低时延网络：架构，关键场景与研究展望

左旭彤，王莫为，崔勇
（清华大学信息科学技术学院，北京 100084）

摘　要：随着时延敏感型应用和超低时延场景的出现，低时延网络的研究受到了学术界、工业界和标准组织的广泛关注。了解时延产生的原因并设计相应的降低时延的技术使新兴应用的发展成为可能。首先按照网络的分层体系架构对时延来源进行分析，并以此为基础对降低时延的技术进行了综述。然后，针对数据中心网络、5G 以及边缘计算 3 种典型的低时延关键场景及优化时延的技术进行分析。最后，从网络架构革新、数据驱动优化时延算法及新协议设计 3 个方面展望了低时延网络发展面临的机遇与挑战。

关键词：低时延；网络架构；时延来源；数据中心网络；第五代移动通信技术；边缘计算
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019175

Low-latency networking: architecture, key scenarios and research prospect

ZUO Xutong, WANG Mowei, CUI Yong
School of Information and Technology, Tsinghua University, Beijing 100084

Abstract: With the advent of delay-sensitive applications and ultra-low latency scenarios, research on low-latency networking is attracting attention from academia, industry, and standards organizations. Understanding the causes of latency and designing corresponding techniques to reduce latency enable the development of emerging applications. The sources of latency according to the layered architecture of the network was analyzed, and summarizes the techniques for reducing the latency. After that, three typical low-latency key scenarios and delay optimization techniques for data center network, 5G and edge computing was analyzed. Finally, the opportunities and challenges that may be encountered in the development of low latency networks were presented from the perspectives of network architecture innovation, data-driven latency optimization algorithm and the design of new protocols.

Key words: low latency, network architecture, causes of latency, data center network, 5G, edge computing

1　引言

新应用和新场景的出现使网络空间更加复杂，不同的应用和场景对于网络的性能要求不尽相同，时延是影响性能的重要评价指标之一。对于游戏、直播等应用，时延是影响其用户体验的决定性因素；而对于物联网（IoT, Internet of things）、自动驾驶等场景，时延则决定其能否正常工作。

计算机网络体系采用分层架构，网络功能被解耦并分配在不同层，每一层按照不同的协议实现各自的功能，共同完成数据传输过程。然而，每层协议或功能的完成可能会使应用的时延增加，比如在 TCP（transmission control protocol）中通过重传机制来保证可靠传输，但是这可能会增加数据分组的

收稿日期：2019–05–23；**修回日期**：2019–06–12
通信作者：崔勇，cuiyong@tsinghua.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2018YFB1800303）；国家自然科学基金资助项目（No.61872211）
Foundation Items: The National Key Research and Development Program of China (No.2018YFB1800303), The National Natural Science Foundation of China (No. 61872211)

传输时延，因此低时延的实现需要每一层的努力。同时，对于特定的低时延场景，如数据中心网络、5G 网络和边缘计算，传统的分层体系架构针对一般网络提出的低时延技术可能不再适用，需要结合场景不同特点进行时延优化。

此前有综述工作按照时延的来源对网络中降低时延的技术进行详尽的分类，提供了对时延产生原因的全面分析[1-2]，但是这些工作提出的技术可能无法适用于特定场景的超低时延需求或者新架构。对于不同的低时延关键场景，也有分别针对数据中心网络[3-4]、5G 网络[5-6]及边缘计算[7-8]的综述，但是它们仅关注某一个具体的场景，无法提供对网络体系结构时延来源的整体分析与技术的泛化迁移。

本文在分析了应用时延需求的基础上，按照传统的网络层次架构阐述了时延的来源及多种影响时延的因素，如网络负载、路由决策等，并介绍了能够减少协议机制与网络功能引入时延的相关技术。

除了对传统体系结构中时延进行分析，本文还将低时延分析具体化至数据中心网络、5G 网络、边缘计算等关键场景。它们分处“云、管、端”的不同位置，互相配合共同构建整个低时延网络架构。与广域网不同，数据中心具备更高带宽更低时延的特性，并可以灵活部署。利用其特性设计传输协议和优化网络拓扑结构可以降低数据中心中任务处理的时延。随着 5G 的发展，超高数据率和超低时延成为可能，5G 为优化时延在架构调整和关键技术上做出努力。物联网的兴起使处在网络边缘的设备产生的数据量急剧增加，这加重了云端计算和网络传输负载。边缘计算通过将计算与存储下移至网络边缘，降低传输及云端计算负载，避开网络传输瓶颈并缩短传输距离，为用户提供高带宽低时延的服务。

2　应用时延需求

时延敏感型应用和超低时延场景的出现对时延提出了严格的要求。几种典型应用的时延需求及需要低时延的原因如表 1 所示，本节对每一种应用进行简单介绍。最后以直播场景为例，分析该应用时延的组成。低时延网络的研究使这些应用性能得到提升，进而优化用户体验。

在直播、虚拟现实（VR, virtual reality）、增强现实（AR, augmented reality）等用户参与度较高的场景下，时延需求主要来自人与人或者人与设备之间的流畅交互。在互动直播场景下，时延超过 1 s 会极大影响主播与观众的用户体验[9]。VR、AR 场景中要求设备能对人给出的信号做出及时的反应，因此，人机交互体验的优化对时延提出了 10 ms 的高要求[10]。

表 1　　几种典型应用的时延需求及其原因

应用/场景	时延需求	需求原因
直播[9]	<1 s	人与人互动
VR/AR[10]	10 ms	人机交互
自动驾驶[11]	1~10 ms	快速应对环境变化人机交互
物联网/工业互联网[12]	1~10 ms	设备与设备通信及协调工作

在自动驾驶场景下，车辆要在复杂的交通环境中及时感受到环境的变化并做出反应，这要求车辆与车辆、行人、道路设施之间进行低时延通信，其时延需求达到几毫秒[11]。另外，自动驾驶设备需要进行人机交互，尤其是切换驾驶模式的情况下，要求设备模式切换在复杂的环境中不出差错，这也需要超低时延。

在 IoT 的场景下，人与设备或设备与设备在协调工作时需要通过网络进行通信。在工业互联网的场景下，工厂实现高效率的自动化生产需要完成实时的操作控制，如果生产的某些步骤因未及时接收到指令而出现滞后便会影响产品质量甚至导致系统崩溃，因此工业互联网对于时延也提出了较高的要求，达到 1~10 ms[12]。

根据本节之前所述，不同的应用和场景对网络时延有不同的需求，低时延网络的研究使这些应用成为可能。为了满足应用的低时延需求，首先需要探究时延的可能来源及降低时延的技术，本节接下来会以流媒体直播应用为例进行简要阐述。

随着视频直播的兴起，直播的质量与用户体验受到学术界与工业界的广泛关注，时延是其中一个重要的评价指标。目前在比较流行的交互式直播中，观众会与主播进行互动并希望得到及时的回应。为了实现更加流畅的交互体验，在实际应用中会采用时延较低的实时消息协议（RTMP, real-time messaging protocol），此时主播与观众之间的端到端时延可以划分为 3 个部分，分别是主播端视频内容上传到服务器的时延、将视频从服务器下载到客户端的时延和下载内容在客户端缓冲的时延[13]。

对于直播来说，播放流畅和低时延都有助于获

得良好的用户体验。为防止因网络抖动造成视频卡顿，客户端中往往会设置视频缓冲区，但这会导致下载到客户端的视频帧不能被立即播放。视频帧到达客户端的时间与该帧被播放时间的差值就是在客户端缓冲区的时延，测量和研究表明客户端缓冲的时延在端到端时延中占有最大的比例[13]。有众多研究关注客户端缓冲时延优化问题，调整客户端播放逻辑、优化传输等方式从网络架构的不同层次优化了缓冲时延，同时不会导致视频卡顿。

调整客户端的播放逻辑以适应网络状况的变化需在应用层做出优化。在客户端，调整预缓冲的数据量[14]及在播放过程中根据网络抖动情况进行缓冲区大小的调整可以降低客户端缓冲时延[15]。缓冲区的历史长度可以反映网络抖动状况，在网络抖动小时，缩短缓冲区长度可以实现低的客户端缓冲时延。对于下层的数据传输，也需要明确时延的来源并进行优化。以传输协议选择为例，在直播中应选择不需要切片的低时延的 RTMP 协议而非 HLS（HTTP live streaming）协议，但是上述 2 种协议均基于 TCP，为了进一步降低时延，可以设计基于 UDP（user datagram protocol）的专用传输协议。

通过上述的分析可知，降低应用感受到的时延需要网络架构从上层到下层的共同努力。在第 3 节，会按照网络的分层体系架构，详细地分析时延的来源并介绍相应的降低时延的技术。

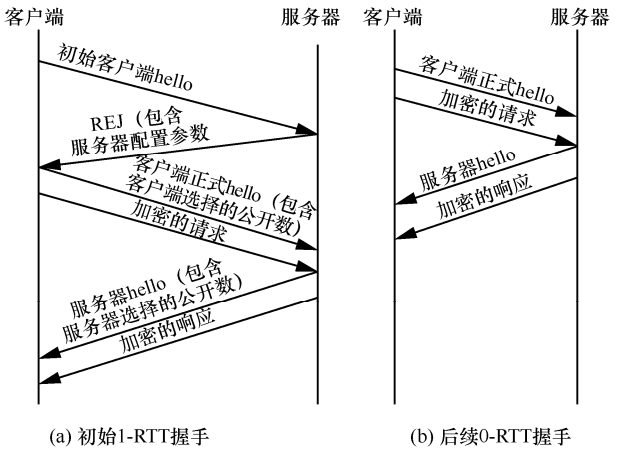
3 分层模型中的低时延

为使复杂的计算机网络系统简化，网络体系结构被设计为分层模型，每一层都有特定的功能，以完成数据通信的过程。网络体系结构中每一层通信协议和处理机制的设计极大地影响了数据传输的性能，本节将按照网络体系结构的分层模型对网络中的时延来源进行分类，介绍为了降低时延而设计的协议或技术并分析其优缺点。

3.1 传输层时延及降低时延的技术

第 2 节提到的 RTMP、HLS 等协议在传输层都基于 TCP，相比于非可靠传输的 UDP 协议，保证端到端之间可靠且按序传输会引入较多的时延，包括连接建立的时延、慢启动的时延、保证数据分组到达的丢失恢复时延、队头阻塞时延等。接下来将以 TCP 为例，对上述时延以及优化技术进行分析。

在使用 TCP 进行数据传输之前需要完成 3 次握手以与接收端建立连接，若使用安全加密的 Web 服务则还存在 SSL/TLS 握手。降低握手（控制信息的交互）的次数，减少数据分组往返时间（RTT, round-trip time）个数，可以有效地降低传输层时延，这种方式对降低短流时延具有显著效果。谷歌[16]于 2011 年提出 TFO（TCP fast open），目标是在握手的同时进行数据传输，它通过使用 cookie 实现，在确认字符（ACK, acknowledge character）回到接收端之前发送数据，该方案在 2014 年被 IETF（The Internet Engineering Task Force）组织标准化，但由于兼容性等问题并未被广泛使用。此后，谷歌公司提出的快速 UDP 网络连接（QUIC, quick UDP Internet connection）[17]中采用了类似 TFO 的技术，将传输握手和加密同时完成，实现一个 RTT 时间完成握手，如图 1(a)所示。在恢复会话时，客户端缓存的 cookie 和已经被加密的数据会直接发送到服务器端，服务器端利用此次的传输信息对客户端进行验证，如果验证通过就接收数据，从而完成零 RTT 的握手，握手过程如图 1(b)所示。



(a) 初始1-RTT握手 (b) 后续0-RTT握手

图 1 QUIC 中握手时延优化

对于新建立的 TCP 连接，初始拥塞窗口的设置以及增长速度对于短流的流完成时间非常关键，有很多针对于此的研究。Allman 等[18]于 2002 年提出可以将 TCP 初始窗口从一个最大分组长度（MSS, maximum segment size）增加到 4 个 MSS 而不会导致拥塞崩溃。Dukkipati[19]于 2010 年提出将初始窗口数从 4 提高到大于或者等于 10，相较于 Allman 等提出的方法，此方案更加激进，在高 RTT 和高带宽时延积的网络中，HTTP 请求的响应平均时延降低了 10%左右，在低带宽的网络中时延也有一定程度的提升。但是之前的方案对初始窗口的设置相对固定，灵活度低，Wang 等[20]提出在大带宽高时延

场景，对慢启动阶段做修改，自适应地反复重置慢启动阈值。通过在启动阶段适应网络条件，发送方能够快速增长拥塞窗口，而不会引起缓冲区溢出和多次丢失分组的风险。

在保证按序传输的协议中，若在传输过程中因分组丢失或选路不同导致先发送的数据分组后到达接收端时，位于缓冲区的后续数据分组就被阻塞而不能提交到上层，队头阻塞现象发生进而增加应用感受到的时延。

如果队头阻塞是由于分组丢失引起的，可以通过尽快通知发送端分组丢失来加快重传，降低重传时延，从而缓解了队头阻塞的情况。在 TCP 的一些增强版本中加入快速重传机制（fast retransmit）[21]，重复 ACK 指示分组丢失，发送端可以在了解到网络发生拥塞的同时加快此丢失分组的重传过程。但是重复 ACK 的接收在检测分组丢失时仍有较长时延，可通过有效载荷方法（cutting payload）[22]在此方面进行优化。在 CP 方案中，如果数据分组到达交换机之后发现交换机上缓冲区不足，即会发生分组丢失，交换机可将此数据分组的有效载荷裁去，仅将数据分组头部传送到接收端，接收端收到数据分组头部后会通知发送端分组丢失及发生拥塞，而不需要等待多次 ACK 的到达甚至发生超时重传。

队头阻塞也可能是由于数据分组无序到达引起的，在多径 TCP 中数据分组无序问题是一个重点关注的问题。为了解决多径 TCP 中数据分组的乱序到达引起的阻塞，可以在数据开始发送前或者传输过程中通过合理调度来缓解阻塞的情况。滑动多径调度器（STMS, slide together multipath scheduler）[23]在数据发送前对数据分组进行调度，STMS 为 RTT 小的快速路径预分配数据分组并使序号大的数据分组在 RTT 大的慢路径上传输。设置通过走快速路径和慢速路径的数据分组的分界点，STMS 可以实现分组顺序到达并缓解因无序到达而造成的队头阻塞。如果在传输中出现队头阻塞，可以使用机会重传（opportunistic retransmission）[24]机制，将造成队头阻塞的数据分组在可能会有可用拥塞窗口的子流上重传。

近几年提出的新协议中也有解决队头阻塞的技术。比如，在 Langley 等[25]提出的 QUIC 中，一个连接支持多个流，每个 QUIC 数据分组都是由属于若干个流的数据帧组成的，在这种情况下，如果数据分组丢失或者先发送的数据分组晚到达，只会影响该分组中包含的数据流，其他数据流并不会发生队头阻塞。QUIC 以其众多优势而被广泛采纳和使用，并被 IETF 标准化为 HTTP 3.0[26]。

拥塞控制是传输层的重要功能，一个设计良好的拥塞控制算法不仅要最大化吞吐量，还应该实现低的排队时延。拥塞控制可以防止队列生成，进而直接减少排队时延。拥塞控制算法中有一类是基于时延的算法，它们将时延作为信号来管理拥塞，时延信号会比分组丢失和显示拥塞通知更及时地反映网络的队列状况。

目前，很多基于时延的拥塞控制算法被提出[27-29]，比如 Copa[28]方案，在该方案中，目标发送速率被设置成所测量到的排队时延的倒数，并且按此发送速率调整拥塞窗口，当发送速率超过目标速率就减少拥塞窗口，从而阻止了队列的生成。谷歌拥塞控制算法（Google congestion control）[29]将（单向）排队时延的梯度作为推断拥塞的信号，排队时延的梯度（导数）可以反应缓冲区的变化情况，提供了对缓冲区大小预测的能力。同时，GCC 使用卡尔曼滤波来估计排队时延梯度，设置该梯度的自适应阈值以控制增加或减少的速率，从而实现最小化缓冲区及时延的目标。

3.2　网络层时延与降低时延的技术

网络层的核心功能是选择路径，不同路径的端到端时延可能会有较大差别。除了端到端时延，路由算法还有可靠性、通信开销等性能指标，路由算法完成性能指标之间的权衡。例如，后压路由算法[30]可以实现最优的网络吞吐量，但是被证明端到端时延随路径跳数呈平方式增长。在该算法中，每个节点需要为每个流维护一个队列，而且一次只能服务一个队列，时延性能有待提升。针对此问题，Bui 等[31]提出减少每个节点维护的实际队列的方法，提升了后压式路由算法的端到端时延性能。

流量工程是优化网络流量分配方式的技术，其目标是通过负载均衡降低网络拥塞，不仅可以实现网络带宽利用率的提升，而且可以降低在拥塞节点的排队时延，进而降低数据分组到达接收端的时延[32]。流量工程的概念最初是在多标签交换网络中提出的[33]，在 2000 年被引入 IP 网络中[34]。根据流量需求的可用性以及进行流量调整的操作时间尺度可以将流量工程分为在线和离线两大类[32]。

对于 IP 网络中的离线流量工程算法，一般是修改内部网关协议中的链路上的权重，然后根据最短

路径进行路由[35]，以期获取端到端低时延。如果两点之间有不止一条最短路径，可以利用等价多路径路由（ECMP, equal-cost multi-path routing）对流量进行平均分配。但是 ECMP 方法无法考虑当前链路负载、时延等因素，因此平均分配可能不是最优策略。有研究提出非均等流量分配，可以在一个合理的最短路子集上进行流量均等分配[36]，或者设置指数级的代价函数以使对长路径设置的惩罚更高[37]，最终实现端到端时延的优化目标。然而，离线的算法无法根据网络负载的变化而进行动态变化，在线的流量工程弥补了这个缺点，但是动态地更新链路权重可能会导致路由振荡的问题[32]。

路由和流量工程共同决定了网络拓扑及拓扑内流量的分配，进而影响网络的拥塞情况与时延[38]，目前，关于拓扑及其上的路由策略的关系的研究较少，而拓扑与路由未合理匹配同样会导致拥塞或者高时延[39]。针对此问题，Gvozdiev 等[39]提出评估拓扑可用性的指标、将流量路由到可用的低时延路径的方案。该方案中的指标是可选路径可用性大于某个阈值的入网点对数与总入网点对的比值，此指标高表示可以在更多的链路周围路由到更短的路径而不引入过多的时延，所以此指标高的拓扑与路由方案匹配更易实现低时延和少拥塞的选路。与此同时，为防止突发流量造成的链路拥塞，会在链路上剩余部分容量，但是剩余容量的增加会使路径时延增加，所以剩余容量成为避免拥塞和降低路径时延的关键点。该研究探究了拓扑与路由的匹配关系，设计剩余容量的大小，提出能应对流量变化并可以利用拓扑路径多样性的方案，将时变的流量低时延无拥塞的路由到接收端。

3.3 链路层时延与降低时延的技术

数据链路层可解决共享介质的访问问题，信道接入时延是数据链路层时延的重要组成部分。对于非争用型的静态信道分配，会存在信号发送前等待时间长、信道利用率低等问题。对于争用型的信道分配方案，由于没有控制器进行统一管理，所以控制和计算开销低，但是会存在信道争用产生的冲突和等待的时延，尤其是在信道负载较高时，竞争的时延开销不能被忽略。

802.11n 中采用 CSMA/CA，帧聚合是一种降低此方案中冲突的技术。帧聚合[40]通过将多个帧聚合之后再发送，可以降低争用的概率，从而降低等待时延和冲突之后的重传开销，但是计算聚合也需要花费时间，需要做两者的权衡。在最新确定的下一代 Wi-Fi 标准 802.11be 中，时延和抖动被定为与高吞吐并行的项目优化目标，在数据链路层需要设计新的分布式的 CSMA/CA 机制，以优化信道接入并保证与部署在其中的独立接入点公平共存[41]。

对于需要超低时延通信的场景，如工业互联网，时延要求是几微秒到几毫秒。IEEE 802.1 时间敏感网络（TSN, time-sensitive networking）标准和相关研究已经寻求为超低时延通信网络提供链路层支持，以解决特定业务时延抖动大，时延范围无法确定等问题。在 TSN 数据链路层中引入帧抢占技术（802.3br[42]和 802.1Qbu[43]），为帧分配优先级。为了传输高优先级的帧，低优先级帧的传输可以被抢占，从而保证高优先级的帧不会被阻塞。另外，在帧调度方面，增加基于时隙的调度（802.1Qbv[44]），遵循时分多址（TDMA, time division multiple access）规则。该方案将不同优先级的帧分派到不同队列，并利用开关门机制决定帧传输。首先在每一个时隙根据门控情况及队列优先级情况决定可以传输帧的队列，之后在每一个队列中采用各自队列的帧调度策略，这样可以确保时延敏感的队列有确定的调度时间，使时延敏感的业务得到有保证的时延。

3.4 网络时延测量

不同层的时延都是端到端时延的一部分，而对包含单向时延及 RTT 的网络时延的测量是研究者了解并分析网络行为及性能的重要部分。对于单向时延的测量，需要解决的一个关键问题是收发两端本地时钟的同步问题，有很多针对于此的研究，如软件时钟同步法（如网络时间协议 NTP（network time protocol））与硬件时钟同步法（如利用全球定位系统 GPS 接收机），以及一些优化的时钟同步算法，如通过双向测量检测时钟调整与估计时钟偏差的 Paxson 算法[45]及 Moon 等[46]提出的线性规划算法，也可通过测量 RTT 绕过时钟同步的问题。

网络时延的测量方法众多，主动与被动时延测量是常用的测量方法。相比于主动向网络中发送测试数据来测量网络时延，被动时延测量更加节省网络带宽资源。然而目前在被动时延测量 RTT 方向时通常会利用传输层信息，如基于 TCP 时间戳的被动 RTT 测量[47]及利用拥塞控制或者流控特性的 RTT 测量方法[48]。但是如 QUIC 等正在部署的传输协议隐藏了被动 RTT 测量的信息。针对这个问题，De

Vaere 等[49]于 2018 年提出代替 TCP 时间戳的轻量级时延信号，此信号使用传输协议头部的三位，支持单流、单点及单向的 RTT 被动测量，使被动网络时延测量方法与传输独立。不过主动时延测量也有高灵活性等优势，所以目前在实际应用中也会采用主动与被动时延测量结合的方式。

4　低时延关键场景

第 3 节按照网络层次架构分析了网络中的时延及优化技术，本节将时延分析具体到 3 个关键的低时延场景：数据中心网络、5G 网络和边缘计算。这些场景因具有不同的特性而采用了不同的低时延技术。

4.1　数据中心网络

相比于广域网，数据中心网络具备更高带宽和更低时延的特性，而且可以灵活部署，这些特性使数据中心网络可以完成大量数据的快速存储和处理，成为大数据和云计算重要的基础设施。本节介绍数据中心网络中实现低时延的技术，主要分为两部分：一部分是传输层优化，优化拥塞控制与流调度以降低传输时延；另一部分是网络层拓扑结构优化，从而设计合理的网络拓扑结构来降低数据中心网络的时延。

4.1.1　拥塞控制与流调度

数据中心网络因为处理一些分布式的任务或者 Web 请求而存在大量的短流，这些短流一般数据量较小但是希望可以获得快速响应。数据中心网络也存在数量较少但是数据量较大的长流，这些长流对时延要求较低但是希望可以实现高吞吐量。关于满足数据中心网络中短流低时延与长流高吞吐需求的研究一直在进行。

DCTCP（datacenter TCP）[50]是专为数据中心设计的类似 TCP 的传输层控制协议。DCTCP 利用显式拥塞通知（ECN, explicit congestion notification）向终端提供多比特反馈。当遇到拥塞时，中间交换机对数据分组进行 ECN 标记，经由接收端通知发送端网络拥塞。发送端根据被 ECN 标记的数据分组比例，即网络的拥塞情况，来调节拥塞窗口，而不是直接将拥塞窗口减半，这样可以提升窗口恢复速度，并使交换机的缓冲队列维持在较低水平，大大降低短流时延的同时满足了长流的高吞吐量需求。

在数据中心网络中，为了防止时延敏感的短流被长流阻塞，可以使用优先级队列调度程序来提高它们的优先级，从而降低短流的完成时间及平均流完成时间。pFabric[51]就是采用了上述思想的数据中心流调度算法，该算法将流调度和速率控制解耦、流简化调度和速率控制，并最终提供了一个接近理论最优值的流完成时间。流调度是基于优先级的，交换机可以利用很小的缓冲区来实现，因此可以降低数据分组的排队等待时间。在速率控制方面，pFabric 不需要进行慢启动，开始时是线速发送，当发生长时间大量的分组丢失时再利用速率控制来降低发送速率。

传统的 TCP/IP 协议栈越来越不能满足新一代数据中心网络工作负载的超高吞吐超低时延的需求，CPU 目前处理数据分组的开销是不能被接受的[52-53]。第 3 节的分析针对传统的网络体系结构，无法从本质上解决端侧的处理时延开销，该问题的解决需要利用新的技术如远程直接内存访问（RDMA, remote direct memory access）、数据平面开发套件（DPDK, data plane development kit）等方式，在不需要端侧的操作系统参与的情况下，直接实现不同主机内存数据的传输和访问。

RDMA 是为解决网络传输中服务器端数据处理的时延而产生的，因为网络 I/O(input/output)存在瓶颈，所以数据中心网络的大带宽无法被充分利用。RDMA 的零拷贝技术，绕过内核处理，省去了中断处理和各种拷贝的时间[52]。数据在发送端与接收端进行处理时，会完成硬件设备与应用层的直接交互而不经过内核的处理，从而实现服务器端处理的低时延，具体如图 2 所示。

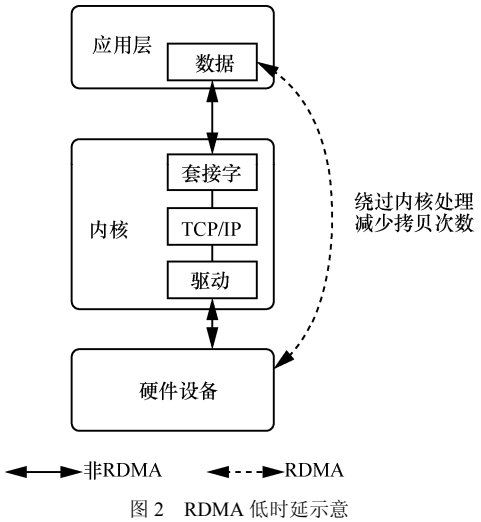


图 2　RDMA 低时延示意

RDMA 已经被部署到数据中心，目前有基于无

线宽带（InfiniBand）的 RDMA 网络，还有基于以太网的 RDMA 网络，即 RoCE （RDMA over converged ethernet）。基于 InfiniBand 的 RDMA 需要更换智能网卡和交换机，成本较高，而基于以太网的 RDMA 网络只需要更换网卡，成本较低。

在基于 IP 路由的数据中心网络上，RDMA 使用 RoCEv2 协议部署[53]，为实现数据链路层的无损传输，RoCE 采用基于优先级的流控（PFC, priority-based flow control），但是直接利用 PFC 操作粒度太粗，可能会面临队头阻塞问题，导致拥塞控制效果不足，若使用流级别的拥塞控制则可以解决此问题。量化拥塞通知（QCN, quantized congestion notification）被提出以解决上述问题，但是 QCN 是一个两层的协议，无法用于网络层。在这种情况下，Zhu 等[53]提出 DCQCN（data-center QCN）以解决 PFC 引入的问题。DCQCN 是一个为 RoCEv2 协议设计的流级别的拥塞控制算法，而且有不需要慢启动等优势，可以将队列长度维持在较低的水平，降低了端到端时延。

4.1.2 数据中心网络拓扑结构优化

数据中心网络拓扑结构对时延的影响主要体现在 2 个方面：网络拥塞和路由路径长度。一方面，数据中心网络的流量具有高动态性，易产生拥塞热点。当持续时间较长的大流发生拥塞时，交换机的缓冲区被填充，从而导致短流的时延增大。加之每类业务发生拥塞时往往同时涉及多条链路，进一步加重了热点拥塞对传输性能的影响。另一方面，端到端过长的路径会增加传输过程中的传播、处理和排队时延，这就要求拓扑结构要具有较小的网络直径。因此，除了传输协议外，合理的拓扑架构和路由方案同样有助于通过消除热点、实现负载均衡的方式降低数据中心网络的时延。对数据中心网络拓扑架构的研究大体经历了 3 个阶段：有线数据中心网络架构、光电交换混合架构和无线数据中心网络架构。

有线数据中心网络往往采用固定的分层树状结构，其传输性能受限于上层交换机的聚合效率，扩展性较差。在实际中，数据中心网络通常采用高超额订购比的结构来缓解流量高峰期的拥塞情况。之后出现了以 Fat-tree[54]和 VL2[55]为代表的新型树状拓扑结构，以 Dcell[56]和 BCube[57]为代表的分层递归拓扑结构以及以 SWDC (small-world datacenter)[58]和 JellyFish[59]为代表的随机小世界拓扑结构。其中，SWDC 通入增加随机链路将小世界模型引入拓扑设计，有效降低了网络直径，然而其采用的基于最短路的贪心路由算法会导致最差情况下网络极低的吞吐量和不佳的负载均衡。

上述网络拓扑结构的网络容量和传输效率与传统结构相比已有较大提升，但并未解决静态网络拓扑与动态热点流量之间的根本矛盾[60]。为了实现对热点流量的动态适配，以 C-through[61]、Helio[62]和 XFabric[63]为代表的光电交换混合架构提出引入光路交换（OCS, optical circuit switching）以实现可变拓扑。由于光电线路交换机往往有更大的网络带宽，当电交换机部分出现拥塞时，可以将流量导入光交换网络中，实现热点消除。另一方面，通过调整链路可以动态调整端到端路径长度，从而实现适应性的时延优化。然而，由于价格高昂，且切换开销较大，进行大规模实际部署商用 OCS 交换机将面临巨大的成本和效率考验。

除采用 OCS 交换机外，无线设备也可作为可变拓扑结构的组成部分。无线数据中心网络架构目前往往采用 60 GHz 无线电模块[64-65]或者空间激光收发器（FSO, free space optical）[66-67]作为基础模块进行搭建，从而实现高度灵活的链路调配。但当前的无线数据中心网络架构设计中，无线设备往往被部署在机架顶部，受到无线设备干扰和阻塞的限制，实际可使用的链路数量非常受限。虽然有工作提出在天花板安装平面镜或球状反射镜的方案来提高反射效率和精度，但这一方案对机架顶部的空间要求过于理想，很难部署。

为了解决这一问题，Wang 等[60]使用多反射环拓扑重新设计了无线数据中心网络架构，其俯视图如图 3 所示。利用部署在服务器上的无线网卡，无线信号可以被多次反射，实现与目标服务器的直连而不需要经过多跳，有效降低了传输路径长度，从而避免了中间设备中的排队和处理时延。

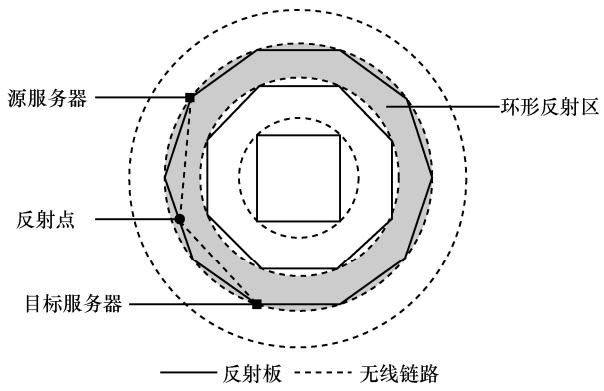


图 3 Diamond 无线反射示意

可变拓扑数据中心网络架构为消除热点、降低时延提供了可能，但仍需高效的拓扑自适应算法才能充分发挥其潜力。拓扑配置问题的离散性决定了基于整数线性规划的方法缺乏扩展性，而启发式算法往往性能不佳。为此，xWeaver[68]提出可以采用深度学习方法针对流量负载对拓扑进行动态配置。其将流量需求作为输入，学习输出近似最优拓扑配置，同时该方法可以灵活地支持流级别或应用程序级别的多样优化目标，包括流完成时间或 Hadoop 任务完成时间等。

4.2　5G 网络

与前四代移动通信不同，5G 旨在提供更高数据率、高可靠低时延及更广连接的服务。5G 的目标空口时延要求达到 1 ms[69]，其低时延特性使自动驾驶、设备到设备通信等应用成为可能。在高可靠低时延场景下会存在时延与可靠性的权衡，比如在自动驾驶、工业自动化等应用中，除了低时延外，高可靠性亦是上述应用正常工作的重要前提。因此，在这些应用中不应采用分组丢失等非可靠方式实现低时延，而应采取其他技术完成时延与可靠性的优化。本节将从网络架构调整和低时延关键技术这 2 个方面分析 5G 网络为实现超低时延的设计，这些设计在优化时延的同时不会造成可靠性的降低。

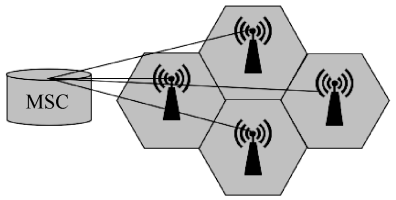
4.2.1　网络架构调整

移动通信网络架构演进呈现分离的趋势，功能的不断分离使部分功能可以灵活部署并下沉至更靠近用户的位置，这样可缩短用户与服务端的距离，进而降低网络时延。对于核心网，在 3G 网络中引入直接隧道技术（DT, direct tunnel）[70]，将控制面与用户面分离，数据传输时绕过服务 GPRS（general packet radio service）支持节点（SGSN, serving gprs support node），利用 DT 将基站与网关 GPRS 支持节点（GGSN, gateway GPRS support node）直接相连，这是核心网分离的开始。DT 技术的采用避免了 SGSN 对数据的处理与转发过程，缩短了时延。

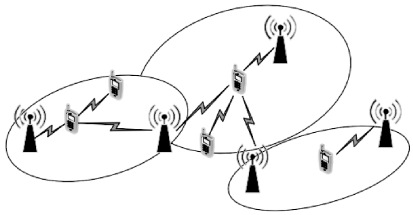
在 5G 时代，基于 4G 核心网，5G 核心网继续进行更完全的分离，SGSN、服务网关（SGW, serving gateway）、 PDN 网关（PGW, packet data network gateway）等网元被分为用户面与控制面两部分[71]。分离核心网用户平面并将用户面下沉到回传网之前可以减轻回传网传输压力与核心网集中处理负担。计算与存储的下沉与分布式架构可以使用户数据无需到达远距的核心网，从而实现毫秒级的时延目标。

对于接入网，5G 将基站分为集中单元（CU, centralized unit）、分布单元（DU, distributed unit）和有源天线单元（AAU, active antenna unit）3 个部分[71]，其中 CU 对应于 4G 网络中室内基带处理单元（BBU, building base band unit）的实时性低的部分，DU 对应实时服务。将 BBU 分离后，可以根据场景和需求灵活地对 CU 和 DU 进行部署，以满足 5G 中不同应用的需求，对于时延需求高的应用需求，可以将 DU 部署于离用户更近的地方。

为了提升数据传输速率并降低时延，提升带宽是一个直接的选择。目前的移动通信使用的是 3 GHz 以下的频段[72]，若提升带宽可以探索使用高频毫米波频段[73]。由于高频信号传播范围小，原来移动通信中以基站为中心的架构可以被调整为以用户为中心[74]，具体如图 4 所示。此时，用户不仅是网络中的节点，而且将参与网络中的中继、传输等任务。在这种情况下，基于 5G 毫米波和之前的 4G 组网方案，研究者们提出 5G-LTE 混合组网方案及仅基于毫米波基站的独立组网方案[75]。



(a) 以基站为中心的架构



(b) 以用户为中心的架构

图 4　从以基站为中心转变为以用户为中心

4.2.2　低时延关键技术

除了架构上的调整外，5G 融合了多种关键技术来优化时延，如利用波束赋形技术，大规模多输入多输出技术及新型正交多址接入技术提升频谱利用效率，优化用户与基站之间的空口时延，实现高速数据传输。5G 架构对于数据传输时延的优化是众多技术融合共同作用的结果。

在移动通信中多址接入是一个关键特征，其目

标是实现用户随时随地接入信道，多址接入技术一直在研究当中。第四代移动通信（4G）采用正交频分多址（OFDMA, orthogonal frequency division multiple access），为保证信号正交而设置大信号传输时间间隔（TTI, transmission time interval），存在频谱效率低、数据传输时延大等不足，无法满足新兴业务对时延的需求，新型的多址技术需要被提出。5G 中包含多种新型的多址接入技术，比如非正交多址（NOMA, non-orthogonal multiple access）、稀疏码多址（SCMA, sparse code multiple access）等。这些新的多址技术可以提升频谱利用率，减少竞争等待时间，从而降低时延。

NOMA[76]与传统的正交多址方案有着本质的区别，在 NOMA 中，功率域被引入，多个用户可以在同一时间、同一码型、同一频率、不同功率等条件进行传输。发送端非正交发送，接受端使用干扰消除进行解调，接收端的解调过程复杂性增加，但是避免了正交信号间隔时间，提升了频谱利用率。同时，在 NOMA 中不需要采用争用型信道分配方式，竞争等待时间缩短。SCMA[77]是一种基于码本与码字映射的非正交多址接入技术，数据经过信道编码之后会按照码本中的对应码字进行高维调制。SCMA 使用码分多址，通过使用多个载波组，实现频谱利用率的提高。

4.3 边缘计算

在传统的云计算架构中，云端完成数据存储与计算[78] 并通过网络将服务提供给用户。但是随着智能化普及，边缘设备产生数据急剧增加，云端计算和网络传输负载加重。同时由于云计算架构中数据传输距离长，可能无法满足时延敏感型应用的需求。为了保证数据处理低时延，降低云计算与传输负载，研究者提出了边缘计算架构。本节介绍边缘计算中的关键技术计算卸载以及利用边缘计算的具体应用。

随着移动设备性能的提升，开发者正在开发愈加复杂的应用程序，如自然语言处理或人脸识别等，需要大量的计算与存储资源，同时，低时延也是这些应用重要的评价标准之一。有研究发现这些应用程序通常由许多可组合组件组成[78]，所以可以利用计算卸载技术，确定如何进行组件的分配，使这些应用能在移动设备上运行，此问题被称为代码分割问题，有众多针对于此的研究[79-82]。

CloneCloud[80]在代码运行前计算分割情况，它通过对目标手机和云上的进程二进制文件的不同运行条件进行离线静态分析来确定这些卸载到云上的部分。但是，这种方法只考虑离线预处理中的有限输入/环境条件，无法涵盖真实的网络状况下的所有情况。MAUI（mobile assistance using infrastructure）[79]是在运行时进行分割决策的，它将此问题建模为一个整数线性规划问题，但是解此线性规划是一个 NP-hard 问题，求解时间不能忽略。Hermes[82]提供一种多项式时间近似方法，以最小化卸载请求的时延，但是它只适用于一个卸载请求情况。

不过上述工作都是针对通用的计算工作负载，面向特定的应用，可以根据应用的特点进行计算卸载。VR 系统有严格时延要求，有研究提出 VR 内容呈现可以分为交互式的前景和可预测的后景两部分，基于此提出一种基于计算卸载的渲染方案[10]，并证明了该方法的可行性。将可预测但渲染负载重的背景的预渲染和预取任务卸载到云端，而轻量级前景交互的渲染在移动端本地的 GPU（graphics processing unit）上完成，以绕过网络传输瓶颈，从而降低时延，优化用户体验。

实时视频分析是边缘计算的另一个重要的应用场景。视频分析的结果需要用来与用户进行交互或者启动下一个系统，所以需要低时延的支持。同时，传输高清视频需要高带宽，此时将大量的视频数据都传输到云端处理是不可行的[83]。利用边缘设备的计算与存储能力及地理上分布式的特点进行视频分析成为一项新兴并且必要的任务，进而为道路流量控制、安全监控等提供便利。

有研究提出地理上分布的公有云、私有集群和边缘的架构是唯一能够满足大规模实时视频分析的方法[83]，具体如图 5 所示，具体的视频分析任务可以分配到公有云、私有云或者边缘设备。

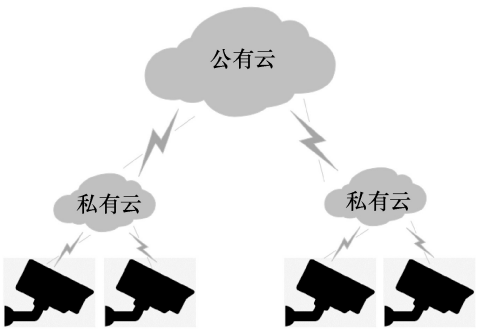


图 5 地理上分布的公有云、私有云与边缘架构

有许多研究工作关注实时视频分析领域。VideoStorm[84]系统中视频分析是在私有云上完成的。此系统由一个中央管理器和一组执行视频分析查询任务的工作机组成，每一个查询被连续输入到集群中，其中包含多个对视频进行处理的转换。Lavea[85]是一个建立在边缘计算平台之上的系统，它可以卸载客户端和边缘服务器之间的计算。同时，边缘服务器之间可以完成协作，以低传输时延或者低排队时延为目标。通过上述操作在靠近用户的地方提供低时延的视频分析。

5　机遇与挑战

5.1　基于 SDN 的网络架构革新

为了优化数据传输时延，满足应用的实时性要求，网络架构的革新与演进一直是研究热点之一。在保持传统网络基础架构不发生变化的前提下，可以利用基于 SDN 的软件化形式构造更加灵活的网络架构，使网络中策略的优化与部署成为可能，为时延优化提供机遇。下面将以基于 SDN 架构的数据中心广域网中的时延优化为例进行介绍。

连接数据中心的广域网对在线服务提供商来说是重要的基础设施，数据中心之间的低时延和高吞吐量数据传输可以提供良好的用户体验和高可靠的服务[86]。为了提供长距离的大容量链路，数据中心之间的广域网资源花费很高成本，但是由于防止分组丢失而设置的冗余链路[87]及分布式资源分配模型[86]造成的次优流量路由等原因使数据中心间广域网链路利用率并不高，很多企业对于本公司的数据中心之间的广域网连接进行改进以期望提升链路利用率进而优化时延，使用 SDN 进行改进是其中的一种方式。

以谷歌公司为例，谷歌公司提出 B4 网络实现其数据中心的互联[87]。B4 架构包含三层，从下到上依次是硬件设备、局部控制器和全局控制器。B4 采用定制的交换机及机制，将现有的路由协议集成在一个 SDN 的环境中，局部控制器控制物理层设备并将收集的链路拓扑信息发送到全局控制层的服务器。当需要进行路径选择时，全局控制器会对所需带宽进行预估并选择一条最优的路径，从而提升链路利用率，使可用带宽提升，阻塞减少，从而降低了时延。为实现基于 SDN 的网络架构革新，除了提出的 B4 之外，谷歌公司于 2015 年提出通过 SDN 实现的数据中心互联架构[88]，于 2017 年提出基于 SDN 的对等边缘路由基础架构[89]，于 2018 年提出网络功能虚拟化堆栈[90]。

虽然 SDN/NFV 技术提供了机遇，但是以虚拟化软件实现硬件功能的方式也同样使 SDN/NFV 新型网络架构难以避免地增加核心网络和数据中心网络的处理时延。有研究指出服务链时间延迟可能随着链的长度线性增长[91]，所以有研究提出网络功能并行化加速的算法 NFP（network function parallelism in NFV）[92]。该方法可智能识别 NF 依赖关系，并自动将策略编译成高性能服务图。然后该框架内的基础设施执行轻量级分组复制、分布式并行分组交付和分组副本的负载均衡合并，以支持 NF 并行性，最终实现在基本没有资源浪费的情况下优化时延。如何处理好软件化与低时延的权衡仍处在研究当中，这是推动 SDN/NFV 技术进一步发展的关键所在。

5.2　数据驱动的低时延优化算法

面对愈加复杂的需求和计算任务，以机器学习、尤其是深度学习为代表的数据驱动方法已经在各领域展开了广泛应用，也成为了网络领域工业界和学术界的关注热点[93-94]。为了降低网络时延，需要对大量复杂的优化和调度问题进行求解。但传统算法往往根据不真实的假设条件或者不准确的建模，采用基于固定策略的启发式算法，从而很难在动态多变的复杂网络环境下保持稳定的性能。机器学习算法则能够直接从数据中学习问题特征或者直接通过与环境交互学习进行决策，从而为解决这一问题提供了新的方向。

网络中大量的问题（如拥塞控制、流量调度等）可以建模为序列决策问题，擅长解决此类问题的强化学习技术也被应用到网络系统的多个方面。以 Skype 为代表的互联网电话（Internet telephony）发展迅猛，但却面临网络抖动带来的巨大性能挑战，因此如何选择合适的中继节点来降低通话时延这一问题亟待解决。Jiang 等[95]将这一问题建模为多臂老虎机问题，采用上限置信区间算法（UCB, upper confidence bound）实时为每对通话选择当前最优的中继节点，从而有效降低了通话的时延。

拥塞控制对减少网络拥塞、降低排队时延至关重要。在动态变化的网络环境下，基于规则的相对固定的窗口调节策略将不可避免的产生性能下降。Remy[96]首次将拥塞控制问题建模为马尔可夫决策过程，采用强化学习思想学习网络状态到窗口调节

方式的映射，从而实现细粒度的精确调节。另一方面，PCC（performance-oriented congestion control）[97]和 PCC vivace[98]采用在线学习方法，对网络环境进行在线适应，一定程度上缓解了机器学习方法的泛化问题。

数据中心中的流量调度对应用性能影响巨大。当前算法往往依赖于手工参数调节，从而导致网络环境与算法参数不匹配。Chen 等[99]提出利用深度强化学习根据当前的流量负载情况动态决策算法阈值，在保障处理效率的条件下，大幅降低了应用的流完成时间。

机器学习算法虽然优化了网络的时延，但也对网络系统提出了新的挑战，尤其是深度学习算法往往需要利用梯度下降算法进行模型更新。而对大规模机器学习任务，训练和计算往往以分布式的方式分发到多台计算机共同完成。这时，机器间的通信开销巨大，严重时会影响模型的训练速度。因此，未来如何进行架构和同步算法设计，如何利用 RDMA 和 DPDK 等技术进行传输优化都是亟待解决的问题。

5.3 低时延新兴协议设计

网络协议作为网络中通信的标准，是低时延网络构建中的重要一部分。然而，互联网在最初设计时提供的是尽力而为的服务，并未对网络时延等指标提供保证或作出优化，为了优化时延，针对不同的应用，学术界和工业界的研究者们提出了多种新兴协议，在协议层面为低时延通信做出努力。

为提升网页访问的速度，HTTP 2.0[100]与 QUIC[25]协议被提出，在数据传输的握手、连接、复用等方面进行改进。为加速数据中心数据传输，使服务器端系统处理速度匹配数据中心网络带宽，RDMA 技术及相匹配的 RoCE 等协议被提出。为使浏览器支持实时音视频传输，传输层采用 RTP 协议的 WebRTC[101]被提出。

与此同时，各企业为优化自身传输系统，根据自身的业务提出了特定的协议，提供了更低的数据传输时延。如 IBM 旗下的 Aspera 公司提出的广域网上海量数据传输的 FASP 传输技术，它避免了 TCP 在分组丢失率高时延高的链路上无法充分利用网络带宽的问题，优化了链路吞吐量提升了文件传输速度。快手研究者提出的基于 UDP 的 KTP（Kuaishou transport protocol）[102]传输协议，将码率和帧率自适应加入，并融合了网络性能估计与拥塞控制等，优化了传输时延及分组丢失等其他指标。为满足不同的应用需求，未来可能会有更多的专用协议出现，这可能成为一个不可或缺的研究点。

6 结束语

低时延网络对新兴应用的性能提升有重要意义，低时延技术是目前的研究热点。本文分析了 TCP/IP 网络架构各层时延的来源，并总结了实现低时延的各层技术。同时，对数据中心、5G 和边缘计算这 3 个关键场景时延优化进行了分析，希望本文的分析可以对该方向的研究提供一些启发。此外，低时延网络可以促进新型协议设计和数据驱动新方法的产生与发展。然而，机遇与挑战并存，所以希望仍有针对这一问题的持续和深入的研究，这需要学术界和工业界的共同努力。

参考文献：

[1] BRISCOE B, BRUNSTROM A, PETLUND A, et al. Reducing internet latency: a survey of techniques and their merits[J]. IEEE Communications Surveys & Tutorials, 2016, 18(3): 2149-2196.

[2] ZUO X, CUI Y, WANG M, et al. Low-latency networking: architecture, techniques, and opportunities[J]. IEEE Internet Computing, 2018, 22(5): 56-63.

[3] ZHANG J, REN F, LIN C. Survey on transport control in data center networks[J]. IEEE Network, 2013, 27(4): 22-26.

[4] XIA W, ZHAO P, WEN Y, et al. A survey on data center networking (DCN): infrastructure and operations[J]. IEEE communications surveys & tutorials, 2017, 19(1): 640-656.

[5] 张平, 陶运铮, 张治. 5G 若干关键技术评述[J]. 通信学报, 2016, 37(7): 15-29.
ZHANG P, TAO Y Z, ZHANG Z. Survey of several key technologies for 5G[J]. Journal on Communications, 2016, 37(7): 15-29.

[6] GUPTA A, JHA R K. A survey of 5G network: architecture and emerging technologies[J]. IEEE access, 2015(3): 1206-1232.

[7] MAO Y, YOU C, ZHANG J, et al. A survey on mobile edge computing: the communication perspective[J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2322-2358.

[8] MACH P, BECVAR Z. Mobile edge computing: a survey on architecture and computation offloading[J]. IEEE Communications Surveys & Tutorials, 2017, 19(3): 1628-1656.

[9] PANG H, ZHANG C, WANG F, et al. Optimizing personalized interaction experience in crowd-interactive livecast: a cloud-edge approach[C]//ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 1217-1225.

[10] LAI Z, HU Y C, CUI Y, et al. Furion: engineering high-quality immersive virtual reality on today's mobile devices[C]//The 23rd Annual International Conference on Mobile Computing and Networking. ACM, 2017: 409-421.

[11] SAMII S, ZINNER H. Level 5 by layer 2: time-sensitive networking for autonomous vehicles[J]. IEEE Communications Standards Maga-

zine, 2018, 2(2): 62-68.

[12] FETTWEIS G, BOCHE H, WIEGAND T, et al. The tactile internet-ITU-T technology watch report [J]. Geneva: ITU, 2014.

[13] WANG B, ZHANG X, WANG G, et al. Anatomy of a personalized livestreaming system[C]//The 2016 Internet Measurement Conference. ACM, 2016: 485-498.

[14] LIANG G, LIANG B. Balancing interruption frequency and buffering penalties in VBR video streaming[C]// IEEE International Conference on Computer Communications. IEEE, 2007: 1406-1414.

[15] WANG J, LEI W, XU P, et al. Adaptive media playout buffer management for latency optimization of mobile live streaming[C]// IEEE International Conference on Multimedia & Expo Workshops. IEEE, 2017: 369-374.

[16] CHENG Y, CHU J, RADHAKRISHNAN S, et al. TCP fast open: RFC 7413 [S].(2014–10)[2019–05–23].

[17] CUI Y, LI T, LIU C, et al. Innovating transport with QUIC: design approaches and research challenges[J]. IEEE Internet Computing, 2017, 21(2): 72-76.

[18] ALLMAN M, FLOYD S, PARTRIDGE C. Increasing TCP's initial window: RFC 3390 [S].(2002–10)[2019–05–23].

[19] DUKKIPATI N, REFICE T, CHENG Y, et al. An argument for increasing TCP's initial congestion window[J]. ACM Sigcomm Computer Communication Review, 2010, 40(3): 26-33.

[20] WANG R, PAU G, YAMADA K, et al. TCP startup performance in large bandwidth networks[C]//IEEE International Conference on Computer Communications. IEEE, 2004(2): 796-805.

[21] STEVENS W R. TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms: RFC 2001[S]. (1997–02)[2019–05–23].

[22] CHENG P, REN F, SHU R, et al. Catch the whole lot in an action: rapid precise packet loss notification in data center[C]// USENIX Conference on Networked Systems Design and Implementation. USENIX Association, 2014: 17-28.

[23] SHI H, CUI Y, WANG X, et al. STMS: improving MPTCP throughput under heterogeneous networks[C]// USENIX Annual Technical Conference. USENIX, 2018: 719-730.

[24] PAASCH C, FERLIN S, ALAY O, et al. Experimental evaluation of multipath TCP schedulers[C]//The ACM SIGCOMM workshop on Capacity sharing workshop. ACM, 2014: 27-32.

[25] LANGLEY A, RIDDOCH A, WILK A, et al. The quic transport protocol: design and internet-scale deployment[C]// The Conference of the ACM Special Interest Group on Data Communication. ACM, 2017: 183-196.

[26] BISHOP M. Hypertext transfer protocol version 3 (HTTP/3)[S]. (2019–07–09)[2019–05–23].

[27] TAN K, SONG J, ZHANG Q, et al. A compound TCP approach for high-speed and long-distance networks[C]// 25th IEEE International Conference on Computer Communications. IEEE, 2006: 1-12.

[28] ARUN V, BALAKRISHNAN H. Copa: practical delay-based congestion control for the Internet[C]//15th Symposium on Networked Systems Design and Implementation. ACM, 2018: 329-342.

[29] CARLUCCI G, DE CICCO L, HOLMER S, et al. Analysis and design of the Google congestion control for Web real-time communication (WebRTC)[C]//The 7th International Conference on Multimedia Systems. ACM, 2016: 13.

[30] TASSIULAS L, EPHREMIDES A. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multi-hop radio networks[C]//29th IEEE Conference on Decision and Control. IEEE, 1990: 2130-2132.

[31] BUI L X, SRIKANT R, STOLYAR A. A novel architecture for reduction of delay and queueing structure complexity in the back-pressure algorithm[J]. IEEE/ACM Transactions on Networking, 2011, 19(6): 1597-1609.

[32] WANG N, HO K H, PAVLOU G, et al. An overview of routing optimization for internet traffic engineering[J]. IEEE Communications Surveys & Tutorials, 2008, 10(1): 36-56.

[33] AWDUCHE D, MALCOLM J, AGOGBUA J, et al. Requirements for traffic engineering over MPLS: RFC 2702[S]. (1999–09) [2019–05–23].

[34] FORTZ B, THORUP M. Internet traffic engineering by optimizing OSPF weights[C]// IEEE Conference on Computer Communications, Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE, 2000(2): 519-528.

[35] HARMATOS J. A heuristic algorithm for solving the static weight optimisation problem in OSPF networks[C]// IEEE Global Telecommunications Conference. IEEE, 2001(3): 1605-1609.

[36] SRIDHARAN A, GUÉRIN R, DIOT C. Achieving near-optimal traffic engineering solutions for current OSPF/IS-IS networks[J]. IEEE/ACM Transactions On Networking, 2005, 13(2): 234-247.

[37] XU D, CHIANG M, REXFORD J. Link-state routing with hop-by-hop forwarding can achieve optimal traffic engineering[J]. IEEE/ACM Transactions on networking, 2011, 19(6): 1717-1730.

[38] AL-FARES M, RADHAKRISHNAN S, RAGHAVAN B, et al. Hedera: dynamic flow scheduling for data center networks[C]//The 7th USENIX Symposium on Networked Systems Design and Implementation. USENIX, 2010: 281-295.

[39] GVOZDIEV N, VISSICCHIO S, KARP B, et al. On low-latency-capable topologies, and their impact on the design of intra-domain routing[C]//The 2018 Conference of the ACM Special Interest Group on Data Communication. ACM, 2018: 88-102.

[40] SKORDOULIS D, NI Q, CHEN H H, et al. IEEE 802.11 n MAC frame aggregation mechanisms for next-generation high-throughput WLANs[J]. IEEE Wireless Communications, 2008, 15(1): 40-47.

[41] LÓPEZ-PÉREZ D, GARCIA-RODRIGUEZ A, GALATI-GIORDANO L, et al. IEEE 802.11 be - extremely high throughput: the next generation of Wi-Fi technology beyond 802.11 ax[J].Cornell University: arXiv:1902.04320, 2019.

[42] IEEE. IEEE standard for ethernet amendment 5: specification and management parameters for interspersing express traffic: 802.3br-2016[S]. 2016: 1–58.

[43] IEEE. IEEE standard for local and metropolitan area networks—bridges and bridged networks—amendment 26: frame preemption: 802.1Qbu-2016 [S]. 2016: 1–52.

[44] IEEE. IEEE Standard for local and metropolitan area networks—bridges and bridged networks—amendment 25: enhancements for scheduled traffic: 802.1Qbv-2015 [S]. 2016: 1–57.

[45] PAXSON V. On calibrating measurements of packet transit times[J]. ACM Sigmetrics Performance Evaluation Review. ACM, 1998, 26(1): 11-21.

[46] MOON S B, SKELLY P, TOWSLEY D. Estimation and removal of clock skew from network delay measurements[C]//IEEE Conference

on Computer Communications, Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE, 1999(1): 227-234.

[47] STROWES S D. Passively measuring TCP round-trip times[J]. Communications of the ACM, 2013, 56(10): 57-64.

[48] CARRA D, AVRACHENKOV K, ALOUF S, et al. Passive online RTT estimation for flow-aware routers using one-way traffic[C]//International Conference on Research in Networking. Springer, 2010: 109-121.

[49] DE VAERE P, BÜHLER T, KÜHLEWIND M, et al. Three bits suffice: explicit support for passive measurement of internet latency in QUIC and TCP[C]//The Internet Measurement Conference. ACM, 2018: 22-28.

[50] ALIZADEH M, GREENBERG A, MALTZ D A, et al. Data center TCP (DCTCP)[J]. ACM SIGCOMM computer communication review, 2011, 41(4): 63-74.

[51] ALIZADEH M, YANG S, SHARIF M, et al. PFABRIC: minimal near-optimal datacenter transport[C]//ACM SIGCOMM Computer Communication Review. ACM, 2013, 43(4): 435-446.

[52] GUO C, WU H, DENG Z, et al. RDMA over commodity ethernet at scale[C]//The 2016 ACM SIGCOMM Conference. ACM, 2016: 202-215.

[53] ZHU Y, ERAN H, FIRESTONE D, et al. Congestion control for large-scale RDMA deployments[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 523-536.

[54] AL-FARES M, LOUKISSAS A, VAHDAT A. A scalable, commodity data center network architecture[C]//ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. ACM, 2008: 63-74.

[55] GREENBERG A, HAMILTON J R, JAIN N, et al. VL2: a scalable and flexible data center network[J]. Communication of the ACM, 2009, 39(4): 51-62.

[56] GUO C, WU H, TAN K, et al. Dcell: a scalable and fault-tolerant network structure for data centers[J]. ACM SIGCOMM Computer Communication Review. ACM, 2008, 38(4): 75-86.

[57] GUO C, LU G, LI D, et al. BCube: a high performance, server-centric network architecture for modular data centers[J]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 63-74.

[58] SHIN J Y, WONG B, SIRER E G. Small-world datacenters[C]//The 2nd ACM Symposium on Cloud Computing. ACM, 2011, 2:1–2:13.

[59] SINGLA A, HONG C Y, POPA L, et al. Jellyfish: networking data centers randomly[C]//The 9th Symposium on Networked Systems Design and Implementation. 2012: 225-238.

[60] CUI Y, XIAO S, WANG X, et al. Diamond: nesting the data center network with wireless rings in 3-D space[J]. IEEE/ACM Transactions on Networking, 2018, 26(1): 145-160.

[61] WANG G, ANDERSEN D G, KAMINSKY M, et al. C-through: part-time optics in data centers[J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 327-338.

[62] FARRINGTON N, PORTER G, RADHAKRISHNAN S, et al. Helios: a hybrid electrical/optical switch architecture for modular data centers[J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 339-350.

[63] LEGTCHENKO S, CHEN N, CLETHEROE D, et al. XFabric: a reconfigurable in-rack network for rack-scale computers[C]//The 13th Symposium on Networked Systems Design and Implementation. USENIX, 2016: 15-29.

[64] HALPERIN D, KANDULA S, PADHYE J, et al. Augmenting data center networks with multi-gigabit wireless links[J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 38-49.

[65] ZHOU X, ZHANG Z, ZHU Y, et al. Mirror mirror on the ceiling: flexible wireless links for data centers[J]. ACM SIGCOMM Computer Communication Review, 2012, 42(4): 443-454.

[66] HAMEDAZIMI N, QAZI Z, GUPTA H, et al. Firefly: a reconfigurable wireless data center fabric using free-space optics[J]//ACM SIGCOMM Computer Communication Review, 2014, 44(4): 319-330.

[67] GHOBADI M, MAHAJAN R, PHANISHAYEE A, et al. Projector: agile reconfigurable data center interconnect[C]//The ACM SIGCOMM Conference. ACM, 2016: 216-229.

[68] WANG M, CUI Y, XIAO S, et al. Neural network meets DCN: traffic-driven topology adaptation with deep learning[J]. The ACM on Measurement and Analysis of Computing Systems, 2018, pp. 1–25.

[69] INTELLIGENCE G. Understanding 5G: perspectives on future technological advancements in mobile[R]. 2014: 1-26.

[70] CISCO. Direct tunnel for 3G networks [EB].(2016–10–27) [2019–05–23].

[71] IMT-2020 (5G) 推进组. 5G 网络架构设计白皮书 [R]. 2016. IMT-2020 (5G) PROPULSION GROUP. White paper of 5G network architecture design[R]. 2016.

[72] ANDREWS J G, BUZZI S, CHOI W, et al. What will 5G be?[J]. IEEE Journal on selected areas in communications, 2014, 32(6): 1065-1082.

[73] RAPPAPORT T S, SUN S, MAYZUS R, et al. Millimeter wave mobile communications for 5G cellular: it will work![J]. IEEE access, 2013, 1: 335-349.

[74] AGIWAL M, ROY A, SAXENA N. Next generation 5G wireless networks: a comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2016, 18(3): 1617-1655.

[75] PI Z, KHAN F. System design and network architecture for a millimeter-wave mobile broadband (MMB) system[C]// The 34th IEEE Sarnoff Symposium. IEEE, 2011: 1-6.

[76] SAITO Y, KISHIYAMA Y, BENJEBBOUR A, et al. Non-orthogonal multiple access (NOMA) for cellular future radio access[C]// IEEE 77th vehicular technology conference (VTC Spring). IEEE, 2013: 1-5.

[77] NIKOPOUR H, BALIGH H. Sparse code multiple access[C]//2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications. IEEE, 2013: 332-336.

[78] ZHANG Y, LIU H, JIAO L, et al. To offload or not to offload: an efficient code partition algorithm for mobile cloud computing[C]// IEEE 1st International Conference on Cloud Networking. IEEE, 2012: 80-86.

[79] CUERVO E, BALASUBRAMANIAN A, CHO D, et al. MAUI: making smartphones last longer with code offload[C]//The 8th international conference on Mobile systems, applications, and services. ACM, 2010: 49-62.

[80] CHUN B G, IHM S, MANIATIS P, et al. Clonecloud: elastic execution between mobile device and cloud[C]// The sixth Conference on Computer Systems. ACM, 2011: 301-314.

[81] RA M R, SHETH A, MUMMERT L, et al. Odessa: enabling interactive perception applications on mobile devices[C]// The 9th International Conference on Mobile Systems, Applications, and Services.

ACM, 2011: 43-56.

[82] KAO Y H, KRISHNAMACHARI B, RA M R, et al. Hermes: Latency optimal task assignment for resource-constrained mobile computing[J]. IEEE Transactions on Mobile Computing, 2017, 16(11): 3056-3069.

[83] ANANTHANARAYANAN G, BAHL P, BODÍK P, et al. Real-time video analytics: the killer app for edge computing[J]. Computer, 2017, 50(10): 58-67.

[84] ZHANG H, ANANTHANARAYANAN G, BODIK P, et al. Live video analytics at scale with approximation and delay-tolerance[C]//The 14th Symposium on Networked Systems Design and Implementation. USENIX, 2017: 377-392.

[85] YI S, HAO Z, ZHANG Q, et al. Lavea: latency-aware video analytics on edge computing platform[C]//The Second ACM/IEEE Symposium on Edge Computing. ACM, 2017(15):1–13.

[86] HONG C Y, KANDULA S, MAHAJAN R, et al. Achieving high utilization with software-driven WAN[C]//ACM SIGCOMM Computer Communication Review. ACM, 2013, 43(4): 15-26.

[87] JAIN S, KUMAR A, MANDAL S, et al. B4: experience with a globally-deployed software defined WAN[C]//ACM SIGCOMM Computer Communication Review. ACM, 2013, 43(4): 3-14.

[88] SINGH A, ONG J, AGARWAL A, et al. Jupiter rising: a decade of clos topologies and centralized control in Google's datacenter network[J]. ACM SIGCOMM computer communication review, 2015, 45(4): 183-197.

[89] YAP K K, MOTIWALA M, RAHE J, et al. Taking the edge off with espresso: Scale, reliability and programmability for global internet peering[C]//The Conference of the ACM Special Interest Group on Data Communication. ACM, 2017: 432-445.

[90] DALTON M, SCHULTZ D, ADRIAENS J, et al. Andromeda: performance, isolation, and velocity at scale in cloud network virtualization[C]// The 15th Symposium on Networked Systems Design and Implementation. USENIX, 2018: 373-387.

[91] KUMAR S, TUFAIL M, MAJEE S, et al. Service function chaining use cases in data centers[EB]. IETF, 2017.

[92] SUN C, BI J, ZHENG Z, et al. NFP: enabling network function parallelism in NFV[C]// The Conference of the ACM Special Interest Group on Data Communication. ACM, 2017: 43-56.

[93] WANG M, CUI Y, WANG X, et al. Machine learning for networking: Workflow, advances and opportunities[J]. IEEE Network, 2018, 32(2): 92-99.

[94] JIANG J, SEKAR V, STOICA I, et al. Unleashing the potential of data-driven networking[C]//International Conference on Communication Systems and Networks. Springer, 2017: 110-126.

[95] JIANG J, DAS R, ANANTHANARAYANAN G, et al. Via: improving internet telephony call quality using predictive relay selection[C]// The 2016 ACM SIGCOMM Conference. ACM, 2016: 286-299.

[96] WINSTEIN K, BALAKRISHNAN H. TCP ex machina: computer-generated congestion control[C]// The SIGCOMM Conference. ACM, 2013: 123-134.

[97] DONG M, LI Q, ZARCHY D, et al. PCC: re-architecting congestion control for consistent high performance[C]// The 12th Symposium on Networked Systems Design and Implementation. USENIX, 2015: 395-408.

[98] DONG M, MENG T, ZARCHY D, et al. PCC vivace: online-learning congestion control[C]// The 15th USENIX Symposium on Networked Systems Design and Implementation. USENIX Association, 2018: 343-356.

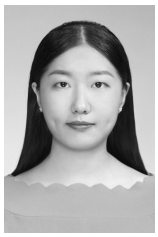
[99] CHEN L, LINGYS J, CHEN K, et al. AuTO: scaling deep reinforcement learning to enable datacenter-scale automatic traffic optimization [C]// The SIGCOMM Conference. ACM, 2018: 191-205.

[100] BELSHE M, PEON R, THOMSON M. Hypertext transfer protocol version 2 (Http/2): RFC 7540 [S].2015.

[101] HOLMBERG C, HAKANSSON S, ERIKSSON G. Web real-time communication use cases and requirements[J]. RFC 7478, 2015.

[102] INFOQ. 快手多媒体传输算法优化实践[EB]. (2019–01–09) [2019–05–23].
INFOQ. Optimization and practice of Kuaishou transmission protocol.[EB]. (2019–01–09)[2019–05–23].

[作者简介]



左旭彤（1995− ），女，河北沧州人，清华大学博士生，主要研究方向为低时延网络、流媒体传输优化。

王莫为（1995− ），男，河北石家庄人，清华大学博士生，主要研究方向为数据驱动网络、流媒体传输优化。

崔勇（1976− ），男，新疆乌鲁木齐人，博士，清华大学教授、博士生导师，主要研究方向为数据驱动网络、低时延网络及应用。