数据中心网络流量调度的研究进展与趋势

李文信[1)] 齐 恒[1)] 徐仁海[2)] 周晓波[2)] 李克秋[2)]

[1)](大连理工大学 计算机科学与技术学院 大连 116024)

[2)](天津大学 智能与计算学部 天津 300350)

摘 要 近年来,流量调度已经发展成为网络领域的热点研究问题.该问题主要决定何时以及以多大速率传输网络中的每条数据流,其对网络性能和应用性能都具有十分重要的影响.然而,在托管着许多大规模互联网应用的数据中心中,流量调度问题正面临着流量矩阵多变、流量种类混杂、以及流量突发等与流量模型相关的挑战.此外,随着数据中心规模的不断壮大,流量调度问题还面临着网络带宽动态化、网络拥塞随机化、以及网络目标多样化等与网络模型相关的挑战.为了进一步提升对数据中心流量调度的关注和理解,推动流调度技术在实际应用中的不断发展,本文分别从调度目标、调度方式和调度对象这三个维度对数据中心网络流调度的相关研究工作进行了分析和对比,并概括出如下结论:现有研究主要以分布式、集中式或混合式的调度方式对数据中心内、数据中心间或数据中心与用户间的流进行高效地调度,从而达到带宽保障、时限保障、最小化流完成时间、最小化Coflow完成时间、公平性保证、最小化流传输成本等目标.本文最后还指出了四个数据中心流调度的未来发展方向,并相应提出尚未解决的研究问题.

关键词 数据中心网络;流调度;Coflow调度;完成时间;优先级队列;带宽保障;时限保障

中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2020.00600

Data Center Network Flow Scheduling Progress and Trends

LI Wen-Xin[1)] QI Heng[1)] XU Ren-Hai[2)] ZHOU Xiao-Bo[2)] LI Ke-Qiu[2)]

[1)](*School of Computer Science and Technology, Dalian University of Technology, Dalian* 116024)

[2)](*College of Intelligence and Computing, Tianjin University, Tianjin* 300350)

Abstract Modern data center serves as the underlying infrastructure for many applications, including online internet services, data-parallel computing, machine learning, and cloud computing. A common denominator of these applications or services is that they will generate massive amounts of data flows in the network. From the perspectives of both the network operator and the applications/users, data center networks must be utilized effectively and efficiently. Flow scheduling is a promising technique to enhance the performance of the datacenter network, and hence has recently gained much research interest. Flow scheduling mainly determines when and at what rate to send each flow in the network, such that the desired objectives (e.g., minimum flow completion time (FCT), guarantee deadline) can be achieved. In this survey paper, we first illustrate the fundamental problems and challenges of scheduling flows in data center networks. It has the following two major challenges. First, the datacenter network flow model is complicated due to the dynamics, burst, and mixture in the traffic. Second, the network model is also full of complexity because of the dynamic network bandwidth, random network congestion, and diverse network optimization objectives. Bearing those challenges in

收稿日期:2019-06-18;在线出版日期:2020-02-07.本课题得到国家重点研发计划课题(2016YFB1000205)、国家自然科学基金重点项目(61432002)资助.**李文信**,博士,主要研究领域为数据中心网络、云计算.E-mail:toliwenxin@gmail.com.**齐 恒**(通信作者),博士,副教授,主要研究领域为软件定义网络、数据中心网络.E-mail:hengqi@dlut.edu.cn.**徐仁海**,硕士,主要研究领域为数据中心网络与云计算.**周晓波**,博士,副教授,主要研究领域为软件定义网络、数据中心网络.**李克秋**,博士,教授,主要研究领域为无线网络、云计算、软件定义数据中心网络.

mind, this paper compares and summarizes the related research work on data center flow scheduling from three dimensions: scheduling optimization goals, scheduling methods, and scheduling entities. At the level of scheduling optimization goal, existing work on flow scheduling can be divided into six categories: bandwidth guarantee, deadline guarantee, minimum FCT, minimum coflow completion time (CCT), fairness guarantee, minimum traffic transmission cost. At the level of scheduling method, existing work mainly falls into three kinds: distributed scheduling, centralized scheduling, and hybrid scheduling. At the level of scheduling entity, they can further be classified into three kinds: intra-datacenter flow scheduling, inter-datacenter flow scheduling, datacenter-client flow scheduling. Though many flow scheduling solutions have been proposed in existing work, most of them are still in the research stage, and are far from being adopted by the industry. The low complexity, low cost, and high-performance flow scheduling schemes need further exploration. Therefore, at the end of this paper, we point out four potential research directions of data center flow scheduling as well as the corresponding unresolved research problems involved in flow scheduling. First, most flow scheduling schemes rely on limited switch function (e. g., priority queues), while modern switches have much more flexibility to support more complex network functions due to its programmability. Hence, programming flow scheduling on switches is one potential research direction. Second, existing solutions need to hook packets in the end-host network stack to tag priorities in the packet header to perform flow scheduling, while such packet tagging incurs substantial overhead, making them inapplicable to high-speed networks. Since 40G and 100G or even 200G networks are coming, scheduling flows at such high-speed networks is another direction. Third, traditional model-based flow scheduling is sub-optimal; machine learning provides a new choice for high-efficiency flow scheduling. Hence, machine learning assisted flow scheduling is the third potential direction. Finally, as the geo-distributed machine learning and federated learning become important workloads, scheduling inter-datacenter flows (especially tiny flows) with security constraints to reduce FCT is also one of the potential directions.

Keywords　datacenter networks; flow scheduling; coflow scheduling; completion time; priority queues; bandwidth guarantee; deadline guarantee

1　引　言

近十几年来,数据中心已经发展成为当今互联网信息化建设的基础设施. 如图1所示,数据中心网络由三层或两层交换机或路由器互联而成[1-2]. 三层架构由接入层交换机,汇聚层交换机及核心层交换机构成;两层架构仅包含接入层交换机和核心层交换机,两层架构能互联5000到8000台服务器,而三层架构则能支撑更多数量的服务器. 此外,为了给不同地域的用户提供低延迟高可靠性的服务,大型互联网公司在世界各地部署数据中心. 例如,Google[3]在世界各地拥有超过30个数据中心,这些数据中心内的服务器数量高达45万台.

随着规模的壮大,数据中心也承载了许多种类繁多的在线服务和应用,包括视频点播应用、存储和文件共享、Web搜索、社交网络、云计算、金融服务、推荐系统及交互式在线工具等等. 这些服务和应用可以根据需要在数据中心中动态扩展,从而为服务提供商节省成本. 并且,服务提供商还可以将多种服务和应用部署在同一数据中心中,进而实现更好的资源利用. 此外,为了节约构建和维护数据中心的成本,它们也可以依赖大型云公司来构建服务和应用,如亚马逊EC2云平台.

托管在数据中心之上的应用会产生巨大的流量. 一方面,在数据中心内,传统"南北流量①"模式逐渐被"东西流量②"所取代. Cisco③的报告表明,截

① 南北流量指的是数据中心外部用户和内部服务器之间交互的流量.

② 东西流量指的是数据中心内部服务器之间交互的流量.

③ Cisco Global Cloud Index: Forecast and Methodology. http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf

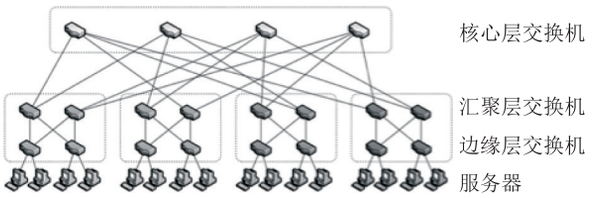

图1 胖树数据中心网络拓扑结构

至2020年，数据中心内“东西流量”将比“南北流量”的3倍还多. 另一方面，数据中心与数据中心之间的流量也急剧增长. 据统计，数据中心间的流量每月能够达到330 TB至330 PB[①]，这给数据中心供应商产生了数以亿计的经济成本[4].

在这种背景下，流量调度成为数据中心中的热点研究问题，其主要指何时以及以多大速率传输每一条网络数据流的问题. 该问题对数据中心供应商和用户都具有重要意义. 一方面，如果网络数据流低速运行，链路带宽未被占满，造成资源浪费，这终将影响数据中心供应商的收益；另一方面，用户应用产生的网络数据流如果没有及时地被调度，那么用户服务质量将会受到很大的影响，这反过来还会进一步影响供应商的收益. 例如，在Google搜索服务中，每增加400毫秒的延迟会使得用户搜索量降低6%[②]；在亚马逊服务中，增加100毫秒的延迟也会使收益降低1%[③]. 此外，随着流量的与日俱增，流量调度的作用和地位还将进一步被凸显. 因此，大量流调度相关的研究工作相继涌现.

本文主要介绍数据中心流量调度问题的研究现状. 首先，介绍流量调度问题的本质及其面临的挑战. 然后，本文从不同维度对现有研究进行分类总结. 具体来讲，按照不同的调度优化目标，现有研究工作可以分为六类，包括带宽保障、截止时间保障、最小化流完成时间(Flow Completion Time，FCT)、最小化Coflow完成时间(Coflow Completion Time，CCT)、最小化流量传输成本、公平等优化目标. 根据不同的调度管理方式，现有研究又可分为分布式调度、集中式调度以及混合式调度这三类. 最后，根据不同的调度对象，现有研究又可分为面向数据中心内流量的调度、面向数据中心间流量的调度以及面向用户与数据中心之间的流量的调度这三类. 最后，本文展望了数据中心流量调度问题的未来发展趋势与方向. 总体而言，本文试图在数据中心网络流量调度的背景下，对现有研究工作的设计原则、目标和特点进行深入地研究. 通过讨论它们的优缺点、相似性和差异性，我们对数据中心网络中的带宽流量调度给出了一个有见地的概述，以期望在未来激励更好的调度设计方案.

2 问题与挑战

本节首先阐述数据中心流量调度问题的本质及含义，然后分析该问题所面临的挑战.

2.1 流量调度问题

给定一个数据中心网络和多条网络数据流. 每条流从数据中心中的某一台服务器发出，终止于另一台服务器. 流量调度主要决定何时在网络中传输每条流的每个数据包，从而实现不同的优化目标，如最小化流的平均FCT和为流提供时限保障等.

在单条链路场景中，若以最小化平均FCT为目标，那么最优的流调度策略是最短流优先. 这里用一个示例进行说明. 如图2所示，假设有一条链路和三条流：A、B和C. 链路的带宽容量假设为1. 流A、B和C的数据量分别为1，2和3. 假设所有流在同一时刻到达. 默认采用公平共享策略，即公平地将链路带宽分配给并发传输的流. 如图2左侧子图所示，在公平共享策略下，流A、B和C分别在第3，第5，和第6个时刻完成，因此流的平均FCT为4.67. 另一方面，如果采用最短流优先策略进行调度，那么流的平均FCT就能够被降低到3.33，如图2右侧子图所示.

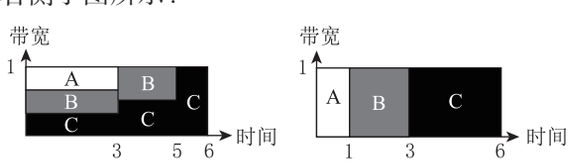

图2 单条链路下的公平共享和短流优先调度策略

以上示例仅包含一条链路. 然而，数据中心网络有多条链路. 在任何一个时刻，任何一条链路都可能发生拥塞，这使得多链路的流量调度成为NP难问题. 针对以平均流FCT最小化为目标的多链路流调度问题，最短流优先策略在理论上被证明与最优解最接近，其近似度为2.

2.2 面临的挑战

在实际生产环境中，多链路流量调度问题的理论结果还是很难获取，这主要归结于两方面的挑战：流量模型复杂和网络模型复杂. 表1总结了流量调

① The future of datacenter wide-area networking. http://www.forrester.com.

② Speed matters for Google web search. https://services.google.com/fh/files/blogs/google_delayexp.pdf

③ http://sites.google.com/site/glinden/Home/StanfordDataMining.2006-11-28.ppt

表1　流量调度所面临的挑战

挑战		描述	影响
流量模型复杂	流量矩阵多变	流量矩阵代表着数据中心中各个主机对之间通信的数据量,其呈现出高度动态化的特性.	加大了流调度的复杂度
	流量种类混杂	由于托管着多种类型的应用,数据中心网络中也充斥着各种类型的流量,包括延迟敏感的流、吞吐率敏感的流、数据大小已知/未知的流,截止时间敏感的流、以及Coflow等.	对流调度提出了更高需求,因为要满足不同类型流的不同需求.
	流量突发严重	流量突发指短时间内流量突增的现象,其包含两方面:单条流的流量突发和到达流突发. 在分布式机器学习应用下,流量突发会显得更严重.	流量突发会造成交换机队列堆积现象,严重时还将造成大量丢包
网络模型复杂	网络带宽动态化	数据中心链路由很多条流共享,每条流观测到的可用带宽动态变化. 在数据中心间的网络中,这种带宽动态变化现象更严重.	对快速流调度决策提出了挑战
	网络拥塞随机化	任何时刻任何交换机都可能发生队列拥塞现象.	队列拥塞增加数据包的排队延迟,拖慢数据流的完成
	网络目标多样化	网络优化目标多样化,包括降低流的完成时间,提高资源利用率,降低流量成本等等.	多样化的优化目标增大了流量调度的困难

度问题所面临的挑战.

2.2.1　流量模型复杂

(1)流量矩阵多变. 数据中心网络中大部分的网络流都是小流. 小流一般由极少个数据包组成,甚至很多小流只有一个数据包[5]. 而数据中心网络中绝大多数流量是由大流所造成. 除了大小流现象,数据中心网络中流量还存在高并发特性. 据统计,数据中心中单个机器每秒产生100到500条流[6],单个交换机中每秒到达的流的个数能够达到10000条[7],单个集群中每秒能够产生100000条流[7]. 流量矩阵表示网络中各个主机对之间通信的数据量. 在数据中心网络高并发流量且大部分流量是小流的环境下,流量矩阵也呈现出高度动态化的特点,这对流量调度提出了巨大的挑战.

(2)流量种类混杂. 数据中心网络流量种类繁多. 知晓不同流量的需求和特征有助于优化流量调度的设计. 按照需求来划分,数据中心流量主要包含延迟敏感的流、吞吐率敏感的流. 延迟敏感流一般由交互式应用产生,如Web搜索. 这种流一般都是短流,大小在几十KB到1MB之间[8]. 延迟敏感流对在线服务至关重要,因为其直接影响服务响应度,进而影响用户量和服务提供商的收益. 根据相关研究表明,在Google服务中,每增加400 ms的延迟将降低6%的用户搜索量①;在Amazon服务中,增加100 ms的延迟也会使收益降低1%②. 吞吐率敏感流对延迟没有很高的要求,但是它们需要很多带宽. 在数据中心网络中,吞吐量敏感的流一般由并行计算作业所产生,如MapReduce,其大小一般在1 MB到100 MB之间. 此外,它们还可以由跨数据中心传输的数据备份流所产生,其大小能够达到GB量级[9]. 虽然对延迟要求没有那么高,但是加快它们的传输也能提升应用的性能. 在有些时候,以上两种类型的流还会有时限需求. 时限表明网络流需要在指定的截止时间前完成. 时限分为软时限和硬时限[10]. 软时限表明流在截止时间后完成对应用仍然存在价值,不过这种价值会随流的FCT离截止时间越远而越小. 相反,硬时限则表明超过截止时间还未完成的流将对应用毫无价值. 另一方面,有些流在传输之前就能知晓其数据量,而有些流则只有在传输完成后才能知晓其准确的数据大小[11]. 例如,在虚拟机迁移和数据备份应用中,流的大小在传输之前就已知;而对于数据库存取和HTTP分块传输等应用,流的大小无法事先被知晓. 最后,从应用的角度出发,单个流的传输性能的提升不一定会给应用性能带来显著提升,在这种背景下,数据流组Coflow[12]的概念应运而生. Coflow主要用于描述一个并行计算作业的两个计算阶段间传输的流的集合. 在一个Coflow中,仅当其内部所有的流传输完成,该Coflow才可被视为结束. 因此,当数据中心网络中包含所有延迟敏感、吞吐量敏感、截止时间敏感、数据大小信息已知或未知、以及Coflow等多种流的时候,流量调度将会变得更加复杂.

(3)流量突发严重. 数据中心流量突发现象严重. 传统硬件卸载、中断调节以及TCP慢启动等均

① Speed matters for Google web search. https://services.google.com/fh/files/blogs/google_delayexp.pdf

② http://sites.google.com/site/glinden/Home/StanfordDataMining.2006-11-28.ppt

有可能引起流量突发现象. 另外,上层应用也有可能一次性发送大量的数据,进而导致流量突发. 例如,在分布式机器学习应用中,算法通常需要经历多轮迭代达到收敛. 在多轮迭代的训练过程中,流量往往会在每一步迭代步骤结束后产生,这就造成突发流量现象[13]. 在突发流量情况下,应用性能被大打折扣. 一方面,突发流量会造成中间网络设备拥堵,并形成队列堆积,进而使得网络流需要经历很长的排队延迟. 另一方面,队列堆积过多,引发丢包. 这种情况下,现有协议会触发频繁地丢包重传,这不仅浪费带宽,还会加剧网络的负担.

2.2.2 网络模型复杂

数据中心网络模型复杂,主要体现在网络带宽动态化、网络拥塞随机化以及网络目标多样化:

(1)网络带宽动态化. 在单个数据中心中,链路带宽容量固定不变. 但是,在每条流发送数据前,其传输路径中链路可用带宽是随时间变化而动态变化的,因为其传出路径中的每条链路会由很多其它的网络流共享,且其它的这些流量又呈现出高度动态化的特性. 因此,很多被动式拥塞控制策略会慢慢地探测网络中可用带宽,如DCTCP[8]. 另一方面,在数据中心间的网络中,可用带宽也是动态变化的. 据文献[14-15]的测量结果表明,数据中心间的网络可用带宽最小仅有30 Mbps左右,而最大可以达到500 Mbps以上. 在带宽动态变化的场景下,当前流量调度决策有可能在部署到实际环境后就已经失效. 这主要是因为大多调度算法基于当前信息进行,而调度决策的计算是需要耗时的,在调度决策计算的这段时间内,可用带宽有可能已经发生变化.

(2)网络拥塞随机化. 在数据中心网络中,任何主机对之间都存在多条路径. 尤其是在leaf-spine这样的网络拓扑下,这些路径距离或跳数一样. 因此,ECMP(Equal-Cost Multi-Path)等价多路径路由被广泛地用于流级别的负载均衡中,其通常基于数据包包头五元组的哈希值为网络流选择路劲. 正是由于流量和ECMP的随机性,加上数据中心网络交换机浅缓存的现状,导致交换机队列会随机地发生拥塞. 换句话说,也就是在任何时候任何交换机都有可能发生队列拥塞. 据文献[16]的测量结果表明,任何时候都存在10%的交换机队列发生拥塞. 队列拥塞会使得数据包经历很长的排队延迟,增加了流的完成时间. 在数据中心间的网络中,拥塞随机化现象更为严重. 数据中心间的RTT往返时延能够达到100毫秒量级,而数据中心内RTT仅有几百个微秒. 在网络拥塞随机变化场景下,流量调度做决策时很难捕捉准确的拥塞信号,进而导致最优解无法被获取.

(3)网络优化目标多样化. 一般而言,数据中心运营商希望其网络能够承载不同的功能. 数据中心网络的最主要的功能就是降低流的传输时间,从而提升上层应用的性能. 除此之外,提高带宽资源利用率也是运营商比较关心的目标. 在某些应用环境下,如股票交易,网络还需要确保零丢包的目标. 在数据中心间的网络中,带宽十分稀缺,而且带宽大都是从Internet服务供应商(Internet Service Provider,ISP)处通过付费租用方式来获取的. 这就给运营商造成了大量的带宽成本,每年大约花费数亿美元. 在这种情况下,降低数据中心间带宽成本对运营商显得极为重要. 多样化的网络优化目标对流量调度提出了挑战,而且不同的优化目标还会相互排斥. 例如,在数据中心间的网络中,降低带宽成本会使拖慢流的完成,影响应用性能.

3 研究现状分析

本节从不同维度对现有数据中心流量调度研究工作进行分类对比和总结. 具体如图3所示.

3.1 调度目标

如表2所示,根据调度目标的不同,现有流量调度相关研究工作可划分为以下6类:

表2 数据中心流调度优化目标概述

调度优化目标	描述
带宽保障	为网络流预留或分配带宽,实现云租户间性能隔离,并为云租户应用的网络性能提供下限保障.
时限保障	在截止时间前完成数据流的传输,能够提升应用的时效性.
最小化FCT	降低流完成时间能够为应用提供低延迟,尤其是对运行在数据中心之上的在线业务有很大的帮助.
最小化CCT	加快Coflow的传输对分布式并行计算应用完成有促进作用.
公平性	公平调度是保障云租户间公平资源竞争的有效措施.
最小化流量传输成本	数据中心间流量传输成本是数据中心供应商的一项重要开支,降低这种流量传输成本意义重大.

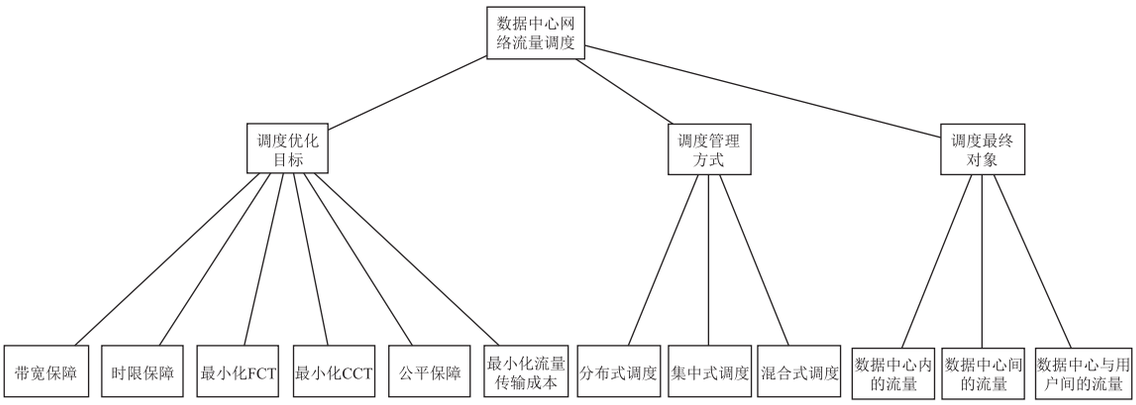

图3　数据中心网络流调度现有研究工作分类

3.1.1　带宽保障

在云数据中心中,为网络流提供带宽保障是实现云租户之间网络性能隔离的重要方法,因为它能给租户应用的网络性能提供下限保障. 在现有提供带宽保障的流调度研究可通过以下两种方式实现.

第一种方式是静态带宽预留. 这种方式一般通过虚拟机放置以及静态速率控制的手段来实现从虚拟网络模型到物理网络的映射,它能给用户提供可预测网络性能的保障,并且还能隔离云租户之间的网络性能. 目前主要有三种虚拟网络模型:Pipe模型[17]、Hose模型[18-21]以及TAG模型[22]. Pipe管道模型是由Guo等人[17]在其发表CoNext 2007的会议论文中首次提出. 基于Pipe模型,用户能够显式地表述其VM-pair之间的带宽需求. 与Pipe管道模型不同,Hose软管模型将所有虚拟机连接到一个大的虚拟交换机中. 虚拟机与虚拟交换机通过虚拟专用链路互连[18]. Ballani等人[19]基于Hose软管模型提出了虚拟集群及虚拟超额认购集群这两个模型,并利用虚拟机放置算法为用户提供确定的带宽保障. 但是,该算法没有考虑虚拟机差异化的带宽需求. 为此,Zhu等人[20]对Hose软管模型进行扩展,使其能够兼顾租户的差异化带宽需求. 最后一种模型是租户应用图TAG模型[21],其不仅能表示租户带宽需求,还能表示租户应用通信模式. 静态预留带宽的方式对具有确定带宽需求的用户有利,而对带宽需求量动态变化的用户,会造成带宽过保障或欠保障.

另外一种方式是提供最小或最低带宽保障. 与这类研究相关的现有工作主要分为两类. 第一类是在服务器端进行带宽分配. 例如,在Rodrigues等人[22]提出的Gatekeeper及Jeyakumar等人[23]提出的EyeQ方法中,用户最小带宽保障通过服务器端限速实现. 但是,它们假设数据中心核心链路不会发生拥塞,这种假设不甚合理. Guo等人[24-25]首先保障虚拟机最低带宽需求,然后将剩余网络带宽通过基于纳什博弈理论模型的方法公平地分配给虚拟机. 第二类是实现对全网链路的带宽分配. Liu等人[26]将每条链路中虚拟机的流速竞争过程抽象为逻辑斯蒂回归模型,然后通过设计分布式链路带宽分配方法,动态控制虚拟机流速. 但是,该方法需要确保虚拟机的带宽需求总量必须小于物理链路带宽. 为此,他们进一步提出SoftBW[27]方法,该方法通过网络定价实现对租户性能保障,并且其能在超约场景中为租户提供最低带宽保障. Popa等人[28]设计了FairCloud带宽分配方法,但是该方法具有局限性:一方面,它仅在树形网络拓扑下提供最小带宽保障,另一方面,它不易实现. Popa等人[29]还设计了名为ElasticSwitch的弹性带宽分配方法,其能够直接实现在虚拟机监视器hypervisor中. 但是,它不能同时实现带宽保障与资源的高效利用. 以上方法仅考虑租户内部的虚拟机通信. 其实,租户与租户之间也会有通信需求. 为此,Ballani等人[30]提出了分层软管模型,用以表示租户与租户之间通信的带宽需求以及通信依赖关系. 他们还设计了一个按权速率控制机制,从而实现租户间流量的带宽分配.

3.1.2　截止时间保障

对于数据中心中的许多应用,网络流有时限需求,也即,网络流必须要在指定的截止时间前完成. 这种截止时间敏感的流应该优先被调度. 例如,Wilson等人[31]提出了D^3方法,该方法采用显示速率控制技术进行带宽分配,它能使网络流刚好在截止时间前完成. 然而,由于D^3给流分配带宽过程遵守先进先出的顺序,在这种情况下,如果没有足够的带宽,势必会造成某些流的时限需求得不到满足.

Hong等人[32]提出了PDQ方法，该方法允许时限需求紧急的流抢占时限宽松的流的带宽. 但是，PDQ需要专用交换机的支持，在实际中不可行. Vamanan等人[33]提出D²TCP方法，该方法根据流的紧急程度和网络拥塞状态调整滑动窗口大小，使得紧急的流不易错过其截止时限. Chen等人[34]提出了MCP方法，该方法基于李雅普诺夫理论模型计算流的发送窗口调整函数，并采用ECN来近似控制窗口大小. MCP允许瞬时流速小于期望速率，但是长期平均流速必须大于等于期望速率，这样以保证流的时限需求.

另一方面，数据中心间的流也具有时限需求. 为此，Kandula等人[9]设计了TEMPUS方法，该方法通过求解一个混合packing-covering线性规划问题来调度具有时限需求的流. 但是TEMPUS不确保任何流能够完整地在其截止时间前完成，因为其只试图让每条流在截止时间前完成的比例最大化. Zhang等人[35]提出了Amoeba方法，从而给数据中心间的数据传输提供截止时间保障. Amoeba的核心是一个基于截止时间的网络抽象模型DNA. DNA模型允许用户明确表述其时限需求或待传输数据量需求. 基于该DNA模型，Amoeba为流分配其时限所需的最低带宽. Amoeba能够显著提升时限得到满足的流的比例，但与此同时，它需要用户显式地表达其时限需求，不可避免的存在用户恶意谎报需求以获取更多资源的现象. Jalaparti等人[36]发现很多用户都愿意为更好的服务付费，例如，更高优先级、更紧迫的时限等. 因此，他们将定价与流量工程结合起来，并提出了Pretium方法. 通过动态地调整链路带宽价格来引导用户对其服务的报价，Pretium能够为用户提供服务保障. 此外，它还能在流量高突发时，通过对链路高定价的措施来缓解网络拥堵的现象.

虽然时限保障重要，但是如果流在截止时间前就已经完成，会损伤延迟敏感流的FCT. 因此，混合类型流调度也是学术界比较关心的问题. 此处不再展开，具体可参考文献[11,37].

3.1.3 最小化FCT

对于延迟敏感的流，降低它们的FCT是网络优化的重要目标之一. 从理论层面来看，最短流优先(Shortest Flow First，SJF)在单链路多流场景中能够最小化平均FCT，但是其前提是所有流及其大小事先已知. 然而，在很多在线系统中，流可能在任意时刻达到. 因此，在这种场景下，最短剩余处理时间优先(Shortest Remaining Processing Time，SRPT)的调度方法被广泛用于减少平均FCT，但其仍然需要知道流的剩余数据量大小. 为此，最少被服务优先(Least Attained Service First，LAS)的调度方法被提出，该方法无需知晓流的大小信息就能近似地取得SJF的调度结果.

虽然以上调度方法在理论中表现较好，但它们受限于诸多实际因素. 例如，数据中心包含多条链路且交换机中优先级队列个数有限. 在这种情况下，Alizadeh等人[5]提出pFabric. 在数据包在发送之前，pFabric将对应流的剩余数据大小信息嵌入数据包包头. 交换机在接收到数据包后，按照剩余数据大小依次发送数据包，从而近似地实现SRPT调度策略. PDQ结合EDF(Earliest Deadline First)和SJF策略，以给截止时间敏感流高优先级，同时降低剩下的流的FCT. pFabric和PDQ都需要对交换机作较大改动，且它们都假设流的数据量大小事先已知，这在实际场景中不可行. 为此，Bai等人[38]提出PIAS方法，该方法无需事先知晓流的大小信息，就能够做到让小流优先于大流进行传输. PIAS的核心是一个多级反馈队列MLFQ(Multilevel-Feedback-Queue). PIAS首先将流放入最高优先级队列，并根据流的已发送数据量逐渐将其移动到低优先级队列中. PIAS仅关注数据大小不可知的流，而在数据中心网络中，有些网络流是可以事先知道数据大小的. 为此，Chen等人[11]提出了Karuna方法以在混合网络流场景中进行调度优化，使得不同类型流的需求都能够被满足. Munir等人[39]提出PASE方案，以将DCTCP、pFabric以及PDQ这三种方法进行融合，从而充分发挥出它们各自的优势. Li等人[40]发现现有基于SRPT的调度方法都假设流的剩余时间由剩余数据大小决定，他们认为这种假设是不合理的. 为此，他们利用从用户空间向内核空间的数据拷贝速率来估算数据生成的速率，并结合剩余数据大小来准确地估算一条流的剩余处理时间，从而实现更好的SRPT调度.

此外，对于许多延迟敏感的应用，例如Web搜索，尾延迟也是一个重要的度量指标，因为它决定了用户体验. 一种有效的降低尾流FCT的方法是复制流[41]，使用最快完成的流，并终止其余的未完成的副本流. 另一种方法是仅仅复制短流[42]，从而在快速响应和资源消耗中实现一个很好的权衡.

3.1.4 最小化CCT

随着数据并行计算应用往数据中心迁移，传统

的流模型已不再适用,因为加快单个流传输并不意味着应用性能就能因此而被提升.在这种情况下,针对并行计算作业的高效数据传输,Chowdhury等人[43]进行了一系列的探索和研究.2011年,他们设计了Orchestra系统来调度并行计算作业产生的网络数据流,Orchestra将主机端可建立的总的TCP连接数按照权重公平方法分配给每个应用.2012年,他们提出Coflow[12],以表示跨服务器运行的并行计算作业所涉及的通信阶段.具体来讲,一个Coflow就是一个并行计算作业中两个不同计算处理阶段之间的网络数据流的集合,例如MapReduce中的Shuffle.为了提升分布式并行计算应用的性能,应该优化Coflow的完成时间CCT,因为并行计算作业依赖于其中间的所有流的完成,而非单个流的完成.

Coflow抽象模型的提出开启了一个新的研究方向,即Coflow调度优化.大多数现有Coflow调度相关研究工作都以最小化CCT为优化目标.例如,Baraat[44]采用先进先出FIFO机制对Coflow进行调度,同时利用公平共享FS策略进行动态地带宽分配,从而避免队头阻塞现象以及减少Coflow的CCT.Chowdhury等人[45]提出了Varys方法,其首先利用SEBF最小有效瓶颈优先方法对Coflow进行调度,然后利用MADD方法为每个Coflow分配最少所需带宽.Varys的调度方法假设所有Coflow的大小已知.为了消除该假设,Aalo[46]方法被提出.Aalo根据Coflow已经发送数据量的大小来决定Coflow的优先级,并保证短Coflow优先于长Coflow完成.以上方法都假设网络数据流与Coflow的从属关系事先已知,而实际上这种从属关系需要对应用做较大改动才能够被获取.为此,Zhang等人[47]提出了CODA调度方法,该方法首先利用无监督机器学习方法(DBSCAN)来自动识别出网络流与Coflow的从属关系,然后还设计了识别出错情况下的错误可容忍的调度方法.Qiu等人[48]关注加权Coflow的平均完成时间,并为此设计了一个多项式时间的近似调度算法.在不考虑路由优化的情况下,仅仅依赖Coflow调度技术所能带来的优化空间十分有限.因此,Zhao等人[49]提出了RAPIER方法,该方法同时考虑Coflow调度和路由,其中路由决策由SDN控制器下发,调度决策由Linux限速实现.但是在Coflow的平均完成时间方面,RAPIER不能提供理论的性能保障.为此,Li等人[50]提出了OMCoflow,该方法在执行Coflow调度和路由时,能够保障其实现的平均CCT有确定的理论上限.

3.1.5 公平性

数据中心网络内运行着许多条流,它们共享数据中心链路带宽以及交换机缓存等资源,如何提供公平性是一个热点研究问题.在带宽分配,交换机队列排队,Coflow优化中都会涉及到公平性.

首先,在链路带宽方面,公平分配在一定程度上,能够为用户/应用/虚拟机/网络流提供性能隔离保障.例如,Brosoce等人[51]在其发表在SIGCOMM 2007的论文中先关注per-flow的公平.随后,他们进一步提出了ConEx模型①,该模型通过控制租户丢包数实现source-destination-pairwise级别的公平.另一方面,Lam等人[52]关注租户之间的公平性.他们设计了NetShare方法,其核心是控制租户在争用链路带宽的时所持有的权重.Shieh等人[53]提出了Seawall的方法,该方法在服务器主机之间建立具有拥塞控制功能的隧道,虚拟机之间的通信流量都要通过该隧道,该方法能够实现跨租户的最大最小公平性.FairCloud[28]中的PS-L方法可以实现租户级别的公平,该方法利用权重公平队列(Weighted Fair Queue, WFQ)机制将每一条链路的带宽分配给租户.Chen等人[54]重点关注并行计算应用间的公平性,他们提出以应用性能为导向的公平带宽分配方法,保证不同应用的网络传输的公平.总体来看,公平带宽分配有助于提高资源利用率,但是无法提供带宽保障.

其次,在交换机队列中,公平排队也得到了学术界的关注.2017年,Nagle[55]首次引入公平排队概念,其核心准则是:为了实现流级的公平排队,交换机中所有活跃的流都应该有相同优先级使用带宽.经典的公平排队方法包含Nagle[55]、BR[56](逐位循环法)以及AFQ[57].Nagle[55]是第一个公平排队工作,它将相同源数据包映射到特定的独立队列,并循环地为这些队列提供服务.实际上,Nagle的公平排队是基于每个源的公平性,并且可以很容易地被修改以支撑其它级别的公平.

最后,在Coflow优化中,Chen等人[58]认为不同Coflow应该被区分对待.为此,他们研究了面向Coflow的Max-Min公平.另一方面,Wang等人[59]同时考虑Coflow的性能及公平性,并提出了Coflex

① Briscoe B, Sridharan M. Network performance isolation in data centres using congestion policing. https://tools.ietf.org/id/draft-briscoe-conex-data-centre-01.html, 2013.

方法，该方法不仅能够有效地降低平均CCT，还能达到Coflow间的性能隔离．以上Coflow调度方法都假设Coflow大小信息已知．为此，Wang等人[60]提出NC-DRF方法，以在Coflow信息不可知场景下，减少平均CCT，同时提供Coflow间的性能隔离保障．

3.1.6 最小化流量传输成本

为了支撑跨数据中心传输的流量，供应商往往斥巨资从Internet服务供应商处租用带宽，每年耗费了数亿美元[4]．因此，服务供应商需要考虑的其中一个重要目标就是尽可能地降低跨数据中心流量的传输成本或带宽成本．另一方面，云平台供应商也为租户制定了不同的数据中心间流量传输定价策略，减少这种流量传输成本对云租户也是极为重要的．降低成本的一种方式就是尽可能的提高带宽资源利用率．例如，Laoutaris等人[61]关注跨数据中心的大块数据流，并相应地提出了NetSticher方法．该方法首先将数据存储在中间节点的数据中心中，待带宽充足时，再将其转发出去，从而在不影响其它流的情况下传输大块数据流，提升了带宽资源利用率．NetSticher能够提升带宽利用率的主要原因是不同数据中心在不同区域，导致正常业务流量对带宽的占用呈现出了很大的差异，这给资源优化带来了机会．NetSticher仅考虑单个文件传输的情况．Feng等人[62]针对不同数据中心的出入口带宽的单位成本存在差异的特点，设计了有效的路径规划及流调度算法，减少了跨数据中心的视频流的传输成本．在社交网络场景中，用户数据需要备份，也会引发跨数据中心流量．为此，Liu等人[63]提出了SD3方法，该方法根据好友访问频率以及个人主页更新的速度来进行数据备份，从而减少数据中心间流量．SD3方法专用于社交网络应用，通用性不高．Li等人[64]提出了TrafficShaper方法，其核心是利用一套基于李雅普诺夫优化技术的在线带宽分配和流量调度方案，将流量更多地整形到95%定价模型中的免费时段．

3.2 调度方式

如表3所示，根据调度管理方式的不同，现有流量调度相关研究工作又可划分为分布式，集中式调度和混合式调度．

表3 不同数据中心流量调度方式的优缺点

挑战	优点	缺点
分布式调度	计算通信开销低，高可靠与高可扩展	调度决策欠优，调度策略管理和下发难
集中式调度	网络全局可控，调度决策可达理论最优	鲁棒性低，全局控制开销大
混合式调度	高可靠，高可扩展，更优的调度决策	复杂度高，难以达到集中式的最优调度

3.2.1 分布式调度

分布式流调度主要分为发送端驱动和接收端驱动两种方式，具体如下：

发送端驱动的调度指调度决策在发送端进行，而不涉及对中间设备网络功能的修改．例如，在TCP、DCTCP[8]和Hull[65]中，发送方在不知道其它发送方的调度决策时仅根据其本地信息及部分网络状态信息作出流速率控制的决策．为了获取部分网络状态信息，它们仅需要交换机支持显示拥塞通知（Explicit Congestion Notification，ECN）功能，而这种功能在绝大多数商业交换机中已经被兼容．但是ECN对短流没有太大帮助，因为短流很小，它们没有时间对网络状态的变化做出快速响应．此外，ECN不足以反馈交换机队列信息，属于粗粒度的信息反馈，也因此使得流调度决策次优．另一方面，pFabric[5]利用网络内的优先级为短流提供低延迟，并同时提高网络带宽利用率．但是，pFabric需要改变网络基础设施，因此在实际场景不可行．

总体而言，发送端驱动的流调度可扩展性及可靠性高，能够完全地在主机端进行实现．但是其在本地盲目地根据最短流优先或最短剩余数据量优先策略进行调度会使得数据中心网络内部发生拥堵．因为发送端看不到其它地方的流量信息，使得流虽然已经发送出去，但是被堵在了中间交换机或者接收端队列中．另外，发送端驱动的调度一般只能等拥塞发生后才能作出被动地响应，这在高速网络环境下将会变得不实用．

接收端驱动的调度指调度控制由接收端接管．与被动的发送端驱动方式不同，接收端驱动的调度大多属于主动式队列拥塞解决方案，且它们能根据当前可用网络资源对网络数据包的传输进行控制．典型的接收端驱动流调度方案有pHost[66]、NDP[67]、ExpressPass[68]、Homa[69]等．在pHost中，发送端首先发送RTS给接收端以向其请求数据的发送，然后接收端可感知到所有要向它传数据的主机，据此，接收端便决定给哪台主机发送令牌，从而允许相应主

机的数据传输.NDP通过在接收端维护一个PULL队列来限制所有incast发送方的总传输速率,当新数据包从发送方到达时,该队列将与附加的PULL请求一起加载,其中PULL请求包含一个计数器,该计数器决定与其关联的发送方所能允许发送的数据包的个数.然后PULL请求会被发送给发送端,从而确保接收端的整体传入速率不大于接口线速.ExpressPass通过在交换机和主机端控制credit报文来管理网络拥塞.credit报文由接收端发送,发送端每接到一个credit报文则发送一个数据报文.其核心是通过credit报文实现对整个网络的反演,进而反向控制数据报文的传输.在Homa中,每条流被分成可调度和不可调度部分,不可调度的数据包在流一启动就携带着流剩余数据大小信息发送出去,接收端根据其接收到的数据包解析出与其相关的所有流的信息,然后决定该给每条流怎样的优先级,并将此信息放在Grant报文中,返回给发送端.发送端每接到一个Grant报文后,则进行相应可调度数据包的发送.

接收端驱动的流调度比发送端驱动的流调度能更快地感知网络状态变化.但是要实现接收端驱动的调度需要对整个网络协议栈进行修改.另外,接收端驱动的方式往往会引发第一个RTT浪费的现象.Homa对此进行了一点改进,但是其中不可调度的数据包在突发情况下会造成队列堆积,严重时还会造成丢包.最后,接收端驱动的方式假设拥塞只发生在最后一跳的ToR交换机处.

3.2.2　集中式调度

在集中式调度方案中,往往有一个中央控制单元[70],其能协调网络中的传输任务,以避免拥塞.它还可以访问网络拓扑和资源的全局视图、交换机的状态信息和终端主机命令,包括流量大小、截止时间、优先级、交换机的排队状态和链路容量.它可以主动地在时间和空间上分配资源,调度流量的传输.为了进一步提高性能,它还可以将调度问题转化为具有资源约束的优化问题,并采用快速启发式方法进行解决.调度的有效性取决于中央控制单元的计算能力以及它到终端主机的通信延迟.

TDMA[71]、FastPass[72]和FlowTune[73]是典型的集中式调度器.TDMA将时间划分为若干轮,在每一轮中,它收集终端主机的需求.每一轮又被划分为固定大小的插槽.在此期间主机可以以无争用的方式进行通信.所有需求都在一轮结束时处理,并生成调度计划给终端主机.FastPass使用控制器进行包级别的调度和路由,以实现零队列延迟.但是,FastPass在控制器上调度每个流,因此它会增加短流的延迟.此外,它的可伸缩性较差,因为它必须以数据包的粒度调度和路由所有流.FlowTune改进了FastPass的可伸缩性.

集中式调度的优势显而易见.它可以提供网络状态和流属性的全局视图,从而提供更优的调度性能.但是它的缺点也比较突出.首先,中央控制器容易造成单点故障.其次,它还存在从终端主机收集网络状态和流属性的延迟和计算开销,控制器必须快速处理并解析传入的消息,然后再采取相应的行动.另外,它还引发了很大的网络资源分配和调度的开销.最后,在大型网络中,网络更新的一致性也是个问题.例如,不同更新可能会以不同的延迟被应用,从而导致暂时的拥塞或数据包丢失,这可能会损害延迟敏感服务的性能.

3.2.3　混合式调度

混合式调度通常指由集中式参数辅助的分布式调度.它具有分布式调度的高可扩性及稳定性,同时还能根据集中式参数作出更优的调度决策.例如,Fibbing[74]利用中央控制器来计算增量式拓扑,并将假节点插入路由器的路由表中,从而强制它们使用或避免使用某些路径来控制网络负载的分布.Mahout[75]在主机端通过监测套接字缓存来判断大小流,被判定为大流的数据包将会被打上标记.当有标记的数据包经过交换机时,交换机会将它们的元信息反馈给中央控制器,从而告诉中央控制器哪些流是大流.最后,中央控制器会为大流计算路径.Hedera[76]最初允许网络使用分布式方法调度和路由流,然后使用具有全局网络状态视图的中央控制器监控和检测大流,最后为大流重新安排路径,从而实现流量负载均衡.Hedera与Mahout的不同在于它们检测大流的方式.Mahout在终端主机插入一个填充层检测大流,而Hedera利用交换机流量计数器功能来检测大流.

混合式调度能够同时吸纳分布式和集中式调度的优势,并具有高可靠、高可扩以及高性能等特性.与此同时,混合式调度的劣势是中央控制器需要与分布式节点进行频繁地通信,这将会影响系统的性能.此外,虽然有集中式网络状态信息的辅助,但是由于存在通信延迟或者信息的粗粒度,分布式节点只能计算出次优的调度决策.

3.3　调度对象

如表4所示,根据调度对象的不同,现有流量调

表4 不同调度对象描述

调度对象	描述
数据中心内的流	运行在数据中心之上的在线业务，数据并行应用，高性能计算应用，机器学习应用等都将产生流量．不同应用对底层网络流的传输有不同需求，例如时限保障需求，带宽保障需求，低延迟需求等．不同需求将触发不同的流调度设计．
数据中心间的流	周期性的业务数据备份，前端面向用户的服务（例如，Web搜索，电子邮件，在线聊天，游戏，视频等）将引发跨数据中心传输的流量．由于数据中心间的网络带宽资源稀缺且昂贵，减少这类流量的传输成本将对数据中心供应商至关重要．
数据中心与用户间的流	数据中心与用户间的流一般指用户发送的服务请求流．这类服务请求流调度主要指服务请求应该被分配到那个数据中心．这类问题的相关研究主要集中在如何进行负载均衡，如何降低能耗减少二氧化碳排放，如何降低用户请求服务的延迟并保障用户的QoS等几个方面．

度相关研究工作又可划分为数据中心内流量调度，数据中心间流量调度，以及数据中心与用户间的流量调度．

3.3.1 数据中心内的流调度

随着数据中心内流量被东西流量所主导，数据中心内的流调度问题也逐渐受到关注．但是由于数据中心中承载着多样化的应用，人们很难找到一个通用的调度方法，能够适合不同业务场景．在云平台业务场景中，公平以及性能隔离保障是网络优化比较关心的目标，因为它与云租户的服务质量以及SLA直接关联．而在运行大规模在线业务场景中，延迟对上层应用极为重要．因此，降低这种类型应用的流的完成时间是流调度所需要达到的目标．此外，研究者发现很多的流有时限需求，并且在有些业务场景中，尾延迟更重要，因此涌现出了一批与时限保障和尾延迟相关的流调度工作．随着大数据的发展，很多新兴的并行计算应用被部署到了数据中心中，这些应用产生的流可通过Coflow模型来描述．因此，很多研究工作围绕着如何减少Coflow的完成时间进行展开．总体而言，数据中心内的流调度与上层业务需求息息相关．

3.3.2 数据中心间的流调度

为了提高链路带宽利用率，Google采用软件定义网络（Software Defined Network，SDN）的技术优化其数据中心间的广域网络，并于2013推出了B4[3]．在B4中，大部分链路的利用率达到了100%，而所有链路的平均带宽利用率也达到70%．至此，数据中心间的网络也开始得到了广泛的关注．数据中心间的流量也呈现爆炸式的增长．相比于数据中心内部网络，数据中心间的网络带宽十分稀缺，而且数据中心间的网络大都基于Internet网络构建，供应商对中间网络管控能力有限．因此，数据中心间的流调度发展也较为缓慢，且大都围绕公平带宽共享来展开．例如，Kumar等人[77]提出BwE方法，该方法关注广域网带宽分配问题．BwE是一个层次化的分配方法，其允许在不同层面采取不同的分配方法．BwE的核心是一个可以描述应用优先级以及不同实体效益的带宽函数．基于该函数，BwE不仅能够为存在竞争关系的应用/服务提供一定的性能隔离，还能为网络操作者提供可定制化的带宽分配策略．另外，相比于数据中心内的网络，数据中心间的带宽昂贵．因此，一些研究工作致力于设计有效的流量调度技术或方法来减少数据中心间的带宽成本或流量传输成本．

3.3.3 数据中心与用户间的流调度

数据中心与用户间的流一般指用户发送的服务请求或数据中心对用户服务请求的响应信息流．用户的服务请求需要被指引到合适的数据中心．如何合理地在数据中心间分配用户服务请求，以达到不同优化目标是现有研究重点关注的问题．第一类方法是从用户利益出发．例如，Xu等人[78]关注用户服务的公平性，他们希望地域较远的用户服务请求也能被公平地对待．为此，他们提出了一个基于纳什均衡博弈理论模型的公平性概念，并在满足实际容量约束前提下设计了分布式的用户服务请求分配算法．Zhang等人[79]认为用户应该就近选择服务接入点．Wong等人[80]采用请求式的网络探测技术来为就近服务接入机制的实现提供设计依据．但是，这种就近选择服务接入点不利于地理位置较远的用户，且还会给他们带来较大的延迟．另一类方法就是在分配用户服务请求时从供应商的利益出发．例如，Wendell等人[81]旨在降低供应商服务用户请求的网络，并在不同数据中心间达到负载均衡．Qureshi等人[82]关注供应商电力成本，并设计了电价感知的用户服务请求的分配策略，其优化的核心依据是不同地理位置的电价不同，这样可以让电价低的数据中心承载更多的请求，但这也容易造成过载问题．Mathew等人[83]在CDN场景中通过开关CDN服务器的方式减少能耗，并实现负载均衡．Mohsenian-Rad等人[84]在电网场景中，通过设计合

理的请求分配策略,实现负载均衡.Liu等人[85]关注绿色节能问题,并设计了绿色负载均衡的请求分配方法.Xu等人[86]同时考虑带宽效率以及能耗成本去设计请求分配算法.Boloor等人[87]利用先进先出调度策略加上动态请求分配方法来提升云服务供应商的全局收益.Le等人[88]联合考虑CO_2排放和能耗.Polverini等人[89]基于时空差异的电价设计了在线分配算法,从而降低能耗.Xu等人[90]基于交替方向乘子法(Alternative Direction Method of Multipliers,ADMM)的请求分配算法,以同时考虑用户和供应商的利益.Gao等人[91]同时考虑用户接入服务的延迟、能耗电力成本、二氧化碳排放等指标进行用户服务请求的分配.

4 发展趋势与展望

4.1 可编程数据平面环境下的流量调度

随着可编程交换机及可编程语言的推广和普及,传统固化的网络也开始变得灵活.PISA①是典型可编程交换机[92].PISA在match-action管道中处理数据包,并在入口和出口处注册阵列.经过可编程入口管道后,数据包进入与所选出端口关联的队列;退出队列后,数据包进入出口管道.此时,交换机可对数据包进行简单的处理.例如,在数据包的包头嵌入队列长度、数据包入队出队时间等信息.交换机还可以做出流量管理的决策,例如标记或丢弃数据包,或向入端口或控制软件发送反馈信息,以影响未来流量的处理.

在可编程交换机的支持下,传统基于主机端的流调度可以获得更精细的网络状态信息,为细粒度流调度设计提供了设计依据.此外,流调度还可直接在可编程交换机中进行.相比于主机端上的流调度,基于交换机的流调度可获得更多的流信息,从而制定更优的调度决策.

4.2 RDMA高速网络环境下的流量调度

传统数据中心应用大都基于TCP/IP进行通信的.在TCP/IP通信协议中,数据需要从用户应用空间的缓存复制到内核空间的Socket缓存中,然后通过一系列多层网络协议的数据包处理工作,数据才会被推送到NIC网卡中的缓存进行网络传输.频繁的数据移动和数据复制操作会带来很大的CPU开销,造成应用性能的下降.为了解决网络传输中服务器端数据处理延迟问题,远程直接内存存取(Remote Direct Memory Access,RDMA)技术诞生.RDMA是一种新型网络协议,它通过网络把资料直接传入计算机的存储区,将数据从一个系统快速移动到远程系统存储器中,而不对操作系统造成任何影响.它消除了外部存储器复制和上下文切换的开销,因而能解放内存带宽和CPU周期用于改进应用系统性能.RDMA有三种不同的硬件实现,分别是InfiniBand[93],RoCE[94]和iWARP[95].在RDMA网络环境下,为了确保无损,需要在交换机中启用基于优先级流量控制PFC技术,而PFC需要为每个队列设置Headroom缓存空间,以防止PAUSE信号已发出但还未到达上游节点的这段时间内发生丢包.在这种情况下,交换机队列可用缓存将更加有限.传统依赖交换机多优先级队列的调度方法将变得不可行,如何设计RDMA高速网络环境下的流调度成为亟待解决的问题.

4.3 机器学习辅助的流量调度

数据中心流调度是一个很难的在线决策问题.以前的方法大都依赖于启发式算法,而且还需要运营商对数据中心流量负载和其环境有很深的理解才能得到较为不错的调度结果.设计和实现合适的调度算法往往需要经历几个星期,这在大规模数据中心运营场景中肯定是不可行的.而近年来,随着机器学习(Machine Learning,ML)在生物信息学、语音识别和计算机视觉等多种应用领域取得了突破,机器学习也为流调度提供了潜在的机会.一般而言,机器学习试图构建算法和模型,这些算法和模型可以直接从数据中学习如何做出决策,而无需遵循预先定义的启发式规则.现有的机器学习算法一般分为三类:监督学习(Supervisored Learning,SL)、无监督学习(Unsupervisored Learning,USL)和强化学习(Reinforcement Learning,RL).更具体地说,SL算法从有标记的数据中进行分类或回归任务决策的学习,而USL算法则专注于用未标记的数据将样本集分类为不同的组.USL主要用于流量分类问题中.例如,Zhang等人[47]利用DBSCAN算法将网络流分类到不同Coflow中,并根据分类后的结果进行Coflow调度.SL最近被用于流量大小预测.例如,Đukić等人[96]运用GBDT(梯度下降树)对网络流的数据大小进行预测,然后再将预测出的流大小信息用于流调度中.在RL算法中,代理通过与环境的交互,学会寻找最佳的动作序列来最大化累积

① The P4 Language Consortium. P416 Language Specification. https://p4.org/p4-spec/docs/P4-16-v1.1.0-spec.html, 2018.

奖励(即目标函数).目前,RL算法被广泛地应用于解决网络问题,例如,拥塞控制[97-98]、异常检测[99-100]、流量工程[101-104]等.至于将RL用于数据中心流调度,Chen等人提出了AuTO[105].AuTO分为本地调度器和集中调度器,本地调度器运行中服务器主机端,它们根据本地信息作出快速的流量调度决策.集中调度器基于强化学习学习如何设置各个优先级队列的阈值,并周期性地将学到的阈值发送给服务器主机端去响应.但是AuTO对流量分布变化的反应不够快:在它们的实验中学习阈值需要8小时,这大大降低了它们的方案的灵活性.因此,如何利用机器学习设计高效、灵活,实用的流调度是当前亟待解决的难点之一.

4.4 数据中心间的流量调度

现有数据中心间的流调度大都针对备份数据流的传输延迟,传输成本等问题展开.对于数据中心间的小流,相关的研究甚少.实际上,随着跨数据中心机器学习的兴起[106],数据中心间还将存在着大量的小流.跨数据中心机器学习与传统分布式机器学习基本都采用参数服务器架构,但不同的是跨数据中心机器学习作业产生的参数需要经过低效且不可靠的数据中心间的广域网进行传输.在这种情况下,如何设计高效的流调度算法,对跨数据中心机器学习的性能提升尤为重要.

此外,由于数据中心分布在不同地理位置,受限于当地国家的政策,数据中心与数据中心间也许不能很自由地进行数据传输,在这种情况下,联邦学习兴起.联邦学习[107]是一种新兴的人工智能基础技术,其核心思想是在保障大数据交换时的信息安全、保护终端数据和个人数据隐私、保证合法合规的前提下,在多参与方或多计算结点之间开展高效率的机器学习.相比于传统跨数据中心的分布式机器学习,联邦学习对数据加密提出了要求.这对数据中心间流量调度的设计提出了更大的挑战.

5 相关工作

本节概述数据中心流量控制优化相关的综述文献.Liu等人[108]针对低延迟数据中心网络的实现方法给出了一个综述.Ren等人[109]主要对TCP Incast问题解决方案进行综述.Chen等人[110]对多租户云数据中心带宽分配问题进行了综述.Zhang等人[111]对数据中心网络传输层所涉及的六类问题进行了综述,包括TCP Incase,低延迟,虚拟化数据中心反常性能,多租户数据中心带宽共享,带宽利用不足,以及无损以太网TCP等.与上述综述文献不同,本文只关注数据中心流调度.

在流调度方面,Hu等人[112]对独立流和网络流组的调度方法进行了综述.他们主要从调度机制的实现方式对各种调度方法进行分析,忽略了调度优化目标以及调度对象这两个维度的对比和分析.Noormohammadpour等人[113]对数据中心网络流量优化的各种机制进行了综述,包括流调度,多径传输,负载均衡,流量整形等等.在流调度方面,他们分析了七种调度技术:带宽预留,流复制,截止时间感知,调度准则,抢占,抖动和ACK控制.他们分析的这些调度技术未能涵盖最小化流量传输成本这一优化目标.在未来研究展望中,他们仅考虑数据中心间流量优化这一个研究方向,而我们对可编程数据平面下的流调度、RDMA高速网络环境下的流调度、机器学习辅助的流调度和数据中心间流调度都作出了展望.

6 结 论

许多在线服务,并行计算应用,机器学习应用等都依赖于数据中心基础设施.这些应用在数据中心网络中产生大量流量,使得流量调度成为提高资源利用率,保障公平共享,以及保障应用性能的重要手段.本文回顾了数据中心流量调度问题所面临的挑战,并从不同维度分析归纳对比和总结了现有研究工作.尽管学术界涌现出了大量研究工作,但考虑到复杂性、性能和成本的指标,大多数流调度仍处于研究状态,远未被业界采纳.低复杂性、低成本,高性能的流调度方案还有待进一步的探索.最后,本文还指出数据中心流调度的几个潜在的发展方向,值得学术界进一步关注.

参 考 文 献

[1] Al-Fares M, Loukisas A, Vahdat A. scalableA, commodity data center network architecture//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Seattle, USA, 2008: 63-74

[2] Mysore R, Pamboris A, Farington N, et al. PortLand: A scalable fault-Tolerant layer 2data center network fabric// Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Barcelona, Spain, 2009: 39-50

[3] Jain S, Kumar A, Mandal S, et al. B4: Experience with a

globally-deployed software defined wan//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Hong Kong, China, 2013: 3-14

[4] Greenberg A, Hamilton J, Maltz D A, et al. The cost of a cloud: research problems in data center networks. ACM SIGCOMM Computer Communication Review, 2008, 39 (1):68-73

[5] Alizadeh M, Yang S, Sharif M, et al. pFabric: Minimal Near-optimal Datacenter Transport. ACM SIGCOMM Computer Communication Review, 2013, 43(4): 435-446

[6] Roy A, Zeng H, Bagga J, et al. Inside the Social Network's (Datacenter) Network//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM), London, United Kingdom, 2015: 123-137

[7] Benson T, Akella A, and Maltz D. Network traffic characteristics of data centers in the wild//Proceedings of the ACM SIGCOMM conference on Internet measurement. Melbourne, Australia, 2010: 267-280

[8] Alizadeh M, Greenberg A, Maltz D, et al. Data Center TCP (DCTCP). ACM SIGCOMM Computer Communication Review, 2010, 41(4): 63-74

[9] Kandula S, Menache I, Schwartz R, et al. Calendaring for wide area networks. ACM SIGCOMM Computer Communication Review, 2015, 44(4): 515-526

[10] Zhang H, Chen K, Bai W, et al. Guaranteeing deadlines for inter-datacenter transfers. IEEE/ACM Transactions on Networking, 2017, 25(1): 579-595

[11] Chen L, Chen K, Bai W, et al. Scheduling Mix-flows in Commodity Datacenters with Karuna//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Florianópolis, Brazil, 2016: 174-187

[12] Chowdhury M and Stoica I. Coflow: A networking abstraction for cluster applications//Proceedings of ACM Workshop on Hot Topics in Networks (HotNet). Redmond, USA, 2012: 31-36

[13] Zhang H, Zheng Z, Xu S, et al. Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clusters//Proceedings of USENIX Annual Technical Conference (ATC). Santa Clara, USA, 2017: 181-193

[14] Hu Z, Li B, and Luo J. Flutter: Scheduling tasks closer to data across geo-distributed datacenters//Proceedings of IEEE International Conference on Computer Communications (INFOCOM). San Francisco, USA, 2016: 1-9

[15] Liu S, Chen L, and Li B. Siphon: expediting inter-datacenter coflows in wide-area data analytics//Proceedings of USENIX Annual Technical Conference (ATC). Boston, USA, 2018: 507-518

[16] Zarifis Z, Miao R, Calder M, et al. DIBS: Just-in-time congestion mitigation for data centers//Proceedings of the European Conference on Computer Systems (EuroSys). Amsterdam, Netherlands, 2014: 6

[17] Guo C, Lu G, Wang H, et al. Secondnet: a data center network virtualization architecture with bandwidth guarantees//Proceedings of the ACM International Conference on emerging Networking Experiments and Technologies (CoNEXT). Philadelphia, USA, 2010: 15

[18] Duffield N G, Goyal P, Greenberg A, et al. A flexible model for resource management in virtual private networks//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Massachusetts, USA, 1999: 95-108

[19] Ballani H, Costa P, Karagiannis T, et al. Towards predictable datacenter networks//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto, Canada, 2011: 242-253

[20] Zhu J, Li D, Wu J, et al. Towards bandwidth guarantee in multi-tenancy cloud computing networks//Proceedings of IEEE International Conference on Network Protocols (ICNP). Austin, USA, 2012: 1-10

[21] Lee J, Turner Y, Lee M, et al. Application-driven bandwidth guarantees in datacenters//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Chicago, USA, 2014: 467-478

[22] Rodrigues H, Santos J R, Turner Y, et al. Gatekeeper: Supporting bandwidth guarantees for multi-tenant datacenter networks//Proceedings of USENIX WIOV. Portland, USA, 2011: 784-789

[23] Jeyakumar V, Alizadeh M, Mazieres D, et al. Eyeq: Practical network performance isolation at the edge//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI). Seattle, USA, 2013: 297-311

[24] Guo J, Liu F, Zeng D, et al. A cooperative game based allocation for sharing data center networks//Proceedings of IEEE International Conference on Computer Communications (INFOCOM). Turin, Italy, 2013: 2139-2147

[25] Guo J, Liu F, C. S. Lui J, et al. Fair network bandwidth allocation in iaas datacenters via a cooperative game approach. IEEE/ACM Transactions on Networking (TON), 2016, 24 (2):873-886

[26] Liu F, Guo J, Huang X, et al. eba: Efficient bandwidth guarantee under traffic variability in datacenters. IEEE/ACM Transactions on Networking (TON), 2017, 25(1):506-519

[27] Guo J, Liu F, Wang T, et al. Pricing intra-datacenter networks with over-committed bandwidth guarantee//Proceedings of USENIX Annual Technical Conference (ATC). SANTA CLARA, USA, 2017:69-81

[28] Popa L, Kumar G, Chowdhury M, et al. Faircloud: sharing the network in cloud computing//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Helsinki, Finland, 2012: 187-198

[29] Popa L, Yalagandula P, Banerjee S, et al. Elasticswitch: practical work-conserving bandwidth guarantees for cloud

computing//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Hong Kong, China, 2013: 351-362

[30] Ballani H, Jang K, Karagiannis T, et al. Chatty tenants and the cloud network sharing problem//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI). Seattle, USA, 2013: 171-184

[31] Wilson C, Ballani H, Karagiannis T, et al. Better never than late: Meeting deadlines in datacenter networks//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto,Canada, 2011: 50-61

[32] Hong C Y, Caesar M, Godfrey P. Finishing flows quickly with preemptive scheduling. ACM SIGCOMM Computer Communication Review, 2012, 42(4):127-138

[33] Vamanan B, Hasan J, Vijaykumar T. Deadline-aware datacenter tcp (d2tcp). ACM SIGCOMM Computer Communication Review, 2012, 42(4):115-126

[34] Chen L, Hu S, Chen K, et al. Towards minimal-delay deadline-driven data center tcp//Proceedings of ACM Workshop on Hot Topics in Networks. College Park, USA, 2013: 21

[35] Zhang H, Chen K, Bai W, et al. Guaranteeing deadlines for inter-data center transfers. IEEE/ACM Transactions on Networking (TON), 2017, 25(1): 579-595

[36] Jalaparti V, Bliznets I, Kandula S, et al. Dynamic pricing and traffic engineering for timely inter-datacenter transfers// Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Florianópolis, Brazil, 2016: 73-86

[37] Noormohammadpour M and Raghavendra C. S. Comparison of Flow Scheduling Policies for Mix of Regular and Deadline Traffic in Datacenter Environments. USA: Department of Computer Science, University of Southern California, Technique Report,2017:17-973

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

[39] Munir A, Baig G, Irteza S M, et al. Friends, not foes: synthesizing existing transport strategies for data center networks. ACM SIGCOMM Computer Communication Review, 2014, 44(4):491-502

[40] Li Z, Bai W, Chen K, et al. Rate-aware flow scheduling for commodity data center networks//Proceedings of IEEE International Conference on Computer Communications (INFOCOM). Atlanta, USA, 2017: 1-9

[41] Liu S, Bai W, Xu H, et al. RepNet: Cutting Latency with Flow Replication in Data Center Networks. IEEE Transactions on Services Computing, inpress

[42] Jalaparti V, Bodik P, Kandula S, et al. Speeding Up Distributed Request-response Workflows//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Hong Kong, China, 2013: 219-230

[43] Chowdhury M, Zaharia M, Ma J, et al. Managing data transfers in computer clusters with orchestra//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Toronto, Canada, 2011: 98-109

[44] Dogar F R, Karagiannis T, Ballani H, et al. Decentralized task-aware scheduling for data center networks//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Chicago, USA, 2014: 431-442

[45] Chowdhury M, Zhong Y, Stoica I. Efficient coflow scheduling with varys//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Chicago, USA, 2014: 443-454

[46] Chowdhury M, Stoica I. Efficient coflow scheduling without prior knowledge//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). London, United Kingdom, 2015: 393-406

[47] Zhang H, Chen L, Yi B, et al. Coda: Toward automatically identifying and scheduling coflows in the dark//Proceedings of the ACM Conference on Special Interest Group on Data Communication (SIGCOMM). Florianópolis, Brazil, 2016: 160-173

[48] Qiu Z, Stein C, Zhong Y. Minimizing the total weighted completion time of coflows in datacenter networks// Proceedings of the ACM Symposium on Parallelism in Algorithms and Architectures (SPAA). Portland, USA, 2015: 294-303

[49] Zhao Y, Chen K, Bai W, et al. Rapier: Integrating routing and scheduling for coflowaware data center networks// Proceedings of IEEE International Conference on Computer Communications (INFOCOM). Hong Kong, China, 2015: 424-432

[50] Li Y, Jiang S H C, Tan H, et al. Efficient online coflow routing and scheduling//Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc). Paderborn, Germany, 2016: 161-170

[51] Briscoe B. Flow rate fairness: Dismantling a religion. ACM SIGCOMM Computer Communication Review, 2007, 37 (2):63-74

[52] Lam T, Radhakrishnan S, Vahdat A, et al. Netshare: Virtualizing data center networks across services. USA: University of California, San Diego, Technique Report: CS2010-0957, 2010

[53] Shieh A, Kandula S, Greenberg A, et al. Sharing the data center network//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI). Boston, USA, 2011:23-23

[54] Chen L, Feng Y, Li B, et al. Towards performance-centric fairness in datacenter networks//Proceedings of IEEE

International Conference on Computer Communications (INFOCOM). Toronto, Canada, 2014: 1599-1607

[55] Nagle J. On packet switches within finite storage. IEEE Transactions on communications, 1987, 35(4):435-438

[56] Demers A, Keshav S, and Shenker S. Analysis and simulation of a fair queueing algorithm. ACM SIGCOMM Computer Communication Review, 1989, 19(4): 1-12

[57] Sharma N K, Liu M, Atreya K, et al. Approximating fair queueing on reconfigurable switches//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI). RENTON, USA, 2018: 1-16

[58] Chen L, Cui W, Li B, et al. Optimizing coflow completion times with utility max-min fairness//Proceedings of IEEE International Conference on Computer Communications (INFOCOM). San Francisco, USA, 2016: 1-9

[59] Wang W, Ma S, Li B, et al. Coflex: Navigating the fairness-efficiency tradeoff for coflow scheduling//Proceedings of IEEE International Conference on Computer Communications (INFOCOM). Atlanta, USA, 2017: 1-9

[60] Wang L and Wang W. Fair Coflow Scheduling without Prior Knowledge//Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS). Vienna, Austria, 2018:22-32

[61] Laoutaris N, Sirivianos M, Yang X, et al. Inter-datacenter bulk transfers with netstitcher//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto, Canada, 2011: 74-85

[62] Feng Y, Li B, Li B. Jetway: Minimizing costs on inter-datacenter video traffic//Proceedings of the ACM International Conference on Multimedia. Nara, Japan, 2012: 259-268

[63] Liu G, Shen H, Chandler H. Selective data replication for online social networks with distributed datacenters. IEEE Transactions on Parallel and Distributed Systems (TPDS), 2016, 27(8):2377-2393

[64] Li W, Zhou X, Li K, et al. TrafficShaper: Shaping Inter-Datacenter Traffic to Reduce the Transmission Cost. IEEE/ACM Transactions on Networking (TON). 2018, 26(3): 1193-1206

[65] Mohammad A, Abdul K, Tom E, et al. Less is More: Trading a Little Bandwidth for Ultra-low Latency in the Data Center//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI). San Jose, USA, 2012: 253-266

[66] Gao P X, Narayan A, Kumar G, et al. pHost: Distributed near-optimal datacenter transport over commodity network fabric//Proceedings of the ACM International Conference on emerging Networking Experiments and Technologies (CoNEXT). Heidelberg, Germany, 2015: 1

[67] Handley M, Raiciu C, Agache A, et al. Re-architecting Datacenter Networks and Stacks for Low Latency and High Performance//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Los Angeles, USA, 2017: 29-42

[68] Cho I, Jang K, and Han D. Credit-Scheduled Delay-Bounded Congestion Control for Datacenters//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Los Angeles, USA, 2017: 239-252

[69] Montazeri B, Li Y, Alizadeh M, et al. Homa: A receiver-driven low-latency transport protocol using network priorities//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Budapest, Hungary, 2018: 221-235

[70] Wan K, Luo X, Jiang Y, et al. The flow-oriented scheduling algorithms in SDN System. Chinese Journal of Computers, 2016, 39(6): 1209-1223(in Chinese)
(宛考, 罗雪峰, 江勇, 徐恪. 软件定义网络系统中面向流的调度算法. 计算机学报. 2016, 39(6):1209-23)

[71] Vattikonda B C, Porter G, Vahdat A, et al. Practical TDMA for datacenter Ethernet//Proceedings of the ACM European conference on Computer Systems (EuroSys), Switzerland, Bern, 2012: 225-238

[72] Perry J, Ousterhout A, Balakrishnan H, et al. Fastpass: A Centralized "Zero-queue" Datacenter Network//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Chicago, USA, 2014: 307-318

[73] Perry J, Balakrishnan H, and Shah D. Flowtune: Flowlet control for datacenter networks//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI). Boston, USA, 2017: 421-435

[74] Vissicchio S, Tilmans O, Vanbever L, et al. Central Control Over Distributed Routing//Proceedings of ACM Conference on Special Interest Group on Data Communication (SIGCOMM), London, United Kingdom, 2015:43-56

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

[76] Al-Fares M, Radhakrishnan S, Raghavan B, et al. Hedera: dynamic flow scheduling for data center networks//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI), SAN JOSE, USA, 2010

[77] Kumar A, Jain S, Naik U, et al. Bwe: Flexible, hierarchical bandwidth allocation for wan distributed computing. ACM SIGCOMM Communication Review, 2015, 45(4):1-14

[78] Xu H, Li B. A general and practical datacenter selection framework for cloud services//Proceedings of IEEE International Conference on Cloud Computing (CLOUD). Honolulu, USA, 2012: 9-16

[79] Zhang Z, Zhang M, Greenberg A G, et al. Optimizing cost and performance in online service provider networks//Proceedings of USENIX Symposium on Networked Systems

Design and Implementation (NSDI). San Jose, USA, 2010: 33-48

[80] Wong B, Sirer E G. Closestnode. com: an open access, scalable, shared geocast service for distributed systems. ACM SIGOPS Operating Systems Review, 2006, 40(1):62-64

[81] Wendell P, Jiang J W, Freedman M J, et al. Donar: decentralized server selection for cloud services. ACM SIGCOMM Computer Communication Review, 2010, 40(4):231-242

[82] Qureshi A, Weber R, Balakrishnan H, et al. Cutting the electric bill for internet-scale systems. ACM SIGCOMM computer communication review, 2009, 39(4):123-134

[83] Mathew V, Sitaraman R K, Shenoy P. Energy-aware load balancing in content delivery networks//Proceedings of IEEE International Conference on Computer Communications (INFOCOM). Orlando, USA, 2012: 954-962

[84] Mohsenian-Rad A H, Leon-Garcia A. Coordination of cloud computing and smart power grids//Proceedings of IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). Maryland, USA, 2010: 368-372

[85] Liu Z, Lin M, Wierman A, et al. Greening geographical load balancing//Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). SAN JOSE, USA, 2011: 233-244

[86] Xu H, Li B. Cost efficient datacenter selection for cloud services//Proceedings of IEEE International Conference on Communications in China (ICCC). Beijing, China, 2012: 51-56

[87] Boloor K, Chirkova R, Viniotis Y, et al. Dynamic request allocation and scheduling for context aware applications subject to a percentile response time sla in a distributed cloud// Proceedings of IEEE International Conference on Cloud Computing Technology and Science (CloudCom). Indianapolis, USA, 2010: 464-472

[88] Le K, Bianchini R, Nguyen T D, et al. Capping the brown energy consumption of internet services at low cost// Proceedings of IEEE Intertional Conference on Green Computing. Hangzhou, China, 2010: 3-14

[89] Polverini M, Cianfrani A, Ren S, et al. Thermal-aware scheduling of batch jobs in geographically distributed data centers. IEEE Transactions on cloud computing (TCC), 2014, 2(1):71-84

[90] Xu H, Li B. Joint request mapping and response routing for geo-distributed cloud services//Proceedings of IEEE International Conference on Computer Communications (INFOCOM). Turin, Italy 2013: 854-862

[91] Gao P X, Curtis A R, Wong B, et al. It's not easy being green. ACM SIGCOMM Computer Communication Review, 2012, 42(4):211-222

[92] Bosshart P, Gibb G, Kim H S, et al. Forwarding metamorphosis: Fast programmable match-action processing in hardware for SDN//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Hong Kong, China, 2013: 99-110

[93] Infiniband Trade Association. InfiniBand architecture volume 2, physical specifications, release 1.3, 2012

[94] Guo C, Wu H, Deng Z, et al. RDMA over commodity ethernet at scale//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Florianópolis, Brazil, 2016: 202-215

[95] Recio R, Metzler B, Culley P, et al. A remote direct memory access protocol specification. RFC 5040, 2007

[96] Đukić V, Jyothi SA, Karlaš B, et al. Is advance knowledge of flow sizes a plausible assumption//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI), BOSTON, MA, USA, 2019: 565-580

[97] Dong M, Meng T, Zarchy D, et al. PCC Vivace: Online-learning congestion control//Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI), Renton, USA, 2018: 343-356

[98] Jay N, Rotman N. H, Godfrey P, et al. Internet Congestion Control via Deep Reinforcement Learning. arXiv preprint arXiv:1810.03259, 2018

[99] Limthong K, Fukuda K, Ji Y, et al. Unsupervised learning model for real-time anomaly detection in computer networks. IEICE TRANSACTIONS on Information and Systems, 2014, 97(8): 2084-2094

[100] Ma M, Zhang S, Pei D, et al. Robust and rapid adaption for concept drift in software system anomaly detection// Proceedings of IEEE International Symposium on Software Reliability Engineering (ISSRE), Memphis, USA, 2018: 13-24

[101] Boyan JA and Littman ML. Packet routing in dynamically changing networks: A reinforcement learning approach// Proceedings of Advances in neural information processing systems(NIPS), Denver, USA, 1994: 671-678

[102] Nie L, Jiang D, Guo L, et al. Traffic matrix prediction and estimation based on deep learning in large-scale IP backbone networks. Journal of Network and Computer Applications, 2016, 76:16-22

[103] Valadarsky DA, Schapira M, Shahaf D, et al. Learning to route with deep RL//Proceedings of Advances in neural information processing systems(NIPS), Long Beach, USA, 2017

[104] Xu Z, Tang J, Meng J, et al. Experience-driven networking: A deep reinforcement learning based approach// Proceedings of IEEE International Conference on Computer Communications (INFOCOM), Honolulu, USA, 2018: 1871-1879

[105] Kai Chen Feng Liu Li Chen, LingysJustinas. Auto: scaling deep rein- forcement learning for datacenter-scale automatic traffic optimization//Proceedings of the ACM Special Interest

Group on Data Communication (SIGCOMM). Budapest, Hungary, 2018: 191-205

[106] Hsieh K, Harlap A, Vijaykumar N, et al. Gaia: Geo-Distributed Machine Learning Approaching LAN Speeds// Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI). Boston, USA, 2017: 629-647

[107] Liu Y, Chen T, and Yang Q. Secure Federated Transfer Learning. arXiv preprint, arXiv:1812.03337

[108] Liu S, Xu H, and Cai Z. Low latency datacenter networking: A short survey. arXiv preprint arXiv:1312.3455, 2013

[109] Ren Y, Zhao Y, Liu P, et al. A survey on TCP Incast in data center networks. International Journal of Communication Systems, 2014, 27(8): 1160-1172

[110] Chen L, Li B, and Li B. Allocating bandwidth in datacenter networks: A survey. Journal of Computer Science and Technology, 2014, 29(5): 910-917

[111] Zhang J, Ren F, and Lin C. Survey on transport control in data center networks. IEEE Network, 2013, 27(4): 22-26

[112] Hu Z, Li D, Li Z. Recent Advances in Datacenter Flow Scheduling. Journal of Computer Research and Development, 2018, 55(9): 1920-1930(in Chinese)
(胡智尧, 李东升, 李紫阳. 数据中心网络流调度技术前沿进展. 计算机研究与发展, 2018, 55(9): 1920-1930)

[113] Noormohammadpour M, Raghavendra CS. Datacenter traffic control: Understanding techniques and tradeoffs. IEEE Communications Surveys & Tutorials. 2017, 20(2): 1492-1525

LI Wen-Xin, Ph. D. His research interests include data center networking, software defined networking, cloud computing, parallel and distributed computing, etc.

QI Heng, Ph. D. associate professor. His research interests include software defined networking and data center networking.

XU Ren-Hai, Ph. D. His research interests include data center networking and cloud computing.

Zhou Xiao-Bo, Ph. D. associate professor. His research interests include software defined networking and data center networking.

LI Ke-Qiu, Ph. D. professor. His research interests include wireless network, cloud computing and software defined data center networking.

Background

Flow scheduling, determining when and at what rate to transmit each flow, has become a hot research topic in the field of data center networks. The are many flow scheduling solutions in the literature. This paper conducts a survey study on the research progress of the flow scheduling problem in data center networks, by classifying, comparing and summarizing existing research work along the following three dimensions: scheduling optimization goals, scheduling methods and scheduling entities. Then, future research trends in this field are discussed.

This work is supported by the National Key Research and Development Program of China No. 2016YFB1000205; the State Key Program of National Natural Science of China under Grant 61432002. These projects aim to make advances to software defined cloud data centers. This paper summarizes the research progress of the flow scheduling problem in data centers in recent years.