(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118317366 A
(43)申请公布日 2024.07.09

(21)申请号 202410538262.7

(22)申请日 2024.04.30

(71)申请人 湘江实验室
地址 410000 湖南省长沙市高新区尖山路217号北斗产业园1栋

(72)发明人 史庆宇 蒋芳雪 刘利枚 贺泓睿 黄璜 李沁

(74)专利代理机构 长沙轩荣专利代理有限公司 43235
专利代理师 王丹

(51)Int.Cl.
*H04W 28/08*(2023.01)
*H04W 28/02*(2009.01)

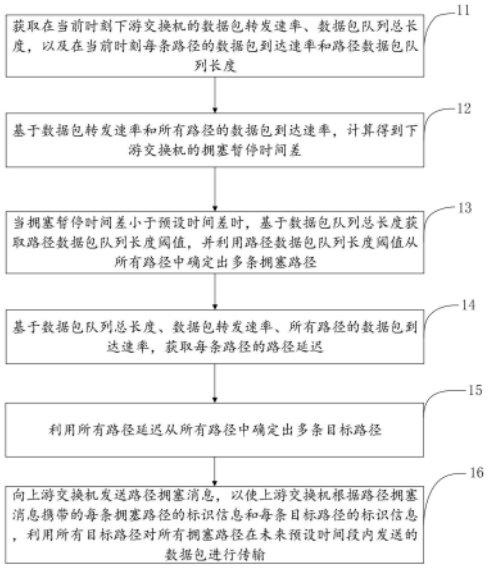
权利要求书3页 说明书11页 附图4页

(54)发明名称

一种RDMA网络数据传输的负载均衡方法及相关设备

(57)摘要

本申请涉及数据传输技术领域,提供了一种RDMA网络数据传输的负载均衡方法及相关设备。该方法包括:获取数据包转发速率、数据包队列总长度、每条路径的数据包到达速率、每条路径的路径数据包队列长度;基于数据包转发速率和所有路径的数据包到达速率计算得到拥塞暂停时间差;当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出拥塞路径;基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率获取每条路径的路径延迟;利用所有路径延迟从所有路径中确定出目标路径;向上游交换机发送路径拥塞消息。该方法能够提高数据传输的负载均衡性能。




CN 118317366 A

1.一种RDMA网络数据传输的负载均衡方法,其特征在于,应用于下游交换机,包括:

获取在当前时刻所述下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度;所述路径为上游交换机向所述下游交换机发送数据包的链路;

基于所述数据包转发速率和所有路径的数据包到达速率,计算得到所述下游交换机的拥塞暂停时间差;

当所述拥塞暂停时间差小于预设时间差时,基于所述数据包队列总长度获取路径数据包队列长度阈值,并利用所述路径数据包队列长度阈值从所有路径中确定出多条拥塞路径;所述拥塞路径的路径数据包队列长度大于所述路径数据包队列长度阈值;

基于所述数据包队列总长度、所述数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟;所述路径延迟用于描述消息传输的延迟和路径进行数据包传输的延迟;

利用所有路径延迟从所有路径中确定出多条目标路径;每条所述目标路径的路径延迟均小于所有其他路径的路径延迟;

向上游交换机发送路径拥塞消息,以使所述上游交换机根据所述路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。

2.根据权利要求1所述的负载均衡方法,其特征在于,所述基于所述数据包转发速率和所有路径的数据包到达速率,计算得到所述下游交换机的拥塞暂停时间差,包括:

通过公式:

$$\Delta t = \frac{Q_{PFC} - \left(\sum_{i=1}^{n} \int_0^{t_d} vi(t)\,dt - t_d \times vr(t)\right)}{\sum_{i=1}^{n} vi(t) - vr(t)}$$

计算所述下游交换机的拥塞暂停时间差Δt;

其中,$Q_{PFC}$表示触发拥塞暂停的数据包队列暂停总长度,$t_d$表示向上游交换机传输消息的当前延迟,n表示路径的数量,vi(t)表示当前时刻第i条路径的数据包到达速率,t表示当前时刻,vr(t)表示所述数据包转发速率。

3.根据权利要求1所述的负载均衡方法,其特征在于,所述基于所述数据包队列总长度获取路径数据包队列长度阈值,包括:

通过公式:

$$F_{cc} = Q \gg \lceil log_2(Q_T[qInd].fNum) \rceil$$

计算所述路径数据包队列长度阈值$F_{cc}$;

其中,q表示所述数据包队列总长度,$Q_T$表示记录所有路径数据包队列长度及所有路径的列表,qInd表示队列索引,fNum表示所述队列索引中的路径的数量。

4.根据权利要求1所述的负载均衡方法,其特征在于,所述基于所述数据包队列总长度、所述数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟,包括:

基于所述数据包队列总长度、所述数据包转发速率、所有路径的数据包到达速率,获取拥塞暂停恢复时间;

基于所述拥塞暂停恢复时间,获取每条路径的路径延迟。

5.根据权利要求4所述的负载均衡方法,其特征在于,所述基于所述数据包队列总长度、所述数据包转发速率、所有路径的数据包到达速率,获取拥塞暂停恢复时间,包括:

通过公式:

$$t_{re} = \frac{Q - Q^*_{PFC}}{-\Delta L} + T_{re\_flight}$$

计算所述拥塞暂停恢复时间$t_{re}$;

其中,Q表示所述数据包队列总长度,$Q^*_{PFC}$表示拥塞暂停结束时的数据包队列恢复总长度,$T_{re_flight}$表示拥塞暂停结束的消息到达上游交换机的时间,ΔL表示队列长度变化梯度:

$$\Delta L = \sum_{i=1}^{n} vi(t) - vr(t)$$

其中,n表示路径的数量,vi(t)表示当前时刻第i条路径的数据包到达速率,t表示当前时刻,vr(t)表示所述数据包转发速率。

6.根据权利要求4所述的负载均衡方法,其特征在于,所述基于所述拥塞暂停恢复时间,获取每条路径的路径延迟,包括:

获取每条路径的传输延迟;

分别针对每条路径,进行以下步骤:

判断所述路径是否为拥塞路径;

若是,则将所述拥塞暂停恢复时间与所述路径的传输延迟之和作为所述路径的路径延迟;

否则,将所述路径的传输延迟作为所述路径的路径延迟。

7.根据权利要求1所述的负载均衡方法,其特征在于,所述利用所有路径延迟从所有路径中确定出多条目标路径,包括:

将所有路径延迟由小到大进行排列,将排列结果中前预设个数的路径延迟均作为目标路径延迟;

分别将每个所述目标路径延迟对应的路径作为一目标路径。

8.一种RDMA网络数据传输的负载均衡装置,其特征在于,包括:

第一获取模块,获取在当前时刻所述下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度;所述路径为上游交换机向所述下游交换机发送数据包的链路;

计算模块,基于所述数据包转发速率和所有路径的数据包到达速率,计算得到所述下游交换机的拥塞暂停时间差;

第一确定模块,当所述拥塞暂停时间差小于预设时间差时,基于所述数据包队列总长度获取路径数据包队列长度阈值,并利用所述路径数据包队列长度阈值从所有路径中确定出多条拥塞路径;所述拥塞路径的路径数据包队列长度大于所述路径数据包队列长度阈值;

第二获取模块,基于所述数据包队列总长度、所述数据包转发速率、所有路径的数据包

到达速率,获取每条路径的路径延迟;所述路径延迟用于描述消息传输的延迟和路径进行数据包传输的延迟;

第二确定模块,利用所有路径延迟从所有路径中确定出多条目标路径;每条所述目标路径的路径延迟均小于所有其他路径的路径延迟;

发送模块,向上游交换机发送路径拥塞消息,以使所述上游交换机根据所述路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。

9.一种终端设备,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时实现如权利要求1至7任一项所述的RDMA网络数据传输的负载均衡方法。

10.一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至7任一项所述的RDMA网络数据传输的负载均衡方法。

# 一种RDMA网络数据传输的负载均衡方法及相关设备

## 技术领域

[0001] 本申请涉及数据传输技术领域,尤其涉及一种RDMA网络数据传输的负载均衡方法及相关设备。

## 背景技术

[0002] 目前,数据中心远程直接内存访问(RDMA,Remote Direct MemoryAccess)网络广泛应用于分布式存储、高性能计算(HPC,High Performance Computing)、分布式AI训练等场景。为了进一步提升数据中心无损网络的传输性能,研究者提出了一系列传输控制和负载均衡方案,以优化数据中心服务器之间RDMA通信性能。其中,与优化端到端单路径通信过程不同,负载均衡方案的目标是将节点间的网络流量均匀分配到多条路径,对提升网络传输性能至关重要。当前数据中心无损网络中RDMA数据流的数据传输负载均衡方法可通过监测不同路径的拥塞程度,将数据流重路由到拥塞程度较低的路径。另外,优先级流量控制(PFC,Priority-based Flow Control)是无损以太网中一种必要的流量控制机制,现有负载均衡机制通过在交换机对PFC暂停/恢复帧进行预测,实现对网络拥塞的提前响应,以避免拥塞导致PFC暂停带来性能损失。

[0003] 然而,尽管现有负载均衡机制在提升无损网络传输性能取得了一定的效果,但仍无法快速感知导致路径拥塞的数据流,将导致性能损失,包括:1)基于链路利用率的负载均衡:PFC暂停的路径由于链路利用率低,被认定为不拥塞路径,导致大量数据流被重路由到PFC暂停的路径,将加剧PFC暂停、PFC拥塞扩散;2)基于链路延迟的负载均衡:由于在源交换机监测路径延迟至少需要1个往返时间,而路径延迟变化快,PFC暂停/恢复的时间周期小,导致当拥塞数据流被重路由到其它路径,而其旧路径PFC暂停可能已经结束,将导致旧路径链路利用率低、新路径PFC暂停和PFC拥塞扩散;3)基于下游交换机PFC暂停预测的负载均衡:根据下游交换机端口队列数据包排队长度预测PFC暂停时间,并主动通知上游交换机该路径为拥塞路径,但由于未精确定位导致路径拥塞的数据流,易将其它正常数据流重路由到其它路径,影响传输效率和带来额外的RDMA数据包乱序。由此可见,现有的负载均衡机制存在RDMA网络数据传输的负载均衡性能低的问题。

## 发明内容

[0004] 本申请提供了一种RDMA网络数据传输的负载均衡方法及相关设备,可以解决RDMA网络数据传输的负载均衡性能低的问题。

[0005] 第一方面,本申请实施例提供了一种RDMA网络数据传输的负载均衡方法,该负载均衡方法包括:

[0006] 获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和每条路径的路径数据包队列长度;路径为上游交换机向下游交换机发送数据包的链路;

[0007] 基于数据包转发速率和所有路径的数据包到达速率,计算得到下游交换机的拥塞

暂停时间差;

[0008] 当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径;拥塞路径的路径数据包队列长度大于路径数据包队列长度阈值;

[0009] 基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟;路径延迟用于描述消息传输的延迟和路径进行数据包传输的延迟;

[0010] 利用所有路径延迟从所有路径中确定出多条目标路径;每条目标路径的路径延迟均小于所有其他路径的路径延迟;

[0011] 向上游交换机发送路径拥塞消息,以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。

[0012] 可选的,基于数据包转发速率和所有路径的数据包到达速率,计算得到下游交换机的拥塞暂停时间差,包括:

[0013] 通过公式:

[0014] $$\Delta t = \frac{Q_{PFC} - \left(\sum_{i=1}^{n} \int_{0}^{t_d} vi(t)\, dt - t_d \times vr(t)\right)}{\sum_{i=1}^{n} vi(t) - vr(t)}$$

[0015] 计算下游交换机的拥塞暂停时间差Δt;

[0016] 其中,$Q_{PFC}$表示触发拥塞暂停的数据包队列暂停总长度,$t_d$表示向上游交换机传输消息的当前延迟,n表示路径的数量,vi(t)表示当前时刻第i条路径的数据包到达速率,t表示当前时刻,vr(t)表示数据包转发速率。

[0017] 可选的,基于数据包队列总长度获取路径数据包队列长度阈值,包括:

[0018] 通过公式:

[0019] $$F_{cc} = Q \gg \lceil log_2(Q_T[qInd].fNum) \rceil$$

[0020] 计算路径数据包队列长度阈值$F_{cc}$;

[0021] 其中,Q表示数据包队列总长度,$Q_T$表示记录所有路径数据包队列长度及所有路径的列表,qInd表示队列索引,fNum表示队列索引中的路径的数量。

[0022] 可选的,基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟,包括:

[0023] 基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取拥塞暂停恢复时间;

[0024] 基于拥塞暂停恢复时间,获取每条路径的路径延迟。

[0025] 可选的,基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取拥塞暂停恢复时间,包括:

[0026] 通过公式:

[0027] $$t_{re} = \frac{Q - Q^*_{PFC}}{-\Delta L} + T_{re\_flight}$$

[0028] 计算拥塞暂停恢复时间$t_{re}$;

[0029] 其中,Q表示数据包队列总长度,$Q^*_{PFC}$表示拥塞暂停结束时的数据包队列恢复总长度,$T_{re_flight}$表示拥塞暂停结束的消息到达上游交换机的时间,ΔL表示队列长度变化梯度:

[0030] $$\Delta L = \sum_{i=1}^{n} vi(t) - vr(t)$$

[0031] 其中,n表示路径的数量,vi(t)表示当前时刻第i条路径的数据包到达速率,t表示当前时刻,vr(t)表示数据包转发速率。

[0032] 可选的,基于拥塞暂停恢复时间,获取每条路径的路径延迟,包括:

[0033] 获取每条路径的传输延迟;

[0034] 分别针对每条路径,进行以下步骤:

[0035] 判断路径是否为拥塞路径;

[0036] 若是,则将拥塞暂停恢复时间与路径的传输延迟之和作为路径的路径延迟;

[0037] 否则,将路径的传输延迟作为路径的路径延迟。

[0038] 可选的,利用所有路径延迟从所有路径中确定出多条目标路径,包括:

[0039] 将所有路径延迟由小到大进行排列,将排列结果中前预设个数的路径延迟均作为目标路径延迟;

[0040] 分别将每个目标路径延迟对应的路径作为一目标路径。

[0041] 第二方面,本申请实施例提供了一种RDMA网络数据传输的负载均衡装置,包括:

[0042] 第一获取模块,获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度;路径为上游交换机向下游交换机发送数据包的链路;

[0043] 计算模块,基于数据包转发速率和所有路径的数据包到达速率,计算得到下游交换机的拥塞暂停时间差;

[0044] 第一确定模块,当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径;拥塞路径的路径数据包队列长度大于路径数据包队列长度阈值;

[0045] 第二获取模块,基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟;路径延迟用于描述消息传输的延迟和路径进行数据包传输的延迟;

[0046] 第二确定模块,利用所有路径延迟从所有路径中确定出多条目标路径;每条目标路径的路径延迟均小于所有其他路径的路径延迟;

[0047] 发送模块,向上游交换机发送路径拥塞消息,以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。

[0048] 第三方面,本申请实施例提供了一种终端设备,包括存储器、处理器以及存储在存储器中并可在处理器上运行的计算机程序,该处理器执行上述计算机程序时实现上述的RDMA网络数据传输的负载均衡方法。

[0049] 第四方面,本申请实施例提供了一种计算机可读存储介质,该计算机可读存储介

质存储有计算机程序,该计算机程序被处理器执行时实现上述的RDMA网络数据传输的负载均衡方法。

[0050] 本申请的上述方案有如下的有益效果:

[0051] 在本申请的实施例中,通过获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度,然后基于数据包转发速率和所有路径的数据包到达速率,计算得到下游交换机的拥塞暂停时间差,当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径,然后基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟,再利用所有路径延迟从所有路径中确定出多条目标路径,最后向上游交换机发送路径拥塞消息,以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。其中,基于当前时刻的数据包转发速率和数据包到达速率计算拥塞暂停时间差,能够对拥塞状况进行实时监控,提高拥塞暂停时间差的实时性和准确性,基于数据包队列总长度、数据包转发速率、数据包到达速率确定目标路径,考虑了路径的数据包队列变化情况,使得获取的目标路径相比其他路径的性能更优,同时,将携带目标路径的标识信息和拥塞路径的标识信息的路径拥塞消息发送给上游交换机,使得上游交换机能够同时识别到拥塞路径和目标路径,便于上游交换机进行精确的负载均衡工作,进而提高数据传输的负载均衡性能。

[0052] 本申请的其它有益效果将在随后的具体实施方式部分予以详细说明。

## 附图说明

[0053] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,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。

[0054] 图1为本申请一实施例提供的RDMA网络数据传输的负载均衡方法的流程图;

[0055] 图2为本申请一实施例提供的RDMA网络数据传输的负载均衡系统的示意图;

[0056] 图3为本申请一实施例提供的RDMA网络数据传输的负载均衡系统的工作示意图;

[0057] 图4为本申请一实施例提供的RDMA网络数据传输的负载均衡装置的结构示意图;

[0058] 图5为本申请一实施例提供的终端设备的结构示意图。

## 具体实施方式

[0059] 以下描述中,为了说明而不是为了限定,提出了诸如特定系统结构、技术之类的具体细节,以便透彻理解本申请实施例。然而,本领域的技术人员应当清楚,在没有这些具体细节的其它实施例中也可以实现本申请。在其它情况中,省略对众所周知的系统、装置、电路以及方法的详细说明,以免不必要的细节妨碍本申请的描述。

[0060] 应当理解,当在本申请说明书和所附权利要求书中使用时,术语“包括”指示所描述特征、整体、步骤、操作、元素和/或组件的存在,但并不排除一个或多个其它特征、整体、

步骤、操作、元素、组件和/或其集合的存在或添加。

[0061] 还应当理解,在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合,并且包括这些组合。

[0062] 如在本申请说明书和所附权利要求书中所使用的那样,术语“如果”可以依据上下文被解释为“当...时”或“一旦”或“响应于确定”或“响应于检测到”。类似地,短语“如果确定”或“如果检测到[所描述条件或事件]”可以依据上下文被解释为意指“一旦确定”或“响应于确定”或“一旦检测到[所描述条件或事件]”或“响应于检测到[所描述条件或事件]”。

[0063] 另外,在本申请说明书和所附权利要求书的描述中,术语“第一”、“第二”、“第三”等仅用于区分描述,而不能理解为指示或暗示相对重要性。

[0064] 在本申请说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此,在本说明书中的不同之处出现的语句“在一个实施例中”、“在一些实施例中”、“在其他一些实施例中”、“在另外一些实施例中”等不是必然都参考相同的实施例,而是意味着“一个或多个但不是所有的实施例”,除非是以其他方式另外特别强调。术语“包括”、“包含”、“具有”及它们的变形都意味着“包括但不限于”,除非是以其他方式另外特别强调。

[0065] 针对现有的RDMA网络数据传输的负载均衡性能低的问题,本申请实施例提供了一种RDMA网络数据传输的负载均衡方法,该负载均衡方法通过获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度,然后基于数据包转发速率和所有路径的数据包到达速率,计算得到拥塞暂停时间差,当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径,然后基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟,再利用所有路径延迟从所有路径中确定出多条目标路径,最后向上游交换机发送路径拥塞消息,以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。其中,基于当前时刻的数据包转发速率和数据包到达速率计算拥塞暂停时间差,能够对拥塞状况进行实时监控,提高拥塞暂停时间差的实时性和准确性,基于数据包队列总长度、数据包转发速率、数据包到达速率确定目标路径,考虑了路径的数据包队列变化情况,使得获取的目标路径相比其他路径的性能更优,同时,将携带目标路径的标识信息和拥塞路径的标识信息的路径拥塞消息发送给上游交换机,使得上游交换机能够同时识别到拥塞路径和目标路径,便于上游交换机进行精确的负载均衡工作,进而提高数据传输的负载均衡性能。

[0066] 接下来对本申请提供的RDMA网络数据传输的负载均衡方法做示例性说明。

[0067] 如图1所示,本申请提供的RDMA网络数据传输的负载均衡方法包括如下步骤:

[0068] 步骤11,获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度。

[0069] 上述路径为上游交换机向下游交换机发送数据包的链路。

[0070] 在本申请的一些实施例中,可以利用iperf等传输性能测试工具获取下游交换机的数据包转发速率、数据包队列总长度,以及每条路径的数据包到达速率、每条路径的路径

数据包队列长度。

[0071] 需要说明的是,上述数据包转发速率为下游交换机将数据包转发给其他设备的速率,上述数据包队列总长度为下游交换机接收到且未转发的数据包的队列长度,上述数据包达到速率为路径中的数据包达到下游交换机的速率,上述路径数据包队列长度为路径中传输到下游交换机且未被转发的数据包的队列长度。

[0072] 步骤12,基于数据包转发速率和所有路径的数据包到达速率,计算得到下游交换机的拥塞暂停时间差。

[0073] 具体的,通过公式:

[0074] $$\Delta t = \frac{Q_{PFC} - \left(\sum_{i=1}^{n} \int_{0}^{t_d} vi(t)\, dt - t_d \times vr(t)\right)}{\sum_{i=1}^{n} vi(t) - vr(t)}$$

[0075] 计算下游交换机的拥塞暂停时间差Δt。

[0076] 其中,$Q_{PFC}$表示触发拥塞暂停的数据包队列暂停总长度,$t_d$表示向上游交换机传输消息的当前延迟,n表示路径的数量,vi(t)表示当前时刻第i条路径的数据包到达速率,t表示当前时刻,vr(t)表示数据包转发速率。

[0077] 需要说明的是,上述拥塞暂停时间差为当前时刻到触发流量控制算法(如PFC暂停)的时间差,如当前时刻为9点,拥塞暂停时间差为10分钟,则在9点10分将触发PFC暂停。

[0078] 示例性的,可以利用MATLAB、Mathematica等数学计算的计算机软件计算得到拥塞暂停时间差。

[0079] 值得一提的是,基于当前时刻的数据包转发速率和数据包到达速率计算拥塞暂停时间差,能够对拥塞状况进行实时监控,提高拥塞暂停时间差的实时性和准确性。

[0080] 步骤13,当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径。

[0081] 上述拥塞路径的路径数据包队列长度大于路径数据包队列长度阈值。上述预设时间差可以为PFC机制中设置的时间差。

[0082] 在本身申请的一些实施例中,当拥塞暂停时间差大于等于预设时间差时,说明此时尚未确定触发PFC机制,如预设时间差为10秒,即10秒后触发PFC机制,而此时拥塞暂停时间差为30秒,大于预设时间差,无需触发PFC机制。

[0083] 具体的,通过公式:

[0084] $$F_{cc} = Q \gg \lfloor log_2(Q_T[qInd].fNum) \rfloor$$

[0085] 计算路径数据包队列长度阈值$F_{cc}$;

[0086] 其中,Q表示数据包队列总长度,$Q_T$表示记录所有路径数据包队列长度及所有路径的列表,qInd表示队列索引,fNum表示队列索引中的路径的数量。

[0087] 需要说明的是,上述利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径的步骤具体为:分别针对每条路径,判断路径的路径数据包队列长度是否大于路径数据包队列长度阈值,若是,则将路径作为一拥塞路径。上述列表为下游交换机维护的数据流表flow-table(flowID,qInd,buff_pkts,sendRate,receiveRate,curr_time,td,Td),其中

flowID为用来区分不同路径的标识符,qInd为分配队列的索引,buff_pkts为flowID路径中占用的队列容量,sendRate为路径中数据包到达下游交换机入端口的到达速率,receiveRate为下游交换机入端口队列数据包的转发速率,curr_time为当前时刻,td为路径的传输延迟,Td为路径的路径延迟。上述队列索引中包括当前时刻正在进行数据包传输的所有路径的标识信息。

[0088] 示例性的,可以利用MATLAB、Mathematica等数学计算的计算机软件计算得到路径数据包队列长度阈值。

[0089] 值得一提的是,根据每条路径的路径数据包队列长度获取拥塞路径,能够精确定位当前时刻需要进行负载均衡的路径,避免出现对正常路径进行负载均衡的情况。

[0090] 步骤14,基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟。

[0091] 上述路径延迟用于描述消息传输的延迟和路径进行数据包传输的延迟。

[0092] 在本申请的一些实施例中,上述基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟的步骤具体为:

[0093] 第一步,基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取拥塞暂停恢复时间。

[0094] 具体的,通过公式:

[0095] $$t_{re} = \frac{Q - Q^*_{PFC}}{-\Delta L} + T_{re\_flight}$$

[0096] 计算拥塞暂停恢复时间$t_{re}$。

[0097] 其中,Q表示数据包队列总长度,$Q^*_{PFC}$表示拥塞暂停结束时的数据包队列恢复总长度,$T_{re_flight}$表示拥塞暂停结束的消息到达上游交换机的时间,ΔL表示队列长度变化梯度:

[0098] $$\Delta L = \sum_{i=1}^{n} vi(t) - vr(t)$$

[0099] 其中,n表示路径的数量,vi(t)表示当前时刻第i条路径的数据包到达速率,t表示当前时刻,vr(t)表示数据包转发速率。

[0100] 第二步,基于拥塞暂停恢复时间,获取每条路径的路径延迟。

[0101] 首先,获取每条路径的传输延迟。

[0102] 示例性的,可以利用iperf等传输性能测试工具获取每条路径的传输延迟。

[0103] 然后,分别针对每条路径,进行以下步骤:

[0104] 判断路径是否为拥塞路径。具体的,若该路径在上述步骤中被确定为拥塞路径,则该路径为拥塞路径。

[0105] 若是,则将拥塞暂停恢复时间与路径的传输延迟之和作为路径的路径延迟。

[0106] 否则,将路径的传输延迟作为路径的路径延迟。

[0107] 示例性的,第1条路径为拥塞路径,则该路径的路径延迟为该路径的传输延迟2秒加拥塞暂停恢复时间3秒,共5秒,第2条路径不为拥塞路径,则该路径的路径延迟等于该路

径的传输延迟3秒。

[0108] 需要说明的是,上述拥塞暂停恢复时间用于描述路径结束拥塞能够正常工作的时间,即下游交换机的入端口的数据包队列总长度达到数据包队列恢复总长度,并将拥塞暂停结束的消息传输给上游交换机的总时间。上述数据包队列恢复总长度可以由支持PFC的交换机进行配置得到,上述拥塞暂停结束的消息到达上游交换机的时间与下游交换机的转发速率、处理时延等的硬件设备相关。

[0109] 值得一提的是,获取每条路径的路径延迟时对路径的拥塞状况、路径的传输延迟进行考虑,使得路径延迟符合路径的实际状况,提高路径延迟的准确性。

[0110] 步骤15,利用所有路径延迟从所有路径中确定出多条目标路径。

[0111] 上述每条目标路径的路径延迟均小于所有其他路径的路径延迟。

[0112] 在本申请的一些实施例中,上述利用所有路径延迟从所有路径中确定出多条目标路径的步骤具体为:

[0113] 第一步,将所有路径延迟由小到大进行排列,将排列结果中前预设个数的路径延迟均作为目标路径延迟。

[0114] 第二步,分别将每个目标路径延迟对应的路径作为一目标路径。

[0115] 示例性的,对3个路径延迟进行排序:3秒,5秒,6秒,取前两个路径延迟作为目标路径延迟,将目标路径延迟对应的路径作为目标路径,即3秒和5秒对应的路径均为目标路径。

[0116] 需要说明的是,目标路径的数量可以根据拥塞路径的数量进行设置。

[0117] 值得一提的是,利用路径延迟确定出目标路径,考虑了路径的数据包队列变化情况,使得获取的目标路径相比其他路径的性能更优。

[0118] 步骤16,向上游交换机发送路径拥塞消息,以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。

[0119] 上述路径拥塞消息携带每条拥塞路径的标识信息和每条目标路径的标识信息。上述预设时间段为所有拥塞路径恢复正常的时间,如:在9点40分上游交换机接收到路径拥塞消息,并利用所有目标路径对所有拥塞路径需要发送的数据包进行传输,在10点所有拥塞路径的路径数据包队列长度均小于等于路径数据包队列长度阈值,说明此时所有拥塞路径均恢复正常,可以继续进行数据传输,则9点40分与10点之间的20分钟为预设时间段。

[0120] 在本申请的一些实施例中,下游交换机发送路径拥塞消息后,上游交换机接收路径拥塞消息,并根据每条拥塞路径的标识信息和每条目标路径的标识信息,在所有路径中识别出所有拥塞路径和所有目标路径,然后利用所有目标路径对所有拥塞路径在未来预设时间段内发送到下游交换机的数据包进行传输。

[0121] 需要说明的是,可以利用PFC机制实现利用所有目标路径对所有拥塞路径在未来预设时间段内发送到下游交换机的数据包进行传输的步骤,如为优先级高的拥塞路径分配路径延迟小的目标路径,使用目标路径将对应的拥塞路径需要发送的数据包进行传输,直到上游交换机接收到该拥塞路径恢复正常,暂停结束的消息。

[0122] 下面结合一具体实例对本申请的RDMA网络数据传输的负载均衡方法进行示例性说明。

[0123] RDMA网络数据传输的负载均衡系统如图2所示,数据流发送到上游源交换机(即上

文中的上游交换机),上游源交换机的输出端口与下游交换机的输入端口相连接,上游源交换机包括重路由模块,用于接受拥塞通告(即上文中的路径拥塞消息)并重路由到指定路径(即上文中的目标路径),下游目的交换机包括流量预测与监控模块,用于预测PFC暂停与恢复,识别拥塞流(即上文中的拥塞路径)与最优路径(即上文中的目标路径),并发送拥塞通告,图中FCN为拥塞通告。

[0124] 上述负载均衡系统进行工作时的数据传输过程如图3所示,图中Spine为上游交换机层,Leaf为下游交换机层,上游交换机与下游交换机之间的直线表示数据传输的路径,箭头表示数据传输或发送消息的方向,$f_1$表示第1个路径,$f_2$表示第2个路径,$f_{n-1}$表示第n-1个路径,$f_n$表示第n个路径,$Q_{PFC}$表示触发拥塞暂停的数据包队列暂停总长度,$F_{cc} = Q \gg [log_2 Q_T[qInd].fNum]$为上文中计算路径数据包队列长度阈值的公式,下游交换机入端口向上游交换机出端口发送FCN,FCN为拥塞通告,上游交换机根据FCN将拥塞流(即上文中的拥塞路径)切换到新路径(即上文中的目标路径)。

[0125] 值得一提的是,基于当前时刻的数据包转发速率和数据包到达速率计算拥塞暂停时间差,能够对拥塞状况进行实时监控,提高拥塞暂停时间差的实时性和准确性,基于数据包队列总长度、数据包转发速率、数据包到达速率确定目标路径,考虑了路径的数据包队列变化情况,使得获取的目标路径相比其他路径的性能更优,同时,将携带目标路径的标识信息和拥塞路径的标识信息的路径拥塞消息发送给上游交换机,使得上游交换机能够同时识别到拥塞路径和目标路径,便于上游交换机进行精确的负载均衡工作,进而提高数据传输的负载均衡性能。

[0126] 下面对本申请提供的RDMA网络数据传输的负载均衡装置进行示例性说明。

[0127] 如图4所示,本申请实施例提供了一种RDMA网络数据传输的负载均衡装置,该RDMA网络数据传输的负载均衡装置400包括:

[0128] 第一获取模块401,获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度;路径为上游交换机向下游交换机发送数据包的链路;

[0129] 计算模块402,基于数据包转发速率和所有路径的数据包到达速率,计算得到下游交换机的拥塞暂停时间差;

[0130] 第一确定模块403,当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径;拥塞路径的路径数据包队列长度大于路径数据包队列长度阈值;

[0131] 第二获取模块404,基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟;路径延迟用于描述消息传输的延迟和路径进行数据包传输的延迟;

[0132] 第二确定模块405,利用所有路径延迟从所有路径中确定出多条目标路径;每条目标路径的路径延迟均小于所有其他路径的路径延迟;

[0133] 发送模块406,向上游交换机发送路径拥塞消息,以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。

[0134] 需要说明的是,上述装置/单元之间的信息交互、执行过程等内容,由于与本申请

方法实施例基于同一构思,其具体功能及带来的技术效果,具体可参见方法实施例部分,此处不再赘述。

[0135] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,仅以上述各功能单元、模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能单元、模块完成,即将所述装置的内部结构划分成不同的功能单元或模块,以完成以上描述的全部或者部分功能。实施例中的各功能单元、模块可以集成在一个处理单元中,也可以是各个单元单独物理存在,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。另外,各功能单元、模块的具体名称也只是为了便于相互区分,并不用于限制本申请的保护范围。上述系统中单元、模块的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0136] 如图5所示,本申请的实施例提供了一种终端设备,该实施例的终端设备D10包括:至少一个处理器D100(图5中仅示出一个处理器)、存储器D101以及存储在所述存储器D101中并可在所述至少一个处理器D100上运行的计算机程序D102,所述处理器D100执行所述计算机程序D102时实现上述任意各个方法实施例中的步骤。

[0137] 具体的,所述处理器D100执行所述计算机程序D102时,通过获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度,以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度,然后基于数据包转发速率和所有路径的数据包到达速率,计算得到下游交换机的拥塞暂停时间差,当拥塞暂停时间差小于预设时间差时,基于数据包队列总长度获取路径数据包队列长度阈值,并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径,然后基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率,获取每条路径的路径延迟,再利用所有路径延迟从所有路径中确定出多条目标路径,最后向上游交换机发送路径拥塞消息,以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息,利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输。其中,基于当前时刻的数据包转发速率和数据包到达速率计算拥塞暂停时间差,能够对拥塞状况进行实时监控,提高拥塞暂停时间差的实时性和准确性,基于数据包队列总长度、数据包转发速率、数据包到达速率确定目标路径,考虑了路径的数据包队列变化情况,使得获取的目标路径相比其他路径的性能更优,同时,将携带目标路径的标识信息和拥塞路径的标识信息的路径拥塞消息发送给上游交换机,使得上游交换机能够同时识别到拥塞路径和目标路径,便于上游交换机进行精确的负载均衡工作,进而提高数据传输的负载均衡性能。

[0138] 所称处理器D100可以是中央处理单元(CPU,Central Processing Unit),该处理器D100还可以是其他通用处理器、数字信号处理器(DSP,Digital Signal Processor)、专用集成电路(ASIC,Application Specific Integrated Circuit)、现成可编程门阵列(FPGA,Field-Programmable Gate Array)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0139] 所述存储器D101在一些实施例中可以是所述终端设备D10的内部存储单元,例如终端设备D10的硬盘或内存。所述存储器D101在另一些实施例中也可以是所述终端设备D10的外部存储设备,例如所述终端设备D10上配备的插接式硬盘,智能存储卡(SMC,Smart

Media Card),安全数字(SD,Secure Digital)卡,闪存卡(Flash Card)等。进一步地,所述存储器D101还可以既包括所述终端设备D10的内部存储单元也包括外部存储设备。所述存储器D101用于存储操作系统、应用程序、引导装载程序(BootLoader)、数据以及其他程序等,例如所述计算机程序的程序代码等。所述存储器D101还可以用于暂时地存储已经输出或者将要输出的数据。

[0140] 本申请实施例还提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行时实现可实现上述各个方法实施例中的步骤。

[0141] 本申请实施例提供了一种计算机程序产品,当计算机程序产品在终端设备上运行时,使得终端设备执行时实现可实现上述各个方法实施例中的步骤。

[0142] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请实现上述实施例方法中的全部或部分流程,可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一计算机可读存储介质中,该计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。其中,所述计算机程序包括计算机程序代码,所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质至少可以包括:能够将计算机程序代码携带到RDMA网络数据传输的负载均衡方法装置/终端设备的任何实体或装置、记录介质、计算机存储器、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、电载波信号、电信信号以及软件分发介质。例如U盘、移动硬盘、磁碟或者光盘等。在某些司法管辖区,根据立法和专利实践,计算机可读介质不可以是电载波信号和电信信号。

[0143] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述或记载的部分,可以参见其它实施例的相关描述。

[0144] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0145] 以上所述是本申请的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明所述原理的前提下,还可以作出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

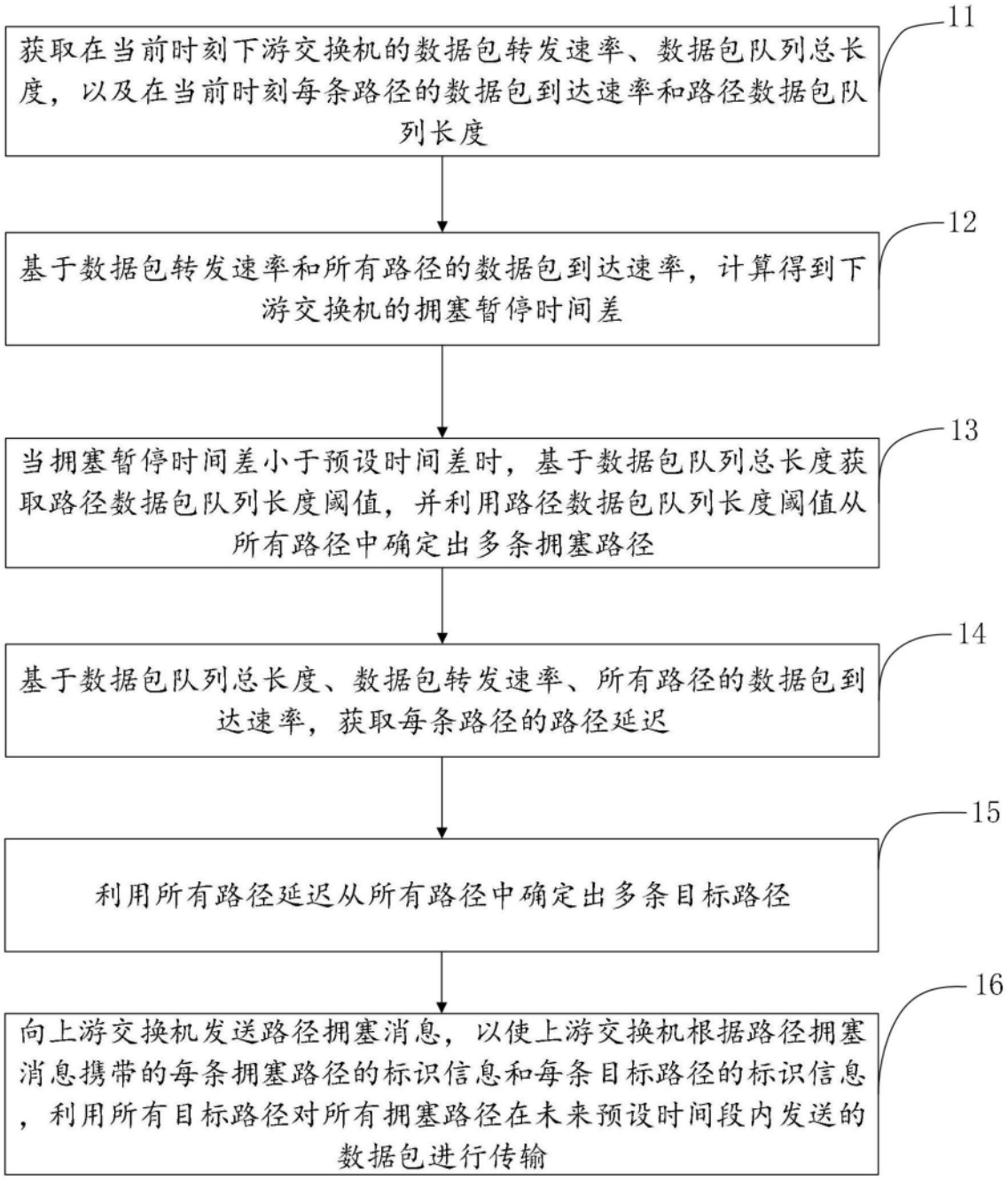

获取在当前时刻下游交换机的数据包转发速率、数据包队列总长度，以及在当前时刻每条路径的数据包到达速率和路径数据包队列长度
11
基于数据包转发速率和所有路径的数据包到达速率，计算得到下游交换机的拥塞暂停时间差
12
当拥塞暂停时间差小于预设时间差时，基于数据包队列总长度获取路径数据包队列长度阈值，并利用路径数据包队列长度阈值从所有路径中确定出多条拥塞路径
13
基于数据包队列总长度、数据包转发速率、所有路径的数据包到达速率，获取每条路径的路径延迟
14
利用所有路径延迟从所有路径中确定出多条目标路径
15
向上游交换机发送路径拥塞消息，以使上游交换机根据路径拥塞消息携带的每条拥塞路径的标识信息和每条目标路径的标识信息，利用所有目标路径对所有拥塞路径在未来预设时间段内发送的数据包进行传输
16

图1

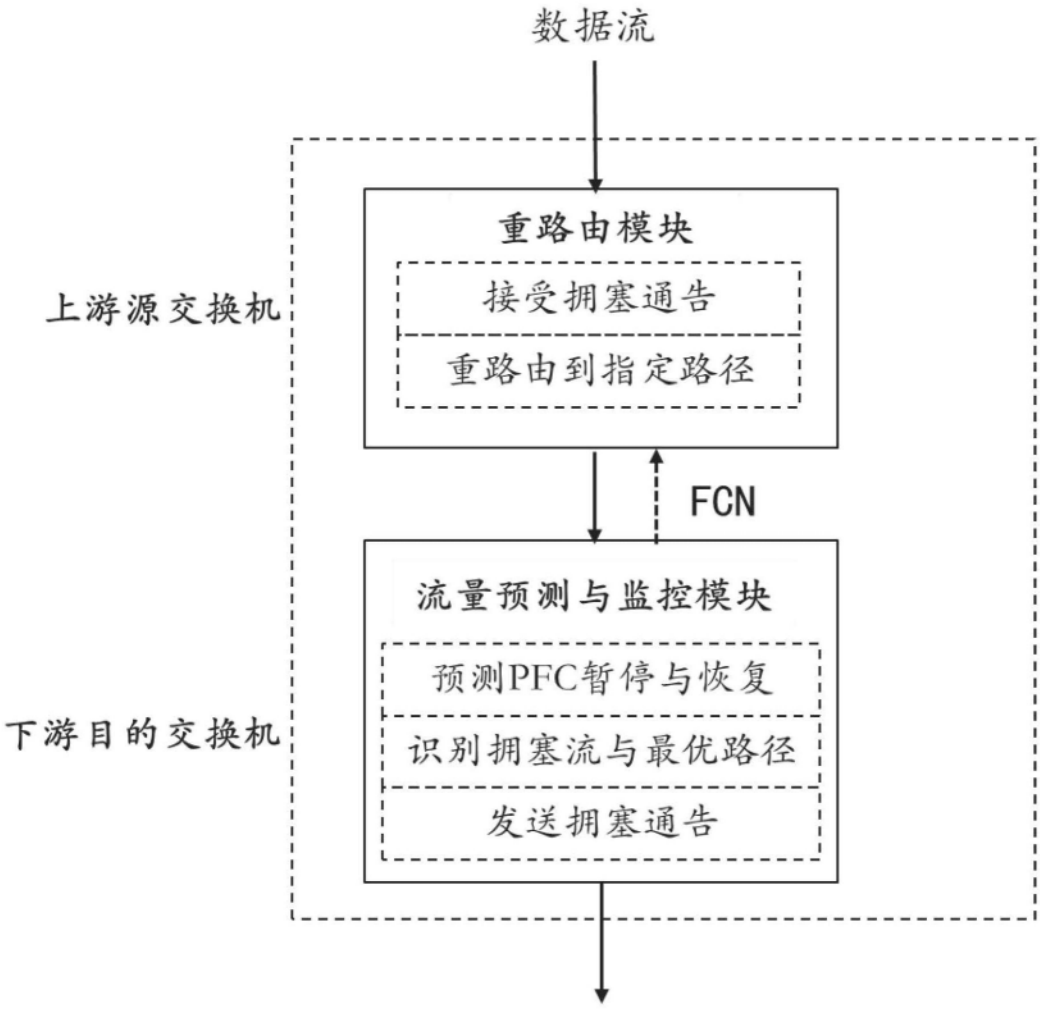

数据流
上游源交换机
重路由模块
接受拥塞通告
重路由到指定路径
FCN
流量预测与监控模块
预测PFC暂停与恢复
识别拥塞流与最优路径
发送拥塞通告
下游目的交换机

图2

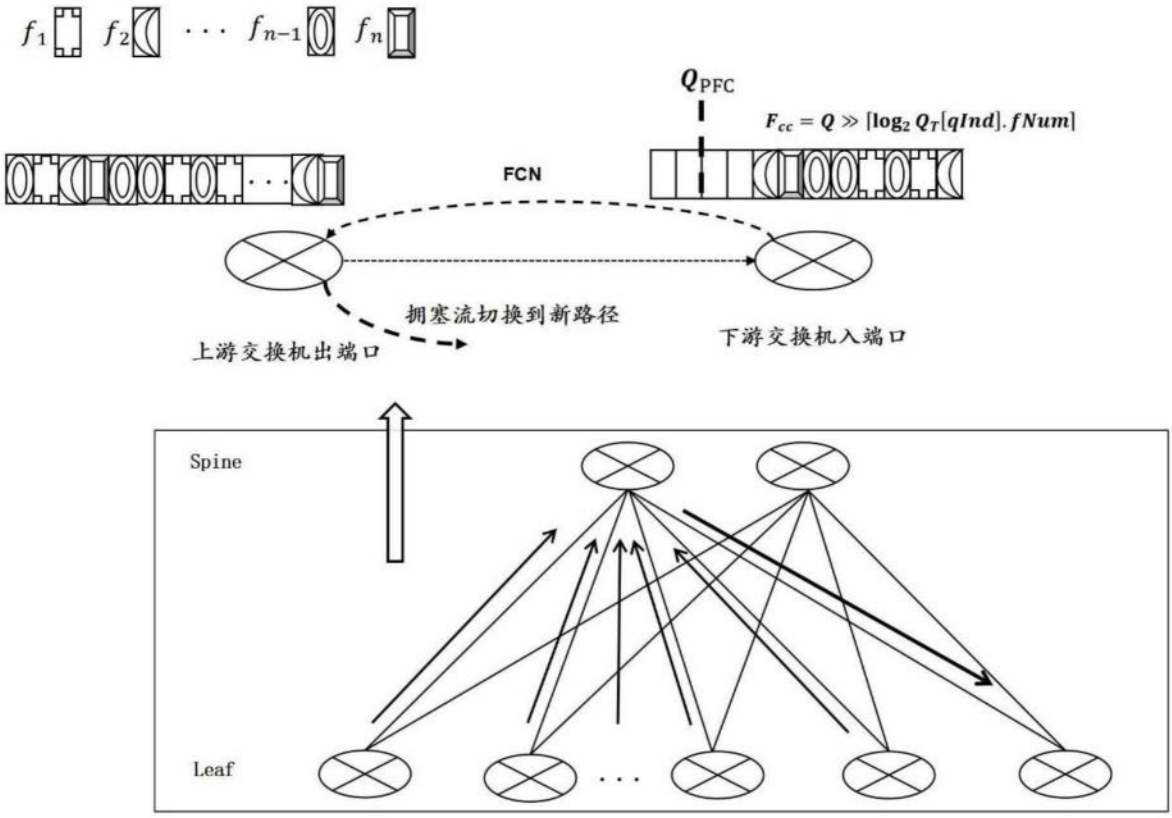

$f_1$ $f_2$ · · · $f_{n-1}$ $f_n$
$Q_{PFC}$
$F_{cc} = Q \gg [\log_2 Q_T[qInd].fNum]$
FCN
拥塞流切换到新路径
上游交换机出端口
下游交换机入端口
Spine
Leaf

图3

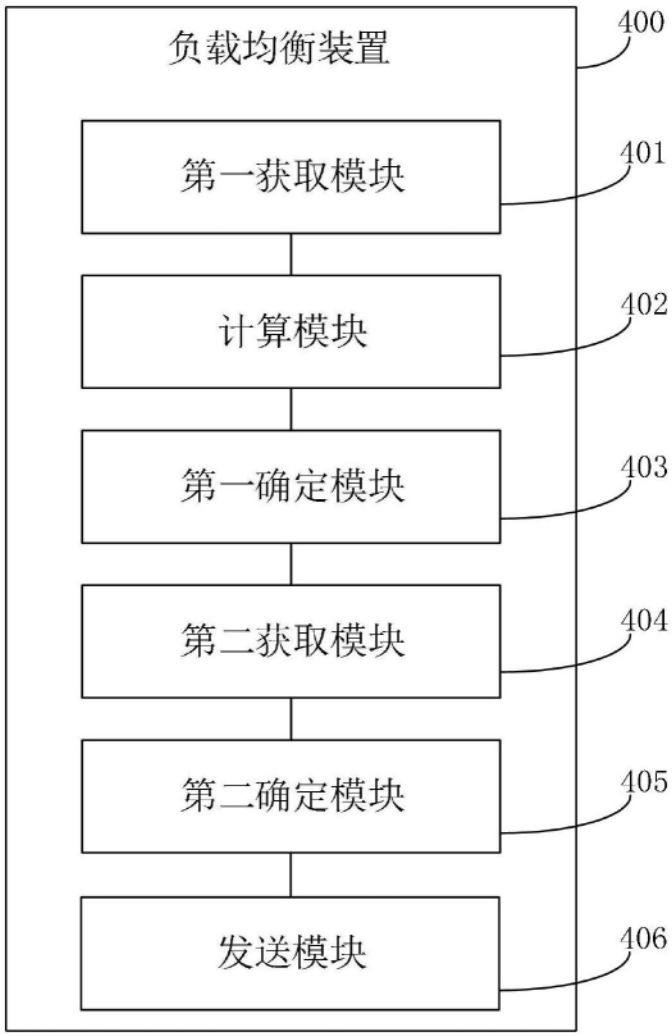

负载均衡装置
400
第一获取模块
401
计算模块
402
第一确定模块
403
第二获取模块
404
第二确定模块
405
发送模块
406

图4

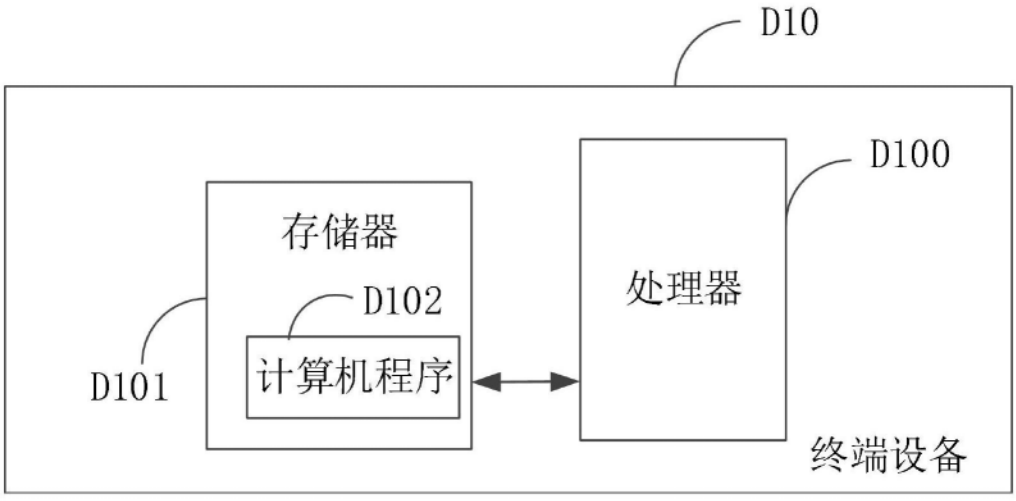

D10
D100
存储器
D102
D101
计算机程序
处理器
终端设备

图5