doi:10.13756/j.gtxyj.2024.240028.
专题:数据中心内光交换
翟锐,李壮志,侯广营,等.基于以太无损网络的智算中心光网络架构研究[J].光通信研究,2024(5):240028.
Zhai R, Li Z Z, Hou G Y, et al. Research on Optical Network of Intelligent Computing Center based on Ethernet Lossless Networking[J]. Study on Optical Communications, 2024(5):240028.

基于以太无损网络的智算中心光网络架构研究(特邀)

翟 锐,李壮志,侯广营,马艺嘉,徐化朗
(中国联合网络通信有限公司山东省分公司,济南 250002)

摘要:【目的】近年来,生成式人工智能(AIGC)掀起了人工智能革命,智算中心(ICC)的网络联接也随之向超高带宽、智能无损和算网融合等方向发展,因此ICC光网络需要降低卡间通信时间,以提升数据访问效率。【方法】文章针对ICC场景光网络的组网架构进行了研究,实现了大带宽、低时延和中央处理器(CPU)效率高的无损网络,满足了ICC的大模型训练和推理需求。文章详细分析了ICC的流量分布特征和人工智能(AI)大模型训练组网场景下的通信流特征,深入研究了基于远程直接内存访问(RDMA)的以太无损传输方案的ICC组网架构,并最终在ICC场景下进行了组网实践和时延测试。【结果】文章提出的基于以太网的RDMA(RoCE)传输方案具备基于优先级的流控制、显示拥塞通知、增强传输选择和数据中心桥能力交换协议(DCBX)等能力,可实现数据中心内基于以太协议的无损传输。测试结果显示,使用RoCE协议的传输时延大约稳定在1 μs,并且显著优于互联网广域RDMA协议(iWARP)。【结论】文章基于智算场景下的流量特征分析,深入研究了ICC的无损以太网络关键特性,利用RDMA技术实现了ICC场景下光交换网络传输效率的提升,并提出了一种在ICC大模型推理场景下的无损以太网络方案,为RDMA技术在智算场景下的应用探索出了可行的方向。

关键词:长距直接内存访问;以太无损网络;智算中心;光交换
中图分类号:TN929 **文献标志码:**A

Research on Optical Network of Intelligent Computing Center based on Ethernet Lossless Networking

ZHAI Rui, LI Zhuangzhi, HOU Guangying, MA Yijia, XU Hualang
(Shandong Branch of China United Network Communications Co., Ltd., Jinan 250002, China)

Abstract:【Objective】In recent years, Artificial Intelligence Generated Content(AIGC) has set off the artificial intelligence revolution. The network connection of the Intelligent Computing Center(ICC) has also developed in the direction of ultra-high bandwidth, intelligent lossless, and computing network convergence. Therefore, the optical network of the ICC needs to reduce the inter-card communication time in order to improve the efficiency of data access.【Methods】The paper addresses the networking architecture of optical networks for ICC scenarios to realize a lossless network with large bandwidth, low latency and high Central Processor Unit (CPU) efficiency, which can satisfy the demand of large model training and reasoning in ICC. This paper analyzes in detail the traffic distribution characteristics of the ICC and the communication flow characteristics under the AI large model training networking scenario. It also conducts in-depth research on the technologies such as Ethernet lossless network based on Remote Direct Memory Access(RDMA) technology and optoelectronic co-encapsulation. Finally it carries out the networking practice and latency test under the ICC scenario.【Results】The RDMA over Converged Ethernet(RoCE)-based transport scheme proposed in this paper has the capabilities of priority-based flow control, displaying congestion notification, enhanced transport selection and data center bridge capability switching protocols, which can realize lossless transmission based on Ethernet protocols in data centers. The test results in this paper show that the transmission delay using the RoCE protocol is approximately stable at around 1 μs and significantly outperforms the Internet Wide Area RDMA Protocol(iWARP).【Conclusion】In this paper, based on the traffic characterization in the intelligent computing scenario, we have studied the key characteristics of the lossless Ethernet network in the ICC, and used the RDMA technology to realize the enhancement of the transmission efficiency of the optical switching network in the scenario of the ICC. We have also put forward a lossless Ethernet network scheme under the large model inference scenario of the ICC, and explored the feasible direction for the application of the RDMA technology in the intelligent computing scenario. The proposed scheme explores a feasible direction for the application of RDMA technology in the smart computing scenario.

Key words: RDMA; Ethernet lossless network; ICC; optical switching

0 引 言

随着生成式人工智能(Artificial Intelligence Generated Content,AIGC)大模型和互联网应用的高速发展[1],智算中心(Intelligent Computing Center,ICC)承载的数据量也与日俱增。ICC的流量越

收稿日期:2024-02-06; 修回日期:2024-03-21; 纸质出版日期:2024-10-10
作者简介:翟锐(1991—),男,湖北黄冈人。工程师,博士,主要研究方向为算力网络、光网络通信和空天信息等。
通信作者:翟锐,博士,E-mail:zhair@chinaunicom.cn

来越大，复杂性也越来越高，已经成为整个互联网流量和业务的制高点，必然会引发技术层面的创新变革。如今，面向 ICC 的技术优化和创新成了新的热潮。ICC 作为存储、处理和分析数据的重要基础设施，其节点算力规模逐渐扩大，对于内部交换网络也提出了新的要求：更低的网络延迟、更高的吞吐量以及更低的中央处理器（Central Processing Unit，CPU）开销。

远程直接内存访问（Remote Direct Memory Access，RDMA）技术是一种新的内存访问技术[2]，可显著提升数据中心网络的性能和效率。RDMA 技术可以在不占用主机 CPU 的情况下，将数据从一个计算节点的内存直接传输到另一个计算节点的内存。RDMA 技术最初是基于 Manalox 公司提出的无限带宽（InfiniBand，IB）架构[3]，IB 是一个为大规模、易扩展机群而设计的网络通信技术协议，可用于计算机内部或外部的数据互连，服务器与存储系统之间直接或交换互连以及存储系统之间的互连。

ICC 的光网络架构得到了世界范围内大学和研究机构的广泛关注，研究者探索了多种流量调度和控制策略。2017 年，Jin 等人提出了一种基于多平面无源光交叉连接网络（Passive Optical Cross-Connection Networks with Multiple Plane，POXN/MP）的新型数据中心网络架构，通过交替放置电子交换机和 POXN/MPs 或多平面和捆绑端口的（POXN/MP and Bundled Port，POXN/MP-BP）来构建树形拓扑，降低构建总成本和维护大型数据中心网络的功耗[4]；2019 年，北京邮电大学的刘爱军关注到可重构数据中心光网络架构的研究，采用光电路交换（Optical Circuit Switching，OCS）和自由空间光（Free Space Optics，FSO）技术，在发送端通过附加耦合器复用集群内光信号，实现了接收端集群内所有架顶（Top of Rack，ToR）交换机均能与附加的波长选择开关（Wavelength Selective Switch，WSS）进行通信[5]；2023 年，谢里夫理工大学的 Rezaei 等人提出了一种动态全光网络体系结构，使用无源元件根据流量负载提供可变带宽，实现了低能耗、低重组延时，消除了电子数据包的缓冲[6]。以上研究均是针对 ICC 光网络的功耗进行优化，而本文针对 ICC 的流量特点，研究了二层光网络架构和物理层流量调度和拥塞控制算法，降低了传输时延，实现了以太无损传输。

1　ICC 无损传输关键技术

ICC 的流量分布具有不均匀性和突发性。现有的光网络架构中，大量的网络流量往往承载于少量节点，导致了大量网络资源的闲置和浪费。另一方面，ICC 中频繁发生小流量的动态爆发，造成对交换机队列长度的冲击和端到端延迟的增长，严重影响了 ICC 网络拥塞控制策略的正常运行。本章深入研究了 ICC 的网络流量特征、无损网络和光模块的关键技术，为下一节提出 ICC 场景下基于以太网的 RDMA（RDMA over Converged Ethernet，RoCE）的光网络传输方案奠定了基础。

1.1　ICC 的网络流量特征

在传统的数据中心场景中，大部分服务器访问需求集中在行业用户的数据存储需求，主要流量是南北向流量，也就是从用户端到数据中心服务器之间的流量。然而随着人工智能（Artificial Intelligence，AI）技术的发展，新型 ICC 的主要流量变为了东西向流量，也就是 ICC 内部服务器之间的流量，南北向和东西向流量呈现“二八定律”，图 1 所示为典型 ICC 的架构和流量流向。在传统的 3 层组网架构下，网络规模和性能受限于汇聚交换机和核心交换机。为了适应大规模分布式的新型 ICC 流量特征，本文针对其组网架构进行研究。

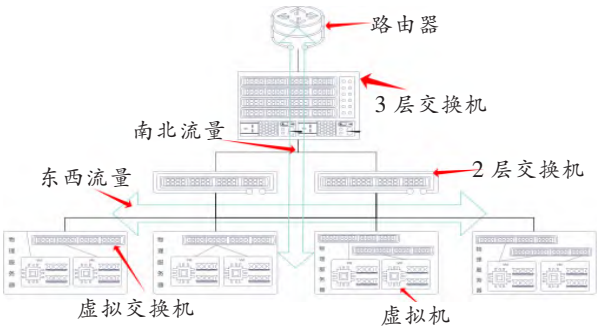


图 1　典型 ICC 架构及流量流向

Figure 1　Typical ICC architecture and traffic Flow

ICC 的业务流量同时具有大象流和老鼠流的特征[7]，且老鼠流的数量占 90％以上，而大象流的数量虽然较少，却承载了网络中 90％以上的数据量。大象流一般指大量且持续的数据传输过程，同时存在于东西向和南北向流量，但主要以东西向流量为主。老鼠流是指少量且短时间的数据传输过程，主要存在于南北向流量。

图 2 所示为 AI 大模型训练组网的示意图，每 8

台服务器一个 Stage，16 个 Stage 对应 128 台服务器，承接训练模型需求。如图所示，在红色标识的流量路径中（服务器→叶子交换机（Leaf）→服务器），同时存在数据并行组（Data Parallism，DP）和流水线并行组（Pipeline Parallism，PP）两种并行模式，DP 流量在 stage 内，不需要跨 Leaf 传输，而 PP 流量需要要跨 stage 计算。图中灰色标识是张量并行组（Tensor Parallism，TP）流量，主要集中于端侧内部。AI 大模型的训练流量特点如下：① 周期性特征，每次训练迭代的通信模式完全一致。② 通信数据流相对较少，包含大小流特征，主要流量为大流。③ 每轮通信总数据量大，以机内 TP 并行的全规约（All Reduce）为主，机间通信大部分流量单跳完成，不经过骨干交换机（Spine）。④ 经过 Spine 的流量特征为多打多的并发中小流（6 Mbit/s），当负载均衡效果不佳时，将因拥塞机制而导致带宽利用率下降。⑤ 机间通信数据量与集群图形处理器（Graphics Processing Unit，GPU）规模呈正相关关系。

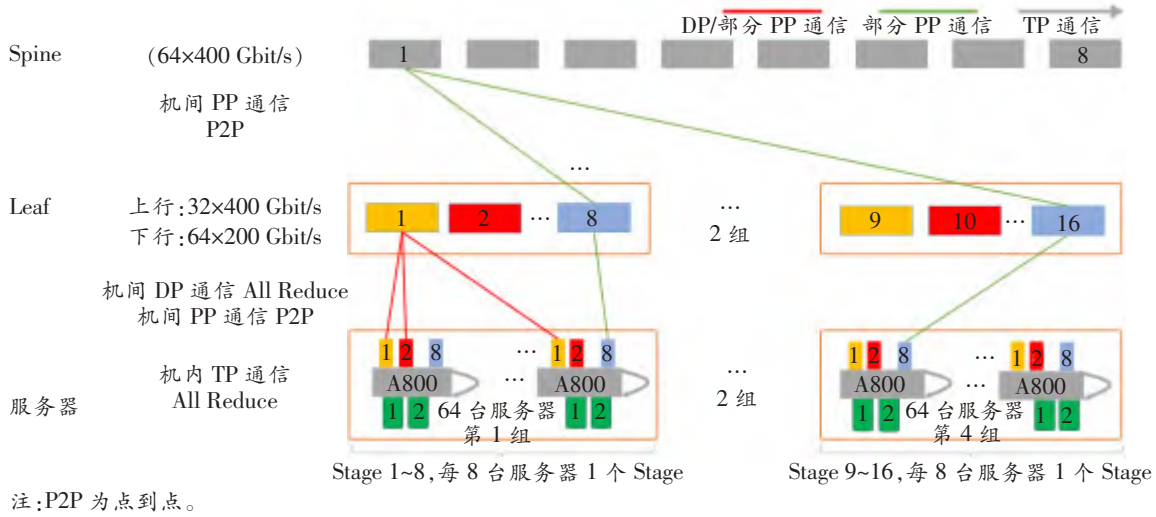


图 2　AI 大模型训练流量组网和通信流

Figure 2　AI large model training traffic networking and communication flows

1.2　无损网络

RDMA 技术的高性能和低延迟依赖于智能网卡和特殊软件架构的支持，通过在智能网卡上固化 RDMA 协议，减少 CPU 的干预，避免数据包的传输和协议处理带来的额外开销，从而提高数据传输的效率和速度[8]。如图 3 所示，传统的基于 socket 传输控制协议/网际协议（Tvamsmission Control Protocol/Internet Protocol，TCP/IP）协议栈的网络通信过程繁琐，通信数据需要在系统内存和各级缓存中多次复制转发，使得 CPU 计算资源和内存总线带宽消耗巨大，也增加了网络延时。RDMA 可以绕过操作系统内的多次内存拷贝[9]，远程节点的 CPU 无需介入，数据直达对端应用缓冲器。

1.3　光模块演进

基于 ICC 中 AI 高算力对数据传输高速率和低功耗的巨大需求，需要匹配更先进的光通信底层算力基础设施，传统可插拔光模块的功耗问题日益明显[10]。传统光/电互连应用板边光模块的方式，以可插拔光模块或有源光缆的形式组装在印刷电路板（Printed Circuit Board，PCB）的边缘，但这种方式

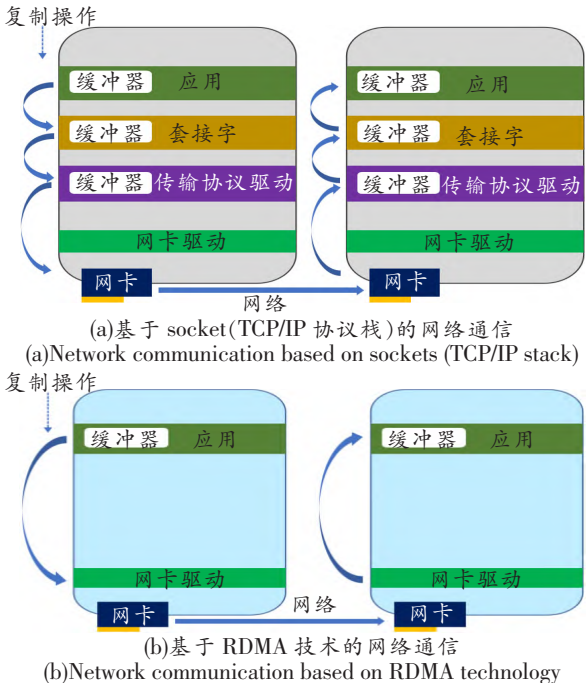


图 3　RDMA 技术对比传统的 TCP/IP 协议栈的资源消耗

Figure 3　Resource consumption of RDMA technology versus traditional TCP/IP stack

存在走线长、模块体积难控制、功耗大和信号完整性得不到保障等缺陷，无法满足大带宽、高传输速率和日益剧增的算力传输需求。为了满足急剧增长的数据量需求，光/电互连从传统板边光模块不断地向着集成度更高、体积更小的方向发展。

光电共封装（Co-Packaged Optics，CPO）是一种新型的光电子集成技术[11]，是光子技术的一个前沿发展方向，即利用硅光的低成本和高速特性，将光收发模块和控制运算的专用集成电路芯片异构集成。CPO 方案降低了交换芯片和光运算单元之间的走线距离，能够满足 AI 算力在数据传输环节的性能瓶颈[12]。

2　基于 RoCE 的传输方案研究

ICC 场景中爆炸式的流量增长导致出现数据阻塞问题越来越严重，本文一方面在网络架构上进行优化，缩短传输路径，另一方面利用 RoCE 协议实现拥塞控制，减少丢包，降低时延，提高 ICC 网络的健壮性，实现无损传输。图 4 所示为 ICC 内的 RoCE 网络架构，采用叶脊（Spine-Leaf）2 层架构，由 Spine 和 Leaf 组成。每台 Spine 和 Leaf 连接，2 层架构相比传统的 3 层架构，简化了东西向流量的传输路径，提高了传输效率。RoCE 可以通过标准化的以太网设备实现基于以太网的数据传输，而不需要使用特殊的网络硬件设备，降低了网络成本。

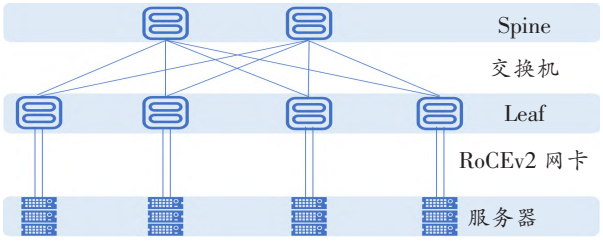


图 4　RoCE 网络架构

Figure 4　RoCE network architecture

本文主要通过基于优先级的流控制（Priority-based Flow Control，PFC）、显式拥塞通知（Explicit Congestion Notification，ECN）、增强传输选择（Enhanced Transmission Selection，ETS）和数据中心桥能力交换（Data Center Bridging Exchange，DCBX）协议等技术对传统以太网进行改造，打造无损以太网络。

PFC 策略能够对网络数据流进行分类和优先级控制，可以基于每个以太网优先级使能暂停（Pause）功能，并发送 Pause 信号抑制上游发送数据。ECN 首先由传输层进行能力协商，协商完毕后发生拥塞的设备对 IP 报文 ECN 功能传输位（ECN Capable Transport，ECT）/经历拥堵位（Congestion Experivnced，CE）标志位置位，接收端接收到 CE 包，向发送端发送拥塞通知（Congestion Notification Packet，CNP），通知发送端降速。在 RoCE 环境中推荐同时使能 PFC 与 ECN，以保证 RoCE 报文无丢包情况下带宽得到保证，PFC 和 ENC 的对比如表 1 所示。为了充分发挥网络高性能转发，一般通过调整 ECN 和 PFC 的 buffer 水线，让 ECN 快于 PFC 触发，即网络还是持续全速进行数据转发，让服务器主动降低发包速率。如果还不能解决问题，再通过 PFC 让上游交换机暂停报文发送。

表 1　PFC 和 ECN 对比

Table 1　Comparison of PFC and ECN

	PFC	ECN
网络位置	3 层	网络层及传输层
作用范围	P2P	端到端
需要全网支持	是	否
被控制对象	网络中上一个节点	发送主机
报文缓存位置	中间网络节点及发送端	发送端
受影响的流量	网络设备中 8 个转发队列中某个队列的所有流量	发生拥塞应用的连接
响应速度	快	慢

注：如果网卡支持 PFC，PFC 对网卡也能生效。

ETS 根据服务类型按照优先级将流量区分成不同的优先级组（Priority Group），并针对不同分组提供最小带宽保障，保证重要流量在传输过程中具有承诺带宽。链路总带宽可以动态地被各优先级组调用，各优先级组的承诺带宽如果出现闲置，其他优先级分组可以动态调用闲置带宽，提高链路利用率。如图 5 所示，链路总带宽为 10 Gbit/s，在 T_1 时间段，应用 A、B、C 的流量带宽均为 3 Gbit/s，接口总流量不超过接口带宽，所有流量都能转发；在 T_2 时间段，应用 C 增长到 4 Gbit/s，但是接口总流量仍没有超过接口带宽，所有流量仍然都能转发；而在 T_3 时间段，总流量超过接口带宽，且应用 C 流量超过给定的带宽，按照ETS的参数进行调度，应用C流

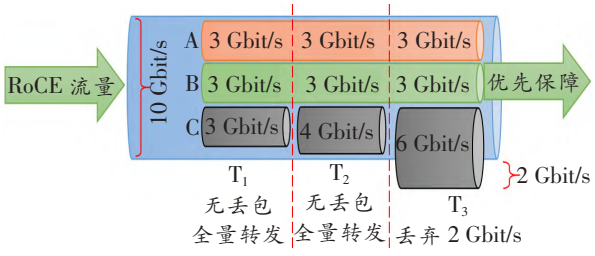


图 5　ETS 带宽控制示意图

Figure 5　Bandwidth control schematic of ETS

量被丢弃 2 Gbit/s。

在 ICC 场景下，为降低网络丢包，链路两端的 PFC 和 ETS 参数需要配置相同；DCBX 链路发现协议可以实现数据中心桥（Data Center Bridging，DCB）能力配置信息都自动协商，减轻管理员的参数配置工作量，并减少人工配置过程中容易出现的配置错误。图 6 所示为 DCBX 的工作流程：

① 交换机 A 和服务器网卡分别本地配置 PFC 参数，打开 DCBX 功能。由 DCBX 模块通知两端封装 PFC 参数，以链路层发现协议（Link Layer Discovery Protocol，LLDP）报文进行发送。

② 交换机 A 定期发送 LLDP 报文，通知服务器自己的 PFC 参数配置。

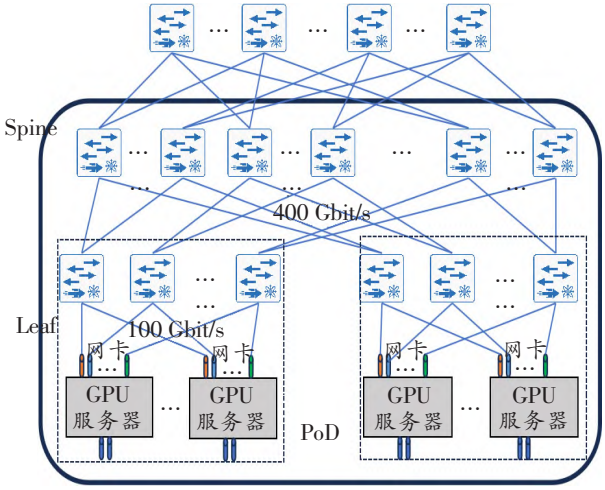
③ 服务器根据接收到的 LLDP 报文得到交换机 A 的 PFC 参数，DCBX 模块比较两段的 PFC 参数，根据自动协商算法生成 DCBX 配置文件，保证两端 PFC 参数相同。

图 6　DCBX 自动协商参数配置流程

Figure 6　DCBX auto-negotiation parameter configuration process

3　ICC 的无损以太网络实践

近年来，伴随着大模型智能水平的提升，AIGC 所需算力不断增长，ICC 作为算力输出底层载体面临着更多的挑战。传统的数据中心 3 层网络架构包括接入层、汇聚层和核心层，这种架构汇聚层和接入层通常直接采用生成树协议（Spanning Tree Protocol，STP），接入交换机的上联链路中只有一条承载流量，会造成带宽的大量浪费，同时 STP 在网络拓扑改变时需重新收敛，因此故障域更大，不适合超大规模的网络。因此本文针对 400 Gbit/s AIGC 场景下的 ICC 无损网络提出了一种可行解决方案，图 7 所示为在 AIGC 推理场景下 RoCE 无损网络的架构。本方案采用 Spine-Leaf 网络架构，接入位置（Leaf）使用具备 48 个 100 Gbit/s 端口＋8 个 400 Gbit/s端口的电交换机，核心（Spine）位置使用 32 个 400 Gbit/s 端口的交换机，网络协议使用 RoCE 协议，实现 GPU 服务器之间大量训推数据的大带宽低时延转发。



注：PoD 为集群中最小可布署管理的基本单元。

图 7　AIGC 推理场景下的 RoCE 无损网络架构

Figure 7　RoCE Lossless Network Architecture for AIGC Reasoning Scenarios

基于上述无损网络架构，我们在 ICC 25、40 和 100 Gbit/s 端口场景下，分别测试了采用 RoCE 协议和互联网广域 RDMA 协议（Internet Wide Area RDMA Protocol，iWARP）传送不同长度数据包时的时延情况，如图 8 所示，在 3 种不同端口能力下，使用 iWARP 和 RoCE 协议的传输时延都会随着数据包长度的增加而变大；如图所示，在数据包＜128 Byte的情况下，使用 iWARP 的传输时延大约稳定在 2.4 μs 左右，而使用 RoCE 协议的传输时延大约稳定在 1 μs左右。因此我们认为在ICC训

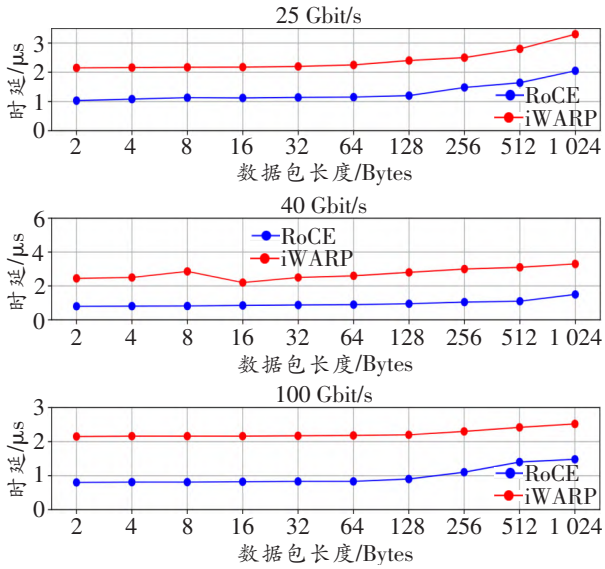


图 8　不同场景下 RoCE 协议和 iWARP 的时延对比

Figure 8　Comparison of delay between RoCE protocol and iWARP in different scenarios

推场景下，使用RoCE协议的传输时延都要显著低于iWARP的传输时延。ICC场景下，传输时延的降低可以有效缓解网络阻塞，进而减少阻塞丢包，实现无损传输。

4　结束语

本文首先详细分析了ICC东西向流量和大象流占比更高的特性，然后研究了智算场景中的无损以太网络架构和基于RoCE的传输方案。最后，本文针对ICC的叶脊2层网络架构进行了RoCE传输方案的实践，实验结果表明，RoCE传输协议的流控制算法和拥塞控制算法有效缓解了网络阻塞，降低了服务器之间的传输时延。下一步，本文所提基于RoCE的ICC无损以太网络架构，可以叠加基于IPv6转发平面的段路由（Segment Routing IPv6，SRv6）和软件定义网络（Software Defined Network，SDN）等技术，构建高带宽、低延迟、智能无损和弹性的ICC网络底座。伴随着高性能计算和通信密集型并行协同计算等需求的不断增加，基于RoCE的无损网络架构在ICC的场景中将获得更为广阔的应用前景。

参考文献：

[1] Bulloch G, Seth I, Lee C H A. ChatGPT in Surgical Research and Practice: a Threat to Academic Integrity, Authorship, and Divergent Thinking[J]. ANZ Journal of Surgery, 2023, 93(9):2270－2271.

[2] RFC 4297-2005, Remote Direct Memory Access (RDMA) over IP Problem Statement[S].

[3] Buyya R, Cortes T, Jin H. High Performance Mass Storage and Parallel I/O: Technologies and Applications[M]. USA: Wiley-ieee Press, 2002.

[4] Jin P, Huang C C. Sandwich Tree: a New Datacenter Network based on Passive Optical Devices[J]. Optical Switching and Networking, 2017, 25:133－148.

[5] 刘爱军. 面向数据中心的光网络架构及资源优化机制研究[D]. 北京：北京邮电大学，2019.
Liu A J. Research on Optical Network Architectures and Resource Optimization Mechanisms for Data Centers[D]. Beijing, China: Beijing University of Posts and Telecommunications, 2019.

[6] Rezaei N, Koohi S. Flat Ball: Dynamic Topology for Energy Management of Optical Interconnection Networks in Data Centers[J]. Optical Switching and Networking, 2023, 48:100730.

[7] 黄宗伟. 基于SDN的数据中心网络大象流负载均衡调度策略研究[J]. 重庆科技学院学报(自然科学版)，2023，25(2)：76－81.
Huang Z W. Research on Scheduling Strategy of Elephant Flow Load Balancing in Data Center Network based on SDN[J]. Journal of Chongqing University of Science and Technology (Natural Sciences Edition), 2023, 25(2):76－81.

[8] 刘通. 支持RDMA的高速网络对大数据与云计算平台效率的影响[J]. 电信工程技术与标准化，2015，28(2)：74－77.
Liu T. Research on Big Data and Cloud Efficiency with High Performance Interconnect[J]. Telecom Engineering Technics and Standardization, 2015, 28(2): 74－77.

[9] 赵精华，郭亮. 智能无损网络：数据中心网络性能优化策略[J]. 中国电信业，2021(S01)：67－72.
Zhao J H, Guo L. Intelligent Lossless Network: Optimization Strategy of Data Center Network Performance [J]. China Telecommunications Trade, 2021(S01): 67－72.

[10] 熊青松，张武平，陈晋敏. 用于以太网的40 Gbit/s CFP光模块设计[J]. 光通信研究，2012(4)：49－51.
Xiong Q S, Zhang W P, Chen J M. Design of 40 Gbit/s CFP Optical Modules for Ethernet[J]. Study on Optical Communications, 2012(4):49－51.

[11] 郭宝增. 硅基光电器件研究进展[J]. 半导体技术，1999，24(1)：19－24.
Guo B Z. Research Development of Silicon-based Optoelectronic Devices[J]. Semiconductor Technology, 1999, 24(1):19－24.

[12] 卞玲艳，曾艳萍，蔡莹，等. 大数据时代光电共封技术的机遇与挑战[J]. 激光与光电子学进展，2024，61(9)：0900006.
Bian L Y, Zeng Y P, Cai Y, et al. Opportunities and Challenges of Optoelectronic Co-packaging Technology in the Era of Big Data[J]. Laser and Optoelectronics Progress, 2024, 61(11):0900006.