(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109861923 B
(45)授权公告日 2022.05.17

(21)申请号 201711242840.9
(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 109861923 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 袁庭球 徐聪 黄韬
(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
专利代理师 王仲凯
(51)Int.Cl.
H04L 47/125(2022.01)
H04L 49/15(2022.01)
H04L 49/10(2022.01)

(56)对比文件
CN 106899503 A,2017.06.27
CN 105227481 A,2016.01.06
CN 107171930 A,2017.09.15
CN 1756233 A,2006.04.05

审查员 朱文君

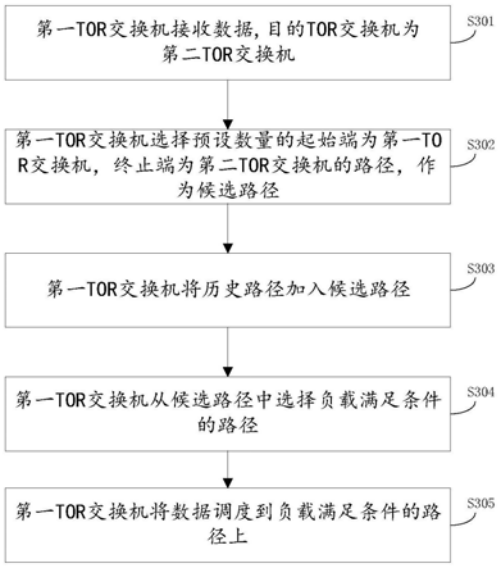
权利要求书3页 说明书8页 附图5页

(54)发明名称

一种数据调度方法及TOR交换机

(57)摘要

本申请的提供了一种数据调度方法，应用于数据中心网络的第一架顶TOR交换机，接收目的地址为第二TOR交换机的数据。随机选择候选路径，候选路径包括历史路径及从数据中心网络中选择的预设数量的、起始端为第一TOR交换机，终止端为第二TOR交换机的路径，历史路径为在历史数据调度过程中选择出的、起始端为第一TOR交换机，终止端为第二TOR交换机、并用于传输历史数据的路径。在候选路径被选择的情况下，将数据调度到候选路径中负载不大于预设的阈值的路径上。因为使用路径的负载作为数据调度的依据，能够动态感知DCN的负载变化情况，因此能够实现大规模DCN网络中多路径的负载均衡。



CN 109861923 B

1.一种数据调度方法,其特征在于,应用于数据中心网络的第一架顶TOR交换机,所述方法包括:

接收数据,所述数据的目的地址为第二TOR交换机;

随机选择候选路径,所述候选路径包括历史最优路径以及从所述数据中心网络中选择的预设数量的、起始端为所述第一TOR交换机,终止端为所述第二TOR交换机的路径,所述预设数量为大于0的整数,所述历史最优路径为在历史数据调度过程中选择出的、起始端为所述第一TOR交换机,终止端为所述第二TOR交换机、并用于传输历史数据的路径,所述预设数量小于以所述第一TOR交换机为起始端,以所述第二TOR交换机为终止端的全部路径的数量的一半;

在所述候选路径被选择的情况下,将所述数据调度到所述候选路径中负载满足条件的路径上,所述条件包括所述负载不大于预设的阈值。

2.根据权利要求1所述的方法,其特征在于,在所述随机选择候选路径之前,还包括:

以第一周期,对于所述数据中心网络中的任意一个TOR交换机,均执行:

随机选择该TOR交换机对应的第一路径和第二路径,所述第一路径和所述第二路径均为起始端为所述第一TOR交换机、终止端为该TOR交换机的路径;

获取该TOR交换机对应的历史最优路径,所述该TOR交换机对应的历史最优路径的起始端为所述第一TOR交换机、终止端为该TOR交换机;

在预先建立的路径状态表中记录所述该TOR交换机对应的所述第一路径、所述该TOR交换机对应的所述第二路径和所述该TOR交换机对应的历史最优路径的信息;

探测并在所述路径状态表中记录所述第一路径、所述第二路径和所述历史最优路径的负载。

3.根据权利要求2所述的方法,其特征在于,所述随机选择候选路径包括:

从所述路径状态表中,查询得到所述第二TOR交换机对应的所述第一路径、所述第二路径以及所述历史最优路径;

在所述将所述数据调度到所述候选路径中负载满足条件的路径上之前,还包括:

从所述路径状态表中,查询所述第二TOR交换机对应的所述第一路径的负载、所述第二路径的负载以及所述历史最优路径的负载。

4.根据权利要求1-3任一项所述的方法,其特征在于,所述随机选择候选路径包括:

在所述数据在预设时间内没有被调度过的情况下,随机选择候选路径。

5.根据权利要求4所述的方法,其特征在于,所述方法还包括:

在所述数据在预设时间内被调度过的情况下,将所述数据调度到在所述预设时间内传输过所述数据的路径上。

6.根据权利要求5所述的方法,其特征在于,在所述数据在预设时间内被调度过的情况下,将所述数据调度到在所述预设时间内传输过所述数据的路径上包括:

查询预设的流表,如果所述预设的流表中所述数据的标识对应的历史最优路径有效,则将所述数据调度到所述流表中所述数据的标识对应的所述历史最优路径上;

其中,以第二周期,维护所述流表的过程包括:

如果所述流表中任意一个数据的标识在上一周期有效,则设置该数据的标识无效;

如果所述流表中任意一个数据的标识无效,则设置该数据的标识对应的历史最优路径

无效。

7.一种架顶TOR交换机，其特征在于，所述TOR交换机为第一TOR交换机，所述第一TOR交换机包括：

接收模块，用于接收数据，所述数据的目的地址为第二TOR交换机；

选择模块，用于随机选择候选路径，所述候选路径包括历史最优路径以及从数据中心网络中选择的预设数量的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机的路径，所述预设数量为大于0的整数，所述历史最优路径为在历史数据调度过程中选择出的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机、并用于传输历史数据的路径，所述预设数量小于以所述第一TOR交换机为起始端，以所述第二TOR交换机为终止端的全部路径的数量的一半；

调度模块，用于在所述候选路径被选择的情况下，将所述数据调度到所述候选路径中负载满足条件的路径上，所述条件包括所述负载不大于预设的阈值。

8.根据权利要求7所述的TOR交换机，其特征在于，还包括：

维护模块，用于在所述选择模块随机选择候选路径之前，以第一周期，对于所述数据中心网络中的任意一个TOR交换机，均执行：随机选择该TOR交换机对应的第一路径和第二路径，所述第一路径和所述第二路径均为起始端为所述第一TOR交换机、终止端为该TOR交换机的路径；获取该TOR交换机对应的历史最优路径，所述该TOR交换机对应的历史最优路径的起始端为所述第一TOR交换机、终止端为该TOR交换机；在预先建立的路径状态表中记录所述该TOR交换机对应的所述第一路径、所述该TOR交换机对应的所述第二路径和所述该TOR交换机对应的历史最优路径的信息；探测并在所述路径状态表中记录所述第一路径、所述第二路径和所述历史最优路径的负载。

9.根据权利要求8所述的TOR交换机，其特征在于，所述选择模块用于随机选择候选路径包括：

所述选择模块具体用于，从所述路径状态表中，查询得到所述第二TOR交换机对应的所述第一路径、所述第二路径以及所述历史最优路径；

所述选择模块还用于：

在所述调度模块将所述数据调度到所述候选路径中负载满足条件的路径上之前，从所述路径状态表中，查询所述第二TOR交换机对应的所述第一路径的负载、所述第二路径的负载以及所述历史最优路径的负载。

10.根据权利要求8-9任一项所述的TOR交换机，其特征在于，所述选择模块用于随机选择候选路径包括：

所述选择模块具体用于，在所述数据在预设时间内没有被调度过的情况下，随机选择候选路径。

11.根据权利要求10所述的TOR交换机，其特征在于，所述调度模块还用于：

在所述数据在预设时间内被调度过的情况下，将所述数据调度到在所述预设时间内传输过所述数据的路径上。

12.根据权利要求11所述的TOR交换机，其特征在于，所述调度模块用于将所述数据调度到在所述预设时间内传输过所述数据的路径上包括：

所述调度模块具体用于，查询预设的流表，如果所述预设的流表中所述数据的标识对

应的历史最优路径有效，则将所述数据调度到所述流表中所述数据的标识对应的所述历史最优路径上；

所述维护模块还用于：以第二周期，维护所述流表：如果所述流表中任意一个数据的标识在上一周期有效，则设置该数据的标识无效；如果所述流表中任意一个数据的标识无效，则设置该数据的标识对应的历史最优路径无效。

一种数据调度方法及TOR交换机

技术领域

[0001] 本申请涉及电子信息领域，尤其涉及一种数据调度方法及架顶(Top of Rank，TOR)交换机。

背景技术

[0002] 作为云计算的核心基础设施，数据中心近年来得到了极大的关注。而数据中心网络(data center network，DCN)是连接数据中心大规模服务器进行分布式计算的桥梁。

[0003] 目前普遍采用的数据中心网络的架构如图1所示，图1所示为胖树(Fat-Tree)型网络结构，其中包括三层拓扑结构，从上至下依次为核心交换机、汇聚交换机和TOR交换机。与传统树型结构不同的是，图1中，TOR交换机和汇聚交换机被划分为不同的集群。在每个集群中，每个TOR交换机与每个汇聚交换机相连，构成一个完全二分图。而每个汇聚交换机与某部分核心交换机连接，使得每个集群与任何一个核心交换机相连。任意两个TOR交换机之间存在着多条端到端的路径，从而保证网络的超额订购率，实现端到端路径的高带宽、无阻塞通信。

[0004] 而目前，现有的用于提升图1所示的大规模DCN的性能的数据流调度方案，均不能实现DCN网络中多路径的实时负载均衡。

发明内容

[0005] 本申请提供了一种数据调度方法及TOR交换机，目的在于解决如何实现大规模DCN网络中多路径的负载均衡的问题。

[0006] 本申请的第一方面提供了一种数据调度方法，应用于数据中心网络的第一架顶TOR交换机，所述方法包括：接收目的地址为第二TOR交换机的数据。随机选择候选路径，所述候选路径包括历史路径以及从所述数据中心网络中选择的预设数量的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机的路径，所述预设数量为大于0的整数，所述历史路径为在历史数据调度过程中选择出的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机、并用于传输历史数据的路径。在所述候选路径被选择的情况下，将所述数据调度到所述候选路径中负载满足条件的路径上，所述条件包括所述负载不大于预设的阈值。因为使用路径的负载作为数据调度的依据，能够动态感知DCN的负载变化情况，对异构以及动态性强的数据中心网络环境具有更强的适应能力，因此能够实现大规模DCN网络中多路径的负载均衡。

[0007] 本申请的第二方面提供了一种架顶TOR交换机，所述TOR交换机为第一TOR交换机，所述第一TOR交换机包括：接收模块、选择模块和调度模块。其中，接收模块用于接收数据，所述数据的目的地址为第二TOR交换机。选择模块用于随机选择候选路径，所述候选路径包括历史路径以及从所述数据中心网络中选择的预设数量的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机的路径，所述预设数量为大于0的整数，所述历史路径为在历史数据调度过程中选择出的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机、

并用于传输历史数据的路径。调度模块用于在所述候选路径被选择的情况下，将所述数据调度到所述候选路径中负载满足条件的路径上，所述条件包括所述负载不大于预设的阈值。所述TOR交换机能够实现大规模DCN网络中多路径的负载均衡。

[0008] 本申请的第三方面提供了一种架顶TOR交换机，所述TOR交换机为第一TOR交换机，所述第一TOR交换机包括：接收器和处理器。其中接收器用于接收数据，所述数据的目的地址为第二TOR交换机。所述处理器用于随机选择候选路径，所述候选路径包括历史路径以及从所述数据中心网络中选择的预设数量的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机的路径，所述预设数量为大于0的整数，所述历史路径为在历史数据调度过程中选择出的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机、并用于传输历史数据的路径，并且，在所述候选路径被选择的情况下，将所述数据调度到所述候选路径中负载满足条件的路径上，所述条件包括所述负载不大于预设的阈值。

[0009] 在一个实现方式中，在所述随机选择候选路径之前，还包括：

[0010] 以第一周期，对于所述数据中心网络中的任意一个TOR交换机，均执行：随机选择该TOR交换机对应的第一路径和第二路径，所述第一路径和所述第二路径均为起始端为所述第一TOR交换机、终止端为该TOR交换机的路径。获取该TOR交换机对应的历史路径，所述该TOR交换机对应的历史路径的起始端为所述第一TOR交换机、终止端为该TOR交换机。在预先建立的路径状态表中记录所述该TOR交换机对应的所述第一路径、所述该TOR交换机对应的所述第二路径和所述该TOR交换机对应的历史路径的信息。探测并在所述路径状态表中记录所述第一路径、所述第二路径和所述历史路径的负载。预先在路径状态表中记录TOR交换机对应的路径的信息和负载，有利于快速获取候选路径。

[0011] 在一个实现方式中，所述随机选择候选路径包括：从所述路径状态表中，查询得到所述第二TOR交换机对应的所述第一路径、所述第二路径以及所述历史路径。在所述将所述数据调度到所述候选路径中负载满足条件的路径上之前，还包括：从所述路径状态表中，查询所述第二TOR交换机对应的所述第一路径的负载、所述第二路径的负载以及所述历史路径的负载。

[0012] 在一个实现方式中，所述随机选择候选路径包括：在所述数据在预设时间内没有被调度过的情况下，随机选择候选路径。

[0013] 在一个实现方式中，所述方法还包括：在所述数据在预设时间内被调度过的情况下，将所述数据调度到在所述预设时间内传输过所述数据的路径上。

[0014] 在一个实现方式中，在所述数据在预设时间内被调度过的情况下，将所述数据调度到在所述预设时间内传输过所述数据的路径上包括：查询预设的流表，如果所述预设的流表中所述数据的标识对应的历史路径有效，则将所述数据调度到所述流表中所述数据的标识对应的所述历史路径上。其中，以第二周期，维护所述流表的过程包括：如果所述流表中任意一个数据的标识在上一周期有效，则设置该数据的标识无效。如果所述流表中任意一个数据的标识无效，则设置该数据的标识对应的历史路径无效。

[0015] 在一个实现方式中，所述预设数值小于以所述第一TOR交换机为起始端，以所述第二TOR交换机为终止端的全部路径的数量的一半，从而获得更优的负载均衡效果。

附图说明

[0016] 图1为DCN的示意图；

[0017] 图2为图1所示的DCN中的数据流的示意图；

[0018] 图3为本申请实施例公开的一种数据调度方法的流程图；

[0019] 图4为本申请实施例公开的又一种数据调度方法的流程图；

[0020] 图5为本申请实施例公开的又一种数据调度方法的流程图；

[0021] 图6(a)和图6(b)为本申请实施例公开的一种数据调度方法的收敛性的效果示意图；

[0022] 图7为本申请实施例公开的一种TOR交换机的结构示意图。

具体实施方式

[0023] 在图1所示的DCN中的交换机(包括核心交换机、汇聚交换机和TOR交换机)，设置有不同类型的应用，不同类型的应用用于执行客户端请求的不同的业务。图1所示的DCN允许不同类型的应用并发执行。应用执行的过程中，在DCN中传输各自的数据，因此，不同类型的应用在并发执行的过程中，DCN中的数据流呈现显著的非均匀分布特性，图2所示。

[0024] 图1以及图2中所示的物理交换机以及链路可能是异构的，性能各异。

[0025] 本申请实施例所示的数据调度方法应用在图1或者图2所示的TOR交换机上。核心交换机和汇聚交换机均不对数据进行再次调度。本申请的实施例中，所述“调度”是指，为数据分配传输路径，并使用分配的传输路径传输数据。在本申请的以下实施例中，调度的颗粒度可以为：数据流(data flow)、数据包(data packet)以及批量数据包(介于数据流和数据包之间的颗粒度)。

[0026] 下面将结合附图，对本申请实施例公开的技术方案进行详细的说明。显然，以下所描述的实施例仅仅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他实施例，都属于本申请保护的范围。

[0027] 图3为本申请实施例公开的一种数据调度方法，包括以下步骤：

[0028] S301：第一TOR交换机接收数据。

[0029] 其中，第一TOR交换机为图1或图2中的任意一个TOR交换机。第一TOR交换机接收到的数据中包括目的TOR交换机的地址信息，为了便于区分，以下将数据中包括的目的TOR交换机称为第二TOR交换机。

[0030] S302：第一TOR交换机随机选择预设数量的起始端为第一TOR交换机，终止端为第二TOR交换机的路径，作为候选路径。

[0031] S303：第一TOR交换机将历史路径加入候选路径。

[0032] 其中，历史路径为第一TOR交换机在历史(例如上一次)数据调度过程中选择出的、起始端为第一TOR交换机，终止端为第二TOR交换机、并用于传输历史数据的路径。

[0033] 将历史路径加入候选路径的目的在于，使得本实施例所述的调度方法在非对称网络架构及非均匀作业流分布下，依然可以保持收敛性(即不会出现某条路径负载过重，而其它某些路径处于闲置的状况)。

[0034] S304：第一TOR交换机从候选路径中选择负载满足条件的路径。

[0035] 具体的,可以使用路径的往返时延(round trip time,RTT)或者带宽表示路径的负载。条件包括:负载不大于预设的阈值。其中,预设的阈值的一个示例可以为:候选路径的负载中的最小值。

[0036] 例如,通过探测选择的路径的RTT,将RTT最小的路径作为负载满足条件的路径。

[0037] 需要说明的是,结合S302和S304可以看出,预设数量的值越大,第一TOR交换机需要选择的路径越多,则后续需要计算的负载数值越多,因此,第一TOR交换机转发数据的时延越长,并且,因为S304需要比较各条路径的负载,因此,第一TOR交换机需要存储计算出的负载。

[0038] 并且,预设数量与负载均衡效果的关系为:当数据中心规模变大时,探测路径状态的耗时增长,导致路径信息的收集时间增长。如果从全部路径中选择负载最轻的路径,因为需要探测的路径的数量巨大,所以探测信息的延迟会造成各个端口的数据积压状况发生巨大的波动,每个时间点会有部分端口的负载极其严重而其它端口的数据积压几乎为零,反而导致负载极其不均衡的状况。

[0039] 因此,实际应用中,可以依据时延需求和/或存储空间的实际情况,确定预设数量。但是,考虑到存储和计算开销,以及能够获得较优的负载均衡的效果,本实施例中,预设数量小于以第一TOR交换机为起始端,以第二TOR交换机为终止端的全部路径的数量的一半。

[0040] 使用本实施例所述的方法,所有端口的数据积压量都严格控制在一个固定的上界之内,即负载波动情况远没有使用探测全部路径时剧烈,负载均衡情况得到了极大的改善。特别地,当每次仅随机探测2条路径时,所有端口的数据积压状况都被限制在5个数据包以内,调度结果的负载均衡状况最佳。因此,当限定择路范围为2时,探测信息的延误对调度结果的影响最小,是最适合大规模数据中心的配置方式。当探测的路径的数量上升5时,与数量为2相比,负载均衡的效果有所下降。因此,申请人通过上述研究过程发现,随机选择2条路径,为最优的方式。

[0041] S305:第一TOR交换机将数据调度到负载满足条件的路径上。

[0042] 完成数据的调度后,第一TOR交换机使用负载满足条件的路径,将接收到的数据传输到目的TOR交换机。

[0043] 从图3所示的步骤可以看出,使用路径的负载作为数据调度的依据,能够动态感知DCN的负载变化情况,对异构以及动态性强的数据中心网络环境具有更强的适应能力,因此能够有效提高DCN的性能。

[0044] 更为重要的是,因为本实施例只选择了部分以第一TOR交换机为起始端,以目的TOR交换机为终止端的路径作为候选路径,所以,能够降低存储和计算开销,并且,还能够保证获得的负载信息相对于数据传输的实时性。同时,候选路径中加入局部最优解(即历史路径),保证路径选择的收敛性。

[0045] 图4为本申请实施例公开的又一种数据调度方法,与图3所示的方法的区别在于:

[0046] 在S301之后,S302之前,第一TOR交换机查询该数据在预设时间内是否被调度过,如果是,第一TOR交换机将该数据调度到在预设时间内传输过该数据的路径上。如果否,第一TOR交换机执行S302-S304。

[0047] 目的在于,在数据在短期内被调度过的情况下,无需重新选择路径,而使用上次传输该数据的路径,以节省第一TOR交换机的资源。

[0048] 图4中与图3中的相同步骤不再赘述。

[0049] 下面将结合图5,以数据流为例,对图4所示的方法中各个步骤的具体实现方式进行详细的说明。

[0050] 图5中,第一TOR交换机中存储流表和路径状态表。

[0051] 其中,流表用于存储数据流的标识信息以及数据流的历史路径的信息。两者存在对应关系。其中,数据流的历史路径为上一次传输同一数据流中的数据包的端到端路径。数据信息的更新周期为T1。

[0052] 流表中任意一条数据流的标识信息以及其对应的历史路径的维护过程为:

[0053] 设置计时器周期为T1。设置数据流的标识信息对应的历史路径的效力标识Valid bit,Valid bit的值为“TRUE”表示历史路径有效,Valid bit的值为“FALSE”表示历史路径失效。设置失效标识Aging bit,Aging bit的值为“TRUE”表示该数据流的信息失效,需要更新,Aging bit的值为“FALSE”表示该数据流的信息有效。在接收到该条数据流的历史路径时,设置Aging bit的初始值为“FALSE”,设置Valid bit的初始值为“TRUE”。

[0054] 以T1为周期,执行以下步骤:

[0055] S11:判断Aging bit的值,如果Aging bit=“FALSE”,执行S12,如果Aging bit=“TRUE”,执行S13。

[0056] S12:设置Aging bit=“TRUE”。执行S13。

[0057] S13:设置Valid bit=“FALSE”。

[0058] 从S11-S13可以看出,数据流的历史路径的有效时长为T1。

[0059] 路径状态表用于存储第一TOR交换机选择出的各个TOR交换机的标识及其对应的最优端到端路径的信息。任意一个TOR交换机对应的最优端到端路径的起始端为第一TOR交换机,终止端为该TOR交换机。最优端到端路径的更新周期为T2。

[0060] 路径状态表的维护过程为:

[0061] 设置计时器周期为T2,设置失效标识Aging bit,Aging bit的值为“TRUE”表示路径状态表失效,需要更新,Aging bit的值为“FALSE”表示路径状态表有效。Aging bit的初始值可以设置为“TRUE”或者“FALSE”。

[0062] 以T2为周期,对于任意一个TOR交换机TOR i,执行以下步骤:

[0063] S21:判断Aging bit的值,如果Aging bit=“FALSE”,执行S22,如果Aging bit=“TRUE”,执行S23-S25。

[0064] S22:设置Aging bit=“TRUE”。

[0065] S23:第一TOR交换机随机选择两条端到端路径r1和r2。r1和r2的起始端为第一TOR交换机,终止端为TOR i。

[0066] S24:获取历史最优端到端路径r3。历史最优端到端路径r3为用于上一次传输数据的、起始端为第一TOR交换机、终止端为TOR i的路径。

[0067] S25:探测并在路径状态表中记录r1、r2和r3的RTT。

[0068] S26:设置Aging bit=“FALSE”。

[0069] 从S21-S26可以看出,最优端到端路径的有效时长为T2,且该路径状态表保证探测到的实时路径状态信息延迟不会超过T2。

[0070] 需要说明的是,在使用RTT表示负载的情况下,满足RTT的上限值≤T2≤T1。而在使

用其它参数例如带宽表示负载的情况下，不限定T2的下限值为RTT的上限值。

[0071] 第一TOR交换机基于流表和路径状态表进行数据流调度的过程为：

[0072] S31：在接收到任意一个数据包(假设数据包的目的地址为第二TOR交换机)后，在流表中查询是否存在该数据包所在的数据流的标识，如果是，执行S32，如果否，执行S34。

[0073] S32：如果数据流的标识对应的Valid bit＝“TRUE”，执行S33，如果数据流的标识对应的Valid bit＝“FALSE”，执行S34。

[0074] S33：将数据包调度到流表中的数据流的标识对应的历史路径上。

[0075] S34：基于路径状态表，调度数据包，具体包括：

[0076] 1、从路径状态表中，查询起始端为第一TOR交换机、终止端为第二TOR交换机的端到端路径r11、r21和r31以及三条端到端路径的RTT。

[0077] 2、将数据包调到r11、r21和r31中RTT最小的路径上(简称为目标路径)。

[0078] 在使用目标路径传输数据包的过程中，目标路径上包括的核心交换机和汇聚交换机仅执行数据包的转发，而不再进行选择路径的操作。

[0079] S35：将目标路径的信息和数据流的标识对应记录到流表中。其中，目标路径作为流程中的历史路径。

[0080] S36：将目标路径的信息作为历史最优端到端路径记录到路径状态表中。

[0081] S35和S36的顺序可以交换。

[0082] S37：第一TOR交换机将r11、r21和r31的RTT反馈给其它TOR交换机。

[0083] 具体的，第一TOR交换机可以将r11、r21和r31的RTT发给DCN的控制器Controller，并由Controller将r11、r21和r31的RTT发送到其它DCN交换机。

[0084] S37的目的在于，如果其它TOR交换机需要探测r11、r21和r31，则无需重新查询路径信息表，以提高调度效率。

[0085] 从图3所示的过程可以看出，第一TOR交换机以T1为周期维护流表，以T2为周期维护路径状态表，并在接收到数据包后，通过查询流表和路径状态表调度数据包，以将数据包调度到负载相对较轻的路径上。与现有的数据流调度方法相比，具有以下优势：

[0086] 1、设计了随机路径状态探测的方法，优化了大规模DCN中状态信息的存储和轮询开销。

[0087] 2、记录了每轮数据调度的局部最优解，并迭代到下一轮的候选解当中，从而保证了异构数据中心环境下调度方法的收敛性。

[0088] 3、将随机择路的范围具体限定为2条，每次仅随机探测2条端到端路径的拥塞状态，此探测范围下，调度策略对过期信息的敏感程度最低，极大优化了过时信息对于调度结果的影响。

[0089] 图6为图5所示的示例的收敛性的效果示意：

[0090] 其中，图6(a)为一个6端口的DCNTOR交换机，在现有的绝对负载均衡的调度策略(利用绝对实时的全局状态信息)下端口数据的积压状况。图6(b)为图5所示的数据调度方法下同一个端口的数据的积压状况。其中，横轴为数据传输的时间(单位为秒)，纵轴为数据包在端口的积压量(单位为个)。

[0091] 可以看到，在绝对负载均衡的调度策略下，TOR交换机各个端口的负载状况变化曲线是完全重合的，说明同一时刻各端口的数据积压量是完全相等的，这是理论上负载均衡

最佳的情况。而在图5所示的方法下，虽然各个端口的负载状况变化曲线出现了一定的波动，但是任意时刻负载最重端口和负载最轻端口之间数据积压量的差值有一个明确的上限(小于20个数据包)。因此，图5所示的方法，虽然仅探测并记录少量的路径状态信息，但依然可以达到较为理想的负载均衡状况。

[0092] 图7为本申请实施例公开的一种架顶TOR交换机，所述TOR交换机为第一TOR交换机，所述第一TOR交换机包括：接收模块、选择模块和调度模块。可选的，还可以包括维护模块。

[0093] 其中，接收模块用于接收数据，所述数据的目的地址为第二TOR交换机。

[0094] 选择模块用于随机选择候选路径，所述候选路径包括历史路径以及从所述数据中心网络中选择的预设数量的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机的路径，所述预设数量为大于0的整数，所述历史路径为在历史数据调度过程中选择出的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机、并用于传输历史数据的路径。具体的，在所述数据在预设时间内没有被调度过的情况下，所述选择模块随机选择候选路径。所述预设数值可以小于以所述第一TOR交换机为起始端，以所述第二TOR交换机为终止端的全部路径的数量的一半。

[0095] 调度模块用于在所述候选路径被选择的情况下，将所述数据调度到所述候选路径中负载满足条件的路径上，所述条件包括所述负载不大于预设的阈值。在所述数据在预设时间内被调度过的情况下，将所述数据调度到在所述预设时间内传输过所述数据的路径上。

[0096] 维护模块用于在所述选择模块随机选择候选路径之前，以第一周期，对于所述数据中心网络中的任意一个TOR交换机，均执行：随机选择该TOR交换机对应的第一路径和第二路径，所述第一路径和所述第二路径均为起始端为所述第一TOR交换机、终止端为该TOR交换机的路径；获取该TOR交换机对应的历史路径，所述该TOR交换机对应的历史路径的起始端为所述第一TOR交换机、终止端为该TOR交换机；在预先建立的路径状态表中记录所述该TOR交换机对应的所述第一路径、所述该TOR交换机对应的所述第二路径和所述该TOR交换机对应的历史路径的信息；探测并在所述路径状态表中记录所述第一路径、所述第二路径和所述历史路径的负载。

[0097] 基于维护模块维护的路径状态表，具体的，所述选择模块随机选择候选路径的方式为：从所述路径状态表中，查询得到所述第二TOR交换机对应的所述第一路径、所述第二路径以及所述历史路径。所述选择模块在所述调度模块将所述数据调度到所述候选路径中负载满足条件的路径上之前，还可以从所述路径状态表中，查询所述第二TOR交换机对应的所述第一路径的负载、所述第二路径的负载以及所述历史路径的负载。

[0098] 所述维护模块还用于：以第二周期，维护所述流表：如果所述流表中任意一个数据的标识在上一周期有效，则设置该数据的标识无效；如果所述流表中任意一个数据的标识无效，则设置该数据的标识对应的历史路径无效。

[0099] 基于维护模块维护的所述流表，具体的，调度模块将所述数据调度到在所述预设时间内传输过所述数据的路径上的方式为：查询预设的流表，如果所述预设的流表中所述数据的标识对应的历史路径有效，则将所述数据调度到所述流表中所述数据的标识对应的所述历史路径上。

[0100] 所述TOR交换机能够实现大规模DCN网络中多路径的负载均衡。

[0101] 本申请实施例还公开了一种架顶TOR交换机，所述TOR交换机为第一TOR交换机，所述第一TOR交换机包括：接收器和处理器。

[0102] 其中接收器用于接收数据，所述数据的目的地址为第二TOR交换机。所述处理器用于随机选择候选路径，所述候选路径包括历史路径以及从所述数据中心网络中选择的预设数量的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机的路径，所述预设数量为大于0的整数，所述历史路径为在历史数据调度过程中选择出的、起始端为所述第一TOR交换机，终止端为所述第二TOR交换机、并用于传输历史数据的路径，并且，在所述候选路径被选择的情况下，将所述数据调度到所述候选路径中负载满足条件的路径上，所述条件包括所述负载不大于预设的阈值。

[0103] 所述处理器还用于：在所述数据在预设时间内被调度过的情况下，将所述数据调度到在所述预设时间内传输过所述数据的路径上。

[0104] 所述处理器还用于：从所述路径状态表中，查询所述第二TOR交换机对应的所述第一路径的负载、所述第二路径的负载以及所述历史路径的负载。基于所述路径状态表，所述处理器随机选择候选路径的具体实现方式为：从所述路径状态表中，查询得到所述第二TOR交换机对应的所述第一路径、所述第二路径以及所述历史路径。

[0105] 所述处理器还用于：以第二周期，维护所述流表：如果所述流表中任意一个数据的标识在上一周期有效，则设置该数据的标识无效。如果所述流表中任意一个数据的标识无效，则设置该数据的标识对应的历史路径无效。基于流表，所述处理器在所述数据在预设时间内被调度过的情况下，将所述数据调度到在所述预设时间内传输过所述数据的路径上的具体实现方式为：查询预设的流表，如果所述预设的流表中所述数据的标识对应的历史路径有效，则将所述数据调度到所述流表中所述数据的标识对应的所述历史路径上。

[0106] 以上所述TOR交换机的功能的具体实现方式，可以参见上述方法实施例，这里不再赘述。

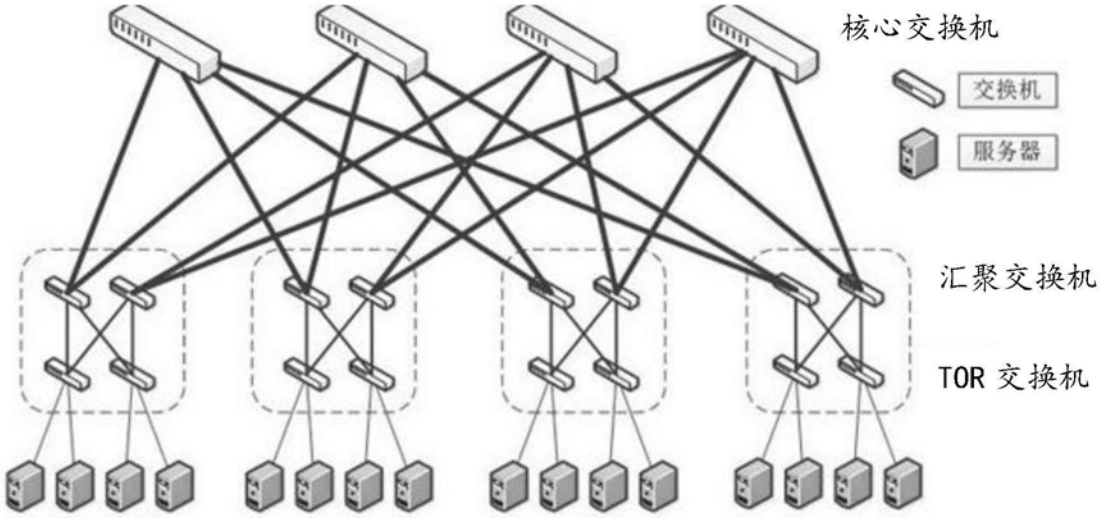

核心交换机
交换机
服务器
汇聚交换机
TOR 交换机

图1

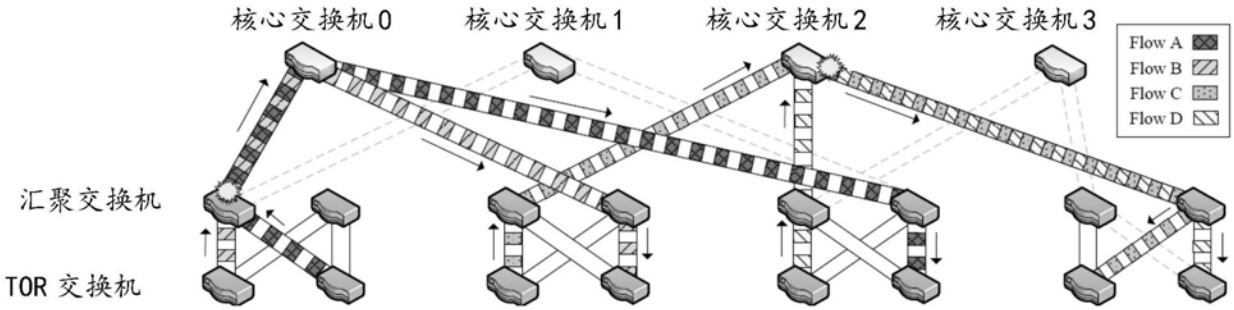

核心交换机0
核心交换机1
核心交换机2
核心交换机3
Flow A
Flow B
Flow C
Flow D
汇聚交换机
TOR 交换机

图2

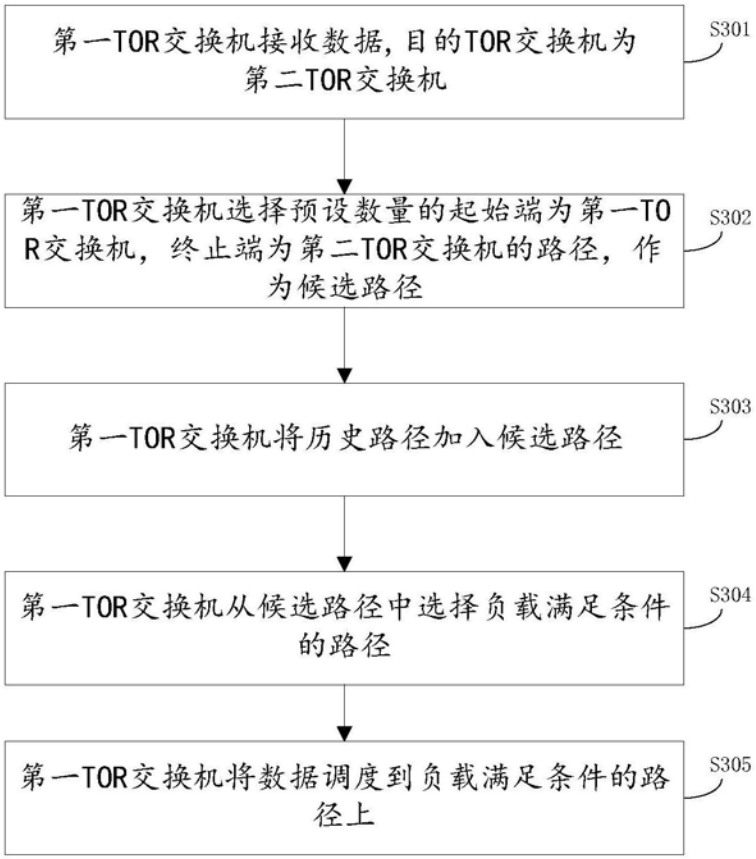

第一TOR交换机接收数据，目的TOR交换机为第二TOR交换机
S301
第一TOR交换机选择预设数量的起始端为第一TOR交换机，终止端为第二TOR交换机的路径，作为候选路径
S302
第一TOR交换机将历史路径加入候选路径
S303
第一TOR交换机从候选路径中选择负载满足条件的路径
S304
第一TOR交换机将数据调度到负载满足条件的路径上
S305

图3

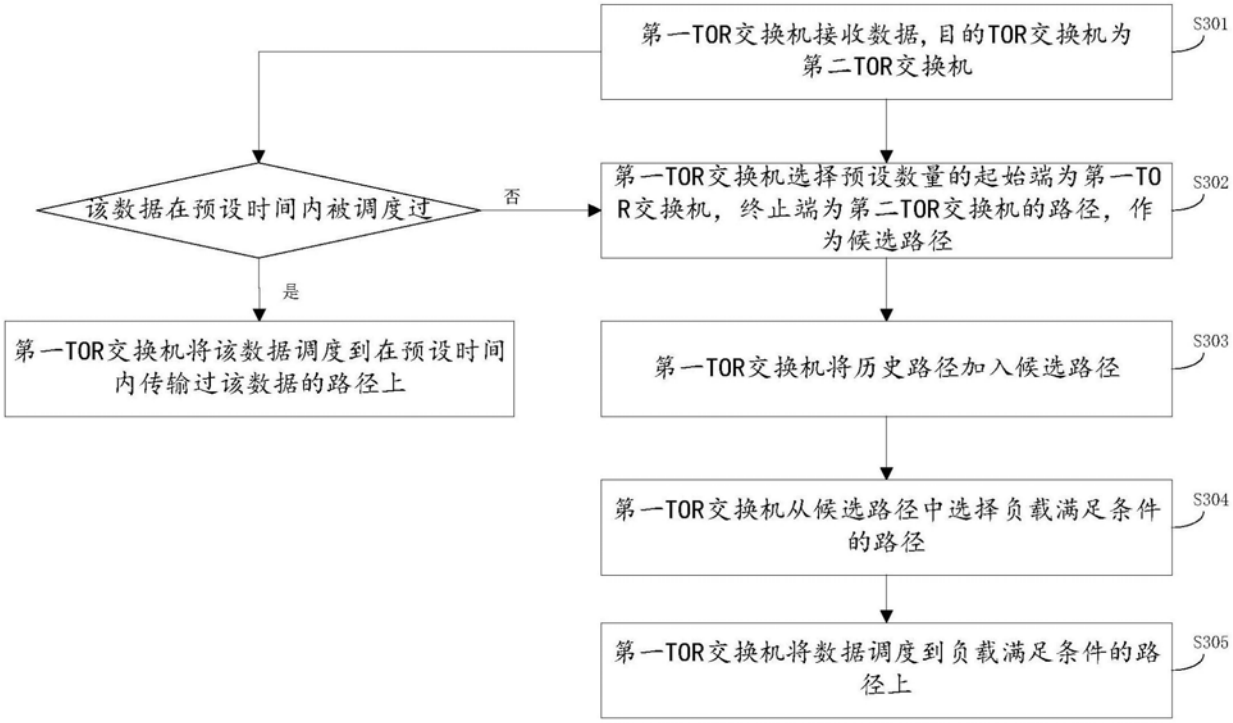

第一TOR交换机接收数据，目的TOR交换机为第二TOR交换机
S301
该数据在预设时间内被调度过
否
是
第一TOR交换机选择预设数量的起始端为第一TOR交换机，终止端为第二TOR交换机的路径，作为候选路径
S302
第一TOR交换机将该数据调度到在预设时间内传输过该数据的路径上
第一TOR交换机将历史路径加入候选路径
S303
第一TOR交换机从候选路径中选择负载满足条件的路径
S304
第一TOR交换机将数据调度到负载满足条件的路径上
S305

图4

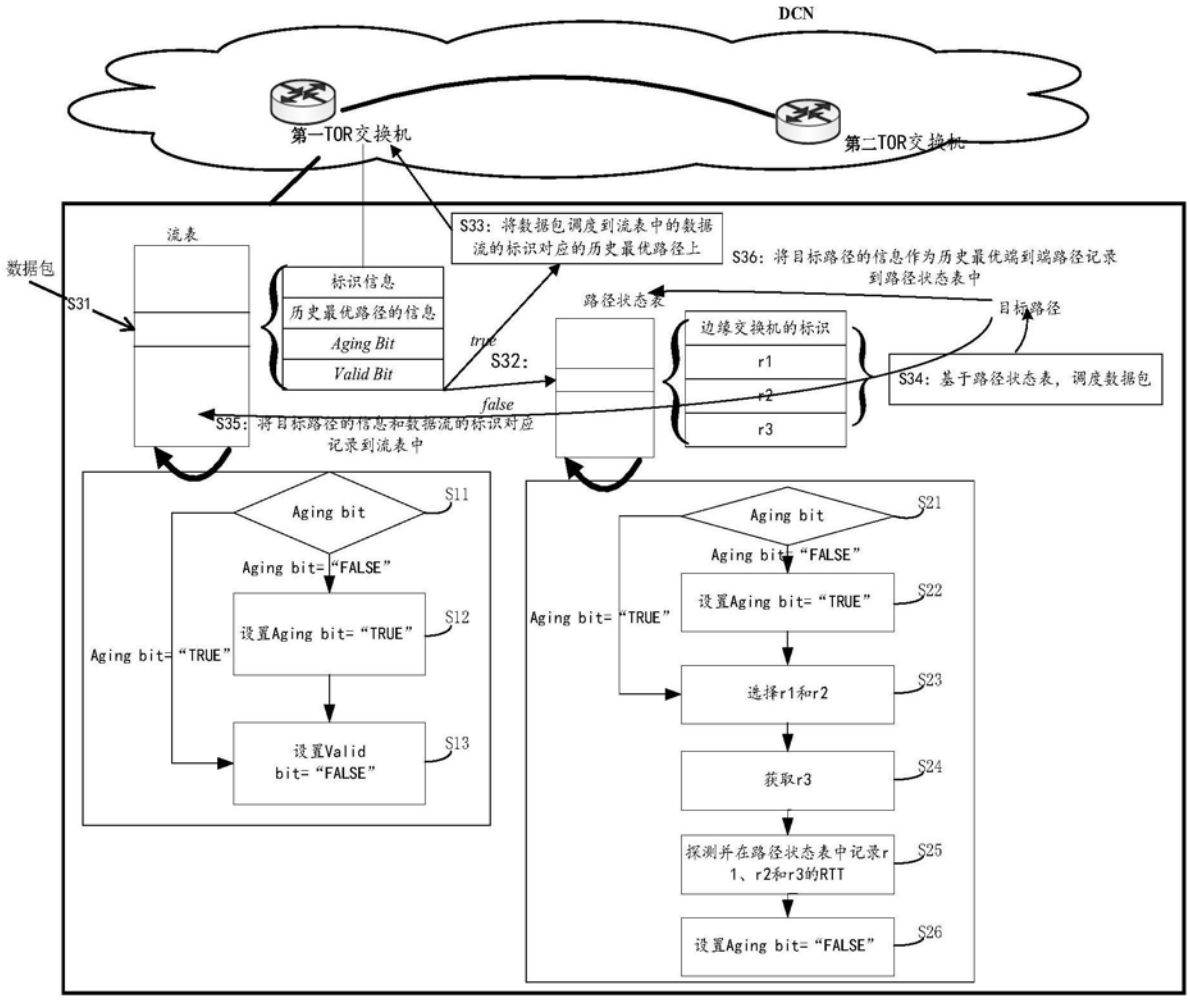

DCN
第一TOR交换机
第二TOR交换机
流表
数据包
S31
标识信息
历史最优路径的信息
Aging Bit
Valid Bit
S33：将数据包调度到流表中的数据流的标识对应的历史最优路径上
S36：将目标路径的信息作为历史最优端到端路径记录到路径状态表中
路径状态表
目标路径
true
S32:
false
边缘交换机的标识
r1
r2
r3
S34：基于路径状态表，调度数据包
S35：将目标路径的信息和数据流的标识对应记录到流表中
Aging bit
S11
Aging bit="FALSE"
设置Aging bit="TRUE"
S12
Aging bit="TRUE"
设置Valid bit="FALSE"
S13
Aging bit
S21
Aging bit="FALSE"
设置Aging bit="TRUE"
S22
Aging bit="TRUE"
选择r1和r2
S23
获取r3
S24
探测并在路径状态表中记录r1、r2和r3的RTT
S25
设置Aging bit="FALSE"
S26

图5

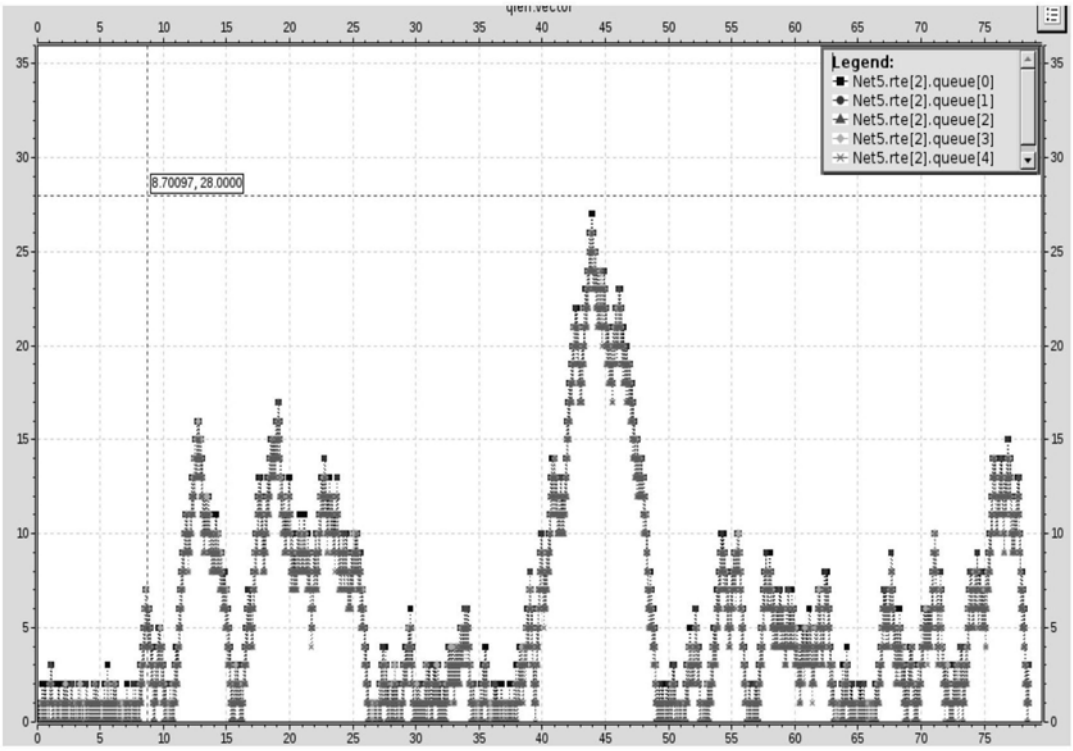

Legend:
Net5.rte[2].queue[0]
Net5.rte[2].queue[1]
Net5.rte[2].queue[2]
Net5.rte[2].queue[3]
Net5.rte[2].queue[4]
8.70097, 28.0000

图6(a)

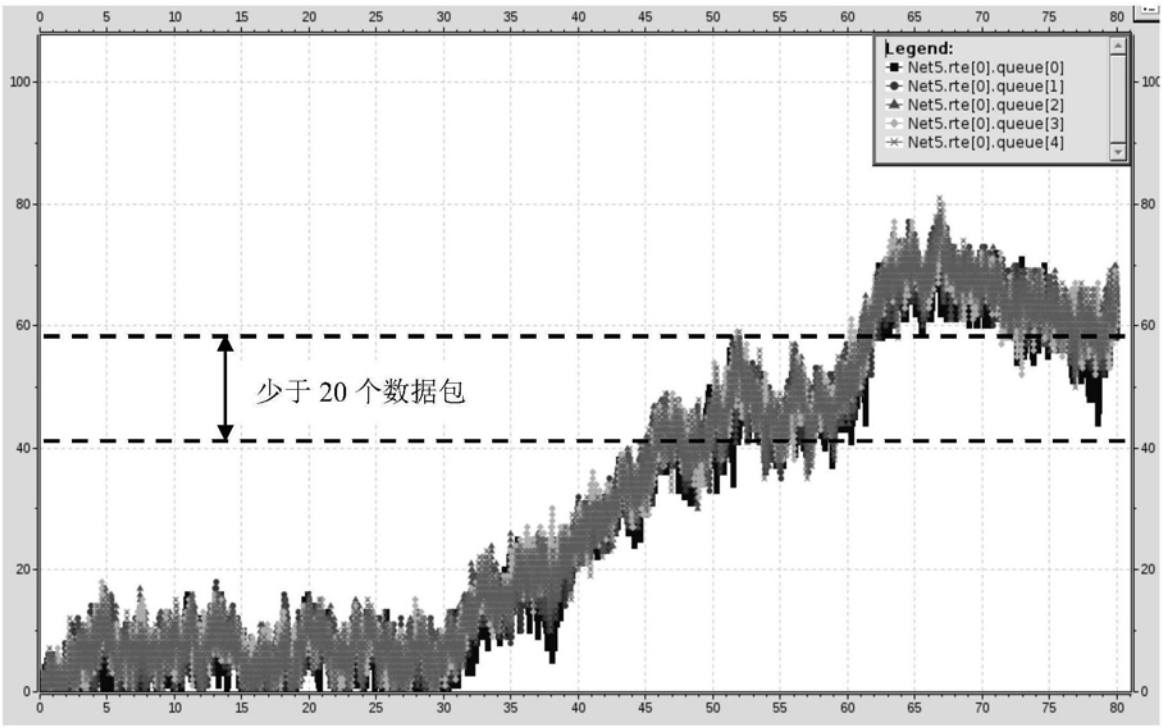

Legend:
Net5.rte[0].queue[0]
Net5.rte[0].queue[1]
Net5.rte[0].queue[2]
Net5.rte[0].queue[3]
Net5.rte[0].queue[4]
少于 20 个数据包

图6(b)

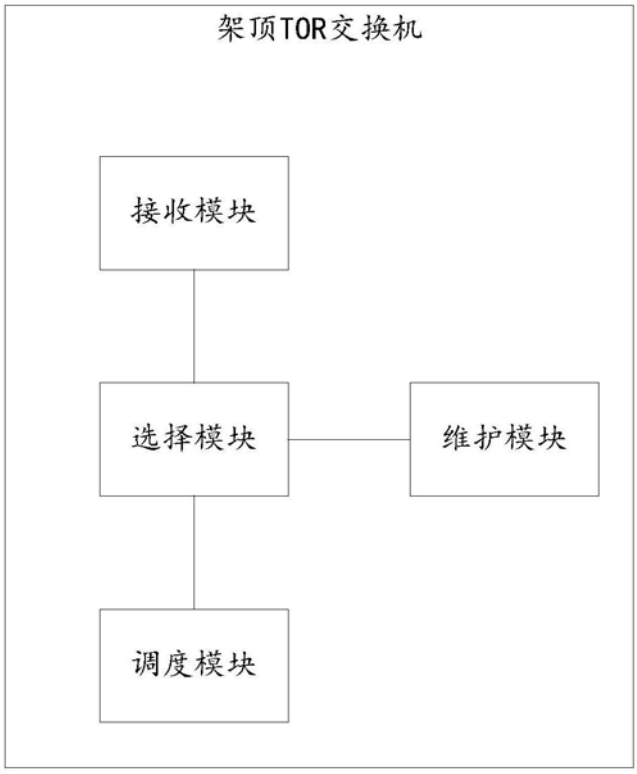


图7