广域无损智算网络关键技术研究与实践

Research and Practice of Wide-Area Lossless Intelligent Computing Network Key Technologies

雷波/Lei Bo，潘永琛/Pan Yongchen，唐静/Tang Jing
（中国电信股份有限公司北京研究院，中国 北京 102209）
(China Telecom Corporation Limited Beijing Research Institute, Beijing 102209, China)

DOI：10.12142/ZTETJ.202603003
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260626.1350.012
网络出版日期：2026-06-26
收稿日期：2026-03-25

摘要：智算中心跨域互联已成为支撑规模化人工智能服务的关键网络新形态。系统分析了广域智算网络面临的传输需求与挑战，提出了构建广域无损智算网络的核心概念，以精准流量控制和队列缓存增强为关键技术路径，推动广域智算互联从尽力而为传输向无损承载演进。广域无损智算网络面向存算分离训练、云边协同推理等多类智算服务场景，实现了网络拥塞条件下的无损高效传输，为规模化智算服务构建了稳定可靠的网络底座。

关键词：智算网络；广域网；拥塞丢包；无损传输

Abstract: Cross-domain interconnection of intelligent computing centers has emerged as a critical network paradigm for large-scale artificial intelligence services. The transmission requirements and challenges of wide-area intelligent computing networks are analyzed, and the core concept of building a wide-area lossless intelligent computing network is proposed. Leveraging precise flow control and enhanced queue buffering as key technical enablers, this network drives the evolution of wide-area intelligent computing interconnection from best-effort to lossless transport. Targeting diverse service scenarios such as storage-compute decoupled training and cloud-edge collaborative inference, the wide-area lossless intelligent computing network achieves lossless and efficient transmission under network congestion, thereby establishing a stable and reliable foundation for large-scale intelligent computing services.

Keywords: intelligent computing network; wide area network; congestion-induced packet drop; lossless transmission

引用格式：雷波，潘永琛，唐静．广域无损智算网络关键技术研究与实践 [J]. 中兴通讯技术，2026, 32(3): 10-14. DOI: 10.12142/ZTETJ.202603003
Citation: Lei B, Pan Y C, Tang J. Research and practice of wide-area lossless intelligent computing network key technologies [J]. ZTE technology journal, 2026, 32(3): 10-14. DOI: 10.12142/ZTETJ.202603003

人工智能（AI）已成为引领新一轮科技革命与产业变革的核心驱动力。构建自主可控、高效普惠的智能算力互联基础设施，是中国把握新质生产力发展机遇、提升全球科技竞争优势的关键战略部署。2025年，《国务院关于深入实施“人工智能+”行动的意见》[1]明确提出，要强化智能算力统筹调度，推进智能算力互联互通与供需精准匹配，创新智能算力基础设施运营模式。

当前，算力主体主要集中于多个独立的大规模智算中心，部分算力分散于企业边缘自建节点，整体资源布局相对割裂。随着人工智能的持续快速发展，单一智算中心已难以满足大规模、弹性化的预训练、微调、推理等业务需求。面对千行百业对人工智能服务的普及化需求，我们亟需依托广域网实现智算中心之间、边缘算力与云端智算中心之间的高效网络互联互通，从而达成算力的灵活调度与资源的弹性供给，从而满足用户在隐私、成本、性能等方面的多样化需求。

本文首先重点剖析人工智能服务在广域智算网络中的互联场景及其面临的挑战，随后提出构建广域无损智算网络的核心观点，通过采用精细化流控与队列缓存增强等技术，保障高效无损传输，最后结合现网开展测试验证，以验证应用成效。

基金项目：国家重点研发计划项目（2024YFB2908700）；国家自然科学基金项目（U25B2038，62394325）

1 广域智算互联场景

随着以Transformer架构为核心的大语言模型技术加速迭代，在模型训练与推理部署等核心场景中，当前智算基础设施面临多重制约。一方面，单一智算中心的物理边界日趋刚性：在基础模型预训练上，大模型参数规模持续攀升，算力、电力、存储等资源供给已逼近上限。另一方面，边缘智算节点在成本、运维与规模化方面亦存短板：面向千行百业的模型推理与微调场景中，企业AI应用已从算力一体机单点部署阶段，迈入规模化落地应用阶段，但私有化部署模式正严重制约着微调、推理服务的规模化扩展。私有化的算力部署面临三大挑战：一是算力购置成本高，扩容成本难以承受；二是边缘侧小规模算力难以高效管控，企业缺乏边缘集群运维能力，扩容与故障恢复困难，易造成服务中断；三是自建算力场景有限、调度不充分、算力闲置严重，造成大量资源浪费。

在预训练场景下，多智算中心协同训练已成为突破超大模型训练瓶颈的关键路径。跨地域、跨集群整合多个大规模智算中心的算力资源，可有效打破单一智算中心在算力、存储等方面的物理限制。智算中心间采用数据并行（DP）、流水线并行（PP）等方式调度分散算力节点，构建统一的分布式训练集群，实现训练任务的高效拆解与协同执行。

在云边协同的推理、后训练与存算分离服务场景下，随着推理算力需求持续增长，企业用算模式正加速向云边协同架构转型，通过协同边缘智算节点与云端智算集群，为千行百业提供多样化AI服务。其中，存算分离的微调服务将训练语料数据集保留在边缘智算节点，通过高性能广域网络向智算中心同步传输批量数据，由智算中心负责梯度计算与参数更新；在隐私增强的推理服务中，按推理模型的流水线结构对模型进行切分，将推理计算实例分别部署于边缘节点和智算中心，通过网络传输高维度隐藏层变量，实现用户信息安全保护；在推理预填充–解码（PD）分离架构中，将预填充与解码阶段分别部署在不同云边节点，以充分利用异构与分散的算力资源，并借助分离架构提升端到端推理性能；对于海量数据入算场景，边缘节点存储的数据通过高性能网络传输至云端智算中心，完成集中处理、计算与数据备份。图1展示各类广域智算互联典型场景业务。

2 广域智算网络面临的挑战

2.1 广域智算网络加剧丢包效能损耗

远程直接内存访问（RDMA）作为智算服务的核心网络底座技术，凭借内核旁路、协议栈硬件卸载与零拷贝三大核心特性，能够实现高吞吐、低时延、低中央处理器（CPU）开销的高性能网络通信，已成为当前智算中心网络的关键支撑技术。其中，基于以太网架构的第二代RDMA融合以太网（RoCE v2）协议已成为数据中心内部RDMA规模化部署的主流方案[2–3]。然而，RoCE v2的稳定运行高度依赖无损网络环境。RDMA传输对丢包极为敏感，仅1‰的丢包率即可导致吞吐率减半[4]。在广域长距传输场景中，链路往返时延（RTT）的大幅增加会进一步放大丢包的负面影响，从而引发传输性能的断崖式劣化。

在长距传输场景中，当链路突发流量超出设备缓存承载能力时，会引发拥塞丢包。当前主流RDMA网卡的丢包重传算法正从回退N帧（GBN）向选择性重传（SR）演进，但两类算法在广域长距场景下均存在明显局限。对于GBN算法而言，由于广域链路端到端距离长、带宽时延积（BDP）较大，单次丢包所触发的全量重传会占用大量链路带宽，导致尾时延增加至毫秒级，极易造成有效吞吐骤降甚至业务中断。SR算法虽通过选择性重传机制减少了重传数据量，提升了重传效率，但由于广域链路的重传环路时延较长，多点丢包易引发端侧位图溢出，最终仍将退化为GBN模式。此外，广域链路固有的长时延、高抖动特性，极易触发端侧超时重传机制，产生大量无效重传，进一步挤占有效带宽，加

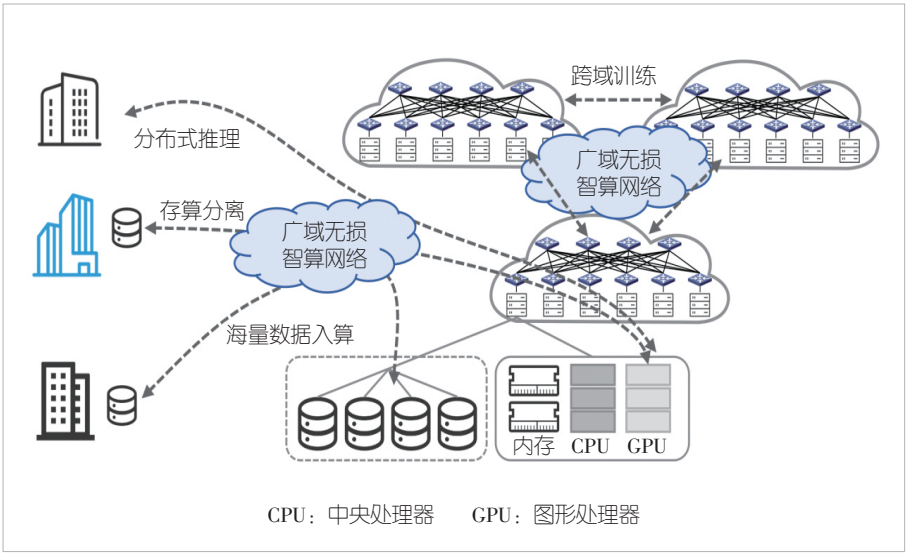


图1 广域智算互联场景

剧性能损耗。如图2所示，在500 km的云边协同推理服务中，RDMA传输性能随丢包率上升而下降，每Token输出时间（TPOT）显著增加，输出Token吞吐量相应下降；当丢包率升至0.1‰时，业务将完全中断。

为避免拥塞丢包，RoCE v2协议在智算中心内依托优先级流量控制（PFC）技术，构建端到端的无损传输网络。当前，学术界尝试基于PFC机制实现广域流控能力的优化[5-6]，但PFC本身存在的瓶颈制约了其在广域智算传输中的应用。PFC仅支持8个优先级队列，易引发拥塞扩散、流量风暴与死锁等问题，产生受害者流，影响网络的公平性与稳定性。若将PFC直接延伸至广域链路，单点拥塞将被扩散至跨地域的整个网络，从而导致大规模业务中断。因此，数据中心内的无损网络技术体系无法直接复用于广域智算网络，难以成为其全局解决方案。

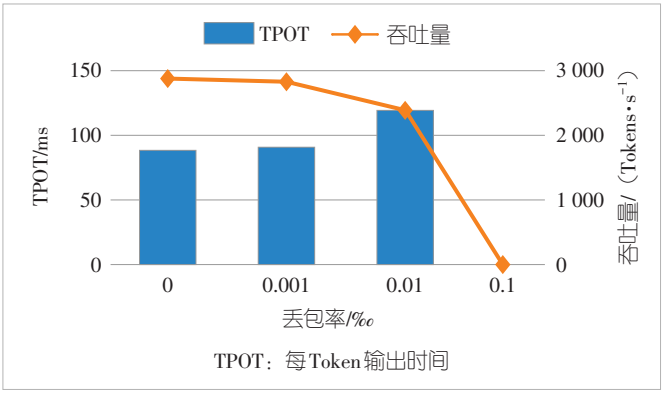
2.2 拥塞控制优化无法保障无损传输

RDMA在广域网中面临拥塞控制环路长、控制决策严重滞后的问题。智算中心内的RoCE v2通过PFC与端到端拥塞控制算法的协同实现无损传输保障，其中主流拥塞控制算法依赖RTT、显式拥塞通知（ECN）等端到端拥塞信号的反馈。然而在广域长距离场景下，链路RTT约为每百公里1毫秒（1 ms/100 km），拥塞信号反馈存在严重滞后，进而引发拥塞收敛缓慢、网络震荡剧烈等一系列问题。更关键的是，当前主流RDMA网卡普遍采用速率驱动的发送机制而非窗口控制机制，无法对网络中的在途流量进行精准管控。一旦发生拥塞，毫秒级的拥塞反馈时延会导致大量过量流量持续注入网络，进一步加剧链路拥塞，甚至引发丢包。

当前数据中心内大规模商用的数据中心量化拥塞通知（DCQCN）算法[4]本质上仍属于端到端的拥塞控制机制，已难以适配广域网络高延迟反馈环路的场景特性。PACC[7]、BiCC[8]、ATC[9]等现有优化方案虽通过提前向源端反馈拥塞信号实现快速速率控制，从而保障拥塞的快速感知，但在面对广域网络内的远端拥塞场景时，这些方案即便将拥塞点转移至远端数据中心出口的大缓存设备上，仍需要依赖端到端的拥塞信号来降低队列缓存，因此难以实现端到端的无损传输保障。

3 广域无损智算网络解决方案

针对广域智算网络中存在的上述挑战，本文提出了广域无损智算网络的观点，并在网络交换设备上采用关键技术，构建了相应的解决方案。如图3所示，关键技术包括：1）通过采用反压式的流控算法构建无损网络，并借助网络切片能力实现精细化流控，以避免受害者流的产生；2）通过部署多队列大缓存的网络设备来容纳突发流量，防止缓存溢出。



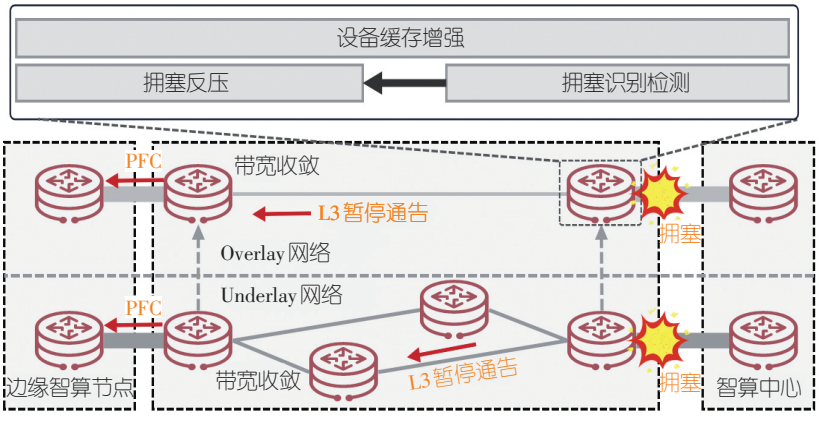
TPOT：每Token输出时间

图2 广域智算网络丢包性能损耗加剧

3.1 精准反压式流控

精准流控反压采用分段反压方式，在智算中心与广域网两个层级部署差异化的反压机制。

在智算中心管控域内，无损智算网络沿用PFC机制。智算中心与广域网的交汇点存在带宽收敛特性，使其成为端到端传输路径上最易发生拥塞的瓶颈点。为保证智算中心出口的流量无损传输，仅在智算中心及边缘的转发设备上使能PFC机制，利用其链路层微秒级快速反压能力，避免出口收敛瓶颈处发生缓存溢出丢包。同时，将PFC的作用边界严格限定在智算中心边缘，以防止智算中心内的拥塞扩散，阻断



PFC：优先级流量控制

图3 广域无损智算网络精准反压式流控机制

内部局部拥塞向广域网的传导。

在广域网管控域内，无损智算网络架构采用一种三层反压通告机制来完成反压式流控。在拥塞识别方面，与PFC检测入端口队列深度的机制不同，本文提出的精准流控基于随机早期检测在出端口执行拥塞判断。当因多对一通信模式或出入带宽不匹配导致队列深度达到预设暂停水线时，系统触发流量暂停操作。识别拥塞队列后，通过互联网控制报文协议（ICMP）报文向上游设备发送暂停通告，报文携带租户切片ID与暂停时长，按基于IPv6的段路由（SRv6）路径路由至上游节点，以暂停对应队列的发送。反压支持逐跳与透传两种模式：逐跳反压需设备升级支持；透传反压基于Overlay的SRv6隧道转发，可兼容存量设备，降低改造成本。传统PFC仅支持8个优先级，易引发租户干扰、拥塞扩散与受害者流。本文所采用的精准流控依托网络切片实现细粒度队列隔离，按租户分配独立切片与专属队列，反压仅在本切片内生效，不影响其他RDMA流量，从而在切片间实现强租户性能隔离，彻底避免受害者流。

3.2 无损网络设备缓存增强

广域网链路时延由智算中心内的微秒级增大至毫秒级，导致带宽时延积（BDP）数量级显著增大。为适配这一特性，广域网设备通常配置大容量缓存资源，用于缓解链路传输波动并容纳突发数据包。尽管精准反压式流控限制了流量突发、维持了队列深度的稳定，但仍需设备缓存能力的配合，以构建无损的网络链路。如图4所示，在ICMP反压通告报文发送至上一跳的时间段内，链路中最大可能存在大量在途数据包。因此，设备需保证至少具备与理论最大在途报文相匹配的缓存空间，以吸收反压过程中的数据包，避免在反压通告间隔期间发生拥塞丢包。在网络部署时，需根据链路BDP选择合适的设备缓存配置。此外，转发设备在大容量缓存的基础上，借助智能化缓存管理、动态水线配置等管理调度算法，可进一步吸收流量突发、平滑队列波动，从而与精准流控机制协同保障全链路的无损传输。

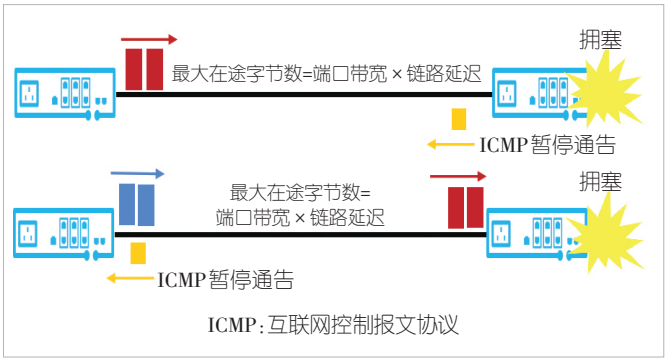


图4　无损网络缓存增强原理

4 现网实践验证

中国电信对智算广域网核心技术开展落地现网实践，完成无损广域智算网络全场景验证工作，在广域云边协同推理服务和存算分离微调训练服务场景中均实现了成本与效率的精准平衡。现网实践中，方案成功将带宽收敛比提升至320:1，大幅降低按算力卡能力建网带来的带宽成本，通过租户级精准流控与设备缓存增强构建了无损广域智算网络，实现多租户性能隔离与无损转发，云边协同训推算效不低于95%。

为验证广域无损智算网络带来的性能收益，中国电信基于500 km现网环境搭建了无损广域智算网络试验床，开展云边协同推理服务和存算分离微调服务的测试，并分别评估了无损网络与容损网络在不同带宽瓶颈下的性能表现。如图5所示，在DeepSeek-lite-v2-16B模型的存算分离微调服务测试中，广域无损智算网络借助精细化流控与设备能力增强，可将互联带宽压缩至1 Gbit/s。而在无无损保障的容损网络下，当带宽瓶颈为1 Gbit/s时，发生拥塞丢包，算效损失达20%。如图6所示，在Qwen3-32B模型的云边协同推理服务测试中，无损环境下可将广域切片带宽压缩至5 Gbit/s，推理性能仍可保持在95%以上；而在容损广域网络中，丢包导致TPOT增加3.4倍，输出Token吞吐率劣化超过60%。上述结果表明，对于当前跨广域网的人工智能服务而言，构建无损的智算网络至关重要。

面向低成本轻量化运营的诉求，边缘网络侧部署轻量化智算客户终端设备（CPE）形态设备，通过中国电信息壤平台进行能力集成，将智算网关与CPE统一纳入集约化运维管理体系。该架构可灵活适配不同层级带宽的汇聚，配套租户粒度精细化流量管控策略，同时联动企业侧实现协同反压联动调度。依托SRv6与网络切片技术，为业务提供可量化可

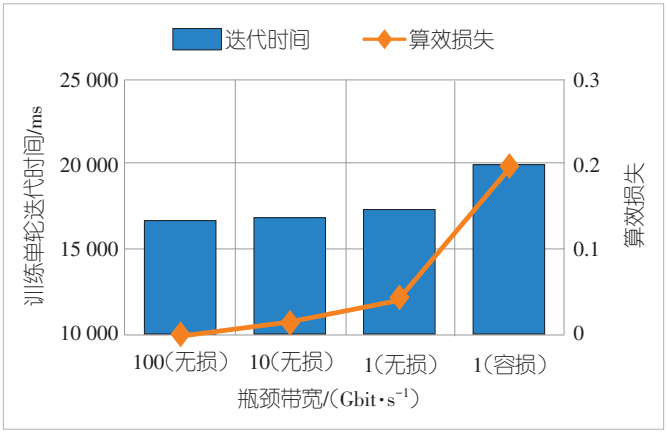


图5　存算分离微调服务性能评估

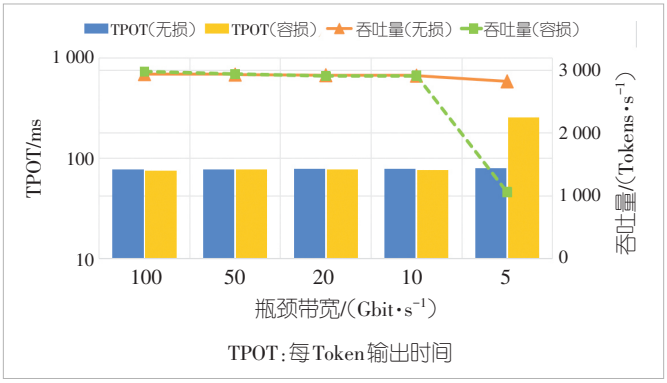


图6　云边协同推理服务性能评估

保障的高品质传输服务，全面支撑企业训练推理算力资源与RoCE无损网络的一体化部署，实现业务极简开通上线与全流程轻量化运维，有效压降企业整体运营成本与管理负担。

面向低成本、轻量化运营的诉求，边缘网络侧部署轻量化智算CPE形态设备，并通过中国电信息壤平台进行能力集成，将智算网关与CPE统一纳入集约化运维管理体系。该架构可灵活适配不同层级带宽的汇聚需求，配套租户粒度的精细化流量管控策略，同时联动企业侧实现协同反压与联动调度。依托SRv6与网络切片技术，该架构可为业务提供可量化、可保障的高品质传输服务，全面支撑企业训练推理算力资源与RoCE无损网络的一体化部署，实现业务的极简开通上线与全流程轻量化运维，从而有效压降企业整体运营成本与管理负担。

5 结束语

在算力泛在化与智算广域互联的发展趋势下，构建高可靠、高效率、可隔离的广域无损智算网络，已成为支撑AI时代智算服务的关键底座。面向跨域智算服务规模化部署的产业需求，亟需持续强化算力、网络与存储的一体化协同设计，加快高性能传输核心技术的工程化落地，推动广域智算网络从试验验证迈向规模商用。需要指出的是，广域无损智算网络的技术成熟与部署推广是一个循序渐进的过程，在方案设计、现网试验与运营优化中需不断总结经验、迭代完善。当前，广域智算网络仍处于技术探索与试点验证的关键阶段，亟需围绕用户需求、建设成本与网络性能开展深入研究，加速关键技术的成熟与方案的落地，从而为构建自主可控、高效普惠的智能算力基础设施提供坚实支撑。

参考文献

[1] 中华人民共和国国务院. 国务院关于深入实施“人工智能+”行动的意见: 国发〔2025〕11号 [EB/OL]. (2025-08-21)[2026-04-20]. https://www.gov.cn/zhengce/content/202508/content_7037861.htm
[2] 李宝嘉, 何春志, 夏寅贵, 等. 星脉网络: 面向GPU集群集合通信与集中式路由的协同优化 [J]. 中兴通讯技术, 2025, 31(2): 3-13. DOI: 10.12142/ZTETJ.202502002
[3] 段威, 于浩, 李和松, 等. 智算中心组网技术及应用 [J]. 中兴通讯技术, 2025, 31(2): 63-71. DOI: 10.12142/ZTETJ.202502009
[4] Zhu Y B, Eran H, Firestone D, et al. Congestion control for large-scale RDMA deployments [J]. ACM SIGCOMM computer communication review, 2015, 45(4): 523-536. DOI: 10.1145/2829988.2787484
[5] Yu P W, Xue F Y, Tian C, et al. Bifrost: extending RoCE for long distance inter-DC links [C]//Proceedings of IEEE 31st International Conference on Network Protocols (ICNP). IEEE, 2023: 1-12. DOI: 10.1109/ICNP59255.2023.10355634
[6] Chen Y Q, Tian C, Dong J Q, et al. Swing: providing long-range lossless RDMA via PFC-relay [J]. IEEE transactions on parallel and distributed systems, 2023, 34(1): 63-75. DOI: 10.1109/TPDS.2022.3215517
[7] Zhong X L, Zhang J, Zhang Y L, et al. PACC: proactive and accurate congestion feedback for RDMA congestion control [C]// Proceedings of IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. IEEE, 2022: 2228-2237. DOI: 10.1109/INFOCOM48880.2022.9796803
[8] Wan Z R, Zhang J, Yu M X, et al. BiCC: bilateral congestion control in cross-datacenter RDMA networks [C]//Proceedings of IEEE INFOCOM 2024 - IEEE Conference on Computer Communications. IEEE, 2024: 1381-1390. DOI: 10.1109/INFOCOM52122.2024.10621412
[9] Wan Z R, Zhang J, Su Y Z, et al. Re-architecting traffic control in cross-datacenter RDMA networks [J]. IEEE transactions on networking, 2025, 33(5): 2632-2647. DOI: 10.1109/TON.2025.3569630

作　者　简　介

雷波，中国电信股份有限公司北京研究院互联网技术研究所副所长（主持工作）；主要从事算力网络、智算网络、智能体网络等新型网络技术研究；发表论文50余篇，参与制定标准10余项，发布著作7部，拥有发明专利33项。

潘永琛，中国电信股份有限公司北京研究院互联网技术研究所研究员；主要从事智算网络传输控制协议与高性能网络优化方面的研究工作；发表论文10余篇。

唐静，中国电信股份有限公司北京研究院互联网技术研究所团队总监；主要从事算力网络、智算网络、数据中心网络等未来网络技术研究，深度参与ITU-T、CCSA等标准化工作；发表论文10余篇。