广域抗损高吞吐URDMA技术

URDMA Technologies for Wide-Area High-Throughput Network

段晓东/DUAN Xiaodong，陆璐/LU Lu，孙滔/SUN Tao，
李志强/LI Zhiqiang，杨红伟/YANG Hongwei，
杜宗鹏/DU Zongpeng

（中国移动通信有限公司研究院，中国 北京 100053）
(China Mobile Research Institute, Beijing 100053, China)

DOI：10.12142/ZTETJ.202406005
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250121.1330.002.html
网络出版日期：2025-01-21
收稿日期：2024-10-12

摘要：随着国家“东数西算”战略实施以及智算、超算业务的快速发展，海量数据广域传输需求不断增多。提出一种广域抗损高吞吐超远程直接内存访问（URDMA）技术方案，通过对传输控制协议/互联网协议（TCP/IP）协议栈的完全卸载，消除中央处理器（CPU）对网络高吞吐性能的限制。采用拥塞控制、丢包恢复、丢包重传等技术增强标准第2代基于融合以太网的远程直接内存访问（RoCEv2）协议，使其在广域有损网络下保持高吞吐性能。测试结果表明，在往返时延（RTT）时延为20 ms、丢包率0.1%的网络环境下，TCP协议吞吐性能仅为0.02 Gbit/s，标准RoCEv2性能接近为0，URDMA协议吞吐性能为88.26 Gbit/s；当RTT时延增加到80 ms时，TCP和RoCEv2协议吞吐基本衰减为0，URDMA协议吞吐性能为83.12 Gbit/s，仍然保持较高的性能。

关键词：广域抗损高吞吐；数据快递；远程直接内存访问；RoCEv2

Abstract: With the implementation of the national "East Data West Computing" strategy and the rapid development of intelligent computing and supercomputing services, the demand for large-scale data transmission is constantly increasing. A wide-area high-throughput ultra remote direct memory access (URDMA) technology solution is proposed, which mitigates the limitation of the central processing unit (CPU) on high-throughput network performance by completely offloading the transmission control protocol/Internet protocol (TCP/IP) protocol stack. By adopting congestion control, packet loss recovery, packet loss retransmission, and other technologies to enhance the 2nd version of remote direct memory access based on converged ethernet (RoCEv2) protocol, URDMA enables high-throughput performance in wide-area lossy networks. The test results show that in a network environment with a round-trip time (RTT) of 20 ms and a packet loss rate of 0.1%, the TCP protocol throughput performance is only 0.02 Gbit/s, the standard RoCEv2 performance is close to 0, and the URDMA protocol throughput performance is 88.26 Gbit/s. When the RTT increases to 80 ms, the TCP and RoCEv2 protocols basically decay to 0, and the URDMA protocol throughput performance is 83.12 Gbit/s, still maintaining high performance.

Keywords: high-throughput in wide-area network; data express; remote direct memory access; RoCEv2

引用格式：段晓东, 陆璐, 孙滔, 等. 广域抗损高吞吐URDMA技术 [J]. 中兴通讯技术, 2024, 30(6): 23-30. DOI: 10.12142/ZTETJ.202406005
Citation: DUAN X D, LU L, SUN T, et al. URDMA technologies for wide-area high-throughput network [J]. ZTE technology journal, 2024, 30(6): 23-30. DOI: 10.12142/ZTETJ.202406005

2022年国家发展和改革委员会、中央网信办、工业和信息化部、国家能源局联合启动了“东数西算”战略。随着东数西算战略的实施，东数西存、东数西训、东数西渲等场景对海量数据跨广域网数据快递需求日益凸显。随着产业数字化、云计算、分布式人工智能（AI）的发展，数据异地上云、云迁移、云灾备、跨智算中心互联等时空大尺度数据搬迁场景中数据规模越来越大，对网络吞吐的要求越来越高。

网络带宽也从10G发展到25G、100G、200G、400G、800G甚至1.6T。与网络带宽快速增长形成鲜明对比的是，后摩尔时代中央处理器（CPU）算力增速远低于网络带宽增速，并且差距还在持续增大。如何充分利用网络带宽破解海量数据广域传输瓶颈，如何以低算力损耗满足高速网络处理和传输要求，对立体泛在算力网络整体算效提升及分布式AI训练、推理性能提升至关重要。

本文中，我们将重点分析数据快递业务对广域网络的高吞吐需求与挑战，并给出数据快递广域抗损高吞吐超远程直接内存访问（URDMA）解决方案及其初步测试结果。

1 数据快递广域高效传输需求与挑战

1.1 数据快递广域高效传输需求

东数西算、数据异地上云、云间灾备、广域智算互联等场景大多涉及海量数据跨省传输。自动驾驶数据上云需要传

输大量数据至智算中心进行训练，每辆车每天生成的数据量可达几TB至十几TB，完成L3级别的训练可能会产生8 EB的数据；天文数据计算，FAST每年约200多个[1]观测项目，单项目产生观测数据量TB至PB量级，年产数据约15 PB。为了缩短数据传输的时间，需要借助高带宽网络。但高带宽不等于高（有效）吞吐，端到端高吞吐才是确保数据时效性和减少传输成本的关键。在面向连接的可靠传输技术中，距离越长，确认报文回复耗时越长，对业务发送端和接收端服务器的缓存要求越高，实现精确丢包检测的难度越大，实现长距离、高吞吐的可靠传输挑战越大。

摩尔定律放缓使得通用CPU性能增长的边际成本迅速上升。数据表明，现在CPU的性能年化增长（面积归一化之后）仅有3%左右[2]，这导致带宽性能增速比（RBP）失调。网络的带宽年化增长在2010年前大约是30%，2015年微增到35%，近年达到45%。相应地，CPU的性能增长从10年前的23%下降到12%，并在近年直接降低到3%。在这3个时间段内，RBP指标从1左右上升到3，并在近年超过了10。网络带宽的剧增对业务发送端和接收端服务器的数据收发处理能力提出了更高要求，基于CPU的传输控制协议/互联网协议（TCP/IP）等传统处理方式逐渐成为端到端高吞吐数据传输的瓶颈。

1.2 数据快递广域高效传输挑战

长距高吞吐是数据快递业务的重要目标。2020年中国移动完成了全球最大的云专网商用部署，基于软件定义网络（SDN）和段路由（SR）技术的云专网实现跨数据中心云资源池的整合，骨干段覆盖全国所有直辖市、省、自治区；在云骨干网中，网络物理带宽已不是瓶颈，如何提升端到端高有效吞成为关键。目前主流业务多采用TCP/IP协议进行海量数据广域传输。由于现有传输协议、拥塞控制算法、丢包冗余恢复机制、选择性重传机制及算力损耗等方面的限制，现有协议和机制无法满足“长肥”网络下的高吞吐数据传输需求。

1）协议。互联网工程任务组（IETF）RFC1323[3]标准规定，TCP理论窗口最大值为1 GB（2^{30} bytes）。依据包守恒定律（理想情况下的窗口大小和Inflight数据相同），当吞吐量为400 Gbit/s时，单流最远传输约为1 000 km；当吞吐量为800 Gbit/s时，最远传输仅500 km，无法满足广域跨智算中心分布式AI训练等少数大象流高吞吐长距离传输需求。

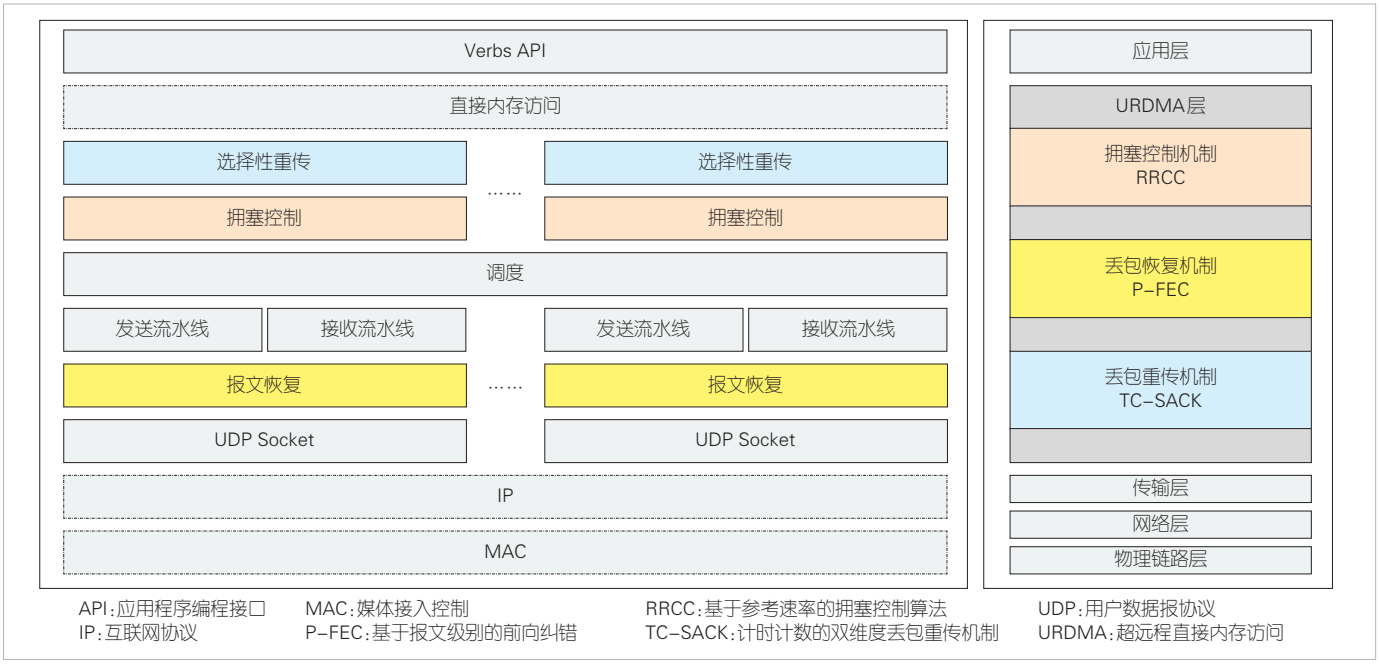
2）拥塞控制算法。拥塞控制算法按照拥塞判断依据大致分为丢包类、时延类和带宽类。丢包类算法如Reno[4]、CUBIC[5]等，依据网络是否丢包来判断拥塞，但因易误判，发送速率会出现过度调整，从而限制了吞吐；时延类算法如FAST[6]、Vegas[7]等，依据网络环回时延的变化来判断拥塞，但可能因为时延突变或回路时延变大造成拥塞误判，导致发送速率过度调整；带宽类算法如瓶颈带宽和往返传输时延（BBR）[8]等，通过检测可用带宽来调整发送速率，但公平性较差，并可能因为发送速率调整不及时而导致大量丢包。

3）丢包冗余恢复机制。丢包冗余恢复机制常采用纠错码。纠错码是一种编码技术，通过在数据传输过程中添加冗余信息来保护数据的完整性。通过选择合适的算法和参数（例如，块大小、纠错码长度、冗余度等），该机制能够抵御多个数据包的丢失或损坏。现有基于纠错码的丢包恢复方法存在不足，例如，接收端在发现丢包后才通知发送端进行纠删码编码和校验块发送，接收端必须等待其所请求的校验块被完全接收后，才能实现丢失恢复。这会产生极大的额外传输时延。对丢包的判断、对数据的编码和对数据的解码均需CPU持续参与控制，会占用大量的CPU时间，从而产生额外的CPU负担。

4）选择性重传机制。TCP有超时重传（RTO）和快速重传两种机制。RTO根据往返时延（RTT）来估算，广域高RTT环境等待时间长、效率低；快速重传在收到的重复的确认字符（ACK）达到3个之后就进行重传，效率高，但需要充足的后续报文。TCP/IP协议目前都依赖发送端的流量控制，发送端通过ACK推测网络的情况，发送速率严重依赖接收端的ACK反馈，不利于广域长距离网络的高吞吐传输。

5）算力低损耗。传统的TCP/IP协议栈运行在操作系统内核空间，主机间通信需要用户空间、内核空间、硬件网卡之间多次交互[9]。数据拷贝、用户态和内核态的切换、数据包收发中断响应等都需要CPU参与，大量的CPU资源被消耗。处理10G网络数据包需要大约4个Xeon CPU核，即仅是网络数据包收发处理就占用了一个通用8核CPU一半的算力[10]。当网络带宽增大到100G、200G甚至400G时，CPU性能将会成为高吞吐传输的瓶颈。因此出现了很多卸载方案，这些方案将协议栈的全部或部分数据包处理工作卸载到硬件网卡上。这样可以充分发挥硬件网卡对数据包的高速处理能力，降低CPU损耗，提升数据收发端（例如业务服务器）的数据包处理能力。其中有代表性的包括TCP/IP协议卸载引擎（TOE）[11]、数据面开发套件（DPDK）[12]，但卸载效果都不是很理想。

RDMA使用内存零拷贝、内核旁路、CPU卸载等技术，将协议栈全卸载到网卡处理，允许用户态的应用程序直接读取和写入远程主机内存，避免了数据拷贝和上下文切换，实现了高吞吐量、低时延和低CPU算力损耗。RDMA有3种技术路径：InfiniBand[13]、基于融合以太网的RDMA



▲图1 URDMA技术架构

（RoCE）[14]和互联网广域RDMA协议（iWARP）[15]。其中，RoCEv2[16]因其兼容传统TCP/IP协议、易于部署管理等优点，在数据中心网络广泛应用。RDMA协议对丢包容忍度较低，要求在无损网络环境中运行，1%的丢包会使其吞吐下降至0。广域网难以实现真正的无损。

2 URDMA架构与关键技术

针对TCP/IP及标准RDMA在广域高效传输场景面临的问题，我们提出了数据快递广域抗损高吞吐URDMA解决方案：包含广域抗损高吞吐协议如URDMA、反向快启动速率控制机制如基于参考速率的拥塞控制算法（RRCC）、数据块丢包恢复机制如基于报文级别的前向纠错（P-FEC）及收发解耦多维重传机制如计时计数的双维度丢包重传机制（TC-SACK）。下文中，我们首先给出URDMA的整体架构，并针对URDMA协议及拥塞控制、丢包恢复、丢包重传3个方面的关键创新展开介绍。

2.1 URDMA技术架构

URDMA架构的设计充分考虑了兼容与平滑演进。一方面，该架构中的Verbs应用程序编程接口（API）与标准RDMA保持一致，便于存量应用平滑迁移；另一方面，该架构与现有TCP/IP协议簇兼容，仅对标准RDMA传输层协议进行增强，避免对广域网中网络设备进行升级改造，降低方案部署门槛。该方案涉及以下关键技术：

1）1套URDMA协议。扩展RoCEv2报文，支持RTT内生测量，为RRCC提供精准网络状态；支持精准内存地址投递机制，逐包携带含内存地址信息的扩展传输头（RETH），为TC-SACK直接“落存”提供访问内存的虚拟地址信息。

2）3个创新机制。反向快启动速率控制机制如RRCC，通过快速拥塞发现以及发送速率调节机制，确保高吞吐传输；数据块切分的丢包恢复机制如P-FEC利用前向纠错机制，实现数据包冗余度、带宽利用率与丢包恢复精度之间的综合最优；收发解耦多维重传机制如TC-SACK，通过优化发送和接收端滑动窗口大小和重传阈值，精确判断丢包，实现数据包选择重传。

2.2 URDMA关键技术

本章节中，我们对URDMA协议及创新机制展开介绍。

2.2.1 URDMA

URDMA协议的设计目标是基于标准RoCEv2协议，在高带宽时延积（BDP）广域网环境下，实现高吞吐性能的同时减少CPU算力消耗。其核心设计原则如下：

1）全卸载协议处理，吞吐性能和CPU利用率无关；

2）极简协议设计，高载荷比报文格式，状态少，易于硬件实现。

URDMA对标准RoCEv2的基础传输头（BTH）、确认扩展传输头（AETH）进行扩展，对RETH及BTH头的A字段的使用方式进行重新约束。

RTT的精度决定RRCC的反应灵敏度，因此URDMA协

议增强了对RTT精确测量的支持力度。针对BTH头的扩展，两个预留字段分别用来指示时间戳信息或时间戳信息在payload中的偏移起始位置。如果预留字段用来指示偏移起始位置，T_1和T_r的最高bit置为1，否则置为0。在payload中的时间戳长度为32 bit。其中T_1为报文离开发送端网卡时间，T_r为接收端到发送端的单向时延。针对AETH头的扩展，URDMA协议增加T_3、T_S、包序号（PSN）3个扩展字段，其中T_3为报文离开接收端网卡时间，T_S为发送端到接收端的单向时延，BTH头的A字段用来触发接收端反馈携带AETH头的ACK报文，发送端可根据RRCC等机制按需对A字段进行置位，触发对RTT的精确测量。

TC-SACK等机制需要报文在接收端直接“落存”。按照标准RDMA Write Request等操作机制，每条队列对（QP）连接的首包携带RETH内存地址信息，报文一旦发生首包丢失或乱序，后续报文将无法正确找到接收端存放此报文的虚拟内存地址，进而无法写入正确的内存物理位置。URDMA协议中采用逐包携带RETH头的方式，确保每个报文都能直接写入接收端内存正确物理位置。

2.2.2 RRCC

URDMA利用多参数协同判断拥塞状态，不断探测链路瓶颈带宽和时延，通过瓶颈链路带宽时延积精确计算和调整发送窗口大小，使在飞数据包维持在合理的范围内，从而在确保最大传输速率的同时减少网络传输的排队延迟，保证数据广域传输具有高吞吐和低丢包的性能。瓶颈带宽指的是端到端传输的网络路径上速率最慢的那段链路的带宽，该带宽决定了端到端传输的带宽上限。测量时延的目的是得到网络路径的最小RTT，即光/电信号从发端到收端的最小时延，具体大小取决于物理距离。

如图2所示，用于广域拥塞控制的RRCC具体实现如下：

1）QP之间建立可靠连接（RC），测量该链路的RTT作为初始最小RTT，然后进入启动阶段，此时发送速率呈指数增加。

2）当发送速率达到或超过瓶颈带宽时，降低发送速率，并排空缓存队列。

3）进入瓶颈带宽探测周期（每5～10个RTT为1个周期），发送速率稳定在瓶颈带宽，并周期性探测当前链路的瓶颈带宽。如果该链路的瓶颈带宽出现变化，则下一个周期的发送速率随之变化。

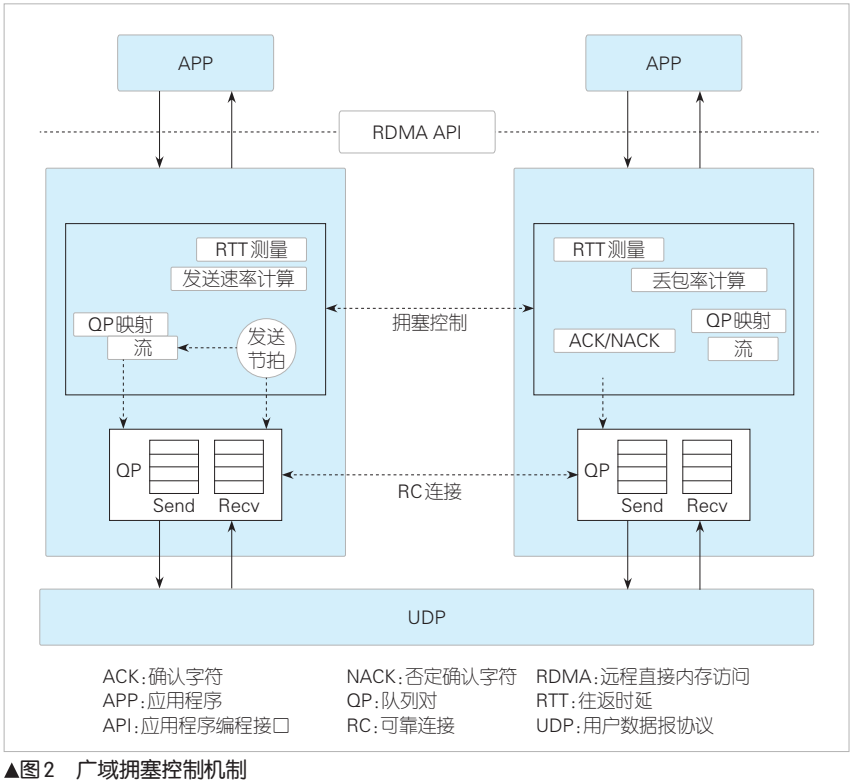
4）当接收端检测到丢包时，接收端启动丢包率统计，并将结果反馈给发送端，发送端判断丢包率大于丢包率阈值，则按比例降低发送速率。

5）设定一个固定周期，将发送速率降低至适当大小，用于测量当前链路的最小RTT。

为了进一步提升传输效率，在受控网络如中国移动云专网中，URMDA的拥塞控制机制还可以支持按照规划的参考速率进行传输，以降低端侧对广域网复杂网络情况的误判率。

在RRCC机制中，广域网的网络带宽资源将在数据中心（DC）出口路由器进行规划和分配，在数据快递的流量启动传输之前，可以从管控模块得到一个参考速率进行传输。这个参考速率可以是基于历史统计的可用带宽，或基于专线规划的带宽。一方面，端侧可以省略慢启动的带宽探测操作，直接按照参考速率开始传输；另一方面，在拥塞控制算法判定需要降速时，可以将测量到的新的瓶颈带宽与参考速率比较后取最大值作为瓶颈带宽，即锁定实际发送速率的下限。

在数据快递的场景中，相关的流通常是大象流。大象流持续时间长，带宽需求较大，但是流数较少，且部分流量可以容忍一定的发送延时如隔日达等。受控网络可根据晚上的网络负载历史数据，在网络负载较低时启动RRCC传输机制，以较快速率发送数据快递流量，达到削峰填谷的效果。



▲图2 广域拥塞控制机制

2.2.3 P–FEC

在丢包恢复机制中，接收端在发生少量丢包时，通过发送端发送的冗余数据实现快速包恢复，从而可以减少丢包重传，降低重传时延。URDMA使用基于前向纠错码（FEC）[17-18]的丢包恢复技术。

如图3所示，发送端根据RDMA协议栈发出的原始数据包进行前向纠错编码，产生冗余修复数据包并随原始数据包一同发送至接收端。

接收端根据RDMA原始数据包头部的PSN字段判断原始数据包是否发送丢失，并通过冗余修复数据包即时地恢复，最终传输至RDMA协议栈模块。

在RDMA协议栈的数据传输过程中，同一数据流内的所有RDMA原始数据包的PSN，在没有发生丢失的情况下，保持连续递增。为了有效地进行冗余修复数据包的生成，发送端采用按分组的冗余计算方式，将数据包根据其PSN分组。分组数据包的数量可根据网络丢包率动态调整，在低丢包率网络中，减少每组数据包数量，降低发送端和接收端编解码资源占用和恢复时长；在高丢包率网络中，增加每组的数据数量，提升丢包恢复成功率。这一机制有助于接收端对原始数据包和冗余修复数据包进行有序组织和识别，以维护数据传输的完整性和可靠性。

每个分组包含连续的m个RDMA原始数据包，而每k个原始数据包生成一个相应的冗余数据包。在传输过程中，这m个原始数据包按照它们的PSN序号经过RDMA协议栈和网卡递增地发送。在前向纠错编码阶段，发送端采用数据包级别的异或运算，对分组内的每个数据包进行操作，以创建冗余修复数据包。

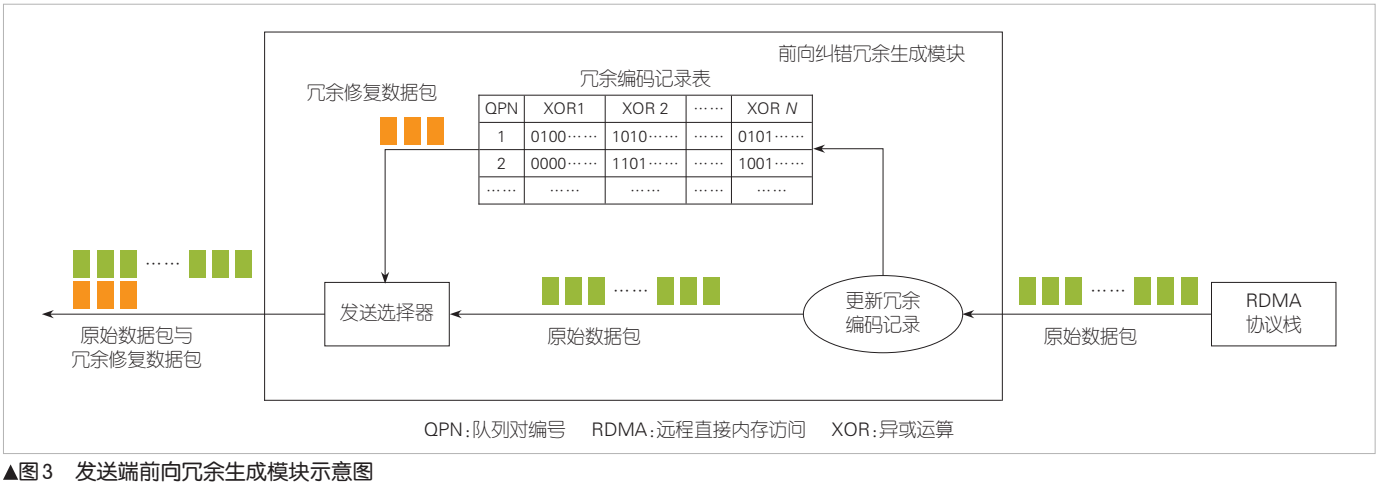
由于异或运算的属性，异或运算可以逐步执行，而无须在运算过程中记录参与异或运算的RDMA原始数据包。具体来说，每个RDMA原始数据包都可以根据其PSN确定其在分组内的位置，进而确定与之对应的编码冗余数据包，即第i个RDMA原始数据包DATA（i）对应的冗余修复数据包必然为XOR（$i \bmod r$）。因此，要计算每个冗余修复数据包的最终结果，只需要在发送第i个RDMA原始数据包时，让其与XOR（$i \bmod r$）执行异或运算即可。当该分组最后一个RDMA原始数据包发送完成时，所得到的异或计算值即所要求的冗余修复数据包。

接收端的解码原理与发送端的编码过程保持一致，都充分利用异或计算的可分步性质，将异或冗余的修复计算分散到每次单独的数据包接收过程中，如图4所示。

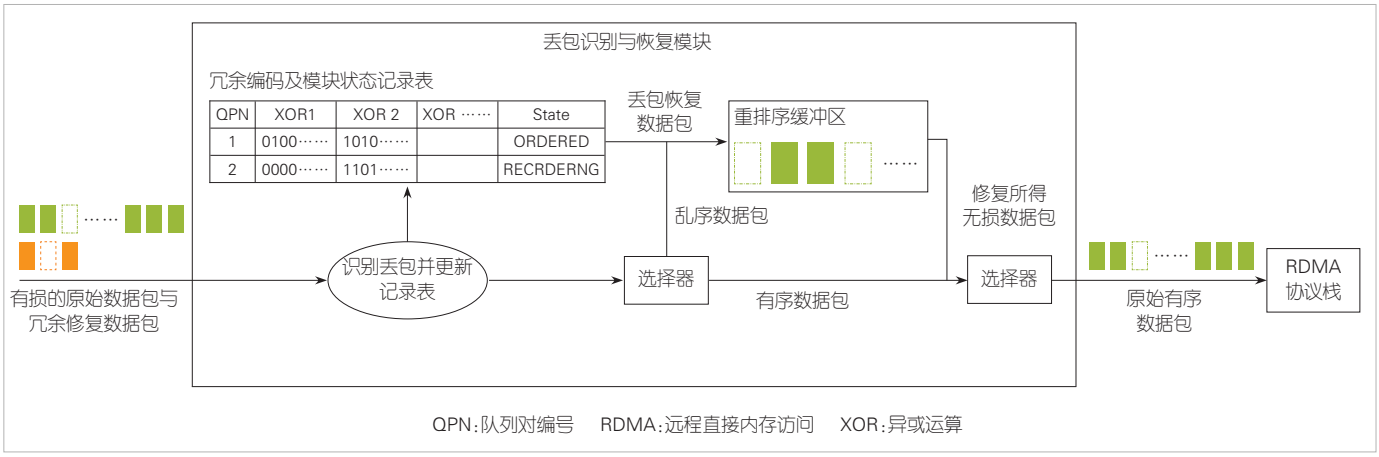
2.2.4 TC–SACK

广域拥塞控制机制能降低丢包，提升网络吞吐，为RDMA协议的运行提供良好的通路。然而，广域网无法在任意情况下都实现0丢包。丢包恢复技术仅能恢复突发性少量丢包，如果丢包数超出恢复范围，就需要启动重传机制。RDMA协议默认采用Go–Back–N丢包重传机制，如图5所示。当接收端检测到丢包时，接收端通知发送端从该丢失包之后的所有包都需要重传，即使有些包已经送达接收端。这种机制好处是接收端不需要缓存数据包，节省接收端的存储空间，且减少乱序重排的时间。但在广域有损网络条件下，Go–Back–N机制对吞吐限制巨大。

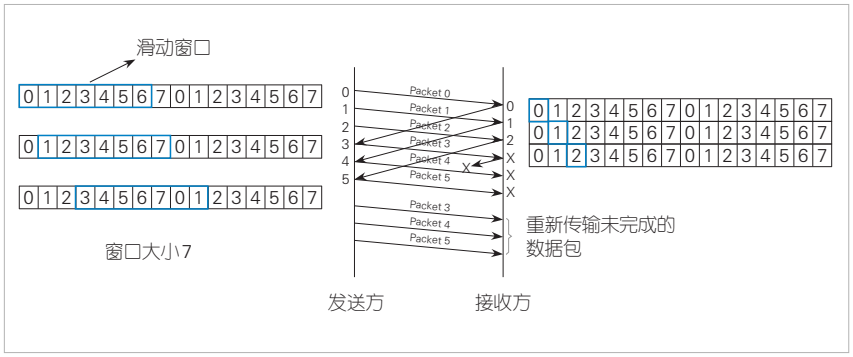
针对数据快递的高吞吐需求，URDMA提出TC–SACK机制，其主要特点为：a)无须维护接收窗口，接收端收到报文后将其直接内存访问（DMA）到内存；b)发送速率与丢包重传解耦，接收端依据预设时间或数据包数量，触发丢包重传通知。丢包信息携带在TC–SACK的ACK报文中，需要扩展扩展传输头（ETH）以支持携带丢包信息。TC–SACK的启用需要



▲图3　发送端前向冗余生成模块示意图



▲图4 接收端丢包识别与恢复模块示意图



▲图5 Go-Back-N丢包重传

在连接建立时进行协商。

在TC-SACK中，达到预设时间或数据包数量时，接收方触发确认机制。在图6中，7个数据包回复1个ACK，该ACK携带未收到的数据包的PSN，其他的乱序或顺序收到的数据包直接DMA到内存。具体流程如下：

1）发送端按照规划的速率发送数据包，接收端收到PSN=1的数据包，直接DMA到内存。这时接收端启动计时器开始计时，同时记录数据包数量为1、丢包数量为0。

2）接收端收到PSN=3的数据包，直接DMA到内存，相关的数据包需要携带目的内存地址。接收端发现PSN=2的数据包丢失，并不立刻反馈ACK。接收端记录数据包数量为3，丢包数量为1。

3）接收端收到PSN=6和PSN=7的数据包，操作类似，接收端记录数据包数量为7，丢包数量为3。到达预设的数据包数量，触发ACK，其中携带PSN2/4/5。同时接收端重置计数器和数据包计数。

4）发送端重发PSN=2/4/5数据包。

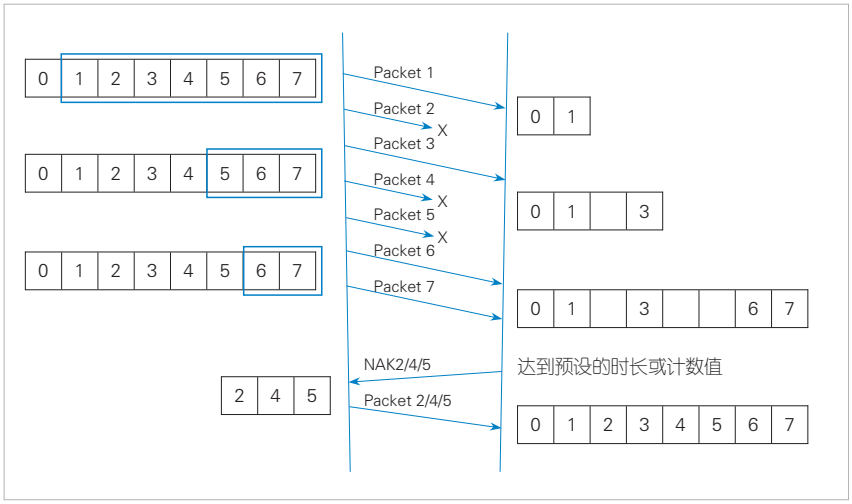
3 性能评估

3.1 测试配置

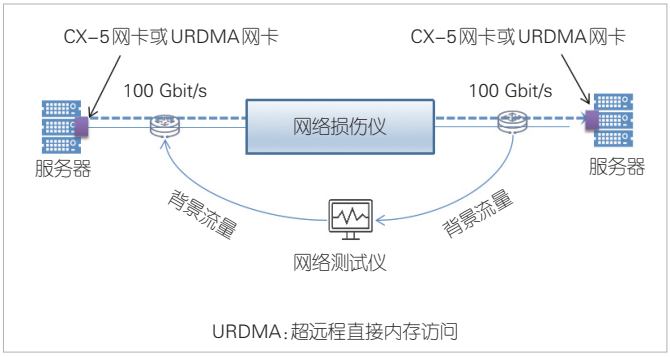
为验证URDMA网卡系统性能，我们搭建了如图7所示的测试环境，对标准RoCEv2技术、TCP协议以及URDMA网卡分别进行吞吐性能测试。该测试通过网络损伤仪模拟广域网的丢包率和RTT时延，其中包含了多种丢包率、RTT时延情况。同时该测试使用网络测试仪按需增加测试背景流量，模拟网络带宽竞争场景。

3.2 测试结果

为公平对比，3种协议都测试单流下的吞吐性能。如图



▲图6 新型重传机制



▲图7　测试组网拓扑

8所示，随着时延（对应数据传输距离）和丢包率的增大，标准RoCEv2协议和TCP类协议吞吐急剧下降，尤其是丢包对两者的吞吐影响很大，而URDMA协议的吞吐相对稳定。

1）TCP协议的性能随时延和丢包率的变化如图8（a）。TCP在极低时延和0丢包下吞吐性能为47.6 Gbit/s。这证实其在数据中心网络中可以保持较高性能，性能瓶颈在于主机CPU和内存的配置。

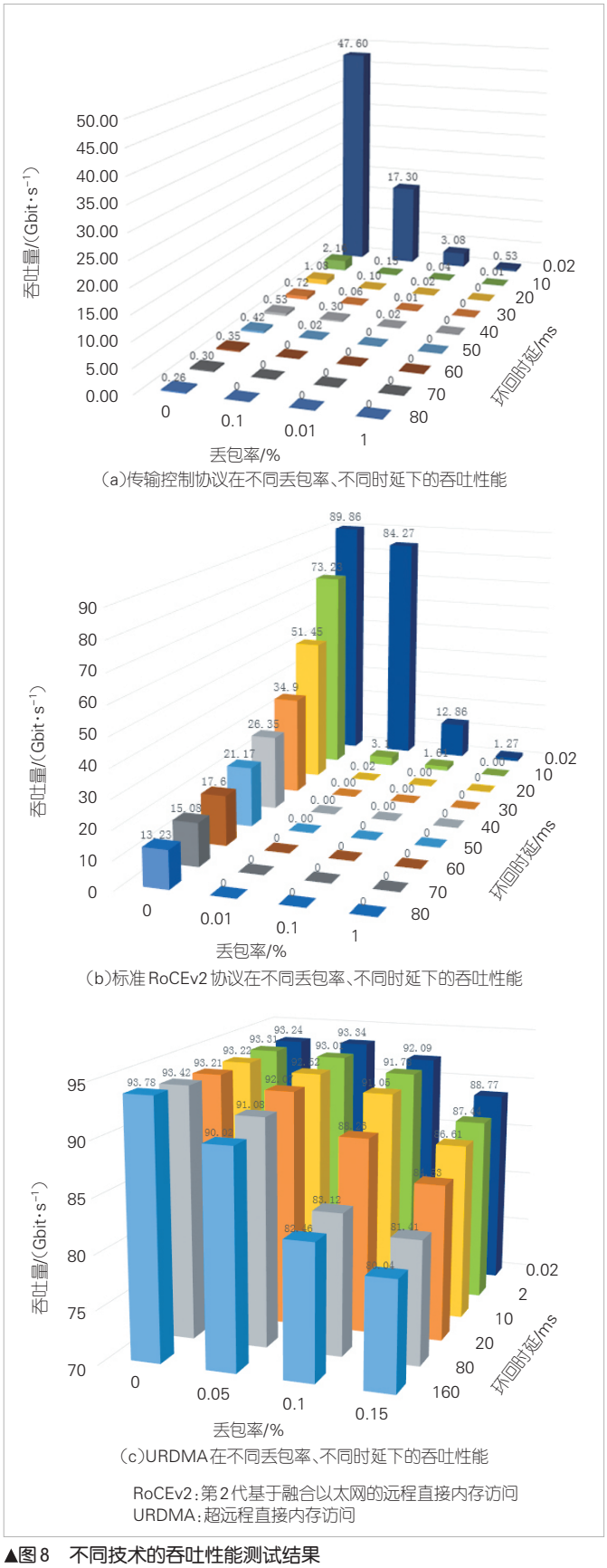
2）标准RoCEv2的性能如图8（b），其在极低时延和0丢包下也有较高性能（吞吐量达到89.86 Gbit/s）；在0丢包但时延变大的情况下，吞吐量会缓慢降低，在80 ms时衰减到13.23 Gbit/s。但如果增加丢包，RoCEv2的性能会急剧衰减，直至不可用。这表明标准RoCEv2对丢包非常敏感，只能在无损网络下使用。

3）URDMA协议性能如图8（c），其性能虽然也会随着时延和丢包变化，但衰减幅度较小。在RTT为20 ms、丢包率为0.1%的网络条件下，TCP类协议吞吐仅为0.02 Gbit/s，标准RoCEv2性能衰减到几乎为0，URDMA协议吞吐性能为88.26 Gbit/s。在RTT为80 ms时，TCP和RoCEv2协议吞吐都降为0，已不可用，URDMA协议吞吐性能为83.12 Gbit/s，仍然保持较高的性能。

4 结束语

RDMA技术已广泛应用于高性能存储、分布式训练等数据中心网络，其高吞吐、低时延性能已得到证实。RoCEv2协议凭借其良好的兼容性逐渐成为高性能网络的主流技术，但其在广域网的应用还在探索中，尚无成熟的商用方案。

本文中，我们从广域网高吞吐数据快递需求出发，分析了当前长距高吞吐协议与技术机制的不足，提出数据快递广域抗损高吞吐技术体系，并研发了URDMA算力低损耗网卡。实际测试结果表明：在广域有损网络环境下，基于广域抗损高吞吐技术的URDMA网卡吞吐性能相较于标准RoCEv2协议和TCP协议有上百倍的提升，解决了数据快递、



▲图8　不同技术的吞吐性能测试结果

跨智算集群分布式AI训练等为典型应用场景海量数据广域高效传输瓶颈问题，提升了数据流通效率。

参考文献

[1] FAST获批项目 [EB/OL]. [2025-10-09]. https://fast.bao.ac.cn/cms/category/approved_projects/

[2] 中国科学院计算技术研究所, 鄢贵海. 专用数据处理器(DPU)技术白皮书 [R]. 2021

[3] JACOBSON V, BRADEN R T, BORMAN D. TCP extensions for high performance [EB/OL]. [2025-10-08]. https://www.semanticscholar. org/paper/TCP-Extensions-for-High-Performance-Jacobson-Braden/a5fc067bca0ee49e047fffd89fc8cd2686f3be21

[4] Jacobson V. Modified TCP congestion avoidance algorithm [EB/OL]. [2024-10-09]. https://www. semanticscholar. org/paper/Modified-TCP-Congestion-Control-Algorithm-for-in-Roy/68507f828ac07a7051150785ebbcd3cfa7e3bbbe

[5] HA S, RHEE I, XU L S. CUBIC: A new TCP-friendly high-speed TCP variant [J]. ACM SIGOPS operating systems review, 2008, 42(5): 64 – 74. DOI: 10.1145/1400097.1400105

[6] TAN L S, YUAN C, ZUKERMAN M. FAST TCP: fairness and queuing issues [J]. IEEE communications letters, 2005, 9(8): 762 – 764. DOI: 10.1109/lcomm.2005.1496608

[7] BRAKMO L S, PETERSON L L. TCP Vegas: end to end congestion avoidance on a global Internet [J]. IEEE journal on selected areas in communications, 2006, 13(8): 1465 – 1480. DOI: 10.1109/49.464716

[8] CARDWELL N, CHENG Y, GUNN C S, et al. BBR: Congestion-based congestion control [J]. Communications of the ACM, 2017, 60(2): 58 – 66. DOI: 10.1145/3009824

[9] RICHARD S, FENNER B, RUDOFF A M. UNIX网络编程卷1：套接字联网API [M]. 北京: 人民邮电出版社, 2009

[10] 李博杰. 基于可编程网卡的高性能数据中心系统 [D]. 合肥: 中国科学技术大学, 2019

[11] 任宏. 关于TOE技术的发展及概况的研究 [J]. 红外, 2005, 26(3): 19-26. DOI: 10.3969/j.issn.1672-8785.2005.03.005

[12] 英特尔亚太研发有限公司. Linux开源网络全栈详解: 从DPDK到OpenFlow [M]. 北京: 电子工业出版社, 2019

[13] InfiniBand℠ Trade Association. InfiniBand™ architecture specification release 1.4 [EB/OL]. [2024-10-03]. https://www.infinibandta.org/tag/infiniband-architecture-specification/

[14] InfiniBand Trade Association. Supplement to infiniBand architecure specification volume 1 release 1.2.2 annex A 16 [EB/OL]. [2024-10-03]. https://www.infinibandta.org/tag/infiniband-architecture-specification

[15] Intel. Understanding iWARP [EB/OL]. [2024-10-05]. https://www.intel.com/content/dam/support/us/en/documents/network/sb/understanding_iwarp_final.pdf

[16] InfiniBand Trade Association. Supplement to infiniBand architecure specification volume 1 release 1.2.2 annex A 17 [EB/OL]. [2024-10-05]. https://www.infinibandta.org/tag/infiniband-architecture-specification

[17] ROCA V, BEGEN A. RFC8680 forward error correction (FEC) framework extension to sliding window codes [EB/OL]. [2024-10-06]. https://www.rfc-editor.org/rfc/rfc8680.html

[18] LUBY M, VICISANO L. RFC3695 compact forward error correction (FEC) schemes [EB/OL]. [2025-10-06]. https://www.rfc-editor.org/rfc/rfc3695.html

作 者 简 介

段晓东，中国移动通信有限公司研究院副院长、“新世纪百千万人才工程”国家级人选，教授级高级工程师；长期从事下一代互联网、算力网络、5G网络架构、6G网络架构、SDN/NFV等技术研究工作。

陆璐，中国移动通信有限公司研究院基础网络技术研究所副所长、中国通信标准化协会TC5核心网组组长；长期从事算网一体，以及移动核心网策略、演进、标准和技术研究工作，主要涉及未来网络架构、智能管道、边缘计算、算力网络等领域。

孙滔，中国移动集团首席专家，正高级工程师，中国科学技术协会第十届全国委员会委员；长期从事移动通信网架构、IP新技术研究和标准化工作。

李志强，中国移动通信有限公司研究院基础网络技术研究所高级工程师；长期从事未来IP网络演进、标准和技术研究工作，涉及下一代IP网络、算力网络、云网融合、SDN/NFV、5G核心网等。

杨红伟，中国移动通信有限公司研究院基础网络技术研究所研究员，高级工程师；长期从事下一代IP网络的技术和应用研究工作。

杜宗鹏，中国移动通信有限公司研究院基础网络技术研究所研究员，高级工程师；研究方向为算力网络、未来IP网络、确定性网络等；已发表论文10余篇。