CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 44 卷第 4 期 2022 年 4 月
Vol. 44, No. 4, Apr. 2022

文章编号:1007-130X(2022)04-0631-14

数据中心网络路由研究进展*

段　晨,彭　伟,王宝生
(国防科技大学计算机学院,湖南 长沙 410073)

摘　要:随着云计算的迅速发展,运营商对数据中心的需求与日俱增。作为数据中心网络的关键技术,路由负责在数据中心内部以及数据中心之间为流量选路,为不同服务质量要求的流量提供差异化的路由转发服务。当数据中心规模比较大时,由于应用不可预估的通信流量以及数据中心网络的拓扑特点,传统因特网路由方法不能提供令人满意的高吞吐率和资源利用率,网络拥塞使得无法为具有服务质量要求的数据流提供带宽或时延保证。首先对数据中心网络的路由问题进行分类分析,然后着重介绍了单播路由方法的研究进展,进一步对拥塞感知的路由方法进行了介绍,最后讨论了新型数据中心网络的路由技术。

关键词:数据中心网络;路由协议;拥塞控制;联合优化
中图分类号:TP393　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2022.04.007

Research progress of routing techniques for data center networks

DUAN Chen, PENG Wei, WANG Bao-sheng
(College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Due to the rapid development of cloud computing, the demand of Internet service providers for data centers is increasing rapidly. As a key technology of data center networks, routing is responsible for selecting routes for network traffic within or across data centers. It can also provide differential services for traffic with different QoS (Quality of Service) requirements. However, for large-scale data centers, due to the unpredictable traffic of applications and the topology characteristic of data centers, the traditional routing methods in the Internet cannot provide satisfactory high throughput and resource utilization. Network congestion make it hard to provide guaranteed bandwidth or transmission delay for flows with the demand of QoS. This paper classifies and compares the routing problems in data centers, and then emphatically introduces the research progress about the unicast routing methods. Further, the congestion-aware routing methods are introduced. Finally, the routing technology of the new data center networks is discussed.

Key words: data center network; routing protocol; congestion control; joint optimization

1　引言

路由是数据中心网络的关键技术之一,负责在数据中心内部以及跨数据中心为流量选路。数据中心由服务器和交换机 2 种设备组成,服务器主要运行各种云计算应用,交换机实现组网通信。数据中心网络通常使用传统 TCP/IP 协议栈进行通信,也有研究提出了一些不采用 TCP/IP 的专用协议,以更好地利用数据中心网络的拓扑结构和流量特

* **收稿日期**:2021-06-10;**修回日期**:2021-12-01
基金项目:国家重点研发计划(2018YFB0204301)
通信地址:410073 湖南省长沙市国防科技大学计算机学院
Address: College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, Hunan, P. R. China

点。

传统因特网中的路由主要考虑网络可达性,提供尽力而为的服务。而运行在数据中心的应用服务普遍具有更严格的低时延和高可靠性要求,因此数据中心网络路由需要为数据流提供更精确的选路策略。

数据中心网络路由研究中常常需要考虑以下4个方面的问题:(1)路由方法的可扩展性:能否适应数据中心网络规模的不断增长;(2)路由方法的部署成本:能否直接部署在现有的商用交换机上;(3)能否为不同服务质量要求的流量提供差异化的路由转发服务,以及能否实现自适应负载均衡;(4)路由方法的容错能力:能否从网络故障中快速恢复。当前,数据中心网络路由的研究大都集中在流量感知、多路径计算和负载均衡的流量分发等方面,目标是提高网络整体吞吐率或带宽利用率,缩短数据流的传输完成时间等。

在数据中心网络路由相关的综述中,Habib等[1]对数据中心网络路由技术进行了分类,但分类粒度较大,对已有研究工作的分析对比不够清晰,缺少对近年来最新研究成果的分析;Quttoum等[2]对数据中心互联拓扑结构、路由与流量工程进行了介绍,其中路由与流量工程部分侧重阐述了面临的问题和协议设计原则,但缺少对路由已有研究成果的详细梳理和分析。

本文对数据中心网络路由技术的当前研究进展进行综述。首先对路由问题和路由方法进行分类,然后重点阐述数据中心网络单播路由技术的研究进展。接着进一步阐述了拥塞感知的单播路由方法的研究成果,从多个维度进行分析对比,为数据中心网络拥塞控制与负载均衡问题提供了研究思路和方向。本文还关注了基于光交换的数据中心网络和有线无线混合的新型数据中心网络中的路由问题,为未来的研究工作提供了研究方向。

2 数据中心网络的路由模型

数据中心是支持各种网络应用、政企业务和科学计算的基础设施,内部网络设备规模巨大且通信流量复杂。面临的挑战之一是如何能够构建一个可扩展的数据中心网络,为应用提供大规模的聚合网络带宽,并在此基础上提供高效的负载均衡路由。

从路由的角度看,数据中心网络的主要目标是实现服务器之间大量的互联互通,且具有高效传输与容错机制。数据中心网络通信模型可以分为一对一、一对多和多对多3种,相应地数据中心网络路由也可以分为单播路由与组播路由;根据路由服务所在的层级,可以分为网络层路由与应用层路由。

2.1 单播路由 vs 组播路由

目前大多数研究都集中于一对一通信模型的单播路由,主要目标是利用多路径传输解决数据中心网络单播流量的负载均衡问题。存在的主要问题是面对大量的不可预估流量时,如何利用网络中存在的冗余路径进行合理的调度,满足应用服务的需求,减少网络拥塞,提升网络整体性能。数据中心网络的单播路由可以借用已有的因特网路由协议来完成,也可以采用新的技术途径,设计新的负载均衡路由协议。

常用的因特网单播路由协议有开放式最短路径优先OSPF(Open Shortest Path First)、中间系统到中间系统IS-IS(Intermediate System-to-Intermediate System)和边界网关协议BGP(Border Gateway Protocol)[3]等。OSPF和IS-IS是链路状态协议,当网络中有链路状态发生改变时,链路状态改变信息需要以泛洪方式传送给所有路由节点,洪泛的过程会消耗网络带宽资源,因此不适用于大规模的数据中心网络。并且OSPF和IS-IS在路由策略支持上不如BGP灵活。结合实际部署经验,OSPF通常被用于中小规模的数据中心网络,BGP则被用于大型数据中心网络。BGP设计的初衷是用于因特网自治系统AS(Autonomous System)之间的域间路由互通,并不直接适用于数据中心,因此Internet协会颁布的RFC 7938中提出了大规模数据中心中使用BGP路由协议需要考虑的问题以及相应的建议[4]。RFC 7938选择外部边界网关协议eBGP(extern Border Gateway Protocol)作为路由协议,主要原因是eBGP比内部网关协议iBGP(internal Border Gateway Protocol)更容易部署和管理。采用BGP实现组网,会面临BGP私有自治系统号不足的问题,因此建议使用4字节AS号或在数据中心网络核心层重用AS号。由于数据中心网络存在大量的点到点直连链路,在路由通告中宣告这些链路可能会带来大量的转发表FIB(Forwarding Information Database)和路由表RIB(Routing Information Database)开销。RFC 7938提出了2种解决方案,分别是不宣告链路地址和宣告链路地址并进行路由汇总。不宣告链路地址不会对网络设备通信产生影响,但运维和

监控会变得更复杂;宣告链路地址并进行路由汇总需要对网络地址进行合理规划,给路由器之间的链路地址分配连续可汇总的地址段。在数据中心网络中存在许多等价多路径 ECMP(Equal-Cost Multi-Path),数据中心 BGP 路由可以实现基于多 AS 号的 ECMP,允许 AS_PATH 长度相同但内容不同的等价多路径生效。针对 BGP 在因特网中会出现路由黑洞和收敛慢的问题,RFC 7938 指出如果支持 BGP 对等会话快速失效(BGP Fast Peering Session Deactivation),链路失效时 RIB 和 FIB 及时更新,就可以实现亚秒级的路由收敛。同时也可以通过配置 BGP 连接的保持时间(holdtime)、存活时间(keep-alive)和更新报文定时器等参数,实现路由快速收敛。

很多研究人员设计了新的数据中心单播路由方法,本文第 3 节将重点阐述这些方法,并对其进行比较分析。

组播路由将源节点发送的数据复制给多个加入组播组的节点,属于一对多通信模型。它存在的主要问题是如何构建并维护一个组播树,以实现高效的组播数据包分发。传统的 IP 组播需要运行复杂的组播路由协议,同时还要维护每条流的转发状态,由于其可扩展性低而没有得到大规模的使用。随着数据中心内部点到多点通信流量的快速增加,部署 IP 组播的吸引力在不断增加。随着软件定义网络 SDN(Software Defined Network)技术的发展,使用集中式控制器优化交换机和路由器转发组播流量的方法备受关注。由 Li 等[5]提出的数据中心高效可扩展组播路由机制 ESM(Efficient and Scalable Multicast routing scheme)将入包布隆过滤器(Bloom Filters)与交换机内转发表相结合。一方面,入包布隆过滤器用于小规模组播组,节约交换机内部路由表空间,同时将大的组播组路由条目下发给交换机,减少带宽开销。模拟实验结果表明,与接收端驱动的组播路由相比,ESM 可以减少 40%～50%的网络流量且应用吞吐量提高了 2 倍,入包布隆过滤器和交换机转发表的组合路由机制显著减少了交换机所需的路由条数。

2.2　网络层路由 vs 应用层路由

在大规模数据中心中,大多数的网络架构都是由企业网或因特网演化而来的,系统运行基于传统的网络层路由模型。网络层提供分组转发和路由选择功能。当分组从发送方流向接收方时,网络层为这些分组决定所采用的路径,其中计算路径的方法称为路由选择算法。在网络层路由模型中,网络管理员控制路由策略并下发配置,业务服务器与网络核心相互隔离。从数据中心运营商的角度看,网络层路由模型便于管理,并且能够在多租户场景中更好地满足客户需求,提高数据中心网络整体性能。然而从数据中心应用的角度看,网络内部是一个黑盒,应用服务无法控制数据包的转发决策。

应用层路由允许终端服务器参与到路由中,可以根据应用层服务的需求,定制路由协议或决定如何转发。英国微软研究院的 Abu-Libdeh 等[6]在 CamCube 项目中提出一种应用层实现路由的方法,应用可以根据自身需求在路由的不同目标之间进行不同的平衡,从而解决了单一路由协议优化目标单一的问题,提高了应用的路由性能。2014 年 Chen 等[7]提出一个将路由作为服务、租户可以直接定向路由控制的框架 RaaS(Routing-as-a-Service)。在多租户数据中心环境下,路由控制可看作是一个售票过程,租户向运营商提交路由控制请求,运营商为租户完成路由传输服务,租户与运行商之间紧密耦合。RaaS 将路由视作服务,只需要很少的管理开销,租户可以独立地进行路由控制,且可以直接利用现有技术运行在商用交换机上,不需要改变数据中心的网络硬件基础设施。

3　数据中心网络的路由方法

数据中心可以被定义为能提供大规模计算和多种网络服务的基础设施,随着因特网应用的不断增加,软件开发者和用户都需要更加高效的数据存储平台,因此人们对数据中心的性能需求也不断提高。

路由协议决定了路由器之间的通信过程,也决定了最优路由。因为数据流量可能流向数据中心内部、外部或交换机之间,因此路由协议必须能在所有服务器之间路由并转发数据,保证网络的连通性。为了有效利用数据中心基础设施,需要设计比传统因特网路由协议更加高效的路由协议,因此近年来提出了很多数据中心网络路由方法。

根据关注要素的不同,路由方法可以划分为拓扑结构相关的路由方法、流量相关的路由方法和能量感知的路由方法。拓扑结构相关的路由方法通常依赖于某种特定拓扑,在此基础上设计编址与路由转发方法,支持最短路径、非最短路径和多路径路由等。流量相关的路由方法主要关注数据中心内部流量调度,优化流量转发路径实现全局负载均衡,对拓扑结构不敏感。能量感知路由方法的主要

目标是在不影响性能的前提下降低数据中心能耗，主要面临2个问题：首先是如何筛选出冗余链路，其次是如何在动态拓扑的情境下为流量选路，因此能量感知的路由方法对效率和可靠性要求较高，需要在通信性能和节能之间达到合理的平衡。

3.1 评价指标

为了更直观方便地对不同路由方法进行比较，Besta等[8]提出了以下6个路由方法的评价指标，本文根据路由方法对交换机的要求添加了第7个指标。

(1)最短路径SP(Shortest Paths)的支持：是否可使用任意的最短路径。

(2)非最短路径的利用NP(Non-minimal Paths)：是否可利用非最短路径。

(3)多路径支持MP(Multi-Path)：是否可同时使用多条路径。

(4)不相交多路径DP(Disjoint Paths)：是否支持不相交的多路径。

(5)自适应负载平衡ALB(Adaptive Load Balancing)：是否具有自适应负载平衡的能力。

(6)拓扑相关性TD(Topology Dependency)：是否依赖于特定的网络拓扑。

(7)交换机可编程SPA(Switch Programm Ability)：是否要求数据中心交换机可编程，如采用白盒交换机。大多数商用交换机可以支持OpenFlow协议，因而可以利用交换机的可编程性来实现功能更强的路由控制方法。

3.2 拓扑结构相关的路由方法

借助已知的拓扑结构知识，优化路由转发的路由方法称为拓扑结构相关的路由方法，这类路由方法又分为以交换机为中心(Switch-Centric)的路由方法和以服务器为中心(Server-Centric)的路由方法。

当采用交换机来构建数据中心网络时，根据交换机的开放可编程情况，一般采用2类路由协议：一类是直接利用商用交换机提供的路由协议，如常见的OSPF和BGP路由协议；另一类要求交换机是开放可编程的白盒交换机，支持用户设计实现自己的路由协议。

3.2.1 以交换机为中心的路由

在传统的因特网中，网络终端只负责数据包的发送与接收，数据包转发都交给了网络核心的交换机与路由器完成。类似地，在数据中心网络中，在交换机上运行路由协议，通过交换机来构建数据中心网络，如典型的Clos结构网络，基于这种网络的路由方法称之为以交换机为中心的路由方法。

以交换机为中心的路由可以在已有拓扑的基础上，添加对虚拟机迁移前后不改变地址的支持或提供统一的高容量传输。如Greenberg等[9]提出的虚拟二层VL2(Virtual Layer 2)数据中心架构，由低开销的专用交换机组成Clos拓扑结构。这种网络架构可以提供所有服务器之间统一的高容量、同一台服务器上的服务之间性能隔离，还支持以太网二层语义，管理员可以给服务分配任意服务器并配置任意IP地址。VL2的一个限制在于它不能提供服务器之间的带宽保证，对于很多要求实时性的应用来说是不足的。Al-Fares等[10]提出了一个可扩展的商用数据中心网络架构，支持任意主机之间充分利用本地网络接口的带宽进行通信。该架构使用胖树(Fat-tree)拓扑，根据服务器所处的机架以及接入交换机的位置进行编址，设计了二级查表匹配的路由方法，一级表使用前缀匹配，二级表使用后缀匹配，在存在多路径的情况下，可以将相同目的地的连续数据包沿相同路径转发，避免数据包重排序。

Besta等[8]提出了一个简单、通用、健壮的以太网路由架构FatPaths，能充分利用数据中心的路径多样性，同时支持最短路径和非最短路径。通过形式化分析，实现了非常高的以太网性能提升，可以应用于超算系统和数据中心。FatPaths重新设计了传输层协议，改善了TCP启动慢等性能障碍，且使用流粒度交换，能避免TCP报文重排序，因此可以很简单高效地实现负载均衡。另外，作者使用了最多一百万个终端节点进行模拟，结果表明使用FatPaths的短直径拓扑比Fat-tree结构性能更好。

3.2.2 以服务器为中心的路由

以服务器为中心的路由方法在服务器上运行路由协议，目的是充分利用服务器的硬件资源，减小核心交换机的压力，因此服务器在多跳通信中起到了中继节点的作用；并且以服务器为中心的路由方法常常将拓扑设计为由递归或模块化的方式得到，具有很好的扩展性，如BCube[11]和Dcell[12]等。

BCube整个结构采用递归定义，结构虽然工整但是比较复杂。BCube的拓扑结构中一个BCubek是由n个$BCube_{k-1}$和n^k个n端口交换机组成。Guo等[11]针对BCube结构设计了BCube源路由算法，数据源服务器会持续发送探测报文探

测链路可用带宽,同时还使用宽度优先搜索算法进行替代路径查询,具有很好的容错性。

类似地,DCell 也以递归定义的方式来连接服务器[12],即高维的 DCell 由许多低维 DCell 递归连接而成,具有很好的容错性,且提供高了吞吐量。Wang 等[13]提出的 SprintNet 拓扑结构也是以服务器为中心的路由方法,它使用分割转发单元的方法提供可扩展的网络架构,且能够保证转发路径较短。

3.2.3　拓扑结构相关的路由方法小结

基于上述评价指标,表 1 对与上述拓扑结构相关的路由方法的研究工作进行了比较。从表 1 中可以得知,拓扑结构相关的路由方法通常依赖于某种特定拓扑,在此基础上设计编址与路由转发方法,从而实现比胖树(Fat-Tree)和叶脊(Spine-Leaf)架构更好的性能。然而,由于它们与底层路由机制相互独立,因此在实践中难以实现理论的高带宽,同时还需要对网络拓扑和网络协议栈进行修改或重新设计,增加了实际部署的难度。

Table 1　Comparison among topology-aware routing methods

表 1　拓扑感知的路由方法比较

路由方法	交换机为中心/服务器为中心	流量感知	SP	NP	MP	DP	ALB	TD
VL2	Switch	否	是	否	是	否	有限	高度依赖 Clos
可扩展商用交换机	Switch	否	是	否	是	是	是	Fat-Tree
BCube	Server	是	否	是	是	是	是	模块化
DCell	Server	否	否	是	否	否	否	仅适用递归模块化
SprintNet	Server	否	否	是	是	是	否	模块化
FatPaths	Switch	是	是	是	是	是	是	任意拓扑

3.3　流量相关的路由方法

流量相关的路由方法是指能够根据数据中心内部流量特点进行调度的方法,路由算法根据不同的优化目标做出满足服务需求的转发决策,总而言之以提升网络整体带宽利用率、实现负载均衡等为主要目标。Benson 等[14]指出,传统的流量相关的路由方法只能达到最优路由机制 80%～85%的性能。主要原因包括 3 点:(1)没有利用数据中心多路径特性;(2)无法自适应动态流量负载;(3)进行路由决策时没有掌握全局视图。本节将参考以上 3 个因素,对流量相关的路由方法进行分析与评估。

流量相关的路由方法可以从 2 个维度进行分类。首先,根据转发决策所掌握的信息和决策范围可以分为分布式路由方法和集中式路由方法。分布式路由方法使用局部或全局信息进行局部决策,集中式路由方法使用全局信息进行全局决策。其次,根据流量调度是否能提前避免拥塞事件的发生可以分为主动式流量调度和被动式流量调度。主动式流量调度能在网络故障或突发流量发生时,在网络拥塞导致丢包前,一定程度上避免网络拥塞,但灵活性不足;被动式流量调度通常依赖网络负载信息做出判断,具有较强的灵活性,但不可避免地会引入反馈时延。

3.3.1　分布式路由方法 vs 集中式路由方法

分布式路由方法中网络设备使用局部或全局信息进行局部决策。Wu 等[15]提出了数据中心网络分布式自适应路由框架 DARD(Distributed Adaptive Routing architecture for Datacenter networks)。该算法的目标是不对现有的基础设施进行修改或升级,且轻量可扩展,能充分利用数据中心对分带宽(Bisection Bandwidth),同时能在大流之间提供公平性。DARD 运行在终端主机上而不是交换机上,对网络中的动态流量负载敏感,可以将超载路径上的流量高效分担到空闲路径。DARD 的设计目标之一是解决大流之间的公平性问题,因此不可避免地忽略了小流的调度。在数据中心网络中,小流产生的流量占比小于 10%,流数量却占总流数量的 99%。每条小流都必须由交换机进行转发决策,因此对数据中心网络性能也产生了很大的影响。Li 等[16]提出了任务级性能优化专用调度器 OPTAS(OPtimize TAsk-level performance Scheduler),这是一个分布式小任务感知的流监控和调度系统,主要解决目前很多研究中小流被忽略或小流给集中式调度带来的开销过高的问题。OPTAS 的架构包括运行在数据接收端的任务监控器(Task Monitor)和决策计算模块(Policy Calculator),其中任务监视器用来识别任务的开始与结束;还包括运行在发送端的缓存监控器(Buffer Monitor)和决策执行模块(Policy Enforcer),可以根据缓冲区待发送数据大小区分大任务和小任务。OPTAS 会给小任务分配更高的优先级。实验结果表明,OPTAS 的调度速率比公平调度快 2.2 倍,比仅分配小任务最高优先级快 1.2 倍。

在数据中心网络中,网络设备由运营商统一管理,具有网络拓扑的全局视图,因此集中式路由方法被认为是可行的数据中心路由方法。集中式路

由方法将数据平面与控制平面分离,将 OpenFlow 协议视作网络设备数据平面的标准化编程接口。一些研究为了实现功能更强的路由控制,使用可编程的白盒交换机。

Hedera 是由 Al-Fares 等[17]提出的一个集中式可扩展、动态流调度器,可以有效利用数据中心网络中的聚合网络资源。它的主要目标是以最少的调度开销最大化对分带宽的利用率。Hedera 实现了路由和流量信息的全局视图,调度器从边缘交换机收集流信息来挑选无拥塞路径,然后交换机负责转发流量。作者提出了全局首次适配(Global First Fit)和模拟退火(Simulated Annealing)2 种启发式算法来解决流调度的 NP 完全问题。与 Hedera 相似,农黄武等[18]提出了基于 SDN 的胖树数据中心网络的多路径路由算法。针对 Hedera 提出的全局首次适配算法无法达到全局最优,而模拟退火法收敛速度可能很慢的问题,该算法使用基于深度优先查找的思路计算缓存网络中所有节点对的多路径信息,以提高响应速度。当新的数据流到来时,使用最差适应法确定备选路径,即在所有的路径中,挑选出瓶颈链路负载最小的路径,实现了快速有效的流量负载均衡。

另外,Hedera 忽略了数据中心网络流数量中占据大多数的小流,无法精确地实现负载均衡。杨洋等[19]提出了一个基于多路径传输的动态路由算法 Dramp。Dramp 对链路权值进行优化,在节点对之间的多条有效路径上合理地分配流量,确保关键链路不会成为产生拥塞的瓶颈链路。并且 Dramp 在完成细粒度流量均衡的同时,能很好地控制控制器的额外计算开销,也不需要对目前的通信协议进行任何改动。彭大芹等[20]提出了基于 SDN 的胖树数据中心网络多路径路由算法,将数据流分为大流和小流,大流根据路径权重值进行路由,小流数量多但处理的复杂性要求较低,选择可用剩余带宽最大的路径,这样能够提高胖树数据中心网络的平均链路利用率和网络吞吐量。

集中式控制方法面临单点失效、控制器负载过重等问题。如果使用冗余备份控制器,则会增加部署成本和控制器同步开销。为了改善如 Hedera 使用一个控制器掌握全局视图导致的负载过重问题,Tam 等[21]引入了下放控制器的概念,为集中式路由方法提供了新的思路。每个小控制器只负责局部的网络视图,但是所有的控制器可以覆盖整个网络,并提出了路径分区启发式算法和分区路径算法为流分配路径。Tam 等[21]在 Spring 等[22]测量得到的 ISP 拓扑结构上,对这 2 种算法进行评估,结果表明这 2 种算法都可以很好地限制控制器负载,虽然找到的不是全局最优解,但找到的解和模拟退火算法找到的解相近且速度更快。

也有研究对单个控制器进行优化,通过减少与控制器的通信过程,减小控制器负担。Ramos 等[23]提出了 SlickFlow 可靠的源路由方法,使用 OpenFlow 协议实现,把源路由与备用路径信息搭载在数据包头部,实现支持快速故障恢复的弹性源路由机制。当首选路径出现故障时,数据包可以通过交换机重路由到备用路径而不需要询问控制器,但需要交换机支持对数据包备用路径的解析。Wang 等[24]提出了一个低开销的、负载均衡路由管理框架 L2RM(Low-cost, Load-balanced Route Management framework),通过持续监控网络流量并计算负载偏差参数来检验是否有交换机链路负载过重,如果有则触发路由修改机制,将流分发到不同的链路实现平衡负载。与此同时还采用轮询的方式查询交换机的状态,减少控制器的开销。

以上 2 种方式都是尝试减少与控制器的通信,提高调度实时性,但必然会面临可扩展性的问题,即随着网络规模的增大,要保证在可接受的时间内完成路由优化。陈松等[25]提出了一种新颖的面向数据中心网络流量的路由优化算法,其主要优化对象是具有语义相关性的 Coflow 流(指数据中心中传输的具有相关性的若干个流),并提出了面向海量流采用 GPU 并行计算的路由优化算法。相比于 CPU,GPU 可以支持数百倍数量的线程。算法还使用了拉格朗日松弛方法分解路由优化问题,利用 Bellman-Ford 最短路径算法为每个业务计算路由,结果表明该算法可以提高流量调度的实时性。

3.3.2 主动式流量调度 vs 被动式流量调度

路由方法根据流量调度是否能提前避免拥塞事件的发生,可以分为主动式流量调度和被动式流量调度[26]。主动式流量调度主动进行流量调度,从而达到避免拥塞发生的目的;被动式流量调度是在拥塞事件发生后,再相应地进行流量调度,以缓解网络拥塞。

ECMP 是最常见的主动式流量调度方法,它是一个具有主动特性且无状态调度的负载均衡方法。通过对流的头部字段进行哈希运算,根据哈希值决定流的传输路径。ECMP 的缺陷在于难以适应非对称拓扑和大小流混合的场景。尽管如此,研究表明[27],随着流的数量增加以及流的大小分布的方差减小,ECMP 的缺陷可以得到缓解并获得

近似最优解。He 等[26]提出了一个基于软件边缘的主动式流量调度方案 Presto,利用边缘虚拟交换机(Virtual Switch)将每条流整形成相似大小的子流,并且将它们均匀地分发到网络中,从而实现负载均衡。Presto 可以提升网络吞吐率和流调度公平性,降低网络延迟,同时能够缩短小流的流完成时间。

被动式流量调度是当调度器或交换机感知到网络拥塞事件后,再进行流量调度。如果突发流量的持续时间很短,被动式流量调度则难以发挥很好的作用。很多集中式流量调度机制处理网络拥塞时都是采用被动方式,并且由于它们的控制环路存在时间延迟,因此只能实现粗粒度优化,如上文提到的 Hedera[17]和 Dramp[19]。

3.3.3　流量相关的路由方法小结

流量相关路由方法主要关注数据中心内部流量调度,针对不同的流量模型设计调度策略。上述流量相关的路由方法都不依赖于特定网络拓扑,也不保证对所有数据流都支持使用非最短路径或不相交路径。因此,表 2 主要关注路由方法对交换机的可编程要求(SPA),并使用是否支持多路径(MP)、自适应负载均衡(ALB)和拓扑依赖性(TD)3 个指标进行对比。

Table 2　Comparison among traffic-aware routing methods

表 2　流量相关的路由方法比较

路由方法	SPA	MP	ALB	TD
DARD	否	是	是	无
OPTAS	否	否	是	无
Hedera	否	是	是	无
文献[18]方法	否	是	是	无
Dramp	否	是	是	无
SlickFlow	是	否	是	无
L2RM	否	否	是	无
文献[25]方法	否	是	是	无
Presto	是	是	否	无

3.4　能量感知的路由方法

大型数据中心全天候运行数十万台网络设备,为了节约能量,可以让不参与通信的设备在不影响网络性能的前提下进入休眠状态,因此需要有高效的能量感知路由方法。目前有很多研究对数据中心网络能量感知路由方法进行探索,由于该优化问题属于 NP 难问题,因此大多数研究都选择启发式算法求解该问题。

Shang 等[28]提出了数据中心网络绿色路由,尝试从路由的角度减少数据中心能源消耗。该方法建立了能量感知的路由模型,证明了问题的 NP 难性质并设计了一个启发式算法,将网络吞吐量视为性能的度量值。该算法首先在所有交换机上计算最短路由,称之为基本路由。然后逐一地将流量负载最低的交换机设置为睡眠状态,直到网络吞吐量下降到某个阈值,达到了网络管理者对网络吞吐量的最低要求,算法结束,并更新拓扑结构,移除处于睡眠状态的交换机。类似地,Xu 等[29]提出了具有吞吐量保证的能耗感知路由算法,主要目标是减少密集连接数据中心中网络设备的能耗。该启发式算法首先将拓扑结构、流量矩阵和吞吐量阈值等输入路由生成(Route Generation)模块,为流量矩阵中的每条流计算一条传输路径;然后经过吞吐量计算(Throughput Computation)模块得到网络吞吐量,再判断是否需要移除交换机或链路;最后输出新的网络拓扑。同时,为了提高可靠性,用户可以在可靠性适应(Reliability Adaptation)模块中通过配置可靠性参数确定为每条流保留备选路径的数量。如果由于网络故障等原因,导致之前移除交换机和链路后产生的路径不可达,流可以切换到备选路径中。

这 2 种能量感知路由协议都需要预先知道流量矩阵。如果流量矩阵不准确,可能会出现网络无连接的问题。何荣希等[30]提出了软件定义数据中心网络多约束节能路由算法 MER(Multi-constrained Energy-saving Routing),考虑网络连通性和时延等因素,选择代价最小的路径传输数据流。

由于能量感知路由方法的目的是降低数据中心能耗,而 3.1 节中的 7 个指标主要关注路由方法的功能,因此表 3 从能量感知路由方法的目标和主要思路 2 方面进行比较。

Table 3　Comparison among energy-aware routing methods

表 3　能量感知的路由方法比较

路由方法	目标	主要思路
绿色路由	在不影响网络性能的前提下,以尽可能少的网络设备提供路由服务	为能量感知路由问题建模并提出启发式路由算法,移除冗余交换机
能耗感知路由	以最小化网络能耗提供路由服务	提供吞吐量保证路由,移除冗余交换机和冗余链路
MER	在保证时延和可靠性的前提下,尽可能多地休眠冗余交换机和链路,以降低网络能耗	提出等效节点、最小网络连通子集、孤岛交换机和无效链路等概念以及辅助图模型和软件定义数据中心网络连通条件,给出多约束节能路由优化模型

根据对上述能量感知路由方法的相关研究比较分析,可以得出能量感知的路由方法的本质是在变化拓扑下的流量调度问题。与流量相关的路由

方法的不同之处在于,能量感知的路由方法需要根据网络流量状况休眠部分网络设备,因此关键问题是如何筛选出网络中存在的冗余设备,而流量调度问题则可以交给流量相关的路由方法负责解决。

4 拥塞感知的路由方法

数据中心网络时常发生网络拥塞事件,传统的TCP/IP协议栈使用TCP协议实现拥塞控制,而网络层路由对网络拥塞状态是无感知的。因此,TCP拥塞控制机制无法充分利用数据中心网络的冗余传输路径,且对时延敏感的流量不友好。一些研究工作提出了适用于数据中心的传输层协议,根据拥塞控制的位置分为发送端拥塞控制和接收端拥塞控制。发送端拥塞控制是指发送端根据拥塞状况调节发送速率,比较典型的有数据中心传输控制协议DCTCP(Data Center Transmission Control Protocol)[31]、流截止时间感知的数据中心传输控制协议D2TCP(Deadline-aware Datacenter Transmission Control Protocol)[32]、基于延迟测量的传输层协议TIMELY(Transport Informed by MEasurement of LatencY)[33]和数据中心量化拥塞通知DCQCN(Data Center Quantized Congestion Notification)[34]。接收端拥塞控制是指接收端根据拥塞状况来调节控制发送端的发送速率,如Homa[35]和ExpressPass[36]。DCTCP、D2TCP和DCQCN借助拥塞信号完成闭环拥塞控制,TIMELY通过往返时间RTT(Round Trip Time)调节发送速率,但是由于一个RTT内数据包往返转发路径可能不同,因此无法精确感知拥塞状况。在Homa和ExpressPass中,接收端为发送端发送数据授权并指派优先级,从而控制发送端的发送速率。以上拥塞控制协议都发生在终端主机的传输层,网络层路由器的路由转发决策对传输层而言是透明的,无法充分利用数据中心网络的多路径特性。因此,路由协议能够根据网络拥塞状态为流量选择不同的路径,成为提高数据中心网络通信性能的重要途径,这样的路由方法称为拥塞感知的路由方法。

根据现有研究,拥塞感知的路由方法可以分为集中式拥塞感知路由与分布式拥塞感知路由。集中式拥塞感知路由通常使用集中控制器获取全局拥塞信息,为流分配传输路径并将转发表项下发到交换机上。分布式拥塞感知路由通常使用全局或局部拥塞信息在网络局部实现优化,局部优化的决策设备分为网络边缘设备和网络交换设备。

4.1 集中式拥塞感知路由

Kanagevlu等[37]提出了一个针对大小流混合的边缘到边缘重路由机制,当发生拥塞时,考虑到小流存在周期短,重路由会增加它们的延迟和开销,因此只将大流重路由到备用路径。Kanagevlu等[38]还提出了一个局部重路由机制,基于SDN技术在数据中心网络中高效管理链路拥塞或故障。不同于其它研究中通过通知源节点进行适当的速率调整,以及在源节点进行重路由来处理拥塞,局部重路由机制是在拥塞发生点或拥塞发生点的上一跳,根据流分类机制把数据流重路由到其它可能的路径。类似地,Fastpass也是一个集中式控制系统[39],其目的是能够实现交换机零队列(或非常浅的队列)。它使用集中式控制器决定每个数据包的传输时间和传输路径,因此该系统的关键在于时间片分配算法和路径分配算法。控制器需要快速响应用户发送数据的请求。同时控制协议要非常稳定,否则数据没有得到分配的时间片,无法发送数据,因此Fastpass在实现上受到了可扩展性与可靠性的限制。与上述2种方法不同,Vissicchio等[40]提出的Fibbing是一个在分布式路由之上使用集中式控制的联合优化架构。Fibbing不需要可编程交换机,可以直接部署在商用交换机上,集中控制器会生成一个增强型拓扑,其中包含一些实际不存在的节点和链路。例如,这些实际不存在的链路可以组成一条新的最短路径,且第一跳与实际拓扑中的非最短路径重叠,则集中控制器下发伪造的最短路径后,在实际转发过程中流量会沿着实际拓扑中的非最短路径转发,从而达到网络路由优化的目的。生成增强型拓扑的关键在于能够避免环路和路由黑洞,且生成的拓扑要尽可能小。如果给定希望得到的转发条目和路由协议,Fibbing会计算出应当通告给路由器的消息,可以实现流量工程、负载均衡和故障恢复。

以上3种集中式拥塞感知路由都是使用SDN对数据流的传输进行集中式控制,以此减少或避免拥塞,但存在集中式方法的单点失效等问题。

4.2 分布式拥塞感知路由

CONGA(CONGestion-Aware load balancing scheme)是一个分布式拥塞感知的路由方法[41],使用全局拥塞信息以流片(flowlet)粒度决策,比流粒度调度的方式更精确。一条流包含了一系列的突发流量,当一条流中2个数据包到达的时间间隔超

过某一阈值,则将之前的数据包划分为一个流片[42]。CONGA将拥塞控制的功能从网络核心设备卸载到网络边缘的接入交换机上,数据包会携带该条路径上的最大负载并抵达目的交换机,目的交换机将负载信息搭载在返回的数据包中,从而实现一个拥塞感知的反馈机制。CONGA在二层叶脊拓扑中可以达到很好的效果,但在复杂拓扑中,因为一条路径由多条链路组成,降低了CONGA的拥塞控制效果。与CONGA类似,逐跳带宽利用率感知的负载均衡框架HULA(Hop-by-hop Utilization-aware Load balancing Architecture)也是一个使用全局信息实现拥塞控制的路由机制[43]。但是,HULA以逐跳的方式实现反馈,每一个交换机节点会周期性地向其它节点发送自身负载信息。这引入了不必要的探测数据包开销,尤其是当网络处于拥塞状态时,探测数据包不仅无法抵达目的地,而且还可能会加剧拥塞程度。Fan等[44]提出了一个基于全局信息决策的分布式数据驱动的拥塞反馈机制,对CONGA和HULA进行改进,可以在大规模多级胖树拓扑中实现很好的性能。它将网络划分为多个独立的路由域,不同路由域之间的数据包经过核心交换机时会携带网络接口的负载信息,因此路由域具有网络的全局负载信息。路由域内的交换机负责维护域内和域外负载矩阵,以流片粒度为数据包选路。一般情况下路由算法若要掌握全局信息需要巨大的控制报文开销,但Fan等提出的机制将负载信息搭载在数据包中,大大减少了控制报文开销。

对于基于全局拥塞信息进行优化的方法,研究认为其控制环路所消耗的时间可能大于大多数拥塞事件的持续时间,因此无法对网络拥塞做出及时的响应。

Zhang等[45]提出了分布式拥塞感知自适应转发方法CAAR(Congestion-Aware Adaptive foRwarding),交换机仅依赖于局部信息(邻居节点的队列长度)就可以实现流量负载均衡,其中心思想是选择链路带宽未被充分使用的路径进行数据包转发。如果所选择的路径在传输过程中发生拥塞,流可以重定向到另一条未充分利用的路径。模拟实验显示,CAAR可以在多种流量模型下自适应流量变化,相比于ECMP可以很好地提高冗余带宽利用率,但它需要交换机可编程,具有一定的部署难度。Ghorbani等[46]提出了一个针对Clos网络拓扑的微秒时间尺度的网络内部分布式随机本地负载均衡DRILL(Distributed Randomized In-network Localized Load balancing)机制。每个数据包仅根据局部交换机队列占用情况和随机算法分发负载,就能够快速对拥塞做出反应。DRILL使用了类似power of two choices[47]的调度算法,选路时会随机选取2个可用的出接口,并与上一次决策时选择的最低负载出接口的队列占用情况进行对比,选择3个接口中负载最小的出接口发送数据包。DRILL还对常常面临的数据包重排序与非对称拓扑问题进行研究并提出了解决方案。与CONGA相比,DRILL可以在队列长度开始增加时就迅速调整流量分发。在80%的流量负载下,DRILL可以实现99%的流完成时间比CONGA的快1.4倍。Huang等[48]提出了队列延迟感知的数据包分发负载均衡机制QDAPS(Queueing Delay Aware Packet Spraying for load balancing)。QDAPS以数据包粒度进行转发,针对数据包粒度转发面临的共同问题——非对称拓扑会引起数据包重排序,QDAPS根据这条流的上一个数据包在队列中的排队延迟决定输出端口,让先到达的数据包比后到达的数据包先发送,从而解决该问题。文中还指出以流粒度、流片粒度转发虽然可以避免数据包重排序问题,但是灵活性不足,会降低链路利用率。

也有一些研究立足于在尽可能不改变交换机硬件的前提下实现分布式拥塞感知的负载均衡路由。Kabbani等[49]提出了一个流级别的拥塞感知路由方法Flicr。Flicr利用ECN感知拥塞,当拥塞发生时,主机修改流的VLAN字段实现重路由。目前商用交换机都支持VLAN功能,因此Flicr不需要改变交换机硬件,只需要修改主机内核、并对交换机进行VLAN相关配置,就可以快速完成部署。Flicr不仅降低了链路故障导致的流恢复时间,而且还适用于非对称网络拓扑。

4.3　拥塞感知的路由方法小结

表4对拥塞感知的路由方法进行比较,主要分析了集中式/分布式、拥塞控制设备、拥塞控制使用的拥塞感知信息和拥塞控制粒度4个特征。总的来说,集中式拥塞感知路由方法普遍掌握全局拥塞信息视图,为交换机下发配置,需要解决的关键问题是实时性和可扩展性问题;而分布式拥塞感知路由方法更加灵活,且获取拥塞信息的方法简单快速,需要解决的关键问题是局部拥塞信息带来的局限性。另外,不同路由方法选取的拥塞控制粒度不同,数据包粒度实现简单,但需要解决数据包重排

序问题;流相关粒度不会面临数据包重排序问题,但需要解决的关键问题是流表规模、查表速度以及在大小流混合的流量模型中如何均匀转发大小流实现拥塞避免。

Table 4 Comparison among congestion-aware routing methods

表 4 拥塞感知路由方法比较

路由方法	集中式/分布式	拥塞控制设备	拥塞感知信息视图	拥塞控制粒度
文献[38]方法	集中式	交换机	全局	flow
Fastpass	集中式	交换机	全局	packet
Fibbing	集中式	交换机	全局	packet
CONGA	分布式	接入交换机/主机	全局	flowlet
HULA	分布式	交换机	全局	flowlet
文献[44]方法	分布式	交换机	全局	flowlet
CAAR	分布式	交换机	局部	flow
DRILL	分布式	交换机	局部	packet
QDAPS	分布式	交换机	局部	packet
Flicr	分布式	主机	全局	flow

5 新型数据中心网络的路由技术

5.1 光交换的数据中心网络

云计算和 Web 服务要求数据中心能提供非常高的通信带宽。相比由商用以太网交换机构成的数据中心网络,光交换网络可以提供更高的吞吐量、减少能量消耗且具有可重构特点,是未来数据中心网络的发展方向。目前已有 DETOUR[50]、c-Through[51]、Helios[52]和 MegaSwitch[53]等光交换数据中心网络,它们使用高容量的光纤和光交换机互联,可以实现光电路交换、光分组交换或光突发交换。光分组交换和光突发交换由于带宽资源粒度较小,可以满足突发性较大的业务流量的交换[54]。

传统交换机存在入接口和出接口缓冲区,负责暂存数据包。对于一个数据包而言,最优的情况是当它经过交换机时,入接口缓冲区和出接口缓冲区都为空,可以直接为它提供服务,如 Fastpass[39]通过集中式调度为所有数据包分配发送时间片和传输路径,避免数据包排队甚至拥塞。然而由于缺乏全网同步机制,Fastpass 也需要很小的缓冲区。在全光或混合数据中心网络,光纤会引入特定的传播延迟,但只要交换机在延迟时间内能够完成数据包选路,就不需要交换机接口缓冲区[52]。因此,光数据中心网络的路由方法需要在非常短的时间内实现选路并完成重构。

传统交换机基于数据包头部内容或数据包元数据为其选路。OpenFlow 协议作为 SDN 网络设备数据平面的标准化编程接口,使用集中控制器为流计算路由并将流表项下发到转发表,并可以基于更多字段为数据包选路。光数据包交换 OPS(Optical Packet Switching)查表可以在有限的时间内完成标签提取和重构操作,光纤延迟线引入的延迟起到了缓存数据包的作用,等待 OPS 查表完成后转发数据包。这种方法非常适用于光数据中心网络。另一种光数据中心网络常用的选路方法是波长查表,即将接收到的具有特定波长的光信号转发到特定端口,因此波长选路不需要提取任何标签字段,直接使用波长实现路由决策,但前提是需要预先建立波长电路[55]。与电路交换类似,波长选路的优势在于选路迅速且选路过程中不需要任何内存和处理资源。Wang 等[56]提出全光交换的数据中心网络端到端调度方法,其主要目标是解决网络内部零缓冲和不可忽视的重构时延,并提出了 2 种具有重构时延的交换机调度算法。

目前光数据包交换查表的路由方法如果出现查表时间超出时间界限,则数据包会被丢弃并重传。未来,光数据包交换查表的路由方法可以考虑的研究方向应当是有上下界决策时间保证的路由方法,能够形式化证明路由方法的决策时间上下界,从而为光数据中心网络提供更高的可靠性。另外,波长查表的路由方法可以充分利用物理特性进行转发,但需要集中控制器参与,类似于 OpenFlow 的方式,为不同波长的光信号分配时间片和路径并下发给交换机。未来,研究波长查表的路由方法可以根据集中式调度引入的额外时延开销,考虑分别对不同阶段所引入的时延进行优化,如发送方上报时延、集中式算法运行时延等,从而降低集中式控制对实时性的影响。

5.2 有线无线混合的数据中心网络

有线数据中心网络面临布线复杂与热点链路 2 个不可避免的问题,导致了很高的部署和运维成本,并且热点链路会产生拥塞事件以致影响全局性能。因此,一些研究工作建议将无线通信技术应用于数据中心网络。其设计的初衷是在不降低链路传输速率的前提下,服务器之间采用无线连接减轻核心交换机负载,降低布线复杂度和运维成本,方便扩展数据中心网络。

目前无线网络的特点是:高频段速率快且稳定,但容易被遮挡和干扰;低频段不易受干扰但速

率慢;另外无线设备的天线方向也对传输性能影响比较大。因此,将无线通信应用于数据中心还面临以下难点:首先必须提供高速率通信技术,60 GHz技术的发展使得高速无线通信技术应用于数据中心成为可能,Kandula 等[57]提出在架顶 ToR(Top of Rack)交换机之间添加 60 GHz 的无线链路来缓解局部热点的拥塞问题。其次为了保证各机柜之间快速建立高速链接,必须优化数据中心网络拓扑结构,传统基于行的数据中心机架排布方式在一定程度上阻碍了机架间建立无线连接[58]。第三还需要设计合理的资源分配机制,为数据流分配互不干扰的信道。清华大学 Cui 等[59]对无线数据中心网络面临的实际部署问题进行了研究,文中将信道分配问题形式化为一个最大化无线传输总利用率的优化问题,为此设计并实现了一个基于 Hungarian 算法[60]的启发式算法。针对以上问题目前已经有许多研究成果,为无线数据中心网络设计提供了参考。

为解决无线数据中心网络中的负载均衡问题,Celik 等[61]提出了应用于光无线数据中心网络的流量聚合路由方法。它只对小流采用三步流聚合 3SFG(Three-Step Flow Grooming)的方法处理。3SFG 方法发生在源服务器节点与交换机上,分为以下 3 个步骤:(1)服务器到服务器 S2S(Server-to-Server):源服务器节点聚合所有去往相同目的服务器的小流,生成 S2S 数据流;(2)服务器到机架 S2R(Server-to-Rack):源服务器聚合所有去往相同机架交换机的 S2S 数据流,并转发给机架交换机;(3)机架到机架 R2R(Rack-to-Rack):机架交换机将收到的 S2S 数据流根据目的机架不同,聚合成多条 R2R 数据流,然后使用 R2R 光无线链路传输数据流。3SFG 方法会根据机架间的长期流量统计数据决定 R2R 数据流的路径容量和路由。由于大流已经有明确的流量需求如流的大小、持续时间等,因此当检测到大流后,不需要流聚合,机架交换机会建立从源到目的的光无线链路,当大流传输完成后终止链路连接。由于 3SFG 方法使用长期流量统计信息,因此被称为长期负载均衡,无法在具有突发流或流量频繁变化的流量模型中实现负载均衡。AlGhadhban 等[62]提出了一个灵活的短期负载均衡机制 SoftFG(Soft-reconfigurations and Flow Grooming)。SoftFG 能够应对变化的流量模型,可以在无线数据中心网络中,对不同类别数据流对应的多个虚拟拓扑实现负载均衡。SoftFG 可以在不对路径容量进行硬件重构的前提下,将拥塞路径上的大流重路由到未完全利用的链路上。考虑到重路由引起的重排序开销对小流影响较大,因此 SoftFG 不对小流进行处理。SoftFG 作为内核模块被安装在虚拟交换机中,它可以基于源和目的之间的协作机制收集数据流的统计信息,同时结合主动探测技术,确保对网络路径的实时监测、早期拥塞感知和快速准确的重路由。网络模拟显示,在部署 SoftFG 的无线数据中心网络中,数据流完成时间比部署 LetFlow(Let the flowlets Flow)[63]和 CONGA[41]分别快了 12 倍和 17 倍。

未来,针对有线无线混合的数据中心,应当提出更多能充分利用无线通信技术特点的路由方法。其次由于无线网络也存在多路径,且链路状况更加复杂,SoftFG 没有解决数据包重排序问题,未来可以借鉴文献[10,48]中解决重排序问题的方法实现更好的路由。最后,由于无线网络通信不需要繁琐的布线且容易改变位置和天线方向,因此可以以绿色节能为优化目标,借鉴文献[28-30]中的方法,实现有线无线混合数据中心能量感知的路由方法。

6　结束语

本文综述了数据中心网络路由技术的研究发展现状,通过对数据中心网络路由模型进行分类比较,讨论了近年来数据中心网络路由方法的相关研究。本文重点关注数据中心网络单播路由方法,分别从拓扑结构相关的路由方法、流量相关的路由方法和能量感知的路由方法 3 个角度,列举分析了现有的研究工作,并使用合理的评价指标对其进行比较。并进一步介绍了拥塞感知的负载均衡路由方法,通过对现有研究工作的整理分析,比较展示了拥塞感知路由方法的一些重要研究成果,为未来的研究提供了思路。最后,介绍了新型数据中心网络——基于光交换的数据中心网络和有线无线混合的数据中心网络,对 2 种数据中心网络路由技术的难点和技术途径进行了分析,阐述了对未来该领域研究方向的一些认识与理解。

参考文献:

[1] Habib S, Bokhari F S, Khan S U. Routing techniques in data center networks [M] // Handbook on Data Centers. New York: Springer, 2015: 507-532.

[2] Quttoum A N. Interconnection structures, management and routing challenges in cloud-service data center networks: A survey[J]. International Journal of Interactive Mobile Technologies (iJIM), 2018, 12(1): 36-60.

[3] Rekhter Y, Li T, Hares S. A border gateway protocol 4 (BGP-4):RFC 4271[S]. New York: The Internet Society, 2006.

[4] Lapukhov P,Facebook,A. Premji,et al. Use of BGP for routing in large-scale data centers:RFC 7938[S]. New York: The Internet Society,2016.

[5] Li D,Li Y,Wu J,et al. ESM:Efficient and scalable data center multicast routing[J]. IEEE/ACM Transactions on Networking,2011,20(3):944-955.

[6] Abu-Libdeh H,Costa P,Rowstron A,et al. Symbiotic routing in future data centers[C]//Proc of the ACM SIGCOMM 2010 Conference,2010:51-62.

[7] Chen C C,Yuan L,Greenberg A,et al. Routing-as-a-service (RaaS):A framework for tenant-directed route control in data center[J]. IEEE/ACM Transactions on Networking, 2013,22(5):1401-1414.

[8] Besta M, Schneider M, Cynk K,et al. FatPaths:Routing in supercomputers,data centers,and clouds with low-diameter networks when shortest paths fall short[J]. arXiv: 1906.10885,2019.

[9] Greenberg A,Hamilton J R,Jain N,et al. VL2:A scalable and flexible data center network[C]//Proc of the ACM SIGCOMM 2009 Conference on Data Communication,2009:51-62.

[10] Al-Fares M,Loukissas A,Vahdat A. A scalable,commodity data center network architecture[J]. ACM SIGCOMM Computer Communication Review,2008,38(4):63-74.

[11] Guo C, Lu G, Li D, et al. BCube:A high performance, server- centric network architecture for modular data centers[C]//Proc of the ACM SIGCOMM 2009 Conference on Data Communication,2009:63-74.

[12] Guo C, Wu H, Tan K, et al. Dcell:A scalable and fault-tolerant network structure for data centers[C]//Proc of the ACM SIGCOMM 2008 Conference on Data Communication, 2008:75-86.

[13] Wang T,Su Z,Xia Y,et al. SprintNet:A high performance server-centric network architecture for data centers[C]// Proc of 2014 IEEE International Conference on Communications (ICC),2014:4005-4010.

[14] Benson T,Anand A,Akella A,et al. MicroTE:Fine grained traffic engineering for data centers[C]//Proc of the 7th Conference on Emerging Networking Experiments and Technologies,2011:1-12.

[15] Wu X,Yang X. Dard:Distributed adaptive routing for datacenter networks[C]//Proc of 2012 IEEE 32nd International Conference on Distributed Computing Systems,2012:32-41.

[16] Li Z,Zhang Y,Li D,et al. OPTAS:Decentralized flow monitoring and scheduling for tiny tasks[C]//Proc of the 35th Annual IEEE International Conference on Computer Communications(INFOCOM 2016),2016:1-9.

[17] Al-Fares M,Radhakrishnan S,Raghavan B,et al. Hedera: Dynamic flow scheduling for data center networks[C]//Proc of the 7th USENIX Conference on Networked Systems Design and Implementation, 2010: 19-20.

[18] Nong Huang-wu, Huang Chuan-he, Huang Xiao-peng. SDN-based multipath routing algorithm for Fat-tree data center networks[J]. Computer Science, 2016, 43(6):32-34. (in Chinese)

[19] Yang Yang,Yang Jia-hai, Qin Dong-hong. Multipath routing algorithm for data center networks[J]. Journal of Tsinghua University(Science and Technology), 2016, 56(3):262-268. (in Chinese)

[20] Peng Da-qin, Lai Xiang-wu, Liu Yan-lin. Multi-path routing algorithm for Fat-tree data center network based on SDN[J]. Computer Engineering, 2018, 44(4): 41-45. (in Chinese)

[21] Tam A S-W, Xi K, Chao H J. Use of devolved controllers in data center networks[C]//Proc of 2011 IEEE Conference on Computer Communications Workshops,2011:596-601.

[22] Spring N T, Mahajan R, Wetherall D. Measuring ISP topologies with rocketfuel[J]. ACM SIGCOMM Computer Communication Review,2002,32(4):133-145.

[23] Ramos R M, Martinello M,Esteve R C. SlickFlow:Resilient source routing in data center networks unlocked by OpenFlow[C]//Proc of the 38th Annual IEEE Conference on Local Computer Networks,2013:606-613.

[24] Wang Y C,You S Y. An efficient route management framework for load balance and overhead reduction in SDN-based data center networks[J]. IEEE Transactions on Network and Service Management,2018,15(4):1422-1434.

[25] Chen Song, Xie Wei. Routing optimization design for data center network traffic[J]. Communications Technology, 2018, 51(8):1883-1887. (in Chinese)

[26] He K,Rozner E,Agarwal K,et al. Presto:Edge-based load balancing for fast datacenter networks[J]. ACM SIGCOMM Computer Communication Review,2015,45(4):465-478.

[27] Sinha S, Kandula S, Katabi D. Harnessing TCP's burstiness with flowlet switching[C]//Proc of the 3rd ACM Workshop on Hot Topics in Networks (HotNets), 2004:1.

[28] Shang Y F,Li D, Xu M W. Energy-aware routing in data center network[C]//Proc of the 1st ACM SIGCOMM Workshop on Green Networking,2010:1-8.

[29] Xu M W, Shang Y F, Li D,et al. Greening data center networks with throughput-guaranteed power-aware routing [J]. Computer Networks,2013,57(15):2880-2899.

[30] He Rong-xi, Lei Tian-ying, Lin Zi-wei. Multi-constrained energy-saving routing algorithm in software-defined data center networks[J]. Journal of Computer Research and Development, 2019, 56(6):1219-1230. (in Chinese)

[31] Alizadeh M,Greenberg A,Maltz D,et al. DCTCP:Efficient packet transport for the commoditized data center[J]. ACM SIGCOMM Computer Communication Review, 2010, 40

(4):63-74.

[32] Vamanan B, Hasan J, Vijaykumar T N. Deadline-aware datacenter TCP (d2TCP)[J]. ACM SIGCOMM Computer Communication Review,2012,42(4):115-126.

[33] Mittal R, Lam V T, Dukkipati N, et al. TIMELY: RTT-based congestion control for the datacenter[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 537-550.

[34] Zhu Y, Eran H, Firestone D, et al. Congestion control for large-scale RDMA deployments[J]. ACM SIGCOMM Computer Communication Review,2015,45(4):523-536.

[35] Montazeri B, Li Y, Alizadeh M, et al. Homa: A receiver-driven low-latency transport protocol using network priorities[C]//Proc of 2018 Conference of the ACM Special Interest Group on Data Communication,2018:221-235.

[36] Cho I, Jang K, Han D. Credit-scheduled delay-bounded congestion control for datacenters[C]//Proc of 2017 Conference of the ACM Special Interest Group on Data Communication,2017:239-252.

[37] Kanagavelu R, Mingjie L N, Mi K M, et al. OpenFlow based control for re-routing with differentiated flows in data center networks[C]//Proc of 2012 18th IEEE International Conference on Networks (ICON),2012:228-233.

[38] Kanagevlu R, Aung K M M. SDN controlled local re-routing to reduce congestion in cloud data center[C]//Proc of 2015 International Conference on Cloud Computing Research and Innovation (ICCCRI),2015:80-88.

[39] Perry J, Ousterhout A, Balakrishnan H, et al. Fastpass: A centralized "zero-queue" datacenter network[J]. ACM SIGCOMM Computer Communication Review, 2014, 44(4): 307-318.

[40] Vissicchio S, Tilmans O, Vanbever L, et al. Central control over distributed routing[C]//Proc of 2015 Conference of the ACM Conference on Special Interest Group on Data Communication,2015:43-56.

[41] Alizadeh M, Edsall T, Dharmapurikar S, et al. CONGA: Distributed congestion-aware load balancing for datacenters [J]. ACM SIGCOMM Computer Communication Review, 2014, 44(4):503-514.

[42] Kandula S, Katabi D, Sinha S, et al. Dynamic load balancing without packet reordering[J]. ACM SIGCOMM Computer Communication Review,2007,37(2):51-62.

[43] Katta N, Hira M, Kim C, et al. HULA: Scalable load balancing using programmable data planes[C]//Proc of the Symposium on SDN Research,2016:1-12.

[44] Fan F, Hu B, Yeung K L. Routing in black box: Modularized load balancing for multipath data center networks[C]//Proc of IEEE Conference on Computer Communications, 2019: 1639-1647.

[45] Zhang J, Ren F, Huang T, et al. Congestion-aware adaptive forwarding in datacenter networks[J]. Computer Communications,2015,62:34-46.

[46] Ghorbani S, Yang Z, Godfrey P B, et al. Drill: Micro load balancing for low-latency data center networks[C]//Proc of 2017 Conference of the ACM Special Interest Group on Data Communication,2017:225-238.

[47] Mitzenmacher M. The power of two choices in randomized load balancing[J]. IEEE Transactions on Parallel and Distributed Systems,2001,12(10):1094-1104.

[48] Huang J, Lü W, Li W, et al. QDAPS: Queueing delay aware packet spraying for load balancing in data center[C]//Proc of 2018 IEEE 26th International Conference on Network Protocols (ICNP),2018:66-76.

[49] Kabbani A, Sharif M. Flier: Flow-level congestion-aware routing for direct-connect data centers[C]//Proc of IEEE Conference on Computer Communications,2017:1-9.

[50] Bao J, Dong D, Zhao B. DETOUR: A large-scale non-blocking optical data center fabric[C]//Proc of Asian Conference on Supercomputing Frontiers,2018:30-50.

[51] Wang G H, Andersen D G, Kaminsky M, et al. c-Through: Part-time optics in data centers[C]//Proc of the ACM SIGCOMM 2010 Conference,2010:327-338.

[52] Farrington N, Porter G, Radhakrishnan S, et al. Helios: A hybrid electrical/optical switch architecture for modular data centers[C]//Proc of the ACM SIGCOMM 2010 Conference,2010:339-350.

[53] Chen L, Chen K, Zhu Z, et al. Enabling wide-spread communications on optical fabric with megaswitch[C]//Proc of Networked Systems Design and Implementation,2017:577-593.

[54] Yang Ting-ting. Study on topology reconstruction strategy for the intra data center optical switching network[D]. Beijing: Beijing University of Posts and Telecommunications, 2017. (in Chinese)

[55] Sankaran G C, Sivalingam K M. A survey of hybrid optical data center network architectures[J]. Photonic Network Communications,2017,33(2):87-101.

[56] Wang C H, Javidi T, Porter G. End-to-end scheduling for all-optical data centers[C]//Proc of IEEE Conference on Computer Communications,2015:406-414.

[57] Kandula J P S, Bahl P. Flyways to de-congest data center networks[Z]. United States: Microsoft, 2009.

[58] Hamza A S. Recent advances in the design of optical wireless data center networks[C]//Proc of Broadband Access Communication Technologies XIII. International Society for Optics and Photonics,2019:109450K.

[59] Cui Y, Wang H, Cheng X, et al. Wireless data center networking[J]. IEEE Wireless Communications,2011,18(6): 46-53.

[60] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly,1955,2(1-2):83-97.

[61] Celik A,Al-Ghadhban A,Shihada B,et al. Design and provisioning of optical wireless data center networks: A traffic grooming approach[C]//Proc of 2018 IEEE Wireless Communications and Networking Conference,2018:1-6.

[62] AlGhadhban A,Celik A,Shihada B,et al. SoftFG: A dynamic load balancer for soft reconfiguration of wireless data centers[C]//Proc of 2020 IEEE Wireless Communications and Networking Conference,2020:1-5.

[63] Vanini E,Pan R,Alizadeh M,et al. Let it flow: Resilient asymmetric load balancing with flowlet switching[C]//Proc of the 14th USENIX Symposium on Networked Systems Design and Implementation,2017:407-420.

附中文参考文献:

[18] 农黄武,黄传河,黄晓鹏. 基于 SDN 的胖树数据中心网络的多路径路由算法[J]. 计算机科学,2016,43(6):32-34.

[19] 杨洋,杨家海,秦董洪. 数据中心网络多路径路由算法[J]. 清华大学学报(自然科学版),2016,56(3):262-268.

[20] 彭大芹,赖香武,刘艳林. 基于 SDN 的胖树数据中心网络多路径路由算法[J]. 计算机工程,2018,44(4):41-45.

[25] 陈松,谢卫. 面向数据中心网络流量的路由优化设计[J]. 通信技术,2018,51(8):1883-1887.

[30] 何荣希,雷田颖,林子薇. 软件定义数据中心网络多约束节能路由算法[J]. 计算机研究与发展,2019,56(6):1219-1230.

[54] 杨婷婷. 数据中心内光交换网络重构策略研究[D]. 北京:北京邮电大学,2017.

作者简介:

段晨(1997 -),男,陕西铜川人,硕士生,研究方向为数据中心网络路由。**E-mail**:duanchen@nudt.edu.cn

DUAN Chen, born in 1997, MS candidate, his research interest includes network routing in data center.

彭伟(1973 -),男,四川大邑人,博士,研究员,CCF 会员(E200015771S),研究方向为网络协议优化和无线自组网。**E-mail**:wpeng@nudt.edu.cn

PENG Wei, born in 1973, PhD, research fellow, CCF member (E200015771S), his research interests include optimization of network protocol, and wireless Ad Hoc network.

王宝生(1970 -),男,河北黄骅人,博士,研究员,研究方向为路由器架构、路由协议和网络空间安全。**E-mail**:bswang@nudt.edu.cn

WANG Bao-sheng, born in 1970, PhD, research fellow, his research interests include router architecture, routing protocol, and cybersecurity.