



中文核心期刊 中国科技核心期刊 中国核心学术期刊
第三届全国期刊奖百种重点期刊 信息通信领域产学研合作特色期刊

ISSN 1009-6868
CN 34-1228/TN

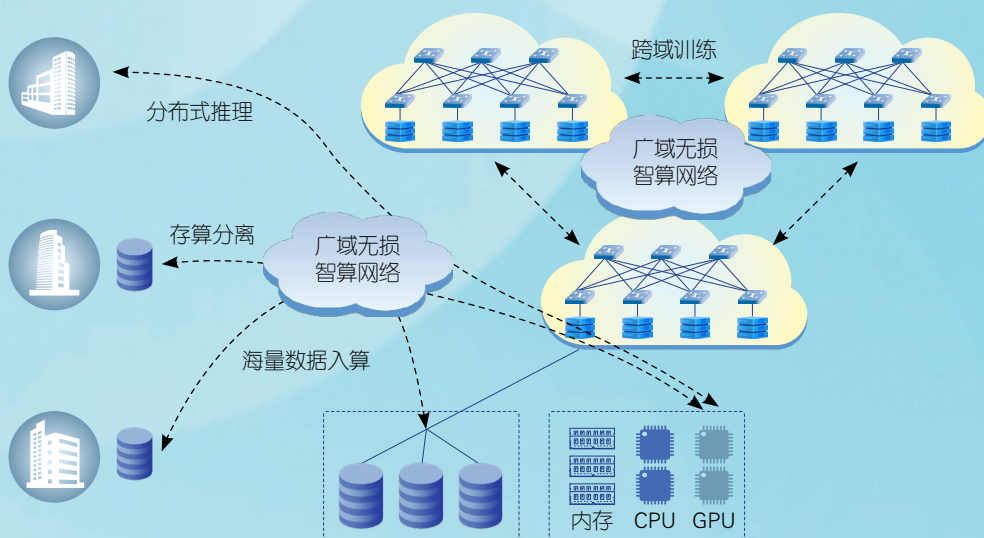
中兴通讯技术

ZTE TECHNOLOGY JOURNAL

<http://tech.zte.com.cn>

第 32 卷 · 总第 189 期 · 2026 年 6 月 · 第 3 期

专题：智算网络



(封面图片详解见 P.11)

ISSN 1009-6868



第二十八届中国科协年会学术论文专刊

《中兴通讯技术》第10届编辑委员会

顾问 侯为贵(中兴通讯股份有限公司创始人) 钟义信(北京邮电大学教授)
糜正琨(南京邮电大学教授) 李自学(中兴通讯股份有限公司前董事长)

主任 陆建华(中国科学院院士)

副主任 方 榕(中兴通讯股份有限公司董事长) 李建东(西安电子科技大学教授)

编委

陈建平	上海交通大学教授	唐万斌	电子科技大学教授
陈前斌	重庆邮电大学教授、副校长	唐雄燕	中国联通研究院副院长、首席科学家
段晓东	中国移动研究院副院长	陶小峰	北京邮电大学教授
方 榕	中兴通讯股份有限公司董事长	汪烈军	新疆大学教授、副校长
高新波	西安电子科技大学教授、校长	王 翔	中兴通讯股份有限公司高级副总裁
葛建华	西安电子科技大学教授	王文博	雄安空天信息研究院院长
管海兵	上海交通大学教授、副校长	王文东	北京邮电大学教授
郭 庆	哈尔滨工业大学教授	王喜瑜	中兴通讯股份有限公司执行副总裁
洪 伟	中国科学院院士、东南大学教授	王耀南	中国工程院院士、湖南大学教授
江 涛	华中科技大学教授	王志勤	中国信息通信研究院副院长
蒋林涛	中国信息通信研究院科技委主任	卫 国	中国科学技术大学教授
金 石	东南大学教授、副校长	邬贺铨	中国工程院院士
李尔平	中国工程院外籍院士、浙江大学教授	吴春明	浙江大学教授
李红滨	北京大学教授	向际鹰	中兴通讯股份有限公司首席科学家
李厚强	中国科学技术大学教授	肖 甫	南京邮电大学教授、副校长
李建东	西安电子科技大学教授	解冲锋	中国电信新一代信息通信专业首席专家
李乐民	中国工程院院士、电子科技大学教授	徐安士	北京大学教授
李融林	华南理工大学教授	徐子阳	中兴通讯股份有限公司总裁
林晓东	中兴通讯股份有限公司副总裁	续合元	中国信息通信研究院首席专家
刘 健	中兴通讯股份有限公司高级副总裁	薛向阳	复旦大学教授
刘建伟	北京航空航天大学教授	杨义先	北京邮电大学教授
隆克平	北京科技大学教授	易芝玲	中国移动研究院首席科学家
卢光跃	西安邮电大学教授、校长	张 杰	内蒙古工业大学教授、学术副校长
陆建华	中国科学院院士、清华大学教授	张 平	中国工程院院士、北京邮电大学教授
马建国	中原工学院教授、学术副校长	张 卫	复旦大学教授
毛军发	中国科学院院士、深圳大学教授	张宏科	中国工程院院士、北京交通大学教授
孟洛明	北京邮电大学教授	张钦宇	哈尔滨工业大学(深圳)教授、副校长
尼玛扎西	中国工程院院士、西藏大学教授	张云勇	中国联通网络信息安全部总经理
石光明	鹏城实验室副主任	赵慧玲	工业和信息化部信息通信科技委常委
史振威	内蒙古大学教授	郑纬民	中国工程院院士、清华大学教授
苏 森	重庆邮电大学教授、校长	钟章队	北京交通大学教授
孙知信	南京邮电大学教授	周 亮	南京邮电大学教授、副校长
谈振辉	北京交通大学教授	朱近康	中国科学技术大学教授
唐 宏	中国电信 IP 专业首席专家	祝宁华	中国科学院院士、南开大学教授

目次

中兴通讯技术 (ZHONGXING TONGXUN JISHU)
第 32 卷 总第 189 期 2026 年 6 月 第 3 期

中文核心期刊 中国科技核心期刊 第三届国家期刊奖百种重点期刊 信息通信领域产学研合作特色期刊 中国知网、万方数据、重庆维普等数据库收录期刊 1995 年创刊

热点专题 ▶

智算网络

01	专题导读	赵慧玲
03	长距高性能网络关键技术研究	王亚晨, 杨峰, 耿竞一
10	广域无损智算网络关键技术研究与实践	雷波, 潘永琛, 唐静
15	Scale-Up 网络内 GPU 卡间互联及 OISA 技术体系	宋晓勇, 李锴, 陈佳媛
22	智算中心光电混合组网下的流量差异化调度技术研究	韩博文, 刘千仞, 廖思忆, 胡雅坤, 曹畅
30	面向边缘场景的高效大模型分层部署机制	万卓然, 谢人超, 王媛, 王芳, 唐琴琴, 黄韬
37	Token 原生的 Scale Across 网络	苏远超
40	高性能智算网络架构探讨与实践	黄光平, 肖敏, 张征
45	面向分布式推理的大模型键值缓存分发网络	宋俊平, 黄天一, 周旭, 李智, 吴茜
51	语义通信技术商用化之路	石光明, 杨旻曦, 高大化, 肖泳
60	大规模智算互联网络仿真技术和平台	熊先奎, 童贤慧, 王勇, 张启明
69	基于特征事件的 PON 网络故障分析诊断与应用	余辰东, 孙志
76	基于 5G 地面公网的网联无人机低空网络覆盖	王俊杰, 王迎, 申奇, 肖腾飞

名家视点 ▶

企业视界 ▶

技术广角 ▶

《中兴通讯技术》2026 年热点专题名称及策划人

1. 6G 关键技术的标准化: Day-1 与未来 中国移动研究院首席科学家 易芝玲	3. 智算网络 工信部信息通信科技委常委 赵慧玲	5. 星地太赫兹高速传输技术 中国科学院院士、东南大学教授 洪伟 电子科技大学教授 唐万斌 东南大学教授 郝张成
2. 广域立体覆盖低空通信技术 东南大学副校长 金石 东南大学教授 刘凡	4. 大模型推理中的存算技术 中国工程院院士、清华大学教授 郑纬民 清华大学副教授 陆游游	6. 智能多天线技术 北京交通大学副校长 艾渤 北京交通大学教授 章嘉懿

MAIN CONTENTS

ZTE TECHNOLOGY JOURNAL
Vol. 32 No. 3 Jun. 2026

Special Topic ►	Intelligent Computing Networks
	01 Editorial Zhao Huiling
	03 Research on Key Technologies for Long-Distance High-Performance Networks Wang Yachen, Yang Feng, Geng Jingyi
	10 Research and Practice of Wide-Area Lossless Intelligent Computing Network Key Technologies Lei Bo, Pan Yongchen, Tang Jing
	15 Intra-Node GPU Interconnection in Scale-Up Networks and OISA Technologies Song Xiaoyong, Li Kai, Chen Jiayuan
	22 Differentiated Traffic Scheduling for Hybrid Optical-Electrical Networking in Intelligent Computing Centers Han Bowen, Liu Qianren, Liao Siyi, Hu Yakun, Cao Chang
	30 An Efficient Hierarchical Deployment Mechanism for Large Models in Edge Scenarios Wan Zhuoran, Xie Renchao, Wang Yuan, Wang Fang, Tang Qinqin, Huang Tao
	37 Token Native Scale Across Network Su Yuanchao
	40 High-Performance AI Computing Network Architecture: Exploration and Practice Huang Guangping, Xiao Min, Zhang Zheng
	45 Large Model Key-Value Cache Distribution Network for Distributed Inference Services Song Junping, Huang Tianyi, Zhou Xu, Li Zhi, Wu Qian
Expert View ►	51 Road to Commercialization of Semantic Communication Technology Shi Guangming, Yang Minxi, Gao Dahua, Xiao Yong
Enterprise View ►	60 Large-Scale Intelligent Computing Interconnection Network Simulation Technology and Platform Xiong Xiankui, Tong Xianhui, Wang Yong, Zhang Qiming
Research Papers ►	69 PON Network Fault Analysis, Diagnosis and Application Based on Feature Events Yu Chendong, Sun Zhi
	76 Low-Altitude Network Coverage for Connected UAVs Based on 5G Terrestrial Public Networks Wang Junjie, Wang Di, Shen Qi, Xiao Tengfei

期刊基本参数: CN 34-1228/TN*1995*b*16*82*zh*P*¥20.00*6500*13*2026-06

敬告读者	本刊享有所发表文章的版权, 包括英文版、电子版、网络版和优先数字出版版权, 所支付的稿酬已经包含上述各版本的费用。 未经本刊许可, 不得以任何形式全文转载本刊内容; 如部分引用本刊内容, 须注明该内容出自本刊。
------	--

智算网络专题导读



专题策划人



赵慧玲

工业和信息化部信息通信科技委常委、中国通信学会理事、中国通信标准化协会（CCSA）网络与业务能力技术工作委员会（TC3）主席、中国通信学会标准工作委员会副主任委员、中国电信科技委常委；长期从事电信网络领域的技术与标准工作；曾荣获多项国家及省部级科技进步奖；已发表论文 100 余篇，出版技术专著 12 部。

智算网络作为算力高效调度与资源协同联动的核心载体，直接决定了算力资源的利用效率，是支撑 AI 大模型训练及“东数西算”工程落地的关键基础设施。当前，智算网络面临的核心技术挑战主要集中在 4 个方面：一是万卡级集群互联扩展性不足，传统网络架构难以支撑高带宽、低时延、无阻塞的大规模组网需求；二是广域无损传输难以保障，远程直接内存访问（RDMA）对丢包极为敏感，在长距场景下，现有拥塞控制与流控机制往往失效；三是大模型流量特征与调度策略不匹配，低熵大象流以及点对多点（P2MP）、全互联（All-to-All）等通信模式易导致负载不均、资源利用率低下；四是存算网边全栈协同缺失，键值缓存（KV-Cache）复用效率偏低，边缘侧资源受限，异构算力与光电混合组网缺乏统一的调度机制。本期专刊特邀互联网和电信运营商、科研院所及高校的专家，聚焦上述核心技术瓶颈，撰写了 8 篇高质量研究成果。文章既关注技术突破，也注重实践落地，系统梳理了行业最新进展，以期为业界同仁提供切实有益的参考。

针对 AI 大模型跨园区训练中的长距组网需求，《长距高性能网络关键技术研究》一文聚焦长距传输时延激增、拥塞反馈不及时及端口负载不均等关键问题，提出了一种端到端

联合优化方案。方案采用分级流控技术，以灵活适配数据中心内部流量与跨园区流量的差异化带宽需求；同时设计了基于单跳信令通道的快速拥塞反馈机制，显著降低反馈延时。实验结果表明，该方案将跨园区算力损耗从 10%~15% 有效控制在 5% 以内，实测效果显著。

面向广域智算网络中 RDMA 传输易丢包、拥塞控制滞后等挑战，《广域无损智算网络关键技术研究与实践》提出了一套广域无损智算网络技术方案。方案通过分段反压精准流控与大缓存增强，实现零丢包传输；并引入基于 IPv6 的段路由（SRv6）切片技术，实现租户级隔离，有效解决优先级流控制（PFC）拥塞扩散问题。文章基于现网实际验证，在云边协同推理和存算分离训练等场景下均能保持高算效，技术路线清晰、工程落地性强，为广域高性能智算组网提供了重要参考。

聚焦于 AI 大模型产业化落地中的算力互联关键问题，《Scale-Up 网络内 GPU 卡间互联及 OISA 技术体系》一文系统梳理了 Scale-Up 卡间互联技术的演进历程与产业趋势，对比分析了私有协议与开放标准的优劣。文章重点介绍了开放互联架构（OISA），从带宽、时延、扩展性及生态兼容性等产业视角，阐述了其在超节点部署、多图形处理器（GPU）协同、大模型训练与推理等场景中的实用价值。该文紧贴产业痛点，兼顾技术落地与生态构建，对智算中心规划、硬件选型与集群建设具有重要指导意义。

针对传统纯电交换网络在带宽、时延、能耗方面存在的

DOI: 10.12142/ZTETJ.202603001
收稿日期: 2026-05-14

瓶颈,《智算中心光电混合组网下的流量差异化调度技术研究》充分利用光电混合组网中“电交换灵活可控、光交换高带宽低时延”的互补优势,构建了“感知—决策—执行—反馈”闭环调度框架。该框架可实现AI训练中数据并行(DP)与流水线并行(PP)流量的精准识别及链路匹配,核心策略是将DP流量优先调度至光链路,而将PP流量动态适配于光或电链路,从而有效提升算力释放效率与链路利用率。

针对边缘大模型推理中显存受限、输入输出(I/O)阻塞、垃圾回收(GC)抖动以及调度粒度过粗等四大瓶颈,《面向边缘场景的高效大模型分层部署机制》提出了Triad-Flow轻量化部署方案。方案采用内存映射(mmap)按需加载技术,破除权重I/O瓶颈;通过静态显存池与指针复用机制,消除运行时抖动;并基于权重矩阵级流水线实现计算与加载的完全重叠。在不增加显存占用的前提下,推理吞吐量提升超过一倍,时延显著降低。

《Token原生的Scale Across网络》一文针对AI时代广域网络互联面临的大规模跨域算力调度挑战,提出以Token为核心调度单元,重构跨域算力互联体系。通过模型与基础设施协同设计、每作业流量工程(Per-job TE)细粒度流量调度、统一IPv6及SRv6寻址,并结合大带宽扁平组网、快速拥塞通知(CNP)、报文裁剪以及链路检测与路由检测(LD/RD)快速倒换等机制,实现了容量扩展能力提升100倍、收敛时间缩短至原来的1/10、成本降低50%的显著效果,为跨域训练推理、存算分离及多可用区(AZ)分布式训练提

供了高性能网络底座。

针对智算网络对高吞吐、低时延及无损传输的严苛需求,《高性能智算网络架构探讨与实践》一文提出了一种端网协同、精准流控、无状态组播与轻量级在网计算一体化的架构方案。具体而言,通过端网速率协商与配额调度,实现主动拥塞规避;基于SRv6开展切片级精准流控;采用比特索引显式复制(BIER)无状态组播技术,优化混合专家模型(MoE)中点对多点(P2MP)通信效率;并利用在网聚合降低多点对一点(MP2P)回传开销。原型验证结果表明,该方案显著消除了调速锯齿现象,提升了传输通量,且兼具标准兼容性与工程落地性。

针对分布式大模型推理中KV Cache复用率低、通信冗余及延迟偏高等瓶颈,《面向分布式推理的大模型键值缓存分发网络》一文提出了键值缓存分发网络(KVCDN),将传统内容分发网络(CDN)升级为张量感知、状态管理的智能缓存架构。方案采用端侧语义规范化、边缘分级缓存与核心全局去重三层设计,通过提示词(Prompt)标准化、语义路由、前缀去重及热点预放置等机制,实现高维张量的高效复用。该方案有效降低了冗余计算与跨网传输开销,为云边协同推理提供了高性能底座。

本次专刊收录的研究成果均立足当前实际技术痛点,为智算网络的技术突破与产业落地提供了重要参考。期望以本专刊为载体,促进业界同仁在智算网络领域深化技术交流与合作,并衷心感谢各位作者的辛勤付出与创新贡献。

长距高性能网络关键技术研究



Research on Key Technologies for Long-Distance High-Performance Networks

王亚晨/Wang Yachen, 杨峰/Yang Feng, 耿竞一/Geng Jingyi

(腾讯科技(深圳)有限公司, 中国 深圳 518038)

(Tencent Technology (Shenzhen) Co., Ltd., Shenzhen 518038, China)

DOI: 10.12142/ZTETJ.202603002

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260626.0951.002>

网络出版日期: 2026-06-26

收稿日期: 2026-05-10

摘要: 针对AI大模型训练规模突破万卡、需跨园区长距组网的新需求, 提出了一套端到端的联合优化解决方案: 分级流控技术将任务完成时间(JCT)降低一个数量级; 快速拥塞反馈机制将拥塞反馈延迟从毫秒压缩至数十微秒; 逐流负载均衡协议消除端口负载不均问题; 采用拓扑亲和的ZeRO 1策略替代ZeRO 3, 大幅减少跨园区通信次数。实验网测试表明, 通过整合上述关键技术, 在千亿参数模型训练场景下可将跨园区训练的算力损失从10%~15%降低至5%以内, 验证了长距网络支撑超大规模AI训练的可行性。

关键词: 长距高性能组网; 大模型训练; 拥塞控制; 负载均衡

Abstract: This paper proposes an end-to-end solution to address the new requirements of AI large-model training at the scale of over ten thousand GPUs and the need for long-distance, cross-campus networking. The hierarchical flow control reduces job completion time (JCT) by an order of magnitude. The fast congestion feedback mechanism compresses congestion feedback latency from milliseconds to tens of microseconds. Per-flow load balancing protocol eliminates port load imbalance. The topology-aware ZeRO 1 strategy instead of ZeRO 3 can greatly reduce the number of cross-campus communications. Testbed experiments show that by integrating these key techniques, in the training of models with hundreds of billions of parameters, the compute efficiency loss of cross-campus training was reduced from 10%–15% to less than 5%, validating the feasibility of long-distance networks in supporting ultra-large-scale AI training.

Keywords: long-distance high-performance networking; large-model training; congestion control; load balancing

引用格式: 王亚晨, 杨峰, 耿竞一. 长距高性能网络关键技术研究 [J]. 中兴通讯技术, 2026, 32(3): 3–9. DOI: 10.12142/ZTETJ.202603002

Citation: Wang Y C, Yang F, Geng J Y. Research on key technologies for long-distance high-performance networks [J]. ZTE technology journal, 2026, 32(3): 3–9. DOI: 10.12142/ZTETJ.202603002

1 跨园区训练的需求

2022年ChatGPT的发布掀起了全球生成式人工智能(AIGC)热潮, 以大语言模型(LLM)为代表的AIGC演变为新一轮全球竞争焦点。在LLM的演进历程中, 模型参数量与算力成为驱动模型能力跃迁的关键因素, 三者的关系被描述成“规模定律”, 即Scaling Law。以OpenAI的GPT系列为例, 从GPT-3到GPT-4, 参数量从1750亿(175B)提升至约1.8万亿(1.8T), 增长了约10倍; 训练数据量则从0.3万亿Token激增至13万亿Token, 增长了约42倍; 配套使用的图形处理器(GPU)卡数量也从千卡规模增长到万卡规模。最新一代的GPT-5更是使用了10万张H100 GPU。

随着大模型训练规模的增长, 从单个园区获取全量GPU卡面临较大的挑战。这些挑战包括两个方面:

1) 电力瓶颈。20万张H100集群总功耗直逼吉瓦级。

单一数据中心受限于供电容量、散热成本及土地资源, 无法承载超大规模GPU部署^[1]。例如, Meta训练Llama 4需要10万卡级集群, 被迫将算力分散至多栋楼宇, 通过聚合层交换机(ATSW)实现跨楼互联^[2]。

2) GPU碎片化问题。由于业务发展对GPU卡有强烈需求, 一个新建成的GPU模块一般会交付多个业务部门。这导致一个业务部门的卡分散在多个园区。虽然通过置换卡可以延缓碎片的产生, 但是仍存在停卡、沟通成本居高不下, 以及最终换无可换等一系列问题, 因此亟需配套技术来整合GPU碎片, 满足百卡规模的训练需求^[3]。

腾讯在2024年搭建了一张长距高性能实验网, 以满足业务探索多模态混训的需求。本文基于该实验网的测试数据和为解决现网问题而研发的原型样机, 阐述长距高性能网络的挑战、关键技术以及性能评估结果。

2 长距高性能网络的挑战

如图1所示,实验网使用了百公里光纤,400G级大端口、数据中心路由器(DR)连接了两个GPU园区。为了降低园区内产生拥塞的概率,本文规划了16:1的收敛比(DR-核心交换机(GPULC)互联带宽是DR-DR带宽的16倍),使出园区的流量能够“一路畅通”地抵达DR,将最为重要的拥塞控制问题交给DR处理,最大程度地降低跨园区流量对园区内业务的干扰。

在未优化的长距实验网上进行集合通信基准测试(NCCLTest)和7B-dense/混合专家(MoE)训练测试后,本研究发现几乎所有的性能指标相对于数据中心内部(DCN)基线都出现了明显下降。实验表明,长距离高性能组网存在以下4个方面的挑战:

1) 消息传输延迟显著增加导致数据传输欠吞吐

在NCCLTest的SendReceive测试中,本研究发现任务完成时间(JCT)出现了10~100倍的显著增长。受物理距离约束的固有传播延迟大幅提升了端到端往返时延(RTT),导致网络带宽-时延积(BDP)显著增加。在DCN窗口机制下,拥塞窗口(cwnd)规模无法匹配膨胀的BDP需求,形成传输瓶颈。这迫使数据传输长期处于欠吞吐状态,无论是流水线并行(PP)下的激活值传输这类小流,还是数据并行(DP)下的梯度同步这类大流,任务完成时间均出现数量级增长。

2) 拥塞反馈不及时导致拥塞丢包和欠吞吐

在perftest中,本研究发现当前实验网存在丢包问题,链路利用率不到理论带宽的80%。这主要由拥塞控制算法数据中心量化拥塞通知(DCQCN)在长距控制器环路下响

应失效引发:当传输距离达到百公里时,拥塞状态反馈产生1 ms以上的滞后,导致端侧无法及时响应动态变化。这种延迟一方面会使得交换机缓存溢出丢包,另一方面会使得发射端因过度降速而出现欠吞吐,导致链路的实际吞吐低于有效带宽理论值。

3) 随机哈希导致端口负载不均和流完成时间增加

根据大数定理,当流数量足够多时,随机等概的出端口调度机制可以实现交换机端口间理想的负载均衡。传统数据中心互联(DCI)流数通常在 $O(10^6)$ 量级,因此DCI交换机通常采用基于五元组的哈希出端口调度方法。LLM训练任务产生的跨园区流(即五元组)数量通常为 $O(10^3) \sim O(10^4)$,呈现所谓的低熵特征。这导致基于随机哈希的出端口调度机制出现显著的端口负载不均衡问题。

7B-dense的小规模测试显示,部分端口的负载超出平均值的20%以上,这意味着高性能DCI需要更精细化的负载均衡控制机制。

4) 显存优化技术(ZeRO 3)导致大量小消息数据跨园区通信

跨园区的算力损失主要取决于跨园区的通信时间。当带宽为非瓶颈因素时,跨园区的通信时间主要取决于跨园区的通信次数。图2展示了跨园区采用的PP分割和DP分割方案。由于DP跨园区通信次数少,所以通常认为跨园区使用DP分割算力损失更小。然而,7B-dense小规模测试发现,DP分割出现了大量小消息数据跨园区通信,实测算力损失高达10%~15%。因此,跨园区训练需要使用拓扑感知的训练框架。

针对上述挑战,本文分别从长距流控技术、快速拥塞反馈、逐流负载均衡和训练框架优化4个方面提出对应的解决方案。

3 长距高性能网络关键技术

3.1 长距流控技术

DCN的RTT只有10 μ s量级,对应的BDP为1 Mbit。因此,DCN通常采用1 Mbit左右的发送窗口。这样既可以充分利用带宽,又可以减少拥塞。而跨DCI场景下,RTT扩大100倍至1 ms量级,对应的BDP增加到100 Mbit。此时如果仍使用1 Mbit的小窗口,发送端会长时间等待接收端的确认而不发送数据,导致较为严重的链路欠吞吐问题。

如图3所示,为了提升长距链路的利用率,本文使用分级的流控方案:对DCN的流,采用小窗口;对跨DCI的流,则使用大窗口,以应对大RTT造成的确认延迟。这一方案需

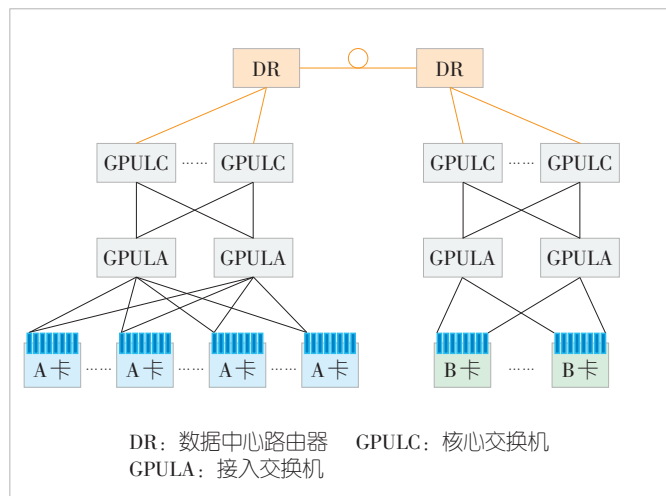


图1 长距高性能实验网架构

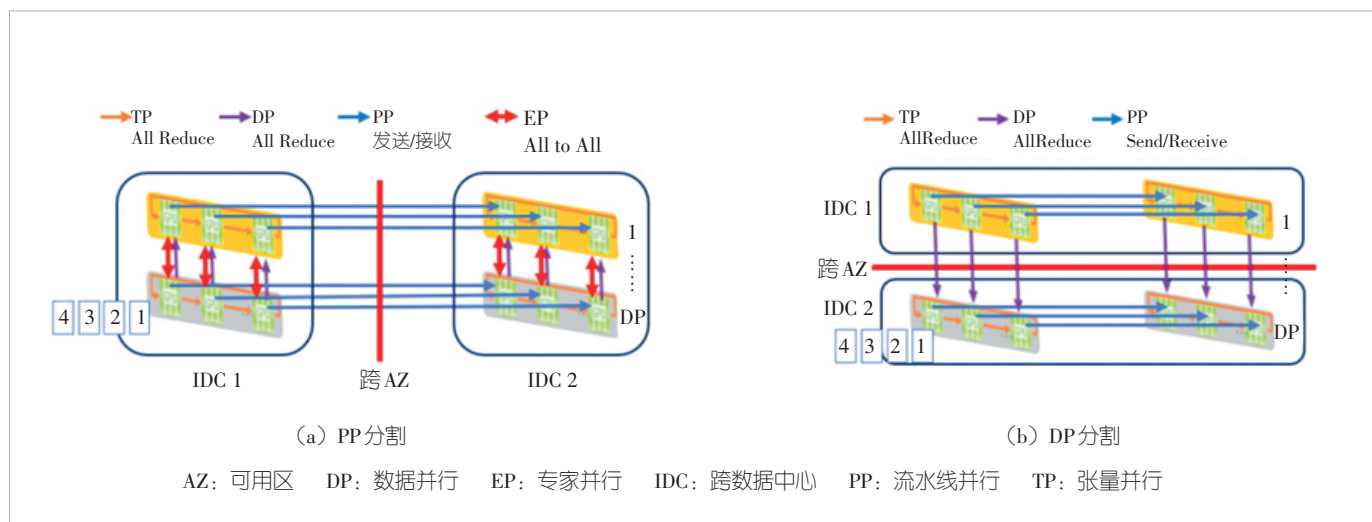


图2 跨园区PP分割和跨园区DP分割

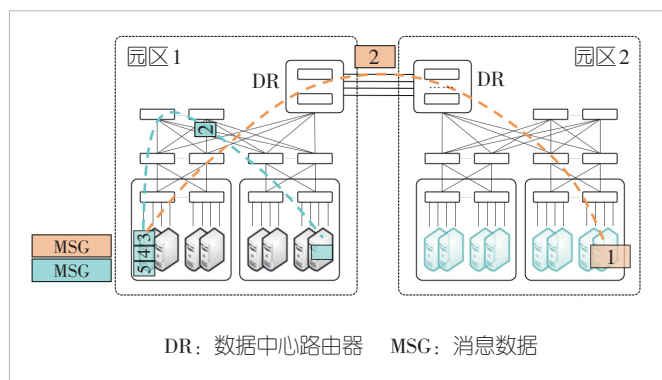


图3 分级流控技术

要配合大容量缓冲区和快速拥塞控制算法，以便动态管理拥塞风险，确保数据传输的可靠性。对于跨DCI的流，发送窗口的大小接近BDP，即以链路带宽和传播延迟的乘积为基础，最大化长距传输的吞吐效率。这种分级方法基于网络拓扑的物理特性，实现了端到端流控的自适应，有效提升了分布式训练的整体性能和长距链路利用率。

3.2 快速拥塞反馈

本文使用快速拥塞反馈技术解决传统端到端拥塞通知机制在长距场景的反馈延迟问题。该技术将传统显式拥塞通知（ECN）的“交换机标记→接收端回传→发送端响应”三级延迟链路，缩短为交换机直连发送端的单跳拥塞通知包（CNP）信令通道以实现近端反馈。

快速拥塞反馈技术的难点包括：

- CNP报文字段队列对编号（QPN）的获取。根据标准^[4]规定，通信双方在建链的时候交换QPN信息，之后的通信中使用对方分配的QPN来标识该连接，因此CNP的QPN无法从数据报文解析中得到。

- CNP报文的实时性。有厂商曾尝试使用CPU产生CNP报文，由于软件难以保证实时性，CNP的延迟抖动超过了1 ms，完全无法满足快速拥塞反馈的要求。

为支持快速拥塞反馈，本文对终端和交换机均进行了“技术改造”。

为了让交换机可以获取QPN，本文采用了基于通信管理（CM）来建立通信链路，其整体流程和包含的步骤如图4所示。

步骤1：在远程直接内存访问（RDMA）连接建立过程

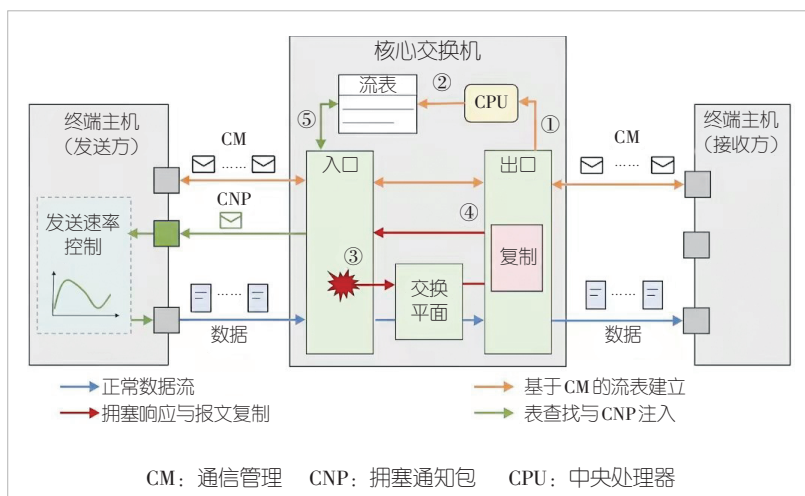


图4 快速拥塞反馈流程

中，客户端解析服务器地址并绑定到连接管理编号（CM ID）上，自动获取并管理 QPN、PSN 和重传上限等关键参数，随后发送请求报文至服务器。服务器接收请求后校验参数，通过应答报文确认连接并回送自身参数。客户端最终发送确认报文，完成链路建立。

步骤2：在 RDMA 建链控制面处理时，交换机“监听”CM 建链请求，应答和确认3类信令报文（①），并将其上传至 CPU（②）。

步骤3：CPU 基于截获的信令报文建立流表，并使用 QPN 标识双向数据流路径。

步骤4：当交换机检测到端口缓冲区超过阈值时，将从数据包中提取三元组：发送源 IP 地址、接收目的 IP 地址、接收端 QPN。

步骤5：获取这些元数据后，交换机 CPU 查询其内部流表，解析出与接收端 QPN 对应的发送端 QPN。借助这些映射，交换机通过硬件流水线生成 CNP 报文，设置发送端 QPN 为目标接收者，从而直接通知发送端 QP 调整发送速率。

在交换机内部，为了确保拥塞通知的实时性，本文通过硬件处理流程快速生成相关报文。当检测到拥塞时，交换芯片会触发专用流程：首先，将正在经过交换机入口的拥塞报文添加生成通知标记；随后，在出口处理时复制该报文，原始报文继续正常转发，而副本则被重新导入入口处理；最后在入口处，根据内部配置修改副本的特定字段，将其转换为标准的拥塞通知控制帧。这种硬件实现方式有效保证了拥塞反馈的快速性。

3.3 逐流负载均衡

3.3.1 负载均衡优化方法

负载均衡在超大规模路径网络（如 Clos 架构）内通过集中式或分布式细粒度调度，将流量分散到不同端口上。调度的粒度由大到小可以分为 flow 级别^[5-6]、flowlet 级别^[7-8]、packet 级别^[9-10]。以上分布式细粒度调度通常需要硬件、协议支持，如数据报文头修改、RDMA 网卡（RNIC）匹配、交换机大缓存区等，难以利用存量机器完成端口级别调度。

此外，近年聚焦大语言模型训练 JCT 优化的负载均衡备受关注。集合通信库多以 JCT 为核心目标设计调度，兼顾链路均衡。例如，技术感知集合通信库（TACCL）^[11]采

用拓扑感知算法最小化 GPU 通信时间；流量工程集合通信库（TECCL）^[12]重建模 MCF 问题，用混合整数线性规划（MILP）生成调度均衡负载。它们在容量约束下分配数据块，平衡传输效率与求解时间。

3.3.2 逐流出口调度

针对跨园区训练流量调度不均的问题，本文提出一种轻量级精细负载均衡方案。

本文以最小化 JCT 为目标，对负载均衡问题进行建模，并提出一种解耦优化算法：在数据中心内部，采用随机哈希结合全局优化路由（GOR）进行局部调整；在 DCI 瓶颈处采用基于排队模型的轮询算法。该框架整合了哈希引导、混合整数规划（MILP）和排队模型，实现了精确建模与求解。实验表明，该方案可将 JCT 缩短约 20%，并通过完全预计算的路由调度，实现了 DCI 与 DCN 协同优化，有效消除了负载失衡问题，并具有显著部署优势。

为实现以上基于轮询的出端口流调度，本文进一步设计了一种端网协同的逐流负载均衡协议。该协议的核心思想是由控制器下发流级别的调度决策，结合用户数据报协议（UDP）字段标记实现流到端口的精准映射，如图 5 所示。

整体协议的实现流程包含以下 4 个步骤：

步骤1：发送端策略标记。发送端交换机根据预计算策略，确定流在瓶颈处的分发端口，并利用 GOR 在 UDP 源端口字段中嵌入 DR 出端口标记，无需修改包头结构就能精准保留路径信息。该方案利用 UDP 头部字段中的源端口 srcport

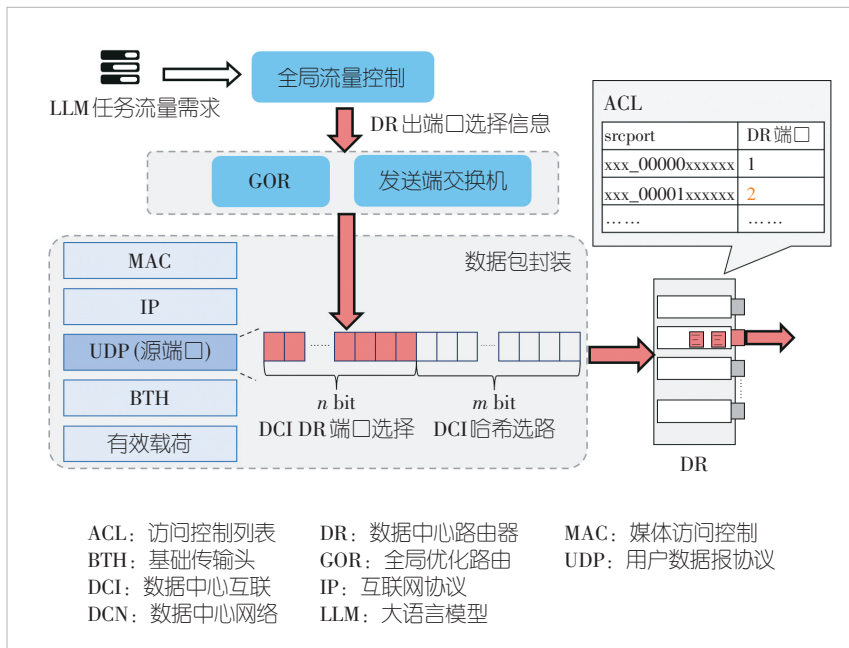


图5 逐流负载均衡流程

字段来承载负载均衡标签。srcport 字段共 2 字节，分为高位 n 比特和低位 m 比特，分别用于 DCI 出端口选取和 DCN 内部哈希选路。

步骤 2：DCN 内部哈希选路。中间交换机建立 srcport 低位到路径的哈希映射，通过低位标记 DCN 内部的选路和到 DR 的哈希。实验表明，随机哈希选路冲突会导致可用带宽下降 25%。为此，本文基于等价多路径路由（ECMP）哈希函数的线性特性，避免同设备对之间 QP 的路径冲突。设 ECMP 哈希函数为 $H(p)$ (p 为流源端口部分字段)，对于源端口字段 p 和 $p \oplus \delta$ ，哈希结果仅取决于 δ 。只要 $H(\delta) \neq H(0)$ ，即可保证两者采用不同哈希路径。

步骤 3：DCI 出口轮询分配。DCI 出园区交换机用 srcport 高位作为流编号，通过 ACL 匹配预计算的轮询结果端口，实现出端口轮询分配。

步骤 4：DCN 负载监测与冲突调整。该协议实时监测 DCN 端口负载状态，对哈希冲突路径进行局部调整，并依托集中控制器 GOR 实时监测故障。收到拥塞信号后，控制器应用 ECMP 算法去除拥塞流所在的下一跳成员，将其调度至低负载路径，以此消除路径冲突。

3.4 训练框架优化

7B-dense 小规模测试结果表明，DP 分割方案中存在大量小消息数据跨园区传输。该问题源于框架利用 ZeRO 3 对 DP 进行优化^[13]。ZeRO 3 同时分割优化状态、参数和梯度，其通信流程在前向和反向传播中均需进行 All-Gather 操作以收集完整参数，最后通过 Reduce-Scatter 同步梯度。ZeRO 3 的本质是以通信换取显存，该技术不适用于跨园区场景。对此，本文采用 ZeRO 1 方式，仅分割优化状态，以显存为代价，减少跨园区通信次数。

4 性能评估

本研究对长距分布式训练的各项性能指标进行了评估，具体实验网架构可参考图 1。

4.1 长距流控的测试结果

在 100 Gbit/s、30 km 的 DCI 基础测试中，当发送数据大小为 16 MB 时，长距窗口优化可以将传输延迟降低至原有的 1/7；当发送数据大小为 256 MB 时，该延迟优化效果提升至 20 倍。

4.2 快速拥塞反馈的测试结果

CNP 响应时间的测试结果表明，优化后 CNP 首包响应时间由 1 ms 缩短至 240 μ s，首包响应时间减少 76%。当拥塞发生时，CNP 频率会迅速增加。发送端触发反压后，速率快速下降，随后维持平稳状态。

图 6 为网卡平均发送速率和瞬时发送速率随时间变化的测试结果，其中网卡线速为 100 Gbit/s，而交换机线速（链路瓶颈）为 25 Gbit/s。结果显示，发送端平均吞吐量由 20.42 Gbit/s 提升至 24.51 Gbit/s，链路利用率接近 100%；同时开启快速拥塞反馈时，网卡瞬时速率不会在 0 ~ 100 Gbit/s 之间剧烈波动，而是稳定在 25 Gbit/s 附近，证实了快速拥塞反馈的有效性。

4.3 负载均衡的测试结果

图 7 为千卡级 GPU 集群测试中随机哈希和逐流负载均衡在 8 个出端口上的负载分布情况。随机哈希无法均分流量，负载最高的 8 号端口和最低的 4 号端口流数差值超过 20%。受木桶效应影响，通信时间将被 8 号端口拉长。逐流负载均衡将所有的流平均分到 8 个端口，此时负载不均差异降为 0，通信时间缩短。随着端口数量增大，随机哈希和逐流负载均衡在不同端口的流数差值超过 40%，逐流负载均衡的优势进一步凸显。

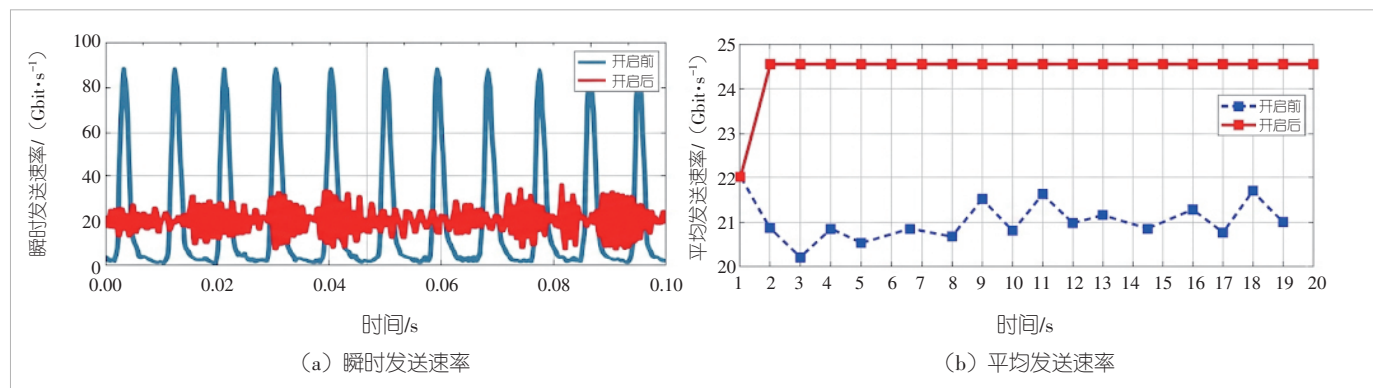


图 6 开启快速拥塞通知包前后网卡瞬时、平均发送速率时序对比

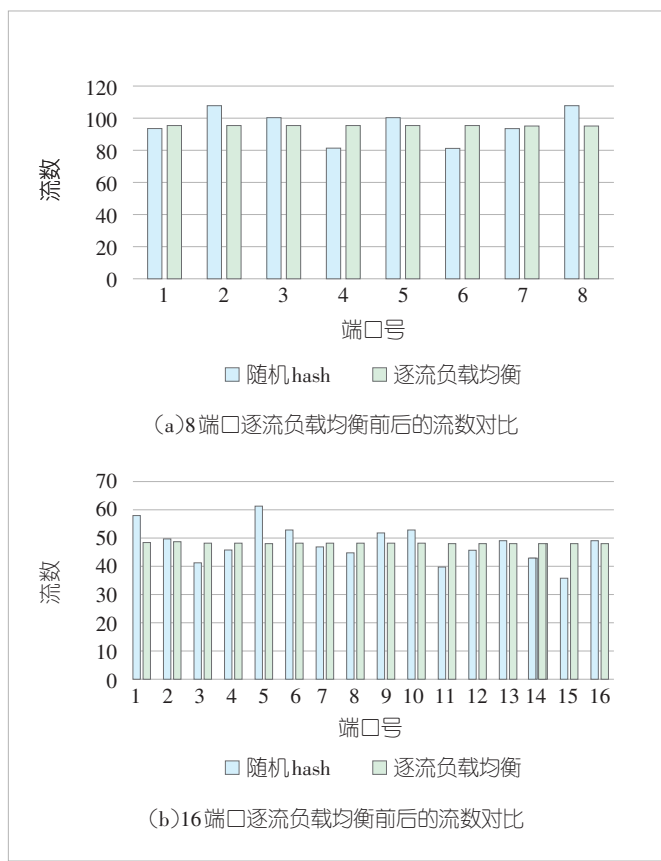


图7 随机哈希和逐流负载均衡的端口负载分布情况

4.4 LLM训练测试结果

1) 小规模基准测试

在7B-dense模型的跨园区分布式训练基准测试中,本文评估了PP与DP两种分割策略的性能。测试发现,跨园区网络的高延迟与小流控窗口会带来显著的额外算力开销。在PP分割中,优化流控窗口大小并启用快速拥塞控制,可将跨DCI引入的算力损失从9%降低至2%以内;在DP分割中,将内存优化策略从ZeRO 3切换为ZeRO 1,可避免大量跨园区参数同步通信,将算力损失从12%同样降低至2%以内,如图8所示。结果表明,针对跨园区网络特性进行通信优化,能有效控制性能损耗,使其与理论预期相符。

2) 大规模基准测试

7B-dense模型的大规模基准测试使用了百卡级和千卡级两种集群配置,跨园区场景采用PP分割策略。如图9所示,随着卡规模的增加,跨园区通信的数据量线性增加,通

信时间也随之增加,混训效率相应降低。在千卡集群下,混训效率从同园区的98%降至95%,整体算力损失为3%。

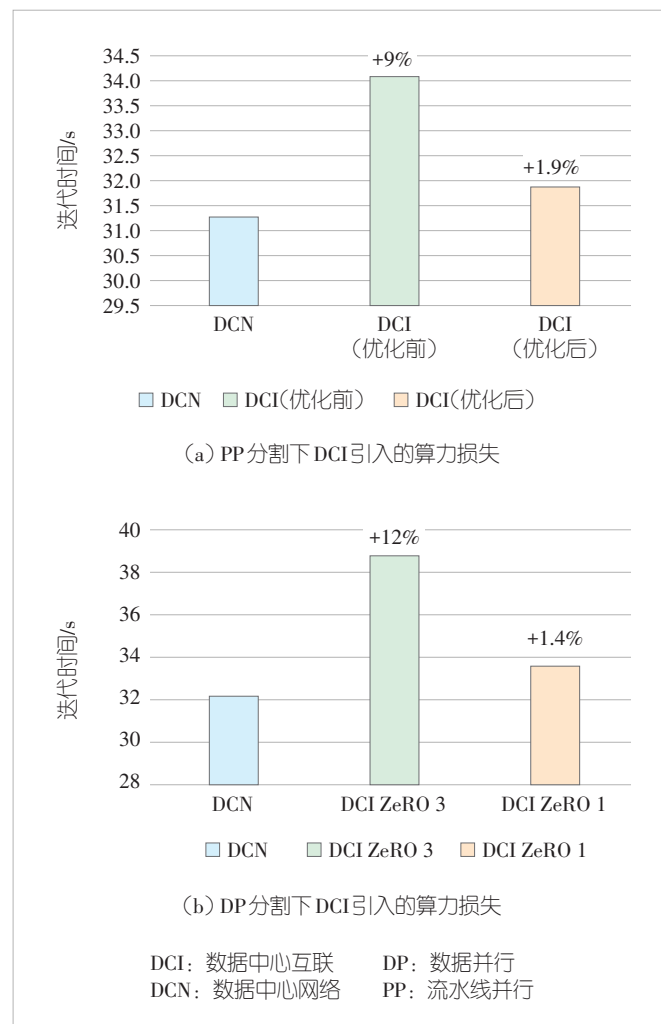


图8 小规模基准测试结果

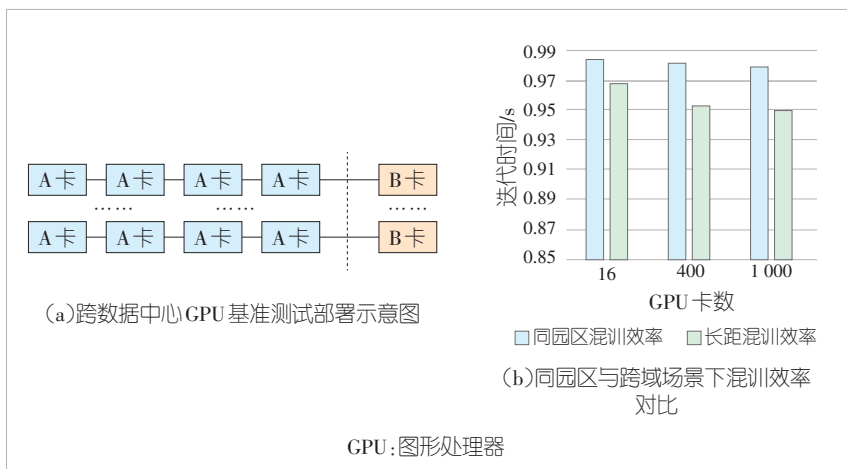


图9 大规模基准测试结果

5 结束语

本文采用长距流控技术、快速拥塞反馈机制和逐流负载均衡算法,并结合拓扑亲和的训练框架,实现了分布式训练端到端的联合优化。实验网测试表明,在千亿参数模型训练场景下,算力损失从10%~15%降低至5%以内,验证了长距网络支撑超大规模AI训练的可行性。

参考文献

- [1] NVIDIA. NVIDIA introduces spectrum-XGS ethernet to connect distributed data centers into giga-scale AI super-factories [EB/OL]. (2025-08-22)[2026-04-16]. <https://nvidianews.nvidia.com/news/nvidia-introduces-spectrum-xgs-ethernet-to-connect-distributed-data-centers-into-giga-scale-ai-super-factories>
- [2] Simonite T. Meta's next Llama AI models are training on a GPU cluster 'bigger than anything' else. [EB/OL]. (2024-10-30)[2026-04-16]. <https://reportwire.org/metanext-llama-ai-models-are-training-on-a-gpu-cluster-bigger-than-anything-else/>
- [3] Weng Q Z, Yang L Y, Yu Y H, et al. Beware of fragmentation: scheduling GPU-sharing workloads with fragmentation gradient descent [EB/OL]. (2023-08-15)[2026-04-16]. <https://www.usenix.org/system/files/atc23-weng.pdf>
- [4] InfiniBand Trade Association. InfiniBand architecture specification Volume 1. Release 1.4 [Z]. 2020
- [5] Al-Fares M, Radhakrishnan S, Raghavan B, et al. Hedera: dynamic flow scheduling for data center networks [EB/OL]. (2010-04-28)[2026-04-16]. https://www.usenix.org/legacy/event/nsdi10/tech/full_papers/al-fares.pdf
- [6] Zhang Z H, Zheng H Y, Hu J Y, et al. Hashing linearity enables relative path control in data centers [EB/OL]. (2021-07-14)[2026-04-16]. <https://www.usenix.org/system/files/atc21-zhang-zhehui.pdf>
- [7] Alizadeh M, Edsall T, Dharmapurikar S, et al. CONGA: distributed congestion-aware load balancing for datacenters [J]. ACM SIGCOMM computer communication review, 2015, 44(4): 503-514. DOI: 10.1145/2740070.2626316
- [8] Vanini E, Pan R, Alizadeh M, et al. Let it flow: resilient asymmetric load balancing with flowlet switching [EB/OL]. [2026-04-16]. <https://people.csail.mit.edu/alizadeh/papers/letflow-nsdi17.pdf>
- [9] Ghorbani S, Yang Z B, Godfrey P B, et al. DRILL: micro load balancing for low-latency data center networks [C]//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. ACM, 2017: 225-238. DOI: 10.1145/3098822.3098839
- [10] Li W X, Liu X Z, Zhang Y X et al. Revisiting RDMA reliability for lossy fabrics [EB/OL]. [2026-04-16]. <https://www.cse.ust.hk/~kaichen/papers/dcp-sigcomm25.pdf>
- [11] Shah A, Chidambaram V, Cowan M, et al. TACCL: guiding collective algorithm synthesis using communication sketches [EB/OL]. (2021-11-09)[2026-04-16]. <https://arxiv.org/pdf/2111.04867>
- [12] Liu X T, Arzani B, Kakarla S K R, et al. Rethinking machine learning collective communication as a multi-commodity flow problem [C]//Proceedings of the ACM SIGCOMM 2024 Conference. ACM, 2024: 16-37. DOI: 10.1145/3651890.3672249
- [13] Rajbhandari S, Rasley J, Ruwase O, et al. ZeRO: memory optimizations toward training trillion parameter models [EB/OL]. [2026-04-16]. <https://aiichironakano.github.io/cs596/Rajbhandari-ZeRO-SC20.pdf>

作者简介



王亚晨, 腾讯云副总裁、腾讯网络总经理, 现任中国通信学会网络专委会和边缘计算专委会副主任委员、算力网络专委会委员; 主要研究方向为智算网络、骨干网络、云网络系统; 获省部级奖励12项; 发表论文10余篇, 拥有中国和国际授权专利30余项, 制定国际标准20余项。



杨峰, 腾讯高级网络架构师; 主要研究方向为高性能DCI网络; 拥有中国和国际授权专利20余项; 发表论文10余篇。



耿竞一, 腾讯DCI网络负责人; 主要研究方向为超大规模DCI网络、高性能网络、网络虚拟化。发表论文10余篇, 拥有中国和国际授权专利20余项, 制定国际标准10余项。

广域无损智算网络关键技术研究与实践



Research and Practice of Wide-Area Lossless Intelligent Computing Network Key Technologies

雷波/Lei Bo, 潘永琛/Pan Yongchen, 唐静/Tang Jing

(中国电信股份有限公司北京研究院, 中国 北京 102209)
(China Telecom Corporation Limited Beijing Research Institute, Beijing 102209, China)

DOI: 10.12142/ZTETJ.202603003

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260626.1350.012>

网络出版日期: 2026-06-26

收稿日期: 2026-03-25

摘要: 智算中心跨域互联已成为支撑规模化人工智能服务的关键网络新形态。系统分析了广域智算网络面临的传输需求与挑战, 提出了构建广域无损智算网络的核心概念, 以精准流量控制和队列缓存增强为关键技术路径, 推动广域智算互联从尽力而为传输向无损承载演进。广域无损智算网络面向存算分离训练、云边协同推理等多类智算服务场景, 实现了网络拥塞条件下的无损高效传输, 为规模化智算服务构建了稳定可靠的网络底座。

关键词: 智算网络; 广域网; 拥塞丢包; 无损传输

Abstract: Cross-domain interconnection of intelligent computing centers has emerged as a critical network paradigm for large-scale artificial intelligence services. The transmission requirements and challenges of wide-area intelligent computing networks are analyzed, and the core concept of building a wide-area lossless intelligent computing network is proposed. Leveraging precise flow control and enhanced queue buffering as key technical enablers, this network drives the evolution of wide-area intelligent computing interconnection from best-effort to lossless transport. Targeting diverse service scenarios such as storage-compute decoupled training and cloud-edge collaborative inference, the wide-area lossless intelligent computing network achieves lossless and efficient transmission under network congestion, thereby establishing a stable and reliable foundation for large-scale intelligent computing services.

Keywords: intelligent computing network; wide area network; congestion-induced packet drop; lossless transmission

引用格式: 雷波, 潘永琛, 唐静. 广域无损智算网络关键技术研究与实践 [J]. 中兴通讯技术, 2026, 32(3): 10-14. DOI: 10.12142/ZTETJ.202603003

Citation: Lei B, Pan Y C, Tang J. Research and practice of wide-area lossless intelligent computing network key technologies [J]. ZTE technology journal, 2026, 32(3): 10-14. DOI: 10.12142/ZTETJ.202603003

人工智能 (AI) 已成为引领新一轮科技革命与产业变革的核心驱动力。构建自主可控、高效普惠的智能算力互联基础设施, 是中国把握新质生产力发展机遇、提升全球科技竞争优势的关键战略部署。2025年,《国务院关于深入实施“人工智能+”行动的意见》^[1]明确提出,要强化智能算力统筹调度,推进智能算力互联互通与供需精准匹配,创新智能算力基础设施运营模式。

当前,算力主体主要集中于多个独立的大规模智算中心,部分算力分散于企业边缘自建节点,整体资源布局相对

割裂。随着人工智能的持续快速发展,单一智算中心已难以满足大规模、弹性化的预训练、微调、推理等业务需求。面对千行百业对人工智能服务的普及化需求,我们亟需依托广域网实现智算中心之间、边缘算力与云端智算中心之间的高效网络互联互通,从而达到算力的灵活调度与资源的弹性供给,从而满足用户在隐私、成本、性能等方面的多样化需求。

本文首先重点剖析人工智能服务在广域智算网络中的互联场景及其面临的挑战,随后提出构建广域无损智算网络的核心观点,通过采用精细化流控与队列缓存增强等技术,保障高效无损传输,最后结合现网开展测试验证,以验证应用成效。

基金项目: 国家重点研发计划项目 (2024YFB2908700); 国家自然科学基金项目 (U25B2038, 62394325)

1 广域智算互联场景

随着以Transformer架构为核心的大语言模型技术加速迭代,在模型训练与推理部署等核心场景中,当前智算基础设施面临多重制约。一方面,单一智算中心的物理边界日趋刚性:在基础模型预训练上,大模型参数规模持续攀升,算力、电力、存储等资源供给已逼近上限。另一方面,边缘智算节点在成本、运维与规模化方面亦存短板:面向千行百业的模型推理与微调场景中,企业AI应用已从算力一体机单点部署阶段,迈入规模化落地应用阶段,但私有化部署模式正严重制约着微调、推理服务的规模化扩展。私有化的算力部署面临三大挑战:一是算力购置成本高,扩容成本难以承受;二是边缘侧小规模算力难以高效管控,企业缺乏边缘集群运维能力,扩容与故障恢复困难,易造成服务中断;三是自建算力场景有限、调度不充分、算力闲置严重,造成大量资源浪费。

在预训练场景下,多智算中心协同训练已成为突破超大模型训练瓶颈的关键路径。跨地域、跨集群整合多个大规模智算中心的算力资源,可有效打破单一智算中心在算力、存储等方面的物理限制。智算中心间采用数据并行(DP)、流水线并行(PP)等方式调度分散算力节点,构建统一的分布式训练集群,实现训练任务的高效拆解与协同执行。

在云边协同的推理、后训练与存算分离服务场景下,随着推理算力需求持续增长,企业用算模式正加速向云边协同架构转型,通过协同边缘智算节点与云端智算集群,为千行百业提供多样化AI服务。其中,存算分离的微调服务将训练语料数据集保留在边缘智算节点,通过高性能广域网络向智算中心同步传输批量数据,由智算中心负责梯度计算与参数更新;在隐私增强的推理服务中,按推理模型的流水线结构对模型进行切分,将推理计算实例分别部署于边缘节点和智算中心,通过网络传输高维度隐藏层变量,实现用户信息安全保护;在推理预填充-解码(PD)分离架构中,将预填充与解码阶段分别部署在不同云边节点,以充分利用异构与分散的算力资源,并借助分离架构提升端到端推理性能;对于海量数据入算场景,边缘节点存储的数据通过高性能网络传输至云端智算中心,完成集中处理、计算与数据备份。图1展示各类广域智算互联典型场景业务。

2 广域智算网络面临的挑战

2.1 广域智算网络加剧丢包效能损耗

远程直接内存访问(RDMA)作为智算服务的核心网络底座技术,凭借内核旁路、协议栈硬件卸载与零拷贝三大核心特性,能够实现高吞吐、低时延、低中央处理器(CPU)开销的高性能网络通信,已成为当前智算中心网络的关键支撑技术。其中,基于以太网架构的第二代RDMA融合以太网(RoCE v2)协议已成为数据中心内部RDMA规模化部署的主流方案^[2-3]。然而,RoCE v2的稳定运行高度依赖无损网络环境。RDMA传输对丢包极为敏感,仅1‰的丢包率即可导致吞吐率减半^[4]。在广域长距传输场景中,链路往返时延(RTT)的大幅增加会进一步放大丢包的负面影响,从而引发传输性能的断崖式劣化。

在长距传输场景中,当链路突发流量超出设备缓存承载能力时,会引发拥塞丢包。当前主流RDMA网卡的丢包重传算法正从回退N帧(GBN)向选择性重传(SR)演进,但两类算法在广域长距场景下均存在明显局限。对于GBN算法而言,由于广域链路端到端距离长、带宽时延积(BDP)较大,单次丢包所触发的全量重传会占用大量链路带宽,导致尾时延增加至毫秒级,极易造成有效吞吐骤降甚至业务中断。SR算法虽通过选择性重传机制减少了重传数据量,提升了重传效率,但由于广域链路的重传环路时延较长,多点丢包易引发端侧位图溢出,最终仍将退化为GBN模式。此外,广域链路固有的长时延、高抖动特性,极易触发端侧超时重传机制,产生大量无效重传,进一步挤占有效带宽,加

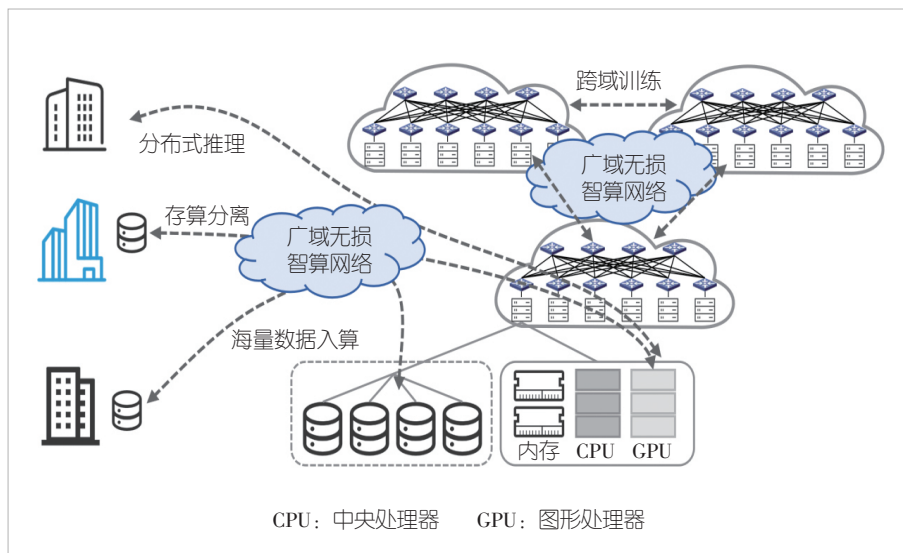


图1 广域智算互联典型场景业务

剧性能损耗。如图2所示,在500 km的云边协同推理服务中,RDMA传输性能随丢包率上升而下降,每Token输出时间(TPOT)显著增加,输出Token吞吐量相应下降;当丢包率升至0.1‰时,业务将完全中断。

为避免拥塞丢包,RoCE v2协议在智算中心内依托优先级流量控制(PFC)技术,构建端到端的无损传输网络。当前,学术界尝试基于PFC机制实现广域流控能力的优化^[5-6],但PFC本身存在的瓶颈制约了其在广域智算传输中的应用。PFC仅支持8个优先级队列,易引发拥塞扩散、流量风暴与死锁等问题,产生受害者流,影响网络的公平性与稳定性。若将PFC直接延伸至广域链路,单点拥塞将被扩散至跨地域的整个网络,从而导致大规模业务中断。因此,数据中心内的无损网络技术体系无法直接复用于广域智算网络,难以成为其全局解决方案。

2.2 拥塞控制优化无法保障无损传输

RDMA在广域网中面临拥塞控制环路长、控制决策严重滞后的问题。智算中心内的RoCE v2通过PFC与端到端拥塞控制算法的协同实现无损传输保障,其中主流拥塞控制算法依赖RTT、显式拥塞通知(ECN)等端到端拥塞信号的反馈。然而在广域长距离场景下,链路RTT约为每百公里1毫秒(1 ms/100 km),拥塞信号反馈存在严重滞后,进而引发拥塞收敛缓慢、网络震荡剧烈等一系列问题。更关键的是,当前主流RDMA网卡普遍采用速率驱动的发送机制而非窗口控制机制,无法对网络中的在途流量进行精准管控。一旦发生拥塞,毫秒级的拥塞反馈时延会导致大量过量流量持续注入网络,进一步加剧链路拥塞,甚至引发丢包。

当前数据中心内大规模商用的数据中心量化拥塞通知(DCQCN)算法^[4]本质上仍属于端到端的拥塞控制机制,已难以适配广域网络高延迟反馈环路的场景特性。PACC^[7]、BiCC^[8]、ATC^[9]等现有优化方案虽通过提前向源端反馈拥塞信号实现快速速率控制,从而保障拥塞的快速感知,但在面对广域网络内的远端拥塞场景时,这些方案即便将拥塞点转移至远端数据中心出口的大缓存设备上,仍需要依赖端到端的拥塞信号来降低队列缓存,因此难以实现端到端的无损传输保障。

3 广域无损智算网络解决方案

针对广域智算网络中存在的上述挑

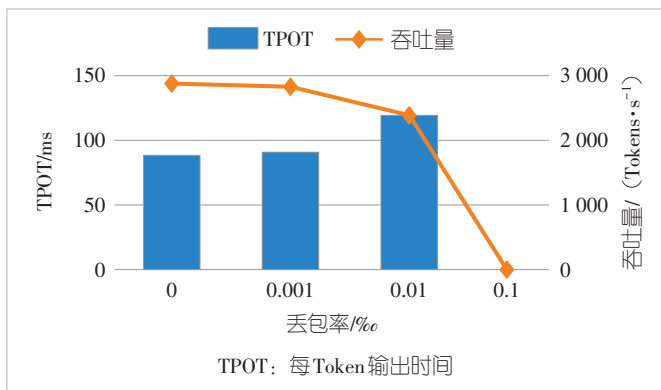


图2 广域智算网络丢包性能损耗加剧

战,本文提出了广域无损智算网络的观点,并在网络交换设备上采用关键技术,构建了相应的解决方案。如图3所示,关键技术包括:1)通过采用反压式的流控算法构建无损网络,并借助网络切片能力实现精细化流控,以避免受害者流的产生;2)通过部署多队列大缓存的网络设备来容纳突发流量,防止缓存溢出。

3.1 精准反压式流控

精准流控反压采用分段反压方式,在智算中心与广域网两个层级部署差异化的反压机制。

在智算中心管控域内,无损智算网络沿用PFC机制。智算中心与广域网的交汇点存在带宽收敛特性,使其成为端到端传输路径上最易发生拥塞的瓶颈点。为保证智算中心出口的流量无损传输,仅在智算中心及边缘的转发设备上使能PFC机制,利用其链路层微秒级快速反压能力,避免出口收敛瓶颈处发生缓存溢出丢包。同时,将PFC的作用边界严格限定在智算中心边缘,以防止智算中心内的拥塞扩散,阻断

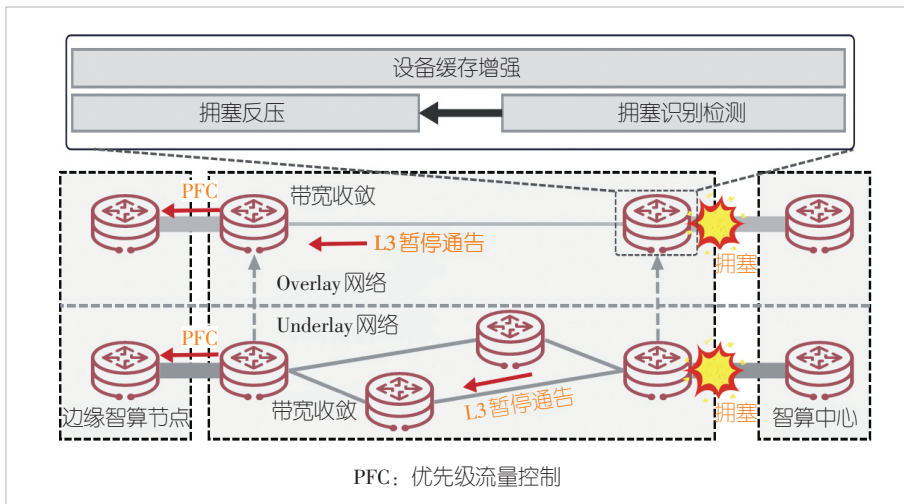


图3 广域无损智算网络精准反压式流控机制

内部局部拥塞向广域网的传导。

在广域网管控域内，无损智算网络架构采用一种三层反压通告机制来完成反压式流控。在拥塞识别方面，与PFC检测入端口队列深度的机制不同，本文提出的精准流控基于随机早期检测在出端口执行拥塞判断。当因多对一通信模式或出入带宽不匹配导致队列深度达到预设暂停水线时，系统触发流量暂停操作。识别拥塞队列后，通过互联网控制报文协议（ICMP）报文向上游设备发送暂停通告，报文携带租户切片ID与暂停时长，按基于IPv6的段路由（SRv6）路径路由至上游节点，以暂停对应队列的发送。反压支持逐跳与透传两种模式：逐跳反压需设备升级支持；透传反压基于Overlay的SRv6隧道转发，可兼容存量设备，降低改造成本。传统PFC仅支持8个优先级，易引发租户干扰、拥塞扩散与受害者流。本文所采用的精准流控依托网络切片实现细粒度队列隔离，按租户分配独立切片与专属队列，反压仅在本切片内生效，不影响其他RDMA流量，从而在切片间实现强租户性能隔离，彻底避免受害者流。

3.2 无损网络设备缓存增强

广域网链路时延由智算中心内的微秒级增大至毫秒级，导致带宽时延积（BDP）数量级显著增大。为适配这一特性，广域网设备通常配置大容量缓存资源，用于缓解链路传输波动并容纳突发数据包。尽管精准反压式流控限制了流量突发、维持了队列深度的稳定，但仍需设备缓存能力的配合，以构建无损的网络链路。如图4所示，在ICMP反压通告报文发送至上一跳的时间段内，链路中最大可能存在大量在途数据包。因此，设备需保证至少具备与理论最大在途报文相匹配的缓存空间，以吸收反压过程中的数据包，避免在反压通告间隔期间发生拥塞丢包。在网络部署时，需根据链路BDP选择合适的设备缓存配置。此外，转发设备在大容量缓存的基础上，借助智能化缓存管理、动态水线配置等管理调度算法，可进一步吸收流量突发、平滑队列波动，从而

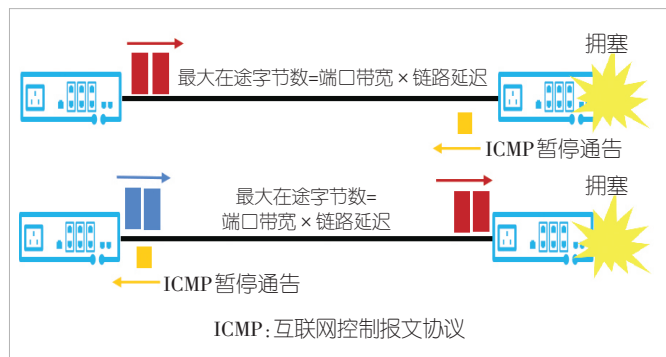


图4 无损网络缓存增强原理

与精准流控机制协同保障全链路的无损传输。

4 现网实践验证

中国电信对智算广域网核心技术开展落地现网实践，完成无损广域智算网络全场景验证工作，在广域云边协同推理服务和存算分离微调训练服务场景中均实现了成本与效率的精准平衡。现网实践中，方案成功将带宽收敛比提升至320:1，大幅降低按算力卡能力建网带来的带宽成本，通过租户级精准流控与设备缓存增强构建了无损广域智算网络，实现多租户性能隔离与无损转发，云边协同训练推算效不低于95%。

为验证广域无损智算网络带来的性能收益，中国电信基于500 km现网环境搭建了无损广域智算网络试验床，开展云边协同推理服务和存算分离微调服务的测试，并分别评估了无损网络与容损网络在不同带宽瓶颈下的性能表现。如图5所示，在DeepSeek-lite-v2-16B模型的存算分离微调服务测试中，广域无损智算网络借助精细化流控与设备能力增强，可将互联带宽压缩至1 Gbit/s。而在无无损保障的容损网络下，当带宽瓶颈为1 Gbit/s时，发生拥塞丢包，算效损失达20%。如图6所示，在Qwen3-32B模型的云边协同推理服务测试中，无损环境下可将广域切片带宽压缩至5 Gbit/s，推理性能仍可保持在95%以上；而在容损广域网络中，丢包导致TPOT增加3.4倍，输出Token吞吐率劣化超过60%。上述结果表明，对于当前跨广域网的人工智能服务而言，构建无损的智算网络至关重要。

面向低成本轻量化运营的诉求，边缘网络侧部署轻量化智算客户终端设备（CPE）形态设备，通过中国电信信息壤平台进行能力集成，将智算网关与CPE统一纳入集约化运维管理体系。该架构可灵活适配不同层级带宽的汇聚，配套租户粒度精细化流量管控策略，同时联动企业侧实现协同反压联动调度。依托SRv6与网络切片技术，为业务提供可量化可

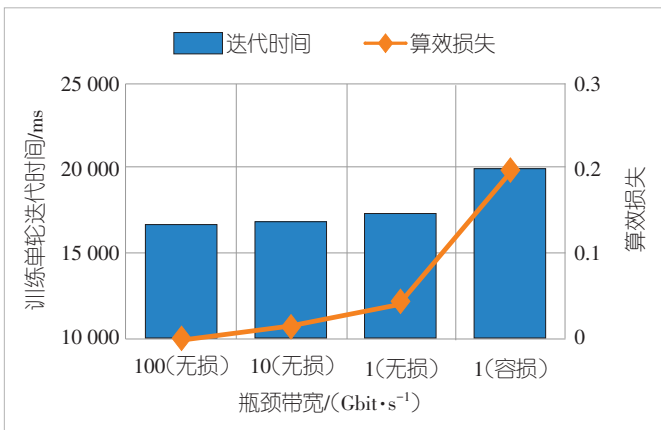


图5 存算分离微调服务性能评估

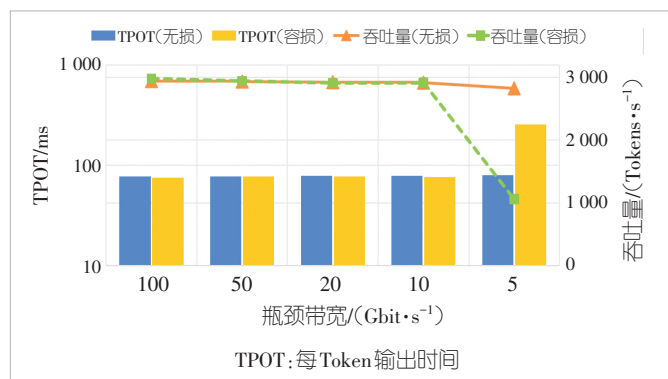


图6 云边协同推理服务性能评估

保障的高品质传输服务，全面支撑企业训练推理算力资源与RoCE无损网络的一体化部署，实现业务极简开通上线与全流程轻量化运维，有效压降企业整体运营成本与管理负担。

面向低成本、轻量化运营的诉求，边缘网络侧部署轻量化智算CPE形态设备，并通过中国电信信息壤平台进行能力集成，将智算网关与CPE统一纳入集约化运维管理体系。该架构可灵活适配不同层级带宽的汇聚需求，配套租户粒度的精细化流量管控策略，同时联动企业侧实现协同反压与联动调度。依托SRv6与网络切片技术，该架构可为业务提供可量化、可保障的高品质传输服务，全面支撑企业训练推理算力资源与RoCE无损网络的一体化部署，实现业务的极简开通上线与全流程轻量化运维，从而有效压降企业整体运营成本与管理负担。

5 结束语

在算力泛在化与智算广域互联的发展趋势下，构建高可靠、高效率、可隔离的广域无损智算网络，已成为支撑AI时代智算服务的关键底座。面向跨域智算服务规模化部署的产业需求，亟需持续强化算力、网络与存储的一体化协同设计，加快高性能传输核心技术的工程化落地，推动广域智算网络从试验验证迈向规模商用。需要指出的是，广域无损智算网络的技术成熟与部署推广是一个循序渐进的过程，在方案设计、现网试验与运营优化中需不断总结经验、迭代完善。当前，广域智算网络仍处于技术探索与试点验证的关键阶段，亟需围绕用户需求、建设成本与网络性能开展深入研究，加速关键技术的成熟与方案的落地，从而为构建自主可控、高效普惠的智能算力基础设施提供坚实支撑。

参考文献

- [1] 中华人民共和国国务院. 国务院关于深入实施“人工智能+”行动的意见: 国发〔2025〕11号 [EB/OL]. (2025-08-21)[2026-04-20]. https://www.gov.cn/zhengce/content/202508/content_7037861.htm

- [2] 李宝嘉, 何春志, 夏寅贵, 等. 星脉网络: 面向GPU集群集合通信与集中式路由的协同优化 [J]. 中兴通讯技术, 2025, 31(2): 3-13. DOI: 10.12142/ZTETJ.202502002
- [3] 段威, 于浩, 李和松, 等. 智算中心组网技术及应用 [J]. 中兴通讯技术, 2025, 31(2): 63-71. DOI: 10.12142/ZTETJ.202502009
- [4] 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. DOI: 10.1145/2829988.2787484
- [5] Yu P W, Xue F Y, Tian C, et al. Bifrost: extending RoCE for long distance inter-DC links [C]//Proceedings of IEEE 31st International Conference on Network Protocols (ICNP). IEEE, 2023: 1-12. DOI: 10.1109/ICNP59255.2023.10355634
- [6] 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. DOI: 10.1109/TPDS.2022.3215517
- [7] Zhong X L, Zhang J, Zhang Y L, et al. PACC: proactive and accurate congestion feedback for RDMA congestion control [C]//Proceedings of IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. IEEE, 2022: 2228-2237. DOI: 10.1109/INFOCOM48880.2022.9796803
- [8] Wan Z R, Zhang J, Yu M X, et al. BiCC: bilateral congestion control in cross-datacenter RDMA networks [C]//Proceedings of IEEE INFOCOM 2024 - IEEE Conference on Computer Communications. IEEE, 2024: 1381-1390. DOI: 10.1109/INFOCOM52122.2024.10621412
- [9] Wan Z R, Zhang J, Su Y Z, et al. Re-architecting traffic control in cross-datacenter RDMA networks [J]. IEEE transactions on networking, 2025, 33(5): 2632-2647. DOI: 10.1109/TON.2025.3569630

作者简介



雷波, 中国电信股份有限公司北京研究院互联网技术研究所副所长(主持工作); 主要从事算力网络、智算网络、智能体网络等新型网络技术研究; 发表论文50余篇, 参与制定标准10余项, 发布著作7部, 拥有发明专利33项。



潘永琛, 中国电信股份有限公司北京研究院互联网技术研究所研究员; 主要从事智算网络传输控制协议与高性能网络优化方面的研究工作; 发表论文10余篇。



唐静, 中国电信股份有限公司北京研究院互联网技术研究所团队总监; 主要从事算力网络、智算网络、数据中心网络等未来网络技术研究, 深度参与ITU-T、CCSA等标准化工作; 发表论文10余篇。

Scale-Up 网络内 GPU 卡间互联及 OISA 技术体系



Intra-Node GPU Interconnection in Scale-Up Networks and OISA Technologies

宋晓勇/Song Xiaoyong, 李锴/Li Kai, 陈佳媛/Chen Jiayuan

(中国移动通信有限公司研究院, 中国 北京 100053)
(Research Institution of China Mobile, Beijing 100053, China)

DOI: 10.12142/ZTETJ.202603004

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260626.1349.010>

网络出版日期: 2026-06-26

收稿日期: 2026-04-05

摘要: 针对图形处理器 (GPU) Scale-Up 卡间互联, 系统梳理了其发展背景、核心需求与现实挑战, 总结了从通用总线、私有协议到开放标准的 3 阶段演进历程及技术趋势。以中国移动提出的全向智感互联架构 (OISA) 为分析对象, 详细阐述了其架构设计、核心技术特征与创新优势。在此基础上, 结合大模型推理阶段键值缓存 (KV Cache) 的存储通信需求, 以及超节点场景下 Scale-Up 互联面临的新挑战, 对其未来发展方向进行了展望与探讨, 为智算网络体系构建及超节点演进提供了参考。研究表明, 各协议技术正逐渐趋同, 开放互联已成为重要发展态势。OISA 依托分层协议栈设计与关键技术创新, 构建了高效、可靠、灵活、开放的 GPU 高速互联体系, 为智算服务器向超节点迭代升级提供了关键支撑。

关键词: 垂直扩展; Scale-Up; 卡间互联; 全向智感互联; 超节点; 智算网络

Abstract: A systematic review is conducted on the background, core requirements, and practical challenges of graphics processing unit (GPU) Scale-Up interconnection, and its three-stage evolution from general-purpose buses and proprietary protocols to open standards, along with corresponding technical trends, is summarized. With the omnidirectional intelligent sensing express architecture (OISA) proposed by China Mobile taken as the analytical object, its architecture design, key technical features, and innovative advantages are elaborated in detail. On this basis, combined with the storage and communication demands of Key-Value (KV) Cache in large-model inference and the emerging challenges faced by Scale-Up interconnection in SuperPod scenarios, its future development directions are discussed and prospected, providing references for intelligent computing network construction and SuperPod evolution. The results indicate that interconnection protocols and technologies are gradually converging, and open interconnection has become a key development trend. Based on hierarchical protocol stack design and key technological innovations, OISA establishes an efficient, reliable, flexible, and open high-speed GPU interconnection system, providing critical support for the iterative upgrade of intelligent computing servers toward SuperPods.

Keywords: vertical expansion; Scale-Up; intra-node GPU interconnection; omni-directional intelligent sensing express architecture (OISA); SuperPod; intelligent computing network

引用格式: 宋晓勇, 李锴, 陈佳媛. Scale-Up 网络内 GPU 卡间互联及 OISA 技术体系 [J]. 中兴通讯技术, 2026, 32(3): 15-21. DOI: 10.12142/ZTETJ.202603004

Citation: Song X Y, Li K, Chen J Y. Intra-node GPU interconnection in scale-up networks and OISA technologies [J]. ZTE technology journal, 2026, 32(3): 15-21. DOI: 10.12142/ZTETJ.202603004

1 Scale-Up 卡间互联

1.1 背景

人工智能技术的爆发式发展, 推动大语言模型 (LLM) 和混合专家模型 (MoE) 向万亿乃至十万亿参数规模演进。在此过程中, 模型训练与在线推理对参数、激活值和梯度的交互提出了极高密度、低延迟和强实时性的要求, 进而对算力、存储及通信效率构成严峻挑战。单张图形处理器

(GPU) 的算力和显存容量已无法独立支撑此类大模型的运行, 必须依托数据并行 (DP)、张量并行 (TP)、MoE 并行、流水线并行 (PP) 等分布式并行策略, 由多张 GPU 协同完成计算任务。然而, 由于 GPU 之间的数据搬运速度明显滞后于其计算速度, 大量算力因等待数据而空转。因此, 内存共享效率与通信效率成为制约多 GPU 协同性能的关键因素。

智算网络是实现多 GPU 协同的关键基础设施, 其性能直接决定模型训练的迭代周期与推理响应速度。智算网络主

要分为3类,即Scale-Up(垂直扩展)网络、Scale-Out(水平扩展)网络和Scale-Across(跨域扩展)网络^[1]。其中,Scale-Up网络专注于单服务器节点或超节点内的GPU高速卡间互联,尤其适配TP、MoE并行等紧耦合任务;这与聚焦同园区内服务器间互联、适配数据并行等松耦合任务的Scale-Out网络不同,也不同于面向跨园区多智算中心广域协同的Scale-Across网络。高性能卡间互联协议及相关技术正是构建Scale-Up网络的核心^[2-3]。

1.2 需求与挑战

Scale-Up网络通过超高带宽、极低时延的互联,突破单节点内多GPU的数据交互壁垒与存储访问边界,构建逻辑一体化的“超级GPU”架构,即超节点。该网络性能持续跃升,推动服务器硬件架构不断演进,算力部署向高密度集约化方向升级;超节点逐步成为新一代智算基础设施的主流形态,同时也对卡间互联技术提出了极为严苛的要求^[4]。

在带宽方面,卡间互联技术需匹配高带宽内存(HBM)的访问速率,GPU单端口带宽需达400~800 Gbit/s,并在高端场景中向更高速率演进,以满足TP海量参数的即时同步需求。在时延方面,GPU间往返时延(RTT)需维持在微秒级,交换拓扑下的单跳转发时延需控制在百纳秒级,以保障协同操作的实时性。在扩展性方面,该技术需突破传统直连或Cube-Mesh等拓扑的限制,基于交换拓扑实现数百张GPU卡的无阻塞全互联。功能方面,该技术需具备无损传输机制及集合通信加速(CCA)能力,保障整体传输效率,并通过统一内存寻址等消除数据拷贝开销,简化多GPU编程模型。

面向上述核心需求,Scale-Up互联面临多重瓶颈。在技术层面,卡间互联需同步满足带宽、时延与互联规模的极致约束;传统互联协议的演进迭代滞后于GPU及HBM硬件的升级,普遍存在固有性能短板,且改造适配成本高。随着组网规模扩大,传统方案易出现时延累积、流量拥塞及拓扑适配等难题,叠加大模型训练与推理差异化的通信特征,进一步增加了架构兼容与设计难度。当前中国GPU厂商多采用私有互联协议,缺乏专用交换芯片的深度适配,阻碍其向超节点架构演进。若针对单一GPU生态定制研发交换芯片,高额研发成本难以通过市场规模实现有效分摊。同时,中国高速率串行器/解串器(SerDes)设计能力与先进芯片制程工艺供给相对不足,叠加高密度集成场景下的散热与能耗管控压力,抬高了超节点一体化设计门槛与工程实现难度。

综上所述,现有Scale-Up互联技术在性能、生态与工程化层面均存在明显局限,亟需一套开放、高性能的互联协议体系,以支撑下一代智算超节点的发展。

2 卡间互联技术演进

2.1 演进历程

GPU Scale-Up卡间互联的演进可划分为以下3个阶段^[5]。

第1阶段:传统通用互联阶段。在此阶段,产业界普遍采用外围组件互连高速(PCIe)总线实现GPU卡间互联。尽管PCIe 6.0及之后版本通过调制编码优化提升了带宽,但其原生设计面向通用外设互联,且采用根复合体(Root-Complex)主从架构,导致其在智算服务器GPU协同场景下存在固有缺陷。具体而言,其带宽与时延难以满足紧耦合并行计算的需求,树状拓扑结构也限制了大规模组网能力,最终成为制约算力释放的瓶颈^[6]。

第2阶段:私有协议主导的定制化互联阶段。为突破PCIe架构的性能限制,GPU厂商围绕自身芯片特性推出了专属互联协议。在国际头部厂商中,NVIDIA于2014年发布NVLink^[7],AMD于2017年推出Infinity Fabric Link^[8],Intel于2020年提出X^e Link;中国厂商也同步推进,壁仞科技在BR100产品中采用BLink,摩尔线程在MTT S4000产品中采用MTLink,均实现了GPU全互联。这些私有协议大幅提升了卡间通信性能,单卡带宽可达200~400 GB/s,点对点时延显著降低。然而,除NVIDIA的NVLink配有NVSwitch专用交换芯片外,绝大多数私有协议缺乏交换芯片支持,仅能实现8卡互联,无法支撑Scale-Up网络的规模化扩展需求。

第3阶段:规模化与标准化互联阶段。面向超节点百卡级乃至更高密度的Scale-Up互联需求,产业界加速构建开放标准协议体系。2024年,中国移动率先提出全向智感互联(OISA)技术体系,并于2025年8月发布2.0版本协议规范;同年,AMD主导联合阿里巴巴、Astera Labs、Intel、Meta、Synopsys等企业成立统一加速器互联(UALink)联盟,并于2025年4月发布UALink 1.0协议^[9];Broadcom于2025年4月发布可扩展统一以太网(SUE)初版规范;华为于2025年9月发布灵衢(UB)基础规范^[10];开放计算项目(OCP)联合AMD、Broadcom、Cisco、NVIDIA等厂商于2025年10月成立增强型扩展网络(ESUN)工作组,并于2026年2月发布首版基础规范^[11]。

这些新型开放标准互联协议的集中涌现,从根本上打破了单一厂商的生态垄断,实现了GPU与交换芯片之间的互联互通,推动了卡间互联从直连拓扑向基于交换架构的全互联模式转变,显著提升了大规模组网能力与拓扑扩展性。

2.2 演进趋势

随着ScaleUp互联进入开放化、标准化阶段,各类新型

卡间互联协议虽然在报文格式与技术细节上存在差异，但其整体演进呈现出清晰的共性趋势。

1) 协议轻量化和高效化。面向GPU高频细粒度通信与内存语义访问模式，各协议以实现超高带宽、超低时延为核心目标，持续精简协议开销，最大化提升有效带宽利用率，支持低时延的Load/Store访问，从而充分释放GPU算力潜能。

2) 互联拓扑向基于交换的无阻塞全互联升级。如图1所示，GPU卡间互联拓扑从传统PCIe树状架构、私有协议的点对点直连，转向以专用ScaleUp交换芯片为核心的全互联架构，可支持百卡至千卡级高密度组网，单跳转发时延控制在百纳秒级，从而满足超节点性能及大规模扩展需求。

3) 传输机制趋于无损化。各协议普遍采用基于信用的流量控制（CBFC）和基于优先级的流量控制（PFC），并结合链路级重传机制，实现无丢包、高可靠传输，以满足大模型训练对通信稳定性的严苛要求。

4) 物理层向高速化方向演进。单端口速率由200 Gbit/s、400 Gbit/s向800 Gbit/s、1.6 Tbit/s升级，高速串行器/解串器（SerDes）与轻量级前向纠错（FEC）成为标配。

5) 生态开放化与多厂商兼容。开放标准协议已成为主流，支持跨厂商GPU与交换芯片互联互通，降低供应链依赖与部署成本，推动ScaleUp技术走向标准化与产业化。

3 全向智感互联OISA技术体系

3.1 总体架构

OISA由中国移动于2024年提出，同年相继发布1.0和

1.1版本协议规范，并于2025年8月发布2.0版本协议规范^[12]。OISA旨在构建高效、智能、灵活、开放的卡间互联体系，基于交换拓扑建立GPU间的对等通信架构，通过原生支持GPU内存语义访问，实现高带宽、低时延的大规模ScaleUp通信，为数据密集型AI应用（如大模型训练、推理及高性能计算）提供支撑。

如图2所示，OISA协议栈采用事务层、数据层和物理层三级分层架构，各层级功能解耦、协同配合，共同支撑高速卡间互联传输。

事务层作为协议栈顶层，直接对接GPU片上网络，主要完成上层访问事务的协议封装与解析，同时集成多模内存访问、集合通信加速、报文聚合分解、即时反馈等核心能力。该层兼容两种内存语义访问模式：一是基于Load/Store指令的同步内存语义（指令直驱模式），二是基于直接内存访问（DMA）的异步内存语义。同时，该层明确了内存栅栏操作要求，以保障多GPU内存访问的一致性与正确性，适配不同业务粒度与性能诉求。数据层位于中间，依托智能状态感知、多种流量控制和链路级重传等关键技术，为上层事务交互提供高效、可靠的传输支撑。物理层采用以太兼容架构，分为逻辑子层（PCS/PMA）与电气子层（PMD），在兼容IEEE 802.3标准、保留生态通用性的基础上，优化高速SerDes、轻量化编解码等底层技术，针对超节点高密度组网场景完成定制化性能优化，实现高带宽、低损耗的物理链路传输。

为适配多样化互联场景，OISA定义了3种工作模式：极致性能访问模式、智能感知访问模式与集合通信加速

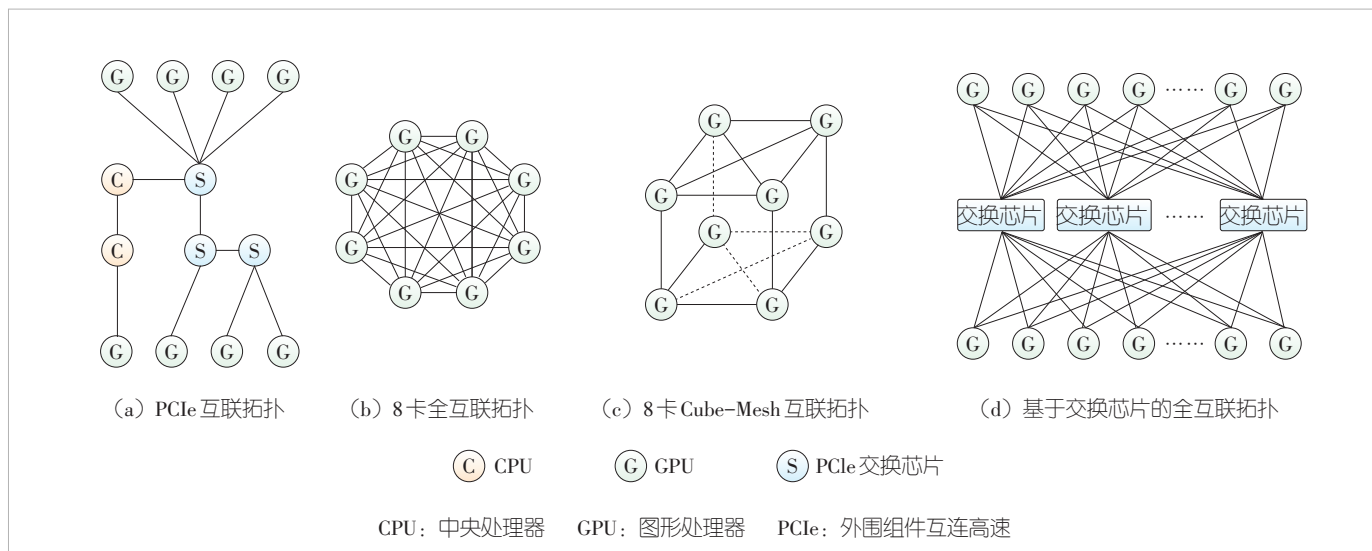


图1 GPU卡间互联拓扑

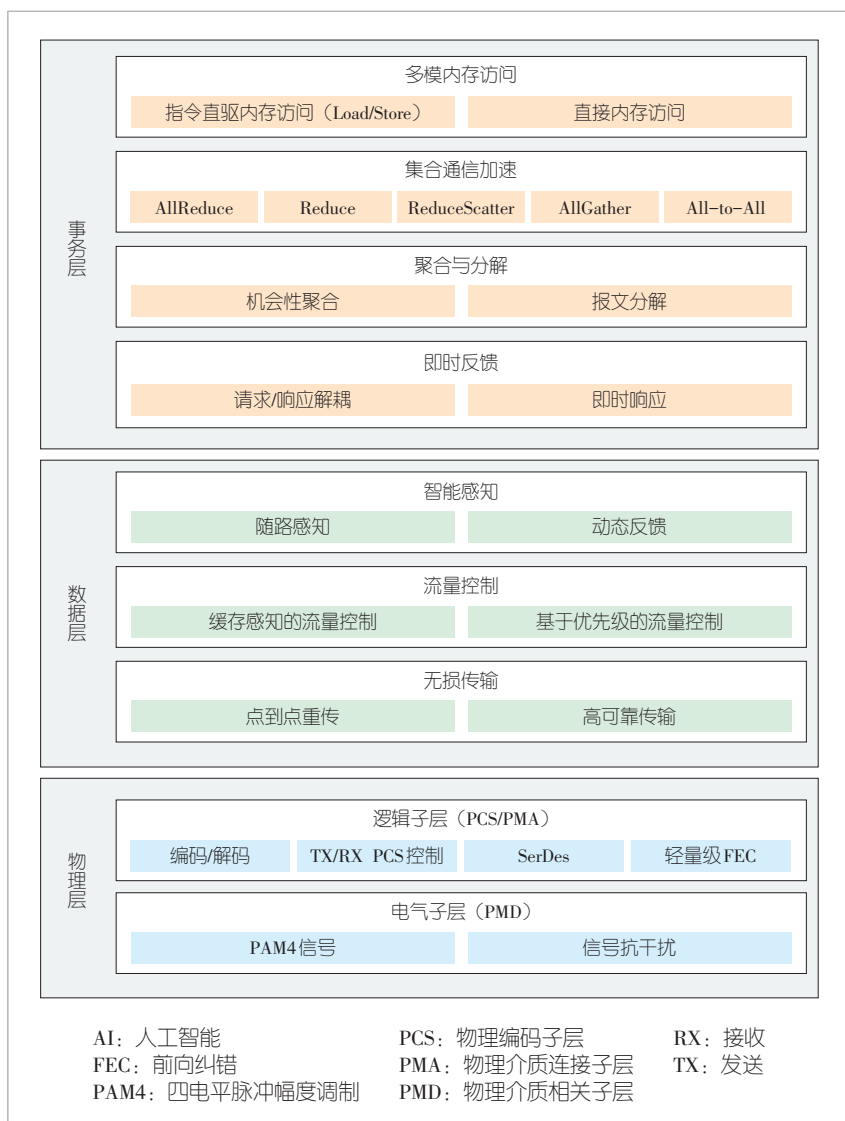


图2 OISA 2.0协议栈

（CCA）模式。其中，极致性能模式面向低时延内存语义访问，通过简化转发流程实现线速处理；智能感知模式支持路径状态采集与动态反馈，实现端到端协同传输优化；集合通信加速模式支持将AllReduce等集合通信操作卸载至交换芯片完成，从而提升分布式训练效率。

3.2 关键技术

OISA构建了一套适配高性能计算与AI场景的高效互联架构。该架构不仅依赖流量控制和重传等传输优化机制，更强调GPU计算单元与卡间互联网络之间的协同，并针对计算侧的业务特性与互联传输痛点进行专门设计。

3.2.1 事务报文动态机会聚合机制

针对推理场景中高频细粒度的通信需求（典型事务大小

为64 B~256 B），OISA在事务层设计了报文动态聚合机制，以降低小型事务传输中协议头部开销的占比。

从技术原理来看，在传统的单事务独立封装模式下，64 B载荷所对应的事务层数据包（TLP）头部开销占比高达20%~40%；对于无载荷的请求或响应类报文，该占比进一步上升。OISA采用动态机会聚合策略，在满足时延约束的前提下，将同目的地址、同类型的多个小型事务报文聚合为单个TLP进行传输，从而有效提升传输效率。

如图3所示，聚合报文将固定协议开销分摊至多个子事务，从而显著提升载荷率与带宽利用率。同时，接收端配备硬件级并行解析单元，依托预定义的载荷分隔标识和流水线处理架构，实现报文的快速解包，确保整体传输效率。

3.2.2 集合通信加速

OISA以“计算卸载”重构分布式通信架构，将AllReduce、ReduceScatter、AllGather、All-to-All等大模型训练场景高频调用的集合通信计算任务，从GPU侧卸载至交换芯片集成的专用计算单元执行。

如图4所示，以AllReduce操作为例，传统直连拓扑通常采用参数GPU集中规约或环形AllReduce等方案。在此模式下，GPU除承担梯度计算任务外，还需参与中间结果的收发、聚合与同步过程。这不仅消耗片上算力资源，亦因多轮端到端数据交互而产生高昂的通信开销。即便借助交换拓扑扩展了互联规模，缺乏计算能力的普通交换芯片仍仅承担报文转发功能，并未改变GPU主导集合通信计算的本质，中间结果反复传输所带来的带宽占用与延迟瓶颈仍未得到有效缓解。

OISA则将AllReduce等集合通信的计算任务直接卸载至交换芯片内部完成。各GPU仅需将本地梯度发送至交换芯片，交换芯片的专用计算单元在报文转发的同时，同步完成求和、求均值等聚合运算，并将最终结果返回至通信组内各节点。该模式从根本上削减了GPU间的中间数据交互量，大幅缩短了数据传输路径与集合通信的整体完成时间。与此同时，GPU算力得以从通信任务中释放，专注于模型训练，从而显著提升分布式集群的整体计算效率。

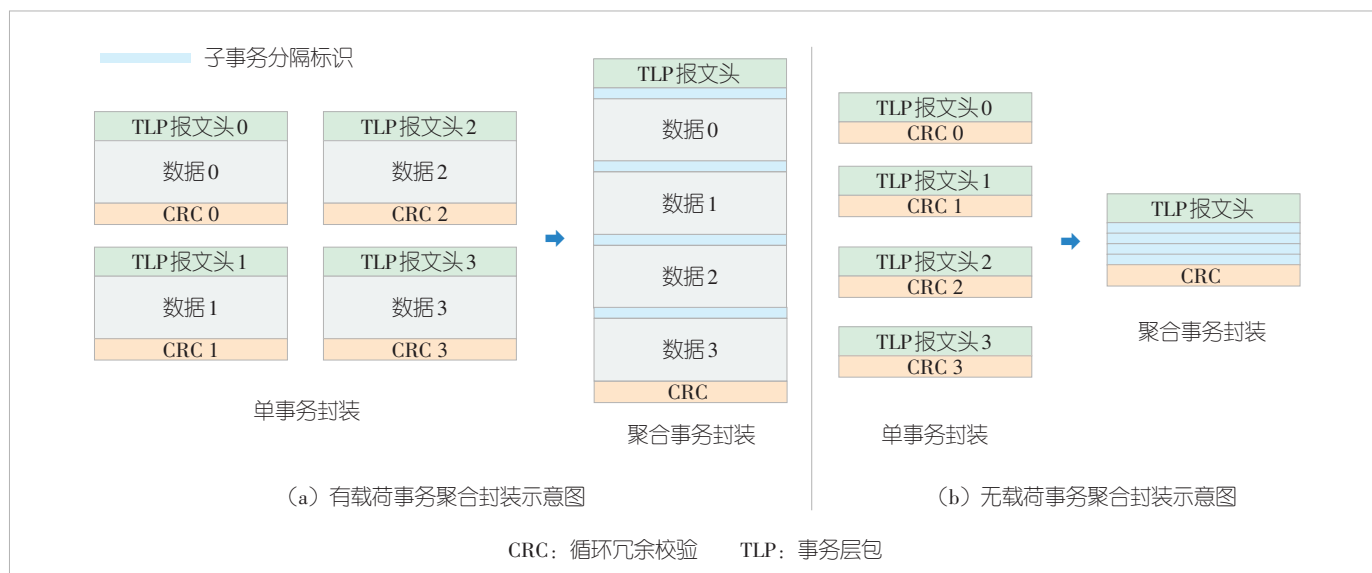


图3 聚合报文格式示意图

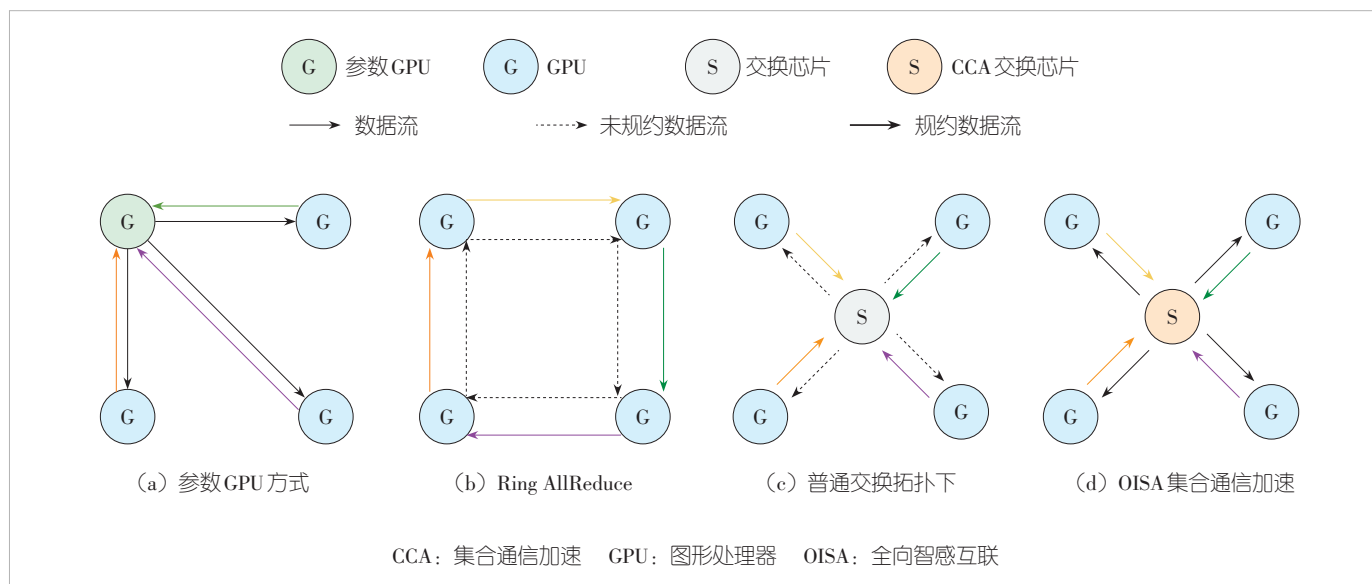


图4 集合通信加速示意图（以规约操作为例）

3.2.3 稀疏感知加速机制

针对 MoE 模型训练推理中因跨 GPU 通信突发性强、分布不均匀、小包占比高等特征所引发的链路拥塞与时延劣化问题，OISA 引入了稀疏感知加速机制。该机制通过在事务报文头部嵌入稀疏感知标签，使 GPU 能够标记当前激活的专家状态；交换芯片则据此仅对活跃数据进行聚合，并动态优化转发调度策略。同时，结合报文聚合技术降低协议开销，有效缓解非对称流量模式所导致的性能衰减，从而充分适配超大参数量 MoE 模型对高效通信的严苛需求。

3.2.4 智能感知

OISA 的智能感知模式通过“发送端主动探测”与“接收端主动告警”两种机制实现随路感知和动态反馈，为传输控制提供精准依据，有效规避拥塞瓶颈。

如图5所示，发送端主动探测机制由发送端 GPU 在报文中嵌入轻量化感知标签，沿数据传输路径随路采集各节点状态信息，并记录潜在瓶颈点。当接收端检测到标签所携带的状态信息达到预设阈值（例如队列占用率超过 70%）时，即向发送端反馈响应信息。发送端收到反馈后，识别并判断拥塞点的具体位置及拥塞程度，进而通过降低发送速率或调整

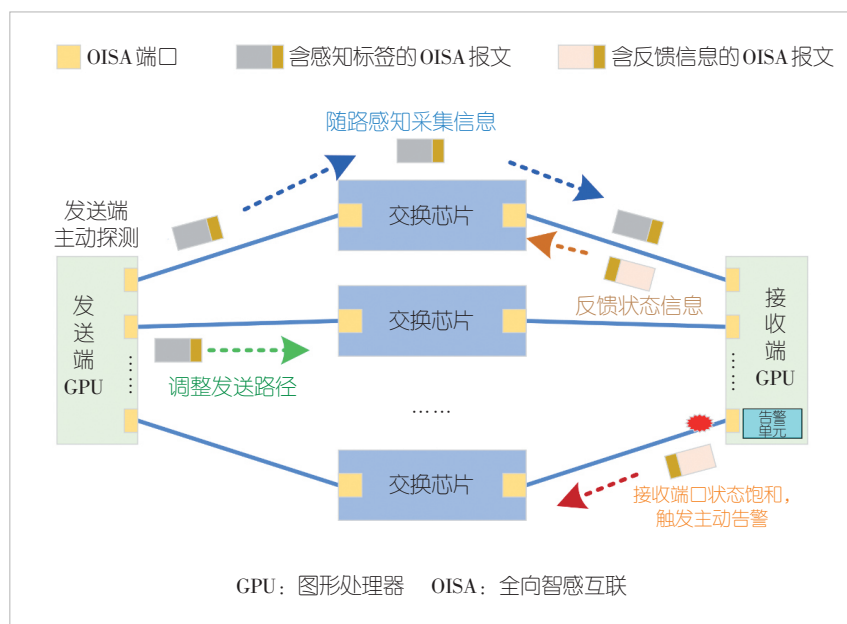


图5 智能感知示意图

转发路径等手段，防止拥塞状况进一步恶化。

接收端主动告警机制则从接收侧提供兜底保障。当接收端 GPU 的接收性能出现持续下降时，该机制主动向发送端反馈状态信息，以应对发送端主动探测未能覆盖的接收端侧拥塞风险。

上述两种机制协同互补，使发送端能够全面感知网络链路及接收端的实时状态，并作出快速响应，从而从源头避免因局部链路拥塞所引发的全局传输瓶颈，切实保障大规模 Scale-Up 集群内数据传输的低时延与高稳定性。

4 未来发展方向

尽管 Scale-Up 网络在互联架构与协议技术方面已日趋成熟，但伴随大模型长文本推理、多模态融合、超大规模 MoE 的演进，现有卡间互联体系仍面临一系列新挑战，亟需从内存架构、传输介质、系统协同等层面实现系统性突破。

随着 GPU 单卡算力和互联规模的持续提升，内存访问效率已逐渐成为制约模型部署效率和系统整体性能提升的核心瓶颈^[13]。在当前万亿参数大模型在线推理阶段，键值缓存 (KV Cache) 占用显存比例可超过 90%，且其占用量随对话轮次与上下文长度的增长呈线性递增趋势，导致单卡本地显存难以支撑高并发长序列业务场景。与此同时，传统存储架构资源碎片化、跨卡共享能力不足、显存分配不均等问题，易引发局部资源过载与频繁数据置换，大幅降低推理吞吐并增大会话时延。为此，Scale-Up 互联的范畴应突破当前 GPU 卡间互联的局限，延伸至内存池、专用存储节点等组件。通

过构建硬件级的一致性缓存域，GPU 可直接经由高速总线访问池化内存中的数据，实现计算资源与存储资源的解耦共享，从而有效满足海量 KV Cache 分布式调度的刚性需求。

在传输介质层面，超节点高密度部署需求与电互联物理极限的共同作用，正加速推动光互联技术的升级演进^[14-15]。传统电互联已逐渐逼近物理极限，在高密度布线环境下存在信号衰减严重、传输距离受限等问题，难以支撑跨机箱、跨机柜的无阻塞互联需求。相比之下，光互联具备高带宽密度、低传输损耗与强抗干扰能力等优势，成为突破电互联瓶颈的关键技术路径。未来，共封装光学 (CPO)、光输入输出 (OIO) 片上光互联等光电融合技术的应用，可进一步提升硬件集成度，降低传输功耗；同时，结合光交换技术支持的动态拓扑重构能力，可灵活适

配 MoE 模型非均匀突发流量特征，构建稳定高效的全光互联底座，从而支撑超节点在远距离、大规模场景下的无阻塞通信需求。

在系统协同层面，超节点的规模化部署对全局资源协同提出了更高要求，Scale-Up 互联正逐步成为统筹算力与存储资源高效协同的核心枢纽。传统计算与网络相分离的设计范式，易导致互联带宽与 GPU 算力、存储带宽之间不匹配，进而制约系统整体效能的释放。因此，后续研究应聚焦全栈协同设计理念，在架构设计阶段统筹多组件间的性能适配，充分发挥交换芯片在集合通信加速与流量调度方面的能力，以提升资源利用效率；同时，应加快推进开放互联协议的标准进程，着力破除跨厂商硬件适配的壁垒，推动超节点从硬件堆叠式部署向一体化全局协同的智算架构演进。

5 结束语

随着大模型参数量与算力需求的持续攀升，基于 Scale-Up 互联架构的超节点已成为新一代高密度智算基础设施的核心形态，相关产业共识正逐步形成。当前，全球虽存在多种新型卡间互联协议并行发展的格局，但整体技术路线与演进方向已呈现收敛态势，基于新协议的超节点正加速研发进程，并稳步推进商用部署与规模化落地。

OISA 立足国产自主技术路线，构建了高性能、开放可扩展的 GPU Scale-Up 互联技术体系，有效突破了传统互联在带宽、时延与组网规模等方面的瓶颈，能够精准适配大模型训练推理等数据密集型 AI 业务的严苛需求。目前，OISA

已完成 2.0 版本核心协议规范的制定及关键功能验证, 硬件研发与产业适配工作正稳步推进。

面向未来发展的持续诉求, OISA 将秉持长期演进策略, 不断迭代完善。在架构层面, 持续拓展互联边界, 构建内存共享池, 并优化报文结构与流量调度机制, 进一步提升集群整体性能与运行稳定性; 在传输层面, 深度融合 CPO/近封装光学 (NPO) 光互连与 Chiplet 等先进封装技术, 构建基于光互连的 OISA 超节点, 进一步降低通信时延与系统功耗。与此同时, OISA 将持续强化多厂商硬件的兼容性与生态开放性, 积极推动协议标准化与产业化落地, 为中国超节点的自主可控建设以及大模型技术的创新突破构筑坚实互联底座, 助力智算基础设施向更高性能、更优能效与更广兼容的方向持续演进。

参考文献

- [1] 段晓东, 程伟强, 张昊. 智算网络发展综述 [J]. 中兴通讯技术, 2025, 31(2): 53–62. DOI: 10.12142/ZTETJ.202502008
- [2] Li A, Song S L, Chen J Y, et al. Tartan: evaluating modern GPU interconnect via a multi-GPU benchmark suite [C]//Proceedings of IEEE International Symposium on Workload Characterization (IISWC). IEEE, 2018: 191–202. DOI: 10.1109/IISWC.2018.8573483
- [3] 崔晨, 吴迪, 陶业荣, 等. 多 GPU 系统的高速互联技术与拓扑发展现状研究 [J]. 航空兵器, 2024, 31(1): 23–31. DOI: 10.12132/ISSN.1673–5048.2023.0138
- [4] 华为技术有限公司, 中国信息通信研究院. 超节点发展报告 [R]. 2025
- [5] Song X Y, Zhou D Y, Li K, et al. Survey of intra-node GPU interconnection in scale-up network: challenges, status, insights, and future directions [J]. Future Internet, 2025, 17(12): 537. DOI: 10.3390/fi17120537
- [6] Das Sharma D. PCI-express: evolution of a ubiquitous load-store interconnect over two decades and the path forward for the next two decades [J]. IEEE circuits and systems magazine, 2024, 24(2): 47–61. DOI: 10.1109/MCAS.2024.3373556
- [7] Foley D, Danskin J. Ultra-performance pascal GPU and NVLink interconnect [J]. IEEE micro, 2017, 37(2): 7–17. DOI: 10.1109/MM.2017.37
- [8] Schieffer G, Shi R M, Markidis S, et al. Understanding data movement in AMD multi-GPU systems with infinity fabric [C]//Proceedings of SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2024: 567–576. DOI: 10.1109/scw63240.2024.00079
- [9] UALink Consortium. UALink consortium releases the ultra accelerator Link 200G 1.0 specification [EB/OL]. (2025-04)[2026-

04-09]. https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-Specification-PR_FINAL.pdf

[10] 华为. 灵衢基础规范 [R]. 2025

[11] Open Compute Project. The OCP ESUN 1.0 specification has been released [EB/OL]. (2026-03-10) [2026-04-09]. <https://www.opencompute.org/blog/the-ocp-esun-10-specification-has-been-released>

[12] 中国移动. 全向智感互联 OISA 2.0 技术规范 [R]. 2025

[13] Liu F X, Zhang Q H, Shen H J, et al. HyperOffload: graph-driven hierarchical memory management for large language models on SuperNode architectures [PP/OL]. arXiv [2026-04-09]. <https://arxiv.org/abs/2602.00748>

[14] Saber M G, Jiang Z P. Physical layer standardization for AI data centers: challenges, progress, and perspectives [J]. IEEE network, 2026, 40(2): 147–155. DOI: 10.1109/MNET.2025.3557812

[15] Liu Y H, Cai Z Z, Chen Y X, et al. InfiniteHBD: building datacenter-scale high-bandwidth domain for LLM with optical circuit switching transceivers [EB/OL]. (2025-02-07)[2026-04-09]. <https://arxiv.org/abs/2502.03885>

作者简介



宋晓勇, 中国移动通信有限公司研究院技术专家; 主要从事 Scale-Up 网络、GPU 卡间互联协议及架构、硬件网络加速等方面的研究工作。



李锴, 中国移动通信有限公司研究院技术经理, 主任研究员; 主要从事通算/智算领域的协议、芯片、部件及整机等方面的研究工作。



陈佳媛, 中国移动通信有限公司研究院技术经理, 主任研究员; 主要从事智算网络、多样性算力、新型智算、GPU 卡间互联等方面的研究工作。

智算中心光电混合组网下的流量差异化调度技术研究



Differentiated Traffic Scheduling for Hybrid Optical-Electrical Networking in Intelligent Computing Centers

韩博文/Han Bowen¹, 刘千仞/Liu Qianren²,
廖思忆/Liao Siyi³, 胡雅坤/Hu Yakun¹, 曹畅/Cao Chang¹

(1. 中国联通研究院, 中国 北京 100048;
2. 中国联合网络通信集团有限公司, 中国 北京 100033;
3. 中国联合网络通信集团有限公司上海市分公司, 中国 上海 200082)
(1. China Unicom Research Institute, Beijing 100048, China;
2. China United Network Communications Group Corporation Limited, Beijing 100033, China;
3. China United Network Communications Group Co., Ltd. Shanghai Branch, Shanghai 200082, China)

DOI: 10.12142/ZTETJ.202603005

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260702.1533.001>

网络出版日期: 2026-06-26

收稿日期: 2026-04-06

摘要: 光电混合组网凭借“电交换的灵活可控”与“光交换的高带宽低时延”双重特性, 已成为智算中心网络的核心发展趋势。聚焦智算中心光电混合组网场景, 深入分析AI训练流量的差异化特征, 并提出一种基于流量特征的差异化调度机制。该机制通过构建“感知-决策-执行-反馈”的闭环调度框架, 实现了对数据并行(DP)流量、流水线并行(PP)流量等多种类型流量的精准识别与链路优势匹配。研究成果为智算中心光电混合组网下的流量差异化调度提供了理论依据与实践参考。

关键词: 智算中心; 光电混合组网; 差异化调度; AI训练

Abstract: Optical-electrical hybrid networking, leveraging the dual advantages of flexible controllability of electrical switching and high-bandwidth low-latency of optical switching, is a key development trend in data center networks for artificial intelligence (AI). Focusing on the optical-electrical hybrid networking scenario in AI data centers, the differentiated characteristics of AI training traffic are analyzed in depth, and a traffic-characteristic-based differentiated scheduling mechanism is proposed. By constructing a closed-loop scheduling framework comprising perception, decision, execution, and feedback, accurate identification of different traffic types—including data parallelism (DP) and pipeline parallelism (PP) traffic—as well as matching with the advantages of optical or electrical links, is achieved. The findings provide both theoretical foundations and practical references for differentiated traffic scheduling in optical-electrical hybrid networks within AI data centers.

Keywords: intelligent computing center; optical-electronic hybrid networking; differentiated scheduling; AI training

引用格式: 韩博文, 刘千仞, 廖思忆, 等. 智算中心光电混合组网下的流量差异化调度技术研究 [J]. 中兴通讯技术, 2026, 32(3): 22-29. DOI: 10.12142/ZTETJ.202603005

Citation: Han B W, Liu Q R, Liao S Y, et al. Differentiated traffic scheduling for hybrid optical-electrical networking in intelligent computing centers [J]. ZTE technology journal, 2026, 32(3): 22-29. DOI: 10.12142/ZTETJ.202603005

数字经济与人工智能(AI)的深度融合, 正推动全球AI算力基础设施建设进入高速扩张周期。AI大模型的训练与推理应用导致算力需求呈现几何级增长。以Seedance 2.0为代表的多模态大模型及OpenClaw等开源AI智能体的广泛部署, 进一步加速了算力需求的攀升。在此背景下, 算力供给能力已成为衡量国家数字经济核心竞争力的关键指标。2026年, 中国政府工作报告^[1]首次将“超大规模智算集群”列为新基建工程。这一举措不仅从国家战略层面确立了智算

基础设施的核心地位, 同时也对其底层网络的支撑能力、资源调度效率与整体能效提出了前所未有的严苛要求。

随着智算集群规模持续扩大, 超大规模AI数据中心的发展瓶颈已从单芯片算力逐步转向集群级互连带宽、能耗密度与系统可扩展性。在此条件下, 网络基础设施成为制约智算集群算力充分释放的核心因素。在大规模智算集群的组网架构中, 传统纯电交换网络虽具备调度灵活的优势, 但在超大带宽扩展、超低时延传输和能耗控制方面存在明显短板;

纯光交换网络虽拥有高带宽、低时延和低功耗的特性，却面临突发流量适配能力差、与现有电交换系统兼容性不足的问题，从而限制了其规模化应用。

作为智算网络技术演进的重要风向标，2026年全球光通信大会（OFC）与图形处理器（GPU）技术大会（GTC）为突破上述瓶颈提供了明确的技术路径。在OFC大会上，Google、Nvidia等厂商展示了光电路交换机（OCS）在张量处理器（TPU）/GPU集群中的规模化应用案例，证实光电混合架构是突破传统电交换网络带宽与功耗瓶颈的重要方向；在GTC 2026上，Nvidia发布了Spectrum-X共封装光学（CPO）交换机，并提出“光铜并行、光电混合”的技术战略。

在此背景下，融合“电交换灵活可控”与“光交换高带宽低时延”双重优势的光电混合组网方式，成为突破智算中心网络瓶颈、支撑下一代高性能智算集群的核心发展趋势。其中，差异化流量调度技术作为光电混合组网的关键支撑，通过为不同智算业务设计定制化调度策略，实现光/电网络资源的协同优化与动态匹配，是提升组网整体性能、释放超大规模智算集群算力的重要途径。基于上述分析，本文聚焦智算中心光电混合组网场景，开展差异化调度技术研究，以期突破现有调度模式的局限，为提升智算中心网络资源利用率与算力释放效率提供技术参考。

1 全球研究现状

1.1 Google

Google作为智算集群光电混合组网技术的先行者，其Jupiter组网架构与第7代自研TPU芯片（TPUv7，代号Ironwood）的深度协同，构建了面向超大规模AI大模型训练的高效网络解决方案，为光电混合组网及调度技术提供了重要实践参考^[2]。

Jupiter架构初期采用三级Clos拓扑，由机架顶交换机（ToR）、聚合块及脊层层级互联构成，然而该拓扑难以适配超大规模数据中心的增量扩展需求与智算负载的通信特征，主要瓶颈体现在4个方面：

1) Spine层预部署导致带宽降级：Spine层需一次性预部署，

后续新代际聚合块的高速链路会因Spine层低速链路被降级，造成算力与网络带宽失衡；

2) 增量升级成本高、易中断：Spine层升级需全网重新布线，操作复杂且影响生产流量，无法实现计算/存储硬件的渐进式更新；

3) 功耗与成本冗余：Spine层交换机及配套光模块带来额外的资本支出（CAPEX）和功耗，且链路利用率低；

4) 无法适配异构速率：不同代际如100G、200G或400G链路难以共存，无法匹配AI集群的差异化带宽需求。

2022年，Google提出Jupiter架构的核心演进方案^[3]（如图1）。该方案通过引入基于微机电系统（MEMS）的OCS，构建了数据中心互联层（DCNI），能够为数据中心网络创建任意逻辑拓扑，实现动态拓扑重构；同时，通过集中式软件定义网络（SDN）控制，进行流量工程优化与自动化网络运维。该演进方案推动Jupiter架构实现了从传统Clos拓扑向块级直连拓扑的范式升级。相较原Clos拓扑，网络带宽与容量提升5倍，功耗降低41%，CAPEX降低30%，充分验证了光电混合组网技术在超大规模数据中心及智算集群中应用的可行性与工业价值。

Google自研的TPU是面向AI大模型训练与推理的专用加速芯片。历经多代技术迭代，TPU已形成规模化集群部署能力。在Jupiter光电混合组网架构的基础上，Google结合TPU自身的芯片互联特性进行定制化创新，打造了适配超大规模智算需求的TPU专属集群组网架构与调度策略。

TPU集群组网的核心思路为“光电分域、分层互联、弹性扩展”，其整体架构如图2所示。该架构以 $4 \times 4 \times 4$ Cube为最小算力组网单元，构建“单元内电交换、单元间光交

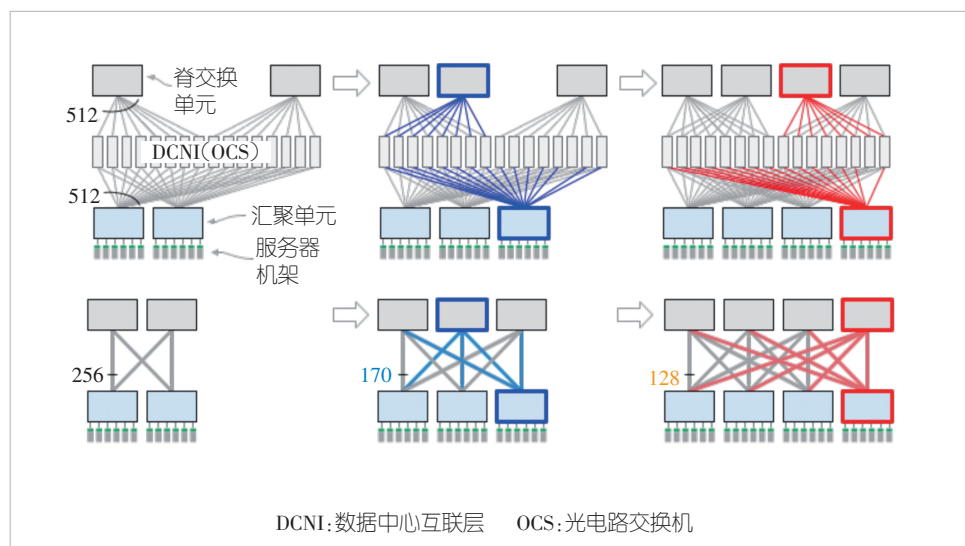


图1 基于DCNI层支持Jupiter架构渐进式扩展

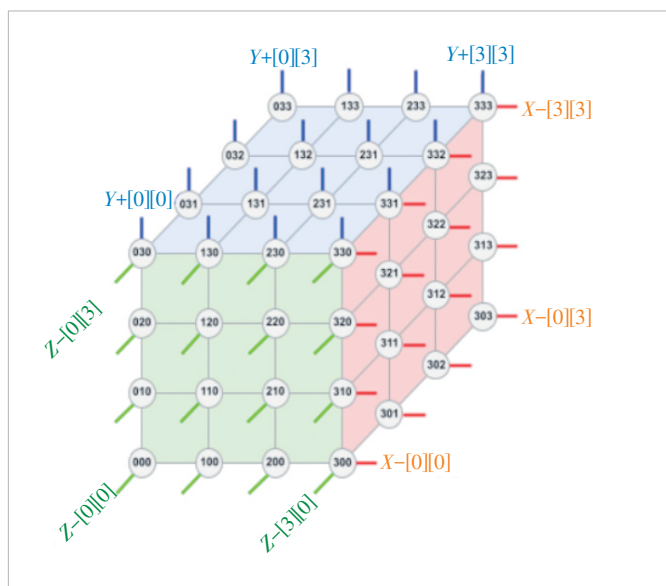


图2 TPU Cube示意图

换”的分层结构：Cube内部通过芯片间互联链路实现全电信号紧耦合通信；多个Cube之间则基于OCS构建三维Torus（3D Torus）拓扑。OCS按三维正交方向部署，支持毫秒级的动态拓扑重构。

TPU集群组网调度策略深度贴合AI大模型训练的通信特征，通过算力切片实现集群资源的弹性分配与多租户隔离；针对稠密/稀疏计算的差异化流量实施分层调度，匹配拓扑高带宽维度与优化拓扑；围绕数据、模型、流水线并行策略动态重构光拓扑，适配不同并行方式的通信需求；依托OCS的快速光路切换能力实现故障自愈，在不中断训练任务的前提下完成流量重路由，最终实现算力与网络资源的精准匹配，最大化释放TPU集群的算力性能。

该调度策略深度贴合AI大模型训练的通信特征，具体体现在5个方面：

- 1) 通过算力切片实现集群资源的弹性分配与多租户隔离；
- 2) 针对稠密计算与稀疏计算

的差异化流量实施分层调度，匹配拓扑的高带宽维度，并优化网络拓扑结构；

3) 围绕数据并行、模型并行与流水线并行策略，动态重构光拓扑，以适配不同并行方式的通信需求；

4) 依托OCS的快速光路切换能力实现故障自愈，在不中断训练任务的前提下完成流量重路由；

5) 最终实现算力与网络资源的精准匹配，最大化释放TPU集群的算力性能。

1.2 Nvidia

在2023年OFC上，Nvidia针对传统数据中心网络物理布线固化、故障自愈机制存在明显缺陷的问题开展了研究。现有故障恢复方案多依赖路由重调整或冗余硬件部署，不仅易导致集群性能降级、硬件资源利用率低下，而且无法解决叶交换机故障引发的服务器断连问题。

针对上述不足，Nvidia提出将SDN控制能力下沉至物理层（L1）的技术方案^[4]，如图3所示。该方案基于OCS构建

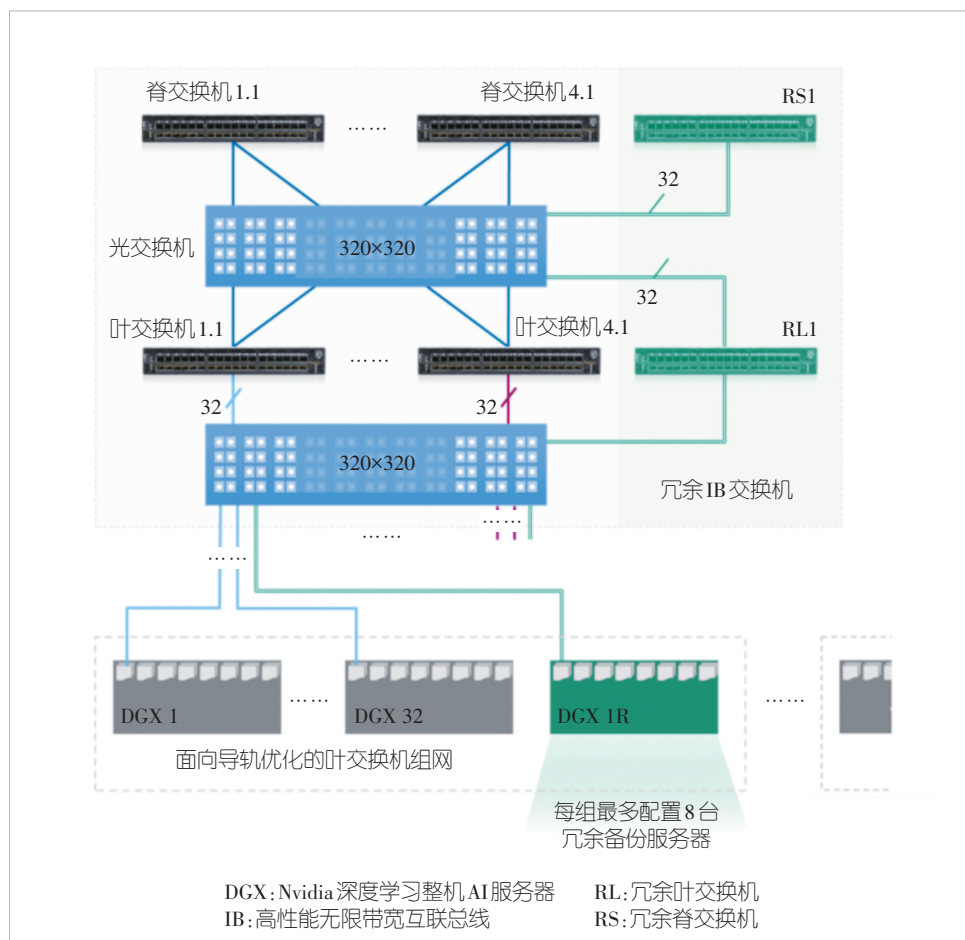


图3 高可用架构图

可编程光交换矩阵，在两层胖树（Fat-tree）拓扑中嵌入OCS并部署冗余叶/脊交换机。同时，通过自研光交换管理器（OFM）实现对OCS的集中控制，完成故障设备的秒级物理链路切换与拓扑重构。

该方案在由4台DGX服务器与14台InfiniBand（IB）交换机构建的测试床中进行了实验验证。结果表明，当叶/脊交换机发生故障时，集群可在数秒之内恢复至满性能状态，有效避免了因叶交换机故障导致的应用崩溃。该方案从根本上解决了传统路由自愈方案中存在的性能降级与冗余硬件利用率低的问题。

总体而言，该方案实现了物理层拓扑的动态可编程，为轻量化冗余硬件达成了高效的故障自愈，为高性能计算（HPC）及AI集群的网络高可用设计提供了全新的技术路径。

2024年，Nvidia在2023年技术成果的基础上，针对大模型预训练集群中存在的三大痛点——传统多级胖树拓扑固化、功耗与成本高昂、难以适配任务通信特征，提出了基于OCS的拓扑动态适配方案^[5]。

该方案利用OCS层替代传统Fat-tree架构中的高层电交换结构。具体实施路径如下：首先，依据大语言模型（LLM）、深度学习推荐模型（DLRM）等典型任务的预期通信模式，通过资源管理工具完成算力与网络资源的分配；随后，利用OCS动态构建环形、全连接等适配物理拓扑，实现网络架构与业务负载的深度适配。

在性能表现方面，该方案不影响单任务的带宽资源，集群资源利用率与传统Fat-tree架构的差距控制在1%以内（如图4）。同时，该方案可降低网络功耗超过50%、网络成

本超过30%，为超大规模AI训练集群提供了高效低耗的组网支撑。

1.3 百度

针对Meta提出的“Rail-only”单导轨组网架构，百度提出了以OCS替换二层以太网分组交换平面（EPS）的改进方案^[6]，如图5所示。该方案将每个TOR组内的流水线并行（PP）流量控制于单台TOR交换机所辖的GPU之间，并将每个Group中编号相同的TOR交换机上联至同一个OCS交换平面。

同时，为支持全互连（all-to-all）通信，百度进一步提出了基于OCS对大规模二层EPS平面进行扩展的网络架构，如图6所示。在该架构中，每个机架顶（ToR）交换机需拆分出部分端口用于连接相应的OCS平面，同时将剩余端口连接至叶交换机。该设计既保留了每个二层EPS平面内的全互连通信能力，又具备了规模扩展以支持更大规模训练任务的能力。

2 流量差异化调度机制

2.1 面临挑战

在光电混合组网场景下，流量调度面临以下核心挑战：

1）流量特征识别与适配难度大。AI训练等高性能计算场景中的流量具有“突发式、高并发、多类型”的特征。例如，数据并行（DP）流量为大带宽、高并发的全局同步流量，流水线并行（PP）流量为低时延、顺序敏感的点对点流量。如何精准识别不同类型流量的特征，并针对性地匹配光链路与电链路的各自优势，是调度面临的核心难点。

2）链路协同与负载均衡难题。光链路与电链路在带宽、时延、转发机制等方面存在固有差异，传统等价多路径（ECMP）负载均衡算法难以直接适配，容易导致链路负载不均。光链路带宽虽高但灵活性不足，若流量调度不当，可能出现光链路闲置而电链路拥塞的状况。同时，光链路的切换延迟可能引发流量中断，进而影响业务连续性。

3）调度策略的动态适应性不足。现有调度方案多基于预设规则或简单的实时流量监测，难以

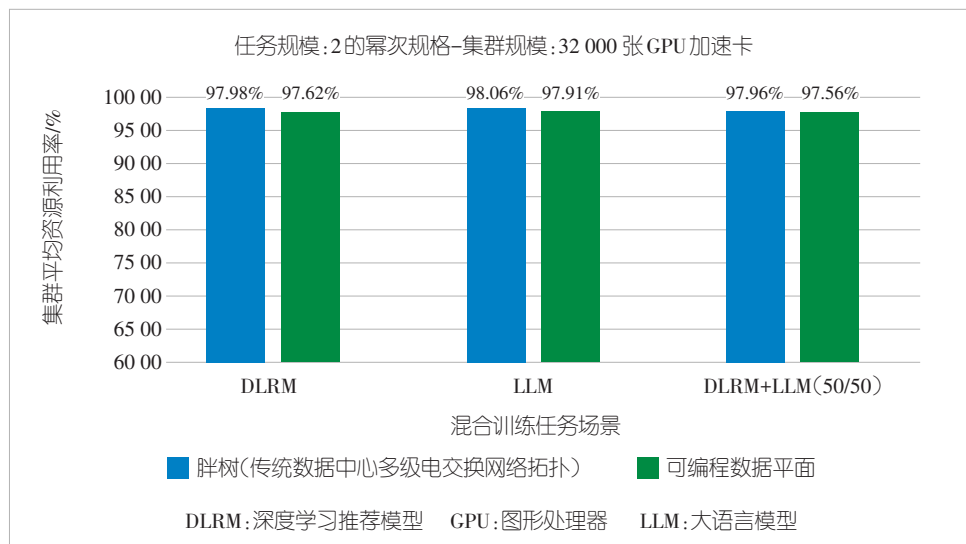


图4 基于DCNI层支持Jupiter架构渐进式扩展

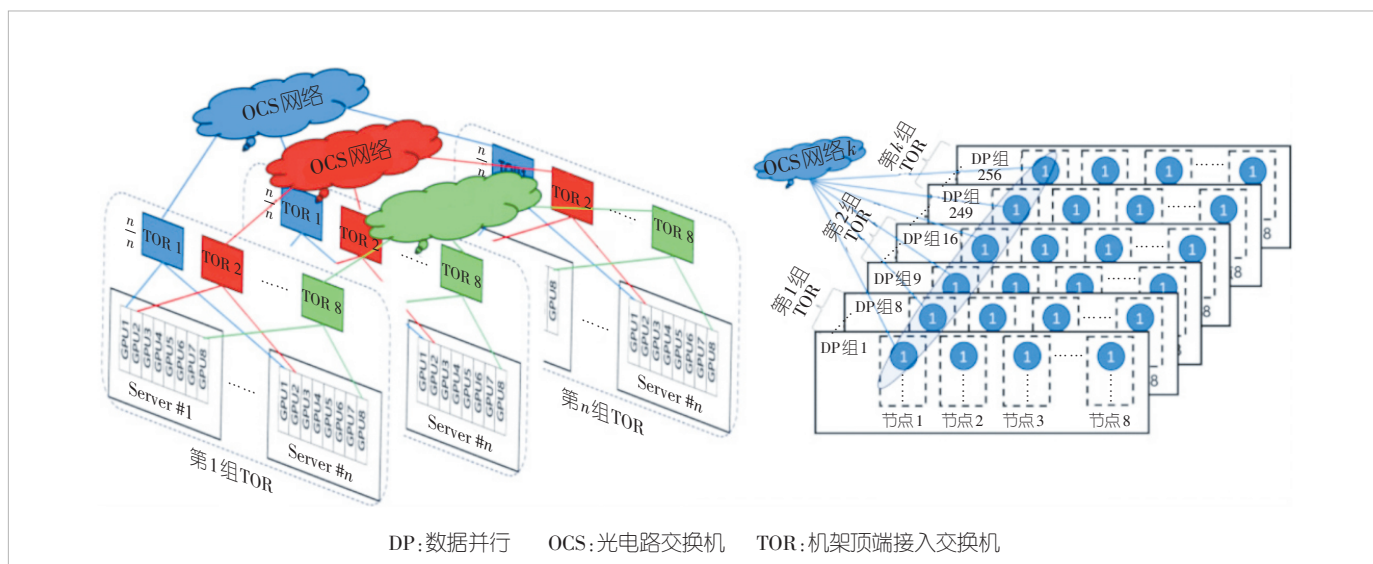


图5 基于DCNI层支持Jupiter架构渐进式扩展

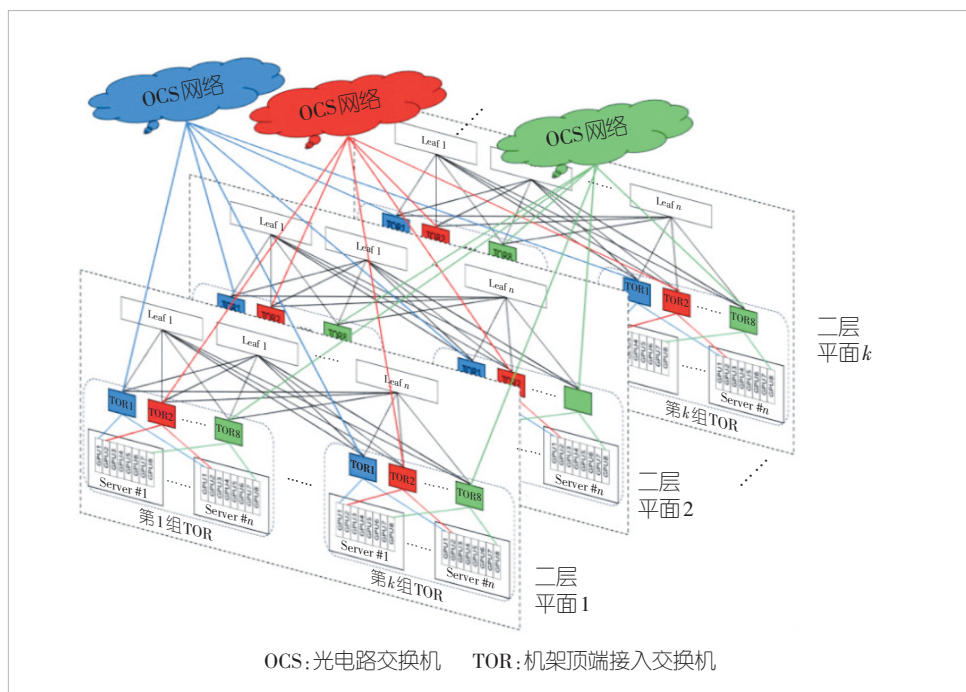


图6 大规模二层EPS平面基于OCS进行扩展

应对复杂场景下流量的动态变化，从而导致网络性能下降。此外，不同业务场景（如训练、推理）的流量特征差异较大，通用调度策略难以兼顾各类场景的差异化需求。

4) 拓扑约束与扩展性限制。基于OCS的光交换虽可实现毫秒级或秒级的物理层光路切换与拓扑动态调整，但在大规模集群中，若所有节点间均通过OCS实现全互连，所需端口数与连接数将随节点规模呈平方级增长，导致光路调度与维护复杂度显著上升。

5) 运维与验证复杂度高。光电混合组网涉及光交换机、电交换机、SDN控制器、业务平台等多设备、多系统的协同，调度策略的配置、部署与验证流程较为复杂。具体而言，如何验证调度策略的有效性、如何定位光/电链路切换过程中的故障、如何评估调度方案对业务性能的实际提升效果，均需依赖专业的运维工具和测试方法，这进一步增加了运维成本。

2.2 设计框架

本方案的核心设计思路为“流量特征驱动、链路优势匹配、动态智能调度”，即针对AI训练场景中的DP、PP等核心流量类型，构建差异化调度机制，以实现

现光链路与电链路资源的最优利用。

方案采用“感知-决策-执行-反馈”的闭环设计框架。该框架包含以下4个核心模块：

1) 感知模块。该模块包括模型切分策略感知与网络状态感知两个子功能。

(1) 模型切分策略感知：获取AI训练任务的参数配置，包括张量并行（TP）/PP/DP切分方式、秩（rank）划分、全局批大小（GBS）设置等，为调度决策提供数据支撑。

(2) 网络状态感知：通过管控平台获取智算网络的链路层发现协议（LLDP）拓扑连接关系（包括交换机与智算服务器之间、交换机与交换机之间的拓扑连接），同时利用 Telemetry 遥测功能实现对网络流量的毫秒级监测。

2) 调度决策模块。该模块基于感知模块采集的数据，构建差异化调度规则库。核心规则如下：

(1) DP 流量调度：DP 流量是 AI 训练中带宽需求最大、对性能影响最为关键的流量类型，其调度机制是本方案的核心内容。

a) 流量识别：通过网络拓扑连接关系与 AI 训练参数（DP 切分方式、rank 划分等），可预先计算出 DP 流量的通信关系，进而实现流量识别。

b) 链路选择优先级：第 1 优先级为 OCS 光链路，第 2 优先级为光电混合链路（采用 ECMP 负载分担），第 3 优先级为纯电链路。

(2) PP 流量调度：PP 流量具有低时延、顺序敏感、数据量相对较小的特点。调度策略根据链路负载动态选择光链路或电链路——若光链路负载较低，则优先选择光链路；若光链路负载过高，则切换至电链路。

3) 执行模块。该模块通过管控平台将调度决策转换为具体的配置指令，下发至光交换机与电交换机。例如，通过配置访问控制列表（ACL）规则实现流量的链路定向转发，或通过边界网关协议（BGP）调整路由策略以实现链路切换。

4) 反馈优化模块。该模块实时监测调度执行后的网络性能（如链路负载、时延、带宽利用率）与业务性能（如 AI 训练的每秒浮点运算次数（TFLOPS）、迭代耗时），并将监测结果反馈至调度决策模块。若网络性能或业务性能未达预期，则动态调整调度规则，优化链路分配策略。

3 试验验证与结果分析

3.1 试验环境与测试方案

试验环境采用典型的 Spine-Leaf 两层架构，在 Spine 层引入异构 OCS 设备，逻辑拓扑如图 7 测试拓扑所示。

1) 计算层：部署 4 台 AI 服务器，每台服务器搭载 8 张 AI 加速卡（共 32 张 GPU），单台服务器

配置 8 个 200G 远程直接内存访问 over 融合以太网（RoCE）网卡，单机总带宽为 1.6 Tbit/s，构成一个标准的训练集群。

2) 接入层：每台 Leaf 交换机下行连接一台服务器上的 8 个网卡，上行通过 400G 链路连接至 Spine 层。上行链路中，一半连接至电交换机，另一半连接至 OCS。Leaf 交换机的上下行带宽保持 1:1 无收敛配置。

3) 核心层：部署 1 台电交换机和 1 台 OCS 光交换机。每台逻辑 Leaf 交换机同时上行连接至电 Spine 与光 Spine，通过 BGP 或静态路由控制流量的转发路径。

3.2 光电性能对比测试

在双机（16 卡）训练场景下，针对二层只走电交换机或二层只走光交换机的场景，我们测试了 Llama2（6.7B）模型的训练性能，逻辑拓扑如图 8 所示。

由于节点数少，网络拓扑结构相对简单，流量冲突概率低。测试结果如表 1 所示。

测试结果表明：在小规模、无阻塞的简单拓扑中，OCS 完全可以替代电交换机，且能提供更优的时延表现。

3.3 调度策略对比测试

在 32 卡集群环境下，我们开展了 Llama2（34B）模型的训练性能对比实验，分别评估了不同调度策略下的性能表现。同时，通过调整全局批大小（GBS），改变了训练过程中的“计算通信比”，从而放大不同网络架构之间的性能差异。表 2 展示了 32 卡环境下、不同 GBS 取值时的训练性能对比结果（TP=8，PP=2，DP=2）。

我们对测试结果进行了分析：

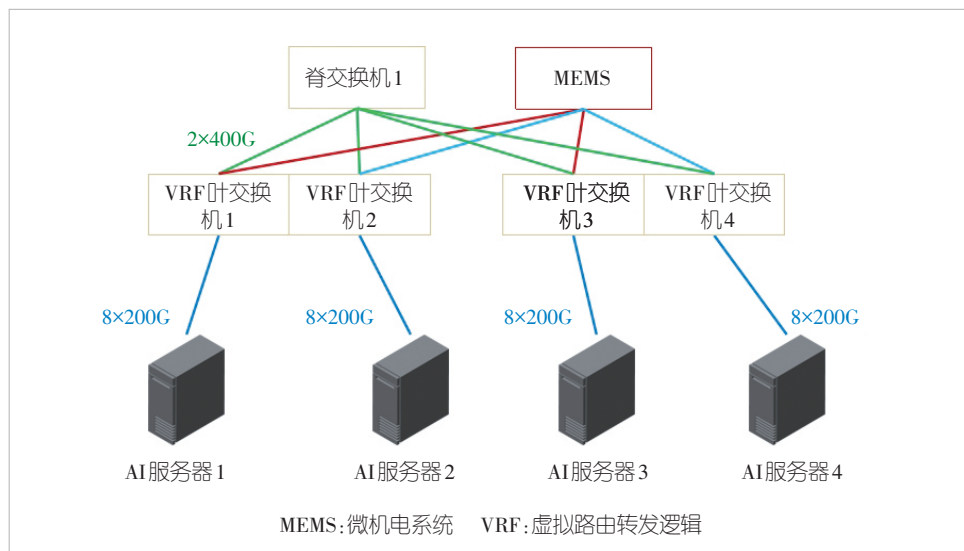


图7 测试拓扑

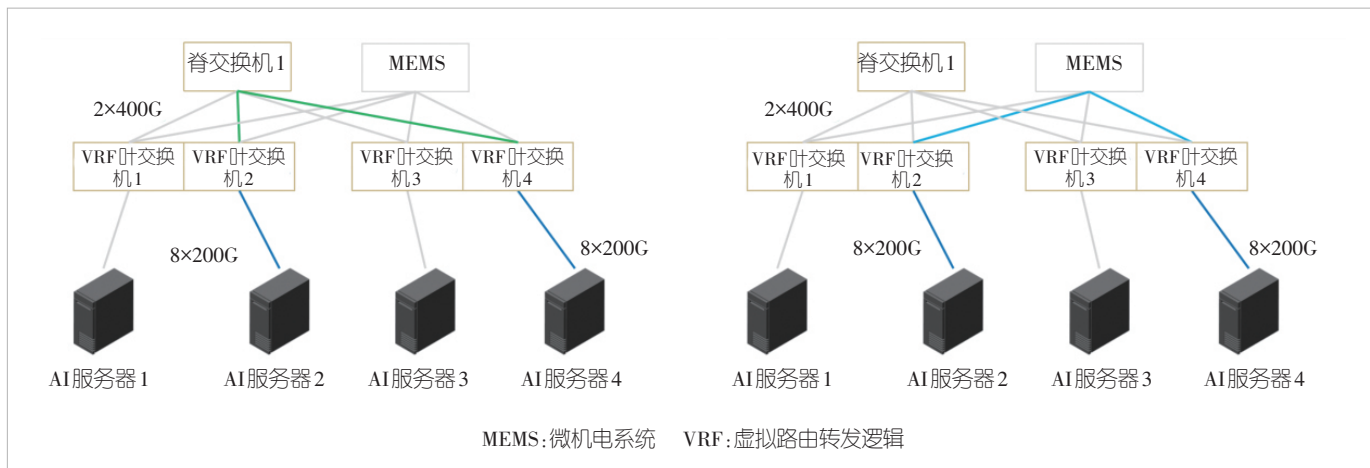


图8 16卡训练示意图

表1 双机训练场景性能对比

配置场景	网络路径	平均耗时/ms	平均TFLOPS
TP8 PP1 DP2	电交换(spine1)	72 076.36	73.53
TP8 PP1 DP2	光交换(mems)	72 061.29	73.56
TP8 PP2 DP1	电交换(spine1)	78 087.56	67.87
TP8 PP2 DP1	电交换(mems)	77 340.46	68.52

DP:数据并行 TFLOPS:每秒万亿次浮点运算
PP:流水线并行 TP:张量并行

表2 不同GBS场景下的性能对比

GBS	DP走光 (TFLOPS)	PP走光 (TFLOPS)	性能差异/%
2	25.82	22.53	14.6
16	86.14	78.94	9.1
32	100.19	96.66	3.7
64	111.39	108.92	2.3
128	117.26	114.87	2.0

DP:数据并行 PP:流水线并行
GBS:全局批大小 TFLOP:每秒万亿次浮点运算

1) 当GBS=2时,系统实际算力仅能达到约25 TFLOPS,约为峰值算力的25%。此时网络通信极为频繁,协议栈处理开销被明显放大。即便如此,DP流量经光链路传输的方案(25.82 TFLOPS)仍优于PP流量经光链路传输的方案(22.53 TFLOPS),性能提升幅度达14.6%。该结果验证了在高频通信压力下,将DP流量优先调度至低时延OCS链路的有效性。

2) GBS=16处于典型的性能“过渡区”,此时计算与通信的时间开销大致相当,网络调度策略的选择对最终性能影响最大。DP流量经光链路传输的方案较PP流量经光链路传输的方案性能高出9.1%。该结果进一步验证了在光电混合

架构中,将DP流量优先调度至OCS链路的核心策略。

3) 当GBS增大至128时,DP流量经光链路与PP流量经光链路的性能差距缩小至2%。此时GPU绝大部分时间处于满负荷计算状态,网络通信时间占比极低,网络架构选择的容错率显著提高。

4 总结与展望

本文聚焦智算中心光电混合组网场景,针对AI训练中流量的差异化调度需求,提出了一种基于流量特征的差异化调度机制。该机制通过构建“感知—决策—执行—反馈”的闭环调度框架,实现了对DP流量、PP流量等不同类型流量的精准识别,并据此完成光链路与电链路的优势匹配。

未来工作将针对混合专家模型(MoE)及推理场景进行专项优化,以进一步提升流量识别精度与调度策略的动态适应性,从而推动光电混合组网技术在智算中心中的持续发展与应用。

参考文献

- [1] 中华人民共和国国务院. 2026 年政府工作报告 [R/OL]. (2026-03-13) [2026-04-07]. http://www.gov.cn/yaowen/liebiao/202603/content_7062620.htm
- [2] 叶通, 胡卫生. 大规模智算中心光电交换网络架构演化综述 [J]. 电信科学, 2025, 31(4): 32-44. DOI: 10.11959/j.issn.1000-0801.2025116
- [3] Poutievski L, Mashayekhi O, Ong J, et al. Jupiter evolving: transforming Google's datacenter network via optical circuit switches and software-defined networking [EB/OL]. [2026-04-10]. <https://web.stanford.edu/class/cs244/papers/poutievski-sigcomm22.pdf>
- [4] Patronas G, Syrivelis D, Bakopoulos P, et al. Software-defined, programmable L1 dataplane: demonstration of fabric hardware resilience using optical switches [C]//2023 Optical Fiber Communication Conference (OFC). IEEE, 2023: Th2A.15
- [5] Bakopoulos P, Patronas G, Terzenidis N, et al. Photonic switched

networking for data centers and advanced computing systems [C]//2024 Optical Fiber Communication Conference (OFC). Optica Publishing Group, 2024: M2G.1

- [6] 朱宸, 周渭, 王佩龙. 可重构 OCS 技术在大模型预训练中的应用 [J]. 光通信研究, 2024 (5): 29–38. DOI:10.13756/j.gtxyj.2024.240049

作者简介



韩博文, 中国联通研究院高级研究员, 工程师; 主要研究领域为智算网络关键技术、管控调度技术等。



刘千仞, 中国联合网络通信集团有限公司建设发展部高级工程师; 主要研究领域为云计算、算力网络等。



廖思忆, 中国联合网络通信集团有限公司上海市分公司云网运营中心高级总监; 主要研究领域为下一代互联网、算力网络、数据中心网络等。



胡雅坤, 中国联通研究院光传输领域研究员; 从事超高速传输与智算光互联方向的理论与工程化应用, 深度参与中国联通算力智联网建设工作, 围绕智算内和智算间光互联开展技术创新。



曹畅, 中国联通研究院下一代互联网研究部总监, 正高级工程师; 主要研究领域为算力网络、IPv6+ 网络新技术、未来网络体系架构等。

面向边缘场景的高效大模型分层部署机制



An Efficient Hierarchical Deployment Mechanism for Large Models in Edge Scenarios

万卓然/Wan Zhuoran¹, 谢人超/Xie Renchao^{1,2},
王媛/Wang Yuan³, 王芳/Wang Fang³,
唐琴琴/Tang Qinqin¹, 黄韬/Huang Tao^{1,2}

(1. 北京邮电大学网络与交换技术全国重点实验室, 中国 北京 100876;
2. 紫金山实验室, 中国 南京 211111;
3. 中国电子科技集团公司第五十研究所, 中国 上海 200063)
(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Purple Mountain Laboratories, Nanjing 211111, China;
3. The 50th Research Institute of China Electronics Technology Group Corporation, Shanghai 200063, China)

DOI: 10.12142/ZTETJ.202603006

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260626.1024.008>

网络出版日期: 2026-06-26

收稿日期: 2026-03-25

摘要: 针对边缘设备资源受限导致大模型推理延迟高、吞吐量低的问题, 提出了一种高效的分层部署优化机制, 构建了“数据供给-空间保障-时间重叠”协同架构。具体而言, 首先利用内存映射(mmap)实现权重按需加载以打破大模型加载的输入/输出瓶颈。其次通过静态内存预分配与指针偏移复用, 彻底规避垃圾回收引发的系统抖动。最后提出基于权重矩阵粒度的流水线调度, 拆解并行计算与加载任务, 有效重叠显存访问延迟。实验表明, 该机制在不增加显存占用的前提下显著提升了推理效率, 推理吞吐量提升1倍以上, 有效支撑了边缘侧大模型的实时推理需求。

关键词: 边缘计算; 大模型部署; 推理优化; 分层推理

Abstract: Addressing the issue of high inference latency and low throughput caused by resource constraints on edge devices, we propose an efficient hierarchical deployment optimization mechanism that establishes a collaborative architecture featuring “data supply-spatial assurance-time overlap”. Specifically, we first leverage memory mapping (mmap) to enable on-demand weight loading, thereby breaking through the input/output bottleneck associated with large model loading. Next, by statically pre-allocating memory and reusing pointer offsets, we completely eliminate system jitter caused by garbage collection. Finally, we introduce a pipeline scheduling approach based on the granularity of weight matrices, which decomposes parallel computation and loading tasks, effectively overlapping the latency of graphics processing unit (GPU) memory access. Experimental results demonstrate that this mechanism significantly improves inference efficiency without increasing GPU memory usage, more than doubling the inference throughput and effectively supporting real-time inference requirements for large models at the edge.

Keywords: edge computing; large model deployment; inference optimization; hierarchical inference

引用格式: 万卓然, 谢人超, 王媛, 等. 面向边缘场景的高效大模型分层部署机制 [J]. 中兴通讯技术, 2026, 32(3): 30-36. DOI: 10.12142/ZTETJ.202603006

Citation: Wan Z R, Xie R C, Wang Y, et al. An efficient hierarchical deployment mechanism for large models in edge scenarios [J]. ZTE technology journal, 2026, 32(3): 30-36. DOI: 10.12142/ZTETJ.202603006

随着智能家居设备的逐步普及, 家庭组网已演进为高度互联、数据密集的智能家居生态。在此背景下, 大模型应用正加速从云端向边缘侧迁移^[1-2]。然而, 依赖云端应用程序编程接口 (API) 的集中式部署在工程落地中面临严峻挑战:

在隐私层面, 家庭敏感数据的云端传输极易引发泄露风险, 且受制于严格的数据本地化监管刚性约束; 在实时性层面, 云端固有的网络延迟与拥塞难以满足跌倒告警、燃气处置等关键任务的快速响应需求, 可能诱发安全隐患; 在成本层面, 高昂的云端算力租赁成本严重制约了该技术的规模化普及^[3-4]。

为突破云端集中式处理在隐私安全、网络延迟及带宽成

基金项目: 国家自然科学基金项目 (92367104)

本上的固有局限，大模型部署的重心正在从云端向边缘侧迁移^[5]。这一技术路径的转向，直接推动了边缘计算范式成为支撑家庭智能化转型的核心架构：边缘计算通过将计算能力下沉至距离数据源头最近的边缘设备，构建了一种具备数据本地化处理与快速响应能力的新型计算体系^[6-7]。

然而，在面向大模型边缘部署时，现有主流框架往往难以兼顾显存效率与推理速度。Hugging Face Transformers^[8]虽生态完善，但原生机制在处理长序列时显存碎片化严重。vLLM^[9]虽通过 PagedAttention 实现了云端高吞吐推理，却因架构复杂难以适配边缘设备。llama.cpp^[10]通过 INT4/INT8 量化实现大模型本地部署，但量化精度损失在复杂任务中较为明显。AirLLM^[11]虽利用分层加载突破了显存瓶颈，但其串行处理机制导致推理延迟过高，仍无法满足边缘智能对实时性的严苛要求。

因此，针对上述显存效率与推理速度难以兼顾的挑战，本文提出了一套高效的边缘大模型分层部署机制：通过构建基于权重矩阵粒度的流水线调度模型，将加载与计算拆解为细粒度并行子任务，重叠显存访问延迟；设计了基于内存预分配的显存管理方案，利用静态预估与指针偏移复用，消除频繁垃圾回收（GC）开销；引入了内存映射（mmap）实现权重按需加载，解决了模型从外存到内存的输入/输出（I/O）瓶颈。所提方案在不增加显存占用的前提下大幅提升了推理吞吐量，有效解决了边缘大模型推理延迟过高导致的各类问题。

1 现有边缘大模型推理框架

面对边缘设备资源受限、算力与显存紧张的现实条件，大模型推理效率与部署可行性面临严峻考验。围绕这一问题，目前业界已提出多条典型技术路线与代表性推理框架，通过差异化方案攻克边缘部署痛点。

Hugging Face Transformers 是行业标准推理框架，其核心机制是依赖显存全量加载模型权重，采用静态显存管理策略，虽具备良好兼容性，但在边缘场景面临严峻挑战。比如，在长序列推理中，KV Cache 随上下文长度线性增长，极易触发内存溢出。显存容量成为部署超大模型的硬性瓶颈，限制了其在资源受限设备上的应用。vLLM 针对高并发场景引入 PagedAttention 机制，通过分页管理优化显存碎片化，显著提升吞吐量。然而，vLLM 高度依赖数据中心级算力与统一计算设备架构（CUDA）优化，复杂架构使得其在边缘侧难以高效运行，无法突破物理显存的绝对限制，部署灵活性差。

llama.cpp 另辟蹊径，采用生成式预训练变换器（GPT）

生成统一格式（GGUF），将模型权重压缩至 4 bit 甚至更低精度，并结合 C/C++ 底层优化提升运算效率。这一策略大幅降低显存占用，使边缘设备得以承载更大模型。但量化带来的精度损失在逻辑推理、复杂语义理解等任务中尤为明显，模型输出质量与原生版本存在显著差异，因此需权衡资源与性能需求。AirLLM 则通过中央处理器-图形处理器（CPU-GPU）分层卸载技术实现“以时间换空间”，动态加载模型层以突破显存限制，使边缘设备可运行超大规模模型。然而，该策略的代价是极高的输入/输出（I/O）开销：频繁的主机-设备数据传输导致加载与显存清理耗时远超实际计算时间，推理延迟显著增加，实时性难以保障。

可以看出，现有主流方案各存在短板：静态加载框架（如 Transformers）受限于显存容量壁垒，难以部署大模型；量化技术（如 llama.cpp）以精度换资源，应用场景受限；动态卸载（如 AirLLM）虽突破显存限制，却因权重加载 I/O 阻塞导致延迟激增。究其根源，现有框架缺乏对推理全生命周期内存行为的动态优化，未能实现计算与内存调度的协同适配。边缘场景下的高效推理，亟需突破传统静态或粗粒度优化范式，构建能够实时感知资源状态、动态调度计算与内存访问的推理框架，在有限资源下实现性能与效率的平衡。这一方向成为当前边缘大模型部署的关键研究命题。

2 高效分层部署机制设计

针对现有分层卸载存在的显存加载延迟、显存清理开销、模型加载 I/O 瓶颈三大性能痛点，本文设计了一套包含映射加载、内存预分配与流水线调度的综合优化框架，旨在通过精细化的资源管理，实现推理过程的“零等待”与“低开销”。

2.1 整体架构设计

本文所述的 TriadFlow 架构由基于 mmap 的按需加载层、基于内存预分配的显存管理层以及基于权重矩阵粒度的流水线调度层构成，如图 1 所示。这 3 层在逻辑上呈现出“数据供给-空间保障-时间重叠”的递进与协同关系，共同构成了面向边缘场景的高效推理引擎。

在边缘设备有限的内存资源下，传统全量加载模式极易导致内存溢出。本文引入 mmap 技术，将庞大的模型权重文件直接映射至进程的虚拟地址空间。这一设计不仅消除了传统 I/O 读写过程中的数据拷贝开销，更实现了权重数据的“按需加载”。它为上层机制提供了逻辑上的“全量数据视图”，解决了超大模型在边缘侧“存不下”的根本矛盾，为后续调度提供了数据源头的供给保障。

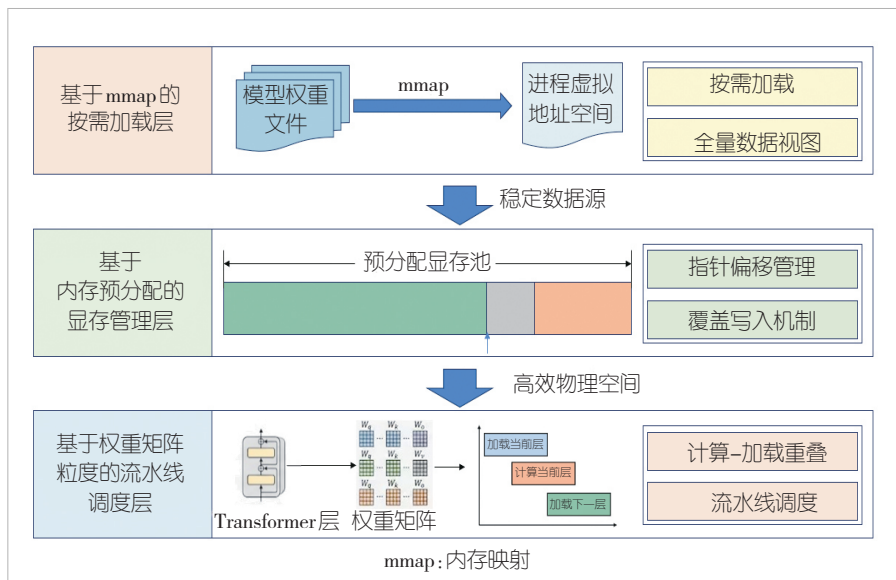


图1 TriadFlow整体架构设计

mmap虽然解决了数据的逻辑存在，但频繁的分层卸载仍需在物理显存中进行高频的数据交换。若采用传统的动态内存分配，频繁的申请、释放与GC将引入巨大的系统抖动。为此，本文设计了静态内存预分配方案。推理引擎在启动阶段即可根据计算图静态预估峰值显存需求，并预先向系统申请一块连续的内存构建“显存池”。通过指针偏移与覆盖写入机制，该方案彻底消除了运行时的动态内存管理开销，为流水线调度提供了稳定、无碎片的物理运行环境。

在具备了稳定的数据源（mmap）和高效的物理空间（预分配显存池）后，调度层致力于解决“计算-访存”失配问题。传统分层卸载以“层”为单位进行串行处理，导致计算单元在等待权重加载时处于空闲状态。本文将调度粒度从“层”细化至“权重矩阵”，将Transformer层拆解为独立的线性运算单元。调度器利用预分配显存池中的指针管理，精确控制数据搬运与计算任务的时序，使得当前层的计算与下一层的权重加载在时间上完全重叠。

综上所述，这3层架构紧密耦合：mmap加载机制为显存管理层提供了按需供给的数据流；显存管理层为流水线调度提供了无锁、低延迟的空间复用环境；而流水线调度层则充分利用前两者的优势，将显存访问延迟完全重叠在计算过

程之中。三者协同，最终可在不增加显存占用的前提下，大幅提升边缘大模型的推理吞吐量。

2.2 基于mmap的模型加载加速

在边缘设备有限的内存资源约束下，传统全量加载模式不仅面临“存不下”的物理限制，更因冗余的数据搬运导致严重的I/O阻塞^[12]。因此，本研究引入mmap机制，旨在通过底层系统调用的优化，构建一个逻辑上全量、物理上按需的权重加载层，为后续推理过程提供高效、稳定的数据源供给。

大模型权重文件通常高达数GB甚至数十GB，且多以二进制格式存储。传统的I/O读取方式采用“磁盘-内存”完整拷贝模式，需将完整的权重数据从磁

盘分区读取至内存缓冲区，再搬运至物理内存的指定地址。这一过程不仅受限于磁盘I/O带宽，还极易因大量数据的集中拷贝导致内存碎片化。实验数据表明，7B模型的内存加载耗时为1.32 s，而72B模型则激增至112.28 s，直观反映出I/O带宽与模型参数量级之间的尖锐矛盾。此外，碎片化的内存空间无法被有效利用，易引发系统颠簸，导致加载过程频繁触发页交换，造成CPU与内存资源的有效消耗，严重影响边缘设备长时间运行的稳定性。

针对上述加载缓慢与内存碎片化问题，本文引入mmap机制，结合边缘设备的存储与内存特性对模型加载流程进行重构，并充分利用该机制的“文件-虚拟内存映射”特性，将大模型权重文件直接映射到推理进程的虚拟地址空间，具体实现过程如图2所示。首先，通过系统调用mmap()函数，将权重文件的磁盘地址与进程虚拟地址空间中一段连续地址区间建立映射关系，该映射关系被记录在进程的页表中。其次，设置映射权限为“只读”，确保权重数据不被意外修改，同时避免写入操作带来的额外开销。最后，将映射后的虚拟地址区间与推理引擎的权重访问接口关联，使引擎可直接通过虚拟地址访问权重数据，无需经过传统I/O的数据拷

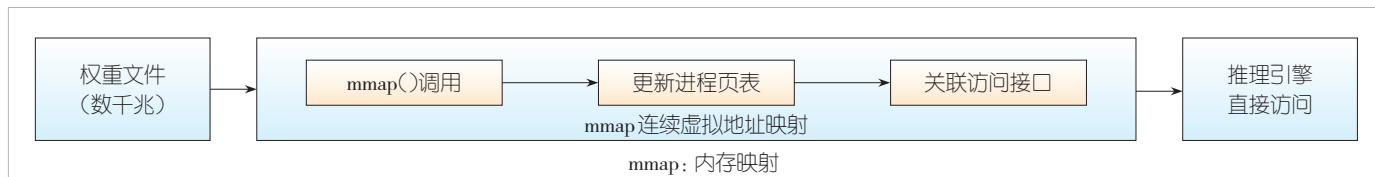


图2 mmap加载机制架构

贝流程。

这一机制使模型推理引擎能够像访问内存变量一样直接读取磁盘上的权重数据，真正实现权重数据的“按需加载”，其加载逻辑与推理流程深度绑定：当推理引擎执行到某一层计算且需要访问该层对应的权重数据时，系统会通过虚拟地址发起访问请求。此时，操作系统会检查该虚拟地址对应的物理内存页面是否已加载。若未加载，则触发页面置换机制，将磁盘上对应的权重数据块（通常为 4 kB 或 8 kB，与系统页大小一致）换入物理内存，同时更新页表。若已加载，则直接从物理内存中读取数据，无须访问磁盘。未被访问的权重数据仍保留在磁盘上，无须占用物理内存资源，有效降低了边缘设备的内存压力。这种加载方式不仅能够大幅提升超大权重文件的加载效率，还可通过“按需加载”减少内存数据的冗余存储，有效规避内存碎片化对推理性能的影响。

基于 mmap 的按需加载方案不仅解决了模型权重在边缘侧的“存与取”问题，更为上层机制提供了逻辑上的全量数据视图。然而，仅有高效的数据供给尚不足以支撑高性能推理。若缺乏物理显存层面的精细化管理，频繁的内存分配与回收仍将引发系统级抖动。因此，下文将基于这种数据供给基础，进一步探讨如何通过静态内存预分配策略构建“无锁、低延迟”的物理运行环境。

2.3 基于内存预分配的显存管理优化

基于上述关于数据供给机制的论述，本节聚焦“空间保障”层面，核心目标是消除运行时显存管理的系统开销。尽管 mmap 解决了权重数据的逻辑存在，但在推理过程中，激活值与中间结果的高频动态变化，往往迫使系统进行频繁的显存申请与释放。针对这一问题，本文设计了轻量化的显存池管理器，摒弃传统的动态分配策略，转而采用静态预分配与指针偏移复用的机制，旨在从根源上切断由 GC 引发的性能抖动，为流水线调度提供稳固的物理空间支撑。

为避免显存溢出，现有分层加载框架需频繁执行显存释放与重新分配操作。这一过程涉及操作系统层面的锁竞争与碎片整理，引入了巨大的时间成本。例如，在 14B 模型推理中，显存清理耗时（12.84 s）已显著高于计算耗时（0.67 s），

成为首要瓶颈。而 72B 模型的清理时间更是达到了 60.48 s，约占用总耗时（302.44 s）的 20%。频繁的显存管理不仅消耗时间，还加剧了系统资源竞争，降低了推理稳定性。

针对显存清理环节的频繁开销问题，本研究提出一种内存预分配优化方案，通过静态管理模式替代传统动态分配策略，从根源上减少显存回收与重新分配带来的系统损耗，如图 3 所示。

现有推理框架因无法提前预判推理过程中的内存需求波动，普遍采用保守的动态分配策略，即根据实时计算需求动态申请和释放显存空间，这一模式导致推理过程中充斥着大量显存回收、重新分配的系统调用，不仅增加推理延迟，还易引发显存碎片，尤其在边缘设备有限的显存资源下，碎片积累会进一步降低显存利用率。该优化方案的核心思路围绕“空间换时间”与“静态管理”展开。本文设计了一款轻量化显存池管理器，该管理器包含内存需求预估、显存区域划分、动态指针调度 3 个核心模块。在推理任务启动前，管理器先通过遍历模型计算图，统计各层权重、激活值、中间特征向量的最大占用空间，结合推理批次大小及边缘设备显存资源上限，精准计算出模型推理过程中的峰值内存消耗。系统据此提前预分配一块连续的、固定大小的显存区域，其容量预留 5%~10% 的冗余空间，确保无需额外动态扩容，同时避免显存资源浪费。预分配的显存区域按功能类型可以划分为权重存储区、激活值存储区、中间结果存储区。各区域边界固定，通过指针标记实现独立管理，以避免不同类型数据的相互干扰。

在推理过程中，本方案采用指针动态管理机制替代传统内存回收机制，具体如图 4 所示。系统为预分配的各显存区域分别设置读写指针与空闲指针，通过移动指针标记显存空间的占用与释放状态，无须对显存块进行物理上的回收与重新分配。权重存储区采用“一次性加载、全程复用”的模式。权重数据加载后固定存储在对应区域，直至推理任务结束。激活值与中间结果存储区采用“覆盖写入”模式。当某一层计算完成后，后续层的激活值的计算结果直接覆盖前一层的闲置空间，通过指针偏移实现空间复用，彻底避免了频繁触发操作系统级的 GC 操作，从根本上规避频繁 GC 带来的性能损耗。

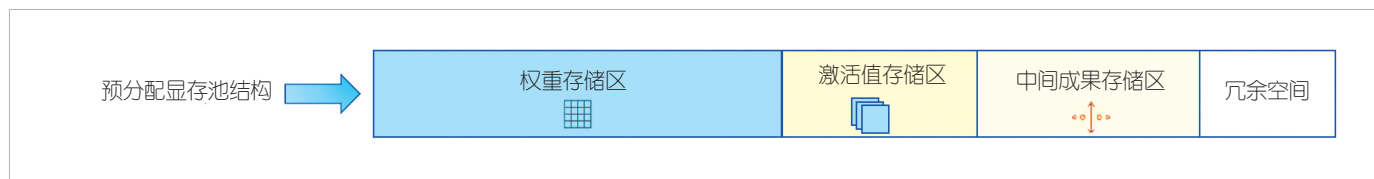


图3 预分配显存机制

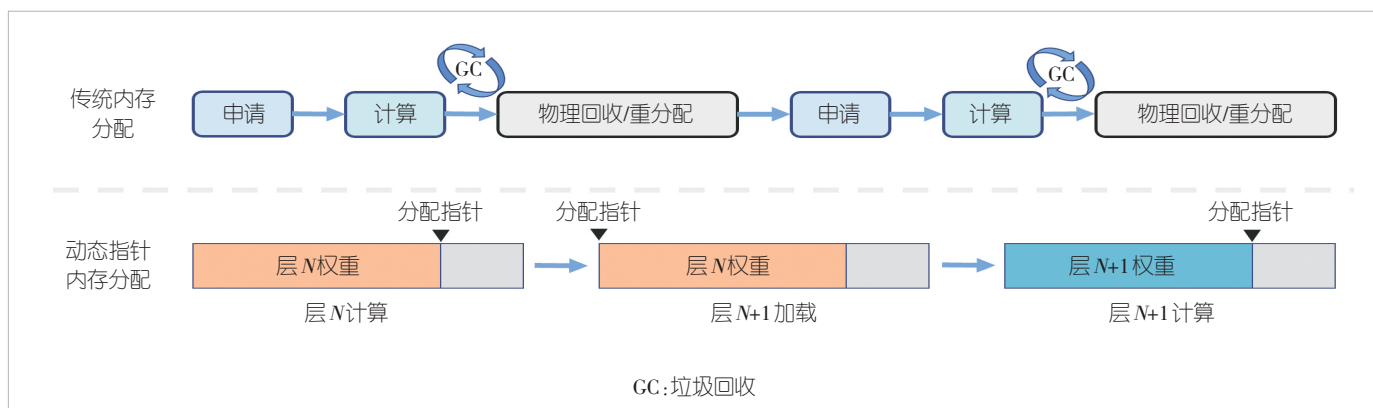


图4 动态指针内存分配机制

然而，仅有静态的空间保障仍无法解决“计算-访存”失配导致的时间效率问题。当显存访问延迟无法被重叠时，计算单元的空闲将直接导致吞吐量下降。因此，下文将围绕“时间重叠”，深入探讨如何通过细粒度的流水线调度，将显存访问延迟完全掩藏于计算过程之中。

2.4 基于权重粒度的流水线调度机制

综合前文所述的数据供给与空间保障机制，本研究构建了边缘推理引擎的基础架构。但要实现极致的推理效率，关键在于“时间重叠”维度的调度优化。在边缘设备有限的算力条件下，传统的粗粒度（层级别）调度往往导致计算单元在等待权重加载时处于空闲状态，形成显著的“气泡”延迟。针对该问题，本文提出了一种基于权重矩阵粒度的流水线调度机制。该机制深度结合Transformer模型，将计算与加载任务拆解为细粒度的并行子任务，旨在通过精确的时序控制，实现加载与计算的完全重叠，从而在不增加显存占用的前提下使硬件利用率达到最大。

现有分层加载框架以Transformer层为单位进行显存交换，导致计算单元在等待下一层权重加载时处于空闲状态，形成了显著的“计算-等待”间隙。虽然AirLLM采用了预取策略进行重叠计算和加载操作，但实验数据显示，加载操作耗时远长于计算操作，简单的层级预取没有办法实现时间覆盖，“计算-等待”间隙需要被优化。随着模型参数量级的增加，显存加载延迟呈非线性增长：7B模型的显存加载耗时为6.83 s，而72B模型则高达123.59 s——时长增加了近20倍，严重制约大模型边缘实时推理能力。

为解决显存加载引发的计算单元空闲问题，本工作摒弃传统“层”级加载粒度，提出基于“权重矩阵”粒度的流水线调度机制。该机制充分适配Transformer模型的内部计算特性，实现加载与计算任务的并行协同。

本工作深入剖析Transformer层^[13]的内部计算结构，结合其“预归一化”或“后归一化”的典型架构，将单一层级完整拆分为多个独立且可并行调度的矩阵运算单元，具体拆分逻辑贴合Transformer的正向计算流程：首先，将自注意力机制模块拆分为查询（Q）、键（K）、值（V）3个独立的权重矩阵运算单元，三者均基于输入嵌入向量与对应权重矩阵完成线性变换，且运算过程相互独立，无依赖关系；其次，将自注意力输出后的线性投影层拆分为独立的权重矩阵运算单元，负责对注意力加权后的特征向量进行维度映射与整合；最后，将前馈网络（FeedForward）层拆分为两个串联的线性变换权重矩阵运算单元，分别对应“升维-降维”的计算过程，同时将层归一化模块的参数提取为独立的参数单元，确保其与矩阵运算单元的调度协同。在此基础上，本文明确各拆分单元的计算时序与显存需求，例如Q、K、V权重矩阵的运算可同步启动，其显存加载需求均为固定维度的浮点数矩阵。而FeedForward层的两个线性变换单元需按“先升维后降维”的顺序依次运算，显存加载需遵循对应时序。优化后的调度器打破“整层加载完成后再启动计算”的传统逻辑，将显存加载任务拆解为与计算任务同步推进的子任务流，形成“计算-加载”并行的流水线模式。在计算单元处理当前批次权重矩阵的运算任务时，数据搬运单元已根据预设时序，提前启动下一阶段所需权重矩阵的显存加载操作，确保计算单元完成当前任务后可立即调用新的权重数据，无需进行等待。

基于这种细粒度流水线设计，本文成功实现了显存加载时间与计算时间的最大程度重叠，有效重叠显存访问延迟，基本消除粗粒度加载模式下普遍存在的计算“气泡”等待时间。具体而言，Q、K、V权重矩阵的加载与运算可实现完全并行，如图5所示。当计算单元处理Q矩阵运算时，数据搬运单元同步加载K、V矩阵；待Q矩阵运算完成后，K、V

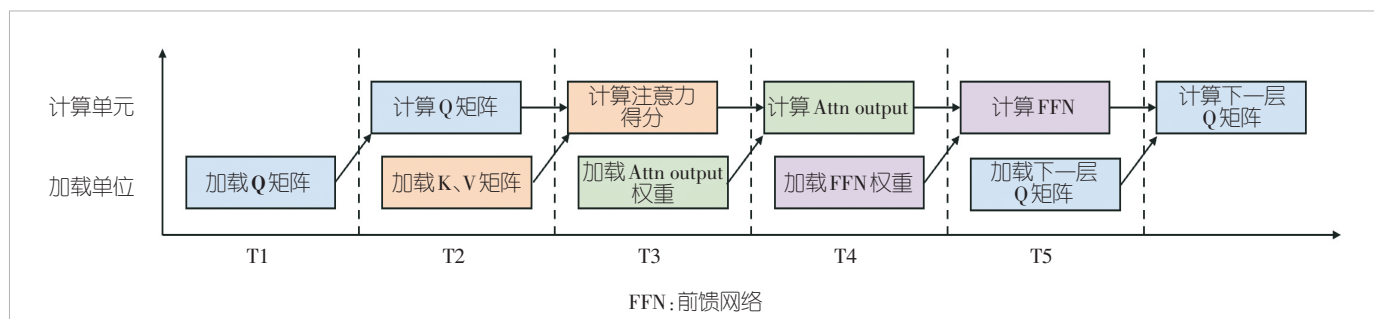


图5 流水线加载重叠显存加载时间与计算时间

矩阵加载完毕，可立即启动注意力得分计算。自注意力输出的线性投影层权重加载，可与注意力得分计算过程并行，进一步压缩无效等待时间。

该机制通过将调度粒度从“层”细化至“权重矩阵”，成功实现了显存加载时间与计算时间的最大重叠，大大减小了计算空泡。综上所述，本章提出的“数据供给—空间保障—时间重叠”三位一体优化框架：依托mmap破除I/O瓶颈，借助预分配消除了GC抖动，采用细粒度流水线重叠访存延迟。这3层机制紧密耦合、互为支撑，共同构成了面向边缘场景的高效大模型分层部署框架。

3 实验验证

3.1 实验设置

为验证本文提出的高效分层部署机制在边缘多智能体场景下的有效性，实验选取了典型的边缘计算设备作为测试平台，并对比了优化前后的推理性能，具体如表1所示。

3.2 性能对比分析

如图6所示，通过引入基于权重粒度的流水线调度与内存预分配优化，本工作提出的TriadFlow成功在单Token推理耗时方面较基准组原版AirLLM的平均耗时在不同尺寸的模型上都有巨大降低，平均降低达到53%。在较小模型上，降低效果更佳，7B模型降低达64%。通过消除显存加载等待

与内存清理开销，推理吞吐量提升了1倍以上。

此外，在显存占用方面，如图7所示，本工作提出的TriadFlow保持了与基准组AirLLM相当的低显存占用水平，均维持在4 GB以下。这说明本文提出的优化策略并未引入额外的显存负担，依然适配边缘设备的资源限制。

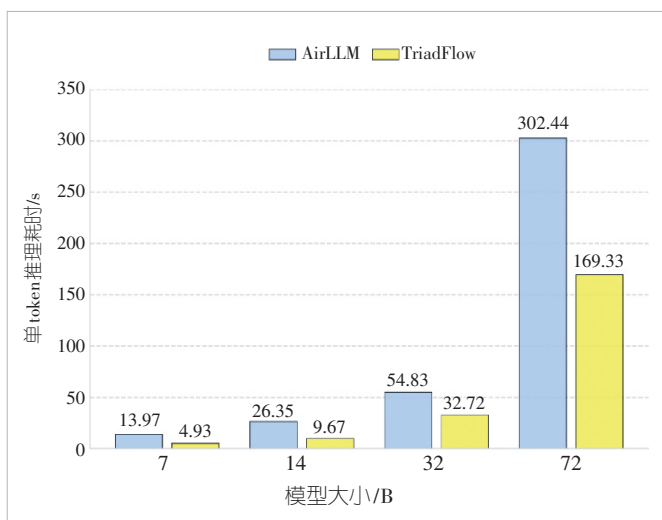


图6 单Token推理平均耗时对比

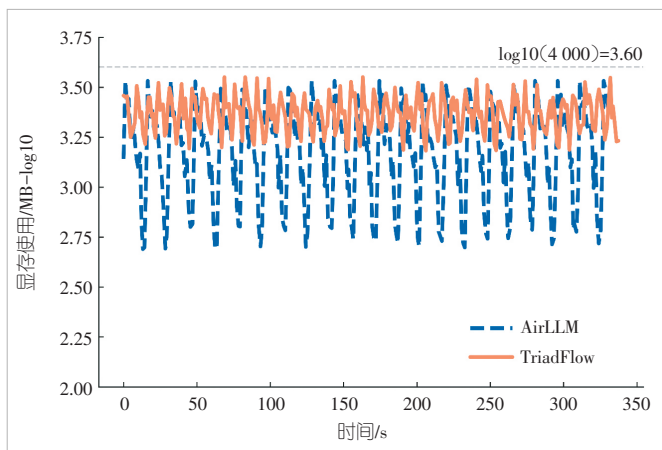


图7 显存占用对比

表1 实验条件设置

硬件环境	模型配置	对比对象
采用NVIDIA L40 (48 GB RAM),使用 torch.cuda.set_per_process_memory_fraction限制显存6 GB,模拟家庭边缘算力环境	选用Qwen2.5-7B、Qwen2.5-14B、Qwen2.5-32B和Qwen2.5-72B ^[14-15] 模型作为基准模型	基准组为原始AirLLM框架,实验组为本文提出的TriadFlow

RAM: 随机存取存储器

4 结束语

针对边缘大模型推理延迟过高的问题,本文提出了一种高效分层部署优化机制。通过剖析现有框架的耗时构成,明确显存加载、内存管理与模型 I/O 是导致延迟的主要原因,并设计了3项核心优化策略:基于权重粒度的流水线调度机制有效重叠了显存访问延迟,内存预分配策略消除了频繁的显存回收开销, mmap 内存映射机制则解决了大模型加载的 I/O 瓶颈问题。实验结果表明,所提机制在不增加显存占用的前提下,将推理延迟降低 50% 以上。未来的工作将探索该机制在异构边缘设备集群中的分布式扩展,并研究如何结合量化技术进一步提升推理效率。

参考文献

- [1] Li Z H, Li T, Feng W J, et al. PRIMA. cpp: fast 30–70B LLM inference on heterogeneous and low-resource home clusters [PP/OL]. arXiv (2025–04–07) [2026–04–16]. <https://arxiv.org/abs/2504.08791>
- [2] Sarhaddi F, Nguyen N T, Zuniga A, et al. LLMs and IoT: a comprehensive survey on large language models and the Internet of Things [J]. IEEE open journal of the communications society, 2026, 7: 3585–3616. DOI: 10.1109/ojcoms.2026.3680064
- [3] Lin Z, Qu G Q, Chen Q Y, et al. Pushing large language models to the 6G edge: vision, challenges, and opportunities [J]. IEEE communications magazine, 2025, 63(9): 52–59. DOI: 10.1109/MCOM.001.2400764
- [4] Huang X Y, Shen L M, Ma Z J, et al. Toward privacy-preserving and personalized smart homes via tailored small language models [J]. IEEE transactions on mobile computing, 2026, 25(7): 9505–9520. DOI: 10.1109/tmc.2026.3654976
- [5] Shi W S, Cao J, Zhang Q, et al. Edge computing: vision and challenges [J]. IEEE Internet of Things journal, 2016, 3(5): 637–646. DOI: 10.1109/JIOT.2016.2579198
- [6] 王睿, 齐建鹏, 陈亮, 等. 面向边缘智能的协同推理综述 [J]. 计算机研究与发展, 2023, 60(2): 398–414
- [7] 张晓东, 张朝昆, 赵继军. 边缘智能研究进展 [J]. 计算机研究与发展, 2023, 60(12): 2749–2769. DOI: 10.7544/issn1000–1239.202220192
- [8] Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing [EB/OL]. [2026–04–14]. <https://github.com/huggingface/transformers>
- [9] Zhu W, Li S, Zheng S, et al. vLLM: high-throughput LLM serving with PagedAttention [C]//Proceedings of the 29th Symposium on Operating Systems Principles. ACM, 2023: 1–16
- [10] Gerganov G. LLaMA model inference in pure C/C++ [CP/OL]. [2026–04–15]. <https://github.com/ggerganov/llama.cpp>
- [11] Lyogvin. AirLLM: efficient inference framework for large language models on low-resource GPUs: Version 2.11.0 [CP/OL]. [2026–04–15]. https://gitcode.com/GitHub_Trending/ai/airllm
- [12] 高赫然, 吴恒, 许源佳, 等. 面向深度学习训练的内存交换机制综述 [J]. 软件学报, 2023, 34(12): 5862–5886
- [13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. ACM, 2017: 5998–6008. DOI: 10.48550/arXiv.1706.03762
- [14] Qwen. Qwen2.5: a party of foundation models [EB/OL]. (2024–09–19)[2026–04–15]. <https://qwenlm.github.io/blog/qwen2.5/>

- [15] Yang A, Yang B S, Hui B Y, et al. Qwen2 technical report [PP/OL]. arXiv(2024–07–15) [2026–04–15]. <https://arxiv.org/abs/2407.10671>

作者简介



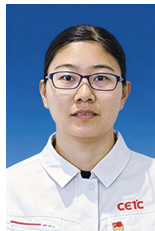
万卓然, 北京邮电大学在读硕士研究生; 主要研究方向为边缘大模型部署与推理任务调度。



谢人超, 北京邮电大学教授、博士生导师; 主要研究方向为算力网络、网络人工智能等。



王媛, 中国电子科技集团公司第五十研究所工程师; 主要研究方向为算力网络、软件定义网络等。



王芳, 中国电子科技集团公司第五十研究所高级工程师; 主要研究方向为软件定义网络、网络人工智能等。



唐琴琴, 北京邮电大学副研究员、硕士生导师; 主要研究方向为算力网络、网络人工智能、网络孪生等。



黄韬, 北京邮电大学教授、博士生导师; 主要研究方向为未来网络、确定性网络等; 先后主持科技部、工信部等重大项目10余项; 获中国通信学会科技奖、中国电子学会科技奖等20余次; 发表论文120余篇, 获授权专利100余项。

Token 原生的 Scale Across 网络



Token Native Scale Across Network

苏远超/Su Yuanchao

(阿里云计算有限公司, 中国 杭州 310030)
(Alibaba Cloud, Hangzhou 310030, China)

DOI: 10.12142/ZTETJ.202603007

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260626.1349.010>

网络出版日期: 2026-06-26

收稿日期: 2026-04-15

摘要: 传统广域网架构的设计初衷是承载普通流量, 并未考虑人工智能 (AI) 工作负载的特殊性。基于第一性原理, 提出一种新的 Token 原生 Scale Across 网络, 支持各类 AI 负载在广域网中的传送。该网络采用模型与基础设施协同设计、每任务流量工程以及统一的互联网协议第 6 版 (IPv6) 寻址方案。相比传统网络, 其容量扩展能力提升 100 倍, 收敛时间缩短至原来的 1/10, 成本降低 50%。

关键词: 模型与基础设施协同设计; 每任务流量工程; 统一的 IPv6 寻址方案

Abstract: Traditional wide-area network (WAN) architectures are originally designed to carry regular traffic, and the specificities of artificial intelligence (AI) workloads are not taken into account. Based on first principles, a new token native Scale Across network is proposed to support the transmission of various AI workloads over the WAN. The network is built upon model-infrastructure co-design, per-task traffic engineering, and a unified Internet Protocol version 6 (IPv6) addressing scheme. Compared with traditional networks, its capacity scalability is improved by 100 times, convergence time is reduced to one-tenth, and cost is reduced by 50%.

Keywords: model-infrastructure co-design; per-job traffic engineering; unified IPv6 addressing scheme

引用格式: 苏远超. Token 原生的 Scale Across 网络 [J]. 中兴通讯技术, 2026, 32(3): 37-39. DOI: 10.12142/ZTETJ.202603007

Citation: Su Y C. Token native scale across network [J]. ZTE technology journal, 2026, 32(3): 37-39. DOI: 10.12142/ZTETJ.202603007

1 Scale Across 全新架构的必要性

随着人工智能 (AI) 技术的飞速发展, 特别是大语言模型和多模态模型的兴起, AI 训练/推理对网络基础设施提出了前所未有的挑战。Agent 技术的爆炸式发展, 不仅催生了 Token 经济体系, 也正推动计算底层架构向全新的范式演进。然而, 传统广域网架构以承载普通流量为设计初衷, 既未考虑 AI 工作负载的特殊性, 如海量参数同步、键值缓存 (KV Cache) 交换, 以及对超低延迟和超高带宽的严格要求, 也未能针对 Token 的传输特征进行优化^[1-3]。

从第一性原理出发, 我们需要回答一个核心问题: “如果 Token 是第一等的用户实体, 那么怎样的 AI 基础设施才是最合适的?” 若聚焦广域网范畴, 问题则转化为: 如何构建最适合 Token 传输的 Scale Across 网络。

本文提出 Token 原生的 Scale Across 架构设计, 旨在构建面向未来 3~5 年的下一代广域网络。该架构并非对现有网络的局部修补, 而是从底层设计到上层应用的系统性重构。

该架构的底层逻辑是面向生产 Token、传输 Token、消费 Token 的全过程, 通过模型与基础设施的协同设计, 构建一张具备 Token 粒度监控与调度能力的超大规模广域专用网络。

Token 原生的 Scale Across 架构设定了 3 项关键技术指

标^[4], 并通过以下技术路径实现:

- 1) 容量可扩展能力: 提升 100 倍——通过路由芯片迭代与 Fabric-router 技术的协同演进, 支持系统容量实现百倍级扩展。
- 2) 故障收敛时间: 缩短至原来的 1/10——综合运用波道保护、链路检测 (LD) 与快速检测 (RD) 等技术, 实现百毫秒级 (100 ms) 故障恢复。
- 3) 单位成本: 降低 50%——通过全面自研芯片与智能流量调度技术的深度融合, 单位带宽成本降低 50%。

2 Token 原生的 Scale Across 网络核心应用场景

Token 原生的 Scale Across 架构主要服务于四大关键场景^[5], 这些场景背后的业务需求与技术挑战, 共同构成了架构设计的直接驱动力。

- 1) 计算到计算互联。该场景涵盖预填充-解码 (PD) 分离架构、强化学习等应用, 覆盖同城与跨城互联需求, 是实现分布式算力协同的基础。
- 2) 计算到存储互联。该场景主要面向 KV Cache 的存储访问优化, 旨在实现同城高速互联。随着模型规模持续扩大, 训练数据集的读取与模型检查点的保存均对存储访问带

宽提出了极高要求。

3) 前端与后端跨可用区互联。该场景旨在打通不同可用区 (AZ) 之间的前端与后端网络, 支撑业务的地理分布式部署, 提升系统容灾与弹性调度能力。

4) 跨可用区训练。该架构以前瞻性设计支撑未来大规模分布式AI训练, 可确保未来3~5年的技术领先性。

上述4个场景从当前迫切需求到未来演进方向层层递进, 形成了完整的应用场景图谱, 同时也为后续技术选型提供了清晰的方向指引。

3 Token原生的Scale Across架构设计

Token原生的Scale Across在架构上继承了阿里巴巴eCore网络的多空间、多平面、多域设计及单栈单片技术体系, 并在此基础上进行了针对性优化^[6]:

1) 采用专用物理网络, 彻底隔离AI业务流量与其他业务流量, 从根本上消除复杂的服务质量 (QoS) 策略配置与相互干扰。

2) 全部采用大带宽端口互联 (端口速率从400 Gbit/s起步) 及大容量单机设备 (交换容量从51.2 Tbit/s起步), 最大程度降低由AI“大象流”引发的哈希冲突, 从而有效减少拥塞丢包。

3) 网络拓扑完全扁平化, 减少转发跳数, 显著提升数据平面转发效率。

4) 路由架构完全扁平化, 端到端统一采用互联网协议

第6版 (IPv6) 地址空间, 减少协议封装与地址转换, 提升流量调度效率。

5) 在运维层面, 支持完善的监控体系, 包括框架层监控、任务级诊断与异常检测、多主机关联分析、网络为中心的定位能力, 以及覆盖计算 (图形处理器 (GPU)、内存)、通信 (远程直接内存访问 (RDMA)、NVLink、外围组件互连高速 (PCIe)) 和网络交换的全栈监控。这一体系确保了Per-job的可观测性和可运维性。Token原生的Scale Across网络拓扑如图1所示。

3.1 模型与基础设施协同设计

模型与基础设施的协同设计, 标志着网络架构正从“被动适配业务”向“主动协同业务”转变。上层模型的结构、并行策略、调度方式及内容传送机制等, 均会对Scale Across网络的流量模型产生深远影响。

因此, 当Scale Across网络出现性能瓶颈并需要调优时, 若仅在网络层面进行优化, 往往既耗时又收效甚微。阿里巴巴的工程实践表明, 需要从两个维度开展协同设计: 一是模型与基础设施的协同设计, 二是端侧 (网卡) 与网络侧的协同设计。

一个典型场景是采用Scale Across网络承载预填充-解码 (PD) 分离架构下的远程直接内存访问 (RDMA) 流量。若不经任何修改, 直接将原本部署在数据中心内的PD分离RDMA流量搬移至广域网传输, 则很可能遭遇首字延迟 (TTFT) 显著恶化的问题, 从而导致业务不可用。该问题的

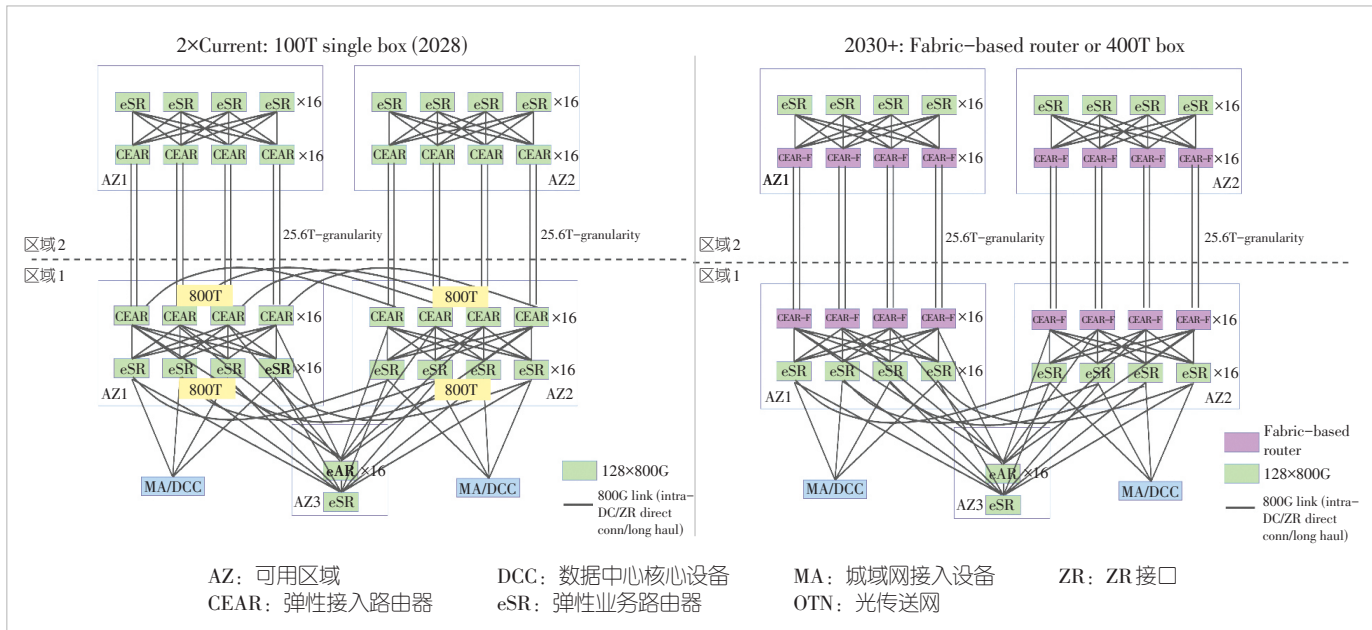


图1 Token原生的Scale Across网络拓扑

根源在于, GPU 直接内存访问 (GDR) 要求发送端与接收端的显存逻辑地址连续, 而 PD 分离场景中张量并行 (TP) 规模往往不一致, 这会触发大量 GDR 小包。在数据中心内, 此类小包仅带来微秒级时延, 影响有限; 但在广域网中, 时延跃升至毫秒级, 将导致网卡侧未完成工作队列元素 (WQE) 快速堆积, 进而造成吞吐率下降, 具体体现为 TTFT 显著增长。通过调整端侧网卡的 WQE 参数, 并结合后文阐述的每任务流量工程机制, 可在基础设施层面优化小包传送性能。然而, 该方案仅能解决约 50% 的此类问题^[7]。

在模型层面, 需解决因 TP 规模不一致而产生大量 GDR 小包的问题。具体做法是在发送端与接收端均执行内存重排操作, 使 RDMA 操作基于连续大块内存进行读写, 以提升 GDR 传输效率。

模型与基础设施的协同设计, 需以任务完成时间 (JCT)、TTFT 及每个输出 Token 的时间 (TPOT) 为共同优化目标, 并在组织层面建立高效的协同机制。

3.2 每任务流量工程 (Per-job TE)

上述端网协同设计是架构最具创新性的核心体现。传统网络通常采用基于五元组的流量识别方法, 但对于远程 RDMA 流量而言, 这种识别远不够精细。本架构通过解析 RDMA 的队列对 (QP) 字段, 并结合目的 IP 地址, 能够精确识别每一条 RDMA 流。新一代网卡支持单 QP 多路径能力, 使每任务流量工程 (Per-job TE) 与网卡多路径机制可正交协同工作, 实现更为灵活的流量调度。

基于上述设计, 即可建立从 Token 到 Job, 再经框架层、通信库、QP, 直至 Scale Across 网络的完整端到端映射链路。这一映射机制构成了 Token Native 架构最为核心的技术基石。

3.3 统一的 IPv6 寻址方案

为支持上述精细化调度, 架构采用统一的 IPv6 地址规划, 覆盖从容器、网卡到高性能网络及 Scale-Across 的完整数据路径。该设计的核心优势在于免去 IPv6-in-IPv6 封装, 从而减少报文处理开销。同时, 基于重新定义的 SRv6 地址格式, 网络可根据业务需求灵活选择默认路径、低时延路径或低成本路径, 以适配不同业务的差异化需求。

3.4 关键技术特性

Per-job 流量工程: 基于 (目的 QP 号、目的 IP) 二元组精确匹配 RDMA 流, 实现更为精细的流量调度粒度。

链路检测 (LD) /接收检测 (RD): 依据 G.709 Annex C 标准, 在网络降级条件下实现无损且快速的 ECMP 故障切换, 确保业务连续性。其中, LD 与 RD 作为光传输系统中的关键

监测机制, 为快速感知链路状态变化提供支撑。

快速拥塞通知包 (CNP): 加速拥塞反馈信号的传递, 使发送端能够更快感知网络状态变化, 并及时调整发送速率, 从而有效缓解拥塞扩散。

Packet Trimming: 当报文因拥塞被丢弃时, 仅保留 L3 与 L4 头部并发送至源端, 支持单包重传。该设计避免了传统模式下需由 CPU 介入处理完整报文的问题, 提升了重传效率。

4 结束语

Token 原生的 Scale Across 架构标志着 Token 时代广域网面向 AI 场景的一次重大演进。这一设计既是对第一性原理的遵循, 也是 Token 内在传输需求的最自然体现。该架构将为未来大规模 AI 训练与推理提供坚实的网络底座, 有力支撑业务的持续创新与增长。

参考文献

- [1] Filsfils C, Camarillo P, Leddy J, et al. Segment routing over IPv6 (SRv6) network programming RFC 8986 [EB/OL]. (2026-05-20) [2026-05-25]. <https://datatracker.ietf.org/doc/rfc8986/>
- [2] Filsfils C, Dukes D, Previdi S, et al. RFC 8754: IPv6 segment routing header (SRH) [EB/OL]. (2020-03) [2026-04-25]. <https://www.rfceditor.org/info/rfc8754/>. <https://doi.org/10.17487/RFC8754>
- [3] Filsfils C, Ketan T, Daniel V, et al. Segment routing policy architecture RFC 9256 [EB/OL]. [(2022-08) [2026-04-20]. <https://datatracker.ietf.org/doc/rfc9256/>. DOI: 10.17487/RFC9256
- [4] Singh R, Agarwal S, Calder M, et al. Cost-effective cloud edge traffic engineering with cascara [EB/OL]. [2026-04-15]. <https://www.usenix.org/conference/nsdi21/presentation/singh>
- [5] Xu Z Y, Yan F Y, Singh R, et al. Teal: learning-accelerated optimization of WAN traffic engineering [C]//Proceedings of the ACM SIGCOMM 2023 Conference. ACM, 2023: 378-393. DOI: 10.1145/3603269.3604857
- [6] Singh A, Ong J, Agarwal A, et al. Jupiter rising: a decade of clos topologies and centralized control in Google's datacenter network [EB/OL]. (2015-08-17) [2026-04-20]. <https://dl.acm.org/doi/10.1145/2829988.2787508>
- [7] Zhong Z Z, Ghobadi M, Khaddaj A, et al. ARROW: restoration-aware traffic engineering [EB/OL]. (2021-08-09) [2026-04-18]. <https://dl.acm.org/doi/abs/10.1145/3452296.3472921>

作者简介



苏远超, 阿里云计算有限公司广域网架构与研发总监; 负责阿里巴巴全球网络规划、建设和开发, 主要研究领域为 IP 和光网络架构、协议及流量工程系统; 发表多篇论文, 拥有多项专利。

高性能智算网络架构探讨与实践



High-Performance AI Computing Network Architecture: Exploration and Practice

黄光平/Huang Guangping, 肖敏/Xiao Min, 张征/Zhang Zheng

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202603008

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260629.1134.002>

网络出版日期: 2026-06-26

收稿日期: 2026-03-25

摘要: 智算网络对超高吞吐、超低时延、超高可用性的极致性能需求, 是传统网络架构的核心挑战和瓶颈。论述了一种新型高性能智算网络架构, 其关键特征包括端网精细化协作、精准流控、无状态组播、轻量级在网计算等。基于智算网络的关键性能堵点, 提出了一种新智算网络架构, 其主要特征是端网双域协同、网侧定量反馈以及无状态组播, 概述并分析了原型测试及主要结论数据。

关键词: 智算网络; 端网协同; 无状态组播; 精准流控; 轻量级在网计算

Abstract: The extreme performance requirements of artificial intelligence (AI) computing networks, namely ultra-high throughput, ultra-low latency, and ultra-high availability, pose critical challenges and bottlenecks for traditional network architectures. A novel high-performance AI network architecture is presented, characterized by fine-grained host-network coordination, precise flow control, stateless multicast, and lightweight in-network computing. To address the key performance bottlenecks of AI networks, a new architectural design is proposed, featuring dual-domain host-network collaboration, quantitative network-side feedback, and stateless multicast. The corresponding prototype tests and primary performance data are outlined and analyzed.

Keywords: artificial intelligence network; host-network coordination; stateless multicast; high accuracy flow control; lightweight in-network computing

引用格式: 黄光平, 肖敏, 张征. 高性能智算网络架构探讨与实践 [J]. 中兴通讯技术, 2026, 32(3): 40-44. DOI: 10.12142/ZTETJ.202603008

Citation: Huang G P, Xiao M, Zhang Z. High-performance AI computing network architecture: exploration and practice [J]. ZTE technology journal, 2026, 32(3): 40-44. DOI: 10.12142/ZTETJ.202603008

相对传统网络业务, 高频且超大体量的数据传输是智算网络最典型的差异化业务特征。同时, 智算网络对丢包和拥塞的容忍度极低, 这对传统数据中心网络及广域网均构成重大挑战。传统网络在架构设计上本质是容损、容拥塞的, 其设计目标仅为“尽可能减少”丢包与拥塞, 因此无法满足智算网络的极致性能需求。

就丢包问题而言, 在多业务共存的流量转发模式下, 网络侧通过缓存与队列机制降低拥塞导致的丢包, 但受限于缓存资源容量及多流间的资源竞争。在端侧, 现有方案通过网络拥塞节点为业务报文添加拥塞标签并反馈至端侧, 端侧据此进行业务流调速, 以减轻网络拥塞压力。然而, 这种依赖性参数的端网粗放式协调机制, 无法充分调度网络可用资源, 难以从根本上解决拥塞问题。

在混合专家模型训练及推理场景中, 高频且体量巨大的全互连 (All-to-All) 通信模式下, 点对多点 (P2MP) 与多

点对点 (MP2P) 流量不仅消耗宝贵的网络节点带宽资源, 还会对端侧输入/输出 (I/O) 造成严重的数据交互瓶颈。然而, 这种资源消耗与性能瓶颈并非不可避免。

基于此, 本文提出一种全新的智算网络架构视角。针对拥塞问题, 将解决范畴从单一网络域拓展至端网精细化协同, 将当前网络对端侧的黑盒模式转变为白盒模式, 即端网协同由定性参数升级为定量参数, 实现业务需求与网络资源的精细化动态匹配, 从而从根本上提前规避拥塞, 而非被动感知拥塞后再作通告。在智算模型训练与推理的 P2MP 与 MP2P 场景中, 分别引入无状态组播机制与轻量级在网计算机制, 将关键性能瓶颈点 (如端侧 I/O) 上的 N 条相同数据流, 通过组播复制与在网聚合计算归并为 1 条数据流, 从而彻底解决性能瓶颈问题。

1 精细化端网协同

面向大容量数据在广域网中的高性能传输, 端网精细化

协同架构引入了端网协商速率机制。该机制向端侧主机动态分配并授权发送流量的配额，以主动防止拥塞；同时，基于配额进行资源调度、准入控制和流量控制，从而满足最优传送时间目标。具体而言，通过端网速率协商，应用端能够以更精细的方式高效、及时地调整发送速率，避免被动依赖端侧传输协议的调速机制，从而有效提升广域传输吞吐量；通过基于配额的资源预留，网络增强了对流量的主动调节与资源调度能力，保障所有任务的传输需求与资源供应；通过流量调度与准入控制，实现多流之间网络带宽资源的合理分配与动态调度，在满足高吞吐要求的同时避免慢增长尾现象，并实现集合通信多流传输的高精度同步。此外，该架构还采用快速拥塞感知及通告机制。当网络拥塞发生时，系统可在近源端进行快速反馈，从而迅速、准确地对流量速率进行通告与调整，及时缓解网络拥塞。

1.1 端网速率协商机制

突发大容量数据流量的传输可能引发网络瞬时拥塞、丢包与排队延迟。在传统拥塞控制机制中，端侧对网络带宽资源状态缺乏感知，导致调速出现锯齿效应，并使吞吐量下降。高性能广域网需引入端网协商速率机制，实现速率协同与拥塞避免。

精细化端网协同机制下包含3种速率协商策略^[1]：

1) 最优速率传输。网络通过资源调度机制为大容量数据提供服务质量(QoS)保障，实现最优速率传输，端侧主机可按协商的最优速率或最优速率范围进行传输。

2) 最小速率传输。网络为大容量数据提供最小资源预留保障，实现最小速率传输，端侧主机可按不低于协商值的速率进行传输。

3) 最大速率传输。网络为大容量数据提供资源预留的上限，实现最大速率传输，端侧主机可按不高于协商值的速率进行传输。

1.2 基于配额的资源预留

在智算场景下，任务式数据传输是主流模式，其主要特征为流量的周期性与可预期性。基于配额(quota)的调度是一种资源管理策略。该策略可扩展至成熟协议，例如资源预留协议-流量工程(RSVP-TE)，使其携带配额信息。配额可定义为一定时间内的可用资源(包括带宽、队列、缓存等)，涵盖动态资源与静态资源两类。静态资源如确定性链路，未申请资源预留的资源仅用于准入判断。网络可根据数据传输任务的需求分配和授权配额，并基于配额实施资源调度，从而实现主动拥塞避免，保障基于配额及对应速率的高

效转发。网络配额保障策略需根据当前网络资源状态进行选择，为此网络需先收集资源信息，再依据具体任务需求选取合适的资源保障策略^[2]。

1.3 流量调度及准入

网络侧在接收流量后，依据协商速率进行流量调度与策略执行，具体包括：对流量进行分类，按业务类型划分优先级，并为关键流量提升服务质量(QoS)等级；对流量进行整形，例如聚合“老鼠流”或分片“大象流”等。网络边缘的流量策略执行能够规范数据流，消除拥塞并最小化流量传输完成时间。基于协商速率实施速率调控时，可依托特定信令协议(如RSVP-TE)进行基于速率的动态资源调度与准入控制。其中，通过最小速率对应的最小带宽配额，保障单流传输完成时间；通过最大速率对应的最大带宽上限，避免多流竞争引发拥塞丢包^[3]。网络节点应依据协商的QoS等级与速率执行准入与流量控制。该节点通过结合准入控制与拥塞控制，可在高效利用网络容量的同时，实现高吞吐量与低时延。

2 精准流控

在传统数据中心网络中，基于优先级的流量控制(PFC)是主流解决方案。拥塞节点通过专用的二层报文，向上游节点发送拥塞反压消息，并逐跳回传，直至到达数据发送源端。然而，随着跨数据中心智算业务的兴起，这类面向无损传输的流控方案需延伸至广域网，这对现有广域网提出了两大挑战：1) PFC逐节点模式要求广域网全网升级，部署代价高昂；2) 广域网需承载多业务共存，必须极力避免特定业务(如智算模型训练与推理)的拥塞对其他业务流量造成严重影响。

相对于PFC基于队列的粗粒度流控模式，为适配广域网多业务、多租户共存共享网络资源的现实场景，本文提出一种面向租户或切片的精细化流控方案，即基于互联网协议第6版段路由(SRv6)的精准流控方案。如图1所示，当拥塞发生时，系统触发向指定源端的反压信号传输。该反压路径可沿SRv6路径向源端通告本地拥塞信息，包括队列状态、缓存深度等。与当前数据中心网络PFC机制不同，SRv6可由路径源端或控制器指定转发路径，因此内在支持插花式组网模式——即无需所有节点均具备PFC报文处理能力，仅需在选定节点(如数据中心边缘节点、长距广域网中继节点)支持PFC段标识(SID)处理能力。PFC报文的目标节点则可基于SID精准指向特定租户的子接口或切片。目的节点收到PFC报文后，可根据具体拥塞状态，选择停止或调低

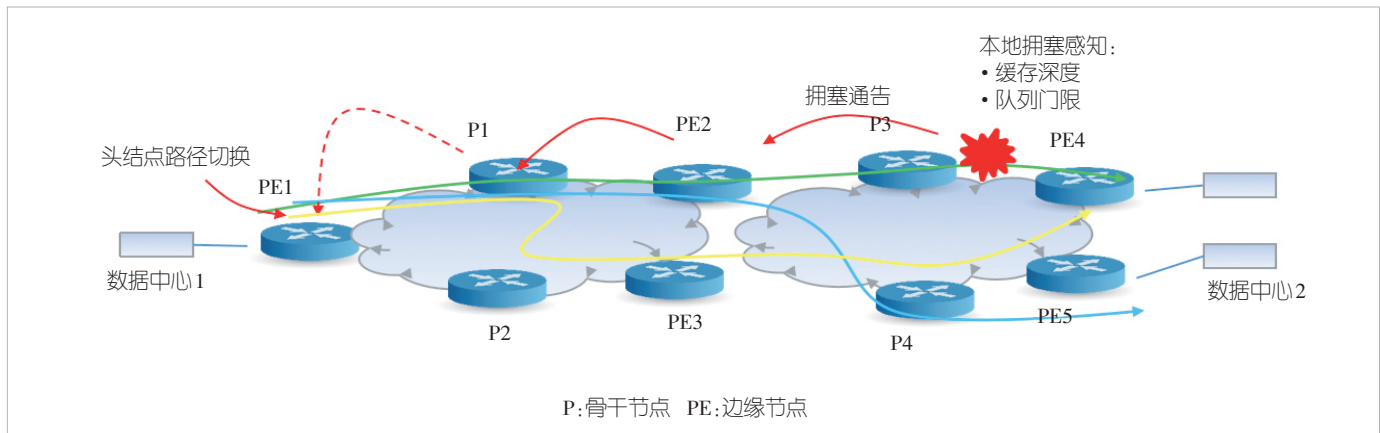


图1 基于SRv6的精准流控方案示意图

发送速率，或切换至备用路径。

拥塞控制的另一个成熟机制是显式拥塞通告（ECN）。在智算场景的极致性能需求下，广域网端到端的往返时延（RTT）成为关键瓶颈。为实现快速拥塞通告与反馈，网络拥塞节点可向就近的代理网关节点发送快速拥塞通告报文，再由该代理网关节点将报文转发至发送端设备。为此，系统需为每个端侧设备指定一个用于快速拥塞通告的代理网关节点，该节点应能够识别并解析端侧设备所支持的拥塞通告报文格式。代理网络节点通过内部网关协议（IGP）或边界网关协议（BGP）向外通告自身的拥塞通告代理能力及其所代理的端侧设备的IP前缀。网络中的其他设备收到该通告后，记录代理节点与端侧设备之间的映射关系表。当网络中发生拥塞时，检测到拥塞的网络节点根据拥塞报文的源IP地址查找对应的代理网络节点，并扩展互联网控制报文协议（ICMP）或用户数据报协议（UDP），向该代理节点发送快速拥塞通告报文，最终由代理节点将拥塞通告报文发送至源端设备^[3-4]。

3 无状态智算组播

极致的传输性能要求使得智算网络同样需要极致的转发效率。在混合专家模型（MoE）下的All-to-All传输场景中，组播技术能够在I/O性能和带宽资源利用上带来显著收益。然而，传统组播技术如协议无关组播（PIM）通常需要在网络节点维护较重的状态信息，因此难以满足智算场景下组播传输的需求。基于位索引的显式复制组播（BIER）因其内生无状态且业务与网络分离的特性，具有明显优势，并能更好地满足动态性需求。

图2展示了MoE组播技术框架。在应用层与网络层的BIER之间，设置了一个网络适配层。该层负责将LLM的组

播请求与BIER组播转发机制进行衔接，从而使BIER技术能够应用于智算MoE组播场景。

3.1 MoE 专家组播实现

MoE通过将超大网络拆分为若干“小专家”，每次输入仅激活少数相关专家，从而在不增加计算量的前提下大幅提升模型参数量。MoE模型的核心流程分为两个阶段：首先，通过路由机制（Gating函数）为每个令牌（token）分配指定数量的专家；其次，在分发（Dispatch）阶段将令牌发送至对应专家，待选定专家完成计算后，再通过汇总（Combine）阶段对计算结果进行整合^[5]。

在Gating过程中，由于专家选择过程极短（通常在毫秒级内），且选择具有高动态性，为实现此场景下的高效组播，BIER是最优的基础方案。

在应用层执行Gating操作并选定需要投递的专家后，将目的专家集合通知组播适配层，由适配层将其封装至网络层BIER报文头中。在BIER中，组播目的地址通过报文头中的位串（BitString）字段标识，BitString中的每一位均可对应一个目的节点。根据转发芯片的实现，BitString可支持64~4 096个目的地址。在MoE组播专家投递过程中，选定专家

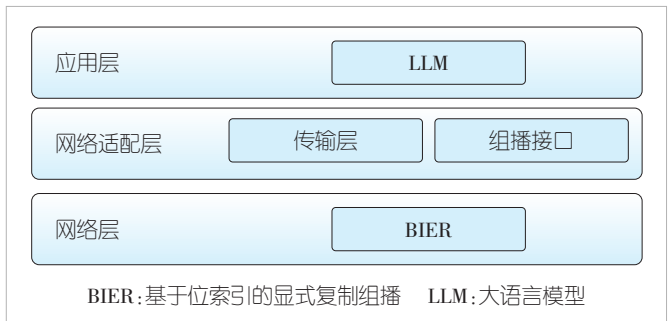


图2 MoE组播技术框架

群作为 BIER 组播转发的终点，数据经 BIER 封装后，Bit-String 中的每一位表示一个目的专家。这些专家可位于同一节点内，也可分布在跨越 Leaf 或 Spine 设备的不同节点上。

在实际 AI 任务开始之前，需先对专家进行编号。例如，对于包含 256 个专家的大语言模型（如 DeepSeek-V3），可将专家编号为 1~256。每个专家 IP 地址与编号的对应关系，可通过两种方式建立：一是在节点内部交换机及 Leaf 交换机上进行静态配置；二是通过路由协议，如 BGP、开放最短路径优先（OSPF）、中间系统到中间系统（IS-IS）或胖树路由协议（RIFT）进行自动通告。随后，节点内部交换机、Leaf 及 Spine 交换机根据这些信息计算并生成 BIER 转发表项。至此，每个专家均可根据转发表项，查找到其对应的下一跳交换机或出口。

每次 Gating 完成专家选择后，被选中的专家编号将直接映射至 BIER 报文的 BitString 字段，对应的位被设置为 1。例如，DeepSeek-V3 模型中若选中 9 个专家（含 8 个路由专家与 1 个共享专家），其编号分别为 1、5、6、12、15、21、22、32、50，则 BIER 报文中相应位（第 1、5、6、12、15、21、22、32、50 位）均被设置为 1。封装完成后的 BIER 报文，经节点内部交换机或 Leaf/Spine 交换机转发，最终到达目的专家。整个过程无需为选中的专家预先建立组播转发树，而是在 Gating 决策完成后直接封装并发送组播报文，从而极大提升了组播效率。

3.2 传输层协作

在智算 MoE 场景中，可能涉及大量内存操作，特别是远程直接内存访问（RDMA）的单边操作（如 READ/WRITE）。这些操作对应于特定图形处理器（GPU）上不同的内存地址。在使用 BIER 技术时，存在两种不同的实现方式：

方式 1：基于映射的转换方案

如果每个 GPU 为该 AI 任务分配的内存地址不同，则 BIER 数据传输后需写入各 GPU 对应的内存地址。当应用程序存在实现限制，无法直接交由网络适配层封装 BIER 报文时，也可采用此方式进行转换。

当网络适配层收到应用发来的报文时（该报文可能为基于 RDMA 协议的单播报文），为实现与 BIER 技术的结合与转换，传输层协议（如 RDMA）可约定某些 AI 任务与操作对应的组播地址。此地址可以是传统组播 IP 地址，但这并非回退使用传统组播技术，而是指 I/O 接口、节点内部交换机或 Leaf 交换机在收到该报文后，可执行 BIER 报文的封装与发送。当数据报文丢失时，传输层可迅速重新发送丢失的报

文，重传仍然采用 BIER 封装，仅需将未收到数据的专家对应的目的位填入 BIER 头即可。

BIER 的封装转换可发生在 GPU 的 I/O 接口、相连的内部交换机或 Leaf 交换机上。这些节点仅需预先保存组播 IP 地址与 BIER 目的位之间的映射关系，收到报文后即可进行封装转换。报文经 BIER 组播转发到达目的后，同样依据映射关系，定位并写入各个目的 GPU 及其对应的内存地址。

方式 2：统一内存地址方案

在开始 AI 任务之前，由参与该任务的所有 GPU 划分出统一的内存地址以供组播操作。例如，在训练或计算时，所有专家所在的 GPU 上均预先划分出一块统一的内存地址，用于执行 READ 或 WRITE 组播操作。此方案省略了映射环节，从而提高了 BIER 组播应用的效率。

3.3 多 BIER 域适配不同应用

BIER 技术支持根据不同应用场景划分独立的子域进行适配。例如，在 MoE 专家投递场景中，可为 MoE 专家划分一个 BIER 子域；在 Agent 场景中，可为 Agent 划分另一个 BIER 子域；对于跨数据中心（DC）的组播数据发送，则可将相关设备划分至第 3 个 BIER 子域。不同子域中的目的编号可以重复使用。BIER 技术本身支持多达 256 个子域，能够充分满足各类场景的差异化需求^[6]。

4 轻量级在网聚合

与内容分发网络（CDN）中的单向组播不同，智算网络 All-to-All 场景下的组播接收端往往需要向发送端返回确认消息，即组播逆向流程。同样，这种 MP2P 的确认消息回传也会给端侧 I/O 系统带来极大的性能压力，并造成带宽资源浪费。若在边缘网络节点对这些结构简单的确认消息进行轻量级聚合处理（例如引入定时限内的目的节点位图机制），则仅需将聚合后的消息发回组播发送端，从而有效避免大量确认消息经由 I/O 系统逐条返回端侧。

5 高性能智算网络原型测试数据及结论分析

高性能智算网络仿真系统由若干端侧服务器与 7 台跨区域路由器构成，路由器采用星型与网状混合组网拓扑。该仿真系统引入了端网精细化速率协商、近源端快速通告及精准流控机制。如图 3 所示，左侧为未引入高性能智算网络机制时的发送速率收敛曲线，锯齿效应和长收敛周期较为明显；右侧为引入本文所述部分方案后的对比曲线，锯齿效应基本消除，收敛周期显著缩短。仿真数据表明，精细化端网速率协同机制及面向广域网的精准流控机制，能够显著改善智算

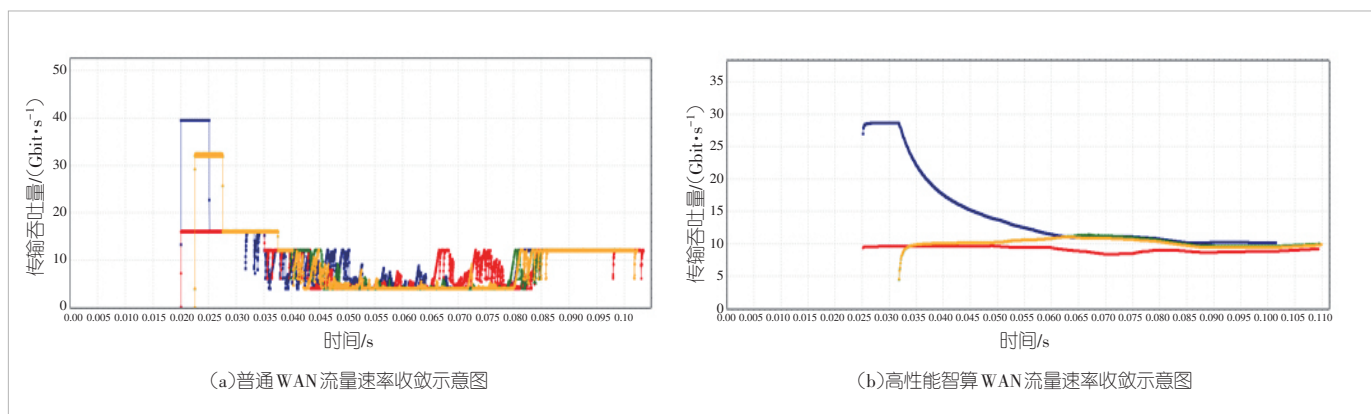


图3 高性能智算网络原型仿真测试速率收敛对比

网络场景下超大数据流量的广域网传输吞吐量,并缩短任务完成时间,从而为跨数据中心模型训练与推理的综合效率提供有力支撑。

6 结束语

高性能智算网络是支撑人工智能应用的核心网络底座,其在跨数据中心场景下的性能瓶颈将严重制约智算网络的规模化部署与拓展。本文提出了全新的架构视角,并针对关键性能瓶颈给出了端到端的解决方案:将端网协同从定性参数层面升级至定量参数层面,将拥塞处理从被动通告转变为主动规避;同时,针对广域网多业务共存的现实特征,提出了基于SRv6的精准流控与快速通告机制;在MP2P和P2MP的智算模型训练与推理场景中,分别引入无状态组播与轻量级在网聚合机制,有效缓解了网络带宽资源浪费及端侧I/O性能瓶颈问题。基于该架构的原型系统测试结果表明,端网精细化协同机制显著提升了数据传输吞吐量与整体传输效率。

参考文献

- [1] Xiong Q, Huang D, Yao K, et al. Framework for high performance wide area network (HP-WAN) draft-xhy-hpwan-framework-03 [EB/OL]. (2026-04-23)[2026-04-25]. <https://datatracker.ietf.org/doc/draft-xhy-hpwan-framework/>
- [2] Xiong Q, Zhu X Y, Lin C W. Signaling solution for HP-WAN draft-xiong-hpwan-signaling-solution-01 [EB/OL]. (2026-03-02)[2026-04-05]. <https://datatracker.ietf.org/doc/draft-xiong-hpwan-signaling-solution/>
- [3] Min X, Zhang K, Hu Z H. Fast congestion notification packet (CNP) with Proxydraft-xiao-rtgwg-proxy-congestion-notification-03 [EB/OL]. (2026-03-15)[2026-04-05]. <https://datatracker.ietf.org/doc/draft-xiao-rtgwg-proxy-congestion-notification/>
- [4] Liu Y, Yao J, Lin C W, et al. Congestion control based on SRv6 pathdraft-liu-rtgwg-srv6-cc-01 [EB/OL]. (2026-02-27)[2026-04-15]. <https://datatracker.ietf.org/doc/draft-liu-rtgwg-srv6-cc/>
- [5] Zhang H, Xu X H, Zhang Z, et al. Optimized use of BIER in AIML

data centers draft-zzhang-bier-optimized-use-in-aidc-01 [EB/OL]. (2026-03-16)[2026-04-25]. <https://datatracker.ietf.org/doc/draft-zzhang-bier-optimized-use-in-aidc/>

- [6] Zhang Z, Duan W, Xu X H, et al. Multicast use case in LLM MoE draft-zhang-rtgwg-llmmoe-multicast-02 [EB/OL]. (2026-04-13)[2026-04-26]. <https://datatracker.ietf.org/doc/draft-zhang-rtgwg-llmmoe-multicast/>

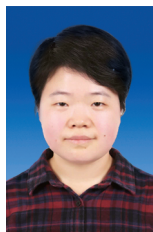
作者简介



黄光平, 中兴通讯股份有限公司有线标准总监;长期从事IP网络、算力网络、确定性网络以及智算网络等领域的技术研究和标准制定工作;发表论文10余篇,申请发明专利50余项,提交国际标准提案100余项。



肖敏, 中兴通讯股份有限公司技术预研资深专家;长期从事数据通信领域的技术研究和标准预研工作;申请发明专利70余项,参与制定标准22项。



张征, 中兴通讯股份有限公司技术预研资深专家;长期从事智算网络、广域网络领域的技术研究和标准制定工作;申请发明专利40余项,参与多项标准的制定。

面向分布式推理的大模型键值缓存分发网络



Large Model Key-Value Cache Distribution Network for Distributed Inference Services

宋俊平/Song Junping¹, 黄天一/Huang Tianyi^{1,2},
周旭/Zhou Xu¹, 李智/Li Zhi³, 吴茜/Wu Qian³

(1. 中国科学院计算机网络信息中心, 中国 北京 100083;

2. 中国科学院大学, 中国 北京 100190;

3. 中国移动通信集团有限公司, 中国 北京 100032)

(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China;

2. University of Chinese Academy of Sciences, Beijing 100190, China;

3. China Mobile Communications Group Co., Ltd., Beijing 100032, China)

DOI: 10.12142/ZTETJ.202603009

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260629.1815.007>

网络出版日期: 2026-06-26

收稿日期: 2026-04-25

摘要: 大模型推理已成为智能算力体系的核心, 传统云端集中式推理面临带宽压力大、推理延迟高、算力资源紧张等瓶颈。现有云边协同方案在通信开销与键值缓存 (KV-Cache) 复用效率方面仍存在明显不足。为此, 提出一种基于内容分发网络 (CDN) 的大模型 KV-Cache 分发网络 (KVCDN), 将传统 CDN 升级为具备张量感知与状态管理能力的分布式缓存架构。该架构采用端侧语义规范化、边缘热点分级缓存与核心全局去重编排 3 层协同结构, 并结合提示词 (Prompt) 标准化、分级存储预放置、全局去重融合与语义路由 4 项关键技术, 从而显著提升 KV-Cache 复用率, 降低冗余计算与推理延迟。

关键词: 大模型推理; 键值缓存 (KV-Cache); 内容分发网络 (CDN); 云边协同; 缓存复用

Abstract: Large language model inference has become a core component of intelligent computing systems. Traditional cloud-centric inference approaches face critical bottlenecks such as high bandwidth consumption, excessive latency, and intensive computational demands. Existing cloud-edge collaborative schemes still suffer from significant communication overhead and low cache reuse efficiency of Key-Value Caches (KV-Cache). To address these issues, a KV-Cache distribution network based on content delivery network (CDN), termed KVCDN, is proposed. In this architecture, conventional CDN nodes are upgraded into a distributed caching system with tensor-aware perception and state management capabilities. A three-layer collaborative structure is adopted, consisting of semantic normalization at the end side, hotspot-aware hierarchical caching at the edge, and global deduplication with orchestration at the core. Combined with four key enabling techniques, i.e., prompt standardization, tiered storage and proactive pre-placement, global deduplication and fusion, and semantic-aware routing, the KVCDN significantly improves KV-Cache reuse ratio, while reducing redundant computations and inference latency.

Keywords: large model inference; Key-Value Caches (KV-Cache); content delivery network (CDN); cloud-edge collaboration; cache reuse

引用格式: 宋俊平, 黄天一, 周旭, 等. 面向分布式推理的大模型键值缓存分发网络 [J]. 中兴通讯技术, 2026, 32(3): 45-50. DOI: 10.12142/ZTETJ.202603009

Citation: Song J P, Huang T Y, Zhou X, et al. Large model key-value cache distribution network for distributed inference services [J]. ZTE technology journal, 2026, 32(3): 45-50. DOI: 10.12142/ZTETJ.202603009

1 研究背景

随着大模型从实验室走向千行百业, 人工智能 (AI) 技术正从以训练为核心转向以推理为核心。国际数据公司 (IDC) 在其发布的《中国人工智能算力发展评估报告》中指出, 到 2027 年, 中国推理算力占整体智能算力的比例将突破 72.6%, 占据绝对主导地位。

第十四届全国人民代表大会第四次会议和中国人民政治协商会议第十四届全国委员会第四次会议的相关提案指出, 中国当前面向海量推理任务的算力集群和网络调度能力仍存在显著缺口。与此同时, 通用大语言模型的广泛部署与海量并发应用带来了极高的算力和访存挑战。传统高度集中的云端单点推理模式不仅给国家骨干网络带来了巨大的带宽压

力,也越来越难以满足海量用户对低延迟交互的严苛要求。

作为互联网内容分发的重要基础设施,内容分发网络(CDN)凭借其广泛的省市级边缘节点部署,天然具备为用户提供低延迟服务的优势。然而,传统CDN主要面向静态文件或流媒体内容进行缓存,直接将大模型部署于CDN之上,在创新性和资源利用率方面均存在明显不足。大模型推理的核心瓶颈在于处理自回归生成过程中产生的海量键值缓存(KV-Cache)。随着大模型对超长上下文处理需求的爆发式增长,中间态张量KV-Cache的管理正从单机显存优化,逐步向广域分布式解耦与跨节点协同架构演进。如何突破单机物理硬件的局限,利用网络拓扑的空间分布特性实现KV-Cache的高效分级与调度,已成为当前网络与系统交叉领域的前沿热点。

2 研究现状

在大模型算力需求爆发的初期,学术界主要沿用了传统深度学习中的模型切分思路,提出了一种基于算力空间转移的云边协同架构^[1]。该架构通过在物理空间上对大语言模型(LLM)的前向计算图进行基于层级或注意力头的静态切分^[2],将计算负载较低的嵌入层及浅层Transformer模块部署于边缘节点,用于提取初级语义特征、承接初始流量并兼顾用户隐私保护^[3];而将对显存需求极高的深层模块与全量权重保留在云端智算中心,以确保复杂推理的精度^[4]。这一策略旨在通过层级间的流水线并行,缓解集中式数据中心在处理海量并发请求时面临的瞬时算力压力^[5]。

然而,这种静态切分方案在LLM特有的自回归推理场景下面临严峻的通信挑战。

由于LLM中间层输出的张量维度远超原始文本,且解码(Decode)阶段频繁的跨节点同步容易受到广域网带宽波动的影响,导致推理时延显著增加。近期研究(如CE-CoLLM^[6])进一步证实,云边协同中LLM上下文数据的跨节点传输已成为核心通信瓶颈,并初步探索了延迟感知的云端上下文管理机制。

为突破静态切分方案带来的“通信墙”瓶颈,学术界开始探索具备环境感知能力的动态协同机制。Splitwise系统^[7-8](ISCA 2024、UCC 2025)提出了一种基于Lyapunov优化理论的动态协同框架。该研究摒弃了粗粒度的层间切割方式,转而通过深度强化学习实时感知网络带宽与节点负载,在注意力头和前馈神经网络(FFN)等细粒度层级上自适应调整云边切分边界。这一突破将云边协同从静态物理划分提升至动态资源调度层面,显著增强了系统在不稳定网络环境下的推理鲁棒性。

在此基础上,Jupiter系统^[9](INFOCOM 2025)进一步针对LLM推理中预填充(Prefill)与Decode阶段的计算不对称性进行了优化。该研究指出,传统方案往往仅注重优化预填充阶段,而严重忽视自回归解码阶段,导致解码过程中因“逐层、逐Token”的串行依赖而使边缘资源长时间空转。为此,Jupiter通过对边缘硬件进行精细化的负载轮廓化建模,提出了一种计算与通信深度重叠的流水线解码机制,即利用云端处理当前Token深层权重所耗费的时间窗口,驱动边缘端同步预执行下一个Token的浅层计算,从而有效消除了流水线中的“计算气泡”。这种从空间切分向时间维度相位解耦的演进,为实现低延迟、高吞吐的生成式服务提供了关键支撑。

尽管上述基于流水线切分的云边协同研究在理论上极具吸引力,但在实际部署于跨省、跨市的广域网环境中时,其内在的结构性缺陷依然明显。首先,通信开销的本质瓶颈尚未消除。模型切分技术在推理运行期必然于切割层产生海量的高维隐状态张量。这类中间计算结果的跨网流转对物理带宽的绝对容量与抖动极其敏感,在网络长尾延迟的影响下,极易引发首字延迟(TTFT)严重超时,难以满足实时交互业务的苛刻需求。其次,更为致命的症结在于,这类架构将LLM推理视作一条无记忆的刚性单向流水线。现有研究普遍忽视了真实业务场景中海量用户提示词(Prompt)之间在语义内容上的强相关性、上下文的重叠度以及计算状态的可复用性,导致严重的重复计算与存储冗余。

为了突破上述物理切分方案带来的跨网张量通信瓶颈,近期的协同推理研究重心正逐渐从单一的“模型空间切割”转向非对称的“云-端协同”与“边-端异步协同”。针对终端算力逐渐增强且广域网条件存在波动的场景,学术界广泛引入了基于投机解码的协同架构^[10-11]。该机制的核心在于,将传统大语言模型逐字生成的串行模式重构为“端侧起草—云/边侧验证”的并行范式:首先,由部署在用户终端或边缘节点的轻量级小模型凭借其极速响应能力,一次性“投机”起草一段候选文本;随后,由算力充沛的云端大模型对该段草稿进行一次性的并行校验。这种“端侧小模型试错、云端大模型定论”的非对称计算结构,从根本上改变了自回归生成中“算一个词、传一个词”的低效模式,同时也降低了大模型推理对网络实时性的严格依赖。该方案在保证输出精度与原生大模型相当的前提下,显著降低了云端交互频率,并提升了端到端的响应速度。

在此背景下,分布式拆分投机解码(DSSD)系统^[12](ICML 2025)针对分布式投机解码中“验证数据量过大”这一瓶颈,展现了较强的创新性。传统方案中,终端小模型每

生成一段草稿,均需向边缘或云端上传庞大的词表概率分布以供校验,导致较高的上行通信开销。DSSD将原本由边缘节点承担的复杂计算逻辑进行逆向重构,使原本需要海量上传的概率数据转化为极其轻量化的单次下行确认信号,从而实现了通信延迟的数量级压缩。这一成果标志着分布式推理从单纯压缩数据体积向重构通信逻辑的重要跨越。

虽然投机解码大幅优化了单次任务的执行效率,但其底层机制仍未有效解决大模型推理的核心难题——推理状态的持久化与语义复用。

在现有系统中,动态上下文KV-Cache虽然占据着庞大的显存空间,却在每次验证或请求结束后被视作无保留价值的临时变量而直接清空。这种无记忆的资源处理模式,导致边缘侧在处理具有高度语义重叠的并发任务时(如多轮对话、共用参考资料)陷入“计算孤岛”困境——每条语义关联的请求都被迫从零开始进行重复的预填充与特征提取,从而造成了严重的计算冗余与存储浪费。因此,如何从“瞬态通信优化”转向“稳态状态管理”,构建具备跨任务记忆能力的云边协同架构,已成为当前制约大语言模型迈向大规模工业化应用的核心科学问题。

随着超长文本推理与检索增强生成(RAG)技术的普及,学术界逐渐形成共识:大模型推理产生的中间态数据亟需引入一种具备持久化、模块化与空间分布式特性的新型基础设施。这一共识直接催生了将传统CDN向“大模型知识网络”演进的理论构想。

芝加哥大学系统研究团队在操作系统原理会议(SOSP)2024及2026年的后续工作中,首次明确提出了“知识分发网络(KDN)”的概念^[13],并开源了核心原型系统LM-Cache^[14]。该研究明确指出,传统的模型微调虽能内化知识,但会严重破坏模型的模块化与响应速度;而常规的上下文学习则会在预填充阶段产生超线性增长的显存与计算成本。为此,KDN提倡将大模型基于外部知识源预计算生成的KV-Cache进行物理剥离,将其视作互联网中的“静态视频或网页资源”进行独立管理与分发。在KDN架构中,系统通过独立的存储模块,以及支持多源KV-Cache动态拼接组合的智能混合模块,使大模型能够像传统CDN分发多媒体流一样,在地理分布的存储节点之间实现推理状态的动态组装。

综上所述,尽管投机解码等技术显著优化了云边协同推理中单次任务的执行效率,现有系统仍将大模型推理视为无记忆的流水线,未能解决推理状态持久化与语义复用这一核心难题。同时,随着超长文本推理及RAG技术的普及,学术界逐渐形成共识:大模型的中间态数据亟需引入一种具备持久化与空间分布式特性的新型基础设施^[15]。正如KDN

概念所倡导的,将大模型预计算生成的KV-Cache视作互联网中的“静态网页资源”进行独立管理与物理剥离,已成为突破“计算孤岛”困境的关键路径。

3 基于CDN的大模型键值缓存分发网络

在广域算网融合与分布式AI技术高速演进的背景下,大模型推理服务正经历从“云端集中式单点处理”向“云边端多级协同”的深刻范式转移。传统CDN主要面向静态文件、图片或流媒体视频的分发而设计,其底层逻辑基于统一资源定位符(URL)精确哈希与就近物理路由。然而,在大模型协同推理服务中,核心性能瓶颈已转移至自回归生成过程中产生的海量高维动态特征张量——KV-Cache。同时,Prompt具有高度非结构化、强口语化及上下文依赖的特征,使得传统CDN的静态文件缓存机制在此类高维、动态、碎片化的智能算力场景下完全失效。

为了彻底打破“缓存击穿”与“管理粗放”的瓶颈,本文提出一种基于CDN的大模型键值缓存分发网络(KVCDN)。该方案在不改变现有算网物理拓扑的前提下,将传统数据分发节点全面升级为具备张量感知与状态管理能力的智能存储单元。整体技术路线依托“端侧规范化重塑—边缘领域热点分级—核心全局去重”3层协同机制,深度融合计算感知路由与可编程网络能力,从而形成一套端到端的大模型缓存优化解决方案。

如图1所示,KVCDN架构在逻辑上划分为3个核心功能层:

1) 端侧/接入侧:作为防范“缓存击穿”的第一道防线,该层通过部署轻量级模型,对异构的口语化输入进行意图对齐、无损压缩与规范化重写,从数学源头上提升请求特征的结构化程度,并生成全局一致的语义指纹。

2) 边缘下级CDN侧:作为“存”的核心承载层,该设计摒弃单机内部的图形处理器-中央处理器-固态硬盘(GPU-CPU-SSD)硬件级存储思维,创新性地提出一种面向“物理空间—社会领域”的多级CDN网络拓扑分级存储架构。结合计算感知理念,利用特定地理区域的社会属性(如专科医院、科研院所),精准定位垂直领域的KV-Cache热点,并实现动态预放置。

3) 核心上级CDN侧:作为全局资源统筹枢纽,该层针对广域网中海量且时空碎片化的冗余张量,构建基于前缀感知的全局字典树索引。通过特征匹配与选择性重计算技术,突破严格前缀匹配的空间限制,实现跨区域的物理去重与融合共享。

面向大模型特征张量分发设计的KVCDN,从端侧请求

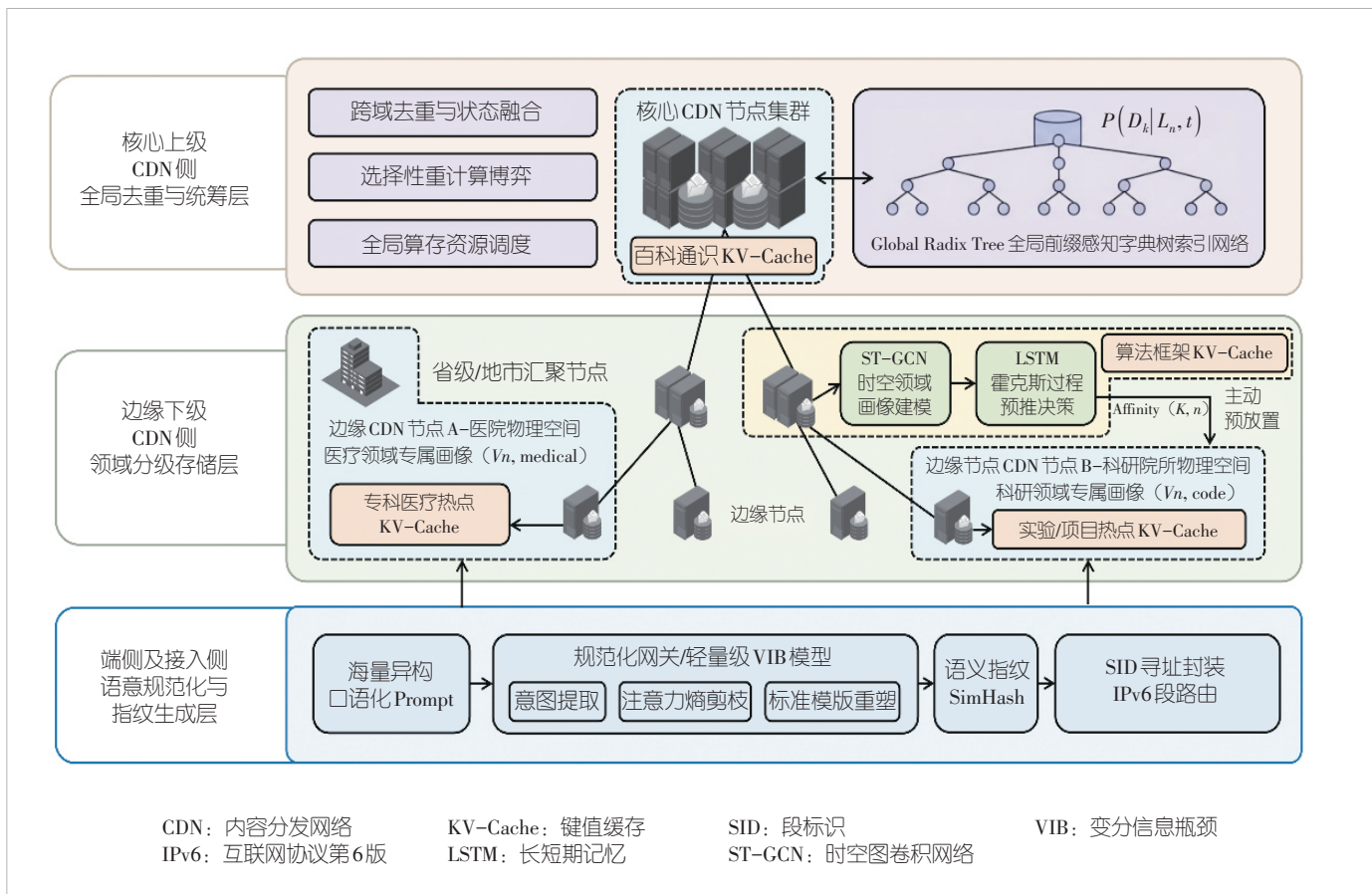


图1 一种基于CDN的大模型键值缓存分发网络架构

重塑到多级张量编排调度，提供了一整套系统性的缓存优化解决方案。针对高度异构且口语化的端侧大模型请求，该系统能够感知并提取核心语义意图，剖析非结构化输入与标准化提示词之间的映射关系，进而精准刻画并重塑具备统一前缀的标准化请求，为后续KV-Cache的高效聚合与全网命中提供源头基础。面对大模型长上下文所带来的海量显存消耗与动态多层次特征，该系统从局部高频响应与全局特征复用的角度，对KV-Cache张量进行编排管理，充分利用边缘与核心等多级算存资源，提升分布式特征张量的处理与调度效率。同时，结合广域网中多并发请求的高度时空变化特征，系统能够感知并预测局部长上下文热点态势，并根据冗余态势与前缀相似度，动态调整多层级间的存储与去重融合方案。

4 关键技术

为了提升基于CDN的大模型键值缓存分发网络的大规模分布式推理性能，解决其面临的云边端推理协同困难、缓存管理粗放、计算性能低下等问题，我们提出以下核心关键

技术思路：

1) 语义等价性驱动的端侧 Prompt 规范化与无损压缩机制

广域网中的海量大模型推理请求普遍存在极强的个体表达异构性。即便多个用户在核心逻辑与所需知识库上完全一致，诸如语气词、停用词或句法结构等方面的微小差异，也会导致CDN节点在进行底层特征张量匹配时产生严重偏差。这种输入异构性直接使本可复用的KV-Cache被系统判定为“未命中”，从而引发显著的算力浪费与冗余传输。

在端侧（或接入侧网关）引入基于语义等价性的规范化机制：基于信息瓶颈（IB）理论的轻量级序列提纯算法，构建轻量级编码器，将用户的原始口语化提示词序列 $X = [x_1, x_2, \dots, x_N]$ 映射为紧凑表示 Z ，在压缩原始序列表示（最小化 $I(X; Z)$ ）的同时，最大程度保留与核心意图相关的信息（最大化 $I(Z; Y)$ ）。在提纯意图的基础上，进一步对 Prompt 进行字粒度的无损压缩，利用预置在端侧的小型语言模型（如极小参数量的蒸馏模型）的注意力掩码机制，计算

每个Token的信息熵；利用局部敏感哈希（LSH）机制，将高维的语义嵌入向量映射为低维特征空间中的紧凑比特串 $F(X_{sid}) \in \{0,1\}^d$ ，作为后续网络路由层在IPv6扩展报头中封装的标识符，从源头确保具备相同语义的请求能够被精准汇聚。

2) 基于时空热度与空间拓扑感知的KV-Cache分级存储与动态预放置方法

大模型推理所产生的上下文数据体量庞大，且不同地理区域的用户请求呈现显著的地域特征与“语义热点”集聚现象。例如，中关村地区的科技类Prompt较为集中，而关于某些全球性社会议题的询问则分布极广。在此背景下，如何在CDN分级存储架构下，根据区域特征实现分布式KV-Cache的精准预热与动态替换，并进而基于语义与缓存热点完成精准路由适配，以最大化缓存命中率与复用率，已成为亟待解决的复杂网络调度问题。

考虑到广域网中大模型推理请求的地域聚集性与圈层性特征，以及多级CDN网络中不同层级节点在覆盖范围、并发承载力与存储资源等方面的差异，需建立基于时空特征的动态张量热度刻画模型。进一步地，从网络拓扑与用户分布的角度出发，设计面向KV-Cache的“拓扑分级”存储策略：将具备强地域性且局部极热的KV-Cache精准下沉并锚定至最接近端侧的边缘节点；将访问用户分散、通用性强的系统级KV-Cache上移至上层汇聚节点存储。同时，在基于KV-Cache的结构化合并树（如LSM-tree）底层引擎之上，结合局部区域的热度演化预测，通过跨边缘节点的热点主动预放置与负载均衡备份机制，消除突发高并发场景下的单点性能瓶颈，从而实现局部高通量推理请求的就近极速响应。

3) 基于前缀感知与选择性重计算的全局编排与去重融合方法

在广域网环境中，大量冗余张量并非完全一致。在多轮智能对话、RAG或多文档阅读等典型应用场景中，用户常基于相同的背景知识库（如一份长篇学术报告）插入不同的提问引导语，或对文档序列进行不同的拼接组合。这些碎片化操作打破了自回归模型所依赖的“严格绝对位置前缀匹配”规则，导致传统哈希缓存直接判定缓存失效，进而使不同CDN节点内堆积了大量实际高度同源却因非严格前缀限制而彼此独立的冗余KV张量碎片。

提出跨节点的KV-Cache深层解耦与动态去重融合方法。该方法在上层核心CDN节点构建全局前缀字典树索引目录，用于感知跨域节点间的冗余张量态势。针对非严格前缀序列，引入“选择性重计算”机制以恢复交叉注意力，从而根据全局冗余态势做出即时合并反馈，并迅速消除各节点

间的碎片化张量，最终实现全局空间去重与资源公用，最大化广域算存资源的统筹与利用率。

4) 基于KV-Cache缓存的语义路由技术

在现有CDN边缘网关处引入轻量级语义解析与动态调度能力，将高维自然语言请求转化为网络路由友好的低维标识，并融合全网算力与存储状态进行多目标联合寻优调度。

关键技术思路包含以下核心维度：

(1) 高效注意力裁剪抽取方法。面向大模型前缀匹配与长文本检索场景，设计高效注意力裁剪与抽取方法，将冗长的Prompt转化为紧凑的特征向量，在保障语义精度的同时，满足微秒级网络路由的时延要求。

(2) 分布式语义路由表。以语义向量和LSH为核心，构建分布式语义路由表。该表将稠密语义向量降维映射为定长（如64位或128位）的二进制哈希签名，并通过计算海明距离来刻画语义相似度。

(3) 云边协同分级路由与联合调度架构。在云边协同环境下，设计全局与局部路由表分级架构。综合考量KV-Cache缓存复用率（即预填充计算降低率）、CDN边缘节点实时算力队列负载以及云边网络传输时延，构建联合调度模型。利用深度强化学习或在线凸优化算法，实现高并发场景下的动态路由决策策略，有效防止热点缓存引发的局部流量风暴。

5 结束语

面向大模型推理从云端集中式向云边端协同演进的趋势，以及KV-Cache复用不足、广域通信开销大、推理延迟偏高的核心痛点，本文提出基于CDN的大KVCDN。通过将传统CDN升级为具备张量感知、状态管理与语义调度能力的智能分发架构，该方案实现了从“单机硬件调度”向“广域网络拓扑协同”的演进，以及从“被动等待命中”向“主动语义重塑”的转变。本文系统梳理了所提架构面临的核心关键挑战及可行的技术解决思路，旨在为大模型时代新型网络基础设施的发展提供参考方向。

参考文献

- [1] Qu G Q, Chen Q Y, Wei W, et al. Mobile edge intelligence for large language models: a contemporary survey [PP/OL]. arXiv (2025-03-20)[2026-04-20]. <https://arxiv.org/abs/2407.18921>
- [2] Kafetzis D, Khalili R, Koutsopoulos I. Large language model partitioning for low-latency inference at the edge [C]// Proceedings of the 23rd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks

- (WiOpt). IEEE, 2025: 1–8
- [3] Cunningham M. Privacy-aware split inference with speculative decoding for large language models over wide-area networks [PP/OL]. arXiv (2026-02-18) [2026-04-22]. <https://arxiv.org/abs/2602.16760>
- [4] Sung M, Palakonda V, Im S, et al. Memory- and latency-constrained inference of large language models via adaptive split computing [PP/OL]. arXiv (2025-11-06) [2026-04-20]. <https://arxiv.org/abs/2511.04002>
- [5] Feng Z D, Lu L, Li Q, et al. Distributed inference optimization for large language model in edge-cloud collaborative networks [C]// Proceedings of ICC 2025 – IEEE International Conference on Communications. IEEE, 2025: 6161–6166. DOI: 10.1109/icc52391.2025.11160773
- [6] Jin H P, Wu Y Z. CE-CoLLM: efficient and adaptive large language models through cloud-edge collaboration [C]// Proceedings of IEEE International Conference on Web Services (ICWS). IEEE, 2025: 316–323. DOI: 10.1109/icws67624.2025.00046
- [7] Patel P, Choukse E, Zhang C J, et al. Splitwise: efficient generative LLM inference using phase splitting [C]// Proceedings of ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA). IEEE, 2024: 118–132. DOI: 10.1109/ISCA59077.2024.00019
- [8] Younesi A, Shabrang M A, Oustad E, et al. Splitwise: collaborative edge – cloud inference for LLMs via Lyapunov-assisted DRL [C]// Proceedings of the 18th IEEE/ACM International Conference on Utility and Cloud Computing. ACM, 2025: 1–11. DOI: 10.1145/3773274.3774267
- [9] Ye S Y, Ouyang B, Zeng L K, et al. Jupiter: fast and resource-efficient collaborative inference of generative LLMs on edge devices [C]// Proceedings of IEEE INFOCOM 2025 – IEEE Conference on Computer Communications. IEEE, 2025: 1–10. DOI: 10.1109/infocom55648.2025.11044734
- [10] Chen C, Borgeaud S, Irving G, et al. Accelerating large language model decoding with speculative sampling [PP/OL]. arXiv (2023-02-02)[2026-04-18]. <https://arxiv.org/abs/2302.01318>
- [11] Leviathan Y, Kalman M, Matias Y. Fast inference from transformers via speculative decoding [C]// Proceedings of the 40th International Conference on Machine Learning. ACM, 2023: 19274–19286. DOI: 10.5555/3618408.3619203
- [12] Ning J, Zheng C, Yang T. DSSD: Efficient edge-device LLM deployment and collaborative inference via distributed split speculative decoding [PP/OL]. arXiv (2025-07-16) [2026-04-18]. <https://arxiv.org/abs/2507.12000>
- [13] Cheng Y, Du K, Yao J, et al. Do large language models need a content delivery network [PP/OL]. arXiv (2024-09-16)[2026-04-15]. <https://arxiv.org/abs/2409.13761>
- [14] Liu Y, Cheng Y, Yao J, et al. LMCACHE: an efficient KV cache layer for enterprise-scale LLM inference [PP/OL]. arXiv (2025-12-05)[2026-04-18]. <https://arxiv.org/abs/2510.09665>
- [15] Team TAib. AlBrix: towards scalable, cost-effective large language model inference infrastructure [PP/OL]. arXiv (2025-02-22)[2026-04-18]. <https://arxiv.org/html/2504.03648v1>

作者简介



宋俊平，中国科学院计算机网络信息中心高级工程师、中科院青年创新促进会会员、中国通信标准化协会 TC630 语义智简通信技术与标准推进委员会资源及模型管理组副组长；主要研究方向为算力网络、人工智能；获省部级奖励 1 项；发表论文 30 余篇，申请国家发明专利 24 项。



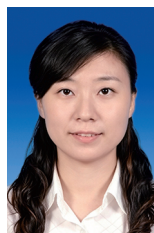
黄天一，中国科学院计算机网络信息中心在读硕士研究生；主要研究方向为分布式推理、算网融合。



周旭，中国科学院计算机网络信息中心研究员，担任中国通信标准化协会 TC614 标准推进委员会副主席、中国通信学会算力网络委员会委员、中国通信学会信息通信网络技术委员会委员、重庆大学通信学院兼职教授/博导等学术职务，国家重点研发计划项目首席科学家，国家级人才计划入选者；主要研究方向为计算机网络体系结构、人工智能、5G/6G 移动网络；近年来作为项目负责人完成重大项目 10 余项，获得部级科学技术奖一等奖 4 次、二等奖 2 次；发表论文 80 余篇，申请专利 60 余项（其中已授权 25 项），撰写专著 1 本，主持/参与制定国际标准/行业标准 15 项。



李智，中国移动通信有限公司网络事业部高级技术专家，高级工程师；主要研究方向为互联网技术、CDN、算力网络、边缘计算等；获得北京市科技进步奖 1 项，中国通信学会科学技术奖 1 项，工业和信息化部中国信息通信研究院科技进步奖 9 项，互联网、CDN、云、人工智能产业协会等科技进步奖 4 项，中国移动科技进步奖 2 项；发表论文 18 篇，拥有专利 36 项，参与研发标准 24 项。



吴茜，中国移动通信有限公司网络事业部高级工程师；主要研究方向为 CDN 的云化、智能化关键技术；获 2020 年中国通信学会科学技术奖三等奖、2023 年工业和信息化部中国信息通信研究院“华彩杯”算力应用创新大赛北区总决赛一等奖、2024 年全国信息通信领域安全生产和网络运行安全成果一等奖等。

语义通信技术商用化之路



Road to Commercialization of Semantic Communication Technology

石光明/Shi Guangming^{1,2}, 杨旻曦/Yang Minxi^{1,3},
高大化/Gao Dahua^{1,3}, 肖泳/Xiao Yong^{2,3,4}

(1. 西安电子科技大学人工智能学院, 中国 西安 710126;

2. 鹏城实验室, 中国 深圳 518055;

3. 琶洲实验室(黄埔), 中国 广州 510555;

4. 华中科技大学, 中国 武汉 430074)

(1. School of Artificial Intelligence, Xidian University, Xi'an 710126, China;

2. Pengcheng Laboratory, Shenzhen 518055, China;

3. Pazhou Laboratory, Huangpu, Guangzhou 510555, China;

4. Huazhong University of Science and Technology, Wuhan 430074, China)

DOI: 10.12142/ZTETJ.202603010

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260626.1023.006>

网络出版日期: 2026-06-26

收稿日期: 2026-05-20

摘要: 随着人工智能技术的快速发展, 传统以比特传输为核心的通信范式正面临根本性变革。系统梳理了语义信息处理与通信从理论起源到技术演进的发展脉络, 阐明其本质内涵与核心优势, 分别从通信与智能两个维度, 剖析其商用化面临的关键挑战并研判发展前景。语义通信的商用化仍受制于理论体系不完善、全链路技术成熟度不足、标准化进程滞后三重瓶颈; 而智能化的持续推进, 特别是智能体技术的发展, 将有力推动语义通信网的落地部署。在工业物联网、应急通信、深海水声通信等具有明确语义任务和场景约束的垂直行业中, 语义通信技术有望在3~5年内实现局部试商用; 伴随6G技术和标准化的演进, 语义编解码和语义认知网络架构等有望在5~10年纳入全球6G标准框架; 而实现全网络语义内生智能化的大规模部署, 则预计需要10~20年。本研究可为学术界与产业界开展语义通信研究及应用部署提供参考。

关键词: 语义通信; 商用化; 语义通信网; 通信体系架构; 智能体通信; 下一代移动通信

Abstract: With the rapid development of artificial intelligence technology, the traditional communication paradigm centered on bit transmission is undergoing fundamental changes. This paper systematically reviews the development route of semantic information processing and communication from theoretical origins to technological evolution, and clarifies its essential connotation and core advantages. From the two dimensions of communication and intelligence, it analyzes the key challenges faced by its commercialization and forecasts its industrial prospects. The commercialization of semantic communication is still restricted by three major bottlenecks: imperfect theoretical system, insufficient maturity of full-link technologies, and lagging standardization process. The continuous advancement of intelligence, especially the development of agent technology, will effectively promote the practical deployment of semantic communication networks. In vertical industries with clear semantic tasks and scenario constraints, such as industrial Internet of Things, emergency communications, and deep-sea underwater acoustic communications, semantic communication technology is expected to achieve partial pilot commercialization within 3–5 years. With the evolution of 6G technologies and standardization, semantic encoding and decoding as well as semantic cognitive network architectures are expected to be incorporated into the global 6G standard framework in 5–10 years. In contrast, the large-scale deployment of full-network endogenous semantic intelligence is estimated to take 10–20 years. This work can provide references for academic and industrial communities to carry out research and application deployment of semantic communication.

Keywords: semantic communication; commercialization; semantic communication network; communication system architecture; agent communication; next-generation mobile communication

引用格式: 石光明, 杨旻曦, 高大化, 等. 语义通信技术商用化之路 [J]. 中兴通讯技术, 2026, 32(3): 51–59. DOI: 10.12142/ZTETJ.202603010

Citation: Shi G M, Yang M X, Gao D H, et al. Road to commercialization of semantic communication technology [J]. ZTE technology journal, 2026, 32(3): 51–59. DOI: 10.12142/ZTETJ.202603010

回顾人类文明的技术演进脉络, 工业革命解放了人类的体力, 互联网构建了全球化的信息通道, 而人工智能正在规模化地生产智力。当今智能时代, 信息技术的服务对象正从人本身扩展到智能机器。机器不仅是工具, 更成为智力的生产者和消费者。在这一背景下, 大约10年前, 少数先驱学者开始倡导语义通信技术, 主张通信系统应从传输比

特转向传递语义。经过近年来的快速发展, 语义通信已成为通信与人工智能交叉领域公认的研究热点^[1-2]。当前不论是高校、研究所, 还是华为、中兴通讯等设备厂商, 以及中国移动、中国电信、中国联通等通信运营商, 都在投入资源从事语义通信技术的研究。国际上, 著名电信企业和研究机构也在密集跟进, 例如: Nokia Bell Labs 已联合 KT 启动 6G 语

义通信研发^[3]；Samsung Research在2025年发布的6G白皮书中，将语义/任务导向通信列为关键方向^[4]；NTT DOCOMO在6G架构中纳入流式语义通信^[5]。不同于大数据、区块链、虚拟现实/增强现实（VR/AR）等先在国际形成热点再被中国追赶的技术轨迹，语义通信走出了一条反向路径：先在中国通信学术圈形成热点，再由国际学者跟进研究，最后企业参与研发落地。

一项信息技术从萌芽到商用，需要技术成熟度、市场迫切性、生态完整性与政策支持度多维度共同支撑，周期通常为10~20年。语义通信从萌芽到成为热点已近10年。那么语义通信何时以何种方式进入商用阶段？许多信息技术企业正在关注此事并作出各自预测，本文就此问题给出系统分析与判断。

1 通信目的与通信技术目标

约10万年前，人类最早靠手势、眼神、简单音节传递信息，逐渐形成口语，实现面对面的高效即时沟通；约3万年前，洞穴壁画、岩画、结绳记事用于记录事件、传递简单信息，形成最早的“符号编码”；公元前，中国烽火狼烟传递军情、非洲击鼓传信、希腊火炬接力实现远距离预警；1844年莫尔斯电报机与莫尔斯电码首次用电信号远距离即时传输文字，华盛顿至巴尔的摩电报线路的开通标志着电气通信时代开启。至此，用电子信号通信成为人类社会不可或缺的要素。无论是远古的洞穴壁画、结绳记事、烽火狼烟、击鼓传信，还是当今的电磁波信号，通信的目的始终是给接收方传递信息、传授知识、下达指令或布置任务。通信的本质是要“达意”^[1]——接收者要能收到发送者的本意，这个本意（信息、知识、指令、任务）被人类用各种模态的信号承载并能跨越时（隔时）空（远距）传播，人脑收到这些信号是要对其处理用于智能决策。当电子技术出现后，所有模态的信号都转变成了数字信号，数字通信成为当下隔空跨时的主流手段，占比达99%以上，几乎成为唯一通信模式。传统通信技术目标就定位于无损快速地把信源数据传递给接收方，使得接收方能够复现发送方的消息信号，也就是信号搬运；至于数据中承载的信息、知识、指令、任务等内容，并未成为传统通信技术的关注点。传统电子通信技术对比特保真的投入重于对内容保真的投入，因为它不知道信号内容在哪，只能坚守“信号保真保证内容保真”的信条。这导致传统网络在智能化时，如AI赋能的网络流量预测、切片优化等，只能基于外挂式的人工智能模型实现“黑盒”数据分析。而语义通信实现了网络数据流的语义化，使网络节点“理解”流量意图与

任务需求，最后可实现真正的内生智能调度与网络自愈。

2 语义通信的起源与发展

随着信息技术和人工智能的发展，移动机器人即将广泛融入人类社会的方方面面，人类社会活动需要传输的信息量越来越大。传统以频带和能耗资源换取通信容量性能的模式受到了挑战。特别是智能体之间的通信，需要一种低代价、高性能、直达意的通信模式。语义通信正是人工智能发展的产物^[1]。当智能终端需要理解视频内容、解读语音意图、判断图像情感时，传统的比特传输方式不得不将海量的语义信息转换为比特流进行传输，造成了严重的“语义鸿沟”——发送端编码的比特序列与接收端理解的语义内容之间存在巨大落差。

早在1988年，北京邮电大学钟义信教授在《信息科学原理》中系统提出“全信息”理论框架^[6]，2017年在*China Communications*上进一步发表语义信息理论^[7]，2018年在《智能系统学报》中构建“信息—知识—智能”转换机制^[8]，关联语义通信与通用人工智能。可以说钟义信先生是中国语义信息、语义通信研究的先驱。同期，北京邮电大学张平院士也在探讨语义通信的可行性，及其相较于传统通信的变革性。2017年12月，西安电子科技大学石光明教授在第S39次香山会议上首次讨论类脑语义信息技术，并于2018年发表文章指出“语义通信是智能时代的产物”^[11]。国际学者在2019—2020年也对语义通信进行了探索^[9-12]。在2020年之前，语义通信或语义信息并不为众多学者所熟知，也不完全被认可，因为很多基本概念尚未刻画清楚和准确。需要特别指出的是，张平院士在当时各类学术会议与报告中多次阐述语义通信的优势和前景，并首次提出“智简通信系统”^[13]。在他的倡导下，语义通信受到了众多学者的关注。据统计，语义通信领域的论文发表数量在2021—2026年间呈快速增长态势。语义通信已成为通信和人工智能领域的研究热点^[14-15]。此外，陆建华院士、张文军院士团队也针对语义通信开展了大量研究工作，取得了丰硕成果。截至目前，从基础理论（语义的数学刻画与度量^[16]、信号内容的表达^[17]），到核心技术（信源语义编解码方法^[2,18]、信源信道联合语义弹性编解码^[19]、语义知识构建^[20]、语义通信系统^[21]），以及语义通信网络^[22-26]，都已经有了相应的解决方案。

3 语义通信的本质与优势

语义通信是从“传信号比特”升级为“传语义比特”的新一代达意通信范式，针对信号中代表内容的重要比特进行

传输。其本质是以知识库为先验,降低信源熵,从而用很少的比特或特征表达信号中的信息内容。早期研究语义通信的学者普遍认可其两大优势。一是频带利用率高,或者说传输信息的密度高:其本质是基于语义知识库提取信源中的语义内容,然后用另一种简洁模态表达语义,在极高的数据压缩比下保留信源中的语义信息。它突破了传统通信传输所有比特的范式,提取并传输有效信息密度更高的语义符号,可在窄带条件下实现大容量传输。也就是说,它在信源压缩方面兼具内容保真与传输高效的优点。二是抗干扰能力强:传统抗干扰信道编码通过增加冗余比特实现可靠传输,但降低了频带资源利用率;而语义通信是利用学习得到的离线先验,在线智能纠正误码,其本质是用算力换取带宽^[13]。语义通信的信源压缩能力与信道抗误码能力,除了在海洋水声、战场等特殊场景中有利外,在当前5G带宽充裕、误码率较低的通信环境中,似乎不再具有明显优势——当前持这种看法的学者很多。然而,该观点仅局限于传统通信视角,存在一定片面性。

其实从智能的角度看,语义通信的一项突出能力——对信源语义内容的直接表达——尚未得到学界广泛重视,而正是这一能力使其成为网络内生智能的基础。具体而言,传统网络智能采用包括大模型等在内的人工智能技术,会产生巨大的传输和计算开销。而语义通信在信息源端就完成了“数据到特征/知识”的转变。由于传输的是直接可供认知和决策使用的语义流,网络节点不需要使用“先重构信号再提取语义”的传统方法,而是能直接根据接收到的语义信息进行任务协同与决策。这种“通信即计算”的新范式,大幅降低了整个智能系统的端到端时延与算力资源消耗。所以说,语义通信是智能时代的产物^[1]。

4 语义通信理论与技术成熟度

4.1 语义通信理论成熟度

语义数学刻画与语义通信的理论研究已从零散探索迈入体系化、可证明、可工程化阶段,核心进展集中在同义性语义信息论、语义数学公理/拓扑刻画、语义通信架构与编码定理、多模态与量子语义拓展四大方向,中国团队(北京邮电大学、西安电子科技大学、华为、中兴通讯等)处于全球引领地位。北京邮电大学张平院士团队构建语义熵、语义互信息、语义信道容量度量体系,该体系兼容香农理论;同时提出ComAI系统架构,支撑6G“原生智简”设计,并证明语义无损信源编码、语义信道编码、语义率失真编码定理,奠定了语义通信的理论极限。石光明教授领导西安电子科技

大学与鹏城实验室研发团队突破了信号-语义鸿沟^[17,27],对信号内容进行了刻画,提出从信号空间到语义空间的跨层映射数学模型^[16],突破了传统信息论仅处理语法信号的局限。石光明教授与华中科技大学肖泳教授合作,推导出任务引导的语义信息容量上界。中兴通讯研发团队发表了《基于熵函数的语义信息边界研究》^[28]。华为实验室提出了编码器速率与解码器准确率的权衡、语义熵的定义与计算等语义信息论基础问题。海南大学段玉聪团队量化刻画了语义对齐偏差与多维度语义歧义问题^[29]。

4.2 语义通信技术成熟度

为系统评估语义通信的技术成熟度,本文按端到端通信链路的功能分层与近年研究脉络,将语义通信划分为5个相互衔接的技术模块,并依次展开分析。

4.2.1 语义信源编码

语义信源编码是语义通信的基础,负责将原始数据转换为语义表示,按传输类型可分为文本、语音、图像3类。

文本语义编码最为成熟。Xie等提出DeepSC框架^[2],利用自然语言处理提取文本语义表示并完成传输,接收端通过深度网络重建文本,在语义准确率相当的情况下,将传输带宽压缩至传统方案的十分之一。

语音语义编码由Weng等于2021年率先开展研究^[18]。他们设计神经网络编码器提取语音高层语义特征,接收端通过语音合成恢复语义等价内容。实验表明,3 kbit/s低比特率下,其主观质量评分(MOS)可达到传统4G语音编码(AMR-WB, 23.85 kbit/s)的同等水平。2023年,该团队实现完整闭环系统,涵盖语音识别、语义提取、重建及语音合成四大环节^[21]。

图像语义编码是新兴方向。2022年,北京邮电大学张平院士团队Dai等提出DTSCC框架^[19],将语义通信拓展至图像领域,通过非线性变换前置模块实现端到端信源-信道联合优化,显著提升结构相似性指数(SSIM)指标。2023年,该团队进一步提出语义切片模型^[30],实现语义单元的按需传输与可扩展编码。

任务导向语义通信是重要扩展方向。2022年,Xie等在JSAC发表论文^[31],首次系统研究以任务完成为目标的语义通信,通过优化编解码器,丢弃与任务无关信息,进一步压缩传输比特率。

4.2.2 语义知识库与知识增强

语义知识库是关键使能技术,发送端与接收端共享先验知识库,仅传输缺失或不确定的语义内容^[20]。2023年,

Wang等在 *IEEE Communications Letters* 发表相关研究成果^[20]，在接收端引入知识图谱，以增强语义理解，提升语义一致性。

当前知识库研究面临两大挑战：一是跨领域通用知识库缺失，现有知识库多为专用知识库；二是静态知识与动态场景信息难以融合。此外，Masked VQ-VAE方案^[32]通过学习语义码本和掩码机制，增强语义码字抗信道噪声能力，保障恶劣信道下的语义恢复质量。

4.2.3 语义信道编码与联合编码

语义信道编码负责在语义层面对抗噪声，是语义通信实用化的关键。2025年，北京邮电大学研究团队在 *IEEE Transactions on Cognitive Communications and Networking* 发表论文^[33]，初步建立语义信道容量理论框架，证明其容量可显著高于传统信道容量。

联合信源—语义—信道编码是当前语义通信领域前沿研究热点。2026年，北京邮电大学研究团队在 *IEEE Transactions on Wireless Communications* 发表Token通信联合编码调制方案^[34]，验证了联合设计的可行性。2023年，Qin等提出通用化语义通信框架^[35]，将语义通信统一到联合编码框架下。在协议层面，2025年石光明团队Wang等提出无重传语义传输协议（SITP）^[36]，实现高可靠传输。目前语义信道编码仍处于起步阶段，与成熟比特层编码差距较大。

4.2.4 语义网络与资源管理

语义网络与资源管理将语义通信拓展至多用户场景，是支撑语义通信大规模商用的必要条件。现有研究已尝试将语义相似性引入用户分组、功率控制与带宽分配^[22]，以提升网络语义吞吐率与能效。此外，多任务场景用户体验质量（QoE）度量与资源分配方法也已得到初步验证^[22]。

在隐私保护方面，2024年刊发于 *IEEE Transactions on Wireless Communications* 的相关研究证实^[37]，语义层面可实现更精细的隐私保护，发送端可选择性传输任务相关语义信息，保留无关隐私。

4.2.5 生成式语义通信

生成式语义通信是新一代语义通信范式^[38-41]，发送端传递高度压缩的语义意图，接收端借助生成模型“再创造”语义等价内容。2024年，石光明团队在ICASSP上提出SG2SC框架^[38]，验证了生成式方法在图像语义通信中的优越性。2025年，石光明教授带领西安电子科技大学和鹏城实验室团队实现公里级水声图像视频通信^[40]。崔曙光团队在 *Engineering* 期刊上综述生成式语义通信（GSC）三大技术支柱，

搭建了对应的范式分类体系^[39]。

GSC对语义通信商用化影响深刻，主要表现在3个方面：一是提升接收端算力需求，约束终端部署；二是放弃逐比特精确性，使语义保真度的定义与验证成为新问题；三是在带宽极度受限的场景（如深空、应急通信）中具有“以算力换带宽”的独特优势。

4.3 语义通信系统成熟度

在频带资源稀缺的通信领域，如水下声通信、短波通信、星地通信，已有语义通信成功传输视频、图像和语音的报道。张平院士在多个学术报告中展示了相关研究进展，他的研究团队成功研制并展示了星地语义通信系统。在海洋水声通信领域，石光明教授团队研制的语义通信系统，实现了超百米级远距离多媒体信源语义无线传输。该语义通信系统所需要的算力不足200 TOPS（每秒万亿次运算）。在此之前，国际上未见有类似成果报道。

5 语义通信商用化挑战

尽管语义通信已在多类受限通信场景验证了技术可行性，但面向商业化落地仍存在诸多挑战。目前语义通信技术的推进多数集中在信源处理，在通信网络协议、信道编解码、发射波形等方面，语义赋能的相关研究成果仍较为匮乏。5G/6G的带宽很大，完全能够支撑目前的应用场景，加上语义通信与现有的网络协议的兼容性还没有完全建立，导致设备制造商和通信运营商对语义通信的落地迭代需求缺乏迫切性。这是当前产业推进的现实瓶颈。因此，语义通信从学术研究走向商用部署仍有很长的路要走。

5.1 兼容性挑战：与现有系统的融合

如果单从通信角度看语义通信，语义通信要实现商用，首先需要解决的是与现有通信系统的兼容问题。现实情况是：从成本和运营效益角度分析，无论是设备制造商还是通信运营商，都难以在短期内为语义通信研发专用设备、搭建全新通信网络。因此，语义通信必须能够与现有的4G/5G网络基础设施共存，甚至在某些场景下需要与传统的比特传输模式切换使用。然而，现有语义通信研究几乎未涉及与现有系统的融合方案。本文调研发现，截至目前，尚无文献系统研究语义层与比特层的共存机制、切换策略与协同工作模式。这一研究空白已成为语义通信商用化的重大障碍^[13]。

5.2 全链路语义化：最大的技术瓶颈

即使解决了兼容性问题，语义通信要实现端到端的语义传输，仍需解决全链路各环节的技术成熟度问题。

语义信源编码是当前成熟度最高的环节。文本^[2]、语音^[18,21]、图像语义编码^[19,30]均有较为完整的解决方案。但这些方案大多在理想实验室条件下完成验证,缺乏在真实复杂信道条件下的测试。

语义信道编码在2025年才刚刚起步^[33],目前尚未形成成熟的工程实现体系。传统的信道编码经过数十年的研究已经非常成熟(Turbo码、Polar码等均已标准化),而语义信道编码的编解码算法设计、性能分析、标准制定等均处于探索阶段。其中,石光明教授团队提出的SITP协议^[36]通过利用语义鲁棒性实现无重传高可靠传输,是协议层首个工程化尝试。

语义调制的相关研究目前几乎是空白^[34]。调制作为物理层的核心功能,负责将编码后的比特或符号映射到波形。语义调制需要将语义表示映射到电磁波形,而如何设计语义调制的星座图、检测算法、抗干扰性能分析,目前均缺乏系统性的研究。

语义网络协议研究仅有极少数初步探索工作^[22]。现有的网络协议栈(网络层、传输层等)均是为比特传输设计的,如果要支持语义网络,就需要在协议层面进行革命性的设计。这不仅是技术问题,还涉及标准制定、产业生态等非技术因素。

在基础支撑层面,语义知识库建设工作已取得进展。2026年5月16日,在中国通信学会现代语义通信专委会上,鹏城实验室联合合作团队发布了《语义知识库白皮书2.0》^[42]。这为语义通信的落地应用提供了重要支撑。

5.3 生成式语义通信的额外商用障碍

GSC在带来新能力的同时,也引入了传统语义通信尚未面临的商用障碍^[39]。

幻觉与可追溯性是GSC商用化面临的首要问题^[39]。生成模型可能在接收端“创造”出发送端并不存在的内容——这在扩展现实(XR)/元宇宙等娱乐场景中幻觉尚可接受,但在工业控制、医疗监测等安全关键场景中,任何非预期的内容生成均构成严重的系统可靠性隐患。现阶段,如何建立GSC的输出追责机制,目前尚无成熟方案。

推理时延与算力约束制约着GSC的实时性^[39]。扩散模型的去噪推理通常需要数十至数百步迭代运算,在当前终端算力水平下难以满足实时通信要求;即使在基站侧进行辅助推理,也面临网络时延叠加的问题。这意味着现阶段GSC更适合非实时或对时延不敏感的场景,如视频点播缓存、场景预渲染等。

感知一致性度量的缺失导致GSC质量评价体系不完

善^[39]。传统通信的峰值信噪比(PSNR)/SSIM度量在GSC场景中失效——高PSNR的生成图像可能存在语义错误,低PSNR的生成图像反而语义准确。目前学界尚未形成被广泛接受的GSC语义保真度量标准,这直接制约了GSC系统的公正评测和标准化推进工作。

内容安全、监管合规与隐私保护是GSC产业化不可避免的非技术壁垒^[37,39]。生成模型固有的内容生成能力可能被滥用于虚假信息传播,各国监管机构对AI生成内容的监管框架正在快速演进,GSC系统的商用合规路径尚不明朗。

6 语义通信商用化的动力

当前设备商和运营商对单纯降低带宽的“紧迫感不强”,但对“网络智能化转型和新业务催生”具有较高的内生动力。随着千亿级异构智能体在网络中的大规模部署(大模型具身智能、无人集群),网络发展的核心痛点不再是速率不足,而是异构智能体之间“无法理解与无缝协作”,即语义鸿沟。语义通信作为“智能体网络网络层协议”的潜在颠覆者,其核心商用驱动力将不再是“省带宽”,而是“赋能网络具备连接、协同高阶智能的能力”。

目前加速语义通信商用的有利因素有3点:一是智能体的出现,催生了智能体集群之间的语义通信需求;二是发展海洋经济、开发深海的国家政策已经颁布,资源有限条件下的多媒体通信业务需求旺盛;三是国际电信组织正在推动语义通信相关国际标准的规划制定工作。

6.1 智能体集群之间的语义通信需求增加

智能体是AI技术发展的一个里程碑,它实现了智能从感知、认知、决策到执行的闭环。从单一智能体作业走向群体智能体协同作业是必然趋势。因此,智能体集群之间的语义通信需求快速攀升。智能体之间的协作交互重点是达意,通信的核心随之由“比特传输”转向“意义传递”,以突破传统句法通信在效率、理解与协作上的瓶颈。

一方面,全球智能体数量正迈向千亿级,通信主体从“人人互联”转向“智能体互联”。异构智能体(大模型Agent、工业机器人、无人机、车载单元等)跨厂商、跨平台、跨领域集群化部署,形成开放动态的智能体语义通信网络。智能体本身也从单一功能执行器进化为具备感知、推理、决策、自主学习能力的认知实体。此时交互目的不再是数据复制,而是意图理解、任务协同、知识共享。

另一方面,自动驾驶、智慧工厂、空天地一体化应急救援、数字孪生城市等场景,需要多智能体分布式协作(如无人机集群搜救、多车协同编队等)。这类业务的任务高度耦

合、动态变化,要求智能体实时对齐意图、共享上下文、同步决策逻辑。专业智能体(如医疗诊断、金融风控、工业控制)跨领域协作时,用户指令高度抽象,传统结构化数据交互难以传递抽象语义与隐性知识,必须依赖语义通信实现“共同理解”。

随着智能体全面融入各行各业,低成本跨主体信息交互成为刚需,将持续推动语义通信技术落地商用。为了保障信息安全和提高通信效率,语义通信网必然会向着“多传知识、少传数据”的方向发展^[43]。当单个智能体无力完成某项任务时,无须向协作方传输全量业务数据,转而由能力适配的智能体输出知识与技能,协助其完成本体能力迭代升级、独立完成任务。这样不仅减少了传输的数据量,而且还能有效保护数据隐私。

6.2 资源有限下的大容量通信需求正在增加

随着卫星直连的快速普及,其场景特性与传统通信的局限使得语义通信的需求持续激增。语义通信已成为卫星直连规模化、高质量发展的核心支撑,具体表现在3个方面:

一是卫星直连信道受限,传统通信难以适配。卫星直连面临链路时延长、多普勒频移明显、带宽稀缺且频谱资源紧张等问题。传统比特传输模式需全量传输数据,不仅容易造成资源浪费,在低信噪比环境下还易出现信号衰减、信息丢失,无法满足高效通信需求。

二是语义通信可破解卫星直连核心痛点。语义通信以“意义传递”为核心,可提取各类终端传输数据的核心语义,压缩冗余信息,大幅降低带宽消耗,同时通过语义保真技术确保关键信息准确传递,适配卫星直连的资源约束场景,提升传输效率与可靠性^[44]。

三是卫星直连场景拓展放大需求。随着卫星直连向应急救援、远洋航运、偏远地区通信等场景延伸,叠加低轨卫星星座规模化部署、多终端协同产生海量数据,语义通信可结合天上计算实现数据智能处理,同时打破异构终端互操作壁垒,支撑天地融合协同。

与此同时,中国实施深海经济开发战略,进一步催生和

放大语义通信刚需。深海开发依赖水下机器人、潜航器等智能体集群,声波作为主要通信载体,存在传播慢、衰减强、噪声高、多径干扰等问题,传统比特传输难以应对低带宽、高干扰场景,易出现指令误解、数据丢失,直接制约深海开发进程。语义通信可提取勘探数据、作业指令的核心语义,在有限带宽下实现高效传输,既能支撑精准勘探,也能保障水下智能体与水面平台的指令对齐,避免协同作业出现失误。随着深海经济向深远海推进,“十五五”时期对深海探测、开发能力的要求将进一步提升。语义通信结合边缘计算可破解传输低效难题,其抗衰落、高容错能力也能适配深海极端场景,未来需求将持续攀升。2025年11月,石光明领导的鹏城实验室与西安电子科技大学联合团队已经实现了水平和垂直千米级的水声多媒体信源通信^[40]。

6.3 标准化正在布局

标准化是技术商用的前提条件。国际标准化组织正在积极布局语义通信标准研制^[45]。中国学者和企业陆续在《中兴通讯技术》等国家级期刊发表面向6G标准化路径的语义通信架构与场景研究^[44,46]。2026年3月,鹏城实验室与北京邮电大学提交的技术提案正式写入3GPP TR 22.870规范^[47],多项6G语义通信场景用例提案在3GPP SA1福冈会议上获得表决通过,为后续语义通信关键技术的标准化推进打下了基础。此外,中国移动等通信运营商也在积极推动面向语义视频通话业务的端到端架构标准化工作。

7 商用前景与路径判断

尽管现阶段语义通信的发展面临诸多挑战,本文认为语义通信的商用前景仍然值得期待^[13,44]。本节将从商用条件成熟度、商用路径预测、关键使能因素3个方面展开具体分析。

7.1 商用条件的成熟度分析

如表1所示,评估语义通信的商用条件,需要从服务对象、效能收益和投资保护3个维度进行分析,针对不同应用

表1 语义通信典型应用场景的商用前景分析

应用场景	预期收益	算力需求	主要阻塞因素	最早落地窗口/年
工业物联网	带宽节省80%	低-中	知识库构建不充分	2026—2028
应急通信	极低带宽可用	中	鲁棒性不足	2026—2028
车联网	时延+带宽双降	中-高	端侧算力不足、标准缺失	2028—2030
扩展现实/元宇宙	沉浸感提升	高	端侧算力不足、GSC成熟度低	2029—2032
卫星/深空通信	高压缩比	中	信道建模不完善、验证不充分	2027—2030

GSC: 生成式语义通信

场景作出差异化判断。

从服务对象来看,智能终端的普及为语义通信提供了广阔的应用土壤。2026年,智能手机、智能音箱、智能汽车等终端已经具备强大的边缘AI能力,能够支持语义编码和解码的本地运行^[13]。然而,语义通信并非适用所有场景,对于实时性要求高、带宽资源丰富、终端能力充足的场景(如家庭Wi-Fi下的高清视频传输),传统比特传输仍然足够高效。语义通信更适合于带宽资源受限、终端能力有限的场景,如偏远地区的物联网数据传输、应急通信等。

从效能收益来看^[2,18,31],目前特定场景下的性能增益已经得到验证,但在更多场景下的泛化性能仍有待验证^[48]。本文认为,语义通信的性能优势具有场景依赖性:在文本摘要、语音情感识别等语义密集型任务中优势明显^[18],在高速率音频传输等语义稀疏型任务中优势并不明显。

从投资保护来看,当前运营商已经大规模建设了4G/5G网络基础设施,不可能在短期内对现有网络进行革命性替换。语义通信的商用需要考虑对现有投资的保护,设计渐进式的演进路径^[13]。

7.2 商用路径预测

基于上述分析,本文对语义通信的商用路径作出如下分阶段预测:

1) 近期(3~5年):特定垂直场景局部商用落地。在工业物联网(IIoT)、远程医疗、应急通信和深海无线通信等垂直场景中,语义通信有望实现局部部署^[44]。在这类场景中,带宽资源有限,传输需求以语义理解而非比特精确为目标,因此语义通信具有突出的应用优势^[21,40]。智能体之间的语义通信也将率先在这一阶段实现商用落地。然而,短期内语义通信不会大规模替代现有通信系统,仅作为传统通信技术的补充方案存在。

2) 中期(5~10年):6G标准框架嵌入语义通信特性。6G标准化工作将进入实质阶段^[47]。语义通信有望作为6G候选技术之一被纳入规范级标准化讨论^[33-34],但更可能是以部分特性(如任务导向的语义编码^[31,37])的形式嵌入6G标准,而非完整的全语义化网络框架。标准化工作将推动技术收敛,助力语义通信产业生态初步成型。

3) 远期(10~20年):全网范围规模化商用部署。随着核心技术持续成熟^[33-34]、标准体系逐步完善以及AI算力的大幅提升,语义通信有望实现更大规模的商用部署。然而,即便在这一阶段,语义通信也不太可能完全取代传统比特传输,该阶段更可能形成语义传输与比特传输并存的混合传输新格局^[1,35]。

7.3 关键使能因素

1) 大模型与多模态AI技术成熟^[13,43]。以GPT-4、Claude等为代表的大语言模型,以及多模态模型的快速发展,为语义理解与生成提供了强有力的技术支撑。随着AI模型能力的进一步提升^[14],语义编码器和解码器的性能有望持续改善。

2) 标准化体系稳步推进^[13,47]。第3代合作伙伴计划(3GPP)、国际电信联盟(ITU)等标准化组织逐步将语义通信纳入标准框架,旨在制定统一的技术规范和接口标准^[33]。标准化将使得不同厂商的设备能够互联互通,降低部署成本,促进产业生态形成。

3) 芯片算力持续提升。语义编码和解码涉及复杂的神经网络运算,对终端和基站的算力有较高要求。随着神经网络处理单元(NPU)、张量处理单元(TPU)等专用AI芯片的持续迭代,边缘设备的算力将不断提升^[31],语义通信的端侧部署将更具可行性。

4) 新型信道编码取得突破^[33-34]。语义信道编码是当前语义通信领域最薄弱的技术环节,需要在理论研究和工程应用两个方面取得突破。未来需要研究设计出复杂度可实现、性能可分析、工程可部署的语义信道编码方案^[36]。

综合来看,语义通信的商用化还需要一些时间积累。本文对语义通信在5~10年内进入商用阶段持乐观态度。从长期发展来看,语义通信具有重要战略价值,值得学术界和产业界持续深耕探索。

8 结论与展望

本文系统探讨了语义通信的商用化前景,并围绕三大核心问题展开分析。一是语义通信能否商用?答案是肯定的,但需循序渐进。语义通信在特定场景下已展现显著优势,可实现5~10倍带宽压缩,理论上具备重塑通信范式的潜力。然而,受技术成熟度、标准化进度等因素制约,其商用化将是渐进迭代过程,无法在短期内实现对现有体系的全面替代。二是何时能够商用?本文提出分阶段判断:近期(3~5年)将在工业物联网等垂直领域实现局部商用;中期(5~10年)有望融入6G标准体系;远期(10~20年)可能实现大规模部署。该判断基于当前技术发展与标准化节奏得出,后续可随关键技术突破与产业政策迭代进一步优化调整。三是商用价值何在?语义通信将推动通信从“比特传输”向“意义传递”跃迁,为带宽受限、语义密集、人机协作等场景带来革命性价值。

本文研究仍存在一定局限性:相关文献调研覆盖范围仍有待完善,对语义通信商用前景的阶段性的研判存在一定主观

倾向性,对产业政策、产业链协同、商业化运营机制等非技术关键影响因素的分析不够系统。为此,后续研究将重点关注语义通信与现有系统融合、通用语义知识库构建、语义信道编码、语义网络协议栈以及语义保真度量等,为语义通信的规模化商用落地提供理论支撑与技术保障。

参考文献

- [1] 石光明,李莹玉,谢雪梅.语义通讯:智能时代的产物[J].模式识别与人工智能,2018,31(1):91-99. DOI: 10.16451/j.cnki.issn1003-6059.201801008
- [2] Xie H Q, Qin Z J, Li G Y, et al. Deep learning enabled semantic communication systems [J]. IEEE transactions on signal processing, 2021, 69: 2663-2675. DOI: 10.1109/TSP.2021.3071210
- [3] Nokia Bell Labs. Nokia Bell Labs strengthens semantic communications research with KT collaboration [EB/OL]. (2025-03-03) [2026-04-15]. <https://www.nokia.com/blog/nokia-bell-labs-strengthens-semantic-communications-research-with-kt-collaboration/>
- [4] Samsung Research. 6G white paper: AI-native and sustainable communication [EB/OL]. [2026-04-15]. https://cdn.codeground.org/nsr/downloads/researchareas/Samsung_6G_Vision_2025.pdf
- [5] NTT DOCOMO. 5G evolution and 6G white paper v5.0 [EB/OL]. [2026-04-15]. https://www.docomo.ne.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v5.0.pdf
- [6] 钟义信.信息科学原理[M].福州:福建人民出版社,1988
- [7] Zhong Y X. A theory of semantic information [J]. China communications, 2017, 14(1): 1-17. DOI: 10.1109/cc.2017.7839754
- [8] 钟义信.机制主义人工智能理论:一种通用的人工智能理论[J].智能系统学报,2018,13(1):2-18. DOI: 10.11992/tis.201711032
- [9] O'Shea T, Hoydis J. An introduction to deep learning for the physical layer [J]. IEEE transactions on cognitive communications and networking, 2017, 3(4): 563-575. DOI: 10.1109/TCCN.2017.2758370
- [10] Farsad N, Rao M, Goldsmith A. Deep learning for joint source-channel coding of text [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 2326-2330. DOI: 10.1109/icassp.2018.8461983
- [11] Bourtsoulatz E, Burth Kurka D, Gündüz D. Deep joint source-channel coding for wireless image transmission [J]. IEEE transactions on cognitive communications and networking, 2019, 5(3): 567-579. DOI: 10.1109/TCCN.2019.2919300
- [12] Güler B, Yener A, Swami A. The semantic communication game [J]. IEEE transactions on cognitive communications and networking, 2018, 4(4): 787-802. DOI: 10.1109/TCCN.2018.2872596
- [13] Zhang P, Xu W J, Gao H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks [J]. Engineering, 2022, 8: 60-73. DOI: 10.1016/j.eng.2021.11.003
- [14] Yang W T, Du H Y, Liew Z Q, et al. Semantic communications for future Internet: fundamentals, applications, and challenges [J]. IEEE communications surveys & tutorials, 2023, 25(1): 213-250. DOI: 10.1109/COMST.2022.3223224
- [15] Luo X W, Chen H H, Guo Q. Semantic communications: overview, open issues, and future research directions [J]. IEEE wireless communications, 2022, 29(1): 210-219. DOI: 10.1109/MWC.101.2100269
- [16] 石光明,高大化.语义信息及其信号内容的数学刻画方法[J].中国科学:信息科学,2025,55(8):1950-1970. DOI: 10.1360/SSI-2025-0032
- [17] 石光明,高大化,杨雯曦,等.信号的语义刻画与度量[J].电子学报,2022,50(9):2068-2078. DOI: 10.12263/DZXB.20210238
- [18] Weng Z Z, Qin Z J. Semantic communication systems for speech transmission [J]. IEEE journal on selected areas in communications, 2021, 39(8): 2434-2444. DOI: 10.1109/JSAC.2021.3087240
- [19] Dai J C, Wang S X, Tan K L, et al. Nonlinear transform source-channel coding for semantic communications [J]. IEEE journal on selected areas in communications, 2022, 40(8): 2300-2316. DOI: 10.1109/jsac.2022.3180802
- [20] Wang B Y, Li R P, Zhu J H, et al. Knowledge enhanced semantic communication receiver [J]. IEEE communications letters, 2023, 27(7): 1794-1798. DOI: 10.1109/lcomm.2023.3274562
- [21] Weng Z Z, Qin Z J, Tao X M, et al. Deep learning enabled semantic communications with speech recognition and synthesis [J]. IEEE transactions on wireless communications, 2023, 22(9): 6227-6240. DOI: 10.1109/TWC.2023.3240969
- [22] Yan L, Qin Z J, Li C F, et al. QoE-based semantic-aware resource allocation for multi-task networks [J]. IEEE transactions on wireless communications, 2024, 23(9): 11958-11971. DOI: 10.1109/TWC.2024.3386807
- [23] Shi G M, Xiao Y. An introduction to semantic communication and semantic-aware networking standardization for 6G [J]. GetMobile: mobile computing and communications, 2024, 28(3): 14-19. DOI: 10.1145/3701701.3701707
- [24] Shi G M, Xiao Y, Li Y Y, et al. From semantic communication to semantic-aware networking: model, architecture, and open problems [J]. IEEE communications magazine, 2021, 59(8): 44-50. DOI: 10.1109/MCOM.001.2001239
- [25] Xiao Y, Shi G M, Zhang P. Toward agentic AI networking in 6G: a generative foundation model-as-agent approach [J]. IEEE communications magazine, 2025, 63(9): 68-74. DOI: 10.1109/MCOM.001.2500005
- [26] Xiao Y, Sun Z J, Shi G M, et al. Imitation learning-based implicit semantic-aware communication networks: multi-layer representation and collaborative reasoning [J]. IEEE journal on selected areas in communications, 2023, 41(3): 639-658. DOI: 10.1109/JSAC.2022.3229419
- [27] 石光明,肖泳,李莹玉,等.面向万物智联的语义通信网络[J].物联网学报,2021,5(2):26-36. DOI: 10.11959/j.issn.2096-3750.2021.00209
- [28] 唐雪,许进,冯雨龙.基于熵函数的语义信息边界研究[J].中兴通讯技术,2025,31(6):61-69
- [29] 段玉聪.语义数学概论[M].西安电子科技大学出版社,2026
- [30] Dong C, Liang H T, Xu X D, et al. Semantic communication system based on semantic slice models propagation [J]. IEEE journal on selected areas in communications, 2023, 41(1): 202-213. DOI: 10.1109/jsac.2022.3221948
- [31] Xie H Q, Qin Z J, Tao X M, et al. Task-oriented multi-user semantic communications [J]. IEEE journal on selected areas in communications, 2022, 40(9): 2584-2597. DOI: 10.1109/JSAC.2022.3191326
- [32] Hu Q Y, Zhang G Y, Qin Z J, et al. Robust semantic communications with masked VQ-VAE enabled codebook [J]. IEEE transactions on wireless communications, 2023, 22(12): 8707-8722. DOI: 10.1109/TWC.2023.3265201
- [33] Ma S, Zhang C H, Qi H Y, et al. A theory for semantic channel coding with many-to-one source [J]. IEEE transactions on cognitive communications and networking, 2025, 11(6): 3618-

3627. DOI: 10.1109/tccn.2025.3544275
- [34] Ying J K, Qin Z J, Feng Y L, et al. Joint semantic-channel coding and modulation for token communications [J]. IEEE transactions on wireless communications, 2026, 25: 8179–8193. DOI: 10.1109/TWC.2025.3636174
- [35] Qin Z J, Gao F F, Lin B, et al. A generalized semantic communication system: from sources to channels [J]. IEEE wireless communications, 2023, 30(3): 18–26. DOI: 10.1109/mwc.013.2200553
- [36] Wang Y H, Ma S, Wu Y L, et al. SITP: a high-reliability semantic information transport protocol without retransmission for semantic communication [PP/OL]. arXiv (2025–12–10) [2026–04–15]. <https://arxiv.org/abs/2512.09291>
- [37] Wang Y H, Guo S S, Deng Y Q, et al. Privacy-preserving task-oriented semantic communications against model inversion attacks [J]. IEEE transactions on wireless communications, 2024, 23(8): 10150–10165. DOI: 10.1109/TWC.2024.3369170
- [38] Yang M X, Gao D H, Xie F, et al. SG2SC: a generative semantic communication framework for scene understanding-oriented image transmission [C]//Proceedings of ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 13486–13490. DOI: 10.1109/icassp48485.2024.10446594
- [39] Ren J K, Sun Y P, Du H Y, et al. Generative semantic communication: architectures, technologies, and applications [J]. Engineering, 2026, 56: 45–61. DOI: 10.1016/j.eng.2025.07.022
- [40] Shi G M. World-first kilometer-level underwater acoustic video transmission based on semantic communication [R]. 2025
- [41] Chang M K, Hsu C T, Yang G C. GenSC: generative semantic communication systems using BART-like model [J]. IEEE communications letters, 2024, 28(10): 2298–2302. DOI: 10.1109/LCOMM.2024.3450309
- [42] 鹏城实验室, 北京邮电大学, 琶洲实验室(黄埔), 等. 语义知识库白皮书 2.0 [R]. 2026
- [43] Shi G M, Lan N, Xie X M, et al. Rethinking intelligence in the era of large language models [J]. Journal of communications and information networks, 2026, 11(1): 39–50
- [44] 李东博, 王新宇, 尹志胜, 等. 星地一体化语义通信网络: 探索与展望 [J]. 中兴通讯技术, 2025, 31(5): 57–65. DOI: 10.12142/ZTETJ.202505009
- [45] Shi G M, Xiao Y. Requirements of semantic-aware networking in future networks [R]. 2023
- [46] 张黎明. 基于通信扩展定义的语义通信三层架构 [J]. 中兴通讯技术, 2024, 30(6): 92–99. DOI: 10.12142/ZTETJ.202406014
- [47] 3GPP TSG SA WG1, TR 22.870 (Release 20): Study on 6G use cases and service requirements [S]
- [48] Dai J C, Wang S X, Yang K, et al. Toward adaptive semantic communications: efficient data transmission via online learned

nonlinear transform source-channel coding [J]. IEEE journal on selected areas in communications, 2023, 41(8): 2609–2627. DOI: 10.1109/JSAC.2023.3288246

作者简介



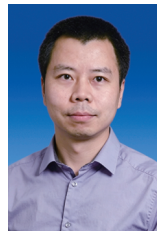
石光明, 鹏城实验室副主任, 西安电子科技大学教授、博士生导师, IEEE Fellow, IET Fellow, 中国电子学会会士, “长江学者”特聘教授, ACM中国杰出学者, IEEE CASS西安分会主席; 主要研究方向为语义通信、语义信号处理等; 获国家自然科学奖二等奖1项、省部级一等奖2项。



杨旻曦, 西安电子科技大学人工智能学院博士后; 主要研究方向为语义通信、生成式人工智能、图像视频压缩等。



高大化, 西安电子科技大学教授、博士生导师, 陕西省中青年科技创新领军人才, 陕西省青年科技新星; 主要研究方向为语义通信、智能信息处理、新体制智能成像、计算光谱成像等。



肖泳, 华中科技大学教授、IMT-2030 (6G) 网络智能组副组长、武汉市(中国-新加坡)智能与通信网络融合“一带一路”联合实验室主任; 主要研究方向为语义认知网络、智能体通信、边缘智能、通信网络博弈理论等。

大规模智算互联网络 仿真技术和平台



Large-Scale Intelligent Computing Interconnection Network Simulation Technology and Platform

熊先奎/Xiong Xiankui, 童贤慧/Tong Xianhui,
王勇/Wang Yong, 张启明/Zhang Qiming

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202603011

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260626.0957.004>

网络出版日期: 2026-06-26

收稿日期: 2026-04-25

摘要: 网络仿真作为高效的评估预测手段, 已成为智算基础设施发展不可或缺的工具。总结了大规模智算场景下网络仿真的三大核心挑战: 规模与复杂性、效率与可扩展性、精度与真实性, 并系统剖析层次化建模、混合仿真、并行仿真、硬件加速、真实数据驱动等仿真关键技术。中兴通讯探索仿真新架构, 设计实现面向数据流特征驱动的网络仿真平台, 并在工程实践中验证了其高效性和可扩展性。

关键词: 网络仿真; 仿真建模; 仿真加速; 仿真校准

Abstract: As an efficient evaluation and prediction method, network simulation has become an indispensable tool for the development of intelligent computing infrastructure. This paper summarizes three core challenges of network simulation in large-scale intelligent computing scenarios: scale and complexity, efficiency and scalability, accuracy and fidelity, and systematically analyzes key technologies, including hierarchical modeling, hybrid simulation, parallel simulation, hardware acceleration, and real-data-driven simulation. ZTE Corporation explores a new simulation architecture, designs and implements a network simulation platform driven by data flow characteristics, and verifies its efficiency and scalability in engineering practice.

Keywords: network simulation; simulation modeling; simulation acceleration; simulation calibration

引用格式: 熊先奎, 童贤慧, 王勇, 等. 大规模智算互联网络仿真技术和平台 [J]. 中兴通讯技术, 2026, 32(3): 60-68. DOI: 10.12142/ZTETJ.202603011

Citation: Xiong X K, Tong X H, Wang Y, et al. Large-scale intelligent computing interconnection network simulation technology and platform [J]. ZTE technology journal, 2026, 32(3): 60-68. DOI: 10.12142/ZTETJ.202603011

现代大语言模型 (LLM) 在 Scaling Laws 的驱动下持续突破规模边界。智算互联网络作为连接海量计算资源和存储资源的核心基础设施, 呈现出新的特征:

1) 高带宽与低延迟需求: 分布式训练要求网络具备太比特每秒级聚合带宽和亚微秒级延迟能力, 推动光互联技术实现突破^[1]。

2) 超大规模与高扩展性: 算力集群中数十万节点通过多层网络拓扑结构实现互联, 驱动规模组网架构技术创新^[2]。

3) 非对称突发流量: All-Reduce、All-to-All 等集合通信操作产生独特的流量模式, 推动传统流量调度与拥塞控制机制革新^[3]。

4) 异构资源协同耦合: 中央处理器 (CPU)、图形处理器 (GPU)、神经网络处理器 (NPU) 等异构计算单元与高速网络、存储设备深度协同, 促进统一资源管理、跨域调度优化等技术应用^[4]。

面对如此复杂系统特性, 物理部署实测的方式难以覆盖全场景、高成本、长周期的验证需求。网络仿真通过构建数字化模型复现网络行为、评估网络性能, 凭借低成本、高灵活性优势, 已成为支撑智算互联网络设计验证的关键手段。

1 网络仿真发展现状

网络仿真技术历经两个阶段发展, 从通用网络场景逐步演进至智算网络场景。

1.1 通用网络仿真

通用网络仿真聚焦传统计算机网络、通信网络的性能模拟与分析,广泛应用于网络协议研发、架构设计与优化等场景。该类仿真主要包括开源与商业两条技术路线:

1) 开源路线:以NS-3^[5]为代表,具备模块化架构与高度可扩展性,但在大规模场景下仿真效率明显不足。Unison^[6]、DONS^[7]等项目通过并行化仿真、事件调度优化等技术提升效率。

2) 商业路线:以OPNET^[8]为典型代表,具备图形化交互界面、高精度模型库以及多维度性能分析能力,但其商业化属性带来使用成本高、难以快速适配特殊场景等问题。

随着网络规模扩大与复杂度提升,通用网络仿真逐渐凸显出效率低、精度不足、兼容性差等突出问题。

1.2 智算网络仿真

智算网络仿真是面向大模型训练、推理等智算场景的定制化技术,其核心目标是精准模拟算力节点与网络的协同特性、工作负载的传输规律以及异构互联架构的性能表现。目前智算网络仿真已广泛采用“以任务为中心”的新范式,其发展经历两个阶段:

1) 基于通用工具扩展阶段,粗粒度集成智算模块,快速构建仿真能力

例如,ASTRA-sim^[9]早期1.0版本仅支持有限集合通信,忽略了计算通信等耦合效应,精度不足;SimGrid^[10]仿真引擎早期在通用仿真框架中集成了加速器网络的通信模型。

2) 专用定制化框架阶段,深度适配智算场景的算网协同需求

此类框架普遍采用计算、存储、网络协同建模。例如,ASTRA-sim在2.0版本中支持层级化网络抽象和计算调度联合仿真;SimAI^[11]面向大模型训练场景,集成了训练框架、计算、集合通信库和网络后端;Vidur^[12]则针对推理场景,构建动态批处理仿真框架;ATLAHS^[13]以应用为中心,采用通用组运算汇编语言(GOAL)格式表达AI、高性能计算(HPC)、存储三类工作负载。

目前,中国头部企业已纷纷开展定制化仿真技术研发并落地应用。例如:阿里云HPN^[3]双平面两层架构设计,借助仿真平台建模训练任务,快速迭代验证通信调度优化机制,最终达成系统利用率与可靠性的双重提升;华为CloudMatrix384^[14]超节点架构,针对Pangu Ultra MoE模型的昇腾亲和特性构建仿真模型,通过评估不同策略的性能表现,确定长序列场景的最优部署方案。

2 网络仿真面临的主要挑战

智算互联网络仿真面临一大核心难题:规模扩展、执行效率、高准确度三者难以兼顾。

2.1 规模与复杂性挑战

规模扩展导致复杂度呈指数级增长,传统建模表达能力已不再适配复杂场景。

1) 互联拓扑复杂。万卡级集群的多层级拓扑结构难以精确建模。SlipStream^[15]相关研究证实,节点间通信关系动态变化进一步加剧了建模难度。同时SplitSim^[16]实验表明,传统的集中建模方式引发状态空间爆炸,显著提升了计算复杂度。

2) 异构性与动态性突出。网络设备、链路带宽、流量模式的异构性,叠加网络拥塞、故障、负载均衡等动态行为,显著增加了仿真建模的适配难度^[8]。

3) 多维度资源耦合。智算任务中计算、存储、网络资源间存在复杂的依赖和竞争关系。每种硬件和网络技术都有其独特的性能参数和行为模式,单一模型缺乏对耦合效应的精准模拟,更难以支撑系统级优化^[17]。

2.2 效率与可扩展性挑战

仿真效率与规模扩展的非线性矛盾突出,传统串行和弱并行架构已接近算力天花板。主要表现在以下3个方面:

1) 资源消耗高。SplitSim^[16]实验表明,10万节点的网络仅拓扑建模就需占用数十GB内存,数据包模拟计算量更是达到千亿次级别。单轮仿真占用CPU资源超过100核/h,远超单节点的承载能力。

2) 仿真时间长。网络仿真的时间复杂度与节点数量的平方成正比。在Multiverse^[18]测试中,模拟一次千亿参数大模型的训练过程,仅参数同步就需要处理数百万次传输事件,整体仿真耗时可达数周。

3) 并行仿真瓶颈。网络事件之间存在复杂的依赖关系,事件同步、状态一致性维护等开销限制了并行效率提升,超大规模场景下更难以实现线性加速^[19]。并行离散事件仿真(PDES)在高动态场景下事件回滚率超50%,反而降低有效吞吐量^[6]。

2.3 精度与真实性挑战

软件层不确定性与硬件特性细节缺失,共同导致仿真结果与实际数据偏差显著。

1) 软件栈影响难以量化。操作系统、设备驱动、通信库等软件层对网络性能的实际开销和影响,在仿真中难以精

准复现^[20]。

2) 工作负载建模不足。现有静态模型难以刻画多任务并发时计算和通信耦合引发的资源竞争与调度交互行为。实验表明^[21]，在LLM训练仿真中，如果忽略通信触发的时序依赖关系，仿真误差将超过30%。

3) 动态网络行为缺失。仿真模型无法复现真实网络的瞬态响应和局部性能波动；同时基于实时流量的路由调整、基于算力节点负载感知调度等动态负载均衡自适应策略未得到精准建模，导致仿真无法反映实际环境的自适应优化能力^[9]。

4) 硬件特性建模简化。现有的模型大多忽略芯片缓存、网卡队列管理、调度算法等底层硬件交互细节。Net-TLMSim^[22]实验表明，网卡队列调度策略会显著影响流量传输的公平性与传输延迟，简化建模会导致10%~40%的性能预测误差。

3 网络仿真关键技术研究

针对前文所述挑战，学术界与产业界开展了一系列系统和硬件的改进优化工作。

3.1 规模与复杂性优化

3.1.1 分层/模块化拓扑建模

建模过程中将复杂的网络按逻辑层级或地域进行拆分：

1) 层次化建模，按逻辑关系等划分层次。例如，ASTRA-sim对核心交换层、算力节点接入层进行建模，通过层级间的接口适配构建大规模拓扑。

2) 模块化建模，按地域等划分多模块。例如，数据中心内采用细粒度建模，精准刻画拓扑端口带宽等细节；跨数据中心采用中粒度建模，重点关注骨干网络的路由特性与传输延迟；广域算力协同采用粗粒度建模，简化地理距离带来的延迟与带宽衰减^[2]。

3.1.2 多粒度混合仿真

多粒度混合仿真通过设置差异化建模精度，兼顾精度与开销，在仿真规模与运行效率之间取得平衡。该仿真的关键在于不同粒度模型之间的协同机制。通过流量映射、状态同步等方式确保模型间的平滑切换。例如，NetTLM-Sim^[22]多粒度混合策略会固定划分建模区域：对加速器网络的核心互联部分采用精细化硬件建模，对边缘网络采用抽象流量建模，通过流量状态同步实现网络高效仿真；SimAI^[11]可动态调整建模粒度，在仿真过程中根据流量类型自动切换细粒度和粗粒度模型，既保证热点区域的仿真精度，

又降低整体开销。

3.1.3 统一资源抽象

统一资源抽象技术对计算、通信、存储等多维度资源进行一体化整合。该技术的关键在于构建统一资源抽象与元数据模型。经典的单步任务执行延时量化模型^[23]能够提取各类资源的核心属性，并将不同资源映射为标准化的抽象实体，具体如公式(1)所示：

$$T_{\text{step}} = \max(T_{\text{comp}}, T_{\text{mem}}, T_{\text{comm}}) + T_{\text{unhidden}} + T_{\text{sync}} \quad (1),$$

其中， T_{step} 为执行总延时；计算时间 $T_{\text{comp}} = \frac{F}{P \times \pi}$ ， F 为单卡浮点运算量， P 为单卡峰值算力， π 为实际利用率；访存时间 $T_{\text{mem}} = \frac{B}{\beta_{\text{mem}}}$ ， B 为访存数据量， β_{mem} 为显存带宽；通信时间 $T_{\text{comm}} = (\frac{M}{\beta_{\text{net}}} + \alpha)N$ ， M 为单次通信数据量， β_{net} 为网络带宽， α 为网络延迟， N 为通信次数； T_{unhidden} 为无法重叠的时间； T_{sync} 为系统同步开销。

该公式依赖静态参数构建，存在动态适应性不强、资源耦合刻画不足等问题。为突破静态模型局限，Meta的Arcadia^[24]将不同维度的参数作为相互影响的系统组件来描述，通过协调器来同步和模拟各种组件的行为，实现状态实时同步与动态参数修正。

3.2 效率与可扩展性优化

3.2.1 基于采样的加速仿真

通过部分样本仿真结果推测整体网络性能，在保证精度损失可控的前提下，可大幅提升仿真效率。

采样策略是有效性的关键保证。合理设计采样策略（如分层采样、逐流采样、重点采样），对智算网络的流量、拓扑、任务等关键要素进行采样，确保采样数据能准确反映整体特征。Network Sampling^[25]实验证明，对智算网络中的热点链路和高并发流量进行优先采样，忽略低流量的网络区域，在采样率仅为30%的情况下，可使仿真精度保持在90%以上。

采样策略的选择取决于网络类型和考量指标。目前还没有一种普遍最优的采样方法，因此，探索特定指标和应用驱动的采样策略是重要研究方向之一^[25]。

3.2.2 高效事件调度与管理

针对智算网络仿真的事件密集型特征，可采用优先级调度、批量处理机制提升执行效率；同时通过合并冗余、筛选

过滤减少事件处理开销。

1) 优先级调度。Unison^[6]设计了分层事件调度器,将网络事件、计算事件、存储事件分别分配到独立的线程池处理,通过优先级队列确保关键事件及时响应。DONS^[7]引入依赖感知机制,通过分析通信计算图来动态调整优先级。

2) 事件合并与过滤。例如SplitSim^[16]采用混合保真度仿真方案,将多个同类事件合并为一个批量事件,减少事件处理次数。粗粒度仿真模式也会过滤部分事件,从而缩短仿真时间。

此外,DeepQueueNet^[26]提出预调度技术,可提前预测关键事件的发生时间与顺序,优化事件处理流程,进一步降低仿真时间开销。

3.2.3 并行与分布式仿真

将大规模仿真任务拆分至多个计算节点并行执行,可突破单节点的计算资源与内存限制。

该方案的核心在于合理的任务划分策略与高效的同步机制。

1) 任务划分。划分过程需保证各节点负载均衡,减少跨节点通信。例如,DONS^[7]将智算网络按照链路连接关系划分为多个独立的仿真区间,支持自动并行能力;SplitSim^[16]通过编排机制将任务分解为多个并行逻辑单元,采用动态负载均衡策略。

2) 同步机制。该机制主要包括混合同步策略、因果依赖调度等技术。例如,NSX^[27]采用混合同步策略,仿真域内部采用乐观同步,域间则采用保守同步。此外,也有研究通过自动识别“低风险异步区”放宽同步约束^[28]。

3.2.4 硬件加速仿真

硬件加速仿真技术利用专用硬件的并行计算能力加速计算密集型任务,主要包括两个部分:

1) 硬件加速架构设计。“软件控制+硬件计算”的混合架构是普遍采用的架构方式。软件负责仿真任务调度、拓扑管理、动态场景控制等逻辑密集型操作,硬件则通过加速器实现数据包转发、延迟计算、拥塞控制等计算密集型任务的并行处理^[22]。

2) 硬件加速模块实现。例如Miniature^[29]在现场可编程门阵列(FPGA)上对队列、内部状态和协议栈进行建模。实验表明,在仿真65 536个节点的智算集群时,相比基于软件的网络仿真,效率提升4 332倍。Multiverse^[18]在GPU上通过CUDA编程实现仿真任务的大规模并行计算,多实验并行仿真的效率提升40倍。

3.3 精度与真实性优化

3.3.1 真实数据驱动仿真

真实数据驱动仿真首先采集真实智算网络中的运行数据,并利用数据挖掘技术提取关键特征。

该技术的核心流程包括3个部分:

1) 数据采集。通过网络探针、设备日志、任务监控系统等手段,获取智算网络的流量特征、拓扑变化、设备性能等真实数据^[26];

2) 数据预处理。通过数据清洗、异常值剔除、特征提取、归一化等操作,将原始数据转化为符合仿真模型输入要求的结构化数据^[30];

3) 模型校准。利用预处理后的数据对仿真模型的关键参数进行校准,提升模型对真实场景的拟合能力^[4-8]。例如,SimAI^[11]通过采集数千次LLM训练的真实通信数据,将仿真精度提升至95%以上。

3.3.2 计算/通信协同建模

该技术融合计算任务特性与网络传输特性,使工作负载的任务执行顺序、调度策略选择能够真实反映算网耦合过程。

该技术的核心是要建立计算与通信的动态关联机制,具体包括两个部分:

1) 时序关联模型。将通信事件与计算任务的关键节点(如层计算完成时刻、批量处理结束时刻)绑定,确保通信触发点与真实场景中一致^[31]。

2) 反馈调节机制。将网络传输延迟、带宽占用等结果实时反馈至计算任务模型,调整计算任务的执行进度。例如,Arcadia^[24]将大模型训练中的层计算与参数同步进行动态绑定,模拟计算与通信的重叠与阻塞关系,计算过程中调整等待通信完成的空闲时间,使得仿真真实度提升25%。

3.3.3 动态场景精确建模

引入实时状态反馈、参数自适应调节等方法,有助于解决动态网络行为建模缺失问题。该技术主要分为两类实现方式:

1) 控制行为建模。精准刻画拥塞窗口调整、信用值分配、流量控制帧交互等动态过程,能够模拟网络过载时的延迟增加、数据包丢失等真实现象。例如,Unison^[6]对以太网远程直接内存访问(RoCE)优先级流控(PFC)帧在多级拓扑中的传播延迟与链路信用耗尽触发条件进行建模,复现了“反压级联”引发的非拥塞丢包机制。

2) 异常行为建模。一方面,构建随机突发流量模型,并依托事件驱动机制触发流量突发。Vidur^[12]借助该方式成

功模拟了 LLM 推理过程中的流量突发场景，仿真结果准确性可达 90%。另一方面，引入故障注入机制^[32]，支持节点故障、链路中断等多种故障类型的模拟仿真。

3.3.4 精细化硬件特性建模

通过解析关键硬件的底层工作机制和性能特征，构建贴合真实硬件行为的硬件仿真模型，精准还原硬件级别的性能差异和行为细节。

1) 网络设备建模。该模块涵盖队列调度策略、缓冲分区机制以及多端口并发处理能力^[26]。例如，ASTRA-sim^[9]实测了不同交换机的队列调度延迟，建立非线性延迟模型，精准模拟不同负载场景下的传输延迟差异。

2) 计算设备通信接口建模。该模块包括其计算能力、内存带宽、数据访问延迟，以及与网卡的直接数据交互特性，可量化硬件卸载对 CPU 开销、传输效率的影响^[16-17]。例如，NetTLMsim^[22]完整还原 RDMA 网卡卸载机制（如零拷贝、远程内存访问流程、缓存同步策略），应用于 Transformer 框架设计时，延迟预测误差降低 52.8%。

3) 存储设备建模。该模块覆盖传统存储与智算专用存储（如内存池化）。重点区分各类存储在输入/输出（I/O）延迟、带宽特性、随机/顺序访问性能上的差异，构建分布式存储系统的元数据管理延迟模型，精准反映存储访问对网络传输的影响^[31]。例如，Arcadia^[24]通过实测不同存储设备 I/O 性能，建立了存储访问模式与网络有效带宽的关联模型，能够准确模拟存储瓶颈导致的网络传输效率下降问题。

4 中兴通讯智算互联网络仿真的探索与实践

4.1 面向数据设计的高性能网络仿真平台

中兴通讯联合清华大学提出了一种数据流特征驱动（DOD）的仿真架构，可针对性解决万卡集群网络仿真的效率与可扩展性问题。

如图 1 所示，通过重构 LLM 训练任务中的计算图与通信依赖关系、优化时间调度逻辑，团队开发了首款面向 GPU 加速的专用仿真器。实测数据表明，相较于主流仿真工具 ASTRA-sim，该方案的网络仿真速率提升 20 倍以上。

4.2 算网存联合仿真平台

中兴通讯智算前瞻团队正推进计算、网络、存储联合仿真平台建设。该平台整合专用加速器、多协议网络设备、高带宽内存等异构资源，构建标准化交互模型。核心目标是支撑三大关键应用场景：算法架构适配性评估、资源需求敏感性分析、集群扩展边界预测。

为应对建模高精度挑战，本平台采用如图 2 所示的架构，采集真实业务数据开展全流程模型校准。基于样本数据集，可进一步挖掘仿真数据与真实运行数据的关联规律，构建智能预测模型，实现对网络拥塞、存储访问异常、算力资源瓶颈等潜在故障的在线预警能力。

4.3 新型互联协议验证与训练部署寻优

自研仿真平台可支撑新型互联协议的快速开发与验证。

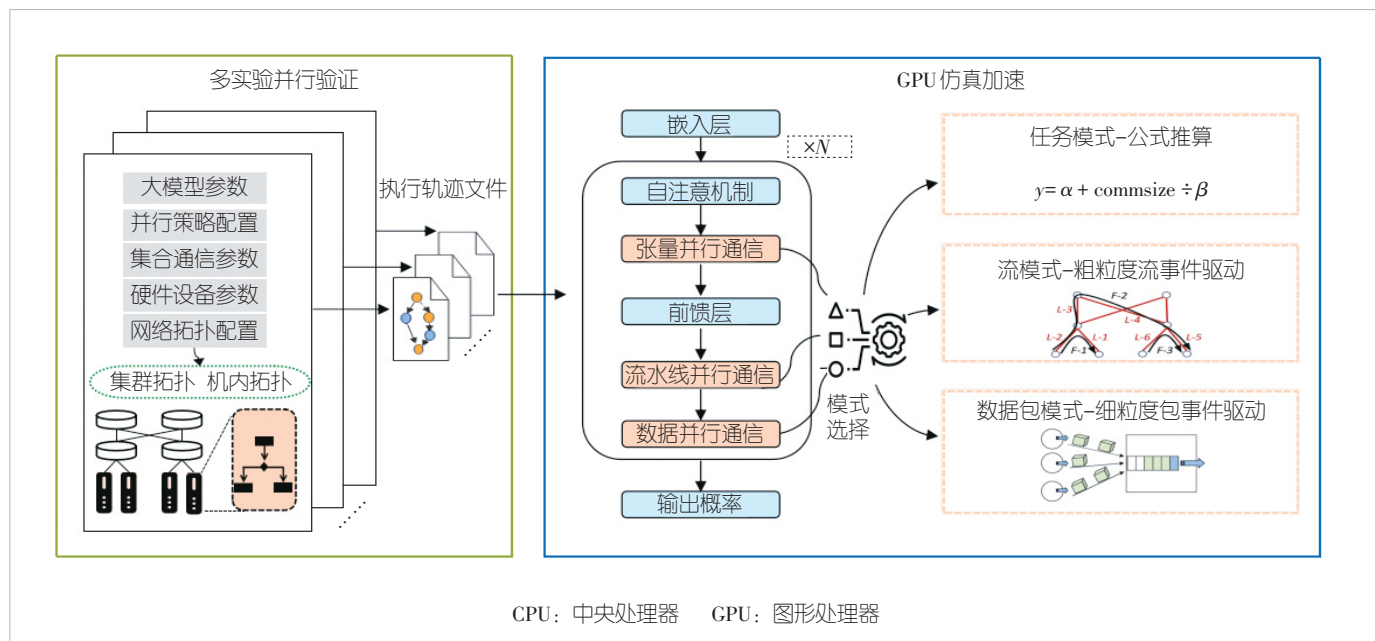


图1 高性能智算网络仿真平台

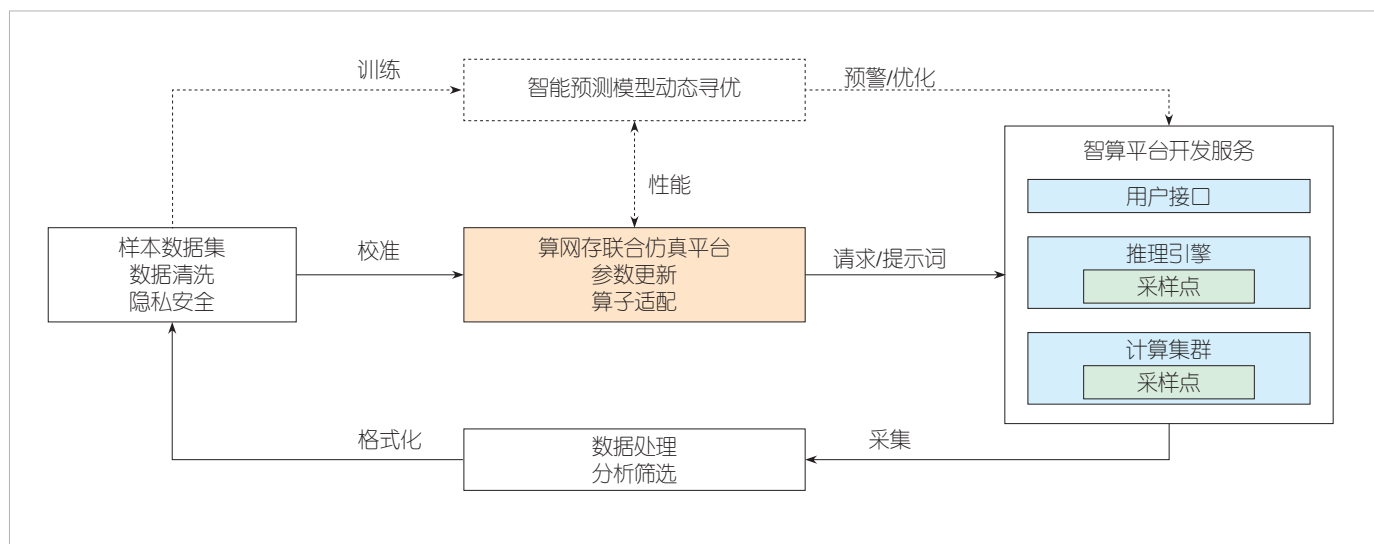


图2 大语言模型推理服务仿真校准流程

由图3灌包测试结果可知，自研逐流拥塞控制机制具备更快的拥塞响应速度、更优的理论速率逼近能力，同时收敛过程中震荡幅度显著低于传统数据中心量化拥塞通知（DCQCN）方案。

此外，该仿真平台还可支撑分布式训练部署方案寻优。如表1所示，借助仿真平台测试不同资源配置与网络拓扑，可量化评估不同并行策略对系统性能的影响，在实际部署前筛选出最优方案。未来该仿真平台将进一步拓展应用边界，覆盖更多异构分布式训练、推理平台，实现更多场景的性能量化预测与设计空间寻优。

5 趋势与展望

未来智算网络仿真将深度融入智算全栈仿真体系，需要从系统工程全局视角统筹仿真体系的架构设计与建模实现。

5.1 跨域一体化仿真

未来智算网络仿真将突破当前仿真的场景局限性，构建支持异构算力、跨域网络、多应用负载的一体化仿真平台。

要打破规模扩展性瓶颈，需要从3个方向实现技术突破：

- 1) 跨域同步机制优化。通过轻量化通信协议等方式解决多域时序一致性问题，保障全局仿真逻辑的准确性。
- 2) 模块化组件设计。通过标准化接口与组件化架构设计，统一仿真模块的交互规范，降低大规模仿真系统的维护成本。
- 3) 动态负载适配。引入资源弹性分配机制，根据仿真负载动态调度系统资源，保障仿真系统稳定性与高效性。

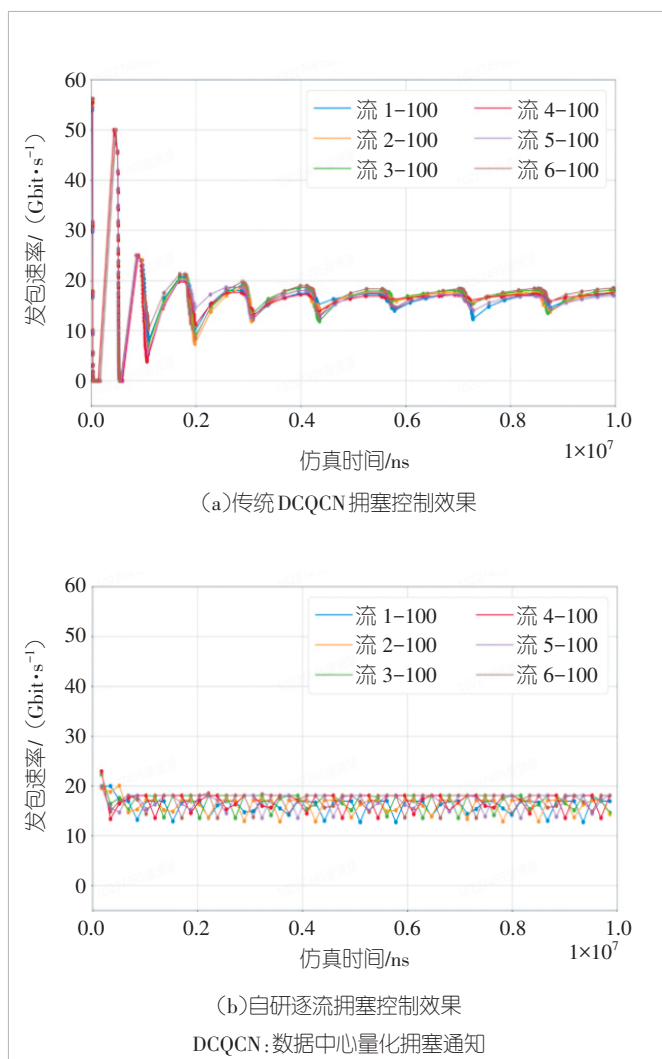


图3 传统DCQCN与自研逐流拥塞控制的灌包测试对比

表1 万卡智算集群Megatron MoE 7B模型32k长序列并行策略对比

	并行策略	计算效率/%	空泡占比/%	TP通信占比/%	EP通信占比/%	PP通信占比/%	DP通信占比/%
1	TP8PP16VPP2EP32DP96	89.38	6.24	3.06	0.90	0.39	0.03
2	TP4PP12VPP2EP64DP256	80.94	15.14	1.97	1.31	0.56	0.08
3	TP2PP16VPP4EP32DP384	73.41	22.20	1.10	2.10	0.91	0.28
4	TP2PP32VPP1EP8DP192	50.85	48.35	0.22	0.33	0.16	0.09

DP:数据并行 EP:专家并行 PP:流水线并行 TP:张量并行

ATLAHS^[13]工具链的设计理念值得参考,该框架通过构建多场景的通用仿真底座,搭配定制化插件适配不同应用的负载特征。此外,还需要解决跨域数据传输延迟精准建模和跨域安全机制建模等问题,以适配更复杂的跨域智算场景^[3]。

5.2 微宏协同全栈仿真

未来智算网络仿真将破除分层割裂的仿真壁垒,逐步构建覆盖“算力芯片微架构、网络传输、算法模型、系统部署”的全栈仿真体系,实现从底层硬件创新到上层系统优化的端到端支持。

相关技术探索已从多个维度开展:微观层面刻画算力芯片微架构的并行访问、指令时序等核心要素特征;宏观层面捕捉任务调度、通信交互等系统级的涌现特性。

微宏协同关键技术包括:

1) 模型间自适应融合,支持全链路上模型跨粒度动态切换。多层次建模和混合仿真是实现不同模型之间的动态切换和数据交换的常用方式。例如,SplitSim^[16]使用的混合保真度方法就是这一技术的早期探索。

2) 跨层参数关联,捕捉全栈各层级间动态交互关系。例如,Transformer框架^[33]梯度压缩算法对网络数据传输量的影响、网络拥塞对硬件算力利用率的影响、芯片通信延迟对算法效率的影响,均可通过该技术实现量化建模。

当前,该技术趋势已初见端倪。中兴通讯智算前瞻团队也自研了全栈仿真平台,一方面持续持续优化平台架构,提升平台效率;另一方面利用仿真平台对存算一体、内存池化、算力池化等新型硬件形态开展前瞻性技术探索。

5.3 实时数字孪生

未来仿真将从“离线设计验证”走向“在线实时优化”,构建“物理网络-数字孪生-决策优化”的闭环系统,最终成为在线运维的核心引擎。

数字孪生与仿真技术的深度融合已形成明确技术路径^[28]:通过高效数据采集与传输实现物理网络的精准映射,依托仿真引擎完成实时分析与优化决策,最终通过执行与反

馈实现闭环控制。两者融合关键在于:1) 高效数据采集与传输,保障物理环境与仿真环境高可靠的数据交互;2) 虚实协同决策机制,保障决策指令快速匹配和实时响应。

当前实时数字孪生技术发展迅速,持续向轻量化仿真引擎演进。仿真引擎不仅是分析工具,更是故障预判、系统自愈、性能优化等技术创新的基石。

5.4 智能化仿真新范式

未来仿真将更多使用大模型技术替代传统的解析建模与离散事件驱动,逐步推动“仿真即服务”的自动化演进。

大模型与仿真技术深度融合,有望实现两大核心突破:

1) 自动化建模。利用LLM的自然语言理解和生成能力,可根据描述性文本自动生成仿真模型。例如,基于系统网络架构文档即可生成对应的网络拓扑、流量模型和参数配置^[31];也可利用物理信息神经网络(PINN)等技术,直接从数据中学习物理规律并构建模型。

2) 智能化决策。依托实时数据流与预测模型,自动输出拓扑调整、资源调度等优化策略^[11],为智算服务的在线调度提供精准实时的决策支撑,从而显著提升系统运维效率。

虽然智能化仿真面临数据采集成本高、模型泛化能力不足、可解释性缺失等技术挑战,但从长远来看,智能化仍是仿真范式演进的关键技术方向。目前,m4^[31]、MimicNet^[34]等项目已初步验证了该范式的技术可行性和复杂场景下的应用潜力。

6 结束语

大规模智算网络的规模化部署与高性能需求,驱动着网络仿真技术从“辅助工具”向“基础设施”转变。同时,智算业务的复杂协同特性,正推动网络仿真深度融入计算、网络、存储全栈仿真体系。尽管当前智算网络仿真仍面临规模、精度、真实性、异构性和标准化等多重挑战,但通过分布式仿真、并行仿真、混合建模、数据驱动、AI辅助以及云原生等关键技术优化,这些痛点正在被逐一突破。未来,网络仿真将摆脱离线运行的局限,演进为与物理环境深度融合的设计、验证与优化核心引擎。行业竞争的焦点也将逐步

转向仿真架构的自主进化能力与集成深度。

参考文献

- [1] 段晓东, 程伟强, 张昊. 智算网络发展综述[J]. 中兴通讯技术, 2025, 31(2): 53–62. DOI:10.12142/ZTETJ.202502008
- [2] 边彦晖, 刘明远, 虞红芳. 面向多算力中心协同的广域智算网络仿真架构设计[J]. 中兴通讯技术, 2025, 31(2): 39–46. DOI: 10.12142/ZTETJ.202502006
- [3] 翟恩南, 操佳敏, 钱坤, 等. 面向大模型时代的网络基础设施研究: 挑战、阶段成果与展望[J]. 计算机研究与发展, 2024, 61(11): 3664–3677
- [4] Akram A, Sawalha L. A survey of computer architecture simulation techniques and tools [J]. IEEE access, 2019, 7: 78120–78145. DOI: 10.1109/ACCESS.2019.2917698
- [5] Riley G F, Henderson T R. The ns-3 network simulator [M]// Wehrle K, Güneş M, Gross J. Modeling and Tools for Network Simulation. Berlin, Heidelberg: Springer, 2010: 15–34. DOI: 10.1007/978-3-642-12331-3_2
- [6] Bai S Y, Zheng H, Tian C, et al. Unison: a parallel-efficient and user-transparent network simulation kernel [C]//Proceedings of the Nineteenth European Conference on Computer Systems. ACM, 2024: 115–131. DOI: 10.1145/3627703.3629574
- [7] Gao K H, Chen L, Li D, et al. DONS: fast and affordable discrete event network simulation with automatic parallelization [C]//Proceedings of the ACM SIGCOMM 2023 Conference. ACM, 2023: 167–181. DOI: 10.1145/3603269.3604844
- [8] Fan W W, Hong L J, Jiang G X, et al. Review of large-scale simulation optimization [PP/OL]. arXiv (2024-03-23) [2025-12-26]. <https://arxiv.org/abs/2403.15669>
- [9] Won W, Heo T, Rashidi S, et al. ASTRA-sim2.0: modeling hierarchical networks and disaggregated systems for large-model training at scale [PP/OL]. arXiv (2023-03-24) [2025-12-26]. <https://arxiv.org/abs/2303.14006>
- [10] Casanova H, Giersch A, Legrand A, et al. Lowering entry barriers to developing custom simulators of distributed applications and platforms with SimGrid [J]. Parallel computing, 2025, 123: 103125. DOI: 10.1016/j.parco.2025.103125
- [11] Wang X, Li Q, Feng Y, et al. SimAI: unifying architecture design and performance tuning for large-scale large language model training with scalability and precision [C]//22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI'25). USENIX Association, 2025: 541–558. DOI: 10.5281/zenodo.14287654
- [12] Agrawal A, Kedia N, Mohan J, et al. Vidur: a large-scale simulation framework for LLM inference [PP/OL]. arXiv (2024-05-08) [2025-12-26]. <https://arxiv.org/abs/2405.05465>
- [13] Shen S Y, Bonato T, Hu Z Y, et al. ATLAHS: an application-centric network simulator toolchain for AI, HPC, and distributed storage [C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2025: 349–367. DOI: 10.1145/3712285.3759838
- [14] Zuo P F, Lin H M, Deng J B, et al. Serving large language models on Huawei CloudMatrix384 [PP/OL]. arXiv (2025-06-15) [2025-12-26]. <https://arxiv.org/abs/2506.12708>
- [15] Gandhi S, Zhao M, Hariharan S, et al. SlipStream: adapting pipelines for distributed training of large DNNs amid failures [PP/OL]. arXiv (2024-05-22) [2025-12-26]. <https://arxiv.org/abs/2405.14009>
- [16] Li H J, Balasubramanian P, Meiers M, et al. SplitSim: large-scale simulations for evaluating network systems research [PP/OL]. arXiv (2024-02-07) [2025-12-26]. <https://arxiv.org/abs/2402.05312>
- [17] Kortas N, Recker T. Large-scale space network simulator for performance-optimized DTNs [C]//Proceedings of the 15th International Conference on Satellite and Space Communications (SPACOMM 2023). 2023: 45–50. DOI: 10.5220/0012178900450050.
- [18] Gui F, Gao K H, Chen L, et al. Accelerating design space exploration for LLM training systems with multi-experiment parallel simulation [C]//Proceedings of the 22nd USENIX Symposium on Networked Systems Design and Implementation. ACM, 2025: 473–488. DOI: 10.5555/3767955.3767980
- [19] Wang H D, Yan H, Rong C, et al. Multi-scale simulation of complex systems: a perspective of integrating knowledge and data [PP/OL]. arXiv (2023-06-17) [2025-12-26]. <https://arxiv.org/abs/2306.10275>
- [20] Svedas J, Watson H, Laubeuf N, et al. A survey of end-to-end modeling for distributed DNN training: workloads, simulators, and TCO [PP/OL]. arXiv (2025-06-10) [2025-12-26]. <https://arxiv.org/abs/2506.09275>
- [21] Kumar S, Temura A, Sharma N, et al. Simulating LLM training workloads for heterogeneous compute and network infrastructure [C]//Proceedings of the 2nd Workshop on Networks for AI Computing. ACM, 2025: 105–107. DOI: 10.1145/3748273.3749212
- [22] Heo J, Kim S, Shin H, et al. NetTLMSim: a virtual prototype simulator for large-scale accelerator networks [C]//The 3rd Workshop on Computer Architecture Modeling and Simulation (CAMS'23). ACM, 2023. DOI: 10.1145/3619895.3623250
- [23] Rico-Gallego J A, Díaz-Martín J C, Manumachu R R, et al. A survey of communication performance models for high-performance computing [J]. ACM computing surveys, 2019, 51 (6): 1–36. DOI: 10.1145/3284358
- [24] Aarwal A, Anand S, Jain A, et al. Arcadia: an end-to-end AI system performance simulator [EB/OL]. (2023-09-07) [2025-12-26]. <https://engineering.fb.com/2023/09/07/data-infrastructure/arcadia-end-to-end-ai-system-performance-simulator/>
- [25] Nguyen Q C. Network sampling: an overview and comparative analysis [PP/OL]. arXiv (2025-04-24) [2025-12-26]. <https://arxiv.org/abs/2504.17701>
- [26] Yang Q Q, Peng X, Chen L, et al. DeepQueueNet: towards scalable and generalized network performance estimation with packet-level visibility [C]//Proceedings of the ACM SIGCOMM 2022 Conference. ACM, 2022: 441–457. DOI: 10.1145/3544216.3544248
- [27] Khashab S, Sezhayan H, Abboud R, et al. NSX: large-scale network simulation on an AI server [C]//Proceedings of the 2nd Workshop on Networks for AI Computing. ACM, 2025: 19–25. DOI: 10.1145/3748273.3749199
- [28] Yang L Y, Luo S, Cheng X, et al. Leveraging large language models for enhanced digital twin modeling: trends, methods, and challenges [PP/OL]. arXiv (2025-03-04) [2025-12-26]. <https://arxiv.org/abs/2503.02167>
- [29] Qian Y C, Shu R, Ma R, et al. Miniature: fast AI supercomputer networks simulation on FPGAs [C]//Proceedings of the 9th Asia-Pacific Workshop on Networking. ACM, 2025: 114–120. DOI: 10.1145/3735358.3735369
- [30] 田海东, 张明政, 常锐, 等. 大模型训练技术综述[J]. 中兴通讯技术, 2024, 30(2): 21–28. DOI: 10.12142/ZTETJ.202402004
- [31] Li C N, Zabreyko A A, Nasr-Esfahany A, et al. m4: a learned flow-level network simulator [PP/OL]. arXiv (2025-03-03) [2025-12-26]. <https://arxiv.org/abs/2503.01770>
- [32] Yu G B, Tan G, Huang H J, et al. A survey on failure analysis and fault injection in AI systems [PP/OL]. arXiv (2024-06-28) [2025-

12–26]. <https://arxiv.org/abs/2407.00125>

[33] 朱炫鹏, 姚海东, 刘隽, 等. 大语言模型算法演进综述 [J]. 中兴通讯技术, 2024, 30(2): 9–20

[34] Zhang Q Z, Ng K K W, Kazer C, et al. MimicNet: fast performance estimates for data center networks with machine learning [C]//Proceedings of the 2021 ACM SIGCOMM 2021 Conference. ACM, 2021: 287–304. DOI: 10.1145/3452296.3472926

作者简介



熊先奎, 中兴通讯股份有限公司无线首席架构师、智算技术委员会前瞻组组长; 长期从事计算系统和体系结构、先进计算范式以及异构计算加速器研究工作; 曾主导过中兴通讯 ATCA 先进电信计算平台、服务器存储平台、智能网卡和 AI 加速器等系统架构设计。



童贤慧, 中兴通讯股份有限公司先进计算存储系统工程师; 长期从事智算系统仿真、安全可信计算等方向的技术研究。



王勇, 中兴通讯股份有限公司预研工程师; 长期从事网络协议、可编程网络以及加速器相关的技术研发工作。



张启明, 中兴通讯股份有限公司系统架构师; 长期从事智算网络及系统架构前瞻预研工作。

基于特征事件的PON网络故障分析诊断与应用



PON Network Fault Analysis, Diagnosis and Application Based on Feature Events

余辰东/Yu Chendong¹, 孙志/Sun Zhi²

(1. 中兴通讯股份有限公司, 中国 深圳 518057;

2. 中国电信集团有限公司深圳分公司, 中国 深圳 518048)

(1. ZTE Corporation, Shenzhen 518057, China;

2. Shenzhen Branch of China Telecom, Shenzhen 518048, China)

DOI: 10.12142/ZTETJ.202603012

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20241212.1005.002>

网络出版日期: 2024-12-12

收稿日期: 2024-10-16

摘要: 为提升无源光网络(PON)运行稳定性以及相关故障的快速定位能力, 提出了基于事件特征的综合故障处理方法。该方法首先通过事件去噪提升查准率, 然后通过事件聚合提升查全率, 并基于关联数据, 借助算法演进完成相关网络问题的快速感知、预警与定位。以PON网络为例, 通过关键事件(告警)的选取, 并基于上述流程实现PON链路故障的快速高效定位。典型网络分析和仿真表明, 该方案去噪比与聚合查全率都能达到80%以上, 能够显著提升PON网络运维效率。

关键词: 去噪; 聚合; 查准率; 查全率; 无源光网络; 信号丢失; 特定光网络单元信号丢失

Abstract: To enhance the operational stability and rapid fault localization capability of passive optical networks (PON), this paper proposes an integrated fault processing method based on event features. Firstly, the method enhances precision through event denoising, and then improves recall through event aggregation. Combined with auxiliary data, the method employs iterative optimization algorithms to achieve rapid detection, early warning, and accurate localization of network anomalies. For PON, the scheme selects key alarm events and applies the above processing flow to efficiently locate link faults. Typical network analysis and simulation results demonstrate that the proposed scheme achieves over 80% denoising accuracy and event aggregation recall, which can significantly boost PON operation and maintenance efficiency.

Keywords: de-noising; aggregation; precision; recall; PON; loss of signal; LOSi

引用格式: 余辰东, 孙志. 基于特征事件的PON网络故障分析诊断与应用[J]. 中兴通讯技术, 2026, 32(3): 69-75. DOI: 10.12142/ZTETJ.202603012

Citation: Yu C D, Sun Z. PON network fault analysis, diagnosis and application based on feature events [J]. ZTE technology journal, 2026, 32(3): 69-75. DOI: 10.12142/ZTETJ.202603012

无源光网络(PON)技术在接入网领域已得到大规模部署。这主要得益于两大核心优势: 1) 采用无源传输网络, 易于部署维护; 2) 采用点对多点(P2MP)拓扑结构, 单PON口下最多可支持64~128个光网络单元(ONU)接入。中国信息通信研究院发布的《中国宽带发展白皮书(2023年)》指出, 当前中国已建成光纤到户端口10.8亿个, 现网10G PON端口数达2 184.8万个, PON已成为主流光接入网技术。

然而, 实际上随着PON网络规模的快速扩大^[1], 运维压力与运维复杂度也在日益增加, 具体体现在两个方面: 1) 根据《接入网技术要求——基于以太网的无源光网络(EPON)》规范, PON作为二层业务转发的系统, 无法直接

感知其承载业务的变化, 即PON网络故障感知能力仍存在不足; 2) 根据IEEE 802.3ah、ITU-T G.984.x等标准定义, PON上行时分多址接入(TDMA)机制和P2MP拓扑结构, 容易导致PON端口下发生单点故障, 增加问题定位与处置难度^[2]。

落实到PON系统现场运维场景中, 现存痛点主要包括3个方面:

1) 位于用户侧的ONU设备受工程安装环境、线路外设质量、用户使用习惯等多重因素干扰, 影响程度远高于传统通信设备^[3]。相较于局端部署的光线路终端(OLT), ONU运行稳定性偏弱, 易出现工作状态异常, 并直接反映在PON链路拓扑上。

2) 随着光接入技术与各类业务应用持续迭代, PON网络拓扑割接、扩容等工程操作会引发ONU状态频繁、大幅波动;与此同时, OLT与上层运维平台间的信令交互, 同样会改变ONU运行状态。

3) 单PON端口下挂载多台ONU, 使得OLT侧汇聚的ONU状态变更、告警信息数量呈爆发式增长, 容易引发运维信息过载, 导致故障处置效率显著降低^[4]。

综上所述, 当前PON网络亟需提升关键异常识别与故障定界、定位的处理效率。

1 故障处理步骤分解

为实现故障的高效精准定位, 打破传统运维模式的局限, 本文从网络层面分析各类异常事件及其关联特征, 推动故障定位方式由单一事件处理模式, 向多源异常数据综合分析模式转变。具体处理流程如图1所示: 对网络中的关键事件进行提取, 并将其作为核心特征, 通过去噪、聚合、识别与定位等过程, 同时以查全率、查准率为评价指标, 实现对整体故障定界、定位结果的评价。

传统故障处理方式主要依托运维人员人工经验, 通过筛选概率最高的故障诱因完成定位。该方式主观依赖性较强, 预估故障原因往往与实际故障存在偏差, 直接影响故障查全率与查准率。其中, 查准率(Precision)是衡量检索系统信

噪比的一种指标, 如公式(1)所示; 查全率(Recall)是指检出的相关信息量与总量的比值, 如公式(2)所示。查全率与查准率两者之间具有互逆相关性:

$$\text{查准率} = \frac{\text{预测为1且正确预测的样本数}}{\text{所有预测为1的样本数}} = \frac{TP}{TP + FP} \quad (1),$$

$$\text{查全率} = \frac{\text{预测为1且正确预测的样本数}}{\text{真实情况下所有为1的样本数}} = \frac{TP}{TP + FN} \quad (2)。$$

查准率与查全率的计算涉及4个评估分类模型性能参数, 其中:

TP表示: 真实情况为“真”, 模型也正确预测为“真”的样本数。

FN表示: 真实情况为“真”, 但模型错误预测为“假”的样本数。

FP表示: 真实情况为“假”, 但模型错误预测为“真”的样本数。

TN表示: 真实情况为“假”, 模型也正确预测为“假”的样本数。

1.1 去噪

去噪即筛选虚假事件, 以此提高查准率。滤除无效数据的事件类型主要包括: 网络割接、设备升级等主动运维操作产生的事件, PON端口上行光信号丢失(PON LOS)等存在多义性、需要语义转译的告警事件, 以及子接口告警抖动等高频波动事件。

如果去噪后数据质量比较好(查准率评价), 那么异常事件识别结果与实际运行状态之间的偏差就会大幅减小。以主干光纤断裂场景为例, 传统运维模式通常将PON LOS告警直接判定为主干光纤断裂, 然而该告警也可能由小区断电等场景触发。结合公式(1)可知, 传统方法判定的主干光纤断裂预测样本(TP+FP)中, 混入了小区断电导致的误判样本(FP), 通过数据去噪手段剔除该类无效样本, 能够有效减少误判数量, 降低FP值, 优化故障定位精度。

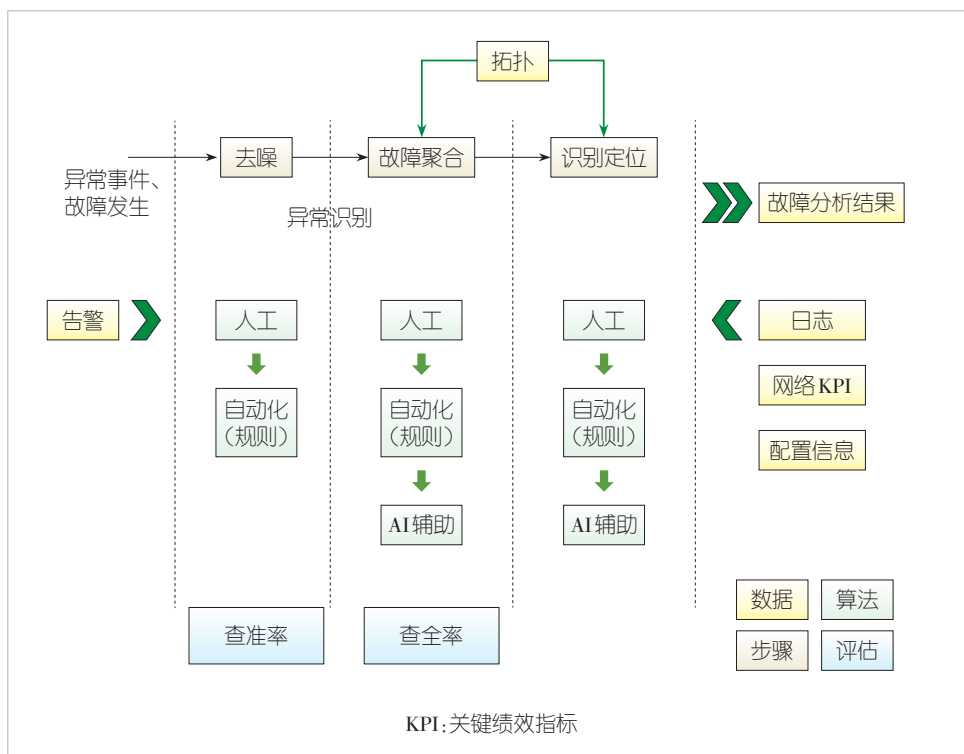


图1 基于事件特征的综合故障处理流程

1.2 故障聚合

外在事件与故障根因之间并非简单映射。故障根因除了会导致本端口异常外,也可能导致多个子端口异常,且各子端口表现为各自独立的事件。为此,可借助聚合策略,将多个相对独立的子事件通过内在逻辑进行关联,以确定根因,有效提升查全率^[5]。

在实际应用中,主干或支路光纤中断会触发告警并推送运维派单任务,需人工维修处理,而终端掉电场景则不需要。若PON端口下ONU掉电(DGi),此时OLT PON端口会检测该ONU上行无光信号(LOSi)。这种情况下的DGi和LOSi实际表征的是同一个故障事件,即ONU掉电。若PON端口下的所有ONU均掉电,此时PON端口检测上行无光信号(LOS)。这种情况下,PON端口下所有ONU的DGi与PON LOS实际表征的也是同一个故障事件,即PON端口下多个ONU掉电。由于PON端口一般会覆盖同一个区域,因此这种掉电也称为小区掉电。由上述描述可知,这两种场景下产生的LOSi或LOS,实际上并非由支路光纤或主干光纤异常(中断)引发PON端口无收光,属于无需运维派单的虚假负样本(FN)。因此,需通过针对性优化降低FN数量,以此提升故障定位查全率,具体原理如公式(2)所示。

事件聚合可实现时间域、空间域与逻辑域的关联事件聚合,将零散、独立的告警事件整合为逻辑统一的事件集合,具体流程如图2所示。首先,通过设定时间范围、关联端口数量等事件集合阈值,完成告警事件的真伪初步判别;其次,以整合后的事件集合为索引开展全网检索,结合检索结果与辅助判定条件,校验当前事件阈值的合理性;最后,依托对应算法完成事件规则的优化,动态调整判别窗口与阈

值大小,实现故障判别精度的持续提升。

以时序关联规则为例,同一条链路产生的LOSi(ONU上行信号丢失)、DGi(ONU断电)、LOS 3类告警通常存在同步触发特征。系统可划定固定时间窗口,依托告警产生先后顺序、触发频次等特征完成多告警聚合,生成统一告警事件集。相较于离散单条告警,该事件集能够更全面、准确地反映对应链路的运行状态与链路质量。除时序维度外,还可基于异常告警所处网络拓扑位置开展关联聚合,从更高物理(逻辑)接口维度完成故障信息整合,实现对故障根因的精准刻画。

1.3 识别定位

借助去噪与聚合手段可使关键事件被快速识别。在此基础上,依托时序维度或拓扑物理维度开展统计对比分析,并结合网络关键绩效指标(KPI)、运行日志、设备配置文件等多源数据,综合研判网络运行状态,完成故障初步定位。通过计算同一时段、同一物理链路内各类事件的关联度,筛选高关联告警组合,综合判定网络异常根因。现有PON链路质差分析便是该类识别定位思路的典型应用。该方案可通过时序网络指标关联分析完成故障研判,例如对OLT侧误码率、前向纠错(FEC)效率、光模块功耗波动等指标开展联合分析。

基于事件关联的故障定位仅能输出PON网络异常现象与潜在诱因,无法直接建立网络异常与上层业务故障的一一映射关系,还需引入外部系统数据交叉核验。例如需要对接宽带远程接入服务器(BRAS)和认证、授权、计费(AAA)服务器的用户账号信息,以便验证业务故障真实发生与否;同时采集故障工单闭环反馈数据,持续迭代优化关联学习算法,不断提升整套故障识别定位流程的处理效率与判定精度^[6]。

2 关键事件选取与定义

告警能够直观反映网络实时运行状态,具备感知时效性强、采集便捷的特征,以各类告警为基础筛选关键事件具备较高可行性。

2.1 关键告警选取

PON网络的关键告警筛选规则需结合自身组网运行特征制定。在PON系统中,ONU需先完

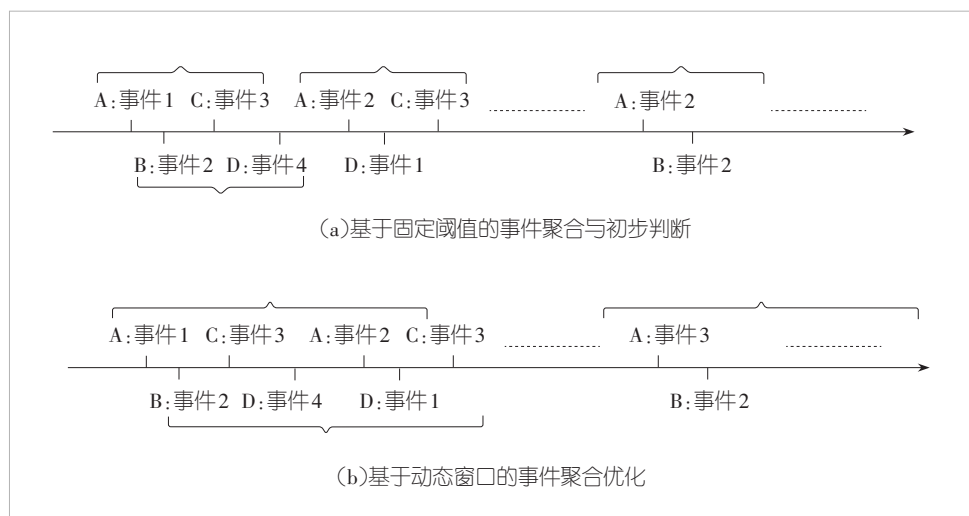


图2 事件合并(聚类)流程图

成OLT注册流程,后续上下行业务数据转发均由OLT统一调度管控,因此ONU稳定注册是PON网络可靠运行的基础条件。PON网络部署海量ONU设备,且终端易出现状态波动,使得ONU注册状态变更成为网络高频发生事件;其中LOS_i(ONU侧信号丢失)告警可直接标识单台ONU运行异常与状态切换。与之对应,PON LOS(PON端口信号丢失)告警能够反映该PON端口下全部ONU的在线连通情况。因此,本文选取PON LOS与LOS_i两类告警作为评判PON网络运行稳定性的核心事件。PON LOS、LOS_i及配套DG_i告警的标准定义如表1所示。

依据设备标准告警定义,当系统上报LOS或LOS_i告警时,代表上行链路无法识别有效光信号,对应PON端口或ONU设备将脱离注册状态。一方面可统计LOS、LOS_i告警的发生频次,以此量化对应物理端口或终端单元的运行稳定程度;另一方面可依托告警持续时长,表征对应接口或ONU的实际业务中断时长。

2.2 LOS/LOS_i告警工程应用

网络运维依托各类告警驱动故障处置流程。在实际工程场景中,LOS、LOS_i告警通常映射为主干光纤故障与分支光纤故障,并自动触发运维工单。为提升用户业务感知,运维对业务中断的管控粒度由多用户群体性故障细化至单台ONU级故障,对应的工单考核标准也同步拓展:PON LOS告警对应主干光纤中断,LOS_i告警对应分支光纤中断。

针对主干光纤中断场景,系统上报PON LOS告警时需附带受影响ONU数量,用于判定故障影响范围并直接生成运维工单;针对分支光纤中断场景,若时间窗口 T_1 内累计产生 N 条LOS_i告警(通常 $N \geq 2$),系统将合并多条告警统一触发工单。

若PON LOS告警与ONU掉电事件DG_i伴随产生,则不生成对应工单。针对LOS_i告警,仅当告

警持续时长超过阈值 T_2 (T_2 大于ONU正常重启时间)时才触发工单,以此过滤终端重启带来的瞬时LOS_i告警。此外,若单台ONU在时间窗口 T_3 内累计上报 M 次LOS_i告警,则判定该终端存在高频闪断故障,系统需对该类异常精准识别并单独标记。

3 LOS/LOS_i告警去噪

对LOS、LOS_i两类关键告警进行去噪处理,可剔除告警中与实际故障无关的干扰信息,保证留存告警与待定位故障具备强关联性。其中,针对PON LOS告警,需区分主干光纤中断与终端断电两类诱因;针对LOS_i告警,则需甄别分支光纤故障与其他非故障类触发因素。

3.1 LOS去噪

PON LOS告警用于表征主干光纤中断故障,其告警判定与原因分析流程如图3所示。ONU通过Ploam通道上传DG_i断电告警,OLT依托DG_i告警状态完成终端告警迁移逻辑:仅上报LOS_i时直接输出该告警;若LOS_i与DG_i同步上报,则将终端状态统一标记为断电DG_i。在此基础上,系统扫描对应PON端口运行状态,结合PON LOS告警时段内全部ONU的状态数据,借助原因分析算法将PON LOS告警划分为7类触发场景^[8]:

1) 主干光纤中断: PON端口下注册多个ONU,主干光

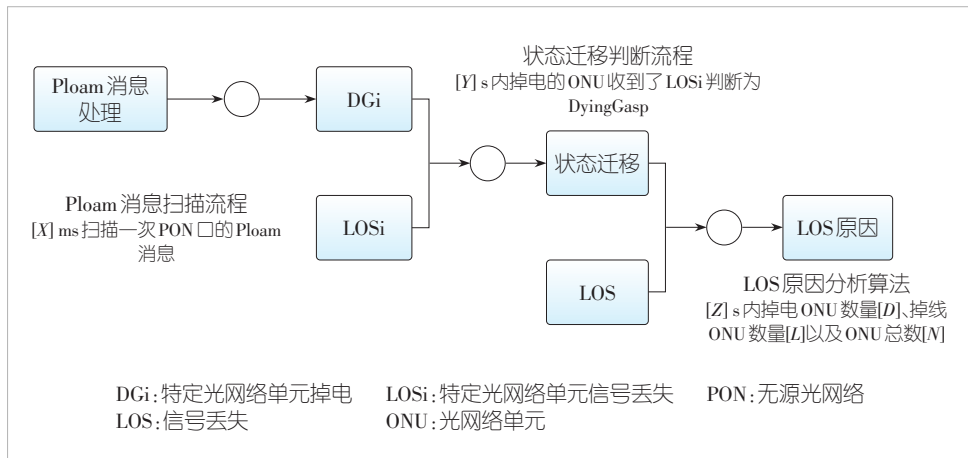


图3 PON LOS告警判定与原因分析流程

表1 LOS/LOS_i的标准定义^[7]

类型	告警产生条件	告警恢复条件
信号丢失	上行方向,OLT PON端口连续4帧未收到预期有效数据	OLT PON口至少接收到一个有效数据
ONU _i 信号丢失	PON端口连续4帧里没有检测某ONU的上行信号	OLT接收到对应ONU的一个有效光信号
ONU _i 掉电	OLT接收到ONU上报的DyingGasp消息	OLT在ONU测距阶段收到一个Ploam消息

OLT: 光线路终端 ONU: 光网络单元 ONU_i: 特定光网络单元 PON: 无源光网络

纤中断;

2) 小区掉电: PON 端口下注册多个 ONU, 所有 ONU 的电源同时关闭;

3) ONU 未上线: PON 端口下注册多个 ONU, 但是 ONU 从未上线;

4) 唯一 ONU 掉电: PON 端口下只注册一个 ONU, ONU 电源关闭;

5) 唯一 ONU 掉线: PON 端口下只注册一个 ONU, 该 ONU 的光纤中断;

6) 最后 ONU 掉电: PON 端口多个 ONU 掉线或断电 Z 秒后, 最后一个 ONU 电源关闭;

7) 最后 ONU 掉线: PON 端口多个 ONU 掉线或断电 Z 秒后, 最后一个 ONU 光纤中断。

在 7 类场景中, 仅第 1 类对应真实主干光纤中断故障, 其余 6 类均为无主干故障的干扰告警, 可完成过滤剔除, 对应去噪识别率可达 85.71%。

3.2 LOSi 去噪(防抖)

ONU 频繁上报 LOSi 告警不仅会增加后台系统处理负载, 还易引发故障根因误判, 因此需引入告警防抖去噪机制^[9]。该机制的核心逻辑为: 系统缓存预设防抖时长内的告警生成、告警恢复事件, 暂不向上层推送; 待防抖计时器超时后, 再统一上报完整的告警状态变更信息。若告警触发防抖过滤规则, 原始告警明细将无法直达上层运维平台, 为此系统增设高频告警专项通知上报逻辑, 保障运维侧能够感知终端持续闪断现象。两种告警防抖方案如图 4 所示, 其中图 4 (a) 为告警生成、告警恢复独立防抖方案, 图 4 (b) 为生成与恢复联动的综合防抖方案。两种方案的差异在于: 后者对成对出现的告警生成、恢复报文同步执行防抖校验, 逻辑约束更严密, 输出告警数量统计结果准确度更高。

合理配置防抖时间窗口与告警频次阈值, 能够高效过滤无业务影响的无效瞬时告警, 显著提升故障识别查准率。现网实测结果表明, 告警防抖机制可过滤 50%~80% 的干扰告警; 极端场景下无效告警占比可达 99.99%, 例如某台 ONU

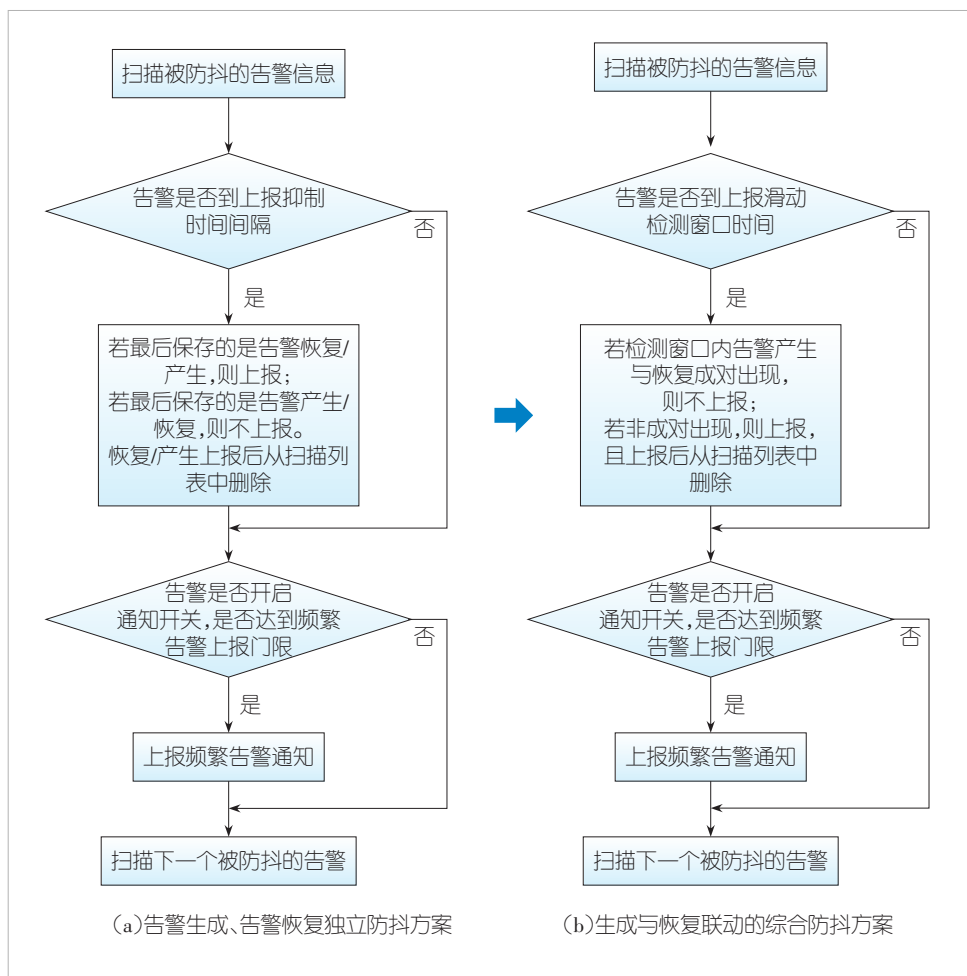


图4 两种告警防抖方案

累计产生 73 787 条 LOSi 告警, 其中 73 785 条均为终端高频闪断引发的无效瞬时告警。

4 告警聚合

完成关键告警筛选与去噪处理后, 留存告警虽可准确反映某一物理/逻辑节点在特定时刻的运行状态, 但仅依靠单条告警无法完整刻画链路、业务乃至整网的实时运行状态。因此, 对多条关键告警开展聚合与关联分析, 从多维度还原网络整体运行状态具备重要工程价值。实践表明, 可从时域、空域两大维度完成告警关联挖掘与聚合处理。

1) 时域关联聚合

以固定物理或逻辑端口为分析对象, 将时间作为分析维度, 依托滑动时间窗口(如图 2 所示)挖掘窗口内多条告警的关联特征。如图 5 所示, 针对单一端口下所有 ONU 告警开展关联性分析: 同一接口产生的多条告警在时间轴上存在区间重叠, 或告警触发顺序呈现固定先后时序。以现网故障为例, 当 ONU 频繁离线时, 经告警聚合可发现 ONU 离线告警

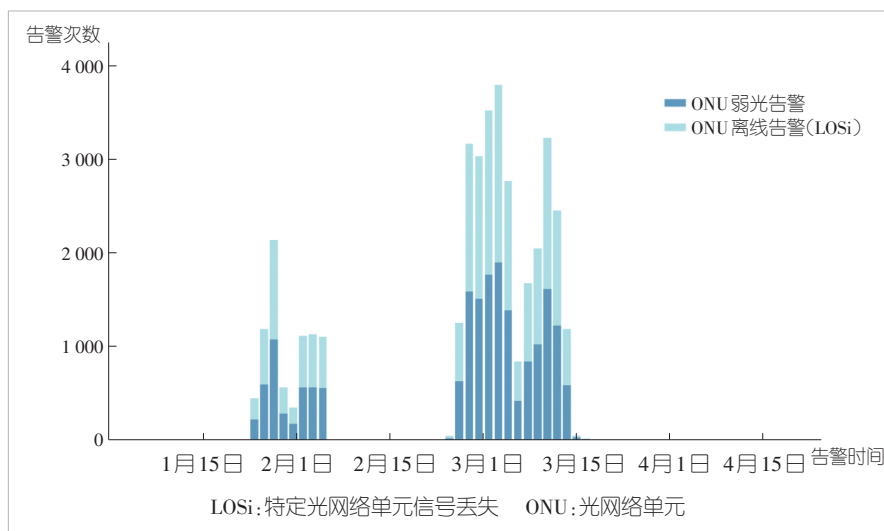


图5 告警时域关联性分析

与弱光告警具备高度时间同步性，由此可判定ONU离线故障由光功率过低引发。

2) 空域关联聚合

锁定固定时间区间，以物理/逻辑端口为分析维度，调整空域分析范围，挖掘存在拓扑关联的各端口告警的同步特征。例如统计同一PON端口下多台ONU同步上报LOSi离线告警的发生概率，并结合其他辅助条件，分析批量终端离线现象与特定时段的关联规律。

整套告警聚合流程分为去噪、聚合、优化3个步骤：第1步借助去噪机制过滤终端高频LOSi干扰告警；第2步统计单位时间内同一PON端口下多台终端产生的同类LOSi告警，结合预设阈值筛选并生成告警事件集合；第3步对告警集合与其他网络事件开展关联分析，推导故障集合对应的潜在根因，各环节详细流程如下：

- 去噪：统计PON端口下LOSi告警数量，设定阈值 k ，剔除任一告警数量 $\geq k$ 的ONU；

- 聚合：统计剩余LOSi告警平均值 μ ，设定系数 α ，记录任一告警数量超过 $\alpha \times \mu$ 的ONU；

- 优化：调整系数 α ，设定影响因子 β ，记录聚合ONU数量超过 $\alpha \times \mu$ 或 β 的ONU。

如图6所示，某PON口下ONU离线事件集合与未注册ONU事件存在某种关联性，在30分钟内ONU离线事件集合

（多个ONU同时离线）发生5次。以1分钟为关联分析窗口判断，离线告警事件与未注册ONU事件相关性为80%；以5分钟为关联分析域，相关性为100%。也即该网络中多ONU出现同时离线问题高概率由链路中的未注册ONU引入。

5 定位验证

依托前文关键告警提取、去噪与聚合全流程，系统能够高效识别网络异常事件并初步推导潜在故障诱因。但各类告警与真实故障根因之间不存在严格一一映射关系，PON LOS告警即为典型案例。仅依靠告警聚合分析不足以完成精准根因定位，还需引入上层多源辅助数

据开展交叉核验，例如对接第三方运维平台、设备运行日志等外部信息。本文结合PON网络组网架构设计一套分层故障研判体系，整体流程如图7所示^[10]，具体执行步骤如下：

第1步，结合PON网络拓扑结构与故障分布特征，将全网故障划分为四大类别：PON局端设备故障、ONU终端设备故障、ODN链路结构故障、ODN链路质量故障。

第2步，针对每一类故障梳理完备的潜在根因清单，并逐条配置故障判别规则完成逻辑校验；针对难以通过固定规则判定的复杂故障场景，引入人工智能学习模型实现智能识别。判别规则所需原始数据取自单位周期内PON端口与ONU终端的告警信息、性能指标统计值。原始数据先经前文去噪、聚合算法完成标准化特征提取，再输入故障判别规

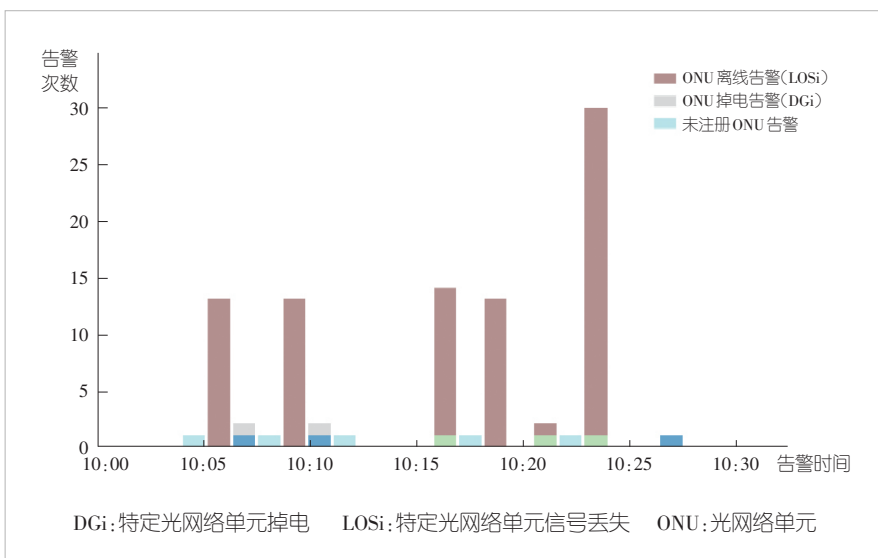


图6 告警空域关联性分析

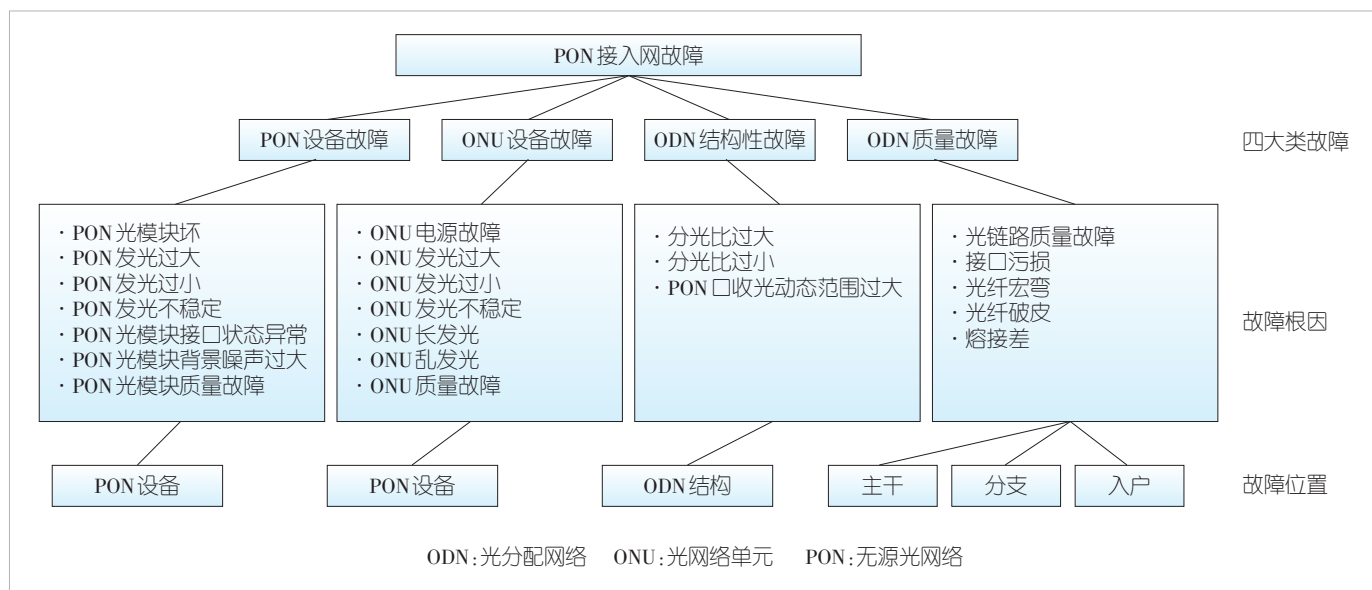


图7 PON接入网故障构成和故障位置设计

则库完成初步根因判定。

第3步，针对规则库无法明确判定状态的样本数据，采用AI故障识别模型二次研判，确认是否存在隐性故障；针对识别为ODN质量类故障的样本，模型进一步智能区分故障发生区段，判别故障位于主干光纤、分支光纤还是入户光纤。

第4步，汇总4类故障的研判结果，综合输出故障精准发生位置与最终故障根因。

6 结束语

综上所述，伴随光接入网部署规模持续扩张，网络智能运维的落地需求愈发突出。基于网元光信号告警完成去噪处理，能够有效提升原始数据洁净度，是后续故障分析高效开展的基础；对去噪后的告警数据实施关联聚合，可完成原始离散告警向结构化网络事件的转换；在此基础上依托可自定义配置的判别规则对事件开展智能分析与根因诊断，能够整体提升PON网络故障排查效率。除告警去噪环节外，文中其余数据处理流程均可与PON网元设备解耦，研究人员能够以故障识别的查全率、查准率作为量化指标，对各模块算法分别迭代优化。

参考文献

- [1] 张伟良, 王霄雨, 黄新刚. 基于动态通道绑定的更高速无源光网络[J]. 中兴通讯技术, 2024, 30(2): 100-106. DOI: 10.12142/ZTETJ.202402014
- [2] 沈成彬. 无源光网络的运行、维护与管理[J]. 中兴通讯技术, 2014(5): 17-18
- [3] 余辰东. PON突发式长发光ONU的防护与检测[J]. 光通信研究, 2017(1): 23. DOI: 10.13756/j.gtxyj.2017.01.008

- [4] 张鹏, 阎阔. FTTx PON技术与应用[M]. 北京: 人民邮电出版社, 2010
- [5] QB-B-102-202001 面向集中运维的无源光网络(PON)OMC系统功能测试规范[S]
- [6] 赵宇粟, 张鹏飞, 金耀辉. 软件定义网络中网络诊断的思考和探索[J]. 中兴通讯技术, 2015, 21(4): 41-44
- [7] YD/T 1949.2-2009 接入网技术要求. 吉比特的无源光网络(GPON). 第2部分: 物理媒质相关(PMD)层要求[S]
- [8] 余辰东, 于金辉. 一种光网络单元的注册方法、装置及光线路终端: CN107295428B [P]. 2020-07-31
- [9] 余辰东, 于金辉. 无源光网络设备告警信息处理方法及光线路终端: CN105991320B [P]. 2020-07-31
- [10] 陈爱民, 陆云, 余辰东. 一种PON系统的故障分析方法、装置及设备: CN118101416A [P]. 2024-05-28

作者简介



余辰东, 中兴通讯股份有限公司固网产品技术支持总工; 主要研究方向为PON智能运维技术、光接入网SDN与自智网络演进; 发表论文22篇, 拥有专利21项。



孙志, 中国电信集团有限公司深圳分公司客户服务调度与资源中心总经理; 主要研究方向为网络资源管理、运营支持系统、确定性网络等。

基于5G地面公网的 网联无人机低空网络覆盖



Low-Altitude Network Coverage for Connected UAVs Based on 5G Terrestrial Public Networks

王俊杰/Wang Junjie¹, 王迪/Wang Di², 申奇/Shen Qi³,
肖腾飞/Xiao Tengfei¹

(1. 中国联合网络通信有限公司, 中国 北京 100033;

2. 中讯邮电咨询设计院, 中国 北京 100048;

3. 中国联合网络通信有限公司智能城市研究院, 中国 北京 100048)

(1. China United Network Communications Group Co., Ltd., Beijing 100033, China;

2. China Information Technology Design & Consulting Institute Co., Ltd., Beijing 100048, China;

3. China Unicom Smart City Research Institute, Beijing 100048, China)

DOI: 10.12142/ZTETJ.202603013

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20240808.0902.002>

网络出版日期: 2024-08-08

收稿日期: 2024-07-10

摘要: 对地面公网对空覆盖性能进行分析, 探索了面向5G网联无人机的低空网络发展思路与方案。仿真及测试结果表明, 测试目标区域内的地面公网对于300 m以下低空区域具有较好的连续对空覆盖能力, 但对于300 m以上低空区域的连续覆盖能力尚存在一定的局限性。针对基于地面公网低空网络的对空覆盖特性, 阐述了面向5G网联无人机的低空网络解决方案及相应的关键技术, 并给出了面向5G网联无人机的低空网络建设思路与建议。

关键词: 地面公网; 低空覆盖网络; 无人机

Abstract: This paper analyzes the air-to-ground coverage performance of terrestrial public networks and explores the development approaches and solutions for low-altitude networks for 5G-connected UAVs. Simulation and test results show that the terrestrial public network in the target area has good continuous air-to-ground coverage capability for low-altitude areas below 300 m, while its continuous coverage capability for areas above 300 m has certain limitations. In view of the air-to-ground coverage characteristics of terrestrial-public-network-based low-altitude networks, this paper elaborates the corresponding solutions and key technologies, and provides suggestions and recommendations for the construction and development of low-altitude networks for 5G-connected UAVs.

Keywords: terrestrial network; low-altitude coverage network; unmanned aerial vehicle

引用格式: 王俊杰, 王迪, 申奇, 等. 基于5G地面公网的网联无人机低空网络覆盖 [J]. 中兴通讯技术, 2026, 32(3): 76-82. DOI: 10.12142/ZTETJ.202603013

Citation: Wang J J, Wang D, Shen Q, et al. Low-altitude network coverage for connected UAVs based on 5G terrestrial public networks [J]. ZTE technology journal, 2026, 32(3): 76-82. DOI: 10.12142/ZTETJ.202603013

根据中国中长期科技发展规划, 空天地一体化网络已成为未来通信网络的重要发展趋势。低空智联网作为空天地一体化网络空基网络基础设施中不可或缺的一部分, 近年来发展迅速, 受到了各行业的高度关注^[1-2]。随着中国5G技术的飞速发展及无人机政策的逐步落地, 网联无人机产业已经迎来了黄金发展窗口期。地面5G网络所具备的大带宽、

低时延、高可靠以及大连接的特点为低空网联无人机通信提供了新的解决方案, 赋予了网联无人机更加多元化的应用, 并加速了网联无人机从消费级向工业级应用的转型。目前, 无人机在各行各业的应用场景中频频作为传统方案的替代者出现, 由此催生出无人机+行业应用的低空智联网新兴产业发展模式, 并且已经在一些典型行业中实现了广泛应用, 如物流运输、能源巡检、水域巡查、地质勘探、应急救援、农业植保、公共安防与消费娱乐等^[1-3]。

低空智联网依托地面蜂窝网络、低轨卫星、各类空基平

基金项目: 国家重点研发计划项目 (2020YFB1808001, 2020YFB1808004)

台及地面互联网基础设施,在3 000 m以下低空空域搭建适配低空业务运营的实体通信网络,是地面移动通信网络向空域的延伸。其中,面向5G网联无人机的低空网络以地面公网为基础,复用5G通信技术为低空飞行器终端提供物联网接入通道,支撑无人机全域管控与行业业务承载,是低空数字产业发展的核心底座^[4-7]。当前该类低空网络产业已进入高速增长阶段:干线与支线物流、中高空巡检的市场规模可达数百亿元;低空巡检、农林植保、测绘安防、末端配送、航拍娱乐等领域的市场规模有望突破1 000亿元^[3-4]。市场机构MNM测算数据显示,2025年中国无人机产业市场规模已达1 800亿元,约占全球市场份额17.5%;其中纯网络连接类业务占比约11%,网络连接配套平台综合服务业务占比68%^[2,4,8]。

现阶段运营商现网宏基站的无线覆盖设计以地面用户为核心,网络规划、参数优化工作均围绕地面移动业务展开,并未针对空域覆盖性能做专项调整。虽然宏基站定向天线主瓣增益指向地面,但低空信道以视距传播为主,现有公网基站仍具备一定空域覆盖能力^[1,9-10]。与传统地面蜂窝网络相比,5G网联无人机低空通信场景存在干扰源繁杂、空中小区覆盖形态不规则、邻区拓扑关系复杂等固有特征^[1,11],进而引发空中终端切换频繁、上下行链路自干扰、空地用户互扰等一系列覆盖与通信质量问题^[12]。针对上述痛点,本文搭建仿真模型评估现网基站低空覆盖能力,结合实测数据给出基于5G地面公网的低空组网建设与落地部署方案,并设计适配网联无人机的低空网络综合管理平台架构。

1 地面公网低空覆盖特性分析及试验验证

1.1 基于地面公网的低空网络覆盖场景

依据覆盖高度对地面公网低空覆盖场景分层,可划分为3类:高度低于100 m的地面覆盖层、高度100~300 m的低空覆盖层、高度300 m以上的高空空域覆盖层。地面覆盖层依托基站天线主瓣完成信号覆盖,区域内楼宇、植被遮挡密集,无线传播以非视距(NLOS)为主;100~300 m低空覆盖层依靠基站旁瓣实现空域覆盖,也是无人机主流作业飞行区间,该空域遮挡物稀少,信道多呈现视距(LOS)传播特征^[3,9]。典型的面向5G网联无人机的低空网络覆盖场景如图1所示。

由图1可见,地面覆盖层与低空覆盖层内终端的无线传输环境、基站覆盖特性存在显著区别,具体差异如下^[1]:

1) 无线传输环境存在差异。地面覆盖层受地形起伏、植被遮挡等共同影响,信道传播条件复杂。针对地面用户的

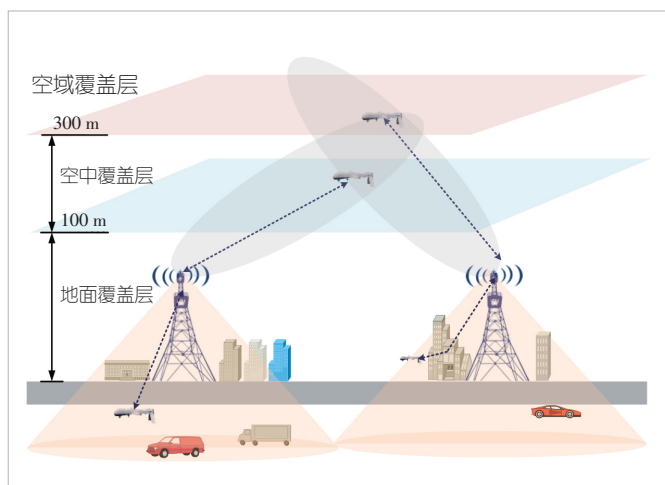


图1 5G网联无人机低空网络覆盖场景

蜂窝网络规划需预留阴影衰落、快衰落、建筑穿透等多重损耗余量,大幅削弱基站对地有效覆盖能力。而100~300 m低空覆盖层几乎无地形与建筑物遮挡,基站与无人机终端之间以LOS传播为主,直射路径信号能量稳定,能够有效减缓空中无线信号衰减速度。

2) 天线覆盖特性存在差异。地面覆盖层依托宏基站天线主瓣辅以少量旁瓣完成对地信号覆盖;低空覆盖层与300 m以上空域覆盖层依靠宏基站天线旁瓣实现空域信号覆盖。

1.2 基于地面公网的低空网络覆盖特性仿真及测试验证

结合图1中宏基站对空覆盖场景,以某行政区环境作为仿真及测试区域。该区域地表形态复杂,其环境呈现出山地、丘陵、平原等多种地貌类型。不失一般性,设宏基站间距为1 000 m,单站水平覆盖半径700 m,天线的机械下倾角为 $6^{\circ}\sim 8^{\circ}$,实际偏离有源天线单元(AAU)/射频拉远单元(RRU)法线方向的向上扩展角为 $22^{\circ}\sim 24^{\circ}$,位于主瓣3 dB外部区域,未进入深衰落和旁瓣区域。本文分别以时分双工(TDD) 64TR AAU设备、频分双工(FDD) 2T/4T RRU设备为例进行分析。FDD方式及TDD方式天线的垂直波瓣增益仿真结果如图2所示。其中, 0° 表示水平方向,正、负角度分别表示天线下倾、上仰的角度。AAU设备具有天线振子较多、垂直扫描波瓣宽度大以及波束赋形能力等特性。此外,RRU设备均为外接天线,且不具备波束扫描能力,其天线的垂直波瓣一般为 $6^{\circ}\sim 7^{\circ}$ 。故TDD方式与FDD方式的天线垂直波瓣差异较大。综上,相比于地面终端,200 m高度无人机的AAU垂直天线增益损失为10~15 dB,但一般公网规划会考虑室内用户的穿透损耗、阴影衰落、NLOS衰落

地面覆盖层

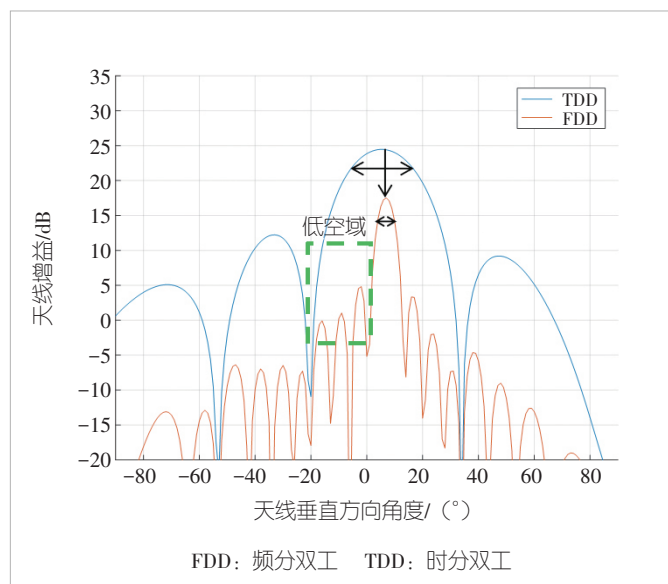


图2 TDD与FDD垂直波瓣增益结果图

余量等,上述因素的叠加会远超天线增益损失。

数值计算结果表明,在相同环境下NLOS和LOS信道模型的覆盖差异超过10 dB,且通信载波频率越高,差异越大。阴影衰落、快衰落和穿透损耗在低空域大幅下降,导致实际低空域用户和地面用户的覆盖差异超过20 dB。

为量化评估地面公网对空覆盖性能,本文依托某运营商北方区域现网工参,完成低空覆盖特性仿真与外场实测。链路预算结果表明:若保障1 km高度无人机终端上行边缘速率20 Mbit/s,宏基站站间距不宜超过7.41 km。基于该站间距约束,从目标行政区现网中选取161座5G宏基站作为空域覆盖测试站点。

不失一般性,本文设定地面公网基站的天线配置、RS发射功率、接收增益及灵敏度分别为64TR、43 dBm、25 dBi及-120 dBm,无人机终端的天线配置、参考信号接收功率(RSRP)分别为2T4R、26 dBm。设定仿真地理精度为20 m,频率为3.5 GHz,传输带宽为40 MHz。基于上述参数配置,对0.1~1 km高度的上行速率与RSRP进行仿真。不同高度下的上行速率与RSRP的仿真结果分别如图3与图4所示。

图3和图4的统计数据显示,理想LOS低空覆盖场景中,随着飞行高度提升,空域内终端上行速率与RSRP均呈现逐步衰减趋势。在0.1~1 km高度区间内,上行速率门限6 Mbit/s、10 Mbit/s、15 Mbit/s、20 Mbit/s、25 Mbit/s对应的最低空域覆盖率依次为94.86%、94.56%、92.88%、89.08%、82.56%。由图4可知,该高度区间的RSRP整体分布于-100~-85 dBm,约95%空域的RSRP高于-100 dBm。

为验证仿真结论的有效性,本文开展连续覆盖空域实地

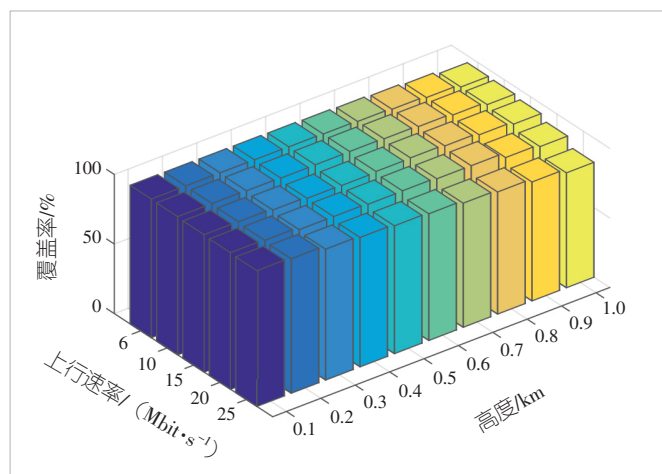


图3 测试区域内0.1~1 km的无人机终端上行速率统计结果

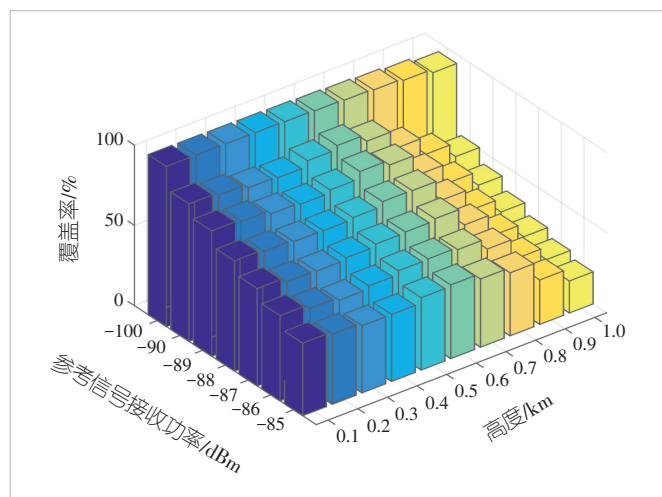


图4 测试区域内0.1~1 km高度的无人机终端RSRP统计结果

测试。为保证测试环境与仿真模型条件一致,测试选取的空域覆盖基站全部复用前文数值仿真的设备与工参配置。在连片覆盖区域搭载5G无人机通信模组,控制无人机沿预设航线往返巡航,完成小区移动性验证测试,测试区域宏基站部署点位与无人机飞行轨迹如图5所示。

结合图5测试场景,无人机单程飞行距离约8 km,分别设置100 m、200 m、300 m共3类典型低空飞行高度,各高度均完成6次往返飞行测试。实测结果表明,无人机低空飞行过程中物理小区标识(PCI)切换现象频发,其中200 m及以下低空区间的小区切换频次更高。测试全程采集无人机终端的多项核心运行指标,包括地面宏网覆盖下的RSRP均值、上行速率均值、Ping时延均值、上行调制与编码策略(MCS)、上行调度资源块(RB)数量、服务小区PCI、邻区RSRP及邻区PCI。基于上述实测指标,可进一步计算得到测试航段内无人机的小区切换成功率与5G网络驻留比,各

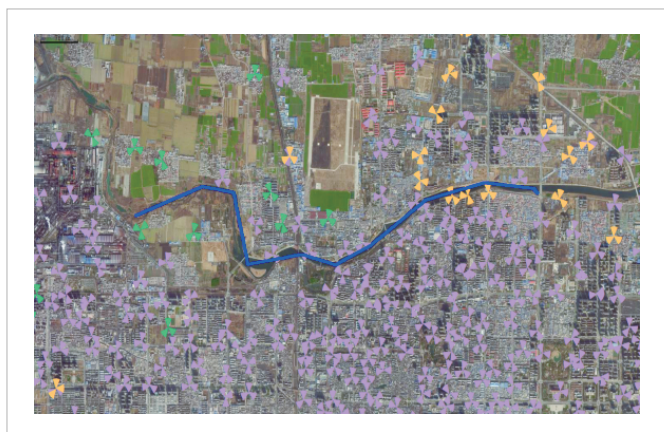


图5 测试区域基站分布与飞行路线图

项测试统计结果如表1所示。

由表1测试结果可知,无人机终端的RSRP与上行速率变化规律与理论仿真高度契合,各项实测指标均落在理论计算区间范围内。无人机通信终端平均Ping时延整体偏高,主要原因是低空飞行过程中频繁的小区切换引发的链路抖动与信令交互延迟。纵向对比不同飞行高度指标可知,300 m高度的无人机上行速率优于100 m、200 m低空区间,这是因为200 m及以下低空空域邻区拓扑关系复杂、信号干扰严重。同时,200 m高度的无人机上行速率与小区切换成功率均较低。该问题源于强遮挡环境(楼宇、山体)与复杂邻区的共同作用,频繁的切换重选导致通信性能劣化。

现场测试数据表明,现有地面5G公网对100~300 m低空区域具备良好的覆盖能力:空域整体RSRP均值约为-85 dBm,上行速率维持在60 Mbit/s左右,平均Ping时延约为35 ms,业务丢包率低于1%。从细分空域来看,测试区域100~300 m低空全域覆盖正常,所有RSRP的采样值均高于-95 dBm。其中,100~200 m高度的Ping时延基本稳定在30~40 ms,300 m高度的时延增至64 ms。此外,经实测发现,低空无人机飞行全程PCI切换频繁,直接造成上下行传输速率明显波动。

基于上述仿真及测试结果分析,现阶段基于地面5G公网的低空网络主要面临下述3个方面的挑战^[1,9,11]:

表1 地面公网基站的空域覆盖性能测试统计结果

高度/m	RSRP/ dBm	Ping时延/ ms	上行速率/ (Mbit·s ⁻¹)	切换成功 率/%	5G驻留 比/%
100	-80.55	32.60	46.11	100.00	100.00
200	-82.63	54.75	26.62	98.04	100.00
300	-86.45	71.43	64.76	100.00	89.08

RSRP: 参考信号接收功率

1) 地面公网低空覆盖主要依托宏基站天线旁瓣实现,是导致无人机小区频繁切换的主要原因。宏基站定向天线存在多组旁瓣结构,且旁瓣之间由天线零位隔离,零位区域无线辐射信号极弱。当无人机飞行至服务小区天线零位覆盖区间时,机载终端极易切换至周边小区旁瓣信号覆盖区域,引发连续PCI切换,增加移动性管理难度。同时,低空视距传播场景下,网络邻区关系随飞行高度提升愈发复杂、难以规划。相较于地面场景,低空终端可接收更多基站的上下行信号,小区重叠覆盖问题突出,信号干扰显著加剧。实测数据显示,低空无人机同步信号块(SSB)的信干噪比(SINR)较地面终端低10 dB左右。

2) 上下行干扰问题。低空LOS场景下,无人机终端将收到多个小区的无线信号,存在较大的上下行干扰。上行方面,地面用户主要由基站主瓣覆盖,无人机低空覆盖则是以地面基站旁瓣覆盖为主。邻区上行也是通过旁瓣接收无人机的干扰信号,其接收到的旁瓣增益远低于接收到的主瓣增益。下行方面,考虑到当前多为非集群式网联无人机通信业务场景,低空无人机对地面邻区用户的干扰较小。综上,空地用户相互干扰主要表现在上行干扰。

3) 空地用户间干扰问题。邻区地面用户的干扰信号被无人机服务小区的主瓣接收,主瓣接收增益相对较大,因此地面用户更容易干扰到无人机。需要说明的是,此种干扰程度与地面用户数量正相关,严重时会导致无人机业务异常。测试表明,无人机对邻区用户干扰小于地面用户对邻区用户干扰,干扰强度相差约为10 dB。

针对上述挑战,本文面向网联无人机提出了低空网络解决方案,主要涉及下述3种技术:

1) 基于广播波束SSB的“1+X”技术方案。该方案将水平设置为一个宽波束,垂直方向设置为X个波束。在水平方向上,“1+X”技术方案采用增强型的宽波束,通过借用小于0.5%的业务功率实现SSB的6 dB功率增强,从而能够实现与水平多波束基本相当的覆盖。在垂直方向上,可以依托智能网规工具及高精电子地图按照特定场景,自适应地规划配置X波束的个数、宽度、倾角、方向等参数,从而提高垂直覆盖高度。也就是说,“1+X”方案在空域覆盖特性方面,以最少的水平SSB波束与按需垂直波束配置的方式,实现三维全空间覆盖;在控制SSB广播邻区干扰方面,通过时域轮询在相邻小区之间配置不同的SSB波束,以便错开邻区之间SSB波束的发送时间,从而降低邻区之间强干扰广播波束的时域空域碰撞概率,能够较好地解决邻区干扰问题。综上,“1+X”技术方案能够将传统单一宽波束转化为三维立体波束,增加垂直维度的自由度,实现空分复用能力,能适

应复杂环境下各种覆盖场景，为基于5G地面网络的对空覆盖奠定了基础。

2) 低空天线调优及邻区关系优化技术。对基站天线工参进行调优处理或者进行PCI邻区关系优化，有助于规避乒乓效应及切换失败等问题。例如，针对网联无人机飞行路线上的邻区关系，可通过准确配置切换迟滞、小区特定偏置等切换参数，并结合有序飞行路径覆盖优化邻区关系或局部无人机飞行线路，从而降低移动性对实时业务的负面影响。

3) 无人机公共网络切片保障技术。针对网联无人机业务差异化资源分配需求，引入端到端公共网络切片技术保障无人机业务质量。该切片由无线接入网、传输网、核心网协同构建，通过统一的端到端切片管理系统实现全域调度管控。网络切片可为无人机用户提供软硬件专属网络资源与定制化保障能力，精准匹配无人机业务相较于地面普通用户的个性化服务质量（QoS）需求，实现低空无人机业务的稳定、高效承载。

2 低空网络建设思路与建议

2.1 公网无人机规划指标建议

参照地面蜂窝覆盖规划标准，本文给出低空网络覆盖指标阈值：密集城区低空场景RSRP不低于-95 dBm、上行边缘速率不低于3 Mbit/s；郊区低空场景RSRP不低于-100 dBm、上行边缘速率为1~2 Mbit/s^[1]。

网联无人机高清视频回传业务以上行流量为主，上行速率需求显著高于下行。其中4K视频回传上行边缘速率建议不低于25 Mbit/s，测绘数据回传与8K高清视频回传上行速率需达到100 Mbit/s及以上^[2-3]。低空网络建设初期可将上下行链路25 Mbit/s作为基础承载标准，满足无人机常规业务运行；后续可依托终端性能迭代、局部基站软硬件升级，平滑演进至100 Mbit/s承载能力。各类典型应用场景对应的业务传输速率指标如表2所示。

2.2 低空网络管理平台体系建设

近年来，网联无人机产业高速发展，已广泛应用于个人消费、农林植保、地理测绘、环境监测、电力巡检、影视航拍等多个行业场景。然而，低空无人机具备数量规模大、单机体积小、任务场景复杂多变等特征，这些特征使得现有空域通信、导航定位等低空管控体系仍存在明显短板。随着中国低空空

域持续开放与无人机产业快速扩容，中国无人机适航管理制度与行业技术标准将持续完善，构建适配低空无人机规模化运行的全域监管体系，已成为低空产业规范化发展的必然趋势^[11,13-14]。

针对当前网联无人机通信业务存在的问题，本文提出一套低空网联无人机整体组网架构方案。整体架构如图6所示，该方案涵盖无线网至核心网的端到端组网逻辑，能够满足多类型低空业务场景的网络部署需求。在该架构下，网联无人机与地面控制终端依托现有地面5G公网完成业务数据回传与控制指令交互，并通过上层业务服务器实现多行业低空应用的业务加载与功能承载。

由图6可见，面向5G网联无人机的低空网络管控平台集成5G机载通信终端、5G地空传输网络、无人机云服务平台、地面控制站及行业业务终端等核心功能单元，可为低空智联网业务提供端到端的网络承载与业务赋能，有效支撑管线巡检、城市安防、生态环保、河湖水域监测、森林消防、低空空域管控等各类低空行业应用。未来，5G网联无人机低空网络管控系统将搭建完备的“端、管、云”一体化架

表2 不同场景业务下的传输速率建议指标^[10-11]

应用场景	业务	上行速率/ (Mbit·s ⁻¹)	下行速率/ (Mbit·s ⁻¹)	覆盖范围
物流	遥测飞行	0.2	0.3	城区、城郊、农村
	高清视频传输	25	0.3	
农业	喷洒作业	0.3	0.3	城郊、农村
	土地勘测	20	0.3	
巡检与 安防	4K视频	25	0.3	巡检覆盖基础设施, 安防覆盖城市
测绘	测绘数据回传	100	0.3	城区、城郊、农村
直播	4K视频	25	0.6	城区、热点区域
	8K视频	100	0.6	

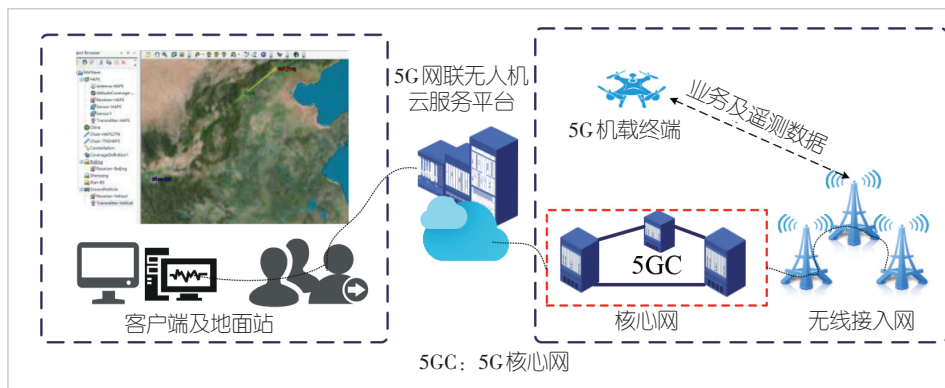


图6 网联无人机组网架构示意图

构,实现低空无人机设备的智能化、自动化全域管控,系统整体功能架构如图7所示。

由图7可知,面向5G网联无人机的低空网络监管系统在功能上分为终端、传输管道、云端三大模块。该系统依托三大模块提供飞行状态监测、数据存储、权限认证、数据调用等基础服务。

低空网络监管平台以无人机为空中业务载体,赋能各类垂直行业。平台对网络资源实施切片处理,针对不同业务提供差异化网络保障,满足各类业务在传输速率、通信时延、定位精度、信息安全等方面的指标需求。平台在云端完成行业应用软件即服务(SaaS)部署与业务能力平台即服务(PaaS)封装,构建无人机“端、管、云”一体化运营服务体系^[1,13-14]。未来,依托5G增强移动宽带(eMBB)、超高可靠超低时延通信(URLLC)、海量机器类通信(mMTC)等面向ToB/ToC的融合通信服务特性,边缘计算能力以及网络切片能力,可以实现网联无人机的用户体验升级和监管系统的自动化,助力无人机产业安全高效发展。

2.3 基于地面公网的低空网络发展建议

5G网联无人机可承载地理测绘、航拍直播、智慧农林、安防巡检、应急通信等多元化业务,是培育数字经济新兴产业的重要载体。未来,可从以下3个方面着手构建面向5G网联无人机的生态体系^[1,9]:

1) 加快民用无人机行业应用基础设施、服务保障体系建设,建立技术应用交流平台、新技术演示验证中心等,推进民用无人机在农林植保、地理测绘、环境监测、应急救援等行业领域创新应用。

2) 推进人工智能在民用无人机领域融合应用,加快提

高民用无人机娱乐性及智能作业水平,支持开发多样化衍生产品和服务。

3) 拓展民用无人机在文化、教育等领域的融合应用,利用“互联网+”开展行业应用创新,发展公共服务供应商,拓展服务领域和价值。

考虑到现阶段仍存在产业发展不均衡的问题,中国无人机行业应当分阶段逐步建立统一技术体制的面向5G网联无人机的低空智联网络。构建面向5G网联无人机低空智联网络,应以地面网络为基础,依托5G网联无人机,打造泛在低空智联网络,实现多种异构网络深度融合。分阶段逐步建立统一技术体制泛在低空通信网络的步骤可概括为:

第1阶段是互联互通与业务融合阶段:搭建一套支持随遇接入、按需调度、安全可信的标准化5G网联无人机低空通信体系,保障用户借助单套多模终端完成低空网络与地面蜂窝网络的一体化接入,进一步拓宽低空网络的接入用户规模与承载业务范围。

第2阶段是混合接入与系统深度融合阶段:构建能够支撑用户随遇接入、支持小区无缝切换的一体化无线网络,使整套网络具备连续稳定通信、高可靠安全传输、区域大容量并发承载、高机动数据回传等综合通信能力,最终完成适配5G网联无人机的完整低空智联网络体系搭建。

3 总结与展望

本文阐述了网联无人机的发展现状及趋势,分析了基于地面公网的低空网络覆盖性能及挑战,探索了面向网联无人机的低空网络的发展思路与解决方案。考虑到国家政策、空域开放程度、应用成熟度与建网成本等因素,对于300 m以下低空区域可基于地面5G公网实现对空连续覆盖;

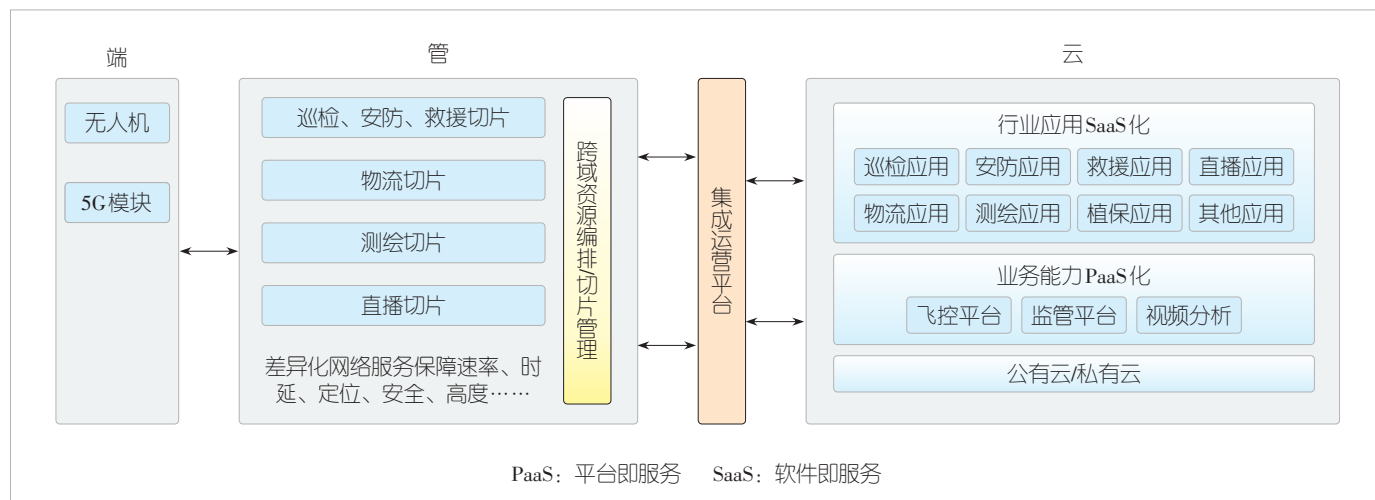


图7 面向5G网联无人机的低空网络“端、管、云”架构示意图

对于300~3 000 m高度的低空区域,其连续覆盖特性尚待进一步验证,或可采用地面对空覆盖专网的建网方式或基于卫星通信的空天地一体化组网方式。现阶段,基于地面公网的低空覆盖网络主要以地面基站的天线旁瓣对空覆盖方式为主。基于地面5G公网的低空网络对空覆盖特性的分析仿真及实测结果表明,地面宏基站对空通信终端的覆盖能力在300 m以下空域内并非全弱于其对地面终端的覆盖能力,且具备连续覆盖条件。需要说明的是,相对于地面通信终端而言,对空通信终端的下行同频干扰也很严重,且其受干扰程度与地面基站的复杂程度、站址密集度正相关。就解决方案而言,基于地面5G网络新空口技术将成为低空通信关键技术,可以采用“1+X”垂直覆盖的技术优化对空覆盖特性,采用三域协同端到端网络切片的方式保障无人机的QoS需求。在网联无人机监管与运营方面,应当以无人机业务能力的原子化、云化,以及产业生态极简调用为目标,构建“端-管-云”架构的无人机监管系统。

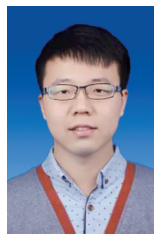
面向5G网联无人机的低空网络应以5G网联无人机为载体,依托地面5G公网的对空覆盖能力,实现对地面区域与低空区域的立体覆盖,同时拓展公共通信业务、政企业务,为客户提供差异化的最优质5G低空通信服务,实现业务体验与系统效率的同步提升。未来,打造“泛在、智能、协同”的低空智联网络还需加速投入技术创新,推动5G无人机业务及产业生态建立,实现“5G+无人机”的方式赋能千行百业,助力各行业产业发展,对打造数字经济新兴产业具有重要意义。

参考文献

- [1] 韩玲,朱雪田,迟永生.基于5G的低空网联无人机体系研究与应用探讨[J].电子技术应用,2021,47(5):1-4+10
- [2] 中国联合网络通信有限公司.中国联通空天地一体化通信网络白皮书[R].2020
- [3] 中国移动通信集团有限公司.基于5G通信技术的无人机立体覆盖网络白皮书[R].2021
- [4] 孙一,吴威,潘文苹,等.5G低空网络部署研究[J].通信电源技术,2020,37(17):183-185. DOI: 10.19399/j.cnki.tpt.2020.17.060
- [5] 刘超,陆璐,王硕,等.面向空天地一体多接入的融合6G网络架构展望[J].移动通信,2020,44(6):116-120
- [6] 林彬,张治强,韩晓玲,等.“空天地海”一体化的海上应急通信网络技术综述[J].移动通信,2020,44(9):19-26
- [7] 李福昌,裴郁杉,王俊杰,等.基于“7·20郑州特大暴雨”对空天地一体化通信的思考[J].移动通信,2021,45(5):81-85
- [8] 中国联通.B5G通信关键技术与发展愿景[EB/OL].(2019-12-30)[2024-10-25].<https://max.book118.com/html/2019/1230/5124023233002213.shtm>

- [9] IMT-2020(5G)推进组.5G无人机应用白皮书[R].2018
- [10] Vanelli-Coralli A, Guidotti A, Foggi T, et al. 5G and beyond 5G non-terrestrial networks: trends and research challenges [C]// Proceedings of IEEE 3rd 5G World Forum (5GWF). IEEE, 2020: 163-169. DOI: 10.1109/5gwf49715.2020.9221119
- [11] 中国联合网络通信有限公司研究院.中国联通6G白皮书(V1.0)[R].2021
- [12] 3GPP TR 38.811: 2020 Technical specification group radio access network; study on new radio (NR) to support non-terrestrial networks (Release 15) [S]
- [13] 张勃,冯毅,马丹,等.5G低空网络解决方案和运营应用[J].电信科学,2020,36(1):28-33
- [14] 中华人民共和国中央人民政府.无人驾驶航空器飞行管理暂行条例[EB/OL].(2023-06-28)[2024-10-25].https://www.gov.cn/zhengce/content/202306/content_6888799.htm

作者简介



王俊杰,中国联合网络通信有限公司工程师;主要研究方向为空地一体化通信网络融合、泛低空通信网络与多目标优化算法等。



王迪,中讯邮电咨询设计院有限公司工程师;主要研究方向为网络大数据分析、数字化运营、云网一体化行业市场创新产品研发。



申奇,中国联合网络通信有限公司智能城市研究院工程师;主要研究方向为智慧城市顶层设计及规划咨询。



肖腾飞,中国联合网络通信有限公司工程师;主要从事国家宏观战略和通信产业政策研究。

中兴通讯技术杂志社

促进产学研合作青年专家委员会

主 任 陈 为 (北京交通大学)

副主任 秦晓琦 (北京邮电大学) 卢 丹 (中兴通讯股份有限公司)

委 员

曹 进	西安电子科技大学	史颖欢	南京大学
陈 力	中国科学技术大学	唐万恺	东南大学
陈 为	北京交通大学	王景璟	北京航空航天大学
陈琪美	武汉大学	王兴刚	华中科技大学
陈舒怡	哈尔滨工业大学	王勇强	天津大学
陈思衡	上海交通大学	温森文	华南理工大学
高 镇	北京理工大学	吴泳澎	上海交通大学
官 科	北京交通大学	武庆庆	上海交通大学
韩 充	上海交通大学	夏文超	南京邮电大学
韩凯峰	中国信息通信研究院	向路平	南京大学
何 姿	南京理工大学	徐梦炜	北京邮电大学
侯天为	北京邮电大学	徐天衡	中国科学院上海高等研究院
胡 杰	电子科技大学	杨川川	北京大学
黄 晨	紫金山实验室	叶迎晖	西安邮电大学
霍佳皓	北京科技大学	尹海帆	华中科技大学
李 昂	西安交通大学	游昌盛	南方科技大学
李 礼	中国科学技术大学	于季弘	北京理工大学
刘 凡	东南大学	张 娇	北京邮电大学
刘俊宇	西安电子科技大学	张宇超	北京邮电大学
卢 丹	中兴通讯股份有限公司	章嘉懿	北京交通大学
陆游游	清华大学	赵毅哲	电子科技大学
梅渭东	电子科技大学	赵昱达	浙江大学
宁兆龙	重庆邮电大学	赵中原	北京邮电大学
潘存华	东南大学	周 伊	西南交通大学
祁 亮	上海交通大学	朱秉诚	东南大学
秦晓琦	北京邮电大学	朱光旭	深圳市大数据研究院
秦志金	清华大学	朱政宇	郑州大学
史 瑶	哈尔滨工业大学 (深圳)		

刊物相关信息



投稿须知



投稿平台



过刊下载



论文索引与
引用指南

办刊宗旨：

以人为本，荟萃通信技术领域精英
迎接挑战，把握世界通信技术动态
立即行动，求解通信发展疑难课题
励精图治，促进民族信息产业崛起

产业顾问：

段向阳、高 音、胡留军、华新海、刘新阳、
史伟强、屠要峰、王会涛、熊先奎、许 进、
闫新成、赵亚军、朱晓光

双月刊 1995 年创刊
第 32 卷 总第 189 期
2026 年 6 月 第 3 期

主管：安徽出版集团有限责任公司
主办：时代出版传媒股份有限公司
深圳航天广宇工业有限公司
出版：安徽科学技术出版社
编辑、发行：中兴通讯技术杂志社

总编辑：王喜瑜
主编：陶善勇
执行主编：吴玲、黄新明
副主编：卢丹
编辑部主任：王萍萍
责任编辑：刘芬、徐烨
编辑：杨广西、朱莉、任溪溪
设计排版：徐莹
发行：王萍萍
编务：王坤

《中兴通讯技术》编辑部
地址：合肥市金寨路 329 号凯旋大厦 1201 室
邮编：230061
网址：tech.zte.com.cn
投稿平台：tech.zte.com.cn/submission
电子信箱：magazine@zte.com.cn
电话：(0551) 65533356

发行方式：自办发行
印刷：安徽添锦印刷科技有限公司
出版日期：2026 年 6 月 30 日
中国标准连续出版物号：ISSN 1009-6868
CN 34-1228/TN
定价：每册 20.00 元