



专题：确定性网络

面向广域 RDMA 的确定性网络需求与技术

赵俊峰¹, 李芳¹, 叶晓峰², 江淞³

(1. 中国信息通信研究院, 北京 100191;

2. 江苏为是科技有限公司, 江苏 苏州 215000;

3. 国网江苏省电力有限公司信息通信分公司, 江苏 南京 210024)

摘要: 随着生成式人工智能 (artificial intelligence generated content, AIGC) 大模型和智能计算应用的高速发展, 面向广域远程直接存储器访问 (remote direct memory access, RDMA) 的确定性网络技术成为近期智算中心互联的研究热点。首先分析了广域 RDMA 技术发展路线、实现方案和技术挑战, 然后研究广域 RDMA 部署时面临的各类网络问题, 通过搭建实验室和外场测试环境开展了一系列测试验证, 分析提出了广域 RDMA 的确定性需求特性和边界性能指标, 最后对面向广域 RDMA 的确定性网络技术进行了总结和展望。

关键词: 广域 RDMA; 确定性网络; 边界性能指标

中图分类号: TP393

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2023248

Research on deterministic networking requirements and technologies for RDMA-WAN

ZHAO Junfeng¹, LI Fang¹, YE Xiaofeng², JIANG Song³

1. China Academy of Information and Communication Technology, Beijing 100191, China

2. Jiangsu Viscore Technologies Co., Ltd., Suzhou 215000, China

3. State Grid Jiangsu Electric Power Co., Ltd., Information and Telecommunication Branch, Nanjing 210024, China

Abstract: With the rapid development of AIGC large models and intelligent computing applications, deterministic networking technology for wide-area RDMA has become a recent research hotspot. The development path, implementation solutions and technical challenges of RDMA-WAN were analyzed, by setting up a laboratory and an out-field testing environment, with a series of testing and verifications. The deterministic demand characteristics and boundary performance indicators of RDMA-WAN were proposed. Finally, the deterministic networking technologies for RDMA-WAN were summarized and prospected.

Key words: RDMA-WAN, deterministic networking, boundary performance indicator

收稿日期: 2023-10-08; 修回日期: 2023-11-07

基金项目: 国家电网公司总部科技项目 (No.5700-202318269A-1-1-ZN)

Foundation Item: State Grid Corporation Headquarters Technology Research Project (No.5700-202318269A-1-1-ZN)



0 引言

作为一种提升业务承载质量、前景广阔的网络技术，确定性网络（deterministic networking, DetNet）引起了学术界和产业界的广泛关注。与 IP 等传统分组交换网络不同，确定性网络可提供可预期的网络服务质量，包括确定性的带宽、丢包率、时延和抖动性能。随着光传送网（optical transport network, OTN）、灵活以太网（flex ethernet, FlexE）、切片分组网（slicing packet network, SPN）、时间敏感网络（time sensitive networking, TSN）、确定性网络、优先级调度队列增强机制、网络演算等各类确定性技术的不断涌现，确定性网络技术体系逐步完善。国内设备商、运营商和行业用户相继开展核心技术攻关研发，并在工业互联网、智能电网、车联网等领域开展了应用示范。近年来，随着生成式人工智能（artificial intelligence generated content, AIGC）大模型和智能计算应用的高速发展，面向智算中心互联的确定性网络技术成为技术研究的新方向。

此外，由于远程直接存储器访问（remote direct memory access, RDMA）具备低时延、高带宽和低 CPU 利用率等特点，可显著提升数据中心网络的性能和效率，国外很早就开展了长距离传输上的应用实践。从 2007 年到 2011 年，12 875 km（即 8 000 英里）以上的长途测试证明 RDMA/IB 比 TCP/IP 以太网具有更好的性能^[1-2]。目前，基于以太网的 RDMA 协议 RoCEv2（RDMA over converged Ethernet v2）在广域网上已有部署实践。由于 Internet 流量的复杂性，且 RoCEv2 基于 UDP 实现，无论在运营商网络还是数据中心内部，Internet 流量都会受到网络限制（如 UDP 限流、端口屏蔽等），甚至会受 TCP 流量的压制。由于确定性网络技术能够针对不同类型的流量和具体应用部署硬隔离切片，因此 RoCEv2 在广域确定性网络环境下面临的问题将大大简化。但是，在广域确定

性网络上部署时仍需解决以下主要挑战^[3-5]。

- 长距离保持链路利用率和高吞吐量能力。
- 提供拥塞控制，避免数据包丢失。
- 在多个并发应用之间公平共享链路带宽资源。
- 数据包确认时延随着距离的延长呈指数增长，高性能应用特别是分布式计算架构面临着挑战。

本文通过建立实验室和外场测试环境，对主流商用 RDMA/RoCEv2 网卡进行了长距离和各种强网、弱网环境下的性能测试研究，包括通过网络损伤仪表仿真不同丢包率、时延和抖动等网络损伤特性。结合对各类广域 RDMA 技术机制的分析，提出影响吞吐量的 RDMA 关键语义和网络性能指标，明确广域 RDMA 在实际部署中对确定性网络的指标需求，为 RDMA 网卡设备、智算中心网络和确定性网络设备的研发、评测和部署提供试验数据支撑和技术参考。

1 广域 RDMA 技术发展

基于不同的广域网络能力、源宿端设备类型、控制机制和算法，广域 RDMA 的技术路线目前主要包括改造网卡、升级广域网和引入网关或交换设备 3 类，广域 RDMA 的技术路线如图 1 所示。其中，技术路线 1 是通过改进 RDMA 网卡实现基于有损网络的长距离、高效的传输；技术路线 2 是通过优化广域网络实现跨数据中心的无损传输，并进一步拓展广域网拥塞控制算法和流控机制；技术路线 3 是在数据中心出口处采用网关设备，隔离或中继数据中心内部和广域网络的流量。对比分析 3 类技术路线，数据中心内均需支持无损 DCN，数据中心间互联可采用有损或无损网络，通用网卡、交换机或独立网关设备实现广域 RDMA 的高性能互联。

1.1 基于无损网络的扩展技术

RDMA 设计最早基于无限带宽（InfiniBand，

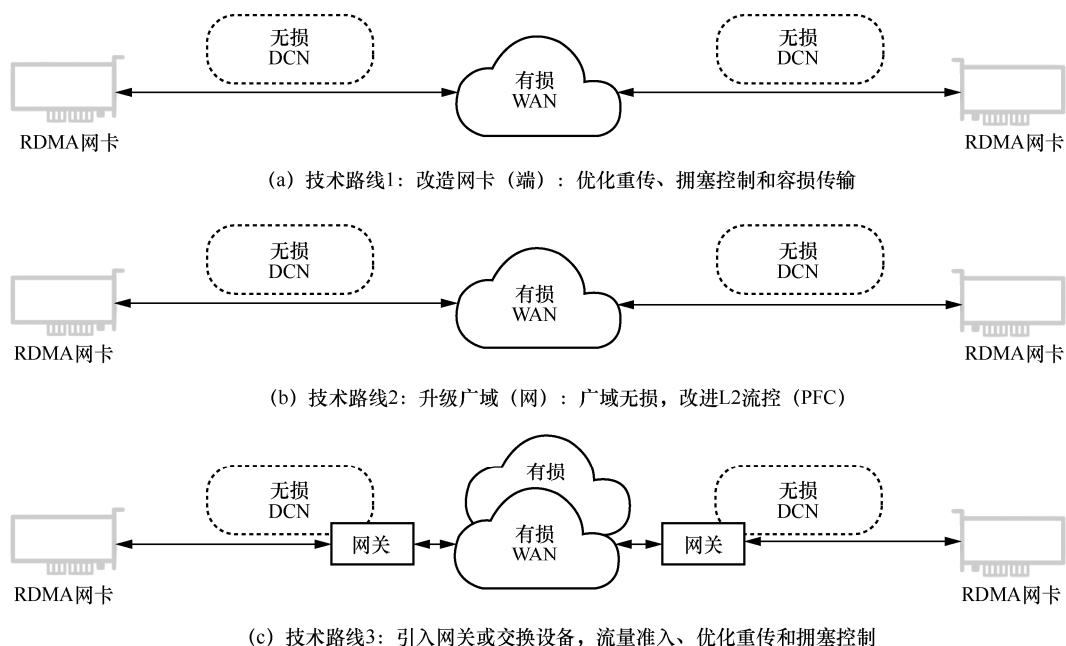


图1 广域 RDMA 技术路线

IB) 网络, IB 链路层采用逐跳的、基于信用的流控机制来保证无损传输。RoCEv2 融合了 RDMA 与以太网的技术特性, 实现了 RDMA 在 3 层 IP 网络上的扩展, 并通过配置优先级流量控制(priority flow control, PFC) 支持无损网络。然而, PFC 是一种基于端口或端口+优先级的粗粒度机制, 无法实现基于业务流的精准流控。PFC 会导致拥塞蔓延(congestion-spreading)现象, 进而引发不公平调度、受害者流、PFC 死锁(deadlock)、PFC 风暴等网络性能问题。因此, 在数据中心内部, RDMA/RoCEv2 大规模组网仍面临诸多挑战^[6]。

针对 PFC 机制缺陷, 流级的精准拥塞控制技术成为业界共识。通过引入 L4 的流级拥塞控制算法, 网络路径上的交换机实现基于每流的拥塞标记, 携带拥塞标记信号的报文到达接收端后, 再被传回发送端由其根据网络拥塞情况进行调速。以上机制既避免了网络过度拥塞, 又充分利用了网络带宽。除了网卡端到端流控和拥塞控制, 目前业界还在研究与交换机配合的端网协同拥塞控制算法。如何与确定性网络设备实现端网协同, 也将是后续研究方向之一。

解决大规模组网下链路利用率的提升问题, 包级负载均衡成为技术研究热点。传统分组交换网络采用基于流的等价多路径(equal cost multipath, ECMP)技术来实现负载均衡。当流数较少且大流量占多数时, 往往会造成路径冲撞, 进而导致链路带宽无法被充分利用^[5]。为此, 业界提出了多路径 RDMA 思路^[3], 以数据包为粒度把流分散到多个等价路径上, 网卡硬件实现多路径传输层逻辑, 包括拥塞感知的流量切分、收端乱序接收等。网络设备将流分散到多个等价路径上实现更优的均衡性能和更高的多链路利用率。目前, 具体的实践方案包括亚马逊的可扩展可靠数据报文(scalable reliable datagram, SRD)^[7]、英伟达的自适应路由以及华为和阿里的智能路由等方案。

在数据中心广域互联场景中, 长距离和大规模组网特性将带来更大的网络性能挑战。在长距离 RDMA 组网时, 无损 RDMA 需要一个深度网络缓冲区。由于 PFC 的响应消息与时延强相关, 数据中心间流量可能无法达到最佳吞吐量。为此, 微软的 Azure 云部署了跨域数据中心之间的基于 RDMA 的分布存储网络^[8], 将 RDMA 无损组网扩



展到 80~100 km 的广域范畴。另一类解决方案是通过专用的中继设备 PFC-Relay^[9], 将高性能无损 RDMA 扩展到长距离链路。

确定性网络通过延续 PFC 信号、长距离拥塞控制、流量控制及负载均衡等技术实现 RDMA 的长距离扩展。2022 年 8 月, 未来网络试验设施 (China Environment for Network Innovation, CENI) 基于确定性网络构建了南京到贵阳的长达 2 450 km 的试验网, 实现了 RDMA 无损流量跨广域传输。

1.2 基于有损网络的优化技术

有损网络环境存在丢包、乱序等网络性能的劣化现象, 这将导致 RDMA 传输吞吐量的大幅下降。因此, 乱序接收、端到端流控、丢包快速恢复和精准重传等无损优化技术受到业界的广泛关注。随着新一代 RDMA 网卡、智能网卡和数据处理单元 (data processing unit, DPU) 片上资源的增加, 不依赖 PFC 的 RoCEv2 的有损网络应用方案成为可能。新一代的 RDMA 网卡可在片上实现更高效的丢包恢复机制和更好的端到端流控, 从而让有损网络环境 (lossy) 的潜力不亚于无损环境 (lossless), 同时这也有可能将 RDMA 推广到更大规模的广域网络中。

RoCEv2 技术规范中 RoCE 拥塞控制允许根据拥塞情况调整速率。因此, 即使底层以太网发生损伤也能最大限度地减少丢包, 并保持高性能传输。目前, 业界提出了通过硬件拥塞控制减少控制处理时延, 从而在拥塞发生时, 更快速、准确地做出响应并调整传输速率^[10]。较短的控制时延还能防止网络缓冲区在拥塞情况下的溢出。此外, 在严重拥塞情况下, 快速硬件重传可显著降低数据包丢弃造成的性能损失, 从而对 L4 拥塞控制起到补充作用。

此外, 基于 RDMA 网卡无损优化的思路, 业界提出了一种改进的 RoCE 网卡 (improved RoCE NIC, IRN) 方案^[15]。相关研究表明, 在典型的网

络场景中 IRN (不含 PFC) 的性能比 RoCE 高 6%~83%^[11]。该方案对当前的 RoCE 网卡做了两个关键优化, 包括改进丢包恢复机制和基于带宽时延乘积 (bandwidth delay product, BDP) 的流量控制。BDP 通过网络的带宽时延乘积来限制在途数据包的数量。该方案既可以发挥链路的最大通信能力, 又降低了网络拥塞发生的概率, 并与其他 L4 拥塞控制机制互不冲突 (如数据中心量化拥塞通知 (data center quantized congestion notification, DCQCN)^[12]和 Timely^[13])。

国内厂商也尝试在硬件上改进并实现对长距离 RDMA 的支持, 通过在现场可编程门阵列 (field programmable gate array, FPGA) 上实现选择性重传、乱序接收和自研拥塞控制算法, 在广域确定性网络上也获得了一定成果^[14]。国内某运营商开展了贵阳天眼到北京国家天文台 2 200 km 的广域 RDMA 测试, 10 Gbit/s 带宽条件下实现的单流性能为 TCP 的数倍^[16]。

1.3 广域 RDMA 潜在的应用场景

RDMA 具备高通量传输、零复制、硬件卸载的特性, 避免了操作系统内核参与数据报文的处理, 节省了大量的 CPU 资源。广域 RDMA 可实现端到端低时延和高吞吐量传输, 能够满足算力网络、大数据传输和浪涌型 I/O 高并发、低时延类应用。国内外相关试验表明, 当广域网带宽超过万兆 (10 Gbit/s) 时, 相较于传统 TCP, 广域 RDMA 具备显著的性能优势, 其潜在的业务应用场景主要包括以下几种。

- 长距离大文件传输。近期, 在国内某运营商发布的数据快递技术白皮书中, 提出了通过广域 RDMA 实现端到端高吞吐数据传输的技术方案。
- 跨域存储访问。在微软 Azure 云的分布存储网络中, 需要将算力和存储资源跨域部署以实现存算分离。在此过程中, 跨数据中心需要实现基于 RDMA 的存储协议访问。

- 多方安全计算。多个数据中心的数据源和计算资源之间存在大量的数据交互,需要通过本地 GPU 直接访问远端内存。多方安全计算是 GPU 访问远程数据的典型应用之一。
- 超高清视频直播和制作。该场景涉及 8K 直播视频的 VR/AR 制作域应用。超高清 VR/AR 视频源需采用无压缩或浅压缩技术进行高质量实时传输,视频流通过 RDMA 可实现网卡到 GPU 和编码设备的直接传输、转发、转码和云端处理。美国亚马逊的云数字接口 (cloud digital interface, CDI) 就是基于 RDMA 实现的一种方案。

2 广域 RDMA 技术挑战

尽管 RDMA 在流量控制、拥塞控制和丢包处理等方面进行了优化,但在长距离、大规模的复杂组网的广域环境下,其仍会面临多方挑战,主要包括以下几种。

- 在 RDMA 无损部署下,经典 PFC 无法扩展到超过 100 km,广域无损需确定性网络提供保证;经典的 DCQCN 等拥塞控制算法在广域部署和算法参数层面仍面临很大局限性,需要研发新型拥塞控制算法。
- 在 RDMA 有损部署下,长距离重传代价较高。即使考虑选择性重传、丢包快速恢复等机制,也很难保证高效率,仍需端侧网卡技术攻关。但这同时带来各厂商网卡之间的算法兼容性问题。
- 在应用部署方面,对适合广域部署的 RDMA 应用,需要配合典型分布式应用场景,从 RDMA 语义和流量模型上做优化分析和全局调度。
- 在网络运营方面,流控和拥塞控制使得 RDMA 在数据中心实现了落地部署,但大规模网络的复杂配置成为新的挑战。运维人员需要根据不同的网络环境和流量模型

进行 RDMA 参数调优,相对确定的网络环境将大幅减小配置的复杂性。

- 在技术研发方面,确定性网络可保证丢包、带宽、时延和抖动等网络指标的可预期,这将为流量控制和拥塞控制算法改进提供准确的边界和判断条件,加快新型优化算法的快速收敛和迭代。

基于以上原因,本文对广域 RDMA 各种网络环境进行实验仿真和研究验证,得出基本的网络性能边界条件,并对广域确定性网络的具体指标给出参考建议。

3 广域 RDMA 确定性需求

业界已开展了关于 13 840 km (即 8 600 英里) 10~40 Gbit/s 的广域 RDMA 的性能验证测试,验证了 RDMA 可实现长距离无损的广域网高性能传输^[4-5]。其中,不可靠语义——不可靠连接 (unreliable connection, UC) 和不可靠数据报 (unreliable datagram, UD) 虽然具备高速和低时延传输特性,但在大规模组网时将面临流量不受控的问题。可靠连接 (RC) 是分布式计算规模组网中的主要方法,且 RC 的写操作性能在长距离下比读操作性能更优。因此,本文的实验室和现网测评主要围绕 RC 中的写操作开展。

首先,通过网络损伤仪在实验室模拟和测试在不同时延、抖动和丢包组合环境下对 RDMA 传输性能的影响,实验室测试框图如图 2 所示。基于以上验证思路,本文在 CENI 的广域确定性网络中对相关测试案例进行了对比评估,得出了满足广域 RDMA 高性能传输的基本网络性能边界条件,广域确定性网络-CENI 测试环境如图 3 所示。

3.1 长距离和时延影响

首先,在无损网络环境下开展了 RDMA RC 的写操作性能测试,作为后续复杂网络环境的基本评估参考。由于目前业界已实现了基于 PFC 的 100 km RDMA 互联部署,本文时延参数的选择主

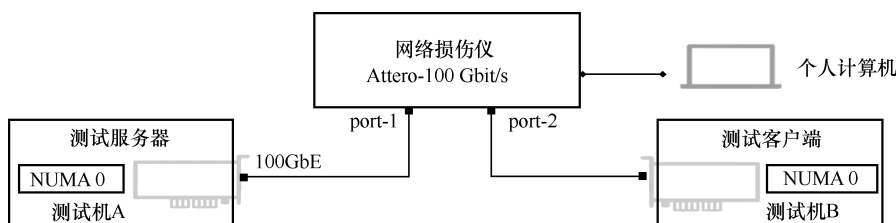


图2 实验室测试框图

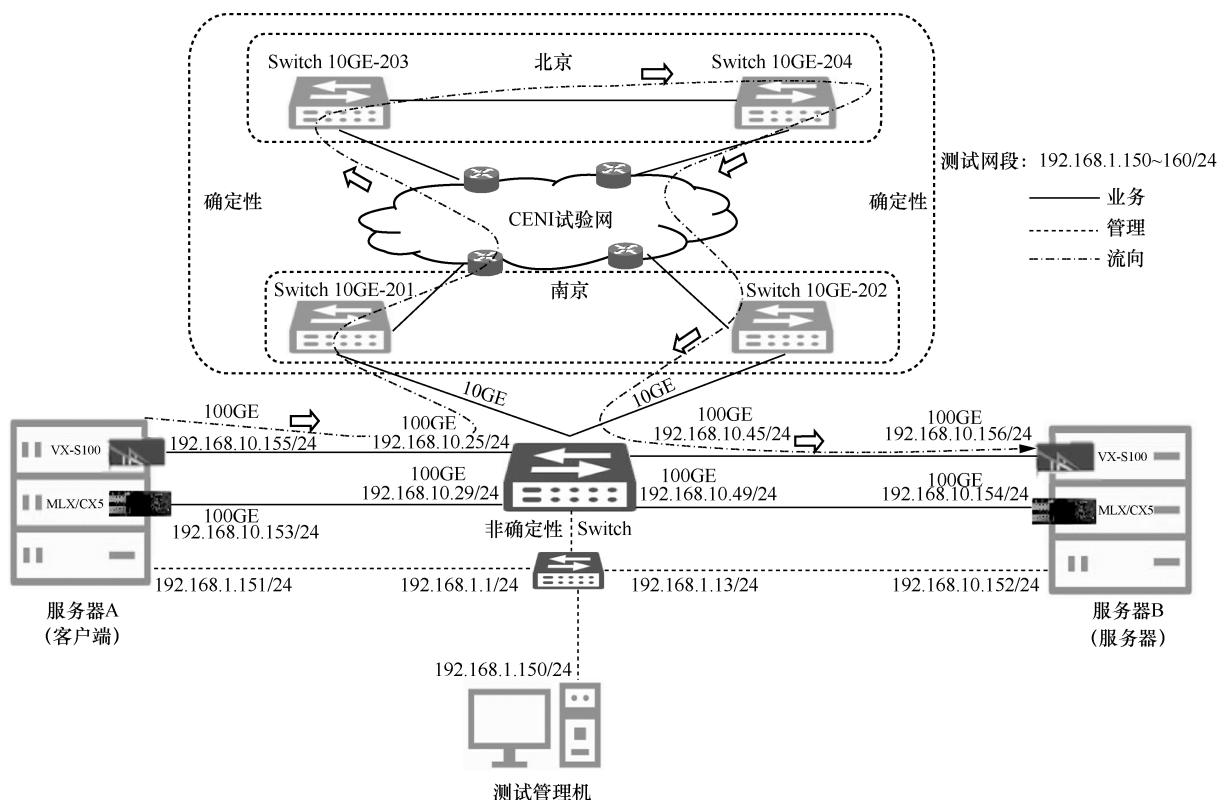


图3 广域确定性网络-CENI 测试环境

要面向 100~2 000 km 的广域场景,即往返时延(round-trip time, RTT)为 1~20 ms。在组合参数选择上,本文以 100 km (1 ms 时延)~2 000 km (20 ms 时延)为基础进行丢包和抖动损伤的叠加验证。时延和消息大小的参数选择见表 1。

表 1 时延和消息大小的参数选择

RTT/ms	消息大小	传输距离/km
<1	<1 024 byte	<100
1	≈256 KB	100
2	≈512 KB	200
5	≈1 MB	500
10	≈8 MB	1 000

通过在实验网络中的损伤仪叠加时延损伤,验证不同网络环境下的 RDMA 传输性能。长距离和时延对 RC 性能影响如图 4 所示,随着时延的增加(1~20 ms),需不断增大 RDMA 消息大小(message size)才能达到 100%吞吐量的高性能传输。由于 RC 支持最大消息大小为 2 GB,在无损情况下可达到 100 Gbit/s 的带宽,满足吞吐量理论式(1),但消息大小设置实际与具体的算网应用、设备缓存空间和缓存阈值设置存在相关性。

$$\text{Throughput} = \text{Window_Size} / \text{RTT} \quad (1)$$

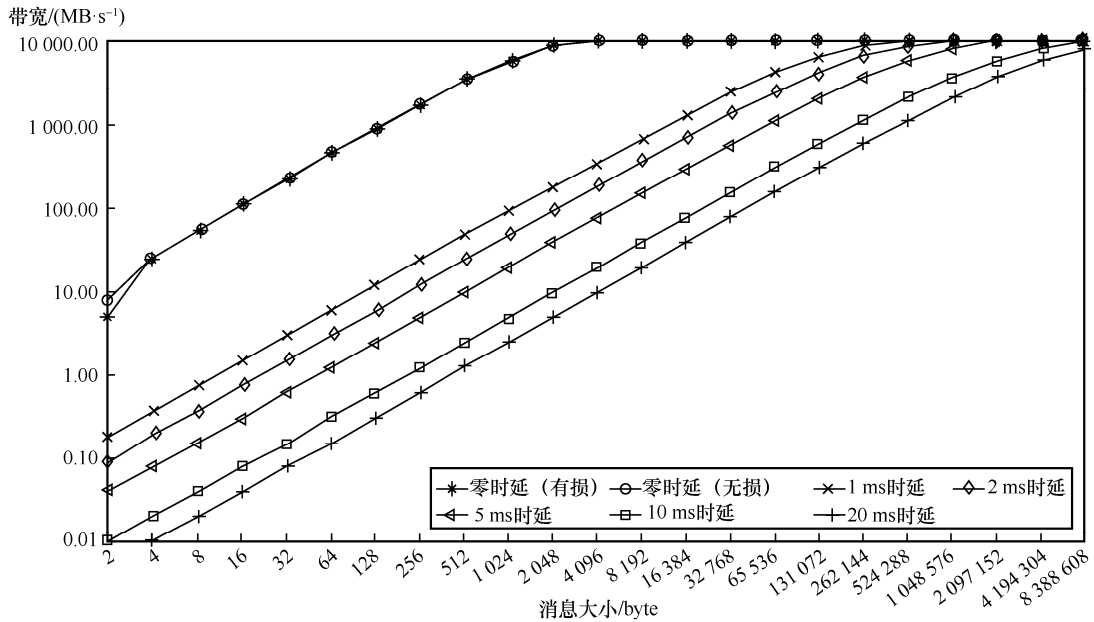


图4 长距离和时延对 RC 性能影响

3.2 时延和丢包影响

传统的 RDMA 采用 Go-Back- N 重传机制, 即重传丢弃的数据包 N 之后的所有数据包, 丢包对 RDMA 会对造成大幅性能下降。而 TCP 只需要重传丢失的单个数据包, 最新的 RDMA 网卡开始采用选择性重传 (selective repeat), 因此, 可参考 TCP 丢包率 (p)、时延和带宽的计算式:

$$BW \leq \frac{MSS}{RTT} C p^{-\frac{1}{2}} \quad (2)$$

RDMA 实际测试性能和 TCP 存在差异, 广域网络的主要影响是时延, 且重传和拥塞控制算法模型基本相似。因此, 通过调整式 (2) 中的参数 C 的取值对 RDMA 的理论速率进行经验判断 (TCP 经验值 $C \approx 1.0$) [15]。

丢包对 RC 性能影响如图 5 所示, 通过网络损伤仪设置不同的丢包率, 当丢包率 $> 0.01\%$ 时, 吞吐量性能低于 86.5 Gbit/s; 当丢包率 $> 0.1\%$ 时, 吞吐量性能下降到 42.6 Gbit/s; 当丢包率 $> 0.2\%$

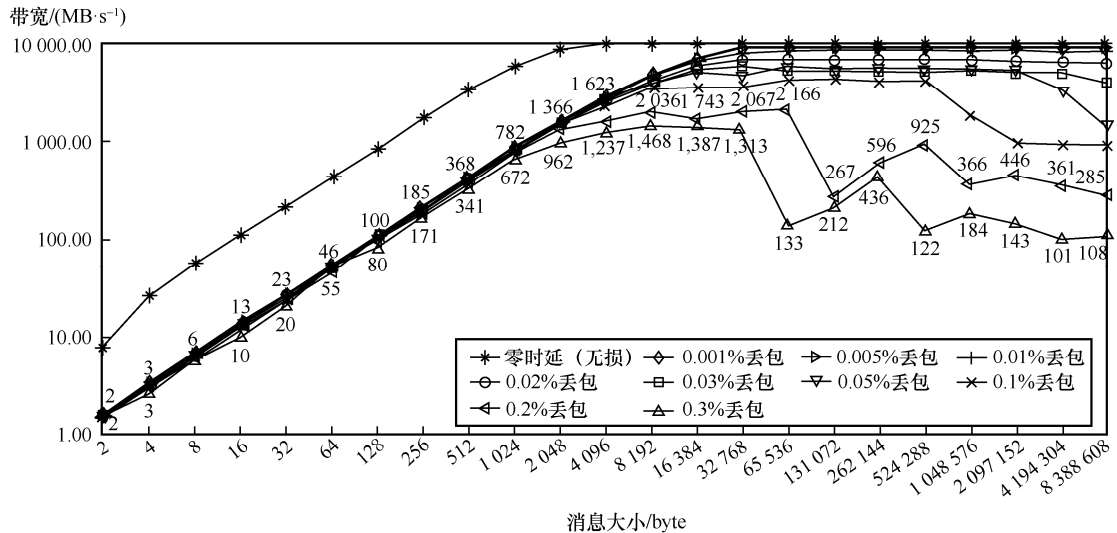


图5 丢包对 RC 性能影响



时, 吞吐量性能下降至 21.6 Gbit/s。

时延和丢包对 RC 性能的影响(一)如图 6 所示, 当 1 ms 时延(100 km 线路距离)叠加 0.001%丢包时, 传输性能最高仅为 70 Gbit/s, 低

于零时延下(相当于数据中心内部)0.02%丢包的情况。依次类推, 分别分析 1 ms 时延与不同丢包率叠加下的吞吐量。

时延和丢包对 RC 性能的影响(二)如图 7 所示,

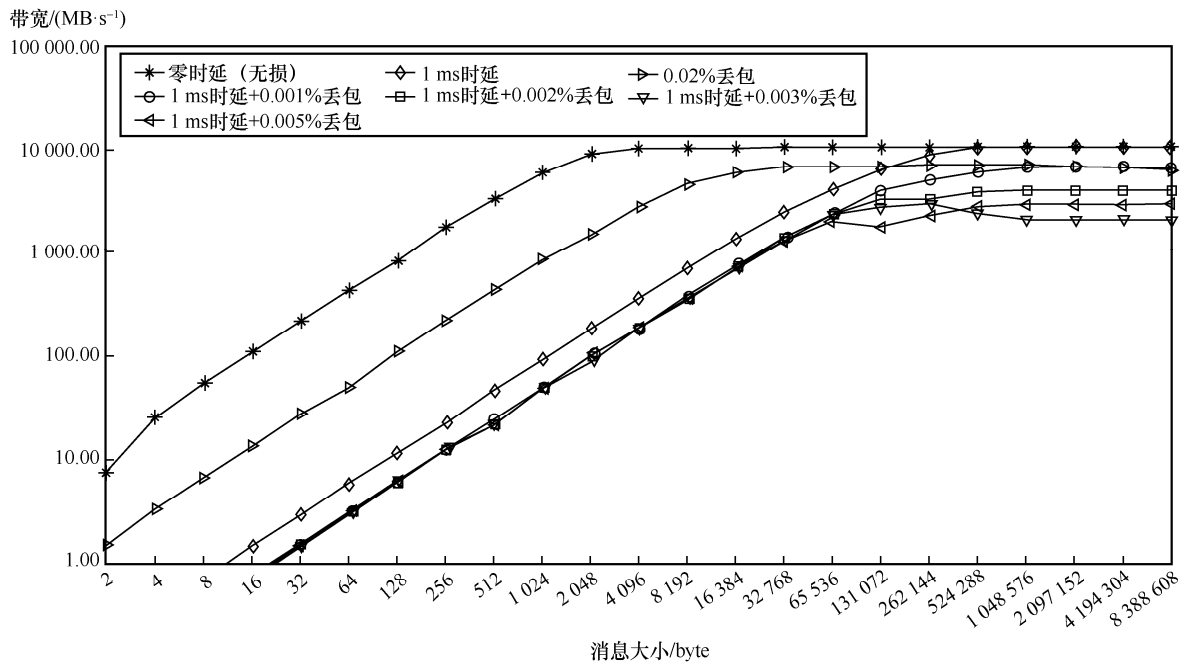


图6 时延和丢包对 RC 性能的影响(一)

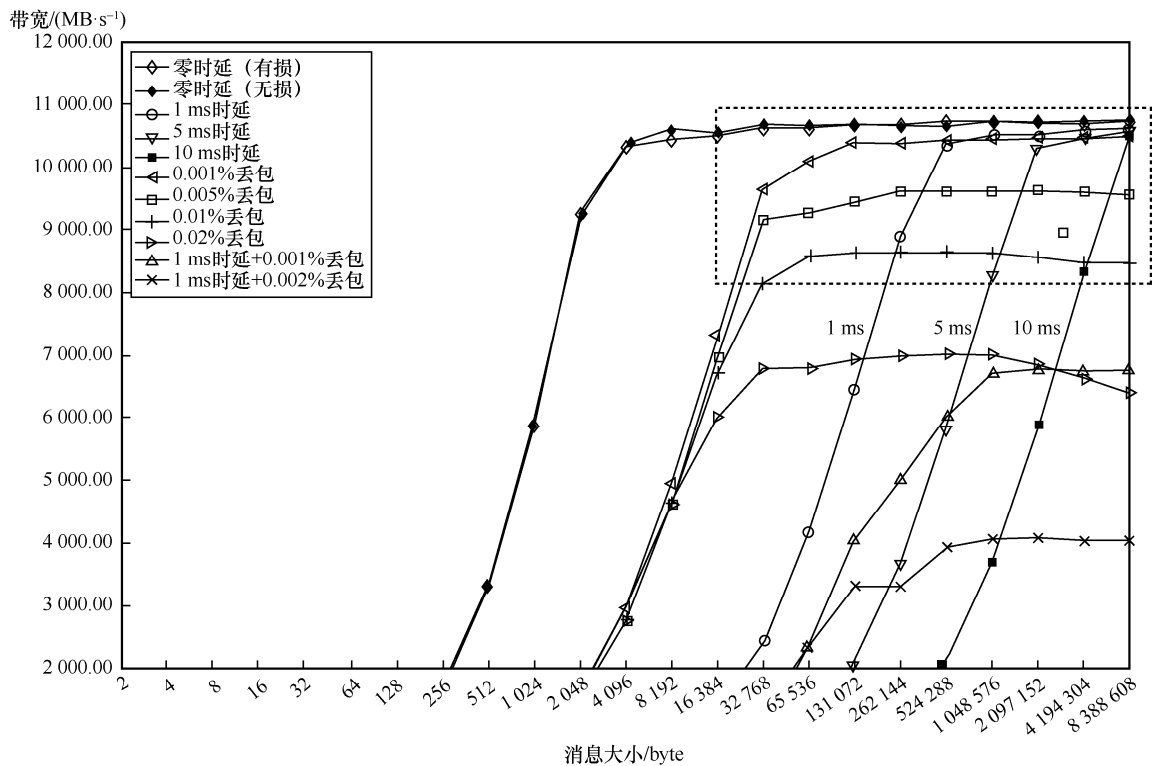


图7 时延和丢包对 RC 性能的影响(二)

当时延和丢包损伤同时存在且达到 80% 的吞吐量目标时, RDAM 参数设置和广域网络确定性性能需在图 7 中的虚线窗口内。数据中心内丢包率需小于 0.005%; 在数据中心广域互联场景时, 由于固有线路时延带来的重传代价和响应时间增大, 相比于数据中心内, 其丢包门限更为苛刻, 要求网络尽量实现无损。在广域场景下, 即使采用选择性重传的优化算法, 在丢包率<0.001%的情况下, 带宽利用率也很难达到 70%以上。考虑 RDMA 的 RC 对丢包的敏感度非常高, 确定性网络的无损保障就非常重要。选择性重传、容损机制和零丢包确定性网络将为高性能的 RDMA 传输提供保障。基于此思路, 微软研究团队^[17]对数据中心内各种 RDMA 网卡的重传算法行为和效果进行了分析。本文后续也会针对各类商用网卡的选择性重传在广域确定性网络下的效果进行深入评估。

3.3 时延和抖动影响

本文验证了不同时延和抖动组合下 RDMA 传输性能。抖动对 RC 性能影响如图 8 所示, 通过对网络损伤仪分别模拟零时延 (0 ms)、1 ms、2 ms, 以及在 1 ms 叠加 100 μ s、200 μ s、1 000 μ s 的网络损伤等情况进行对比分析, 具体如下。

- 1 ms+100 μ s 抖动, 性能表现总体低于 2 ms 无抖动的网络。

- 抖动>200 μ s 时, 大消息类应用 (消息大小>1 MB), 出现显著的性能下降的现象。

在同样的时延条件下, 图 8 选择性地标注了 512 byte 和 512 KB 两个典型消息长度, 测试验证了叠加 100 μ s 抖动将会导致 19%~41%的吞吐量性能下降。

在 RDMA RC 的典型应用场景中, 存储类 I/O (如 NV Me over RoCE) 消息大小为 512 byte~8 KB。时延和抖动对 RC 性能影响 (存储类 I/O 应用) 如图 9 所示, 存储类 I/O 应用叠加时延、抖动损伤的对比分析具体如下。

- 抖动比时延损伤的负作用更为明显。
- 1 ms (100 km): 512 byte 消息、100 μ s 抖动会造成吞吐率下降 50% (46.0~23.3 MB/s) 见图中①。
- 10 ms (1 000 km): 512 KB 消息、100 μ s 抖动会造成吞吐率降低 17.3% (2 128~1 759 MB/s) 见图中②。
- 对于大消息 (消息大小>1 MB), 1% 的抖动会显著降低传输性能。

数据和文件传输类应用的消息大小为 64 KB~4 MB。时延和抖动对 RC 性能影响 (数据和文件传输类应用) 如图 10 所示, 数据和文件传输类应用叠加时延、抖动损伤的对比分析具体如下。

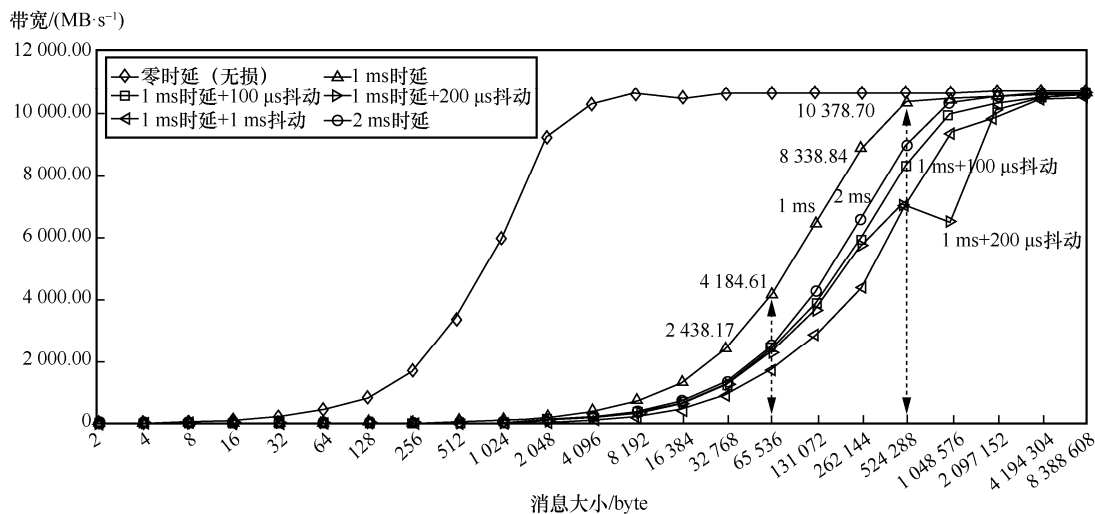


图 8 抖动对 RC 性能影响

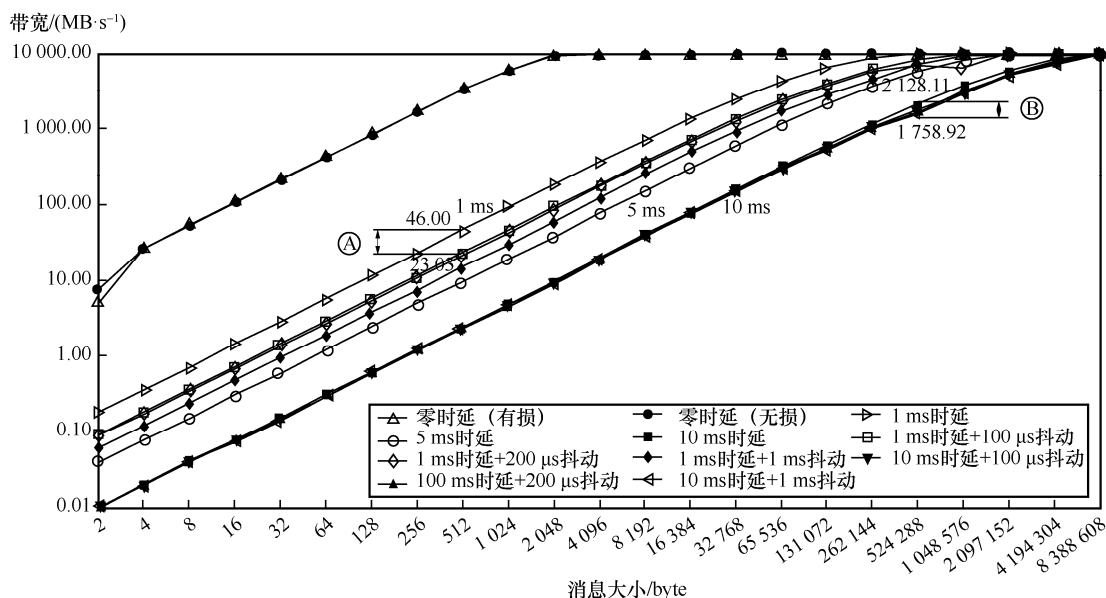


图9 时延和抖动对 RC 性能影响 (存储类 I/O 应用)

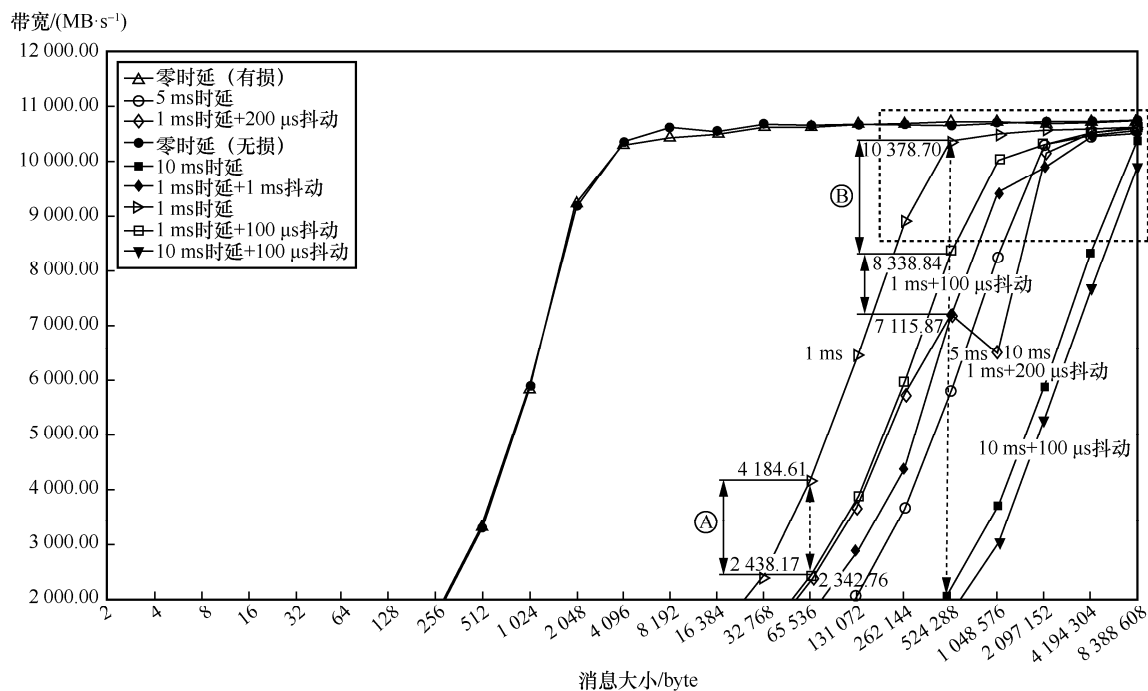


图10 时延和抖动对 RC 性能影响 (数据和文件传输类应用)

- 在 1 ms (100 km) 场景时, 64 KB 消息, 100 μs 的抖动吞吐率降低 41.73% (4 185 MB/s→2 438 MB/s), 见①; 512 KB 消息, 100 μs 的抖动吞吐率降低 19.65% (10 379 MB/s→8 339 MB/s), 见②。

- 长距离需要达到 80%以上带宽利用率的场景, 要求网络环境在虚线窗口内, 消息大小需大于 512 KB, 且随着距离增加而增大; 跨数据中心抖动<100 μs。

基于以上分析可得, 抖动对网络带来影响

更为显著。RDMA 拥塞检测方法包括基于显式拥塞通知 (explicit congestion notification, ECN) 的 DCQCN^[12]和基于 RTT 的 Timely^[13]。基于 ECN 反馈的拥塞控制响应速度慢, 且要求网络交换设备支持 ECN, 对广域组网有条件限制。而基于 RTT 的拥塞检测基于如下假设, 即如果发生拥塞, 数据包会在接收队列中排队等待, RTT 也会相应较高, Timely 通过 RTT 的梯度变化来衡量拥塞情况。以上机制需对时延做精确计算, 要求接收方网卡对每一个接收包回复一个确认信息。但长距离特别是超过 1 000 km (10 ms) 时, 网络抖动并非一定是拥塞造成的, 因此对 RTT 拥塞控制会造成误判或过度响应。目前, 广域确定性网络可以保证 20~30 μ s 级别的抖动控制, 如能确保抖动和拥塞负载的正相关性, 那么基于 RTT 的广域拥塞控制算法的应用就将成为可能。

另外, RDMA 网卡在实现选择性重传算法时, 网络抖动及乱序接收重传对网卡的缓存和内存资源占用影响较大。在长距离场景下, 广域确定性网络和非确定性网络对网卡和主机内存资源的使用差异, 有待进一步量化分析。

3.4 确定性需求特性总结

根据广域 RDMA 的技术发展路线, 结合上述的实验和测试数据分析, 面向广域 RDMA 的确定性需求特性见表 2。

4 面向广域 RDMA 的确定性网络技术展望

由于广域确定性网络具备网络切片、业务硬隔离等能力, 因此 RDMA 流量在广域网传输时可以避免其他流量 (如 TCP/UDP 和流媒体) 的干扰, 更容易实现面向 RDMA 的流量控制算法, 并可通过应用模型和 RDMA 语义进行更细粒度的流量调度。

表 2 面向广域 RDMA 的确定性需求特性

项目	网络有损原因	无损网络特征	无损网络技术	广域无损挑战	确定性需求	
					性能	功能
设备	交换机浅缓存 (MB 级别), 端口共享缓存	零丢包: 避免交换机缓存溢出	流量控制: 发送速率匹配接收速率	流量控制: - PFC 流控粒度大, 影响吞吐量; - 广域端到端时延大, 端到端流控响应慢	(1) 无损零丢包 - DC 内: 丢包率 <0.005%; - DC 间广域互联: 无损零丢包	- 设计更好的广域流控算法, 精细化粒度; - 网络切片和隔离
	Incast 形成突发 (多打一)	低时延: 降低交换机排队时延	拥塞控制: 防止过多数据注入网络, 避免负载过大	拥塞控制: - 基于 ECN 反馈的拥塞控制响应速度慢, 且要求网络设备支持; - 基于 RTT 的拥塞控制对于 RTT 变化敏感, 易响应过度; - 基于 INT 的拥塞控制需要端网协同	(2) 确定性低时延 - 可预期时延简化算法优化和调参; - 低时延减少缓存消耗和响应时间 (3) 确定性低抖动 - DC 间广域互联: 抖动 <100 μs, 消息长度大于 512 KB (80%吞吐性能) (4) 硬隔离切片 RDMA 与 TCP/UDP 流隔离, 避免带宽强占 (5) 高精度时频同步 - 支撑基于精准时延测量的拥塞控制算法	- 确定性设备支持 ECN; - 精确控制抖动和时延固定部分, 使得 RTT 变化和网路拥塞正相关; - 支持逐流 INT/Telemetry; - 发挥确定性时间同步优势, 按时间切片速率控制
链路	流量分布不均, 部分链路过载 (负载均衡)	高吞吐: 提高链路带宽利用率	流量调度: 调度业务流量, 均衡链路负载	流量调度: - 广域端到端多路径时延差异大, 乱序难处理; - 数据平面技术异构, 难以通过改变流标签等实现重路由		端网协同能力 - 端侧: RDMA 支持乱序接收, 兼容自适应路由算法; - 网侧: 广域智能流量调度, 根据应用特性预测流量模型



此外，确定性网络具备跨数据中心高精度时间同步特性。除了可支持在广域部署高性能分布式应用（如基于时间同步的高性能分布数据库），新型拥塞控制算法也可以考虑应用此特性。RTT 精确计算可依赖精准时间戳，实现单向时延的快速计算，从而对上下行不对称的链路进行流控。拥塞控制中速率控制也可按照时间片来设定步长并结合网络切片，实现对网络流量的精准调节。

针对广域无损技术方案，确定性网络需更多地参与到流量控制和拥塞控制机制优化中。确定性网络通过支持 PFC、ECN 和新型流控协议，构建端网协同的 RDMA 高性能互连网络^[18]。

5 结束语

结合对各类广域 RDMA 技术机制和问题挑战的分析，本文对广域 RDMA 的各种网络环境进行实验研究和仿真测试验证，得出了基本的网络性能边界条件，并对广域确定性网络需求特性给出参考建议，明确广域 RDMA 在实际部署中对确定性网络的指标需求。为实现广域智算中心互联和确定性网络的协同部署和应用，后续仍需加快技术研发与标准化推进，提高技术的成熟度和可靠性，为智算中心高质量互联提供数据参考和技术支撑。

参考文献：

- [1] CARTER S, MINICH M, RAO N S. Experimental evaluation of in finiband transport over local-and wide-area networks[C]//Proceedings of the 2007 Spring Simulation Multi Conference-Volume 2. [S.l.: s.n.], 2007: 419-426.
- [2] YU W K, RAO N S V, VETTER J S. Experimental analysis of InfiniBand transport services on WAN[C]//Proceedings of 2008 International Conference on Networking, Architecture, and Storage. Piscataway: IEEE Press, 2008: 233-240.
- [3] LU Y, CHEN G, LI B, et al. Multi-path transport for RDMA in datacenters[C]//Proceedings of 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). [S.l.: s.n.], 2018: 357-371.
- [4] ALIZADEH M, EDSALL T, DHARMAPURIKAR S, et al. Conga: distributed congestion-aware load balancing for data-centers[J]. ACM SIGCOMM Computer Communication Review, 2014, 44(4): 503-514.
- [5] RAICIU C, BARRE S, PLUNTKE C, et al. Improving data-center performance and robustness with multipath TCP[C]//Proceedings of the ACM SIGCOMM 2011 Conference. New York: ACM Press, 2011.
- [6] GUO C X, WU H T, DENG Z, et al. RDMA over commodity Ethernet at scale[C]//Proceedings of the 2016 ACM SIGCOMM Conference. New York: ACM Press, 2016: 202-215.
- [7] SHALEV L, AYOUB H, BSHARA N, et al. A cloud-optimized transport protocol for elastic and scalable HPC[J]. IEEE Micro, 2020, 40(6): 67-73.
- [8] BAI W, ABDEEN S S, AGRAWAL A, et al. Empowering azure storage with RDMA[C]//Proceedings of 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23). [S.l.: s.n.], 2023: 49-67.
- [9] 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.
- [10] SHPINER A, ZAHAVI E, DAHLEY O, et al. RoCE rocks without PFC: detailed evaluation[C]//Proceedings of the Workshop on Kernel-Bypass Networks. New York: ACM Press, 2017: 25-30.
- [11] MATHIS M, SEMKE J, MAHDAVI J, et al. The macroscopic behavior of the TCP congestion avoidance algorithm[J]. ACM SIGCOMM Computer Communication Review, 1997, 27(3): 67-82.
- [12] MITTAL R, SHPINER A, PANDA A, et al. Revisiting network support for RDMA[C]//Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. New York: ACM Press, 2018: 313-326.
- [13] 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.
- [14] MITTAL R, LAM V T, DUKKIPATI N, et al. Timely: RTT-based congestion control for the data-center[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 537-550.
- [15] 刘运渠. 2023 算力互联互通大会贡献示范单位“为是科技”展示长途 RDMA[EB]. 2023.

LIU Y Q. Viscore Tech. demonstrates long-haul RDMA in the 2023 Computility Interconnection Conference[EB]. 2023.

- [16] 陆璐. 网络与 AI 融合的思考和实践[R]. 2023.

LU L. Thinking and practicing network and AI convergence[R]. 2023.

- [17] YU Z L, SU B W, BAI W, et al. Understanding the micro-behaviors of hardware offloaded network stacks with lumina[C]//Proceedings of the ACM SIGCOMM 2023 Conference. New York: ACM Press, 2023: 1074-1087.

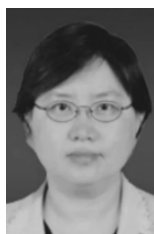
- [18] 李彤. 广域确定性网络中的传输优化[R]. 2023.

LI T. Transmission optimization in wide-area deterministic networks[R]. 2023.

[作者简介]



赵俊峰（1979—），男，中国信息通信研究院高级工程师，主要研究方向为分组传送、4G/5G 移动承载、确定性网络和未来网络等。



李芳（1974—），女，中国信息通信研究院正高级工程师，主要研究方向为高速光通信、分组传送、4G/5G 移动承载和软件定义光网络等。



叶晓峰（1974—），男，江苏为是科技有限公司产品经理，主要研究方向为高性能网络和 RDMA 网卡及智能网卡芯片。



江崧（1983—），男，国网江苏省电力有限公司信息通信分公司高级工程师，主要研究方向为电力光通信网络、语音交换网络等。