XXXX

云边协同分布式训推方案研究

李志勇[1]，郗卓宁[1]，田雯[2]，张明[3]，郑坤松[2]，曹原铭[2]，孙傲[2]，武振宇[2]，丁国强[2]，魏逸哲[2]
（1. 中国移动通信集团有限公司，北京，100033；
2. 中国移动通信集团设计院有限公司，北京，100080；
3. 中国移动通信集团有限公司供应链管理中心，北京，100053）

摘　要：企业智能化转型中，大模型训练与推理需要大量智能算力资源，自建算力往往成本高、建设周期长，而租赁算力服务则面临数据出域安全风险等问题。本文提出云边协同分布式训推方案，通过模型分层部署，边缘企业侧处理敏感数据，云端智算中心承载大规模计算，中间通过支持RDMA的跨域无损传输网络传递敏感性更低的中间变量，在保障企业数据隐私性的基础上，依靠安全网络架构，实现算力弹性扩展，形成适配不同行业场景的高性能组网解决方案。

关键词：云边协同；分布式训推；模型分层部署；跨域无损传输
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.

Research on Cloud-Edge Collaborative Distributed Training and Inference Solution

LI Zhiyong[1], XI Zhuoning[1], TIAN Wen[2], ZHANG Ming[3], ZHENG Kunsong[2], CAO Yuanming[2], SUN Ao[2], WU Zhenyu[2], DING Guoqiang[2], WEI Yizhe[2]
1. China Mobile Group Co., Ltd., Beijing 100033, China
2. China Mobile Group Design Institute Co., Ltd., Beijing 100080,China
3. China Mobile Group Supply Chain Management Center, Beijing 100053,China

Abstract: In the context of enterprise intelligent transformation, large model training and inference require massive intelligent computing resources. Self-built computing power often suffers from high costs and lengthy construction cycles, while rented computing power services face issues such as data exfiltration security risks. This paper proposes a Cloud-Edge Collaborative Distributed Training and Inference Solution. Through hierarchical model deployment, the enterprise edge side processes sensitive data, while the cloud-based intelligent computing center undertakes large-scale computing tasks. Between the cloud and edge ends, non-sensitive intermediate variables are transmitted via an RDMA (Remote Direct Memory Access)-supported cross-domain lossless transmission network. On the basis of ensuring enterprise data privacy, the solution achieves elastic scaling of computing power based on a secure network ar-

收稿日期：2026−XX−XX；修回日期：2026−XX−XX
通信作者：田雯
基金项目：Foundation Items: 无

chitecture, and finally forms a high-performance networking solution that is adaptable to diverse industry scenarios.

Key words: Cloud-Edge Collaboration, Distributed Training and Inference, Hierarchical Model Deployment, Cross-Domain Lossless Transmission

0 引言

模型微调与推理应用正成为企业智能化转型的核心算力需求。与基础大模型不同，企业应用往往基于行业专有数据和企业私有样本数据，通过二次训练或微调[1]形成行业大模型或特定场景应用模型，因此对数据安全要求严苛，通常采用本地化处理方式。然而，企业算力需求具有显著的动态性，在业务高峰期，本地算力资源可能难以同时满足推理业务与二次训练的并发需求。为此，企业亟需一种既能保证数据不出园区，又能实现算力灵活扩展的解决方案。云边协同分布式训推架构与这一需求高度契合，即企业可在本地完成敏感数据的处理，同时通过智算广域专网安全地调用云端智算中心的弹性算力资源。

当前云边协同存算分离方案只是在本地保存原始数据，但在训练或推理过程中，原始数据、提示词、上下文和结果通常均在云端参与计算，敏感信息存在暴露风险；而拆分学习（Split Learning）将模型切割层拆成前后两段，前端浅层部署在客户端，后端深层部署在服务端，主要聚焦训练场景，对企业用户模型微调及推理场景下的全过程数据保护支持有限。针对上述问题，本文提出一种面向企业场景的云边协同分布式训推方案，通过模型首尾层本地部署、中间层云端部署的方式，使敏感数据始终保留在本地，与云端仅交互激活值和梯度值等中间结果，在保障数据安全合规的同时实现云端算力的弹性调度与灵活扩展。

1 当前政企客户使用算力资源面临的问题

在政务办公、医疗应用及金融服务等典型行业场景中，人工智能模型的训练与推理对算力需求日益增长。当前企业主要通过自建私有算力池或租赁智算中心算力来满足业务需求，但两种模式均存在风险与挑战。

采用自建私有算力池的方式，企业虽然能够确保敏感数据的本地化存储与处理，但实际落地存在以下三方面的难度：一是基于现有机房的改造往往面临电力负载不足的问题，且单机柜功率密度提升还会对机架布局、散热系统等配套设施提出更高要求，完成智算资源适配的机房改造往往需要 3~6 个月，制约业务上线速度；二是智算设备造价高昂、耗电需求大，前期建设和后续运营需大量成本投入；三是智算设备运维复杂，设备安装调测和故障定位恢复难度高，多数企业缺乏相应的专业运维能力，可能导致昂贵的智算设备长期处于低效运行状态，影响资源性能的充分发挥。

采用租赁智算中心算力的方式，可有效缓解本地算力不足的问题，但面临隐私数据出域[2]的安全挑战。政务、医疗及金融等行业的核心数据高度敏感，因其涉及公共利益与个人隐私，受到行业法规的严格监管，对数据的物理隔离和本地化处理有固有要求，算力租赁过程中的数据跨域传输行为，可能引发严重的行业违规并产生较大的安全隐患。

2 云边协同分布式训推方案

云边协同分布式训推方案可有效解决上述问题。在企业侧部署训推一体机，满足隐私数据不出园区的安全要求，轻量化部署，上线周期短；同时租赁云端算力，按需扩缩容，轻松应对模型微调和推理业务的潮汐式需求，避免重资产投

入，实现降本增效。二者协同，可有效弥补纯自建或纯租赁方案的短板，助力企业智算业务快速、安全、低成本地部署上线。

云边协同分布式训推方案总体架构如图 1 所示，具体而言，采用模型分层部署[3]的方式，模型的首尾层部署在企业园区本地算力池，用于处理输入输出环节，确保输入的私有数据和训推过程产生的提示词、上下文、结果等输出数据均不出域；模型的中间层则部署于云端智算中心，承载大规模计算。本地与云端之间仅交换梯度值、激活值等敏感性更低的中间变量，从而在降低用户本地资源需求及传输带宽需求的同时，保障核心数据的本地安全存储与处理。

在网络架构方面，本地出口侧需部署用户边缘接入设备（Customer Edge，CE[4]），负责发起智算中心连接请求，进行数据加密与协议转换；云端智算中心出口侧需部署运营商边缘接入设备（Provider Edge，PE[4]），实现企业流量的汇聚、路由转发、带宽调度与安全防护。本地和云端之间通过支持远程直接内存访问（Remote Direct Memory Access，RDMA[5]）技术的智算专线互联，以其高速无损的特性，为大模型训练与推理提供关键网络保障。该方案能显著提升吞吐量并大幅降低时延，从而确保业务服务等级协议（Service Level Agreement，SLA[6]）的稳定性。

2.1　模型分层部署实现方案

主流大语言模型[7-9]通常包含输入层（Embedding[10-12]）、多个中间层（Transformer[13-14]）和输出层（Unembedding[10-12]）组成。Embedding 将输入词元（Token）映射为向量，通过自注意力机制和前馈神经网络形成多个堆叠的 Transformer 层，最终由 Unembedding 将隐藏状态映射回词表大小向量。

在分布式训推方案中，优先考虑数据安全，在跨域数据暴露风险最小化的基础上，结合本地资源最小化部署原则，故将模型的 Embedding 层、Unembedding 层和 Transformer 的首尾层在企业侧部署，Transformer 的中间层在云端智算中心部署。从数据安全角度考虑，Embedding 层直接处理用户输入、企业私有数据以及上下文信息，而 Unembedding 层直接生成最终输出结果，均需保留在本地以满足数据不出域的安全要求，同时在本地部署 Transformer 的首尾层，利用非线性结构和特征变换能力提升还原难度；从资源利用角度考虑，大部分参数量和计算开销集中在 Transformer 中间层，将中间层部署于云端智算中心，可显著降低企业本地资源需求，实现数据安全与资源成本之间的平衡。

（一）　分布式训练模型部署方案

云边协同分布式训练模型分层部署流程如图

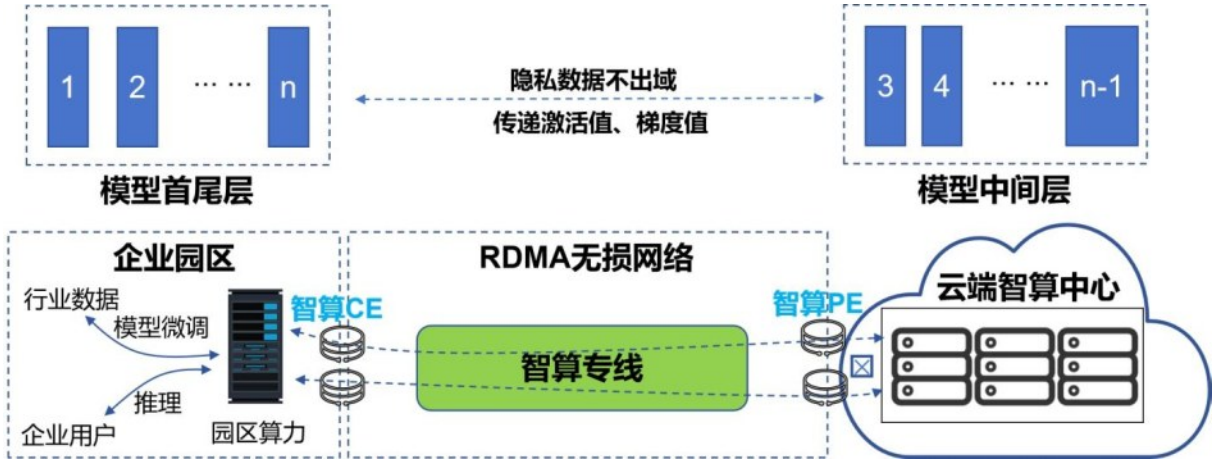

图 1　云边协同分布式训推方案总体架构图

2所示，通过“前向传播传递隐变量与反向传播传递梯度”的双向数据交互，实现云边协同分布式训练下的模型参数迭代更新。企业侧计算节点主要负责原始多模态私有样本（文本、音频、图像、视频等）的采集与预处理，将本地训练样本映射为高维隐变量，并生成首层隐变量数据，通过智算跨域专线传输至云端，云端中间层则基于该隐变量表示进行深度特征建模与复杂表征学习，生成尾层隐变量数据回传至企业侧首尾层，再由企业侧Unembedding层解码隐变量，结合损失函数完成输入样本与标签之间的误差计算，从而确保原始训练样本始终不出本地。

在反向传播阶段，企业侧基于损失函数计算本地部署层的梯度，并将梯度信息传输至云端，云端利用接收到的梯度对中间层参数进行更新，并将更新后的梯度信息反馈至企业侧，从而驱动企业侧模型参数的同步优化，形成从边缘局部训练到云端全局优化的闭环训练机制。

（二）　分布式推理模型部署方案

分布式推理模型分层部署流程如图3所示，在推理阶段，用户首先在企业内部输入文本、音频、图像和视频等多模态提示词（Prompt），由企业侧Embedding层基于神经网络结构实现对输入Prompt的特征映射与向量化处理，生成首层隐变量，该隐变量数据通过智算跨域专线上传至云端中间层。云端智算中心对接收到的隐变量进行深度非线性变换与特征推理，生成尾层隐变量数据并回传至边缘侧首尾层；最终，企业侧Unembedding层对尾层隐变量进行解码，生成业务Token并输出，完成推理服务闭环。

2.2　数据安全解决方案

在云边协同的分布式大模型应用中，数据安全问题贯穿训练与推理的全流程。企业应用的专有数据、私有样本以及知识增强检索体系一旦泄露，将直接威胁知识产权安全。因此，构建一套覆盖分布式训练与分布式推理的系统化数据安全解决方案，成为企业智能化转型过程中的关键任务。

大模型本身具备一定的数据隐私保护能力，以主流大语言模型为例，Embedding层为纯线性映射，token直接对应固定高维向量，无额外非线性变换；Unembedding层结合概率归一化函数（Softmax）与采样算法，将固定输出转为概率随机映射，增大逆向还原难度。Transformer含多层

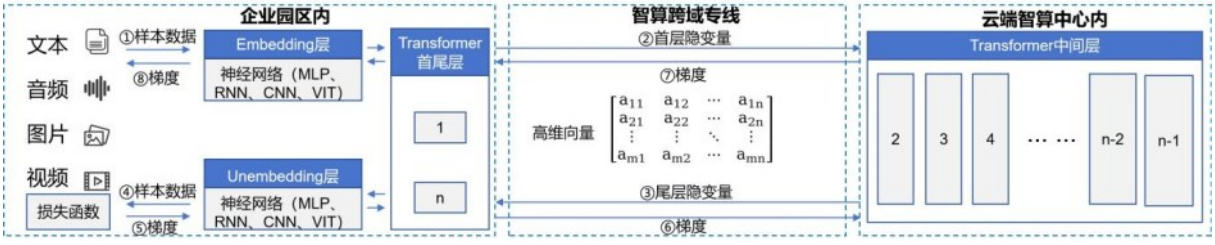

图2　分布式训练模型分层部署流程示意图

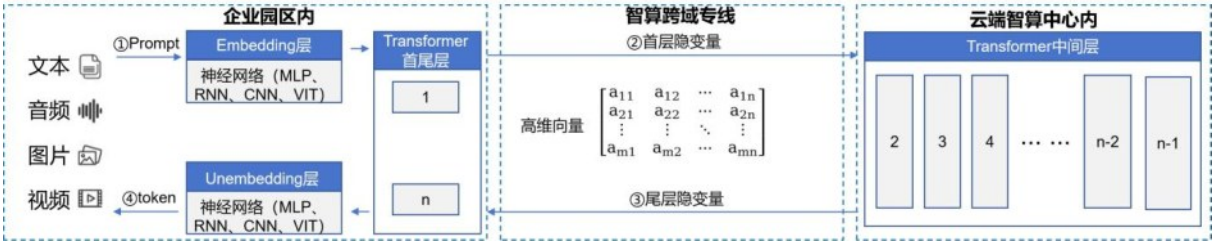

图3　分布式推理模型分层部署流程示意图

堆叠结构，每一层通常包含多头注意力子层（Multi-Head Attention, MHA）与前馈神经网络子层（Feed-Forward Network, FFN）。在计算过程中，通过层归一化（Layer Normalization, LayerNorm）与残差连接（Residual Connection）对子层输入与输出进行融合，可以增强信息传播能力；搭配非线性激活函数（Gaussian Error Linear Unit，GELU）、线性变换（Linear Transformation）以及随机失活机制（Dropout），以提升模型表达能力并抑制过拟合。多层非线性变换叠加使得输入信息在高维特征空间中逐层重构与混合，大幅提升通过中间变量反推原始数据的破解难度，筑牢隐私安全基础。

在云边协同分布式训推场景中，为切实保障数据安全，企业在模型部署和应用阶段可采取以下针对性策略，进一步降低企业隐私数据泄露风险。其一，在模型首轮迭代与初步训练阶段，优先使用脱敏的公共数据集开展模型微调工作，避免企业核心私有样本在训练初期直接参与计算，从源头减少隐私数据被逆向提取、特征恢复的风险；其二，在企业侧边缘节点进行模型部署时，除常规配置 Embedding 层与 Unembedding 层外，额外增设 Transformer 首尾关键层级的本地化部署[15-16]，通过强化模型非线性结构表达与特征变换能力，增加攻击者逆向还原原始输入的难度，显著削弱基于激活值反演等隐私攻击的可行性，提升整体隐私安全防护能力。

3　云边协同分布式训推组网要求

为确保云边协同分布式训推方案的可落地实施性，云端智算中心、跨域专线、企业园区侧分别需要具备不同的组网能力。

在云端智算中心，需要开放参数面与数据面的网络通道，基于 RDMA 相关协议，实现与跨域专线的对接，构建跨域数据交互的网络通道。在网络安全与架构优化层面，“目标方案”在参数面与数据面汇聚（Spine）交换机侧新增边缘网关交换机（Border-Gateway，BGW），连接至运营商 PE 设备，实现边界租户隔离、最小路由发布及白名单流量过滤，从网络接入控制与流量治理方面强化安全防御能力。“简化方案”可基于现有参数面和数据面的 Spine 交换机空余接口，直接连接运营商 PE 设备，并在 Spine 交换机上配置精细化过滤规则，在复用既有网络设施的前提下，保障参数与样本数据传输的安全性与可控性。

跨域专线建设上，运营商各级 PE 设备及用户侧 CE 设备均需升级具备 RDMA 无损传输能力，以满足大模型训推场景下大带宽、低时延、高可靠的数据传输需求。部署方案上，在基于优先级的流量控制技术（Priority-based Flow Control，PFC）基础上，引入精准流控技术，在报文中额外携带用户流量特征信息，依托第六版互联网协议（Internet Protocol Version 6，IPv6）的分段路由切片能力，通过全网部署业务专属切片，可有效解决多租户共享带宽的拥塞扩散问题，满足不同业务对网络带宽、时延、抖动等差异化 SLA 需求。此外，运营商各级 PE 设备还需具备流级负载均衡技术，部署专有网络控制器实现对整网流量路径的信息收集，通过优化算法精准识别大象流并上报控制器，控制器结合整网链路带宽利用率情况，动态调整大象流端到端路径，确保大小流均衡混跑。

同时，运营商侧需规划部署光传送网（Optical Transport Network，OTN）、分组交换以太网或裸纤直连等末端传输链路，实现用户侧的网络接入。企业园区内部的算力设备均接入用户侧 CE 设备，实现与云端智算中心的网络互联，为园区内训推任务提供端到端的网络通路与算力协同接口。

（一）　省内组网方案

在同省业务场景下，云端智算中心内部集成

通算、智算及存储设施以及各级交换机等网络设施，形成完整的云端算力与数据支撑体系。内部逻辑划分为业务面、参数面、数据面三张专用网络，通过BGW交换机及边界租户隔离、流量白名单过滤等技术手段，对参数面与数据面实施网络安全隔离，从网络域保障训推过程中参数与样本数据的机密性和完整性。同时，运营商PE设备作为连接云边算力集群的核心互联节点，承接边缘侧的算力调度请求与数据交互需求，实现云端算力对边缘侧的能力输出。

（二） 跨省组网方案

在跨省场景下，云端智算中心与企业园区侧内部部署与同省业务场景保持一致。不同的是，在跨域专线方面，需通过跨省光传送网构建运营商PE设备之间的广域无损传输通道，实现省间算力资源共享，缓解区域算力分布不均问题。

云边协同训推网络架构如图4所示。

4 测试验证分析

4.1 测试环境搭建

整体测试环境搭建基于大模型推理场景，采用云边协同的方式执行推理任务，边侧处理大模型的Embedding/Unembedding层和Transformer的首尾层，云端处理Transformer的中间层。物理组网上，云端智算中心选用2台国产智算服务器（共16卡），单卡半精度浮点计算（Floating Point 16-bit，FP16）的理论算力为376万亿次浮点运算每秒（Tera Floating Point Operations Per Second，TFlops），内部卡间通信的吞吐速率为双向56吉字节每秒（Gigabytes per second，GB/s），每台设备采用2条25吉比特每秒（Gbps）的链路上行连接至业务面网络，用于模型部署和可视化界面操作，同时采用8条200Gbps链路接入参数面网络，支持RDMA无损转发能力，用于本次云边协同推理任务的测试，参数面和业务面网络均采用100Gbps链路接入云端智算中心侧的PE设备；边侧在用户机房部署1台同等规格的国产智算服务器（共8卡），物理网络环境与云端智算中心保持一致；用户机房侧的CE设备与云端PE设备之间采用10Gbps链路的广域传输专线，实现云边两端设备的接入与数据转发，广域传输专线同样具备RDMA无损转发能力，拉远距离为128公里。

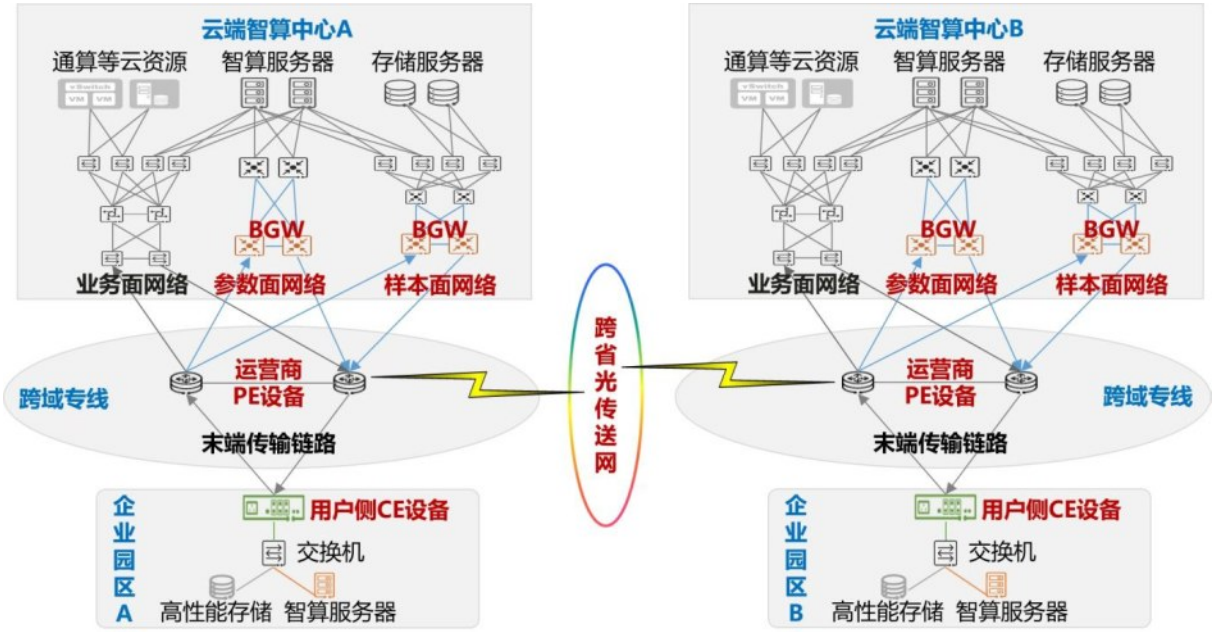

图4 分布式训推组网架构示意图

云边协同推理验证方案组网示意图如图 5 所示。

测试选用 Qwen2.5-32B（320 亿参数量）大语言模型作为测试模型，配置最大输入长度为 4,096 tokens，最大输出长度为 1,024 tokens。负载测试以每秒 1.5 个请求的速率持续发起，累计处理 500 个用户请求。本次主要测试四类典型场景：场景一是云端 16 卡无拉远、无收敛、模型不分层部署场景，本场景作为基线标准场景；二是云端 16 卡无拉远、无收敛、模型分层部署场景；三是云边协同 8+8 卡 128km 拉远、开启传输无损能力、有收敛、模型分层部署场景；四是在场景三的基础上关闭传输无损能力的对比场景。其中场景三与场景四均包含 10/8/6/5/4/3 Gbps 多带宽梯度测试。本次测试主要验证各场景的每 token 平均输出延迟（Time Per Output Token，TPOT）和系统吞吐量（tokens per second，tokens/s）两大指标，并对云边协同场景的劣化情况进行分析。测试场景如表 1 所示。

4.2　测试结果分析

基于上述测试环境和测试场景下的实际测试结果如表 2 所示。

此部分测试结果显示，场景二在云端智算中心引入模型分层部署架构，每 token 平均输出时延略微降低，考虑测试存在少量抖动偏差，可认为与基线水平持平，吞吐性能仅有 0.08% 的微小劣化，算力利用率趋近于 100%。场景三采用云边协同模型分层部署模式，当开启传输无损能力并以 10Gbps 带宽运行时，每 token 平均输出延迟增加约 2.75ms，吞吐劣化 0.21%，整体性能平稳；当带宽从 10Gbps 降至 3Gbps，算力利用率从 99.79% 降至 86.47%，集群服务始终处于可用状态，且带宽保持在 5Gbps 及以上时，算力利用率均维持在 95% 以上的水平。对比之下，场景四在关闭传输无损能力后，性能边界明显收窄，10Gbps 带宽下延迟增加约 2.76ms，与场景三持

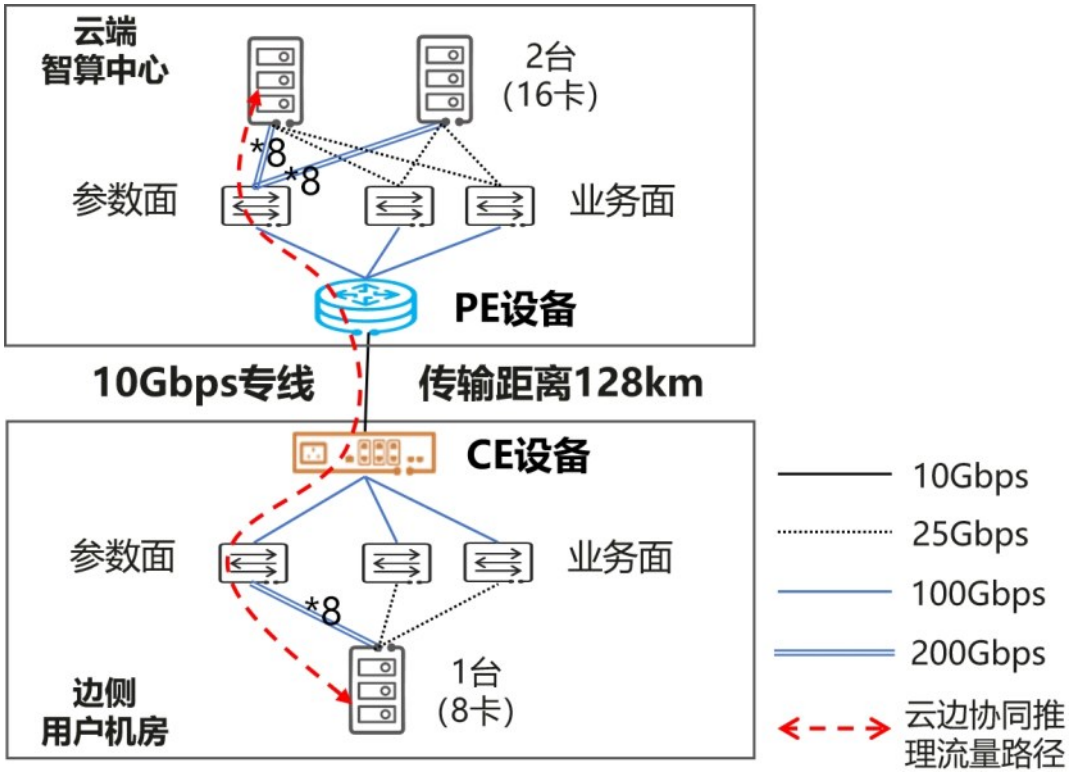

图 5　云边协同推理验证方案组网示意图

表1　测试场景

测试场景	部署架构	卡数配置	拉远距离	网络配置	带宽及收敛比
场景一（基线）	云端集中+模型不分层部署	16卡	无	本地无损网络	1600Gbps，1:1
场景二	云端集中+模型分层部署				
场景三	云边协同+模型分层部署	云端8卡+边侧8卡	128km	远距无损传输	10Gbps，1:160；8Gbps，1:200；6Gbps，1:267；5Gbps，1:320；4Gbps，1:400；3Gbps，1:533
场景四				远距有损传输	

表2　实际测试结果

场景	测试带宽	每token平均输出延迟（TPOT/ms）	系统吞吐量（tokens/s）	劣化情况	
				延迟	吞吐
场景一	10Gbps	82.713	1243.1603	N/A（基线标准）	
场景二	10Gbps	81.4193	1242.1673	-1.56%	0.08%
场景三	10Gbps	85.4672	1240.5241	3.33%	0.21%
	8Gbps	86.9952	1229.8565	5.18%	1.07%
	6Gbps	89.439	1201.1893	8.13%	3.38%
	5Gbps	89.8949	1181.7727	8.68%	4.94%
	4Gbps	90.0335	1141.8024	8.85%	8.15%
	3Gbps	94.5768	1074.953	14.34%	13.53%
场景四	10Gbps	85.5715	1224.4906	3.46%	1.50%
	8Gbps	124.7081	863.7576	50.77%	30.52%
	6Gbps	124.7226	740.3076	50.79%	40.45%
	5Gbps	推理中断			

平，但吞吐劣化升至1.5%；当带宽降至8Gbps，算力利用率骤降至69.48%；当带宽降至5Gbps，推理中断，集群服务完全不可用。经分析，在关闭传输无损能力的情况下，随着链路带宽持续降低，网络拥塞程度逐渐加剧，数据丢包概率显著增加，由于RDMA协议对网络传输质量较为敏感，丢包需通过重传机制恢复丢失数据，频繁重传则导致转发效率大幅下降，增加端到端通信时延，当重传恢复时间超过应用设定的阈值时，RDMA连接失败，从而导致推理任务中断。

经上述测试验证，云边协同分布式推理架构具备工程可行性和实际应用价值，且证实在远距传输保障无损特性且带宽充足的前提下，可以有效抑制长尾延迟的恶性波动，从而确保业务时延指标与算力利用率的双重稳定。

4.3　测试范围及适用性说明

需要说明的是，本次测试以Qwen2.5-32B模型为研究对象，在单一用户环境下开展验证，主要目的是验证模型分层部署策略在云边协同推理业务场景下的可行性。从方法机理来看，模型分层部署策略主要依赖于Transformer类大语言模型普遍具有的层级串行结构，而非特定模型的网络实现，因此对于具有相似架构的大语言模型具有一定参考价值。然而，不同模型在参数量、层级结构等方面存在差异，具体性能劣化情况有待进一步验证分析。此外，本次测试环境未涉及多租户共享网络资源的场景，因此未启用面向多租户场景下的精准流控和流级负载均衡功能，不影响

本文对分层部署策略可行性的评估。但在实际业务环境中，多租户部署对云边协同整体性能的影响仍值得进一步研究。

未来工作将结合多种模型、多租户场景开展更全面的验证，以进一步评估所提方案的普适性。

5　结束语

综上，随着企业智能化转型的加速推进，大模型训练与推理在算力建设与数据安全之间的矛盾日益凸显。本文提出的云边协同分布式训推方案，以模型分层部署为核心机制，即企业侧专责敏感数据的采集与处理，云端智算中心承载大规模计算任务，云边之间仅交换敏感性更低的中间向量，从而在确保数据安全与合规的同时，实现算力的弹性扩展，有效缓解了自建算力成本高、租赁算力存在数据出域风险等行业用算问题。

网络架构依托高性能网络与算力调度扩展机制，省内和跨省采用不同组网方案，能够在多地域、多场景环境下高效支撑大模型训推业务的云边分布式协同，为企业提供安全高效、灵活可扩展的算力服务。展望未来，本文提出的云边协同分布式训推创新解决方案，有望为不同行业的智能化转型升级提供可推广的参考路径，并推动算力网络与人工智能应用的深度融合发展。

参考文献：

[1] Li Z,Wu S,Li L,et al.Energy-Efficient Split Learning for Fine-Tuning Large Language Models in Edge Networks[J]. IEEE Networking Letters, 2025, 7(3): 176-180. DOI: 10.1109/LNET.2025.3530430.

[2] Yang Q,Dong Z,Liang S,et al.Research on Data Privacy and Information Security Evaluation and Optimization for Large Language Models[C]. 2025 6th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), October17-19 2025, Shanghai, China: IEEE, 2025. DOI:10.1109/ICBAIE66852.2025.11326645.

[3] Yang Z,Jin Y,Zhang Y,et al.Research on Large Language Model Cross-Cloud Privacy Protection and Collaborative Training Based on Federated Learning[C]. 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), April 11-13 2025, Shenzhen, China: IEEE, 2025.DOI:10.1109/AINIT65432.2025.11035133.

[4] Internet Engineering Task Force (IETF) . OSPFv3 as a Provider Edge to Customer Edge (PE-CE) Routing Protocol:ISSN:2070-1721[S]. 2012.

[5] Guo C,Wu H,Deng Z,et al.RDMA over commodity ethernet at scale[C]. SIGCOMM '16: ACM SIGCOMM 2016 Conference, August 22-26 2016, Florianopolis, Brazil: Association for Computing Machinery,New York,United States,2016:202-215.ISBN: 978-1-4503-4193-6.

[6] Catherine J. Doughty, Michael H. Long.The Handbook of Second Language Acquisition[M]. Blackwell Publishing Ltd, 2003. Print ISBN: 9780631217541, Online ISBN: 9780470756492, DOI:10.1002/9780470756492.

[7] Zhao W,Zhou K,Li J,et al.A Survey of Large Language Models [EB/OL]. (2023-03-31)[2026-3-26]. arXiv:2303.18223v19 [cs.CL] .https://arxiv.org/abs/2303.18223.

[8] Naveed H,Khan A,Qiu S,et al.A Comprehensive Overview of Large Language Models[J]. ACM Transactions on Intelligent Systems and Technology,2025,16(5):1-72.

[9] Shanahan M.Talking about Large Language Models[J].Communications of the ACM,2024,67(2):68-79.

[10] Subakan C, Ravanelli M, Cornell S, et al. Attention Is All You Need In Speech Separation[C]. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),June 06-11 2021,Toronto, ON, Canada:IEEE, 2021.DOI:10.1109/ICASSP39728.2021.9413901

[11] Lei Y,Shen T,Cao Y,et al.Enhancing Lexicon-Based Text Embeddings with Large Language Models[C]. 63rd Annual Meeting of the Association for Computational Linguistics,July 2025, Vienna, Austria. ACL: Association for Computational Linguistics, 2025:18986-19001. DOI:10.18653/v1/2025.acl-long.930

[12] Matsuda K,Frohlich S,Wang M,et al.Embedding by Unembedding[J]. Proceedings of the ACM on Programming Languages, 2023,7(189):1-47.

[13] M S,Kiruba G,J S,et al.Large Language Model AI Text Generation Detection based on Transformer Deep Fast Quantum Convolutional Neural Networks[C]2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA),June 25-27 2025,Coimbatore, India.IEEE,2025.Electronic ISBN:979-8-3315-2142-4.

[14] Vig J,Belinkov.Analyzing the Structure of Attention in a Transformer Language Model[C]Proceedings of the 2019 ACL Work-

shop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, August 2019, Florence, Italy. Association for Computational Linguistics,2019:63-76.

[15] Yao D, Li B. Is Split Learning Privacy-Preserving for Fine-Tuning Large Language Models?[C].IEEE Transactions on Big Data （Early Access）, 2024: 1-12. DOI: 10.1109/TBDATA.2024.3524101.

[16] Qu W, Zhou Y, Wu Y,et al.Prompt Inversion Attack against Collaborative Inference of Large Language Models[C].2025 IEEE Symposium on Security and Privacy（SP）, May 12-15 2025, San Francisco CA USA. IEEE, 2025. DOI: 10.1109/SP61157.2025.00160.

[作者简介]

李志勇（1986-），男，硕士研究生，中国移动通信集团高级工程师，主要研究方向为云计算、智算中心等算力基础设施技术及工程方案。邮箱：lizhiyongjh@chinamobile.com。手机号：13810256181。

郗卓宁（1980-），男，工程硕士，中国移动通信集团高级工程师，主要研究方向为云计算、算力网络及网络安全。邮箱：xizhuoning@chinamobile.com。手机号：13910232266。

田雯（1987-），女，硕士研究生，中国移动通信集团设计院有限公司高级工程师，高级研究专员，主要研究方向为云计算、智算中心等算力资源的规划咨询及工程方案设计。邮箱：tianwen@cmdi.chinamobile.com。手机号：15810490675。

张明（1991-），男，硕士，主要研究方向为云计算、算力网络。邮箱：zhangmingcg@chinamobile.com。手机号：13810228117。

郑坤松（1998-），男，硕士研究生，中国移动通信集团设计院有限公司助理研究专员，助理工程师，主要研究方向为云计算、智算中心等算力基础设施工程方案设计。邮箱：zhengkunsong@cmdi.chinamobile.com。手机号：15652091288。

曹原铭（1983- ），男，硕士研究生，中国移动通信集团设计院有限公司高级工程师，研究总监，主要研究方向为云计算、人工智能算力基础设施相关。邮箱：caoyuanming@cmdi.chinamobile.com。手机号：13811939667。

孙傲（1998- ），男，硕士研究生，中国移动通信集团设计院有限公司开发专员，助理工程师，主

要研究方向智算工程方案设计，人工智能前沿技术跟踪。邮箱：sunao@cmdi.chinamobile.com。手机号：15910595657。

武振宇（1975- ），男，硕士研究生，中国移动通信集团设计院有限公司高级工程师，研究总监，

主要研究方向为云计算、算力网络、智算中心等。邮箱：wuzhenyu@cmdi.chinamobile.com。手机号：13810020765。

丁国强（1986- ），男，硕士研究生，中国移动通信集团设计院有限公司高级研究专员，高级工程师，主要研究方向为云计算和人工智能技术等。邮箱：dingguoqiang@cmdi.chinamobile.com。手机号：

18301298801。

魏逸哲（1999- ），男，本科，中国移动通信集团设计院有限公司工程设计师，助理工程师，主要

研究方向为云计算、智算中心基础设施相关。邮箱：weiyizhe@cmdi.chinamobile.com。手机号：15810162852。