特色专题

广域长距离高性能传输技术研究与讨论

梁腾[1],杨健[1,2],杨佳宇[1],张宇[1,3],张伟哲[1,2,3*]

摘要 广域长距离高性能传输技术在中国“东数西算”工程构建全国一体化算力网背景下具备重要的战略价值。3个趋势对广域分布式算力协同范式提出新需求:对算力资源要求极高的人工智能(AI)大模型智能应用的兴起;高端高性能图形处理单元(GPU)芯片被禁运限制单中心算力资源;中国各地建设的算力集群形成算力分散分布态势。广域长距离高性能传输技术是上述新范式的关键技术。从支撑广域分布式算力协同新范式、技术路线、承载网络、研究难点、成本5个方面进行讨论,结合深圳到宁夏中卫2100 km实网实验结果,将现有远程直接内存访问(remote direct memory access, RDMA)技术基于广域全光网进行长距离优化的方案是短期内可行性高、成本低且利于开展研究的最佳方案之一,通过优化基于融合以太网的远程直接内存访问(RDMA over Converged Ethernet, RoCE)可以在广域全光网上实现“广域光数直达”逼近物理层通信性能指标。

关键词 广域长距离高性能传输;广域远程内存直接访问(WRDMA);算力网络;东数西算

广域长距离高性能传输技术在中国“东数西算”工程背景下具备重要的战略价值。根据2023年“东数西算”工程实施意见[1],加快构建全国一体化算力网,充分发挥国家枢纽节点引领带动作用,积极推进低时延、高带宽、低抖动的新兴网络技术在“东数西算”工程中应用,打通国家枢纽节点与非国家枢纽节点间网络主干道,提升网络传输性能。国家8大枢纽节点包括京津冀、长三角、粤港澳、贵州、成渝、甘肃、宁夏、内蒙古,节点间物理距离超过1000 km。

广域长距离高性能传输在本研究中指的是2台终端设备在广域长距离尺度上进行端到端高性能网络传输。主要瓶颈点分为网络侧与端侧2部分,前者根据中国联通运营商提供的信息,从终端设备角度,目前单端口常见的商用骨干光通信速度可达到100 Gbps;对于后者而言,常见的商用高性能网卡也能够达到100 Gbps。然而,实网实验表明,现有的商用网卡主要应用于数据中心内部,在长距离线路上性能表现不稳定,如何实现跨域长距离端到端高性能网络传输,以逼近底层通信性能指标,是本研究讨论的重点。

1 支撑广域分布式算力协同新范式

首先,从应用需求进行分析。2022年11月30日,AI应用ChatGPT3.5的问世掀起AI大模型应用热潮,导致相关算力资源的价值和需求量极大提升。以语言大模型GPT3.5[2]为例,GPT3.5拥有1750亿参数,如果用1024张英伟达A100型号GPU进行一次训练,估计需要几周时间。其公司OpenAI[3]后续发布

1. 鹏城实验室,深圳 518000
2. 哈尔滨工业大学(深圳),深圳 518055
3. 哈尔滨工业大学,哈尔滨 150006

收稿日期:2024-08-21;修回日期:2025-04-01
基金项目:新一代人工智能国家科技重大专项(2022ZD0115303);鹏城国家实验室重大攻关任务(PCL2023A06)
作者简介:梁腾,特聘副研究员,研究方向为新型网络架构,电子信箱:liangt@pcl.ac.cn;张伟哲(通信作者),教授,研究方向为并行计算、分布式计算、云计算以及计算机网络,电子信箱:wzzhang@hit.edu.cn
引用格式:梁腾, 杨健, 杨佳宇, 等. 广域长距离高性能传输技术研究与讨论[J]. 科技导报, 2025, 43(9): 31-37; doi:10.3981/j.issn.1000-7857.2024.08.01032

能力更强的 GPT4,其模型参数被外部估计是 GPT3.5 参数的数倍至数十倍。一个发展趋势是发展更大的模型,意味着需要更多的算力。

其次,从现状进行评估。2023 年 10 月 17 日,美国商务部工业和安全局(BIS)公布新的先进计算芯片出口管制规则,限制中国购买高端芯片,根据最新的规则[4],英伟达包括 A100、A800、H100、H800、L40S 及 RTX 4090 在内的芯片对中国出口都将受到影响。如果没有同等算力能力替代出现,该禁运令在短期内将限制中国单集群的算力上限能力。

同时,中国各地建设的算力集群形成算力分散分布态势。根据 2023 年“东数西算”工程实施意见,国家 8 大枢纽节点地理位置分散,枢纽间物理距离尺度大。除此之外,各地也在建设智算中心。

基于上述现状与趋势,探索广域分布式算力协同新范式是高效利用现有算力资源的方式之一。具体应用方面,联邦学习[5]属于典型代表,大模型广域协同训练尚存在争议,其中网络传输是瓶颈,因此,广域长距离高性能传输技术是支撑该新范式的关键技术。同时广域长距离高性能传输技术也将促进新兴应用发展。

2 广域长距离高性能传输技术路线

针对广域长距离高性能传输的需求,简要比较 TCP/IP 传输,以及 RDMA 传输技术路线,并讨论选择哪种 RDMA 传输技术路线作为重要参考,分析其可行性与合理性。

TCP/IP 是互联网的核心传输协议栈[6]。在互联网兴起的年代,TCP/IP 传输协议的设计目的是提供异构设备与网络的互联性和在海量设备与应用互联下兼并可扩展性与可靠性的互联网传输基础协议。因此,TCP 是“尽量而为”与“公平竞争”互联网上的通用传输协议,具有良好的应用生态。同时也意味着 TCP 并非为了高性能而设计,虽然存在大量针对该方向的研究,但通用性并非本文中广域长距离高性能传输的重要目标。并且追求通用性导致 TCP 的复杂的拥塞与传输控制机制不利于高性能实现,这点在 iWARP(RDMA over TCP)传输技术中也有所印证。

远程直接内存访问(remote direct memory access, RDMA)传输技术[7−8]是一种允许网络中的计算机在不涉及操作系统、CPU 或其他中介处理的情况下直接从一台机器的内存到另一台机器的内存交换数据的传输技术。这种数据传输方式显著降低了延迟和中央处理器(central processing unit, CPU)开销是一种高性能传输协议,因此广泛应用于高性能计算(high performance computing, HPC)、大型数据中心和存储网络,同时也是大模型训练的重要传输协议。

RDMA 目前有 3 种实现方式,无限带宽(InfiniBand, IB)[9−10]、基于融合以太网的 RDMA(RDMA over Converged Ethernet, RoCE)[11−12]、互联网广域 RDMA 协议(Internet Wide Area RDMA Protocol, iWARP)[13−14]。IB 需要专用 IB 交换机、网卡和线路,因此,IB 网络的部署与成本较高。RoCE 将 RDMA 引入以太网,并且可以穿越 IP 路由器(RoCEv2),不但成本降低也提供了更优的灵活性。iWARP 是一个旨在将 RDMA 功能扩展到标准 IP 网络的协议,通过在 TCP/IP 协议栈上实现 RDMA 操作,在国际互联网工程任务组 IETF 发布核心标准 RFC 5040[15]与扩展标准 RFC 5041[16]。iWARP 的设计初衷是将 RDMA 与 TCP 打通,将 RDMA 的高性能特性扩展到广域互联网上,然而 iWARP 的市场份额远低于 IB 和 RoCE,因为 iWARP 的性能不如 RoCE 和 IB。iWARP 并未达到预期成功的经验值得反思,将 RDMA 通过基于 TCP 的方式实现广域化是否是一条具有前景的道路,在这种情况下的高性能与通用性是否存在内生的冲突?

RDMA 作为一种成熟的高性能传输技术,目前已经广泛应用于 HPC 和 AI 训练任务中。将其作为重要参考来设计广域长距离高性能传输技术,一方面,可以吸取成熟的经验,另一方面,可以与数据中心内部网络传输技术实现兼容。具体实现方式上,IB 由于其昂贵的成本与专用硬件的封闭特性不适合应用在广域长距离上。虽然 iWARP 的设计初衷之一是为了实现广域网高性能传输,但将复杂通用的 TCP 机制在硬件网卡中实现追求高性能本身存在极大挑战,同时互联网基于因特网协议(Internet Protocol, IP)网络体系架构的“尽量而为”与“公平竞争”特性与高性能本身存在冲突。比较而言,将 RoCE 应用于广域网这一想法成为重要的思考方向。应用的前提是广域网需要具备融合以太网(converged ethernet)特性,下文将

讨论广域长距离高性能传输承载网络选择以支持RoCE。

国内外相关研究方面,2008 年美国橡树岭国家实验室在 13840 km 的 10 Gbps 全光网(OC192)广域网上对 IB 传输服务进行实验分析,研究表明 IB 传输服务中不可靠连接(unreliable connection)和不可靠数据报(unreliable datagram)传输服务在广域网上提供更好的高带宽操作性能,并且针对广域网的特性定制消息传递接口(message passing interface, MPI)实现,可以显著提高性能[17]。2023 年微软在 NSDI(USENIX Network System Design and Implementation)会议上发表其用 RDMA 构建 Azure 云存储服务的工作,文中提及截至 2023 年 2 月,Azure 云中 70% 的网络流量均来自 RDMA,其中城域云中心间长链路可达数十千米[18]。

3 广域长距离高性能传输承载网络

广域长距离高性能传输方案中承载网络的选择包括 IP 承载网[19]和全光网[20−22]。二者逻辑上是上下层关系(图 1),存在互补性。IP 承载网主要关注数据包的处理和路由,支持各种 IP 服务和应用。全光网络专注于提供高速、低延迟的光纤传输,支持大带宽的数据传输需求。

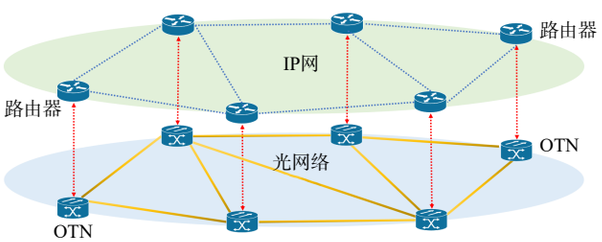

图 1 广域长距离承载网络选择: IP 网与全光网

具体而言,IP 承载网是互联网的基础设施,指用于传输和路由 IP 流量的网络基础设施,通常包括 IP 核心路由器、交换机和相关的网络设备。它承载了各种 IP 服务和应用,如互联网、企业网络和数据中心服务。全光网络是指使用光纤传输技术来提供网络连接的网络架构。全光网络中的信号在光纤中以光波的形式传输,不需要在传输过程中进行电到光的转换(即从光信号转换为电信号,反之亦然),从而实现了高带宽和低延迟的数据传输。

相较于 IP 网,全光网更适合作为 RoCE 广域化的承载网络底座。为了减少 IP 设备增加的队列及相关的资源竞争与管理,全光网可以提供高带宽、低时迟和确定性抖动的网络服务,该条件对实现高性能端侧传输控制提供良好的网络侧条件,降低了高性能实现难度。相比于 IP 网络,全光网的成本更高、灵活性更差,但由于本文中广域长距离高性能传输具备较强的计划性,目前不需要一个通用的解决方案,因此选择全光网,相比 IP 网络降低了实现高性能传输的挑战。

Jin 等[23]利用软件定义网络(software defined network, SDN)和全光网设备来优化广域网上的大规模数据传输效率,该论文证明全光网在广域网传输上的优势,并通过 SDN 集中式联合控制网络层与物理层,动态重构网络拓扑,提升全光网配置效率和灵活性。该工作也证明全光网在广域数据传输上的优势。

4 广域长距离高性能传输研究难点

选取全光网作为广域承载网络可以规避 IP 网络的不确定性带来的挑战,将数据中心内部 2 张 RDMA 网卡通信进行单纯物理距离拉长。在这种情况下,该研究仍然面临至少 2 个可预见的难点,传输控制与高性能硬件实现。

首先,广域长距离高性能传输管道符合长肥网络定义。根据 RFC 1072[24]中的定义,如果一个网络的带宽时延乘积显著大于 10^5 bit,该网络被认为是长肥网络,带宽时延乘积如式(1)

$$BDP = B \times RTT \tag{1}$$

式中,BDP 为带宽时延乘积,B 为带宽(Bandwidth),RTT 为来回通信时延。

以 100 Gbps 通信管道为例,当通道来回通信延迟为 1 ms 时,该通道就属于长肥网络

$$BDP = 10^8\ \text{bit/s} \times 10^{-3}\ \text{s} = 10^5\ \text{bit} \tag{2}$$

来回通信时延 RTT 为 1 ms 的全光网通道有多长呢?答案是 100 km。由于光在光纤中的速度是 200000 km/s[25],单边通信的时间为 0.0005 s,来回通信延迟 RTT=0.001 s。

$$T_{\text{单边}} = \frac{100\ \text{km}}{200000\ \text{km/s}} = 0.0005\ \text{s} \tag{3}$$

$$RTT = T_{\text{单边}} \times 2 = 0.001\ \text{s} \tag{4}$$

因此,对于 100 Gbps 全光网通道而言,当距离大于 100 km 时,该通道即为长肥网络。在长肥网络中实现高性能传输,对传输控制算法、所需资源与资源管理都是不小挑战。

另外,长距离导致高 RTT,会进一步影响端处拥塞控制算法的收敛速率,使得带宽利用值相比数据中心传输场景变低。具体来说,假设初始速率为R_0,带宽上限制为R_t,包的往返时延为RTT,速率调整的增量因子为α,减量因子为β。在没有拥塞时,收敛时间T计算如式(5)

$$T \propto RTT * \frac{R_t - R_0}{\alpha} \tag{5}$$

相比数据中心网络场景,RTT 在长距离传输下会增长上百倍,则其收敛所需的时间也变得更长。此外,当有拥塞出现时,假设增量因子和减量因子的综合作用下调整的速率为$\Delta R = f(\alpha, \beta)$,其中$f(\cdot,\cdot)$为函数。则有

$$T \propto RTT * \frac{R_t - R_0}{\Delta R} \tag{6}$$

可以看出,在长距传输场景下,拥塞控制算法的调整周期也会增长。当网络因拥塞出现丢包时,则接收端处需要的缓存空间也更多。

此外,高性能硬件实现也是广域长距离传输研究的难点。支持至少 100 Gbps 的端到端高性能传输实验环境是研究基础。单纯的软件仿真验证难以证明算法有效性。反之,高性能硬件实现要求对于算法的复杂度又增加新的约束条件,因此广域长距离高性能传输研究是一项系统性工程。

5 广域长距离数据传输成本

本研究讨论 2 种大数据搬运方式,一种是通过交通工具对硬盘存储进行运输,一种是通过光纤进行网络传输。

由于硬盘存储体积小容量大,其规模效益明显,运输的速度和成本取决于物流体系能力,目前是常见的搬运方式。其缺点同样明显,交通工具的速度远低于光速,因此延迟高,无法满足对低延迟有要求的应用服务。

通过光纤传输在速度上具备极大优势,然而成本极为高昂,包括建设维护成本与带宽资源成本。目前租用一条 100 Gbps 的跨省光纤一个月需要花费数十万元。

虽然建设成本高昂,但已经建设的全光网络作为固定资产保留下来,在足够使用的前提下并不需要大规模重复建设。并且根据中国电信与中国联通运营商反馈,当网络流量负载达到当前上限一定比例时,运营商就会对网络进行扩容,因此当前网络存在一定冗余容量,在一定范围内足以支撑广域长距离数据传输需求。

此外,昂贵的光纤带宽租赁费用,有可能通过技术手段降低成本。由于全光网在使用上不够灵活,租赁时间单元目前是按月进行销售计算,导致对使用时间并不充分的用户仍需要付月租。如果通过技术手段实现按需使用、即插即用、用完即释放可以降低光纤租赁的成本。

为了详细对比通过交通工具对硬盘存储进行运输和通过光纤进行网络传输 2 种方案的成本,假设有 2 个数据中心分别位于深圳和北京,且目前需要将深圳数据中心的 500 TB 数据传输到北京数据中心。

以华为云为例[26],目前华为云的数据快递服务提供了 2 种方式来传递数据,分别是:Teleport 和硬盘,其区别主要是用于存储数据的存储介质不同。Teleport 是华为云针对数据加密标准(data encryption standard, DES)服务设计的专用存储介质,其在安全性(防尘防水、抗震抗压、安全锁等)方面要优于普通硬盘,并且支持 NFS/CIFS/FTP 等协议的数据源导入光突发交换(optical burst switching, OBS)。目前支持数据快递服务的区域有中国内地—中国香港、亚太—曼谷、非洲—约翰内斯堡、拉美—圣保罗、华北—北京。以从深圳到北京为例,采用 Teleport 方式传递数据的收费标准是 500 USD/(60 TB·d)(不足 60 TB 的部分按照 60 TB 计算)。因此,采用数据快递的方式完成传输任务的成本为

$$500 \div 60 \times 500 = 4500\ \text{USD} \tag{7}$$

以电信的天翼云 SD–WAN 为例[27],方案中包括 SD–WAN 智能网关和 SD–WAN 网络带宽。SD–WAN 智能网关的作用是将本地网络连接至互联网及天翼云,并且用户可以通过智能网关来管理传输网络。选取 100 G 网络带宽,那么 SD–WAN 方案的成本为

$$45 \times 100 \times 1000 + 510 = 4500061\ \text{CNY} \tag{8}$$

6 实网测量及结果

为了测试现有网卡在广域全光网表现性能,鹏城实验室联合中国联通,在深圳—宁夏中卫 2100 km 的 100 Gbps 全光网线路上对英伟达 5 代 100 Gbps 智能网卡(ConnectX–5)[28]进行了实验。利用 RDMA 常用测量工具 perftest[29],每次实验传输大小相同的文件。对 RDMA 的读操作 Read(图 2)和写操作 Write(图 3)进行多次测试,并抽样选取其中 40 次结果进行展示。测试结果表明现有 RDMA 网卡在长距离线路上性能不稳定,无法充分利用底层通信性能。

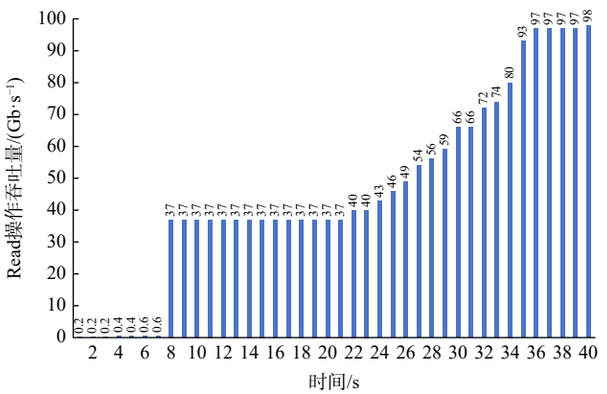

图 2 Read 操作结果

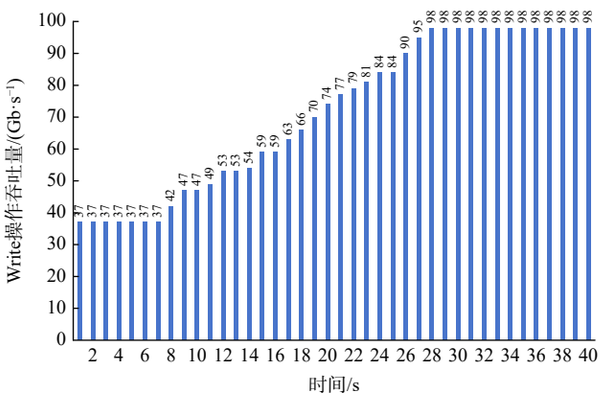

图 3 Write 操作结果

由于现有网卡是黑盒,无法确定具体导致传输性能不稳定的原因。一个推测是现有网卡设计的场景均为数据中心内部,网络时延远远低于跨数据中心的传输时延,因此,其传输与拥塞控制机制不适用于广域长距离线路也属于合理推测范围。以上实网测量结果发现现有商业网卡存在的缺陷,同时也驱动广域长距离高性能传输研究。

为了适应光传输的高带宽、低延迟和长距离特点,将 RDMA 应用在光传输网络上时,需要对协议和算法进行一些关键的修改,有如下 3 个方面。首先,在光传输网络中,由于传输距离增长,传输延迟会远远超出数据中心内传输的微秒级别,而高达几十毫秒。此时,RDMA 系统需要调整其时钟范围以支持长距传输。其次,光传输网络需要长传输距离和更高的往返时间,然而传统的 RDMA 拥塞控制算法(如 DCQCN)通常针对数据中心的短距离高带宽环境,无法很好地适应长距离光传输的特点。为此,需要对拥塞控制算法进行调整,使其能在高延迟环境下有效工作。最后,为了在长距场景下提供超高带宽,可能需要对 RDMA 的协议再进行精简。因为一般来说长距传输为专线传输,因此,可以移除冗余的以太网头部标识,从而支持更大的有效载荷。

广域长距离高性能传输技术在中美竞争和中国国情现状背景下具备重要战略意义,通过现有工作调研,将现有 RDMA 技术基于广域全光网进行长距离优化是短期内可行性高、成本低且利于开展研究的最佳方案之一,通过优化 RoCE 在广域全光网上实现“广域光数直达”逼近物理层通信性能指标。

参考文献(References)

[1] 关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见[EB/OL]. (2023–12–25) [2024–01–02]. https://www.gov.cn/zhengce/zhengceku/202401/content_6924596.htm.

[2] Brown T B, Mann B, Ryder N, et al. Language models are few–shot learners[EB/OL]. [2024–01–02]. https://arxiv.org/abs/2005.14165v4.

[3] OpenAI. GPT–4 is OpenAI’s most advanced system, producing safer and more useful responses[EB/OL]. [2024–02–04]. https://openai.com/index/gpt-4/.

[4] U S Bureau of Industry and Security. BIS updated public information page on export controls imposed on advanced computing and semiconductor manufacturing items to the People’s Republic of China (PRC)[EB/OL]. [2024–03–02]. https://www.bis.doc.gov/index.php/about-bis/newsroom/2082.

[5] McMahan H B, Moore E, Ramage D, et al. Communication–efficient learning of deep networks from decentralized data [EB/OL]. [2024–03–02]. https://arxiv.org/abs/1602.05629v4.

[6] Cerf V, Kahn R. A protocol for packet network intercommunication[J]. IEEE Transactions on Communications, 1974, 22(5): 637–648.

[7] Wikipedia. Remote direct memory access [EB/OL]. [2024–03–13]. https://en.wikipedia.org/wiki/Remote_direct_memory_access.

[8] 刘雨蒙, 唐正梁, 路松峰, 等. RDMA协议应用及安全防护技术综述[J]. 网络与信息安全学报, 2024, 10(2): 22−46.

[9] Wikipedia. InfiniBandInfiniBand[EB/OL]. [2024–03–13]. https://en.wikipedia.org/wiki/InfiniBandInfiniBand.

[10] Beck M, Kagan M. Performance evaluation of the RDMA over ethernet (RoCE) standard in enterprise data centers infrastructure[C]//Proceedings of the 3rd Workshop on Data Center–Converged and Virtual Ethernet Switching. Omaha: ACM, 2011: 9−15.

[11] Wikipedia. RoCE[EB/OL]. [2024–03–13]. https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet.

[12] Beck M, Kagan M. Performance evaluation of the RDMA over ethernet (RoCE) standard in enterprise data centers infrastructure[C]//Proceedings of the 3rd Workshop on Data Center–Converged and Virtual Ethernet Switching. San Francisco: ACM, 2011: 9−15.

[13] Wikipedia. iWARP [EB/OL]. [2024–03–13]. https://en.wikipedia.org/wiki/IWARP.

[14] Peterson C, Sutton J, Wiley P. iWarp: A 100–MOPS, LIW microprocessor for multicomputers[J]. IEEE Micro, 2002, 11(3): 26–29.

[15] Recio R, Metzler B, Culley P, et al. A remote direct memory access protocol specification[EB/OL]. [2024–03–13]. https://www.rfc-editor.org/info/rfc5040.

[16] Shah H, Pinkerton J, Recio R, et al. Direct data placement over reliable transports[EB/OL]. [2024–03–13]. https://www.rfc-editor.org/info/rfc5041.

[17] Yu W K, Rao N S V, Vetter J S. Experimental analysis of InfiniBand transport services on WAN[C]//Proceedings of International Conference on Networking, Architecture, and Storage. Chongqing: IEEE, 2008: 233−240.

[18] Bai W. Empowering azure storage with RDMA[C]//20th USENIX Symposium on Networked Systems Design and Implementation. Massachusetts, USA: Usenix, 2023: 49–67.

[19] Clark D. The design philosophy of the DARPA Internet protocols[C]//Symposium Proceedings on Communications Architectures and Protocols. New York: ACM, 1988: 106−114.

[20] 迈向智能世界白皮书2023—全光网[EB/OL]. [2024–03–13]. https://www-file.huawei.com/-/media/corp2020/pdf/giv/striding-towards-the-intelligent-world/the_intelligent_world_all_optical_network_2023_cn.pdf.

[21] 中国联通研究院. 算力时代的全光底座白皮书[EB/OL]. [2024–03–13]. http://221.179.172.81/images/20220921/28711663727812497.pdf.

[22] 中国电信发布全光网2.0技术白皮书[EB/OL]. [2024–03–13]. https://www.c114.com.cn/topic/117/a1179842.html.

[23] Jin X, Li Y R, Wei D, et al. Optimizing bulk transfers with software–defined optical WAN[C]//Proceedings of the 2016 ACM SIGCOMM Conference. New York: ACM, 2016: 87−100.

[24] Jacobson V, Braden R. RFC1072: TCP extensions for long–delay paths [EB/OL]. [2024–01–02]. https://www.rfc-editor.org/rfc/rfc1072.html.

[25] Hecht J. Understanding fiber optics[M]. Washington DC: Wiley–IEEE Press, 2018.

[26] 数据快递服务_DES_Teleport[EB/OL]. [2024–03–13]. https://www.huaweicloud.com/product/des.html.

[27] 天翼云SD_WAN解决方案[EB/OL]. [2024–03–13]. https://www.ctyun.cn/solutions/10050195.

[28] NVIDIA,Mellanox®ConnectX®–5网卡[EB/OL]. [2024–03–13]. https://www.nvidia.cn/networking/ethernet/connectx-5/.

[29] Github, open fabrics enterprise distribution (OFED) performance tests[EB/OL]. [2024–05–15]. https://github.com/linux-rdma/perftest.

On developing wide–area long–distance high–performance transport techniques in computer networks

LIANG Teng[1], YANG Jian[1,2], YANG Jiayu[1], ZHANG Yu[1,3], ZHANG Weizhe[1,2,3*]

1. Pengcheng Laboratory, Shenzhen 518000, China
2. Harbin Instituted of Technology (Shenzhen), Shenzhen 518055, China
3. Harbin Instituted of Technology, Harbin 150006, China

Abstract The technology for high–performance long–distance transmission has significant strategic value in the context of China's "East Data, West Calculation" project, which aims to construct a nationwide integrated computing network. Three trends are driving new requirements for the paradigm of wide–area distributed computing power coordination: 1) the rise of AI large–scale model intelligent applications that demand extremely high computing resources, 2) the embargo on high–performance GPU chips limiting single–center computing power resources, and 3) the formation of a dispersed computing power distribution pattern due to the establishment of computing clusters across various regions in China. High–performance long–distance transmission technology is crucial for this new paradigm. This paper discusses five aspects: supporting the new paradigm of wide–area distributed computing resource coordination, technical routes, underlay networks, challenges, and costs. Based on the results of a real–network experiment spanning 2100 kilometers from Shenzhen to Zhongwei in Ningxia, the authors believe that optimizing existing RDMA technology for long–distance transmission over wide–area optical networks is one of the most feasible and cost–effective solutions in the short term. By optimizing RoCE (RDMA over Converged Ethernet) over wide–area optical networks, it is possible to achieve "optical direct access to data in WAN" approaching physical layer communication performance indicators.

Keywords wide–area long–distance high–performance transmission; wide–area RDMA (WRDMA); compute–first networks; east–data–west–computing

（责任编辑　王微）