数据中心 Clos 网络负载均衡方案:问题、进展与展望

刁兴龙 余晓杉 刘 勇 黄狄涛 顾华玺

(西安电子科技大学 ISN 国家重点实验室 西安 710071)

摘 要 在云计算、大数据等信息技术的带动下,分布式应用成为数据中心网络内的主要业务模式.数据中心网络内部的流量呈现"东西向为主"、"大小流混合"的特性,实现数据中心网络负载均衡对提升网络带宽利用率、满足大小流需求具有重大意义.得益于数据中心网络 Clos 拓扑提供的并行相等跳数的路径,等价多路径(Equal-Cost Multipath,ECMP)被广泛用于数据中心网络负载均衡.在实际中复杂的流量场景下,ECMP 并未表现出预期的性能,哈希碰撞、鼠流阻塞问题未能满足象流的高吞吐需求以及鼠流的低时延需求,另外,由链路故障导致的网络拓扑非对称性问题也给负载均衡带来了重大挑战.本文首先分析了 ECMP 在实际应用中面临的问题,然后针对具体问题对现有负载均衡方案进行了综述,剖析了各方案的基本思想和实现方法,并对各方案的均衡效果进行了分析.最后,本文从处理粒度、决策方式、拥塞感知范围等多个维度对现有的数据中心网络负载均衡方案做了对比,总结并展望了未来数据中心网络负载均衡技术的发展趋势.

关键词 数据中心网络;负载均衡;ECMP;哈希碰撞;鼠流阻塞;非对称性

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2020.02241

Load Balancing Schemes for Data Center Network: Problems, Progress and Prospects

DIAO Xing-Long YU Xiao-Shan LIU Yong HUANG Di-Tao GU Hua-Xi

(*State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an* 710071)

Abstract Driven by information technology such as cloud computing and big data techniques, distributed applications have become the most popular business model in data centers networks. There are two characteristics of traffic within the data center: east-west traffic accounts for a large proportion of the total traffic within data center, and the traffic within data center is a mixture of large flows and small flows (also called elephant flows and mice flows, respectively). Load balancing in data center networks is of great significance to improve network bandwidth utilization and meet the demand of large flows and small flows. Modern data centers often employ multi-rooted Clos topology which provides multi paths with equal hops between any two hosts, thus Equal-cost Multipath (ECMP) is widely used for load balancing in data center networks. The realistic traffic is very complex, and ECMP does not show the expected load balancing effect. Firstly, this paper analyzes the problems faced by ECMP in the realistic scenario, such as hash collision and mice-flow blocking, which fails to meet the high throughput demand of elephant flows and the low latency demand of mice flows. In addition, frequent link failures in data center networks lead to asymmetric topology, which reduces the number of equal cost paths and brings

收稿日期:2019-02-24;在线发布日期:2020-05-11.本课题得到国家重点研发计划项目(2018YFE0202800)、国家自然科学基金(61634004,61934002,61901314)、陕西省杰出青年科学基金(2020JC-26)、中央高校基本科研业务费专项资金(JB190105)、计算机体系结构国家重点实验室开放课题(CARCH201919)资助.**刁兴龙**,博士研究生,主要研究方向为负载均衡、拥塞控制、数据中心网络.E-mail:xl.diao@stu.xidian.edu.cn.**余晓杉**,博士,讲师,主要研究方向为云计算数据中心光互连网络技术.**刘 勇**,博士研究生,主要研究方向为数据中心网络、软件定义网络.**黄狄涛**,硕士研究生,主要研究方向为拥塞控制、数据中心网络.**顾华玺**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)高级会员,国家重点研发计划项目首席,主要研究领域为网络技术、片上网络、光互连技术.E-mail:hxgu@xidian.edu.cn.

great challenges to load balancing. Then this paper investigates the existing load balancing schemes, analyzes the basic idea and the implementation of each scheme, and discusses the advantages and disadvantages of each scheme and the load balancing effect on the specific problem faced by ECMP. At last, this paper compares the various load balancing schemes in terms of three levels, including granularity, decision-making approach and the range of link-state awareness. Based on the comparison, this paper summarizes and forecasts the trend of future load balancing technology in data center networks. First, the granularity of load balancing tends to fine grained, because flow-level granularity can easily lead to hash collision and mice-flow blocking when elephant flows are transmitted in the network. Though packet-level granularity is the finest granularity and can theoretically achieve the optimal load balancing effect, this granularity faces serious challenges of packet reordering, especially in the asymmetric topology of data center networks. In this paper, we advocate that flowlet-level granularity is a promising technology for load balancing in future data center networks, because this granularity can both refine the granularity for better balancing performance and alleviate packet reordering. Second, the decision-making approach tends to be distributed. Compared to the centralized approach, distributed decision-making approach is more flexible, more timely, and more suitable for the fine-grained granularity of load balancing. Third, the decision-making function of load balancing tends to be deployed inside the network, because performing load balancing at the switch can ensure timely response for traffic burst. The host at network edge has abundant resources, which makes it more efficient to detect traffic information at the host, and the detected information can be used to assist the switch in performing load balancing. In this paper, we hope that the above trends can guide future research on load balancing in data center networks, such as dynamically setting flowlet timeout based on the delay differences of equal-cost paths.

Keywords data center networks; load balancing; ECMP; hash collision; mice-flow blocking; asymmetry

1 引 言

在互联网、大数据、云计算等信息技术的带动下,数据中心逐渐在世界各地普及和建立,已经成为社会各行业的必要基础设施.据思科对全球数据中心流量的预测①,预计 2021 年全球数据中心 IP 流量将达到 20.6 泽字节(ZettaByte),其中,数据中心与用户之间的流量、数据中心与数据中心之间的流量各占比 14.9%、13.6%,数据中心内部的流量占比高达 71.5%.另外,数据中心网络内部的业务需求发生了巨大转变,分布式业务取代了传统业务的主要地位,服务器集群之间的交互更为频繁,流量模式逐渐由"南北向流量"演进为"东西向流量"[1-2].

为应对日益增长的流量负载压力,数据中心网络通过负载均衡来实现高吞吐量和高可用性.对于南北向流量,一般通过负载均衡设备将来自数据中心外部的业务请求分配到集群网络内多个服务器上处理[3-4].对于东西向流量,一般通过静态或动态路由将流量均衡到数据中心网络的路径上.由于东西向流量占据了数据中心网络流量的极大比例,本文仅讨论数据中心网络内东西向流量间的负载均衡.

除了以东西向流量为主的特征外,数据中心网络的流量还具有明显的"大小流"特征.对于流的定义,一般认为五元组信息(源/目的 IP 地址、源/目的端口号、协议类型)相同的数据包为一条流.大流也称为象流,主要由数据存储、数据备份等应用产生,单条流的字节量较大,在网络中持续时间长,对网络带宽需求较大,例如数据存储应用程序交互处理的数据量一般会达到拍字节(PetaBytes)量级[5].小流也称为鼠流,主要由网页搜索、在线游戏等即时通信类客户应用产生,单条流的字节数较小,一般小于

① Cisco Global Cloud Index: Forecast and Methodology, 2016-2021. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html

100 KB，在网络中持续时间短，对网络时延敏感性较强. 在数据中心网络中，鼠流的总数目占比超过 80%[6]，但鼠流的总字节量占比较低，象流的总数目占比较少，但象流的总字节量占比极高，传输时约占据着数据中心网络 80%以上的带宽[7]，因此数据中心网络的负载均衡还应考虑到象流的高带宽需求、鼠流的低时延需求.

负载均衡与网络拓扑结构关系紧密，数据中心网络拓扑虽种类多样[8-16]，但目前大规模数据中心网络多是分层部署的 Clos 网络架构，并提倡横向扩展网络规模[17]. Clos 网络架构通过利用多个小规模、低成本的交换机来构建大规模、复杂的交换网络，常见的 Clos 架构包括两层的叶脊架构①、三层的 FatTree 架构[8-9]、三层的 VL2 架构[10]. Clos 网络架构对主机间的通信模式敏感度低、包容性强，因此也更容易均衡负载. 如无特殊说明，本文 Clos 架构均以 FatTree 架构为例，该结构如图 1 所示.

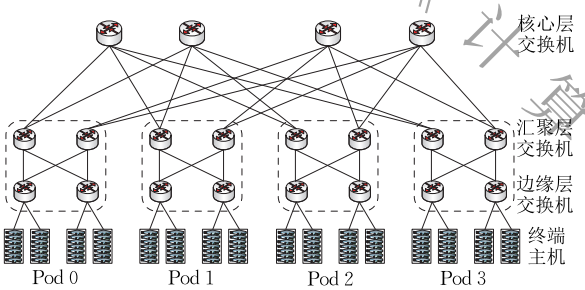

图 1　FatTree 拓扑架构

Clos 网络更容易均衡负载还主要归功于拓扑提供的并行相等跳数的路径，因此等价多路径(Equal-Cost Multipath，ECMP)[18-19]成为数据中心网络主流的负载均衡思想. 开放式最短路径优先(Open Shortest Path First，OSPF)[20]、中间系统到中间系统(Intermediate System to Intermediate System，ISIS)[21-22]等各种路由协议也已明确允许等价多路径路由. 结合 RIP 协议(Routing Information Protocol)[23]，ECMP 在现有商品交换机中已经得到了实现，通过对数据包报头中的字段进行简单哈希运算来选择下一跳地址，将一条流静态映射到一条等价路径上.

ECMP 由于简洁性而得到广泛部署，但 ECMP 在实际应用中也存在诸多不足. ECMP 既未考虑大小流问题，也未考虑网络当前状态的影响，前者会导致哈希碰撞、鼠流阻塞，后者则涉及到网络拓扑非对称性问题. 而一个流在其生命周期中，由于其与转发路径的单一映射关系，网络无法切换所选择的路径，很可能造成数据中心网络中路径带宽的不合理分配，以至于造成链路拥塞，进而恶化网络性能.

现有的负载均衡方案主要围绕实际应用中的问题来针对性地提出解决方法，在处理粒度方面可分为三类：流级别、数据包级别以及流切片级别. 处理粒度为流级别时，仅对一条流的第一个数据包进行路由计算，该流后续数据包的转发路径与该流第一个数据包的路径相同. 处理粒度为数据包级别时，则对流的每一个数据包均进行路由计算. 流切片是一条流的某段连续的数据包，一般按照时间[24]或字节量[25]可将一条流切分成多个流切片. 处理粒度为流切片级时，仅对一个流切片的第一个数据包进行路由计算，同一流切片内的数据包的转发路径相同，不同流切片的转发路径可能不同.

文献[26]从不同级别处理粒度对现有负载均衡方案作了综述，文献[27]从响应方式和处理粒度方面对现有负载均衡方案进行了简单讨论，但均对现存的实际应用问题针对性不足.

本文将首先分析数据中心网络负载均衡在实际应用中的问题，再逐一针对各实际应用问题对现有负载均衡方案进行讨论，详细分析各方案的基本思想、实现方法及均衡效果，总结并展望数据中心负载均衡技术的发展趋势.

2　负载均衡现存问题

ECMP 的处理粒度为流级别，路由结果依据于对数据流五元组的静态哈希运算. ECMP 决策时不考虑流的字节大小以及链路的承载能力，进而导致网络带宽利用率较低. 在大规模数据中心网络中，流量负载巨大、服务器间通信频繁，很容易发生以下情况：

(1) 哈希碰撞

不同象流的转发路径同时经过某一物理链路，即该段物理链路同时负载着多条象流，称这种情况为哈希碰撞. 由于象流承载了网络中 80%以上的数据传输，单条象流的字节量较大，一旦发生哈希碰撞，碰撞链路无法同时满足多条象流的带宽需求，进而产生拥塞.

数据中心网络中的每一跳路径都有可能发生哈希碰撞，如图 2 所示，象流 1 和象流 2 在核心层交换

① Cisco Data Center Spine-and-Leaf Architecture: Design Overview. https://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/white-paper-c11-737022.html

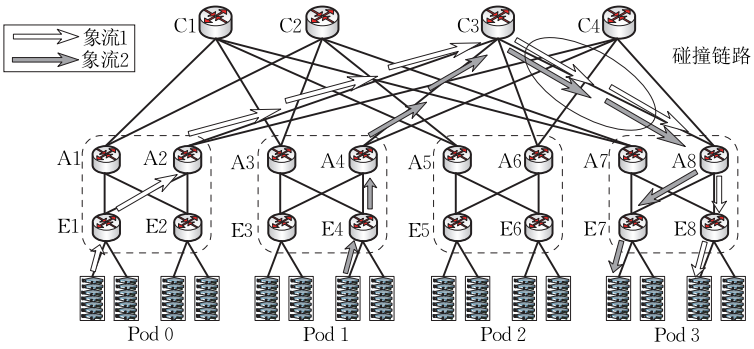

图 2 哈希碰撞示意图

机 C3 至汇聚层交换机 A8 的下行链路上发生路径碰撞.

(2) 鼠流阻塞

ECMP 决策时并不区分象流和鼠流,交换机对缓存中的数据包按序转发,而象流字节量大、数据包数目较多,因此鼠流数据包会排队在象流数据包后面[28-29],造成鼠流发生队头阻塞,这种情况称为鼠流阻塞.

由于象流需要传输的字节数大,在交换机缓冲区中占用的空间大,进而导致缓冲区队列较长.而鼠流多为即时通讯类业务产生,对时延敏感性要求较高,对网络缓冲区的诉求是短队列、低排队时延.一旦发生鼠流阻塞,网络缓冲区长时间被象流占据,则鼠流将会经历较长的排队时延,进而增加了鼠流的流完成时间(Flow Completion Times,FCT),导致较高的长尾时延,最终将导致业务的服务质量下降.

(3) 非对称性

ECMP 路由算法在对称的网络拓扑中理论上能实现较优的负载均衡效果,然而在实际的大规模数据中心中,链路故障以及交换机故障频率较高[30-31],据统计,每天的链路故障数高达 40 次[32],严重破坏了数据中心网络拓扑的对称性[33],进而导致 ECMP 算法的实际负载均衡效果不理想.

链路故障分为部分链路故障和全链路故障,部分链路故障指的是链路容量的下降,而全链路故障指的是链路完全失效,这两种链路故障均会造成网络拓扑非对称.

非对称拓扑具体表现为拓扑带宽非对称以及拓扑时延非对称.如图 3 所示,链路 A2-C4、A4-C4 发生部分故障,假设链路容量从 10 Gbps 降低到 1 Gbps,相对于其他正常链路,链路 A2-C4、A4-C4 无法达到原有的高吞吐量,此时,整个网络拓扑在带宽方面已经失去对称性.带宽非对称会引起网络拥塞,并进一步影响网络吞吐.

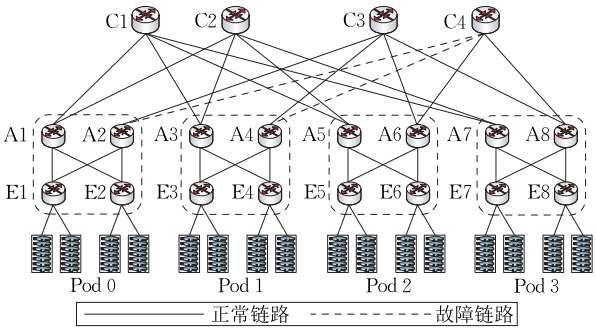

图 3 链路故障示意图

部分链路故障或全链路故障均会导致拥塞产生,破坏负载均衡效果,尤其是在采用多路径路由时会导致路径间的时延差异,造成时延非对称.仍以图 3 所示为例,假设 Pod0 的所有业务流量都发向 Pod3,但此时链路 A2-C4 发生完全故障,业务流量无法在 C4 处中转.相比于由核心层交换机通向 Pod3 的其他下行链路(如链路 C3-A8),链路 C4-A8 上的负载相对较轻.若此时 Pod2 存在通向 Pod3 的业务流量,经由核心层交换机 C3 抵达 Pod3 需利用链路 C3-A8,经由核心层交换机 C4 抵达 Pod3 需利用链路 C4-A8.由于链路 C4-A8 上的负载比链路 C3-A8 上的负载轻,则在链路 C4-A8 上将经历较低的时延,在链路 C3-A8 上将经历较高的时延,进而出现时延不对称现象.时延不对称性对数据包级别的均衡算法影响最为明显,同一条流的数据包在发送端顺序地被转发至多条路径,不同路径间时延非对称,先发送的数据包后到达,在接收端将诱发严重的数据包乱序.数据包乱序将引起发送端拥塞窗口频繁缩减[34-35],导致网络吞吐性能的降低.

在复杂的流量场景下,ECMP 逐渐暴露出一些问题,如表 1 所示,受哈希碰撞、鼠流阻塞以及非对称性的影响,ECMP 无法完全满足“大小流”各自需求.

表 1　ECMP 的问题及影响

问题	原因	影响
哈希碰撞	多条象流被转发到同一路径	无法满足象流的高带宽需求；且会导致拥塞
鼠流阻塞	交换机缓存被象流占据，鼠流等待	无法满足鼠流的低时延需求
非对称性	交换机或链路故障导致路径间带宽、时延非对称	恶化负载均衡效果

3　研究现状分析

本节围绕数据中心网络负载均衡面临的问题，将具体介绍相关的负载均衡方案，并对比分析各方案的基本思想、实现方法及均衡效果.

3.1　针对哈希碰撞的负载均衡方案

学术界提出多种针对哈希碰撞的负载均衡方案，解决思路大概可以分为以下几类：(1) 检测象流，为象流分配路径，如 Hedera[36]、Mahout[37]、FlowBender[38]、SCAA[39]；(2) 采用细粒度均衡，如数据包级别负载均衡方案 RPS[40]、DRB[41] 以及流切片级别的负载均衡方案 Flare[24]；(3) 对象流鼠流区分路由，如 DiffFlow[42].

3.1.1　重路由策略

Hedera[36] 是一种流级别的负载均衡方案，执行过程有三步：(1) 在边缘层交换机处检测象流；(2) 在集中控制器处估算象流的带宽需求以动态分配新路径；(3) 集中控制器将象流路径信息下发到各交换机，迁移象流以减轻拥塞.

边缘交换机检测来自终端主机的流的速率，一旦某条流的速率超过链路物理带宽的 10%，则判定该流为象流，并向集中控制器报告. 在带宽需求的估算中，集中控制器基于最大最小公平算法，建立象流的带宽需求矩阵，经过多次迭代最终得到流的“自然需求”. 分配新路径采用了两种方法：全局最先匹配法(Global First Fit)和模拟退火法(Simulated Annealing). 全局最先匹配法采用贪婪的方式，将象流分配到第一条能够满足带宽需求的路径上，实现较为简单. 模拟退火法先通过将核心层交换机分配到终端主机进行优化以减少搜索空间，再运行模拟退火算法计算得出象流路径.

Hedera 方案的不足之处在于：(1) 该方案基于集中控制器来决策均衡策略，执行周期高达 5 s，而数据中心网络业务流的完成时间大多在毫秒级，该方案无法及时响应网络负载的动态变化情况；(2) 该方案过分依赖象流识别与象流的稳定性，实际中主机硬盘速度会限制数据流速率，主机之间的数据流带宽可能很少会超过链路带宽的 10%，无法触发象流识别门限，从而得不到重路由的机会.

与 Hedera 不同的是，Mahout 方案[37] 在终端主机处检测象流，当某条流的套接字缓存高于设定阈值后则判定为象流，不仅检测效率高而且能降低交换机的计算开销. 终端主机利用带内信令将象流信息报告至集中控制器，集中控制器基于 Increasing First Fit 算法[43] 为象流分配路径. 对于未被判定为象流的其他流，交换机基于 ECMP 对其进行转发.

FlowBender[38] 是一种流级别的负载均衡方案，该方案基于显式拥塞通知(Explicit Congestion Notification，ECN)技术[44]，对遇到拥塞或预期即将遇到拥塞的流进行重路由. 发送端统计在一个往返时延(Round-Trip Time，RTT)内被 ECN 标记的数据包占该流总数据包的比例，如果该比例大于既定阈值(通常设置为 1%到 10%)，则判定该流当前使用的路径发生拥塞，需要对该流进行重路由. 重路由时每一个报文的转发路径不仅取决于五元组信息，还由 TTL 字段决定. 终端若判定一条流遭遇拥塞则修改该流的 TTL 字段，交换机有一定概率为该流后续的报文分配其他的转发路径，但重路由时选路存在随机性，决策结果可能无法收敛.

文献[39]提出一种流级别的基于链路状态感知的动态负载均衡方案(Simple Congestion-Aware Algorithm，SCAA). 该方案根据链路利用率来配置链路权重，将新产生的数据流转发到权重最小的路径上. 该方案定义链路开销为链路利用率的凸函数，并认为网络开销是所有链路开销的综合. 通过建立数学模型描述网络流量，使用链路上流量总和表征链路状态. 以最小化网络开销为目标优化流量模型，根据优化结果将流量调度到链路开销最小的路径上.

该方案可实现渐进性的负载均衡优化效果，算法复杂度较低，并且适用于数据中心网络任意拓扑、任意业务模式. 但该方案不足之处在于：(1) 实际部署的效果取决于数据中心网络中链路的权重更新以及最小权重路径的计算速度. 需要通过实时计算象流权重、对鼠流使用此前已计算的最小权重路径，来提升算法的运行速度；(2) 该算法一旦部署到大规模数据中心时，交换机拥塞信息的存储会加重硬件开销，因此该方案扩展性较差.

3.1.2　细粒度级策略

文献[40]提出在数据中心网络中采用数据包级别的负载均衡方案——随机包扩散(Random Packet Spraying，RPS). 不同于 ECMP 算法中对五元组进行哈希运算来确定整条流的转发路径，该方案对一

条流的每一个数据包单独决策转发路径，并随机地将其转发到候选下一跳，并且该方案不需要对主机进行修改，在很多商用交换机上可以部署.

处理粒度为数据包级别的方案可以从根本上避免哈希碰撞，通过把同一条流的数据包分散到不同的路径上可以提升链路利用率，理想情况下的负载均衡效果趋于最优.但数据包级别处理粒度的方案会面临数据包乱序问题，网络负载不均衡时将导致链路负载程度存在差异，具体表现为交换机队列长度存在差异，由于同一条流的不同数据包采用了不同的转发路径、历经不同的路径时延，可能出现“先发送后到达”的情况.数据包乱序将对 TCP 拥塞控制产生负面影响，这也是商用交换机默认不打开数据包粒度级路由功能的原因.该方案提出基于最小化队列长度差值原则，比如通过限制队列长度最大值等措施，来限制队列长度差值小于最长队列长度，以缓解数据包乱序.

文献[41]也提出一种数据包级别负载均衡方案——数字反转跳变法(Digit-Reversal Bouncing，DRB).不同于 RPS 方案对数据包的随机性转发，DRB 方案为一条流的每个数据包选择交错的转发路径，以避免同一条流相邻的数据包被转发到同一路径上，如图 4 所示.该方案执行步骤可分为：(1)源主机为每个数据包预先选择一个核心层交换机作为跳变交换机，然后对该数据包进行 IP-in-IP 封装，即外部 IP 标头的目标地址设置为所选跳变交换机的地址；(2)当跳变交换机接收到数据包后，将数据包解封装并将数据包发送到目的主机.

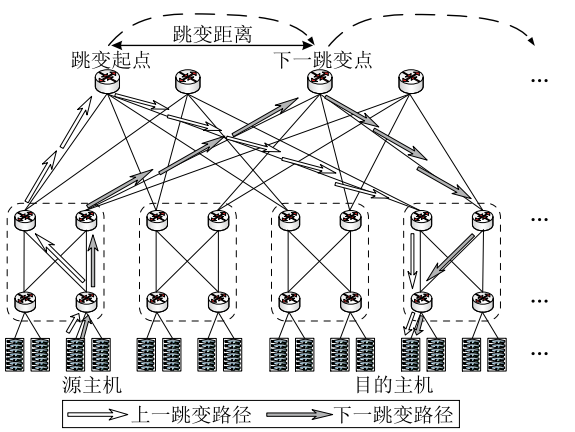

图 4 数字反转跳变示意图

在 FatTree 结构中，核心层交换机和等价路径唯一对应，选定跳变交换机就选定了转发路径.选择跳变交换机的方法基于数字反转跳变法，把核心层交换机依次编号，本次选择的交换机的编号与上次选择的交换机的编号间隔相等，即跳变距离相同.由于路由路径的唯一性，因此可以静态地配置交换机的路由表.

DRB 方案采用静态路由，对交换机路由表条目需求不大，交换机的路由条目数最多仅为终端主机子网数加上跳变交换机的数量，约为数千个，因此现有商用交换机均可满足需求.该方案还针对网络拓扑非对称性提出了解决办法，将在下文进行分析.但该方案采用数据包级别的均衡算法，同样会受到数据包乱序的影响，特别是在网络中存在链路故障的情况下.

文献[45]设计了一个基于排队时延感知的数据包分发策略——QDAPS(Queueing Delay Aware Packet Spraying)，能够有效缓解数据包级别负载均衡面临的数据包乱序问题.该方案根据输出端口排队时延选择路径来实现先发送的数据包先到达目的端.

QDAPS 方案的部署位置是交换机，有三个主要模块：(1)估计排队时延，当一个新的数据包到达时，根据其分配端口的当前队列长度估计数据包的排队时延.对于每个流，该方案仅记录最后达到数据包的排队时延；(2)数据包重新排序，该方案为后发送的数据包选择排队时延较大的输出端口进行转发；(3)重路由象流，象流可能会经历较长的排队时延，当前队列长度大于给定阈值时，重路由数据包到队列长度最短的输出端口.

当流的第一个数据包到达时，选择具有最短队列长度的输出队列，对于之后到达的数据包，选择输出队列时需要确保其排队时延大于前一个到达数据包的排队时延.鼠流通常很快完成传输，不会经历长时间的排队时延.而象流的数据包因为上一个规则的制约会堆积在某几条队列上，最终出现某些队列的堆积而某些队列没有排队.因此当队列长度超过阈值时，对象流进行重路由.一方面，一条流如果持续发送数据并且一直在某个队列上堆积，较大的排队时延会导致缓慢的拥塞窗口增长和网络链路的低利用率；另一方面，重新路由到负载较轻的路径上虽然会降低排队时延、提高链路利用率，但是会导致数据包乱序.因此重路由需要从开销和收益角度来进行权衡.

Flare 方案[24]在流级别处理粒度和数据包级别处理粒度间做了折中，基于流突发特性，按时间将一条流切分成多个流切片，不同流切片被转发到不同的并行等价路径上，能将同一条流的负载拆分到多条路径上.

该方案首先在交换机处周期性检测等价路径间

的最大时延差，并以最大时延差作为切分流切片的时间阈值. 然后再通过判别同一条流的相邻数据包之间的时间间隔来检测流突发，若相邻流突发的时间间隔超过既定时间阈值则将当前流突发内所有数据包切分成一个新的流切片. 该方案在切分流切片时保证了相邻流切片之间的时间间隔不小于等价路径间的最大时延差，因此细化处理粒度提升均衡效果的同时，还能避免数据包乱序问题.

3.1.3　区分调度策略

DiffFlow 方案[42]是一种流级别和数据包级别混合的负载均衡方案，该方案对象流鼠流进行区分路由，处理过程可分为两步：(1) 检测象流鼠流；(2) 区分路由，对象流采用数据包级别算法，对鼠流采用流级别算法. 在具体执行过程中，栈顶(Top Of Rack，TOR)交换机每隔一段时间对数据包抽样，如果检测到有两个数据包来自同一个流，则判定该流为象流，并认为可能存在链路拥塞的风险，同时 TOR 交换机向 SDN 控制器报告象流信息. SDN 控制器保存象流的信息，并将象流信息下发至网络中其他的交换机. 交换机通过比对转发规则，如果是鼠流，就用默认的 ECMP 方法进行转发；如果是象流，则采用 RPS 方案[40]转发处理.

通过检测象鼠流并对象鼠流区分路由，该方案能很好地避免哈希碰撞，提高网络吞吐的同时还能降低流完成时间，执行效果如图 5 所示. 但该方案也

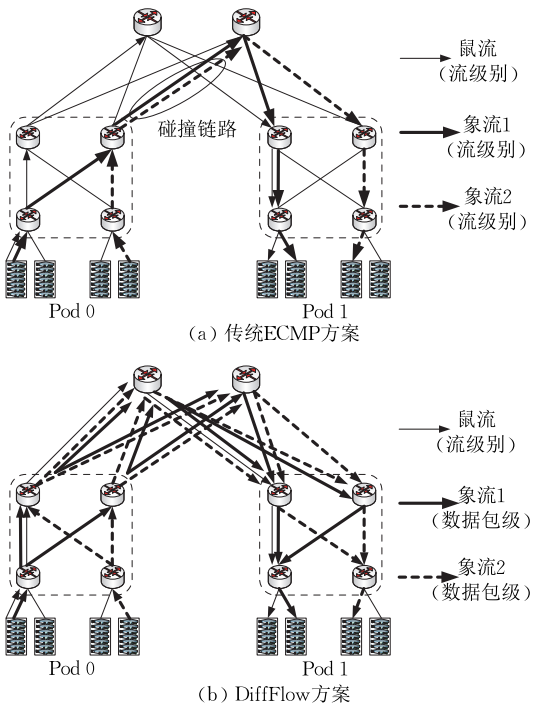

图 5　DiffFlow 方案解决哈希碰撞示意图[42]

存在一些不足：(1) 对象流采用 RPS 进行路由，极易引发数据包乱序问题；(2) DiffFlow 策略依赖于集中控制器来统筹象流的处理过程，包括建立转发规则以及下发转发规则，而集中控制器调度周期一般较长，无法应对网络中的突发流量.

3.2　针对鼠流阻塞的负载均衡方案

近年来学术界对于 ECMP 存在的鼠流阻塞问题也进行了诸多研究，解决方案可概述为：(1) 流量切分式调度，如 TinyFlow[28]、Presto[25]；(2) 对象流鼠流区分调度，如 Freeway[46]、CAPS[29]；(3) 对象鼠流分配优先级，如 DeTail[47]、PIAS[48]、Aemon[49].

3.2.1　流量切分策略

TinyFlow 方案[28]对流进行细粒度划分，通过将象流分解为多条鼠流，并将鼠流动态重路由到不同路径上，可同时解决鼠流阻塞以及哈希碰撞，执行效果如图 6 和图 7 所示. 该方案包括两部分：象流检测和随机重路由. 对于象流检测，通过在边缘层交换机的下行端口对数据流抽样，若一个抽样周期内抽样的数据包属于同一条流的数量超过某个阈值，则标记该条数据流为象流. 对标记的象流，通过随机重路由的方式将象流随机转发到等价路径上.

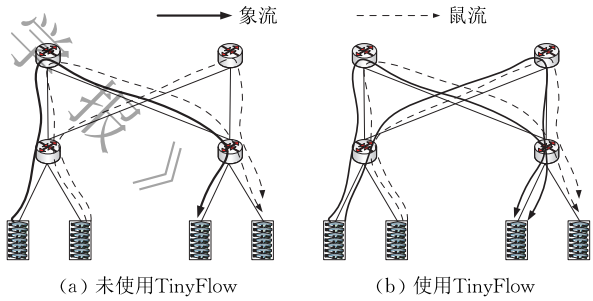

图 6　TinyFlow 解决鼠流阻塞示意图[28]

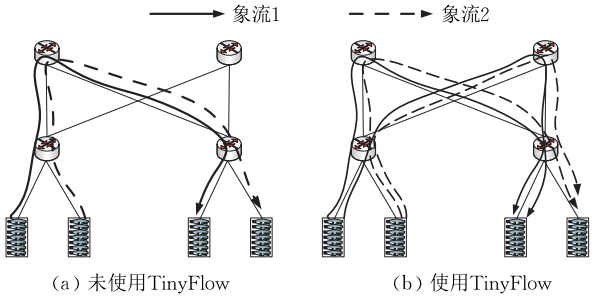

图 7　TinyFlow 解决哈希碰撞示意图[28]

Presto 方案[25]在边缘层交换机处将每一条流按字节量均切分成统一字节大小(64 KB)的流切片，来执行细粒度的负载均衡. 该方案利用集中控制器收集网络拓扑信息，每个虚拟交换机对应的每个生

成树都有一个相应的 MAC 地址，而每个交换机的物理 MAC 地址映射一列对应的影子 MAC 地址.因此流量可通过多个生成树发送到目的交换机，同一对源目的结点对之间会有多个影子 MAC 地址对应，流量就可以通过多条路径达到.

在 Presto 方案中，总字节量小于 64 KB 的流不会发生数据包乱序，但是大于 64 KB 的流可能会导致严重的乱序.该方案通过修改管理程序中的 GRO (Generic Receive Offload)将数据包聚合到更大的段中以应对数据包乱序问题.修改 GRO 后，可以区分数据包丢失还是乱序，一旦乱序发生，该段被保持，并创造一个新的段.因此乱序数据包有足够的时间到达接收端并正确排序.若发生丢包，GRO 立即向上推高当前段，以便 TCP 尽快对丢包作出反应.GRO 在虚拟机内核中实现，由 NIC 驱动直接调用，为了使 NIC 不发生改变，接收端数据缓冲区被置于 GRO 层.发送端一次最多只能发送最大 64 KB 的流切片，同一个流想要继续发送需要切换影子 MAC.每 64 KB 切换线路会导致乱序，为了处理这个问题，该方案引入了流切片 ID，每个流切片的 ID 按顺序增加.流切片 ID 相同的数据包路径相同并且有序，当流切片 ID 相同的数据包序列号出现间隙时默认为丢包而不是发生乱序.当流切片边缘出现缺口而无法判断是丢包还是乱序时，可根据实际调试经验设置超时参数，若超时则判断为丢包.

Presto 方案基于全局控制，每个虚拟交换机维护多个影子 MAC 地址，一旦网络规模拓展，所有交换机需要更新所有的影子 MAC 信息，规模越大造成的网络消耗越大.

3.2.2 区分调度策略

Freeway 方案[46]采用自适应象鼠流区分调度来解决鼠流阻塞问题.该方案将网络中的路径动态划分为低时延路径和高吞吐路径，将鼠流转发到低时延路径上，将象流转发到高吞吐路径上.

该方案分为两步执行：(1) 首先将路径划分为用于转发鼠流的低时延路径和用于转发象流的高吞吐路径，并根据路径上实时的流量情况，动态调整两种路径的数量；(2) 对于象流，以最大化剩余带宽、最小化链路资源利用率为优化目标，通过全局网络视图，将象流转发到最轻负载的高吞吐路径上；以最小化链路传输时延为优化目标，将鼠流转发到最轻拥塞的低时延路径上.

动态路径隔离算法在设计和实现上需要满足两个约束条件：一个是源、目的服务器之间至少有一个低时延路径和一个高吞吐路径；另一个是基于实时的鼠流情况，能够动态添加和动态消减低时延路径.为了同时满足两个约束条件，如图 8 所示，在一个 8 个 Pod 的 FatTree 拓扑架构上，从交换机 A 到交换机 B 之间的序号 1 到序号 4 路径同时被划分为低时延路径，同时路径 13 到路径 16 被同时划分为高吞吐路径，因此至少各有 4 条高吞吐路径和低时延路径满足约束 1 条件.为了满足约束条件 2，路径 5 到路径 13 可根据鼠流需求被动态划分低时延路径或高吞吐路径.实际网络运行过程中，如果低时延路径上平均链路利用率超过临界值，则认定该路径发生拥塞，如果小于临界值，则认定该路径空闲.如果存在一半以上的拥塞路径，则添加一个新的低时延路径；如果存在一半以上的空闲路径，则消减一个低时延路径.

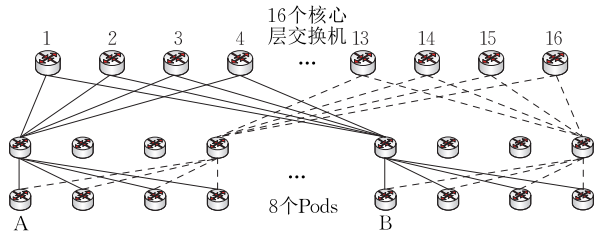

图 8 链路划分示意图[46]

该方案能有效降低鼠流的流完成时间，并提升网络吞吐，但不足之处在于，象流的调度需要通过一个集中控制器并根据全网视图才能将其转发到最优路径上.

CAPS 方案[29]提出一种基于编码的自适应数据包扩散机制.为应对鼠流阻塞问题，该方案对鼠流的处理粒度为数据包级别，采用 RPS 方案将鼠流转发到所有等价路径上.由于鼠流并非全时段均在进行传输，存在“活动/熄灭”状态.因此在鼠流处于“活动”状态时，对象流的处理粒度为流级别，采用 ECMP 将象流转发到某些固定的等价路径上.在鼠流处于“熄灭”状态时，象流采用 RPS 方案执行转发.

针对 RPS 方案的数据包乱序问题，发送端对流数据包进行前向纠错(Forward Error Correction, FEC)编码，若当前发送窗口含有 k 个数据包，则扩展冗余至 $k+r$ 个数据包，并根据阻塞概率动态调整冗余包的数量.即使 $k+r$ 个数据包不能完整按序到达接收端，接收端也可根据其中任意 k 个数据包恢复出发送端原始发送窗口的数据.

当象流基于 ECMP 被转发，象流数据量也仅会影响到一条路径的拥塞情况.鼠流基于 RPS 被转发，则大部分鼠流数据包不会经历拥塞，能够降低鼠

流流完成时间.即使小部分鼠流数据包被阻塞在网络内,也可基于 FEC 技术由正常到达的数据包恢复出失序数据包.

由于 FEC 编码层位于传输层和网络层之间,该方案不需要对 TCP/IP 协议栈做更改.该方案能很好地降低鼠流的流完成时间,但实际数据中心网络中鼠流数目较多,对鼠流采用 FEC 编码技术会增加大量的冗余编码包,会加重网络负载.

3.2.3　优先级策略

DeTail 方案[47]利用底层协议检测拥塞并保证高优先级流优先传输,并通过跨层信息在上层执行自适应路由决策,以均衡网络负载.

该方案首先在应用层通过套接字接口向下层传递数据流的优先级.传统链路层流控策略会导致队头阻塞,而 DeTail 方案采用基于优先级的流控策略(Priority Flow Control,PFC).当交换机输入端口缓存占用率到达某一门限后,前向反馈至上一跳交换机或主机,令其停止发送低优先级的数据包.这样既保证了高优先级流的数据包不被阻塞,也避免了网络拥塞导致的丢包行为.在网络层,该方案执行数据包级别的自适应负载均衡,基于交换机输出端口缓存占用率选择轻拥塞的下一跳地址.由于链路层保证了无丢包传输(硬件故障情况除外),即使传输层采用数据包级别的均衡算法,也不会导致传输层发生数据包乱序重传行为.因此 DeTail 方案中解除了传输层中的快速恢复机制和快速重传机制,来保证网络层自适应决策更大的灵活性,但这样会增加终端主机的缓存开销.当低优先级流的数据包占用较多交换机输出端口缓存时,通过对其设置 ECN 标记反馈至发送端来降低其发送速率,同时还能消除潜在的队头阻塞.

PIAS 方案[48]利用多级反馈队列(Multiple Level Feedback Queue,MLFQ)来实现流量调度以最小化流完成时间,同时也巧妙地解决了负载均衡层面的鼠流阻塞问题.该方案在未知流先验信息的情况下模拟最短任务优先算法①,利用现有商用交换机中可用的多个优先级队列来实现一种多级反馈队列,根据流已发送的字节总量逐渐将流从高优先级队列降级到低优先级队列.

具体执行步骤可分为:(1)在不同队列中的数据包以严格的优先级进行调度.商用交换机中一般会有 4～8 个队列,首先在终端主机处,标记数据包的优先级(如一级到四级),随着该流传输的字节数增多,逐渐降低该流的优先级.然后在交换机处执行严格的优先级排队;(2)在同一队列中的数据包基于先进先出(First Input First Output,FIFO)进行调度.降级到相同的低优先级队列的大流会公平地共享链路,这有助于最小化大流的响应时间,从而减轻大流饿死问题.

优先级降级的阈值应取决于网络负荷以及流大小的分布.该方案使用测量值以估计整个数据中心网络中负载和流大小的分布,为所有终端主机推导出一组通用的降级阈值.该方案能够保证鼠流在前几个高优先级队列中完成,而象流最终会降级为低优先级队列,于是鼠流一般先于象流完成传输,因此能够有效避免鼠流阻塞问题.不足之处在于,当前甚至是过去的流量信息并不能反映此后的流量形式,MLFQ 方案中设定的降级阈值会出现参数失配的情况.

Aemon 方案[49]也能很好地缓解鼠流阻塞问题,该方案定义一条流的紧急程度随时间增加,并采用两级的优先级调度时延敏感流和时延不敏感流.鼠流多属于时延敏感流,随着一条流传输时间的增加,该方案动态提高时延敏感流的优先级,降低时延不敏感流的优先级,在鼠流的截止时间之前优先转发鼠流来缓解被阻塞.在相同的优先级别中,该方案优先发送时延不敏感流来避免该流饿死,但此时可能依然会发生鼠流阻塞现象.

3.3　针对非对称性的负载均衡方案

在大型数据中心中,交换机或链路故障频率较高,导致网络拓扑非对称性,进而恶化负载均衡效果.学术界针对非对称性进行了许多研究,解决方案可以概述为:(1)处理过程区分故障链路与正常链路,如 SAPS[50]、DRILL[51]、WCMP[52];(2)探测可用路径来获取拓扑信息,如 CLOVE[53]、HULA[54]、Hermes[55]、Luopan[56]、DRB[41];(3)故障链路会导致拥塞产生,通过感知并反馈拥塞来规避故障链路,如 CONGA[57]、LetFlow[58].

3.3.1　路径区分策略

SAPS 方案[50]是一种拓扑自适应性的数据包级别均衡方案(Symmetric Adaptive Packet Spraying),该方案将实际中不对称的拓扑划分成多个对称的虚拟拓扑,如图 9 所示,特定数据流将在对应的虚拟拓扑中完成数据传输.该方案可分三部分:(1)将实际中不对称的拓扑映射成多个对称的虚拟拓扑,这种

① 最短任务优先算法(Shortest Job First,SJF)是一种最优的调度算法,单链路上的平均流完成时间最小.

虚拟拓扑是源和目的节点之间具有对称带宽和对称延迟的路径构成的集合；(2) 结合流的大小及虚拟拓扑容量，基于一定概率在流刚产生时就将流映射到一个虚拟拓扑中，较小的流多被映射到半分带宽较小的虚拟拓扑，较大的流将始终被映射到半分带宽较大的虚拟拓扑；(3) 在虚拟拓扑对每条流执行 Packet Spraying 算法进行路由.

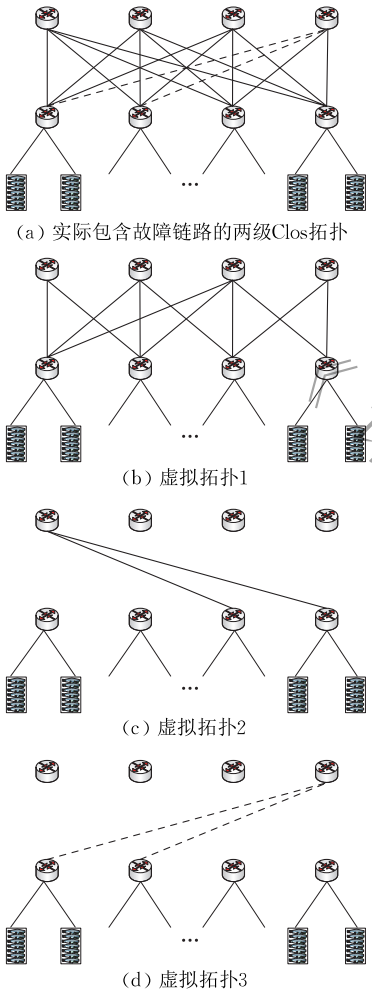

图 9 两级 Clos 网络虚拟拓扑划分示意图

虚拟拓扑的划分要保证带宽对称和时延对称. SAPS 方案采用以下原则来保障带宽对称：容量不同的链路将被划分到不同的虚拟拓扑中，每个虚拟拓扑均由一条条完整的物理链路组成. 当路径发生故障时，交换机向集中控制器发送对应端口的故障状态信息，集中控制器根据转发规则更新流表项，进而抽象出对称虚拟拓扑. 另外，SAPS 方案采用 DCTCP(Data Center TCP)作为传输层协议并通过 ECN 反馈拥塞，这样可以避免网络中交换机队列过长，通过减少排队时延来处理路径时延不对称性.

每个新流在终端主机处产生后，终端主机对流的大小进行标记，标记通过修改所属该流的数据包的 IP 报头 DSCP 字段实现. 在交换机处，根据对数据包的标记决定数据包所属的流映射到何处虚拟拓扑. 大流将始终被映射到半分带宽较大的虚拟拓扑，若流的大小低于某一门限，执行概率映射，则小流有时可能被映射到半分带宽较小的虚拟拓扑.

SAPS 通过将实际不对称的拓扑划分成对称的虚拟拓扑，并在拓扑中采用数据包级别均衡算法，可以实现极佳的负载均衡效果. 但该方案基于 OpenFlow 交换机实现，依赖于集中控制器来划分虚拟拓扑，执行周期较长，且对网络突发流量不敏感. 当部分故障链路的分布在大规模网络中较为分散时，某条部分故障链路可能无法被映射到虚拟拓扑.

DRILL 方案[51]是一种基于链路状态感知的数据包级别的负载均衡方案，根据本地交换机的队列占用情况为每个数据包选择下一跳转发地址，可实现微秒时间尺度上的微负载均衡. 该方案在本地交换机处采样出队列信息，将两个随机端口的队列长度与此前使用的出端口队列长度作比较，将数据包发送到队列长度最小的端口. 为应对数据包乱序问题，该方案在终端主机 GRO 层部署缓存来缓冲等待失序数据包.

在应对不对称性方面，该方案首先将不对称的网络拓扑划分成一个个对称的路径组，然后将流散列到某个路径组，在每个路径组内的等价路径上执行上述数据包级别的微负载平衡. 随着网络不对称的加重，该方案会逐渐趋近于传统 ECMP 方案的处理机制.

文献[52]从路由协议的拓扑视角出发，提出一种流级别负载均衡方案——加权成本多路径(Weighted Cost Multipath，WCMP). WCMP 方案先对下一跳路径对应的交换机出口端口进行权重赋值，选择下一跳转发地址时考虑出端口权重值，以优化流在多条路径中的分布.

WCMP 可在基于商用交换机构建的 SDN 网络上运行，并根据网络拓扑的动态变化，按可用链路容量的比例，将流量均衡到可用的下一跳节点中，方案执行框架如图 10 所示. 任意拓扑可通过运行最大流最小割算法为每个源目的对创建 WCMP 组并分配链路权重，若进一步利用权重优化算法来减少权重数目，还可节省表条目资源.

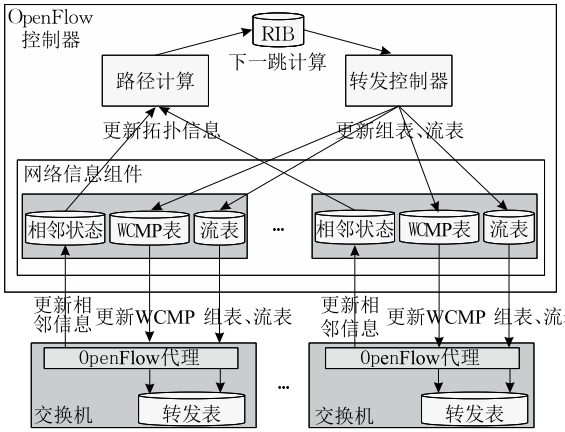

图 10　WCMP 系统框架示意图[52]

交换机通过创建多路径组来实现多路径转发，该多路径组表示目的地的“等成本”出口端口阵列，每个出口端口对应于可用于到达目的地的多个路径之一. 交换机散列到达的包头以确定多路径组中每个包的出口端口. 通过散列数据包报头中的特定字段可确保同一流中的所有数据包都遵循相同的网络路径，从而避免数据包重新排序.

该方案将 WCMP 端口权重映射到较小的整数权重集上，在实现加权散列时，能显著降低交换机转发表条目. 在减少 WCMP 组的权重时，该方案针对两个不同的目标之一进行优化：(1) 给定最大超额订购上限，以最大限度地减少组条目数；(2) 限制组的总条目数，以减少最大超额订购. 这两类优化采用了贪心算法来实现，并能保证一定的可靠性. 该方案通过底层路由协议来报告交换机或链路故障以更新权重，因此开销较小，可以在几毫秒内完成权重更新.

3.3.2　路径探测策略

由于交换机对数据包的五元组执行静态散列确定下一跳转发地址，因此可预先更改包头五元组元素间接为数据包指定路径. 文献[53]提出了一种在虚拟边缘实现负载均衡的方案 CLOVE. 该方案部署于源终端主机管理程序，方案执行流程如图 11 所示，首先通过探针数据包执行路径探测来获取所有路径信息，并基于拥塞状况为路径分配权重. 在终端主机处将流划分为流切片，然后更改每一个流切片中数据包的源端口信息(五元组的元素)来为流切片间接分配路径.

在路径探测中，源主机管理程序先向目的主机发送多条探针流，并为每条探针流随机分配源端口，进而各探针流被转发到不同的路径上. 每条探针流

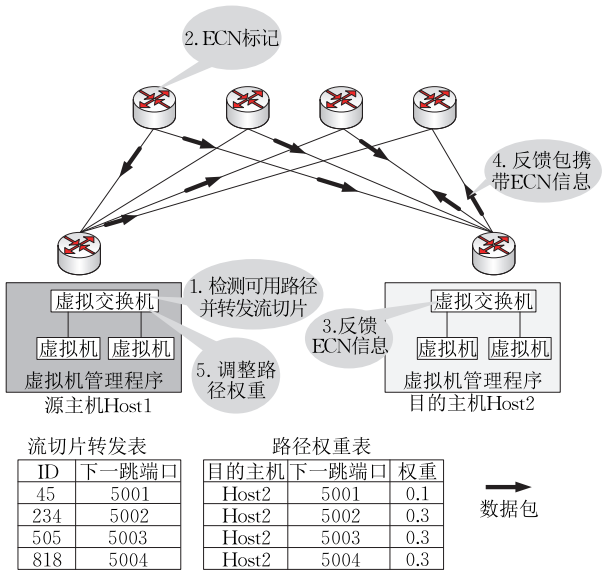

图 11　CLOVE 方案的拥塞信息反馈示意图[53]

由多个数据包组成，各数据包的五元组相同，但 TTL 字段值递增. 由这些数据包组成的一条探针流被转发到网络内，则源主机可以获得该探针流所在路径上每一跳交换机的 IP 地址列表，进而完成路径探测.

基于 ECN 反馈，对探测得到的路径进行权重赋值. 在源主机管理程序中，根据路径权重表，为每个流切片分配路径. 路径分配通过更改流切片中数据包的源端口信息来实现，相当于指定每个流切片中的数据包在物理网络中的路径，进而实现对整条流的负载均衡.

HULA 方案[54]利用探针数据包来检测网络中所有可用路径，并汇总、存储到每一个交换机上，以供决策下一跳地址. 每个交换机都会更新路径探测结果以及最佳下游路径，并将其发送到其他上游交换机. 同时每个交换机只决策下一跳转发地址，交换机只需维护每个目的地址对应的最佳路径.

探针数据包在栈顶交换机生成，每个交换机都会复制，复制机制由控制平面多播组控制. 利用探针数据包可以向所有交换机传播链路利用率信息，当探针数据包到达交换机入端口，交换机也会更新相应的路径信息，有助于及时发现和适应拓扑的变化.

Hermes 方案[55]通过对路径状态的检测结果决定是否为拥塞路径上的流进行重路由. 路径状态检测分为静态触发和主动探测两部分，(1) 静态触发部分. 联合 RTT 和 ECN 的反馈信息来将路径拥塞情况分为三级，并基于重传和超时来检测网络中的故障；(2) 主动探测部分. 使用探针来进一步检测网

络链路拥塞情况，为降低探针开销，源端随机检测两条等价路径，以及重新检测上一时间段拥塞程度最低的等价路径. 静态触发和主动探测的结合保证了对网络拓扑的可视性，降低开销的同时还能高效检测出网络故障，能很好应对网络拓扑非对称性.

该方案并非对所有的流均进行重路由，若某条流的当前路径状态检测结果为故障状态，则进行重路由，并优先为该流分配一条低拥塞级的路径. 若某条流的当前路径状态检测结果为拥塞状态，则联合当前发送速率、重路由后发送速率的估计值以及剩余字节数对重路由的性能提升进行评估，若有较大提升，才进行重路由操作.

Luopan 方案[56]在终端主机缓冲区将数据流分割为固定 64 KB 大小作为调度单元. 该方案首先通过抽样一些等价路径，直接把流切片调度到轻拥塞路径上以实现网络负载均衡. 通过拥塞感知的流调度转发方式提高了流完成时间，相比于传统 ECMP 方案，在非对称拓扑下，该方案具有更强的鲁棒性.

每个 TOR 交换机会周期性地发送一些探测数据包到目的 TOR 交换机，通过抽样一些随机路径，然后把路径信息存储在 TOR 交换机中的路径路由表中. 在实际部署大规模数据中心网络时，相比于目前的全局拥塞感知方案，该方案通过抽样处理的方式降低了网络存储全局拥塞信息带来的开销. 因此该方案采用抽样处理的方式简化了网络协议，相比于目前非对称方案具有更好的伸缩性优势.

该方案通过配置源 TOR 交换机并定期发送 d 个探测包到目的 TOR 交换机，目的是为了抽样 d 条随机的等价多路径，并将其路径信息保存在 TOR 交换机路径信息表中. 当一个源探测包发送到目的交换机时，首先遍历所在路径上所有交换机相应出端口的队列长度信息(队列长度信息定义为路径拥塞指标值)，并将其添加到源探测包头文件信息中. 当目的交换机接收到源探测包信息后，会立即生成一个 ACK 探测包，并将源探测包的包头文件信息拷贝到 ACK 探测包中，然后立即将其发送给源 TOR 交换机. 源 TOR 交换机接收到 ACK 探测包时，解析检索出目的交换机 IP 地址，路径 ID 以及路径的拥塞指标值等信息，并将该路径的拥塞指标值作为拥塞条目信息更新到源交换机中拥塞信息表中.

前文分析了 DRB 方案[41]应对哈希碰撞的解决方法，本节分析该方案如何解决链路故障导致的非对称性问题. 该方案采用拓扑更新协议让所有主机都掌握网络拓扑信息以及跳变交换机的 IP 地址信息. 终端主机获得网络故障信息后，路由将不使用故障链路所在的路径以及所对应跳变交换机. 拓扑更新协议采用以下三条原则来降低网络中的拓扑信息广播数据包数量：(1) 交换机定期或在检测到故障时广播拓扑更新信息，终端主机仅被动接收广播信息，不产生或转发广播消息；(2) 当交换机从其上行链路端口接收广播消息时，将消息转发到其下行链路端口；(3) 当栈顶交换机检测到上行链路故障时，仅将消息广播给所属终端主机.

DRB 在应对链路故障时将故障链路所涉及到的整条路径均排除在路由选择之外，而链路故障可分为部分链路故障和完全链路故障，部分链路故障时，该链路仍有部分带宽可用，因此完全弃用部分故障的链路会浪费网络资源.

3.3.3 拥塞感知策略

CONGA 方案[57]是一种流切片级的分布式负载均衡方案，实际部署于两层的叶脊网络拓扑，通过叶子交换机之间的反馈来收集路径上的实时拥塞状态信息，根据逐跳拥塞信息来计算最优的路径.

叶子交换机需要维护一张量化拥塞程度的二维表，以目的端输出端口为坐标，权值由源叶子交换机出口到目的叶子交换机的路径上的最大拥塞值决定，采用 DRE(Discounting Rate Estimator)调度转发，方案执行流程如图 12 所示.

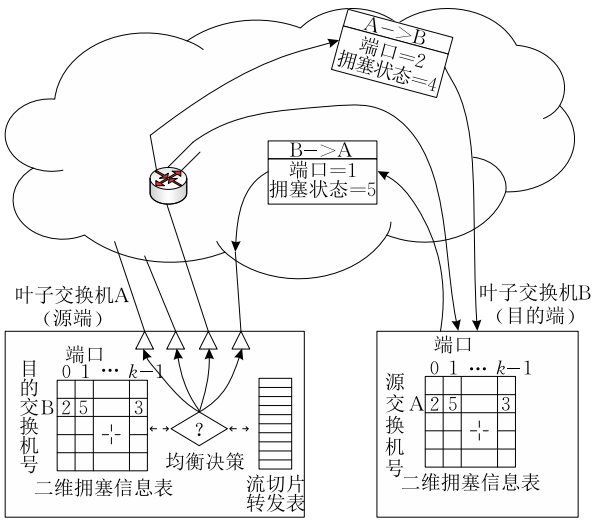

图 12 CONGA 系统框架示意图[57]

当源叶子交换机需要发送流给目的叶子交换机时，源叶子交换机根据拥塞信息表选择一条拥塞程度最小的路径发送. 路径的每一跳链路的拥塞信息

被记录在数据包报头字段中，因此路由过程中可实时更新路径上拥塞程度的最大值. 由于采用了往返捎带转发的方式进行反馈，目的端可能不会马上有数据发送回源端，源端的拥塞信息表更新存在滞后问题.

当目的叶子交换机有数据包传送回源端交换机时，两者的身份互换. 同样先根据本交换机的拥塞信息表选择一条最优的路径，同时携带需要反馈的拥塞信息. 由于两次路由的时间不同，路径拥塞状态情况可能不同，因此，选取的最佳路径也可能不同，拥塞信息在新选取的路径上仍继续更新.

CONGA 方案能够在 RTT 时间量级上对路径拥塞情况做出响应，缓解了数据中心流量不稳定性和突发性带来的影响. 该方案未采用集中式控制，但也能利用全局信息在链路发生拥塞时灵活地改变流量路径，调度周期可达到微秒级，能够很好地处理链路下游拥塞的情况. 但该方案也存在以下不足：(1) 拥塞信息二维表的更新存在滞后，因此该方案是基于原始状态的拥塞感知，而且滞后的时间不确定；(2) 成本开销太大，需要定制芯片；(3) 每个叶子交换机需要维护的数据随着规模的扩大越来越多，交换机内存将限制网络拓展性.

LetFlow 方案[58] 是流切片级别的负载均衡方案，研究者认为流切片级别处理粒度具备较好的弹性. 当流切片被转发到拥塞路径后，由于拥塞控制协议会对路径拥塞情况做出反应，源端拥塞窗口减小、发送速率减缓，则一段时间内到达网络的数据包较少，数据包与数据包之间的时间间隔会扩大，形成流突发，则后续流突发内的数据包会被切分成一个新的流切片. 该方案在交换机处为新的流切片随机分配一个候选输出端口，很大概率上会区别于上一流切片的转发路径，进而新进入网络的负载被转移到其他并行路径上. 该方案利用 TCP 拥塞控制协议来自适应切分流切片，从而达到自适应调谐负载均衡的效果，能很好应对网络拓扑非对称性问题.

LetFlow 方案不需要掌握全局的拥塞信息，相比于 CONGA 等方案，可扩展性更高. 但是由于该方案对流切片转发的随机特性，且流切片字节大小基于路径过去时间段的拥塞情况而变化，该方案无法阻止链路上短时间内的流量不均，从而增加链路的排队延迟，缺乏及时性问题.

此外，不同于 Flare 方案中周期性更新切分的时间阈值，CONGA 方案、LetFlow 方案均基于固定时间周期来检测流切片，虽然仅利用两比特就可实现流切片的划分，节省了交换机硬件开销，但也存在两个主要缺陷：(1) 若流突发的时间间隔在一倍检测周期与两倍检测周期之间，则可能无法被检测识别，将导致新的流突发被默认划分到上一个流切片；(2) 实际数据中心网络内流量负载随时间不断变化，等价路径间的最大时延差也在不断变化，固定的检测周期未必能与实际负载情况相匹配，进而划分后的流切片之间的时间间隔未必与等价路径间的最大时延差相匹配，会诱发严重的数据包乱序.

4　发展趋势与展望

随着云计算等信息技术的发展，未来全球数据中心 IP 总流量将达到几十泽字节，而数据中心内部的流量就占据着超过七成，因此实现数据中心网络内部的负载均衡具有重要意义.

本文调研并分析了现有的数据中心网络负载均衡方案，在表 2 中从处理粒度、决策方式、链路状态感知等维度上对各方案进行了对比，本节基于以上对比维度总结并展望了未来数据中心网络负载均衡技术的发展趋势：

(1) 处理粒度趋向于细粒度级

ECMP 的处理粒度为流级别，静态决策时既不考虑流量特征也不考虑链路状态，转发象流时处理粒度较大会带来哈希碰撞和鼠流阻塞问题.

若不考虑流量特征，仍维持流级别的处理粒度，仅考虑链路状态则不足以显著提升负载均衡效果. 这是因为，链路状态的感知范围为本地或局部时，负载均衡的决策结果非最优. 链路状态的感知范围为全局时，虽然能得出最优的负载均衡决策结果，但决策周期较长，无法满足鼠流的低时延需求，一般仅对象流决策，而象流字节量较大，在网络中的存在时间较长，该最优决策结果对于象流未必能长久保持有效性.

为进一步优化负载均衡效果，在处理粒度上逐渐向细粒度级过渡. 数据包级别是最精细的处理粒度，理论上能实现趋于最优的负载均衡效果，但在实际应用中会引发严重的数据包乱序，因此如何有效解决数据包乱序问题是未来的研究方向之一.

流切片级别的处理粒度是很好的折中方案，在切分流切片时只需保证相邻流切片之间的时间间隔不小于等价路径间的最大时延差，则既能优化负载均衡效果又能避免数据包乱序问题. 现有方案中多基于固定时间阈值来切分流切片，虽节省交换机硬件开销，但难以适应于实际网络中动态变化的流量

表 2 数据中心 Clos 网络负载均衡方案的综合比较

方案名称(年份)	处理粒度	决策方式	链路状态感知范围	突出特点
Hedera (2010)	流级别	集中式	全局	在交换机处检测象流,估算流带宽需求,动态重路由象流
Mahout (2011)	流级别	集中式	全局	在终端主机处检测象流,仅特定为象流分配路径
WCMP (2014)	流级别	集中式	全局	按链路可用容量对链路进行权重赋值
SCAA (2017)	流级别	集中式	全局	动态调整链路权重,优先使用权重小的链路
Freeway (2014)	流级别	集中式	全局	将网络路径动态划分为低时延路径集和高吞吐路径集
FlowBender (2014)	流级别	分布式	局部	重路由遇到拥塞或预期遇到拥塞的流
Hermes (2017)	流级别	分布式	局部	先判别重路由代价,再决定是否为拥塞路径上的流进行重路由;并通过路径探测感知故障
DiffFlow (2016)	流、数据包级别混合	集中式	全局	对象流和鼠流区分调度
CAPS (2018)	流、数据包级别混合	分布式	\	在发端对鼠流数据包冗余编码,在网内采用 RPS 转发,在收端通过冗余编码包恢复失序数据包
RPS (2013)	数据包级别	分布式	\	将数据包随机转发到任意等价路径上
DRILL (2017)	数据包级别	分布式	本地	根据本地交换机的队列占用情况来决策
DRB (2013)	数据包级别	分布式	\	为数据包跳变性选择路径,并基于链路故障的广播信息避开故障路径
DeTail (2013)	数据包级别	分布式	本地	基于优先级流控构建无丢包传输,数据包级自适应负载均衡无乱序
SAPS (2017)	数据包级别	集中式	全局	将实际不对称的拓扑划分成多个对称的虚拟拓扑
Presto (2015)	流切片级别	集中式	\	将所有流转变成统一 64 KB 大小的流切片,基于轮询选择路径
Luopan (2019)	流切片级别	分布式	局部	通过抽样路径,直接把流切片调度到轻拥塞路径上
TinyFlow (2014)	流、流切片级别混合	分布式	\	将象流分解成多条鼠流
CONGA (2014)	流切片级别	分布式	全局	根据网络实时拥塞状态信息来计算最优的分配路径
HULA (2016)	流切片级别	分布式	全局	使用探针探测网络路径并汇总至交换机处以供决策
CLOVE (2016)	流切片级别	分布式	全局	主机管理程序根据路径探测结果提前设定流切片的五元组信息
LetFlow (2017)	流切片级别	分布式	\	基于 TCP 拥塞控制机制对网络状况的反馈,自适应调整流切片大小,自适应均衡负载

注:在“链路状态感知范围”对比维度上,“全局”的感知范围是网络内全部路径下的链路,“局部”的感知范围是网络内部分路径下的链路,“本地”的感知范围是仅与当前决策节点直接连接的链路.

负载场景,因此,如何根据网络路径间的时延差异动态划分流切片也是未来的研究方向之一.

(2)决策方式倾向于分布式

集中式决策方式常用于基于全局链路状态感知的负载均衡,虽能做出最优决策,但仍存在一系列问题:① 基于 SDN/OpenFlow 技术的集中式决策方式的调度周期较长,而实际网络中流量负载随时间动态变化,无法保证决策结果的最优性在下一决策时刻前一致有效;② 集中控制器与交换机之间的简单静态连接关系会受到流量动态变化的影响,进而导致数据中心网络负载不均衡,因此需要集中控制器与交换机之间的动态连接方案[59-60],部署复杂度较高.

当前数据中心网络内流量负载中鼠流数目占比较高,鼠流在网络内的存在时间短,流完成时间在数个毫秒级,因此对负载均衡决策方式的及时性有一定要求.另外,细化处理粒度后也要求负载均衡方案能够快速响应,而集中式决策方式调度周期较长,难

以符合上述需求，分布式决策方式具备灵活度高的特点，能满足及时性需求，是极具潜力的部署方式.

当细化处理粒度后，对决策结果的有效性持续时间要求不高，适合采用分布式决策方式，因此在细粒度级下基于本地或局部链路状态感知的分布式负载均衡方案具备着较大的研究意义.

(3)均衡决策功能倾向于网内，辅助信息标记功能倾向于网络边缘

基于终端主机的方案试图弱化网络内的功能并将均衡决策功能部署至网络边缘，如 MPTCP[61-63] 在终端主机处将数据流划分成多条子流，并通过多个接口进入网络. 但鼠流一般在几个 RTT 内就能完成，若小流遭遇拥塞，上述方案无法及时响应并做出处理，因此将均衡决策功能保留在网内能够及时响应短时间内的负载不均.

针对 ECMP 静态决策时不考虑流量特征的不足，目前越来越多的方案将流量特征作为辅助信息以供更全面地负载均衡决策，将统计辅助信息的功能剥离至网络边缘，并通过数据包报头携带辅助信息至网内，能极大节省交换机的计算开销. 终端主机具备丰富的计算资源[64]，更容易获取流量的属性信息，相比于在网络中检测流量特征，在终端主机处检测的实时性更强[65]，如 Mahout 方案[37]. 如何在终端检测并高效识别流量特征以及如何低代价地将流量特征信息传递至网内交换机也具有极大研究价值，对负载均衡决策起着重要作用.

参　考　文　献

[1] Li Dan, Chen Gui-Hai, Ren Feng-Yuan, et al. Data center network research progress and trends. Chinese Journal of Computers, 2014, 37(2): 259-274(in Chinese)
(李丹，陈贵海，任丰原等. 数据中心网络的研究进展与趋势. 计算机学报，2014，37(2)：259-274)

[2] Yu Xiao-Shan, Wang Kun, Gu Hua-Xi, Wang Xi. The optical interconnection for cloud computing data centers: State of the art and future research. Chinese Journal of Computers, 2015, 38(10): 1924-1945(in Chinese)
(余晓杉，王琨，顾华玺，王曦. 云计算数据中心光互连网络：研究现状与趋势. 计算机学报，2015，38(10)：1924-1945)

[3] Miao R, Zeng H, Kim C, et al. SilkRoad: Making stateful layer-4 load balancing fast and cheap using switching ASICs//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. Los Angeles, USA, 2017: 15-28

[4] Olteanu V, Agache A, Voinescu A, et al. Stateless datacenter load-balancing with beamer//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI). Renton, USA, 2018: 125-139

[5] Chen K, Hu C, Zhang X, et al. Survey on routing in data centers: Insights and future directions. IEEE Network, 2011, 25(4): 6-10

[6] Benson T, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild//Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. Melbourne, Australia, 2010: 267-280

[7] Kandula S, Sengupta S, Greenberg A, et al. The nature of data center traffic: Measurements & analysis//Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. Chicago, USA, 2009: 202-208

[8] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture//Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication. Seattle, USA, 2008: 63-74

[9] Niranjan M R, Pamboris A, Farrington N, et al. PortLand: A scalable fault-tolerant layer 2 data center network fabric//Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication. Barcelona, Spain, 2009: 39-50

[10] Greenberg A, Hamilton J R, Jain N, et al. VL2: A scalable and flexible data center network//Proceedings of the Special Interest Group(SIGSOMM) on Data Communication. Barcelona, Spain, 2009: 95-104

[11] Guo C, Wu H, Tan K, et al. DCell: A scalable and fault-tolerant network structure for data centers//Proceedings of the Special Interest Group(SIGSOMM) on Data Communication. Seattle, USA, 2008: 75-86

[12] Guo C, Lu G, Li D, et al. BCube: A high performance, server-centric network architecture for modular data centers //Proceedings of the Special Interest Group(SIGSOMM) on Data Communication. Barcelona, Spain, 2009: 63-74

[13] Li D, Guo C, Wu H, et al. FiConn: Using backup port for server interconnection in data centers//Proceedings of the 28th IEEE International Conference on Computer Communications(INFOCOM). Rio de Janeiro, Brazil, 2009: 2276-2285

[14] Abu-Libdeh H, Costa P, Rowstron A, et al. Symbiotic routing in future data centers//Proceedings of the ACM Special Interest Group on Data Communication(SIGCOMM). New Delhi, India, 2010: 51-62

[15] Wu H, Lu G, Li D, et al. MDCube: A high performance network structure for modular data center interconnection//Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies. Rome, Italy, 2009: 25-36

[16] Li D, Xu M, Zhao H, et al. Building mega data center from heterogeneous containers//Proceedings of the 19th IEEE

International Conference on Network Protocols (ICNP). Vancouver, Canada, 2011: 256-265

[17] Greenberg A, Lahiri P, Maltz D A, et al. Towards a next generation data center architecture: Scalability and commoditization//Proceedings of the ACM Workshop on Programmable Routers for Extensible Services of Tomorrow. Seattle, USA, 2008: 57-62

[18] Hopps C E. Analysis of an equal-cost multi-path algorithm. Ann Arbor: NextHop Technologies, IETF RFC: 2992, 2000

[19] Thaler D, Hopps C E. Multipath issues in unicast and multicast next-hop selection. Redmond: Microsoft & Ann Arbor: NextHop Technologies, IETF RFC: 2991, 2000

[20] Moy J T. OSPF: Anatomy of an Internet Routing Protocol. 8th printing. Boston, USA: Addison-Wesley Professional, 2004

[21] Oran D R. OSI ISIS Intradomain routing protocol. Littleton: Digital Equipment Corporation, IETF RFC: 1142, 1990

[22] Shand M, Ginsberg L. Reclassification of RFC 1142 to Historic. Milpitas: Cisco Systems, IETF RFC: 7142, 2014

[23] Malkin G. RIP version 2. Billerica: Bay Networks, IETF RFC: 2453, 1998

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

[25] He K, Rozner E, Agarwal K, et al. Presto: Edge-based load balancing for fast datacenter networks//Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM). London, UK, 2015: 465-478

[26] Zhou Yu-Xin, Bao Wei-Dong. Review of load balancing scheme for data center network. Command Information System and Technology, 2018, 9(6): 6-12(in Chinese)
(周昱昕, 包卫东. 数据中心网络负载均衡方案综述. 指挥信息系统与技术, 2018, 9(6): 6-12)

[27] Noormohammadpour M, Raghavendra C S. Datacenter traffic control: understanding techniques and tradeoffs. IEEE Communications Surveys & Tutorials, 2017, 20(2): 1492-1525

[28] Xu H, Li B. TinyFlow: Breaking elephants down into mice in data center networks//Proceedings of the IEEE 20th International Workshop on Local & Metropolitan Area Networks (LANMAN). Reno, USA, 2014: 1-6

[29] Hu J, Huang J, Lv W, et al. CAPS: Coding-based adaptive packet spraying to reduce flow completion time in data center//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Honolulu, USA, 2018: 2294-2302

[30] Gill P, Jain N, Nagappan N. Understanding network failures in data centers: Measurement, analysis, and implications. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 350-361

[31] Liu V, Halperin D, Krishnamurthy A, et al. F10: A fault-tolerant engineered network//Proceedings of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI). Lombard, USA, 2013: 399-412

[32] Shelly N, Tschaen B, Förster K T, et al. Destroying networks for fun (and profit)//Proceedings of the 14th ACM Workshop on Hot Topics in Networks (HotNets). Philadelphia, USA, 2015: 6

[33] Liu J, Huang J, Li W, et al. AG: Adaptive switching granularity for load balancing with asymmetric topology in data center network//Proceedings of the 27th International Conference on Network Protocols (ICNP). Chicago, IL, USA, 2019: 1-11

[34] Zhang Y, Ansari N. On architecture design, congestion notification, TCP incast and power consumption in data centers. IEEE Communications Surveys & Tutorials, 2013, 15(1): 39-64

[35] Bensley S, Thaler D, Balasubramanian P, et al. Data Center TCP (DCTCP): TCP congestion control for data centers. Redmond: Microsoft, IETF RFC: 8257, 2017

[36] Al-Fares M, Radhakrishnan S, Raghavan B, et al. Hedera: Dynamic flow scheduling for data center networks//Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation (NSDI). San Jose, USA, 2010: 89-92

[37] Curtis A R, Kim W, Yalagandula P. Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection//Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM). Shanghai, China, 2011: 1629-1637

[38] Kabbani A, Vamanan B, Hasan J, et al. FlowBender: Flow-level adaptive routing for improved latency and throughput in datacenter networks//Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies (CoNEXT). Sydney, Australia, 2014: 149-160

[39] Shafiee M, Ghaderi J. A simple congestion-aware algorithm for load balancing in datacenter networks. IEEE/ACM Transactions on Networking, 2017, 25(6): 3670-3682

[40] Dixit A, Prakash P, Hu Y C, et al. On the impact of packet spraying in data center networks//Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM). Turin, Italy, 2013: 2130-2138

[41] Cao J, Xia R, Yang P, et al. Per-packet load-balanced, low-latency routing for Clos-based data center networks//Proceedings of the 9th ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT). Santa Barbara, USA, 2013: 49-60

[42] Carpio F, Engelmann A, Jukan A. DiffFlow: Differentiating short and long flows for load balancing in data center networks//Proceedings of the 59th Annual IEEE Global Communications Conference (GLOBECOM). Washington, USA, 2016: 1-6

[43] Correa J R, Goemans M X. Improved bounds on nonblocking 3-stage Clos networks. SIAM Journal on Computing, 2007, 37(3): 870-894

[44] Floyd S. TCP and explicit congestion notification. ACM SIGCOMM Computer Communication Review, 1994, 24(5): 8-23

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

[46] Wang W, Sun Y, Zheng K, et al. Freeway: Adaptively isolating the elephant and mice flows on different transmission paths//Proceedings of the IEEE 22nd International Conference on Network Protocols (ICNP). Raleigh, USA, 2014: 362-367

[47] Zats D, Das T, Mohan P, et al. DeTail: Reducing the flow completion time tail in datacenter networks//Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication. New York, USA, 2012: 139-150

[48] Bai W, Chen L, Chen K, et al. Information-agnostic flow scheduling for commodity data centers//Proceedings of the 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI). Santa Clara, USA, 2015: 455-468

[49] Wang T, Xu H, Liu F. Aemon: information-agnostic mix-flow scheduling in data center networks//Proceedings of the 1st Asia-Pacific Workshop on Networking (APNet). Hong Kong, China, 2017: 106-112

[50] Irteza S M, Bashir H M, Anwar T, et al. Load balancing over symmetric virtual topologies//Proceedings of the IEEE Conference on Computer Communications (INFOCOM). Atlanta, USA, 2017: 1-9

[51] Ghorbani S, Yang Z, Godfrey P, et al. DRILL: Micro load balancing for low-latency data center networks//Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM). Los Angeles, USA, 2017: 225-238

[52] Zhou J, Tewari M, Zhu M, et al. WCMP: Weighted cost multipathing for improved fairness in data centers//Proceedings of the 9th European Conference on Computer Systems (EuroSys). New York, USA, 2014: 5

[53] Katta N, Hira M, Ghag A, et al. CLOVE: How I learned to stop worrying about the core and love the edge//Proceedings of the 15th ACM Workshop on Hot Topics in Networks (HotNets). Atlanta, USA, 2016: 155-161

[54] Katta N, Hira M, Kim C, et al. HULA: Scalable load balancing using programmable data planes//Proceedings of the Symposium on SDN Research (SOSR). Santa Clara, USA, 2016: 10

[55] Zhang H, Zhang J, Bai W, et al. Resilient datacenter load balancing in the wild//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. Los Angeles, USA, 2017: 253-266

[56] Wang P, Trimponias G, Xu H, et al. Luopan: Sampling-based load balancing in data center networks. IEEE Transactions on Parallel and Distributed Systems, 2019, 30(1): 133-145

[57] Alizadeh M, Edsall T, Dharmapurikar S, et al. CONGA: Distributed congestion-aware load balancing for datacenters//Proceedings of the ACM SIGCOMM Computer Communication Review. Chicago, USA, 2014: 503-514

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

[59] Wang T, Liu F, Guo J, et al. Dynamic SDN controller assignment in data center networks: Stable matching with transfers//Proceedings of the 35th Annual IEEE International Conference on Computer Communications (INFOCOM). San Francisco, USA, 2016: 1-9

[60] Wang T, Liu F, Xu H. An efficient online algorithm for dynamic SDN controller assignment in data center networks. IEEE/ACM Transactions on Networking, 2017, 25(5): 2788-2801

[61] Ford A, Raiciu C, Handley M, et al. TCP extensions for multipath operation with multiple addresses. IETF RFC: 6824, 2013

[62] Raiciu C, Barre S, Pluntke C, et al. Improving datacenter performance and robustness with multipath TCP. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 266-277

[63] Kheirkhah M, Wakeman I, Parisis G. MMPTCP: A multipath transport protocol for data centers//Proceedings of the 35th Annual IEEE International Conference on Computer Communications (INFOCOM). San Francisco, USA, 2016: 1-9

[64] Wang T, Xu H, Liu F. Multi-resource load balancing for virtual network functions//Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). Atlanta, USA, 2017: 1322-1332

[65] Ballani H, Costa P, Gkantsidis C, et al. Enabling end-host network functions. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 493-507

DIAO Xing-Long, Ph. D. candidate. His research interests include load lancing, congestion control, data center network.

YU Xiao-Shan, Ph. D., lecturer. His current research interests include optical interconnects for cloud computing data center.

LIU Yong, Ph. D. candidate. His research interests include data center network, software defined networking.

HUANG Di-Tao, M. S. candidate. His research interests include congestion control, data center network.

GU Hua-Xi, Ph. D., professor, Ph. D. supervisor. His current research interests include networking technologies, network on chip, optical interconnect.

Background

With the rapid development of cloud computing and big data technology, data center has become the key infrastructure of information technology. Load balancing for the increasing traffic in data centers is of great significance. Various load balancing schemes have been discussed in both academia and industry.

This paper conducts a survey on load balancing schemes, including the problems induced by ECMP in practical situations, the improved load balancing solutions. The trend of load balancing is also discussed in this paper.

This work was supported in part by the National Key R&D Program of China under Grant No. 2018YFE0202800, and National Natural Science Foundation of China under Grant Nos. 61634004, 61934002 and 61901314, and the Natural Science Foundation of Shaanxi Province for Distinguished Young Scholars under Grant No. 2020JC-26, and the Fundamental Research Funds for the Central Universities under Grant No. JB190105, and State Key Laboratory of Computer Architecture (ICT, CAS) under Grant No. CARCH201919.