匀地进行流量分配.在对 TCP 和 MPTCP 的流传输过程中,能够保证短的流完成时间.在服务器数量较多的情况下,Beamer 能够在短时间内进行新配置的部署.Beamer 设计无状态的负载均衡方案,减轻了负载均衡器的存储和处理开销,但当网络状态变化影响流的数量较多时,其 daisy chaining 机制会导致大量流量通过服务器进行转发,使得服务器开销增大,成为网络传输性能瓶颈.

(6) 应用层方案对比

应用层方案考虑数据中心网络中服务器之间的性能差异,保证服务能够均匀地分配到相应的服务器,同时维护连接的一致性,实现数据中心网络的负载均衡.通过分析现有应用层方案,从方案的关键设计和方案性能两个角度对各方案进行对比,比较结果见表 3.

Table 3 The comparison of load balancing schemes based on the application layer in data center networks

表 3 数据中心网络负载均衡应用层方案对比

方案	关键设计				方案性能			
	特点	架构	状态维护	硬件需求	瓶颈	速度	适应性	复杂度
Ananta	3 层负载均衡结构	软件	服务器	交换机+服务器	软件性能	慢	高	低
Duet	现有商业交换机和 软件交换机共同维护	软件+硬件	交换机	交换机+服务器	分布式存储	一般	高	一般
Maglev	绕过操作系统的软件负载均衡器	软件	服务器	服务器	运行模块	一般	一般	一般
SilkRoad	利用交换芯片代替 软件负载均衡器	硬件	交换机	P4 交换机	特殊芯片性能	快	低	高
Beamer	利用连续空间哈希和 daisy chaining 技术	软件	服务器	服务器	Daisy chaining 技术	一般	高	一般

方案关键设计包括方案特点、设计架构、状态维护方式和方案硬件需求几个方面.从比较结果来看,现有传输层负载均衡方案主要通过软件和硬件负载均衡器的部署实现负载均衡功能,为了提高负载均衡速度通常会采用相结合的部署方式,部分方案还对软件负载均衡器进行优化.硬件负载均衡器的优点在于具有更快的处理速度,软件负载均衡器的优点在于能够维护更多状态并执行更复杂的负载均衡策略.因此,结合软/硬件是现有应用层方案的趋势.大多数负载均衡状态都在软件负载均衡服务器进行维护,但部分利用新型交换芯片的方案会在交换机维护部分状态,这类方案会造成更多的硬件需求.

在方案性能方面,对应用层方案的性能瓶颈、方案速度、适应性和复杂度几个方面进行比较.应用层方案的性能主要受限于负载均衡器的性能,但是对于利用特殊存储结构或状态维护算法的方案,其设计的关键结构和算法也是负载均衡性能的瓶颈.硬件负载均衡方案通常相比于软件负载均衡方案具有更快的速度,但软件负载均衡方案由于其部署难度较低,因此具有更高的适应性.对于能够利用现有网络设备的方案其复杂度相对较低,但对于修改现有网络设备或需要专用设备的方案则具有更高的复杂度,其部署难度也更大.

2.5 综合方案

数据中心网络流量具有高动态性,其与传统网络特点不同,因此优化传统的负载均衡机制只能在一定程度上提升负载均衡效果.影响数据中心网络负载的因素很多,包括拓扑结构、流量特性和调度机制等,针对数据中心网络设计合适的负载均衡方案需要综合考虑这些因素,通过对数据中心网络进行多层设计来达到需要的性能.目前典型的、综合的数据中心网络负载均衡机制包括 DeTail[48]、Fastpass[49]、Hermes[50]、NDP[51]和 AuTO[52]等,其分类结果如图 5 所示.

Fig.5 The classification results of the schemes base on the synthetic design

图 5 综合方案分类结果

(1) DeTail

DeTail 是一种跨层网络栈设计,其在底层网络快速检测拥塞,驱动上层网络执行路由决策以选择更不拥塞的路径,从而实现网内的负载均衡并保证流的时间约束.DeTail 在链路层采用 PFC 机制[53],利用 Pause 和 Unpause 信息控制交换机上下游的包发送来实现无损传输.DeTail 在交换机本地执行每个包的负载均衡,根据端口队列占用反映的拥塞信息动态地选择包的下一跳.由于无损结构网络不会因为拥塞造成包丢失,因此传输层采用 ECN 机制实现拥塞感知.同时根据不同流的时延敏感程度指定流优先级,网络传输过程中保证高优先级的包先处理.DeTail 能够完全避免拥塞相关的丢包,相比于流哈希机制能够极大地缩短流完成时间.同时,其能有效地让包绕过拥塞热点,相比于无损的包随机发送也能实现更小的流完成时间,对不同的流模式,包括序列流和分裂聚合流,都能提升其传输性能.DeTail 通过设计跨层结构实现每个包的负载均衡,然而,由于其利用 PFC 实现无损的 2 层传输,对数据中心网络设备有一定需求,而且在网络拓扑较大的情况下,PFC 的消息通告方式会造成不同位置流的相互影响,从而降低 2 层数据的传输效果.

(2) Fastpass

理想的数据中心网络传输性能包括低时延、高利用率和拥塞避免等,为了实现网络资源的最优分配,Fastpass 提出利用一个中央控制器(arbiter)控制每个发送端发送状态的方案.当每个包需要进行传输时,需要和 arbiter 进行交互来获得包的发送时间和路径.Fastpass 设计 FCP(fastpass control protocol)来实现低带宽开销和低时延的端系统和 arbiter 之间的消息传递协议,还设计多个 arbiter 之间的复制机制以实现故障容忍的恢复策略.Fastpass 设计最大最小公平分配的时隙分配算法来实现包发送时间分配,并利用快速边涂色算法进行包路径选择,保证 arbiter 为每个包进行时隙和路径分配的效率.将 Fastpass 方案与传统基于 TCP 传输的数据中心网络方案进行对比,Fastpass 能够实现近乎零队列的高吞吐量,显著地缩短了交换机队列的平均大小以及时延,同时能够保证公平的、快速的吞吐量分配.Arbiter 中算法的设计有效地提高了其可扩展性,并能实现细粒度的时隙控制.Fastpass 利用集中式控制的优点,实现了全局感知的包调度策略,提高了网络传输性能,并消除了端系统拥塞控制的需求.然而,由于大部分调度工作在集中式控制器完成,因此,在网络负载较大的情况下,集中式控制器性能容易成为瓶颈,限制了其应用场景.

(3) Hermes

Hermes 是一种基于端系统的不需要修改交换机的负载均衡方案,其在端系统设计感知模块进行不同路径的拥塞感知,并设计重路由模块来确定是否进行重路由.Hermes 根据 RTT 和 ECN 共同感知路径拥塞状态,将路径分为 3 种类型,并利用包的重传和超时事件来推测交换机故障情况.为了提高端系统对网络的可见性,Hermes 采用 power-of-two-choices[15]机制探测部分路径时延,从而有效地辅助负载均衡策略.当包对应的流出现超时情况,或包的路径发生拥塞时,触发重路由过程,从最好的路径开始选择,并考虑重路由为传输带来的收益,从而进行重路由决策.在对称拓扑的实验中,Hermes 相比于 ECMP 和 CLOVE-ECN 机制能够实现更小的流完成时间,性能接近 Presto 和 CONGA.在非对称拓扑实验中,Hermes 由于其及时的重路由特性,相比于 CONGA,能够实现更小的流完成时间,利用其主动探测的特性,相比于 CLOVE-ECN 和 LetFlow,也能实现更小的流完成时间.Hermes 相比于其他方案能够及时、有效地检测故障情况.Hermes 在端系统进行拥塞感知和负载均衡减少了部署成本,并实现了相当的性能.然而,端系统感知拥塞有一定延迟,因此无法及时处理突流情况,而且路径感知和重路由算法中参数较多,如何选择合适的参数需要结合网络情况进行选择.

(4) NDP

NDP 提出一种新的网络传输架构,可为短流实现接近最优的完成时间,并在多种场景下提供大吞吐量.NDP 对交换机和端系统协议栈都进行修改.在交换机本地采用浅缓存策略(8 个包)并执行 CP 机制[54],维护两个优先级队列,低优先级队列存储包数据,高优先级队列存储修剪的包头、ACKs 和 NACKs 包.当一个包到达后,如果低优先级队列溢出,则以 50%的概率选择修剪新到达的包或低优先级队列的尾包.在调度过程中,两个优先级队列采用权重调度方式来提高传输能力.NDP 中发送端进行每个包的路径选择,发送端获取到达目的地的所有路径链表并随机组合,发送包时按照组合的链表顺序发送,当所有路径都发送过一个包后重新组合路径链表继续发

送过程.NDP 采用接收端驱动的传输层协议,在第 1 个 RTT 时间内发送一定窗口的包,然后根据确认的 ACKs 和 NACKs 信息,接收端发送 PULL 包实现未收到数据的重新拉取.当交换机高优先级队列溢出时,直接交换新生成的修剪包头的源地址和目的地址,将包返回给发送端.在端系统,NDP 不需要执行拥塞控制策略.NDP 相比于 DCTCP、MPTCP 和 DCQCN[55]能够实现更小的短流完成时间.在网络高负载情况下,NDP 在交换机队列长度为 8 个包的情况下能够实现近乎最优的最大网络能力,在多对一传输场景中能够实现近优的时延和公平性.NDP 通过重新设计网络协议栈,利用接收端驱动机制实现网络传输的负载均衡,保证小流的低时延和大流的高吞吐量.但 NDP 在高负载网络中,没有拥塞控制机制会造成过多的重传开销,在非对称网络中,NDP 无法处理长度不同的路径分配,而且 NDP 对网内和端系统都进行修改,部署难度较大.

(5) AuTO

AuTO 通过模仿生物神经系统结构设计了一种二级的深度强化学习(deep reinforcement learning,简称 DRL)系统[56]以实现数据中心网络中的流量传输优化.AuTO 基于 PIAS[57]的架构,在主机和交换机本地实现多级反馈队列,根据不同优先级队列对应的流大小阈值,将不同流分配到不同优先级中执行严格的优先级调度.AuTO 在主机实现 PS(peripheral system)系统,收集本地流量信息,并根据控制器发送的不同优先级流大小阈值进行本地流量决策;设计一个控制器 CS(central system),根据网络状态信息通过 sRLA(short flow DRL agent)计算最优的多级反馈队列流阈值,并根据主机上报的大流信息,通过 lRLA(long flow DRL agent)实现大流的路径选择、速率控制和优先级设置.在不同负载情况的实验中,AuTO 相比于最短任务优先和最少服务优先机制能够实现更小的流完成时间.在时间和空间同构或异构的环境中,AuTO 能够明显地提升网络传输能力,实现稳定的性能提升.而且,AuTO 对端系统造成的开销小,能够实现 10ms 内的状态更新.AuTO 将人工智能(artificial intellengence,简称 AI)应用于数据中心网络中的流量传输优化,利用 AI 的自学习自优化特性实现阈值自适应以保证传输效果.但是 AI 方法的可行性和适用性仍是一个问题,当网络场景与训练场景不一致时,基于 AI 的方法是否能够适应不同场景的变化仍需要通过进一步的实验进行验证.

(6) 综合方案对比

综合方案对网络中多个层面进行修改,设计多种机制克服单一修改某个层可能产生的缺陷,实现性能更好的负载均衡效果.通过分析现有综合的方案,从方案的关键设计和方案性能两个角度对各方案进行对比,比较结果见表 4.

方案的关键设计包括方案特点、方案实现关键需求、拥塞信息获取、方案对大小流的考虑情况以及是否考虑不同优先级传输等方面.大部分综合方案对网络具有较大的修改,包括特殊的设备需求、特殊的传输策略以及高性能的处理需求等.通常,综合方案会采集网络拥塞信息作为决策参考,但对于全局最优的方案来说,拥塞信息可以不考虑.综合方案根据设计不同对大小流会进行不同的处理,对于有明显大小流需求的场景,会直接考虑大小的流差异化管理.大部分综合方案都会考虑传输数据的优先级,优先处理控制信息从而进行及时响应.

方案性能主要考虑综合方案的速度、适应性和复杂度等几个方面.对网内设备修改较多的方案通常执行本地决策,因此具有更快的处理速度,而集中式管理的方案会由于管理周期影响其速度.对于设计过程中综合考虑网络特性的方案来说,其相对于考虑特殊情况的方案来说具有更好的适用性,不会由于网络情况变化造成性能的明显下降.大部分综合方案都需要修改网络中的多个部分,因此大多数方案复杂度较高,除了少数只修改端侧的方案外,综合方案的部署难度都会相对较大.

Table 4 The comparison of load balancing schemes based on the synthetic designs in data center networks

表 4 数据中心网络负载均衡综合方案对比

方案	关键设计					方案性能		
	特点	需求	拥塞信息	大小流	优先级	速度	适应性	复杂度
DeTail	采用 PFC 技术实现无损传输	PFC 实现无损二层	交换机队列	不考虑	有	快	高	高
Fastpass	为每个包分配路径和时间	高性能中央控制器	无	不考虑	无	慢	一般	高
Hermes	端侧推测感知拥塞并重路由	端侧修改	RTT 和 ECN	不考虑	无	一般	高	低
NDP	接收端驱动协议	交换机和端侧修改	控制包	考虑	有	一般	一般	高
AuTO	利用二级 DRL 实现传输管理	控制器运行 DRL 模块并决策	流状态	考虑	有	慢	一般	高

3 方案对比和讨论

数据中心网络负载均衡方案通过不同角度的设计保证了低时延和高吞吐的网络性能.表 5 从各个不同方面整体分析并评估了各个方案的特点.根据负载均衡机制之间的差异,从方案类型、负载均衡粒度、方案控制结构、负载均衡类型、拥塞感知机制、网络修改位置、方案扩展性和部署难度几个方面进行了对比.

从不同方案对比中可以发现,大部分数据中心网络负载均衡方案通过网络层传输调度直接进行优化,综合方案普遍性能更好,但部署难度较大.从不同负载均衡粒度来看,粒度较小的方案实现的效果较好.在不同控制结构中,集中式方案会造成巨大开销,从而限制其扩展性.不同的负载均衡类型对网络拥塞的控制效果也不同,被动负载均衡在拥塞后进行处理降低了拥塞响应速度,主动方式可以避免拥塞的发生.对于不同的感知策略,全局拥塞信息感知对负载均衡具有明显的促进作用,可以实现全局最优的调度.现有方案中,越复杂且对网络的修改越多的方案其性能相对来说要更好.然而,实际数据中心网络部署中,可部署性是非常重要的问题,无法简单升级或迭代更新的方案在实际数据中心中难以应用.因此,数据中心网络负载均衡方案的设计需要从性能、可扩展性和实用性等角度综合考虑,在满足传输性能的同时还要尽可能地保证低成本和开销.

Table 5 The evaluation of load balancing schemes in data center networks

表 5 数据中心网络负载均衡方案评估

方案	方案类型	粒度	控制结构	均衡类型	拥塞感知	修改位置	效果	扩展性	部署难度
DRB	网络层	包	分布式	主动	无	端系统+交换机	中	中	容易
RPS	网络层	包	分布式	主动	无	交换机	中	好	容易
DRILL	网络层	包	分布式	主动	本地	交换机	好	好	中等
CONGA	网络层	Flowlet	分布式	被动	全局	ToR	好	中	困难
HULA	网络层	Flowlet	分布式	被动	全局	交换机	好	好	中等
CLOVE	网络层	Flowlet	分布式	被动	全局	端系统	中	好	容易
LetFlow	网络层	Flowlet	分布式	被动	无	交换机	中	好	容易
Presto	网络层	Flowcell	分布式	主动	无	端系统	中	好	容易
ECMP	网络层	流	分布式	主动	无	无	中	好	容易
WCMP	网络层	流	分布式	主动	无	交换机	中	好	容易
Hedera	网络层	流	集中式	主动	全局	控制器	中	差	容易
MicroTE	网络层	流	集中式	主动	全局	控制器+端系统	中	好	容易
Mahout	网络层	流	集中式	主动	全局	控制器+端系统	中	好	容易
Devoflow	网络层	流	集中+分布	主动	全局	控制器+交换机	好	好	中等
LocalFlow	网络层	流	分布式	主动	本地	交换机	中	中	困难
FlowBlender	网络层	流	分布式	主动	全局	端系统	中	好	中等
Freeway	网络层	流	集中式	主动	全局	端系统+交换机	中	中	中等
Expeditus	网络层	流	分布式	主动	本地	交换机	好	好	中等
Scheduling Algorithm	网络层	流	集中式	主动	全局	控制器	好	差	困难
MPTCP	传输层	多流	分布式	被动	全局	端系统	中	好	容易
XMP	传输层	多流	分布式	被动	全局	端系统	中	好	中等
RackCC	传输层	多流	分布式	被动	全局	端系统+ToR	好	中	困难
MMPTCP	传输层	多流	分布式	被动	全局	端系统	好	中	中等
VMS	传输层	包	分布式	被动	全局	端系统	中	中	中等
Ananta	应用层	流	分布式	主动	本地	软件 Mux	中	中	容易
Duet	应用层	流	分布式	主动	本地	软件 Mux+交换机	好	好	容易
Maglev	应用层	流	分布式	主动	本地	软件 Mux	中	中	容易
SilkRoad	应用层	流	分布式	主动	本地	P4 交换机	中	中	困难
Beamer	应用层	流	分布式	主动	本地	软件 Mux	好	中	容易
DeTail	综合	包	分布式	被动	本地	交换机	好	中	中等
Fastpass	综合	包	集中式	主动	全局	控制器+端系统	好	差	困难
Hermes	综合	流	分布式	被动	全局	端系统	好	中	容易
NDP	综合	流	分布式	主动	无	端系统+交换机	好	中	困难
AuTO	综合	流	集中+分布	主动	全局	控制器+端系统	好	中	中等

4 总结与展望

数据中心网络是现代网络和云计算的重要基础架构,数据中心网络的传输性能直接影响网络的传输能力

和服务体验.因此,设计高效的数据中心网络负载均衡方案以满足低时延高吞吐的数据中心网络需求十分关键.现有的数据中心网络负载均衡方案可按照不同解决层面分为:网络层、传输层、应用层和综合方案四大类,其他角度则可从负载均衡粒度、方案控制结构、负载均衡类型和拥塞感知策略等方面对负载均衡方案进行归纳.虽然现有方案设计了多种机制以提高网络传输能力,但仍存在一些弊端:(1) 方案基于传统 TCP 协议栈,性能受限于TCP基于窗口的传输策略;(2) 负载均衡方案性能依赖于现有数据中心典型的网络拓扑;(3) 现有负载均衡方案针对某个场景优化,适用性不强;(4) 现有方案未充分利用数据平面控制能力;(5) 没有充分开发新型传输设备的优势.鉴于现有方案的分析,总结了数据中心网络负载均衡未来可能的研究方向.

(1) 设计新型协议栈

由于传统 TCP 传输协议栈不适用于数据中心的高速场景,在负载均衡过程中无法快速达到最优的网络性能,因此重新设计协议栈是解决传统TCP问题的关键.先前设计的新型协议栈NDP是重新设计网络协议栈的探索,近期提出的 Homa[58]也是通过设计新型协议栈来提高网络传输性能.但是,这些新型协议栈的主要目的是通过进行网络的拥塞控制来提高传输能力,对网络负载均衡的效果只能起到部分提升的作用.因此设计兼顾负载均衡目的的新型协议栈是未来重要的研究方向.

(2) 优化数据中心拓扑

现有数据中心网络负载均衡方案主要基于典型的数据中心网络拓扑(Clos 拓扑)进行设计,负载均衡性能受限于拓扑约束,因此优化数据中心网络拓扑结构是提高网络最大传输能力的重要途径.近期提出的 Flat-tree[59]即是利用可变换的交换结构实现不同场景中的拓扑改变来提升传输性能.负载均衡方案设计时可以结合拓扑优化机制实现较高的网络利用率,并设计最优的引流策略以保证负载均衡的效果.

(3) 结合机器学习和人工智能技术

由于很多数据中心网络负载均衡机制与网络应用场景和流量特性相关,而且设计的机制性能依赖于相关参数的选取,因此可利用机器学习和人工智能技术,通过分析历史数据实现自适应的参数调整.先前的 AuTO 方案即是根据网络中流的状态自适应地给定相关阈值,并进行自学习过程.近期提出的 DRL-TE[60]方案则是利用 DRL 实现流的最优分流传输.目前,此类方案在特定应用场景中效果显著,其适用性仍有待进一步验证.因此结合机器学习和人工智能技术的数据中心网络负载均衡方案是未来研究的一个热点.

(4) 设计新型数据中心网络数据平面

随着网络传输需求的不断增大和各种传输模式的涌现,传统数据中心网络数据平面已经成为提升传输性能的瓶颈,因此出现了许多新型数据平面技术,包括智能网卡的应用[61]、新型网络数据面架构[62]以及网内缓存技术[63]等.这些新型数据平面为网络带来更多的管理维度,结合新维度实现数据中心网络负载均衡是进一步提高数据中心服务能力的关键.利用智能网卡提供的管理能力,可以直接提升端系统均衡吐流能力,同时结合新型数据面架构进一步提高网络负载管理性能.尤其是最近提出的网内缓存技术,可以利用网内缓存对网络负载的吸收特点,设计更深层次的差异化管理,从而保证数据中心网络的均衡负载.因此,结合数据平面新技术是设计数据中心网络负载均衡方案的重要发展方向.

(5) 结合可编程网络技术

受限于传统网络硬件更新周期长和定制能力差等缺点,可编程网络提供了一种高效易管理的网络抽象.典型的可编程网络技术包括 P4[24]和各种可编程交换机技术[64,65],为网络管理者提供转发平面控制接口,从而提供更灵活且高性能的调度能力.因此结合可编程网络技术,可以直接在转发平面定制负载均衡功能,直接感知数据平面状态并做出决策,同时还能根据网络和流量状态变化修改转发平面配置,实现动态负载均衡策略.随着可编程网络技术的不断发展,网络中的可编程设备能够维护和存储的信息量不断增大,使得复杂的负载均衡方案能够在数据中心网络中部署,增大了方案的可扩展性和适用性.随着可编程网络逐渐成为未来研究热点,结合可编程网络技术来设计数据中心网络负载均衡方案也成为了未来的研究热点.

References:

[1] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2008. 63–74. [doi: 10.1145/1402958.1402967]

[2] Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S. VL2: A scalable and flexible data center network. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2009. 51–62. [doi: 10.1145/1592568.1592576]

[3] Singla A, Hong CY, Popa L, Godfrey PB. Jellyfish: Networking data centers randomly. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2012. 225–238.

[4] Kandula S, Sengupta S, Greenberg A, Patel P, Chaiken R. The nature of data center traffic: Measurements & analysis. In: Proc. of the ACM SIGCOMM Conf. on Internet Measurement. New York: ACM, 2009. 202–208. [doi: 10.1145/1644893.1644918]

[5] Benson T, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild. In: Proc. of the ACM SIGCOMM Conf. on Internet Measurement. New York: ACM, 2010. 267–280. [doi: 10.1145/1879141.1879175]

[6] Benson T, Anand A, Akella A, Zhang M. Understanding data center traffic characteristics. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2010. 92–99. [doi: 10.1145/1672308.1672325]

[7] Roy A, Zeng H, Bagga J, Porter G, Snoeren A C. Inside the social network's (datacenter) network. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2015. 123–137. [doi: 10.1145/2829988.2787472]

[8] Noormohammadpour M, Raghavendra CS. Datacenter traffic control: Understanding techniques and tradeoffs. IEEE Communications Surveys & Tutorials, 2018,20(2):1492–1525. [doi: 10.1109/COMST.2017.2782753]

[9] Hopps C. Analysis of an equal-cost multi-path algorithm. RFC 2992, 2000. [doi: 10.17487/RFC2992]

[10] Zhou J, Tewari M, Zhu M, Kabbani A, Poutievski L, Singh A, Vahdat A. WCMP: Weighted cost multipathing for improved fairness in data centers. In: Proc. of the European Conf. on Computer Systems. New York: ACM, 2014. 1–14. [doi: 10.1145/2592798.2592803]

[11] Cao J, Xia R, Yang P, Guo C, Lu G, Yuan L, Zheng Y, Wu H, Xiong Y, Maltz D. Per-packet load-balanced, low-latency routing for clos-based data center networks. In: Proc. of the ACM Conf. on Emerging Networking EXperiments and Technologies. New York: ACM, 2013. 49–60. [doi: 10.1145/2535372.2535375]

[12] Dixit A, Prakash P, Hu YC, Kompella RR. On the impact of packet spraying in data center networks. In: Proc. of the Int'l Conf. on Computer Communications. Piscataway: IEEE, 2013. 2130–2138. [doi: 10.1109/INFCOM.2013.6567015]

[13] Ghorbani S, Godfrey B, Ganjali Y, Firoozshahian A. Micro load balancing in data centers with DRILL. In: Proc. of the ACM Workshop on Hot Topics in Networks. New York: ACM, 2015. 17–23. [doi: 10.1145/2834050.2834107]

[14] Ghorbani S, Yang Z, Godfrey P, Ganjali Y, Firoozshahian A. DRILL: Micro load balancing for low-latency data center networks. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2017. 225–238. [doi: 10.1145/3098822.3098839]

[15] Mitzenmacher M. The power of two choices in randomized load balancing. IEEE Trans. on Parallel Distribution System, 2001, 12(10):1094–1104. [doi: 10.1109/71.963420]

[16] Sinha S, Kandula S, Katabi D. Harnessing TCP's burstiness with flowlet switching. In: Proc. of the ACM Workshop on Hot Topics in Networks. New York: ACM, 2004.

[17] Kandula S, Katabi D, Sinha S, Berger A. Dynamic load balancing without packet reordering. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2007. 51–62. [doi: 10.1145/1232919.1232925]

[18] Alizadeh M, Edsall T, Dharmapurikar S, Vaidyanathan R, Chu K, Fingerhut A, Lam VT, Matus F, Pan R, Yadav N, Varghese G. CONGA: Distributed congestion-aware load balancing for datacenters. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2014. 503–514. [doi: 10.1145/2619239.2626316]

[19] Katta N, Hira M, Ghag A, Kim C, Keslassy I, Rexford J. CLOVE: How I learned to stop worrying about the core and love the edge. In: Proc. of the ACM Workshop on Hot Topics in Networks. New York: ACM, 2016. 155–161. [doi: 10.1145/3005745.3005751]

[20] Katta N, Hira M, Kim C, Sivaraman A, Rexford J. Hula: Scalable load balancing using programmable data planes. In: Proc. of the Symp. on SDN Research. ACM, 2016. 10–23. [doi: 10.1145/2890955.2890968]

[21] Vanini E, Pan R, Alizadeh M, Taheri P, Edsall T. Let it flow: Resilient asymmetric load balancing with flowlet switching. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2017. 407–420.

[22] Mahalingam M, Dutt D, Duda K, Agarwal P, Kreeger L, Sridhar T, Bursell M, Wright C. Virtual extensible local area network (VXLAN): A framework for overlaying virtualized layer 2 networks over layer 3 networks. RFC 7348, 2014. [doi: 10.17487/RFC7348]

[23] Pan T, Song E, Bian Z, Lin X, Peng X, Zhang J, Huang T, Liu B, Liu Y. INT-path: Towards optimal path planning for in-band network-wide telemetry. In: Proc. of the Int'l Conf. on Computer Communications. Piscataway: IEEE, 2019. 487–495. [doi: 10.1109/INFOCOM.2019.8737529]

[24] Bosshart P, Daly D, Gibb G, Izzard M, McKeown N, Rexford J, Schlesinger C, Talayco D, Vahdat A, Varghese G, Walker D. P4: Programming protocol-independent packet processors. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2014. 87–95. [doi: 10.1145/2656877.2656890]

[25] He K, Rozner E, Agarwal K, Felter W, Carter J, Akella A. Presto: Edge-based load balancing for fast datacenter networks. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2015. 465–478. [doi: 10.1145/2829988.2787507]

[26] Agarwal K, Dixon C, Rozner E, Carter J. Shadow Macs: Scalable label-switching for commodity ethernet. In: Proc. of the ACM Workshop on Hot Topics in Software Defined Networking. New York: ACM, 2014. 157–162. [doi: 10.1145/2620728.2620758]

[27] Al-Fares M, Radhakrishnan S, Raghavan B, Huang N, Vahdat A. Hedera: Dynamic flow scheduling for data center networks. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2010. :89–92.

[28] Benson T, Anand A, Akella A, Zhang M. MicroTE: Fine grained traffic engineering for data centers. In: Proc. of the ACM COnf. on Emerging Networking EXperiments and Technologies. New York: ACM, 2011. 8–20. [doi: 10.1145/2079296.2079304]

[29] Curtis AR, Kim W, Yalagandula P. Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection. In: Proc. of the Int'l Conf. on Computer Communications. Piscataway: IEEE, 2011. 1629–1637. [doi: 10.1109/INFCOM.2011.5934956]

[30] Curtis AR, Mogul JC, Tourrilhes J, Yalagandula P, Sharma P, Banerjee S. DevoFlow: Scaling flow management for high-performance networks. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2011. 254–265. [doi: 10.1145/2018436.2018466]

[31] Sen S, Shue D, Ihm S, Freedman MJ. Scalable, optimal flow routing in datacenters via local link balancing. In: Proc. of the ACM Conf. on Emerging Networking Experiments and Technologies. New York: ACM, 2013. 151–162. [doi: 10.1145/2535372.2535397]

[32] Kabbani A, Vamanan B, Hasan J, Duchene F. Flowbender: Flow-level adaptive routing for improved latency and throughput in datacenter networks. In: Proc. of the ACM Conf. on emerging Networking EXperiments and Technologies. New York: ACM, 2014. 149–160. [doi: 10.1145/2674005.2674985]

[33] Wang W, Sun Y, Zheng K, Kaafar MA, Li D, Li Z. Freeway: Adaptively isolating the elephant and mice flows on different transmission paths. In: Proc. of the IEEE Int'l Conf. on Network Protocols. Piscataway: IEEE, 2014. 362–367. [doi: 10.1109/ICNP.2014.59]

[34] Wang P, Xu H, Niu Z, Han D, Xiong Y. Expeditus: Congestion-aware load balancing in clos data center networks. In: Proc. of the ACM Symp. on Cloud Computing. New York: ACM, 2016. 442–455. [doi: 10.1109/TNET.2017.2731986]

[35] Correa JR, Goemans MX. Improved bounds on nonblocking 3-stage clos networks. SIAM Journal on Computing, 2007,37(3): 870–894. [doi: 10.1137/060656413]

[36] Shafiee M, Ghaderi J. A simple congestion-aware algorithm for load balancing in datacenter networks. IEEE/ACM Trans. on Networking (TON), 2017,25(6):3670–3682. [doi: 10.1109/TNET.2017.2751251]

[37] Shafiee M, Ghaderi J. An improved bound for minimizing the total weighted completion time of coflows in datacenters. IEEE/ACM Trans. on Networking (TON), 2018,26(4):1674–1687. [doi: 10.1109/TNET.2018.2845852]

[38] Raiciu C, Barre S, Pluntke C, Greenhalgh A, Wischik D, Handley M. Improving datacenter performance and robustness with multipath TCP. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2011. 266–277. [doi: 10.1145/2018436.2018467]

[39] Cao Y, Xu M, Fu X, Dong E. Explicit multipath congestion control for data center networks. In: Proc. of the ACM Conf. on Emerging Networking Experiments and Technologies. New York: ACM, 2013. 73–84. [doi: 10.1145/2535372.2535384]

[40] Zhuo D, Zhang Q, Liu V, Krishnamurthy A, Anderson T. Rack-level congestion control. In: Proc. of the ACM Workshop on Hot Topics in Networks. New York: ACM, 2016. 148–154. [doi: 10.1145/3005745.3005772]

[41] Kheirkhah M, Wakeman I, Parisis G. MMPTCP: A multipath transport protocol for data centers. In: Proc. of the Int'l Conf. on Computer Communications. Piscataway: IEEE, 2016. 1–9. [doi: 10.1109/INFOCOM.2016.7524530]

[42] Li Z, Bi J, Zhang Y, Dogar AB, Qin C. VMS: Traffic balancing based on virtual switches in datacenter networks. In: Proc. of the IEEE Int'l Conf. on Network Protocols. Piscataway: IEEE, 2017. 1–10. [doi: 10.1109/ICNP.2017.8117566]

[43] Patel P, Bansal D, Yuan L, Murthy A, Greenberg A, Maltz DA, Kern R, Kumar H, Zikos M, Wu H, Kim C, Karri N. Ananta: Cloud scale load balancing. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2013. 207–218. [doi: 10.1145/2486001.2486026]

[44] Gandhi R, Liu HH, Hu YC, Lu G, Padhye J, Yuan L, Zhang M. Duet: Cloud scale load balancing with hardware and software. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2014. 27–38. [doi: 10.1145/2619239.2626317]

[45] Eisenbud DE, Yi C, Contavalli C, Smith C, Kononov R, Hielscher EM, Cilingiroglu A, Cheyney B, Shang W, Hosein JD. Maglev: A fast and reliable software network load balancer. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2016. 523–535.

[46] Miao R, Zeng H, Kim C, Lee J, Yu M. SilkRoad: Making stateful layer-4 load balancing fast and cheap using switching ASICs. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2017. 15–28. [doi: 10.1145/3098822. 3098824]

[47] Olteanu V, Agache A, Voinescu A, Raiciu C. Stateless datacenter load-balancing with beamer. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2018. 125–139.

[48] Zats D, Das T, Mohan P, Borthakur D, Katz R. DeTail: Reducing the flow completion time tail in datacenter networks. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2012. 139–150. [doi: 10.1145/2342356.2342390]

[49] Perry J, Ousterhout A, Balakrishnan H, Shah D, Fugal H. Fastpass: A centralized zero-queue datacenter network. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2015. 307–318. [doi: 10.1145/2619239.2626309]

[50] Zhang H, Zhang J, Bai W, Chen K, Chowdhury M. Resilient datacenter load balancing in the wild. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2017. 253–266. [doi: 10.1145/3098822.3098841]

[51] Handley M, Raiciu C, Agache A, Voinescu A, Moore AW, Antichi G, Wojcik M. Re-architecting datacenter networks and stacks for low latency and high performance. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2017. 29–42. [doi: 10.1145/3098822. 3098825]

[52] Chen L, Lingys J, Chen K, Liu F. AuTO: Scaling deep reinforcement learning for datacenter-scale automatic traffic optimization. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2018. 191–205. [doi: 10.1145/3230543. 3230551]

[53] Mittal R, Shpiner A, Panda A, Zahavi E, Krishnamurthy A, Ratnasamy S, Shenker S. Revisiting network support for RDMA. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2018. 313–326. [doi: 10.1145/3230543.3230557]

[54] Cheng P, Ren F, Shu R, Lin C. Catch the whole lot in an action: Rapid precise packet loss notification in data center. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2014. 17–28.

[55] Zhu Y, Eran H, Firestone D, Guo C, Lipshteyn M, Liron Y, Padhye J, Raindel S, Yahia MH, Zhang M. Congestion control for large-scale RDMA deployments. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2015. 523–536. [doi: 10.1145/2829988.2787484]

[56] Silver D, Huang A, Maddison C, Guez A, Sifre L, Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lilicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of Go with deep neural networks and tree search. Nature, 2016,529(7587):484–489. [doi: 10.1038/nature16961]

[57] Bai W, Chen L, Chen K, Han D, Tian C, Wang H. Information-agnostic flow scheduling for commodity data centers. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2015. 455–468.

[58] Montazeri B, Li Y, Alizadeh M, Ousterhout J. Homa: A receiver-driven low-latency transport protocol using network priorities. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2018. 221–235. [doi: 10.1145/3230543. 3230564]

[59] Xia Y, Sun XS, Dzinamarira S, Wu D, Huang XS, Ng TSE. A tale of two topologies: Exploring convertible data center network architectures with flat-tree. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2017. 295–308. [doi: 10.1145/ 3098822.3098837]

[60] Xu Z, Tang J, Meng J, Zhang W, Wang Y, Liu CH, Yang D. Experience-driven networking: A deep reinforcement learning based approach. In: Proc. of the Int'l Conf. on Computer Communications. Piscataway: IEEE, 2018. [doi: 10.1109/INFOCOM.2018.8485853]

[61] Le Y, Chang H, Mukherjee S, Wang L, Akella A, Swift MM, Lakshman TV. UNO: Uniflying host and smart NIC offload for flexible packet processing. In: Proc. of the Symp. on Cloud Computing. New York: ACM, 2017. 506–519. [doi: 10.1145/3127479. 3132252]

[62] Li Y, Wei D, Chen X, Song Z, Wu R, Li Y, Jin X, Xu W. DumbNet: A smart data center network fabric with dumb switches. In: Proc. of the EuroSys Conf. New York: ACM, 2018. 9:1–9:13. [doi: 10.1145/3190508.3190531]

[63] Jin X, Li X, Zhang H, Soule R, Lee J, Foster N, Kim C, Stoica I. NetCache: Balancing key-value stores with fast in-network caching. In: Proc. of the Symp. on Operating Systems Principles. New York: ACM, 2017. 121–136. [doi: 10.1145/3132747. 3132764]

[64] Chole S, Fingerhut A, Ma S, Sivaraman A, Vargaftik S, Berger A, Mendelson G, Alizadeh M, Chuang ST, Keslassy I, Orda A, Edsal T. DRMT: Disaggregated programmable switching. In: Proc. of the Conf. on Special Interest Group on Data Communication. New York: ACM, 2017. 1–14. [doi: 10.1145/3098822.3098823]

[65] Sharma NK, Liu M, Atreya K, Krishnamurthy A. Approximating fair queueing on reconfigurable switches. In: Proc. of the Conf. on Networked Systems Design and Implementation. Berkeley: USENIX, 2018. 1–16.

沈耿彪(1991－),男,硕士生,主要研究领域为数据中心,负载均衡,流量调度,协议栈,软件定义网络,内容分发网络.

汪漪(1983－),男,博士,CCF 专业会员,主要研究领域为未来网络体系架构,信息中心网络,软件定义网络,高性能网络器件设计与实现,智能化网络.

李清(1985－),男,博士,CCF 专业会员,主要研究领域为网络体系架构,路由协议与算法,网络传输调度,边缘计算.

徐明伟(1971－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络.

江勇(1975－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为计算机网络体系结构,下一代互联网,人工智能网络.