面向智算中心互联的光算协同技术研究

Optical-Computing Collaborative Technology for Intelligent Computing Center Interconnection

谭艳霞/TAN Yanxia[1]，满祥锟/MAN Xiangkun[1]，
吴绍辉/WU Shaohui[2]，张贺/ZHANG He[1]，徐博华/XU Bohua[1]
（1. 中国联合网络通信有限公司研究院，中国 北京 100048；
2. 北京市国防动员办公室，中国 北京 100053）
(1. Research Institute of China United Network Communications Co., Ltd., Beijing 100048, China;
2. Beijing Municipal National Defense Mobilization Office, Beijing 100053, China)

DOI：10.12142/ZTETJ.202506003
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20251218.1429.002
网络出版日期：2025-12-19
收稿日期：2025-10-15

摘要：针对智算中心互联对光网络的新需求，结合当前智算网络发展现状，探讨智算中心互联架构及关键技术，以实现高性能算力互联。同时，针对跨智算中心分布式协同训练场景，搭建基于光传送网（OTN）的跨智算中心现网试验环境，在广域收敛比不低于16∶1的场景下，百亿AI大模型跨域分布式训练性能达到95%以上。该试验验证采用单波800G实现300 km的传输，并验证其超高可靠传输能力。

关键词：智算中心互联；光传送网；分布式协同训练；高可靠传输

Abstract: The interconnection architecture and key technologies for intelligent computing centers are explored to address the new demands of optical networks for their interconnection, while considering the current development status of intelligent computing networks, with the aim of achieving high-performance computing power interconnection. Furthermore, focusing on the scenario of distributed collaborative training spanning multiple intelligent computing centers, an optical transport network (OTN)-based experimental testbed for cross-center interconnection is implemented on a live network. Under conditions where the wide-area convergence ratio is no less than 16:1, a performance of over 95% is achieved for cross-domain distributed training of AI large models with 10 billion parameters. Single-wave 800G transmission over 300 km is employed, and its ultra-high reliability and transmission capability are verified.

Keywords: interconnection of intelligent computing centers; optical transport network; distributed collaborative training; highly reliable transmission

引用格式：谭艳霞，满祥锟，吴绍辉，等. 面向智算中心互联的光算协同技术研究 [J]. 中兴通讯技术，2025，31(6)：13-19. DOI：10.12142/ZTETJ.202506003
Citation: TAN Y X, MAN X K, WU S H, et al. Optical-computing collaborative technology for intelligent computing center interconnection [J]. ZTE technology journal, 2025, 31(6): 13-19. DOI: 10.12142/ZTETJ.202506003

随着人工智能（AI）大模型的快速发展，智算中心已成为支撑企业数字化转型的关键基础设施，也是推动AI大模型创新的核心动力。据统计，大模型参数规模每两年增长约10倍，算力需求持续攀升，而当前图形处理器（GPU）芯片算力仅以2～4倍的速度增长，远落后于模型规模的扩张速度。随着模型体量不断增大，单体智算中心在算力、电力、空间等方面面临限制，需要在园区乃至更广范围内整合多个智算中心资源，通过高速互联形成逻辑统一的超级算力资源池，以支持大规模分布式协同训练。

近年来，中国积极实施“东数西算”工程，加快建设全国一体化算力网络，有力推动了智算中心互联的技术演进与业务创新。通过光网络与算力资源的协同调度，实现泛在化算力聚合与一体化服务，形成“以网强算”的发展格局。在智算时代，光网络需满足跨中心数据推理、多数据中心协同训练等多样化任务需求，具备更弹性、更智能的网络能力。因此，智算中心互联对大带宽、低时延、高可靠及智能化的光网络提出了明确要求，也为光通信领域带来了新的发展机遇。

1 智算中心互联光网络

为支撑大模型训练的持续增长需求，AI算卡规模正快速扩张。以OpenAI最新官宣启动的下一代前沿模型GPT-6为例，其训练预计将需要高达70万～80万张算卡支持，这已远超单体智算中心的承载能力。在此背景下，分布式协同计算已成为必然趋势。

智算中心互联光网络正是为应对这一趋势而构建的，旨在打造一个具备大带宽、低时延、高可靠、强智能特性的全光互联底座。该网络通过光算协同等关键技术，实现对分布式算力资源的高效协同与智能调度，能够有效满足“数据入算、模型训练、推理下发”等典型智算场景的需求，并以此驱动新的业务增长。

1.1 智算中心光互联架构及典型场景

随着智能城市、自动驾驶和超高清视频直播等应用的快速发展，AI技术正加速向各行各业深度渗透，区域间以及同城/区域内算力协同的需求急剧增长。智算中心互联架构如图1所示，通过构建算力间全光高速平面，将算力中心的Spine/Leaf节点经智算网关直接与光传送网（OTN）的光传输设备相连。智算网关具备多业务流识别与调度能力，可提供长距离、高效率、智能化的流量调度功能；OTN设备则基于物理层参数与业务侧参数协同，实现高吞吐、长距离、无损的数据传输服务，从而保障智算节点间数据传输质量，推动分布式算力资源的高效互联与协同发展。

从承载网络的视角分析，结合智算中心互联架构，可归纳出四大典型需求场景，如图2所示：数据入算、模型训练、模型下发与模型推理[9]。

1）数据入算。数据入算指将各类数据上传至算力中心，为后续训练与推理提供数据基础，在医疗、政务、金融、影视、科研等领域均有广泛需求。在该场景下，承载网络需具备弹性灵活的传输能力，例如能够识别并高效调度大象流，以应对太字节（TB）/拍字节（PB）级海量数据的传输需求。同时，网络应支持根据业务实际需要动态调配资源，提供弹性可定制、任务化的服务模式。

2）模型训练。模型训练是指通过大量的数据和复杂的算法，使模型学习数据中的规律和特征，从而具备特定的智能决策能力或预测能力。目前主要存在3种训练场景：数据中心（DC）内训练、跨DC协同训练和存算分离拉远训练。不同的训练场景对网络提出了差异化的要求。其中，跨DC协同训练场景是基于多DC协同进行超大模型训练或实现碎片化算力整合出租，对网络的运力服务提出了更大的挑战，需要网络具备广域无损、低时延与负载均衡等能力；存算拉远训练场景主要针对政务办公、医疗应用、金融领域等数据敏感用户，实现用户数据私域存储与AIDC之间的高效拉远训练，需要网络具备广域无损和数据安全保障等能力。

3）模型下发。模型下发是指将训练完成的人工智能模型从开发环境部署至生产环境或各类终端设备，以实现智能推理与决策能力。此场景对承载网络的需求与数据入算相似，同样要求网络具备高效传输大规模数据的能力，能够根据业务需求灵活调度资源，并提供按需定制或任务式的数据

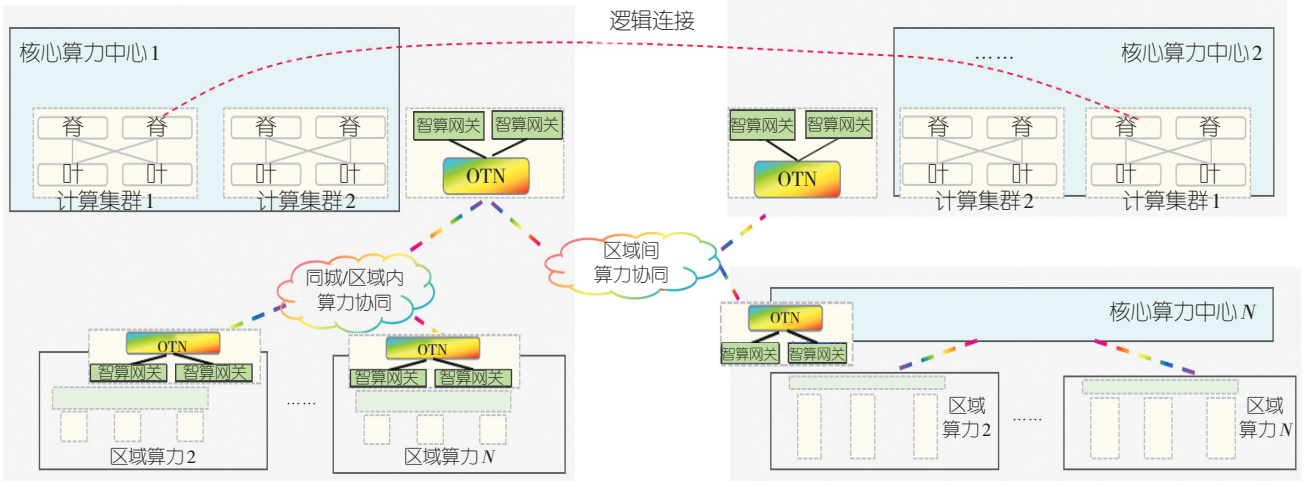


图1 智算中心互联架构示意图

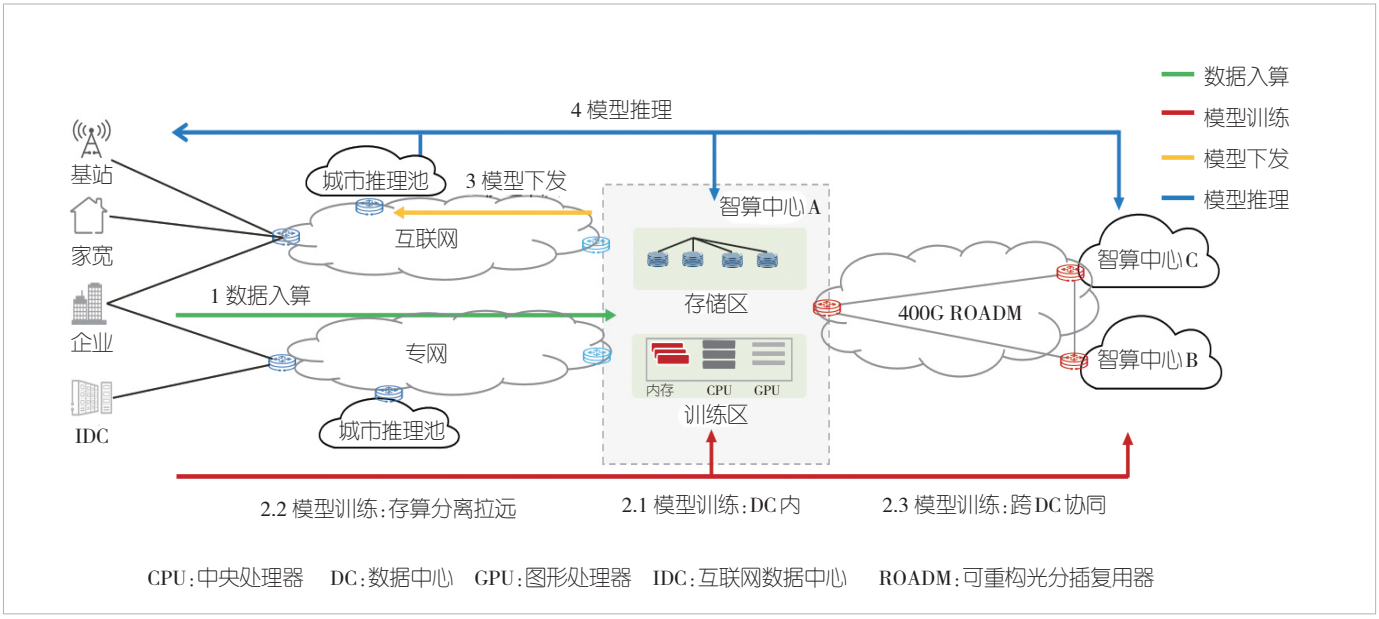


图2　智算业务典型场景[9]

传送服务模式。

4）模型推理。模型推理是指将训练好的模型应用于实际业务场景，对输入的新数据进行实时分析与预测，从而提供决策支持或直接的服务输出。为满足用户泛在接入与实时交互的需求，该场景要求承载网络具备广域覆盖能力及可定制的确定性承载性能。为此，需借助应用感知、算力感知及算网一体化调度等关键技术，实现网络服务的差异化与精准化承载。

1.2 智算中心光互联面临挑战

与智算中心内部互联相比，智算中心之间的长距离互联环境更为复杂。为满足上述智算中心互联典型场景中的用户需求，光网络首先需要具备长距离、大带宽、高速率的基础传输能力。在此基础上，还需进一步提供弹性、按需、可靠、智能的连接服务，并助力运营商实现新业务形态与商业模式的拓展。具体而言，智算中心互联对光网络的关键需求主要包括[8]：

1）大带宽：随着智算卡规模从千卡迈向万卡级别，所需互联带宽常高达数百太比特至超拍比特量级，单纤容量需持续提升。网络需兼顾大客户的超大带宽需求，实现海量数据的快速迁移，同时也要降低中小企业接入算力的成本，提供灵活的带宽服务。

2）低时延：传输距离的增大会带来时延的累积。为满足跨智算中心协同训练对传输时延的严格要求，光网络必须提供稳定且可保障的低时延传输能力。

3）任务式灵活拆建：为适应人工智能计算任务的动态需求，需支持光网络连接的快速构建、灵活调整与及时拆除，实现算力资源与网络资源的协同弹性调度，从而满足业务灵活拆建的需求。

4）可信可靠：在跨智算中心互联场景下，网络丢包、闪断及故障等问题可能影响协同计算的可行性与效率，降低算力利用率，甚至导致协同训练中断。与智算中心内部丢包相比，长距离传输中的丢包重传机制可能引入高达千倍的时延累积[10-11]。因此，光网络需具备高可靠的无损传输能力，并与终端侧协同优化，保障高性能传输协议的效率。

5）统一管理编排：当前光网络与智算中心网络在管控上相互分离，导致跨智算中心的分布式协同计算难以实现性能最优。现有固化的组织架构与生产流程亦无法支持逻辑统一的异地智算中心快速部署。因此，需对现有管理编排体系、生产流程及管控平台进行改造与升级。

2 智算中心光互联关键技术

着眼于智算中心间互联需求，实现算力节点之间的一体化协同、枢纽间/枢纽内算力的任意调度，本文以智算中心间互联给光网络带来的挑战为基础，提出智算中心光互联的关键技术。

2.1 确定性承载技术

智算中心间基于OTN/可重构光分插复用器（ROADM）技术[12-13]，采用单波400G/800G高速传输[13]构建大容量底层

光网络，通过在智算中心之间建立光层直达通道，实现高速、海量数据的传输服务。该长距传输不经过传统2/3层交换机或路由器设备，而是通过光层一跳直达至对端智算中心出口交换机，从而在物理层面避免因队列调度机制引入的额外时延与丢包，构建高效无损的点对点传输通道。

2.2 安全可靠技术

为保证客户数据安全，智算中心间网络需采用更可靠的安全技术以提升客户体验。OTN技术基于光层（L0）的波长波分复用与电层（L1）的光通路数据单元（ODU）固定时隙对用户业务进行物理隔离，确保各项业务资源独享、互不干扰。同时，OTN支持传输加密，并可与量子密钥分发[14]（QKD）等高安全性技术结合，实现保密通信。

OTN可提供多层次、全面的保护机制，具备高可靠保护能力，保障数据传输业务不中断。根据故障层次的不同，可分为电层保护、光层保护及光电协同保护[10]。电层保护面向业务级，提供端到端1+1保护，倒换时间低于50 ms；对可靠性要求极高或光纤中断风险较大的场景，还可支持抗多次断纤的毫秒级保护能力。光层保护针对网络级光线路或节点故障，支持光传输段（OTS）1+1保护或光层自动交换光网络（ASON）保护，并具备确定性重路由能力，以降低同一故障引发的大规模业务中断风险。光电协同保护示意如图3所示，适用于对可靠性有更高要求的业务，通过在电层部署1+1保护、光层配置ASON动态恢复，实现协同防护，可抵御多次光纤故障，满足业务99.999%的可靠性要求。

2.3 任务式带宽技术

为适应智算业务的弹性带宽需求，传输管道需从静态分配模式向灵活拆建模式演进，从以年为单位的长期占用转变为支持小时级、天级的分时复用。为此，光网络应具备“任务式敏捷建链能力”与“弹性带宽调整能力”。

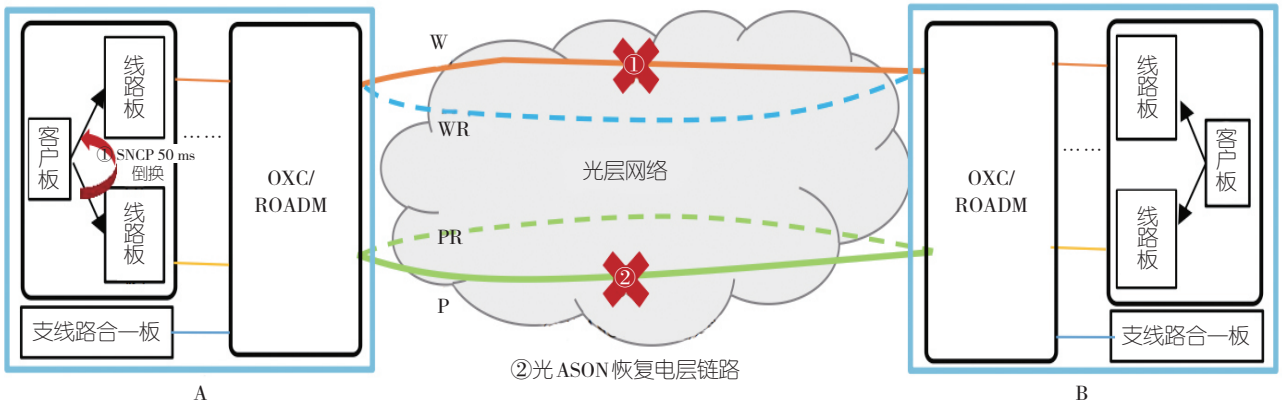
1）任务式敏捷建链，实现波长级传输通道快速开通。任务式敏捷建链的核心在于构建两大关键能力：一是基于电层驱动、实现最优资源规划的光层控制能力；二是实现波长级业务从配置到开通的全流程自动化。

2）基于Flex-O技术的弹性管道按需带宽调整。Flex-O技术通过信号拆分与FlexO映射，将多个标准速率的100G光模块绑定，实现$N\times$100G信号的高速传输。通过动态绑定或解绑定光模块数量，即可灵活调整传输带宽，满足业务对带宽的弹性需求。

2.4 长距离无损传输技术

为了应对超长距传输的挑战，承载网络需具备长距无损确定性传输能力，并与终端侧协同优化，以满足高性能协议对传输效率的要求。长距离无损传输技术主要包括端网协同物理层及保护路径信息、端网协同流控两方面。

1）端网协同物理层及保护路径信息。在长距离场景下，可重构光分插复用器（RDMA）的吞吐量受传输距离、链路误码、保护倒换等因素影响。传输距离越长，吞吐量通常越低。网络可将物理路径的距离信息通过协议传递给网卡侧，网卡据此动态调整发送参数，以实现长距环境下的满速传输。当链路误码触发RDMA重传机制时，网络向终端侧通知



光电恢复优势互补，以更少的资源，实现抗多次断纤更高生存性

ASON：自动交换光网络　PR：保护倒换　SNCP：子网连接保护
OXC：光交叉连接器　ROADM：可重构光分插复用器　WR：波长路由

图3　光电协同保护

误码信息，终端据此判断重传原因属于链路误码还是网络拥塞，进而采取不同的传输策略。网络侧因故障触发的保护倒换可能影响传输效率，网络设备需将倒换后新路径的长度信息通知终端侧，由终端侧调整发送参数，以恢复满速传输。

2）端网协同流控。在长距传输过程中，若远端数据中心发生拥塞，交换机会触发优先级流量控制（PFC）等流控机制，但其缓存能力通常不足以支撑长距无损传输。针对广域长距RDMA承载场景，OTN等传输设备需要具备对PFC流控信号的响应能力，协同缓存反压流量，并支持逐级向上游反压流量，从而确保RDMA流量在拥塞时不丢包，保障业务有效带宽。通过传输设备与交换机的协同机制，最终实现长距离环境下的无损传输[4]。

2.5 算网协同管控编排技术

智算中心大规模部署后，呈现出分布式布局特征，且整体利用率较低。这一现状对算网协同智能调度提出了更高要求，需实现跨地域、跨层级、跨主体的高可靠调度。为保障典型智算业务的稳定运行，如图4所示，需通过算力管理系统与网络管理系统的协同调度，实现对算力与网络资源的统一管控与编排，从而为智算业务提供一体化的服务能力[15]。

1）构建算网业务运营层，实现算网产品的一体化运营服务。该层面向不同任务需求（如入算、训练、推理等），为客户提供基于数据量、时长等参数的产品化服务新模式。

2）构建算网协同编排调度层，实现算力与网络的统一编排。该层向上为运营层提供统一的算网一体业务服务能力，支持跨网络域与算力域的资源调配，还提供端到端业务编排与统一调度功能，支持时延优先、算力优先、成本优先等多种算网协同调度策略。

3）强化网络管控层的动态拆建能力。网络管控系统需通过北向接口开放业务服务能力，将包含用户至各算力节点的连接时延、带宽等信息的“运力地图”同步至算网大脑，以支撑其快速、自动地选择最优算力节点与网络连接。

3 智算中心光互联的典型试验验证

针对多智算中心协同训练的典型应用场景，我们开展了跨智算中心的分布式协同训练现网试验验证。本次验证通过优化智算模型的并行策略，适配广域带宽条件下的超大收敛比，有效降低了超大规模智算中心互联场景中对长距离传输带宽的需求。试验中采用800G光传送网实现了300 km距离的超大带宽传输，并对多种可靠性影响因素进行了测试，包括智算拉远场景下传输带宽下降、链路误码以及保护倒换对计算效率的影响。

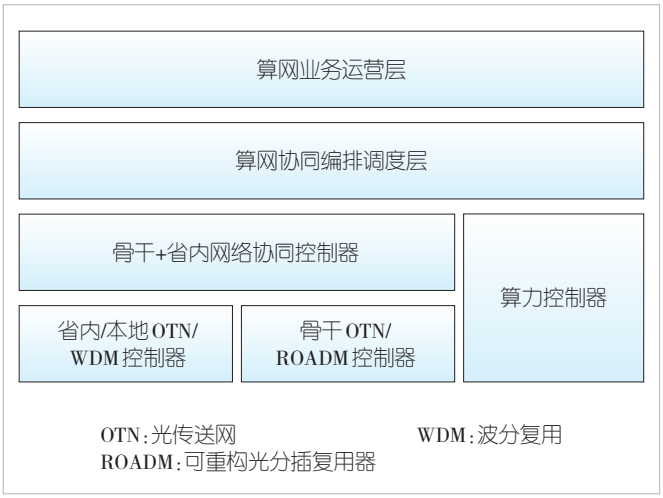


图4 算网协同管控编排架构

3.1 试验验证环境

分布式协同训练技术是指利用多个智算中心的计算资源协同完成同一AI大模型的训练任务，每个数据中心承担部分训练工作，并将每次迭代的结果实时同步，从而实现计算资源的按需分配与高效利用。跨智算中心分布式协同训练试验验证场景如图5所示，在现网验证中构建了包含32张计算卡的双智算中心环境，每个中心配置16张卡。智算中心内部采用Leaf+Spine两层拓扑的基于融合以太网的远程直接内存访问（RoCE）网络，中心之间通过智算网关经800G OTN设备直接互联，并依托OTN设备与长距光纤实现约329 km的远程连接，互联带宽为2×800G。试验验证中，OTN客户侧配置2个400G接口，在物理带宽上实现了4∶1的收敛比。

为评估跨智算中心集群对模型训练效率的影响，在试验验证中部署了开源模型LLaMA 7B与LLaMA 13B，对在不同收敛比及模型参数配置下的计算效率变化情况进行研究与验证。在计算拉远协同训练场景的等效算力时，以相同参数配置下、基于数据中心内部网络进行的基准测试结果为参照，将拉远场景在同一时间内完成的训练进度占比作为等效算力指标。无论是否采用拉远部署，对于相同模型参数及训练样本集而言，其单轮迭代所需的总浮点运算次数保持不变，因此等效算力的计算公式可简化为：

$$\text{拉远等效算力} = \frac{\text{拉远场景总秒均算力}}{\text{基准场景总秒均算力}} = \frac{\text{单迭代总浮点运算次数/拉远场景单次迭代时间}}{\text{单迭代总浮点运算次数/基准场单次迭代时间}} = \frac{\text{基准场景单次迭代时间}}{\text{拉远场景单次迭代时间}}\%。$$

3.2 试验验证结论

本次试验通过引入新一代智算网关、精准流量控制与并行策略优化，在按张量并行（TP）：流水线并行（PP）：数据并行（DP）= 4：4：2的比例划分模型的情况下，采用PP并行方式跨中心拉远部署，实现了不低于16：1的广域收敛比。在此配置下，百亿参数规模大模型的分布式训练性能可达到单智算中心训练性能的95%以上。算力效率测试结果如图6所示，图中数据为多次测试取平均所得。

在光传送网可靠性对算效影响方面，通过在OTN网络中模拟传输带宽下降、链路误码及保护倒换等故障，研究并验证了不同故障场景对计算效率的影响。

在服务器上部署LLaMA2 13B模型，按TP：PP：DP = 4：2：4的比例划分模型，并行方式分别配置为PP拉远与DP拉远。在PP拉远场景下，当传输带宽从800G降至400G时，单轮迭代时间平均为10.840 12 s，对算效基本无影响。在DP拉远场景下，同样将带宽从800G降至400G，单轮迭代时间在一个迭代周期内增加至15.249 3 s，计算效率下降约38.496%，随后逐渐恢复。

我们将网关收敛比设置为4：1，并行方式配置为PP拉远。无误码时单轮迭代时间正常约为10.784 3 s；当出现误码且误码率在10^{-9}至10^{-10}量级间波动时，模型训练效率下降，单轮迭代时间延长至约15.42 s，如图7所示。当误码率升至10^{-8}量级时，模型训练将出现中断。

我们将网关收敛比设置为8：1、并行方式配置为PP拉远。当模拟工作链路发生50 ms级别保护倒换时，仅影响一个训练迭代周期，该周期内计算效率下降约56%，但对整体协同训练过程影响有限。当模拟工作链路触发ASON动态重路由倒换时，同样影响一个迭代周期，但在部分情况下会导致模型训练直接中断。训练中断的出现，可能与倒换过程中关键数据帧丢失有关。

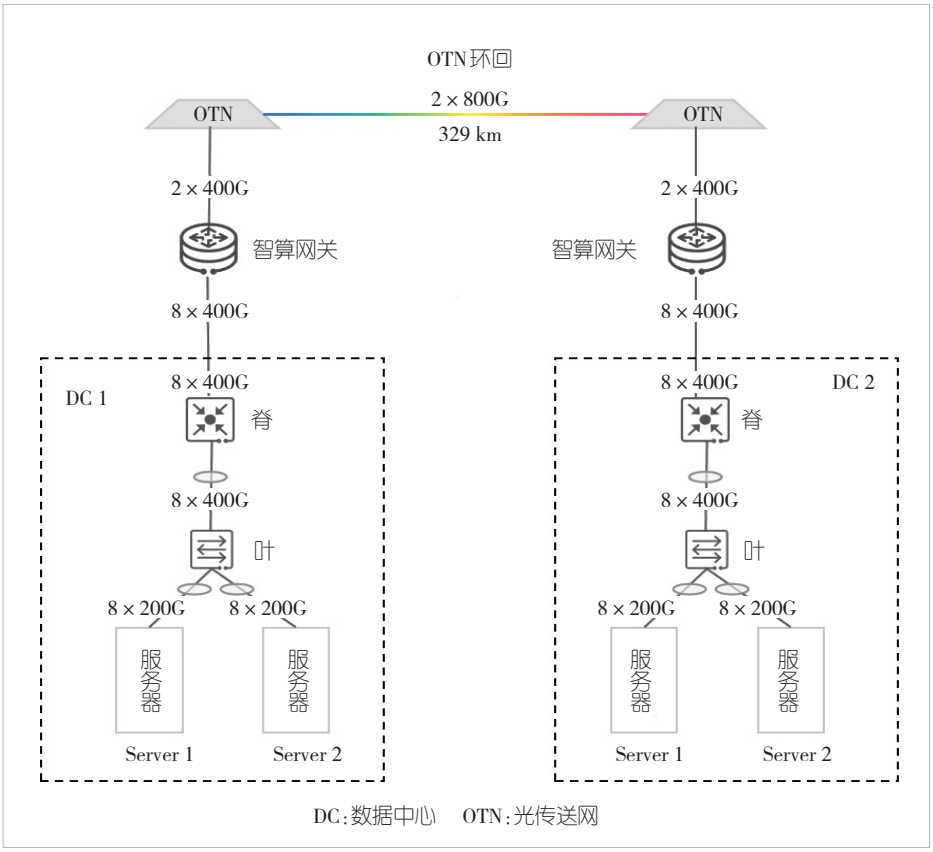


图5 跨智算中心协同训练试验验证拓扑图

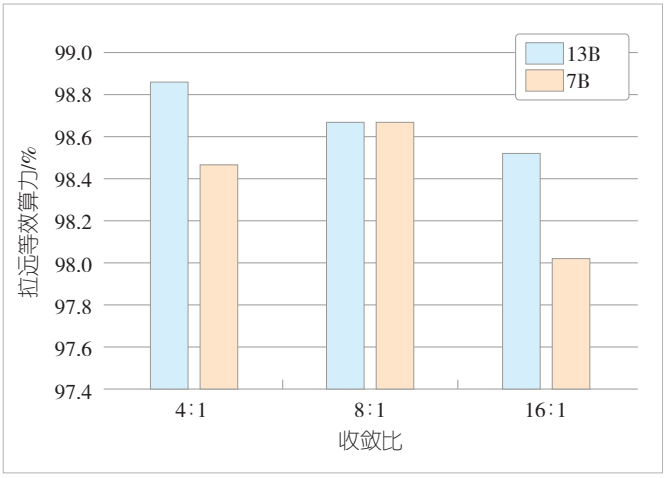


图6 分布式训练等效算力测试结果

4 结束语

面向智算中心互联的光算协同技术，需以光网络为物理底座，通过突破确定性承载、安全可靠、任务式带宽、长距离无损传输以及算网协同管控编排等关键技术，构建“高速无损、灵活智能”的算力互联体系。未来，为应对智算网络面临的多重挑战，将在现有研究成果基础上，积极探索任务式灵活部署产品形态，优化算网协同管控机制与网络计费模式，实现算力在任意地点与时间的灵活调度。同时，持续开

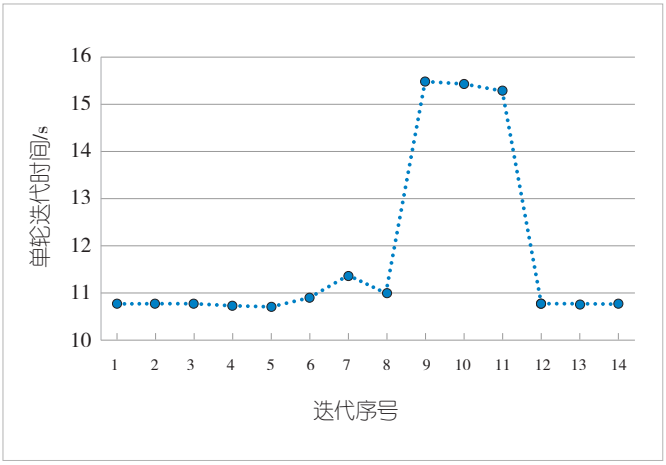


图7 链路误码对训练单轮迭代时间的影响测试结果

展异构算力协同训练应用试验，推动RDMA协议层与光网络物理层之间的感知联动，最终实现超长距离环境下的高吞吐量无损传输。

参考文献

[1] 李国杰. 智能计算技术的历史性突破与巨大挑战 [J]. 集成技术, 2025, 14(1): 1−8
[2] 中国信息通信研究院. 中国算力发展报告(2024年) [R]. 2024
[3] 丁宏庆, 张鹏飞, 牛红, 等. 云化的智算中心万卡集群创新与实践 [J]. 电信科学, 2024, 40(12): 125−135
[4] TAN Y X, MAN X K, WANG G Q, et al. Field trial of long−distance RDMA lossless transmission for wide−area data center interconnection [EB/OL]. (2024−11−05) [2025−11−08]. https://ieeexplore.ieee.org/document/10809882
[5] 张德朝, 孙将, 曹珊, 等. 面向跨智算集群互联的新型HIC−OTN技术 [J]. 电信科学, 2025, 41(4):53−60
[6] LIU Y Y, ZHANG A X, WANG X S, et al. Field trial of multi−datacenter distributed training for LLM based on bandwidth convergence and two parallel strategies over 120km high−reliability 800Gbit/s C+L OTN [EB/OL]. [2025−11−09]. https://ieeexplore.ieee.org/document/11047207
[7] 中国信息通信研究院. 算力时代全光网架构研究报告 [R]. 2024
[8] 中国联通研究院. 基于RDMA的长距无损数据搬移技术白皮书 [R]. 2024
[9] 易昕昕, 张乃瞻, 刘雅承, 等. 算力智联网关键技术研究 [J]. 中兴通讯技术, 2025, 31(2): 31−38. DOI: 10.12142/ZTETJ.202502005
[10] 王光全, 满祥锟, 徐博华, 等. 确定性光传输支撑广域长距算力互联 [J]. 邮电设计技术,2024(2): 7−13
[11] MACARTHUR P, RUSSELL R D. A performance study to guide RDMA programming decisions [EB/OL]. [2025−11−09]. https://ieeexplore.ieee.org/document/6332248
[12] 唐雄燕, 王海军, 杨宏博. 面向专线业务的光传送网(OTN)关键技术及应用 [J]. 电信科学, 2020, 36(7): 18−25
[13] WANG C Y, HU Y K, SHEN S K, et al. Channel power management of 400 G transmission system based on C6T + L6T spectrum and QPSK modulation format [J]. Optics express, 2024, 32(11): 20279. DOI: 10.1364/oe.523644
[14] QU W X, ZHANG Y, LU Y M, et al. Low−cost lightweight−client twin−field quantum key distribution network with wavelength division multiplexing [EB/OL]. [2025−11−10]. https://ui. adsabs.harvard.edu/abs/2022OptEn..61a6102Q/abstract
[15] 中国联通研究院. AI时代的全光底座白皮书 [R]. 2025

作 者 简 介

谭艳霞，中国联合网络通信有限公司研究院高级工程师；主要研究方向为传输网管控技术、智算网络等。

满祥锟，中国联合网络通信有限公司研究院正高级工程师；主要研究方向为光通信网络、智算网络等。

吴绍辉，北京市国防动员办公室工程师；主要研究方向为传输网管控技术、智算网络等。

张贺，中国联合网络通信有限公司研究院正高级工程师；主要研究方向为光纤传输、同步网新技术等。

徐博华，中国联合网络通信有限公司研究院高级工程师；主要研究方向为数据中心网络关键技术等。