不赘述。

[0212] 可理解的,在第2次数据传输之后,还会经过多次数据传输,直至m组数据传输完成。之后的多次数据传输的步骤都和图5所示的第2次数据传输的步骤相同,可参照相关描述,在此不赘述。

[0213] 通过图4及图5所示方法实施例,第一控制器在每一次数据传输时可以确定可用中转交换机,并在预先存储的多个packing序列组中查找到对应的packing序列组,根据packing序列组确定分别中转m组数据的可用中转交换机。

[0214] 这里,由于第一控制器在确定分别中转m组数据的可用中转交换机时,排除了不可用中转交换机,负载大的中转交换机不会被使用,避免进一步加大这些中转交换机的负载,相当于进行了第一次负载均衡。

[0215] 第一控制器在确定分别中转m组数据的可用中转交换机时,每个可用中转交换机都被使用到,且每个可用中转交换机承担的数据量差别不大,相当于在所有的可用中转交换机中实现了第二次负载均衡。

[0216] 综上,本申请的数据传输方法,计算复杂度低,效率高,能够避免部分中转交换机负载过大,实现数据中心网络中各个中转交换机的负载均衡,提高带宽资源利用率,减小传输时延。

[0217] 上述详细描述了本申请的数据传输方法,为了便于更好地实施本申请的上述方法,相应地,下面提供了本申请的相关装置。

[0218] 参见图6,图6为本申请提供的控制器10的结构示意图。控制器10可以实现为图2数据中心网络中的控制器,也可以是图3A所示胖树网络中的POD控制器、图3B所示叶脊网络中的

的叶控制器,还可以是上述方法实施例中的控制器。如图6所示,控制器10可包括:通信接口103、一个或多个控制器处理器101、耦合器111和存储器105。这些部件可通过总线或者其它方式连接,图6以通过总线连接为例。其中:

[0219] 通信接口103可用于控制器10与其他设备,例如图2中的交换机组和网络节点,图3A中的汇聚交换机、边缘交换机、网络节点,图3B中的叶交换机、网络节点等进行通信。具体实现中,通信接口103可以有线的通信接口(例如以太网接口)。

[0220] 存储器105与控制器处理器101耦合,用于存储各种软件程序和/或多组指令。具体实现中,存储器105可包括高速随机存取的存储器,并且也可包括非易失性存储器,例如一个或多个磁盘存储设备、闪存设备或其他非易失性固态存储设备。存储器105可以存储操作系统,例如uCOS、VxWorks、RTLinux等嵌入式操作系统。本申请中,存储器105还可以预先存储多个packing序列组,可参照上述本申请涉及的概念(四)中控制器存储多个packing序列组的相关描述。

[0221] 在本申请的一些实施例中,存储器105可用于存储本申请的一个或多个实施例提供的数据传输方法在控制器10侧的实现程序。关于本申请的一个或多个实施例提供的数据传输方法的实现,请参考图4-图5所示方法实施例。

[0222] 控制器处理器101可以是通用处理器,例如中央处理器(central processing unit,CPU),处理器101还可包括硬件芯片,上述硬件芯片可以是以下一种或多种的组合:专用集成电路(application specific integrated circuit,ASIC)、现场可编程逻辑门阵列(field programmable gate array,FPGA)、复杂可编程逻辑器件(complex programmable logic device,CPLD)。处理器101可处理接收到的数据,本申请中,处理器601还可根据接收到的数据确定数据中心网络中的可用中转交换机,以及,确定中转各个源网络节点发送的数据的中转交换机。

[0223] 本发明实施例中,控制器处理器101可用于读取和执行计算机可读指令。具体的,控制器处理器101可用于调用存储于存储器105中的程序,例如本申请的一个或多个实施例提供的数据传输方法在控制器10侧的实现程序,并执行该程序包含的指令。

[0224] 需要说明的,图6所示的控制器10仅仅是本发明实施例的一种实现方式,实际应用中,控制器10还可以包括更多或更少的部件,这里不作限制。

[0225] 参见图7,图7为本申请提供的控制器20的功能框图。

[0226] 如图7所示,控制器20可包括:确定单元201和指示单元202。其中,

[0227] 确定单元201,用于从数据中心网络的可用中转交换机中,确定出分别中转 m 组数据的可用中转交换机;所述 m 组数据分别为 m 个源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输的数据;所述数据中心网络包括多个中转交换机、所述 m 个源交换机组,所述目的交换机组、所述源网络节点、所述目的网络节点;其中,所述可用中转交换机为所述多个中转交换机中负载不超过第一阈值的中转交换机; m 为正整数;

[0228] 其中,一个可用中转交换机用于中转至少一组所述数据,且任意两个可用中转交换机中转所述数据的组数的差值不超过第二阈值;

[0229] 指示单元202,用于指示所述目的网络节点向所述源网络节点发送路由信息,所述路由信息包括用于中转所述源网络节点向所述目的网络节点传输的数据组的可用中转交换机的标识。

[0230] 在可选实施例中,控制器20还包括获取单元203,用于获取至少一个数据包;确定单元201还用于:在所述数据包中的拥塞显示指示域的取值为第一值的情况下,确定发送所述数据包的中转交换机为所述可用中转交换机;在所述数据包中的拥塞显示指示域的取值为第二值的情况下,确定发送所述数据包的中转交换机在预设时长后为所述可用中转交换机。

[0231] 在可选实施例中,所述至少一个数据包来自所述多个中转交换机,或者,所述至少一个数据包来自上一次数据传输中的可用中转交换机。

[0232] 在可选实施例中,为保证所述任意两个可用中转交换机中转所述数据的组数的差值不超过第二阈值,第一控制器确定的任意一个可用中转交换机中转所述数据的组数不超过 $\lceil mk/v \rceil$;其中,k为中转一组所述数据的可用中转交换机的数量。

[0233] 在可选实施例中,确定单元201根据packing序列确定分别中转m组数据的中转交换机,具体的:确定单元201在预先存储的多个packing序列组中,确定所述m个源交换机组分别对应的packing序列;其中,一个源交换机组对应一个packing序列,所述packing序列包括v个元素,所述v个元素分别对应所述数据网络中的v个可用中转交换机;当一个元素的取值为第三值时,所述元素对应的可用中转交换机为,中转所述源交换机组连接的源网络节点向目的交换机组连接的目的网络节点传输的数据的中转交换机;

[0234] 且,当 $v > k$ 时,所述v个元素中有k个元素取第三值,当 $v \leq k$ 时,所述v个元素取第三值;其中,一个packing序列组包括m个packing序列,在所述多个packing序列组中的每一个packing序列组中,任意一个元素最少取一次第三值,最多取 $\lceil mk/v \rceil$ 次第三值;

[0235] 其中,v为所述可用中转交换机的数量,k为中转一组所述数据的可用中转交换机的数量。

[0236] 在可选实施例中,所述路由信息承载在应答信号中。

[0237] 在可选实施例中,所述数据中心网络为胖树网络,或者,所述数据中心网络为叶脊网络。

[0238] 可理解的,关于控制器20包括的各个功能单元的具体实现,可参考前述图4-图5以及相关描述,这里不再赘述。

[0239] 另外,本申请还提供了一种数据中心网络,所述数据中心网络可以是图2、图3A或图3B所示的网络,可包括:中转交换机、交换机组、网络节点和控制器。其中,所述控制器可以是图4-图5分别对应的方法实施例中的第一控制器。

[0240] 具体实现中,所述控制器可以是图2所示所述中心网络中的控制器,可以是图3A所示胖树网络中的POD控制器,还可以是图3B所示叶脊网络中的叶控制器。

[0241] 具体实现中,所述控制器可以是图6或图7所示的控制器。

[0242] 综上,实施本申请,能够实现数据中心网络中各个中转交换机的负载均衡,提高带宽资源利用率,减小传输时延。

[0243] 在上述实施例中,可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时,可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时,全部或部分地产生按照本申请所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中,或者

从一个计算机可读存储介质向另一个计算机可读存储介质传输,例如,所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线)或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质,(例如,软盘、硬盘、磁带)、光介质(例如,DVD)、或者半导体介质(例如固态硬盘Solid State Disk)等。

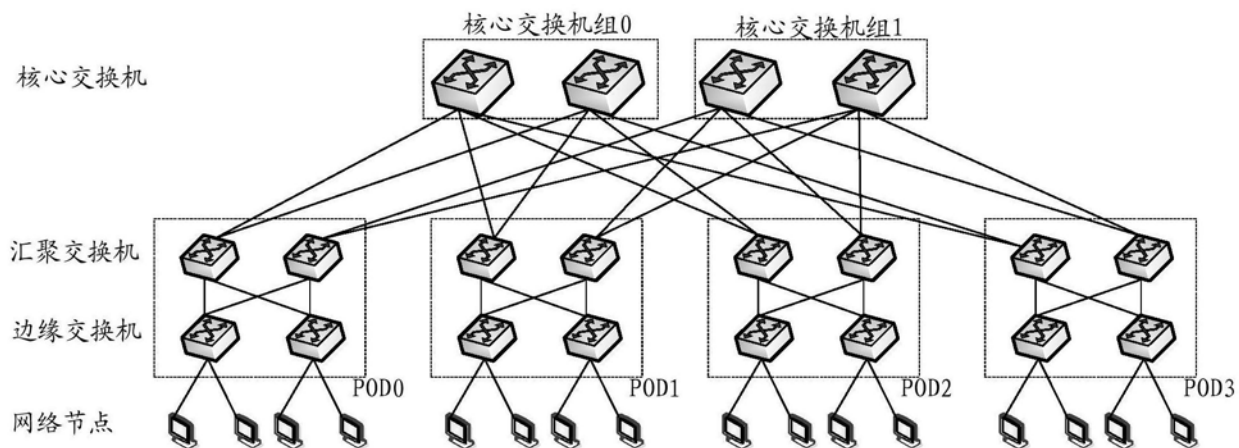


图1A

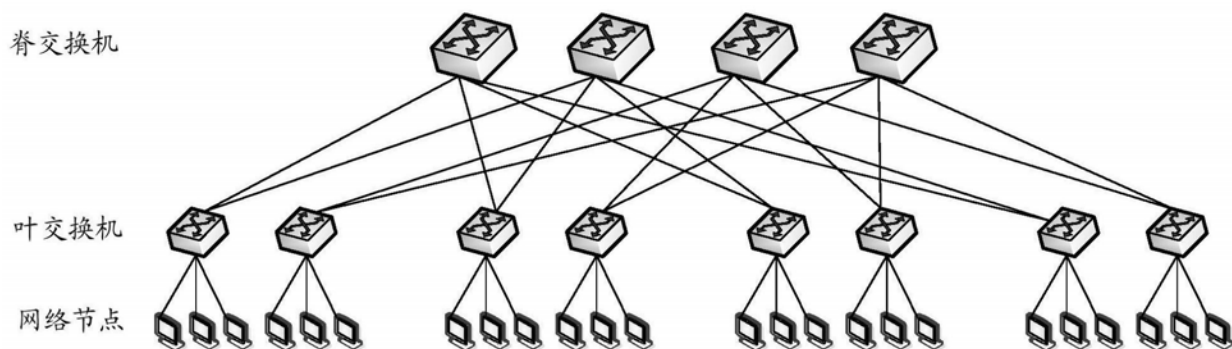


图1B

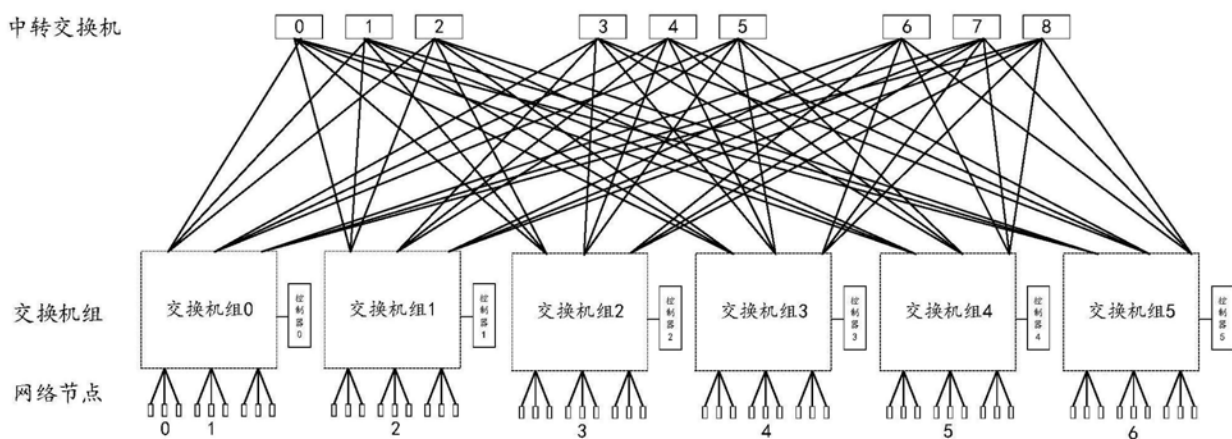


图2

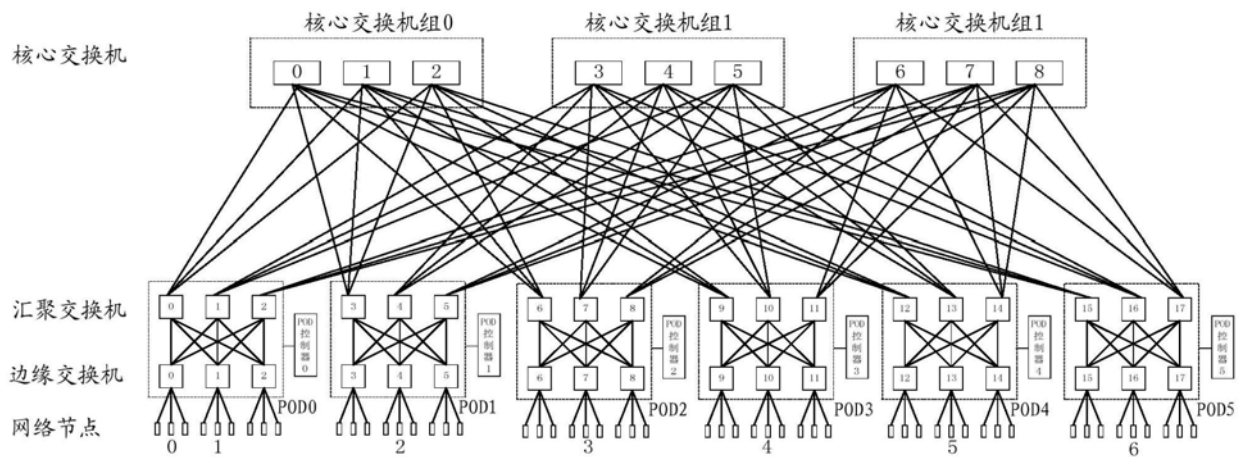


图3A

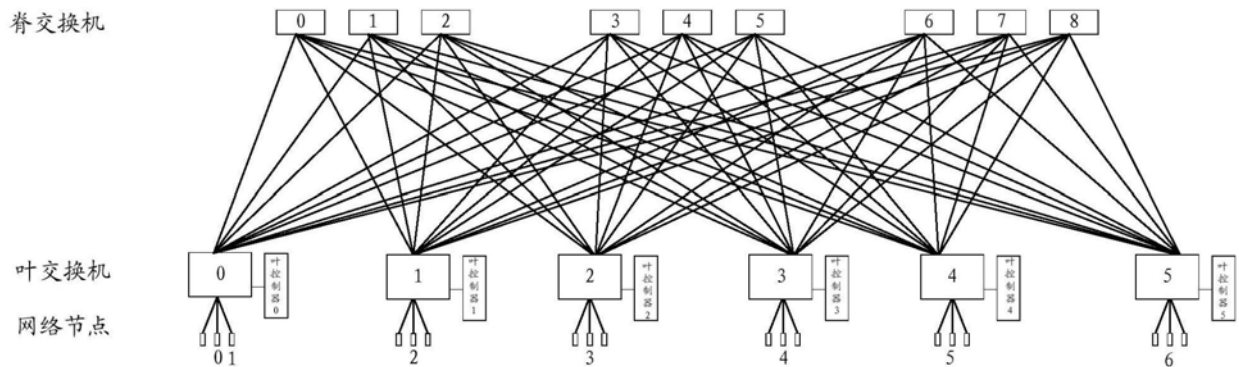


图3B

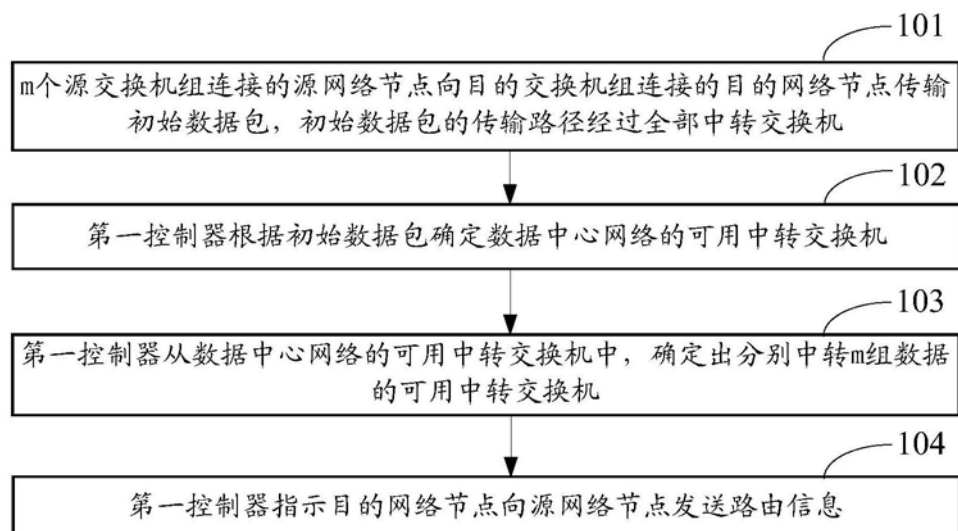


图4

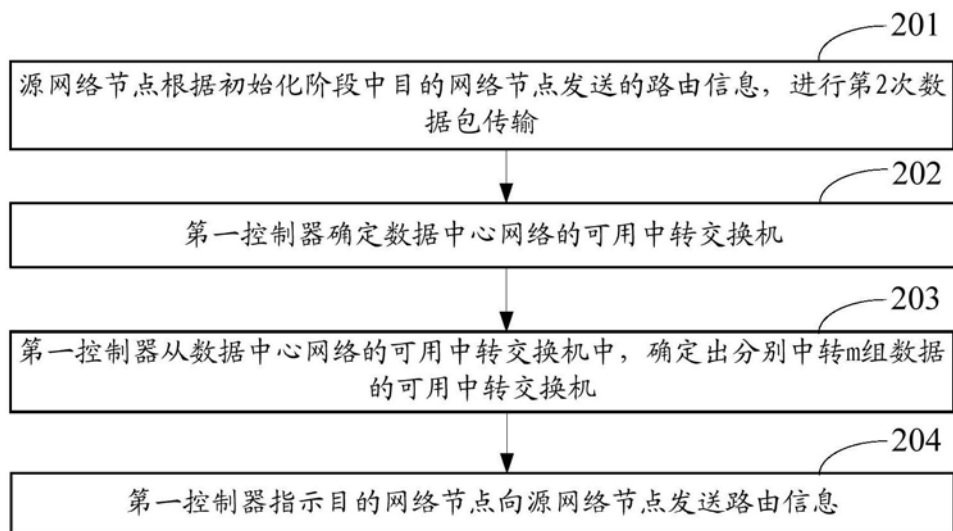


图5

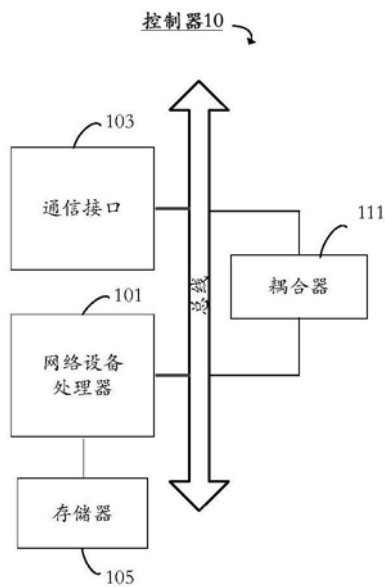


图6

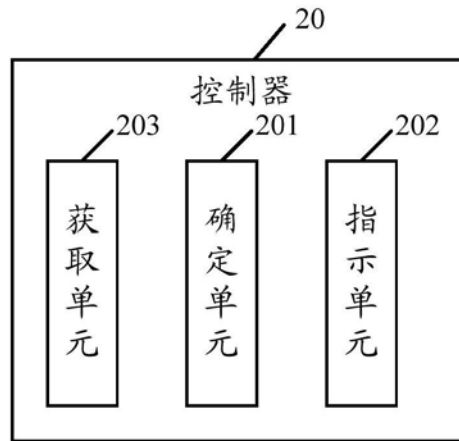


图7