



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

ISSN 1009-6868  
CN 34-1228/TN

# 中兴通讯技术

## ZTE TECHNOLOGY JOURNAL

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

第 31 卷 · 总第 185 期 · 2025 年 10 月 · 第 5 期

**专题：网络中的 AI 技术**



ISSN 1009-6868



9 771009 686250



# 《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	唐雄燕	中国联通研究院副院长
陈前斌	重庆邮电大学教授、副校长	陶小峰	北京邮电大学教授
段晓东	中国移动研究院副院长	汪烈军	新疆大学教授、副校长
方 榕	中兴通讯股份有限公司董事长	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	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	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	北京邮电大学教授
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、深圳大学校长	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
孙知信	南京邮电大学教授	钟章队	北京交通大学教授
谈振辉	北京交通大学教授	周 亮	南京邮电大学教授、副校长
唐 宏	中国电信 IP 领域首席专家	朱近康	中国科学技术大学教授
唐万斌	电子科技大学教授	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 31 卷 总第 185 期 2025 年 10 月 第 5 期

中文核心期刊 中国科技核心期刊 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 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 ▶

### 网络中的 AI 技术

01	专题导读	解冲锋, 孟洛明, 崔勇
03	大模型驱动的网络智能运营管理标准化和应用展望	李文璟, 方宏林, 喻鹏
11	大语言模型赋能智能网络的应用与挑战	牛嘉林, 邢铭哲, 张蕾
19	AI 智能体赋能网络运营的研究与应用	郑雨婷, 程新洲, 王静云
25	检索增强的网络流量预测方法	常远, 吴春鹏, 王峰
30	原生 AI 融合网络数字孪生赋能下一代无线网络自治	王首峰, 郭建超, 边森

## 名家视点 ▶

37	光纤通信技术演进与发展展望: 从基础突破到融合创新	张海懿
----	---------------------------	-----

## 企业视界 ▶

50	下一代 AI 大模型计算范式洞察	熊先奎, 王程晨, 蔡文豪
----	------------------	---------------

## 技术广角 ▶

57	星地一体化语义通信网络: 探索与展望	李东博, 王新宇, 尹志胜, 承楠, 刘劼
66	5G 基站节能面临的关键问题和解决方案	王小锋, 韩茜

## 综合信息 ▶

02	《中兴通讯技术》2026 年专题计划
24	中兴通讯技术杂志社第 30 次编委会议暨 2025 通信热点技术研讨会隆重召开

## 《中兴通讯技术》2025 年热点专题名称及策划人

### 1. 6G 立体覆盖技术

西安电子科技大学教授 李建东  
西安电子科技大学教授 刘俊宇

### 3. 6G 网络安全

北京航空航天大学教授 刘建伟  
北京航空航天大学教授 王景璟

### 5. 网络中的 AI 技术

中国电信新一代信息通信专业首席专家 解冲锋  
北京邮电大学教授 孟洛明  
清华大学教授 崔勇

### 2. 智算网络

中国移动研究院副院长 段晓东  
清华大学教授 李丹  
电子科技大学教授 虞红芳

### 4. 面向 6G 的高时效智能机器通信技术

中国工程院院士、北京邮电大学教授 张平  
北京邮电大学副教授 秦晓琦

### 6. 新一代光传输技术

上海交通大学教授 陈建平  
中国联通研究院副院长 唐雄燕

# MAIN CONTENTS

ZTE TECHNOLOGY JOURNAL  
Vol. 31 No. 5 Oct. 2025

Special Topic ►	<b>AI for Network</b>
	01 Editorial ..... XIE Chongfeng, MENG Luoming, CUI Yong
	03 Standardization and Application Prospects of Large Model Driven-Intelligent Network Op- eration and Management ..... LI Wenjing, FANG Honglin, YU Peng
	11 Applications and Challenges of Intelligent Networks Empowered by Large Language Models ..... NIU Jialin, XING Mingzhe, ZHANG Lei
	19 Research and Application of AI Agent Empowering Network Operations ..... ..... ZHENG Yuting, CHENG Xinzhou, WANG Jingyun
	25 Retrieval-Augmented Network Traffic Prediction Method..... ..... CHANG Yuan, WU Chunpeng, WANG Feng
	30 Native AI-Integrated Network Digital Twin for Empowering Next-Generation Wireless Net- work Autonomy ..... WANG Shoufeng, GUO Jianchao, BIAN Sen
Expert View ►	37 Evolution and Development Prospects of Optical Fiber Communication Technology: From Foundational Breakthroughs to Convergent Innovation..... ZHANG Haiyi
Enterprise View ►	50 Insights into Computational Paradigm of Next-Generation AI Large Model ..... ..... XIONG Xiankui, WANG Chengchen, CAI Wenhao
Research Papers ►	57 Semantic Communication for Satellite-Terrestrial Integrated Networks: Exploration and Prospects..... LI Dongbo, WANG Xinyu, YIN Zhisheng, CHENG Nan, LIU Jie
	66 Key Problems and Solutions of Energy-Saving for 5G Base Stations ..... ..... WANG Xiaofeng, HAN Qian

期刊基本参数: CN 34-1228/TN\*1995\*b\*16\*70\*zh\*P\*¥20.00\*6500\*10\*2025-10

敬告读者	本刊享有所有发表文章的版权, 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。
------	---



# 网络中的AI技术专题导读



## 专题策划人



解冲锋



孟洛明



崔勇

当前，以大语言模型和智能体为代表的人工智能（AI）技术发展迅猛，在网络运维自动化与智能化提升方面展现出巨大潜力，正深刻改变着网络运营管理模式。国际上，以国际电信联盟电信标准化部门（ITU-T）和互联网工程任务组（IETF）为代表的标准化组织已多次围绕“AI for Network”展开专题研讨，并启动相关标准化工作。然而，业界对于AI在网络运维中仍存在诸多疑问，例如：AI在网络中究竟能发挥哪些独特作用？其作用机制为何？目前已有哪些成功应用案例？尚存哪些关键问题亟需攻克？为此，本期以“网络中的AI技术”为主题，聚焦AI技术如何赋能网络运营，邀请该领域的专家学者撰写了5篇文章。这些文章介绍并分析了的最新关键技术进展，从多角度探讨如何在网络中应用AI技术，并对存在的问题和具体的解决方案进行了深入讨论。本专题旨在探索中国网络基础设施发展所需要的AI技术，为进一步的产业化和标准化铺平道路。

《大模型驱动的网络智能运营管理标准化和应用展望》在分析网络运营管理智能化需求的基础上，总结了相关大模型标准化进展，提出大模型驱动的智能运营管理架构，阐述其在自配置、自优化、自治愈等环节的关键技术与挑战，并

通过故障智能运维案例验证了其可行性与实效，最后从标准引领、价值落地与能力演进等角度对未来发展进行展望。

《大语言模型赋能智能网络的应用与挑战》提出基于大模型的智能网络管理框架，梳理其在关键领域的应用路径，分析其在提升决策效率、增强服务适配性和降低运维成本方面的优势，并探讨解空间组合爆炸与NP难问题、多维度不确定性、实时性约束、数据异构性及人机协同与成本平衡等技术挑战，同时总结了现有应对思路。

《AI智能体赋能网络运营的研究与应用》重点阐述智能体赋能网络运营的关键技术，列举共享网络融合规建、基于意图的网络运营服务等代表性应用，结合6G通感算一体与天地一体网络等前沿趋势，探讨智能体在未来网络运营中的发展方向，旨在构建意图驱动、闭环自优的智能化网络新范式。

《检索增强的网络流量预测方法》提出一种融合时序大模型与语言大模型的协同预测框架，实现基于变更事件的动态流量预测。针对变更事件稀疏性与专业语义理解难题，设计了基于检索增强生成的变更影响知识库，通过检索历史相似变更的流量影响特征，构建可解释的上下文提示。实验表明，该方法在含变更事件的预测场景中有效降低了误差。

《原生AI融合网络数字孪生赋能下一代无线网络自治》提出一种融合原生AI与网络数字孪生的一体化架构，涵盖

数据采集、模型构建与管理等关键环节，以及原生 AI 驱动的网络性能预测、AI 用例自生成与网络策略自定制等核心能力。该架构实现了“数据-模型-决策-验证”的内生智能闭环，为应对 6G 网络高复杂度与高动态性环境下的自治挑战提供了系统化支撑。

本期专题的作者来自知名高校、头部企业与科研机构，面向“AI for Network”，从新技术、需求、网络架构、新型技术、应用实践等方面介绍了最新研究成果。期待这些高质量的研究成果能够为基于 AI 的网络运营提供有益的参考和启示，并在此对所有作者和审稿专家的大力支持表示由衷的感谢！

#### 策 划 人 简 介

**解冲锋**，中国电信新一代信息通信专业首席专家，教授级高工，中国通信学会会士，中国互联网协会学术委员会副主任委员，北京市 IPv6 重点实验室主任，曾在美国 UCLA 大学做政府公派访问学者一年；长期从事宽带网络架构、IPv6 下一代互联网、物联网、网络安全、云网融合等方面的研究；曾获得 2023 年度

国家科技进步奖一等奖和 2023 年度中国通信标准化协会科学技术奖一等奖，2019 年获得“国务院政府特殊津贴”；参与制定 IETF RFC 标准 6 项。

**孟洛明**，北京邮电大学教授、《北京邮电大学学报》编委会主任、工业和信息化部通信科技委常委、中国通信标准化协会 TC7 主席、中国通信学会通信软件技术专业委员会名誉主任、国家“973”计划项目“可测可控可管的 IP 网的基础研究”的首席科学家、“长江学者”特聘教授、国家杰出青年科学基金资助获得者、“长江学者和创新团队发展计划”创新团队带头人、国家自然科学基金创新研究群体带头人，并获得国家级有突出贡献的中青年专家等称号；长期从事网络管理领域的科研和教学工作；研究成果获国家科技进步奖二等奖 2 次、中国标准创新贡献奖一等奖 1 次、省部级科学技术奖一等奖 5 次。

**崔勇**，清华大学长聘教授、网络技术研究所所长，教育部“长江学者”特聘教授，首届“青年长江学者”获得者，国家优秀青年基金和教育部新世纪人才获得者，中国互联网协会学术工作委员会秘书长，中国通信学会边缘计算委员会副主任委员，先后担任国际互联网标准化组织 IETF 工作组主席和 4 本 IEEE 期刊编委；主要研究方向为低时延传输技术、视频分析、内容安全、流媒体传输、网络数字孪生、网络 AI 等；先后获国家科技进步一等奖、国家发明二等奖、国家科技进步二等奖，多次获得国家信息产业重大技术发明；发表论文 100 余篇，获国家发明专利 40 余项，完成 RFC 国际标准 10 余项，出版学术著作 4 部。

## 综合信息

### 《中兴通讯技术》2026 年专题计划

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

# 大模型驱动的网络智能运营管理 标准化和应用展望



## Standardization and Application Prospects of Large Model Driven-Intelligent Network Operation and Management

李文璟/LI Wenjing, 方宏林/FANG Honglin, 喻鹏/YU Peng

(北京邮电大学, 中国 北京 100876)

(Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTETJ.202505002

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20250926.1514.008>

网络出版日期: 2025-09-26

收稿日期: 2025-08-16

**摘要:** 大模型的迅猛发展正在深刻变革网络运营管理方式, 推动自智网络从“外挂式智能”迈向“内生式智能”。聚焦大模型驱动的网络智能运营管理, 在分析网络运营管理智能化的发展需求基础上, 总结了网络运营管理大模型标准化进展。在提出大模型驱动的网络智能运营管理架构基础上, 阐述了大模型在网络自配置、自优化、自愈愈等过程的关键技术和挑战。对大模型在网络运营管理智能化中的应用和实例进行了验证, 展望了面向未来“标准引领、价值落地、能力演进”愿景的大模型运维体系, 可为实现真正智能自治的网络管理范式转型提供参考。

**关键词:** 网络智能运营管理; 大模型; 自智网络; 智能体

**Abstract:** The rapid development of large models is profoundly transforming the methods of network operation and management, driving autonomous networks to evolve from "external intelligence" to "embedded intelligence". This paper focuses on large model-driven intelligent network operation and management. Based on an analysis of the development needs for intelligent network operation and maintenance, it summarizes the standardization progress of large models in network operation and management. After proposing a large model-driven architecture for intelligent network operation and management, this paper elaborates on the key technologies and challenges in processes such as self-configuration, self-optimization, and self-healing in networks. Following this, this paper validates the application and examples of large models in the intelligentization of network operation and management. Finally, it envisions a future large model-based operation and maintenance system guided by "standard leadership, value realization, and capability evolution", providing a reference for the paradigm shift toward truly intelligent and autonomous network management.

**Keywords:** intelligent network operation and management; large model; autonomous network; intelligent agent

**引用格式:** 李文璟, 方宏林, 喻鹏. 大模型驱动的网络智能运营管理标准化和应用展望 [J]. 中兴通讯技术, 2025, 31(5): 3-10. DOI: 10.12142/ZTETJ.202505002

**Citation:** LI W J, FANG H L, YU P. Standardization and application prospects of large model driven intelligent network operation and management [J]. ZTE technology journal, 2025, 31(5): 3-10. DOI: 10.12142/ZTETJ.202505002

伴随网络技术的不断演进, 网络运营管理也经历了从人工、半自动到智能化的发展过程, 而智能内生驱动的自智网络成为未来网络运营的主要发展方向。大模型等新的人工智能方法在自然语言处理、多模态信息处理等领域取得了巨大成功。然而, 目前大模型在网络智能运营管理中仍然处于起步阶段。针对上述问题, 本文重点分析网络智能运营

管理需求, 提出大模型驱动的网络智能运营管理架构, 并针对网络运营管理生命周期中的关键技术和挑战进行总结, 为网络智能运营管理的发展提供技术参考依据。

## 1 网络运营管理智能化的发展与标准化需求

### 1.1 网络运营管理的智能化发展历程

网络运营管理的智能化发展深刻反映了通信网络技术与人工智能的融合历程。早期的网络运维主要依赖人工配置和经验判断, 存在响应慢、误判多的问题。随着网络规模的扩

**基金项目:** 国家自然科学基金项目 (U22B2031); 北京市自然科学基金-海淀原始创新联合基金项目 (L232045)

大，传统方式的劣势更加明显，难以应对复杂的运维挑战。

进入21世纪后，网络运营管理逐步迈向自动化。以故障检测、资源调度、配置下发为代表的自动化手段显著提升了运营效率。典型的例子是软件定义网络和网络功能虚拟化的引入，开启了以编程方式控制网络的新时代，为后续的智能化管理奠定了基础<sup>[1]</sup>。

近年来，随着大数据、云计算和人工智能的快速发展，网络运营管理进一步演进至智能阶段。人工智能（AI）技术开始应用于流量预测、异常检测、智能排障和策略优化等场景。例如，基于图神经网络的拓扑建模可实现更精准的网络状态感知<sup>[2]</sup>，强化学习技术已被用于动态路由与资源编排<sup>[3]</sup>等。这一阶段的特征是“数据驱动+知识增强”，推动了网络运营管理从“被动响应”向“主动决策”转变<sup>[4]</sup>。当前，大模型（如ChatGPT、DeepSeek、通义千问）在自然语言理解、知识推理与多模态感知等方面展现出强大能力，正推动网络运营管理从智能化迈向泛在智能化和自治化的新阶段<sup>[5]</sup>。这一趋势预示着：未来网络运营管理将具备更强的环境理解、自主判断与任务执行能力，自智网络将成为网络智能运营管理的未来发展方向。

## 1.2 自智网络智能化特征分析

自智网络以人工智能、大数据与云计算等先进技术为基础，赋予通信网络感知、分析、决策与执行等能力，进而实现“自配置、自优化、自诊断、自恢复”的高度自治目标<sup>[6]</sup>。自智网络旨在应对网络规模持续扩张与业务形态日趋复杂所带来的管理挑战，推动网络运维模式从人工主导向智能自主方向转型，从而实现管理范式的根本变革。相较于传统依赖规则和静态配置的运维方式，自智网络更加强调“意图驱动”与“闭环控制”的融合，以实现网络系统的动态适应与自演进，推动网络智能化迈入全新阶段，其目标架构如图1<sup>[7]</sup>所示。

自智网络的智能能力覆盖网络资源的自动编排、故障的根因识别与预测、服务质量保障以及策略的动态调优等多个维度<sup>[8]</sup>，其功能不再局限于通信承载，而是构建起一个具备泛在智能的协同系统，能够感知环境变化并实现自动调控。尤其在5G/6G发展背景下，自智网络已扩展至业务编排、服务保障与客户体验等全链条场景，推动网络从连接

工具向智能服务平台转变。

从生命周期视角，自智网络贯穿“规划、建设、维护、优化、运营”5个核心阶段。其中，在规划阶段，系统基于意图与历史数据自动生成最优网络拓扑与资源部署策略；建设阶段强调自动化部署与持续集成；维护阶段通过自感知与自愈机制实现快速响应与故障修复；优化阶段依托反馈数据与模型训练动态调整运行参数；运营阶段则以用户体验为中心，实现端到端的智能闭环管控，构建面向未来的高度自治通信体系。

## 1.3 大模型驱动的智能网络发展需求

在网络规模持续扩张与业务形态高度复杂的背景下，传统依赖规则与经验的网络运营管理方式逐渐难以满足高可靠、低时延和多样化业务场景下的智能化需求。大模型凭借强大的语义理解、知识抽取和跨任务泛化能力，为自智网络的发展提供了全新技术支撑，正逐步成为推动网络从自动化向自治化演进的关键引擎。

在产业界，大模型在自智网络中的应用实践不断深化。中国移动、中国电信、爱立信、诺基亚等企业，在智能客服、网络优化与自动运维等场景中积极部署多模态或垂类大模型<sup>[9]</sup>，显著提升了网络服务的智能水平；华为提出以大模

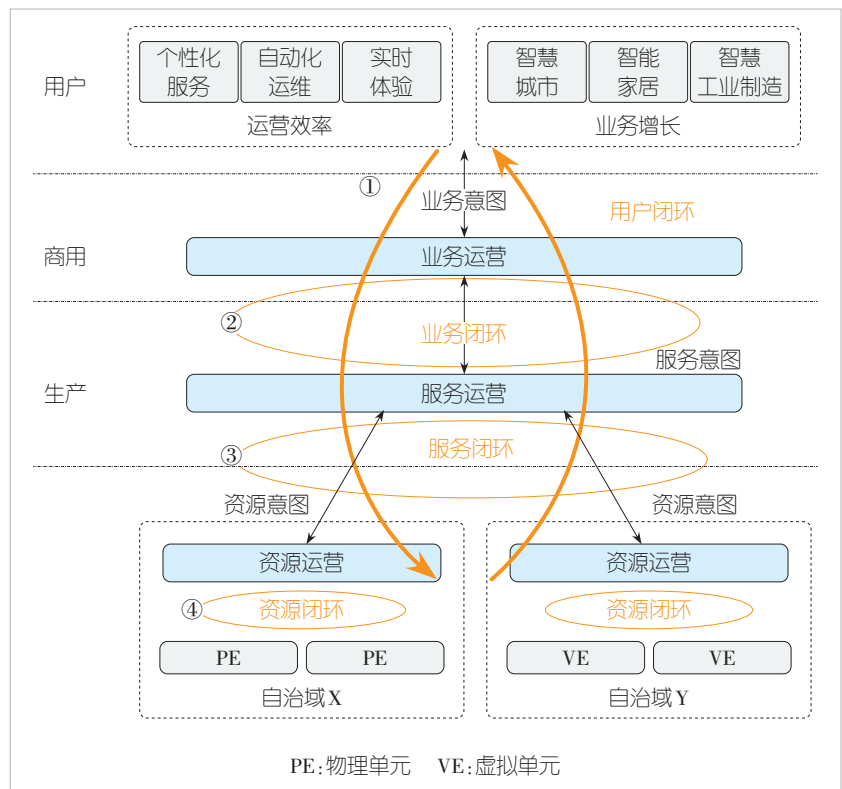


图1 TM Forum的自智网络框架<sup>[7]</sup>



型驱动的 L4 级自智网络架构，构建统一智能中枢以支撑意图识别、根因分析与策略编排等任务，为实现 L4 级别高阶自智目标提供了思路。大模型正从“工具型 AI”向“平台型 AI”演进，成为嵌入式智能控制系统的核心。

学术界则围绕大模型在网络场景中的适配性展开深入研究，聚焦于自然语言意图驱动的网络控制框架、时序行为建模、多模态数据融合与小样本自适应能力提升等方向。一些研究尝试将生成式 AI 结构引入网络策略生成与性能预测任务，提升泛化与协同能力<sup>[10]</sup>；同时也关注模型可解释性、可控性及其在资源受限条件下的压缩与部署问题<sup>[11]</sup>。然而，如何实现高性能与高时效的统一仍是当前研究的技术瓶颈。

## 2 网络运营管理大模型标准化研究

目前，中国通信企业纷纷推出网络运营管理领域的大模型，如中国电信启明大模型、中国联通元景大模型、中国移动九天众擎基座大模型、中兴通讯星云大模型、华为通信大模型等。然而，业界对网络运营管理大模型还缺乏一致的理解，需要进行标准化，为大模型在网络运营管理领域的应用实践提供规范。

中国通信标准化协会（CCSA）开展了“网络运营管理大模型”系列标准的研制工作。该系列标准包括网络运营管

理大模型架构、相关系统、系统间接口、系统对外提供的服务、应用场景与流程、测试与评估以及关键技术和应用等方面。图 2 所示为网络运营管理大模型标准体系规划图。

其中，《网络运营管理大模型总体技术要求》是系列标准中的基础标准，由我们团队提出，规定了网络运营管理大模型的概念、应用基本过程以及架构等。图 3 所示为大模型在网络运营管理领域的应用基本过程。

如图 3 所示，整体流程包括应用需求提出、应用构建与运行、模型训练与发布、模型部署与推理等过程，数据工程则完成各过程中所需数据的准备和预处理等。基于该基本过程，后续标准仍在持续制定过程中。

## 3 大模型在网络运营管理中的挑战 and 关键技术

### 3.1 大模型驱动的网络智能运营管理总体架构

基于上述分析，本文提出了大模型驱动的网络智能运营管理架构，如图 4 所示，在网络多维数据和网络智能体的支撑下实现智能化运营管理应用。

网络运维数据维度多元，涵盖时序流量-时延数据、任务部署参数、历史故障日志等。由于数据格式与特征差异显著，大模型需高效整合多源异构数据，既要应对时序数据动

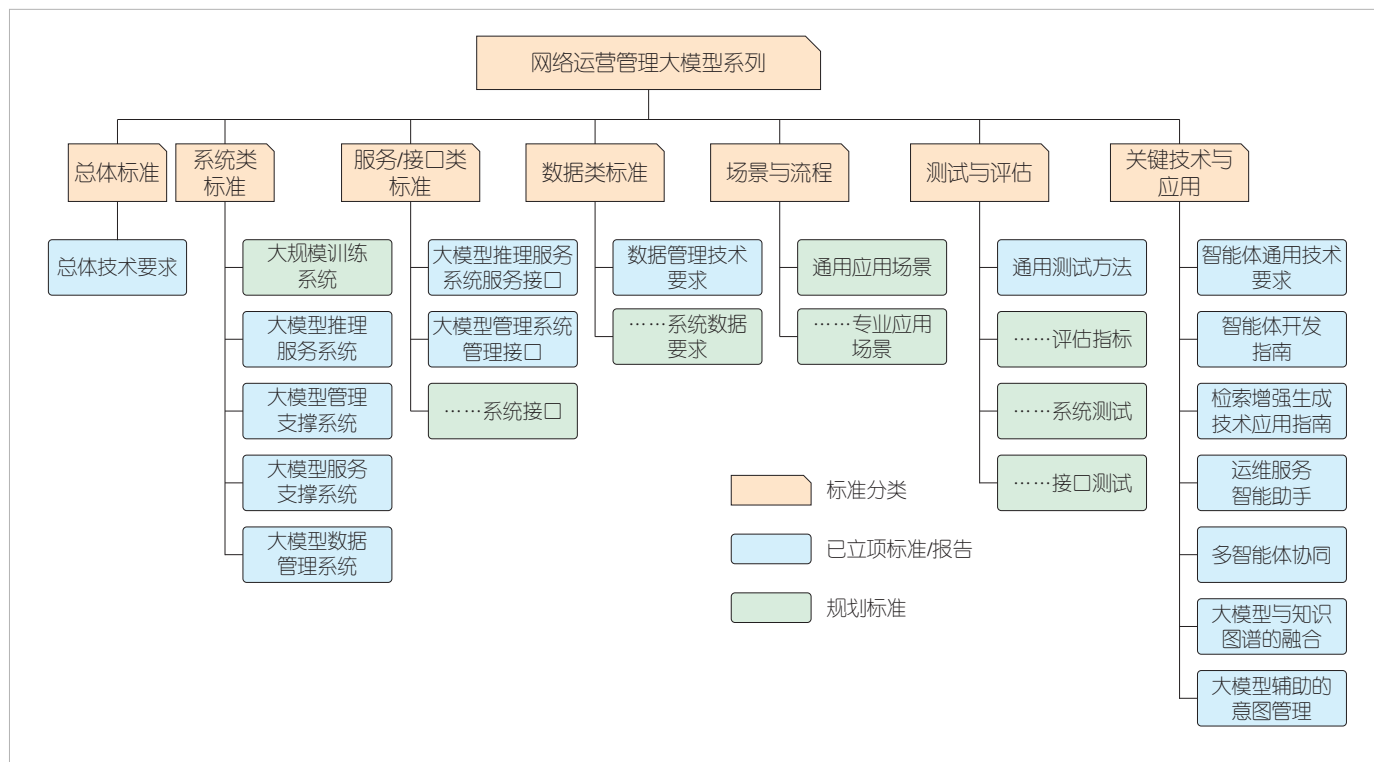


图2 网络运营管理大模型标准体系规划图

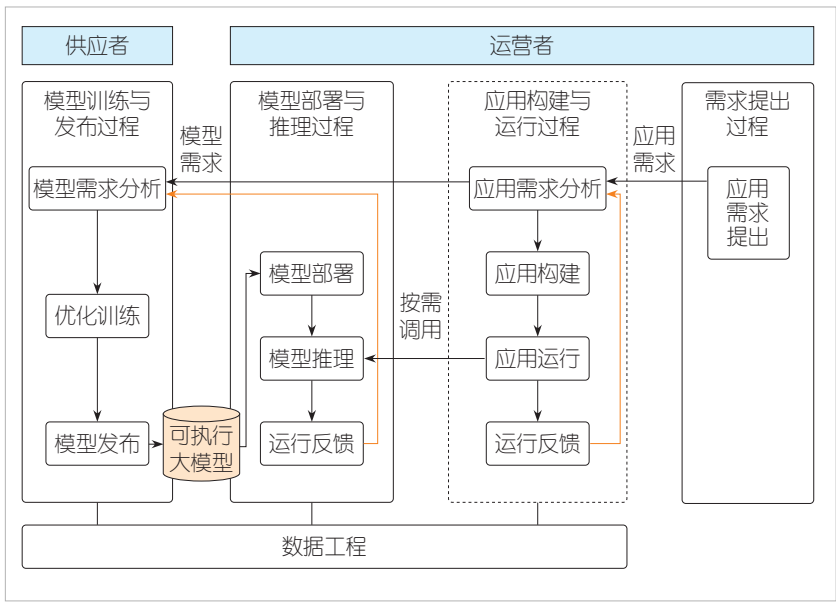


图3 大模型在网络运营管理领域的应用基本过程

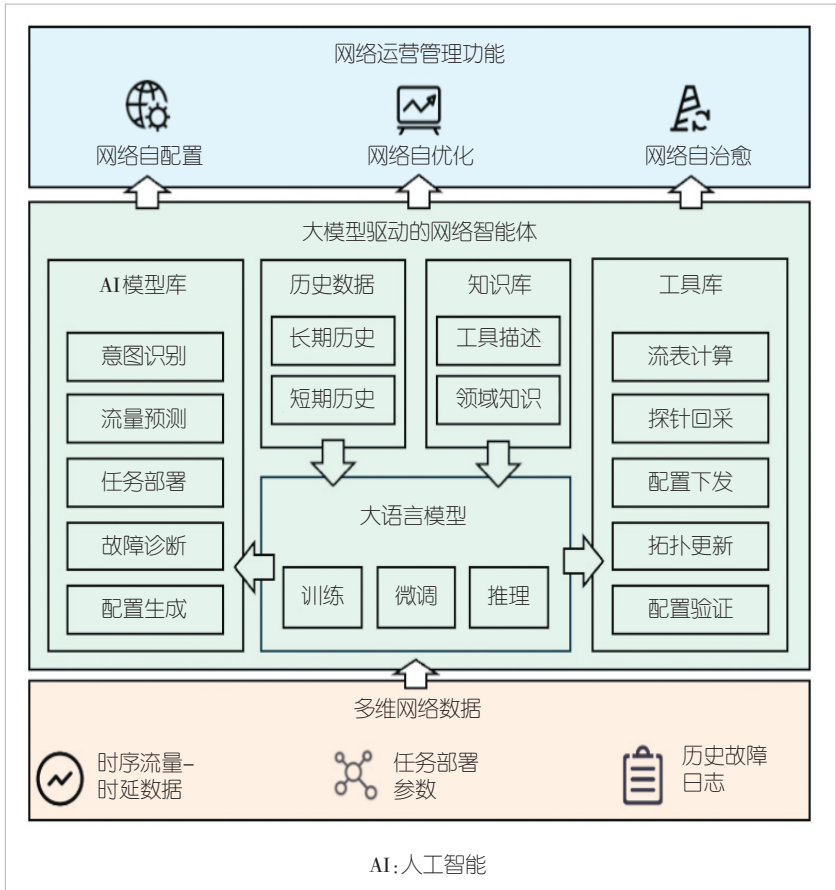


图4 大模型驱动的网络智能运营管理总体架构图

态变化，又要挖掘故障日志潜在关联信息，数据融合与预处理难度颇高。同时，网络运维有工具描述、网络协议等专属领域知识，大模型需融合长短期历史数据中的经验与领域知

识。然而，知识表示形式多样，因此需要构建统一的知识体系，以支撑模型对跨类型知识的精准理解、有效运用和沉淀复用，这一过程具有挑战性。

大模型驱动的网络智能体是运营管理的“智能中枢”，由AI模型库、历史数据、知识库和工具库协同支撑。AI模型库借助意图识别、流量预测等模型，实现智能分析能力；历史数据分为长短两个周期，为训练决策提供经验；知识库整合工具描述与领域知识，构建运维知识体系；工具库含流表计算、探针、配置下发等工具，衔接决策执行。大模型为智能体核心，其运转逻辑清晰体现了各构建模块间的紧密交互：

1) 数据感知与输入：首先，架构底层从网络环境中采集多维数据，如时序流量、设备告警和配置日志等，并将其输入到网络智能体中。

2) 智能体内部处理与决策：接收到数据后，作为核心的大模型开始工作。它会调用AI模型库中的特定模型（如流量预测模型）对数据进行初步分析，同时查询历史数据以寻找相似模式或历史经验。为了深刻理解运维任务，大模型会借助知识库中的网络协议、运维手册等领域知识进行推理，并利用工具库中定义的工具（如探针、配置下发）及其使用说明，评估可行的操作选项。经过这一系列融合分析，智能体最终形成一个具体的运维决策或方案，例如生成一个配置脚本或一条优化指令。

3) 应用功能实现与执行：最后，智能体输出的决策方案被上层的运营管理应用所调用。例如，在“自配置”应用中，该方案可能是一系列配置命令，由应用模块通过工具库中的下发工具执行到网络设备上；在“自愈”应用中，该方案则可能是一个故障恢复流程，指导系统完成故障隔离与业务恢复。通过这个“数据输入-智能决策-应用执行”的闭环，架构实现了高效的智能运营管理。

依托该智能体，可实现网络自配置、自优化、自愈等运营管理应用功能。自配置借助意图识别、配置生成，解析业务需求并自动下发配置；自优化通过流量预测、拓扑更

新,实时分析状态,动态优化资源与拓扑来提升性能;自愈通过故障诊断、配置验证,精准识别故障并执行自愈策略,保障网络稳定运行。网络自配置、自优化、自愈等功能,要求大模型驱动的多智能体间协同运作,不同功能对模型输出需求不同。如何让模型在多任务、多功能间灵活切换、高效协同,保障运维全流程顺畅是研究难点。

为应对这一挑战,我们需聚焦关键技术。大模型在网络智能运维过程中,依赖训练、微调与推理关键技术。由于运维数据具有时序性、多维度特征,训练环节要突破异构融合难题。多模态训练框架为此提供了可行路径,在构建时借助注意力机制强化数据交互,助力模型捕捉网络运行规律。微调环节利用故障日志、配置参数等专属数据微调通用大模型以适配场景,核心是设计高效参数调整策略,如结合检索增强生成(RAG)技术,先从外部知识库检索故障分析信息,再由生成模型构建内容,弥补领域知识的不足,提升对运维流程关系的理解与分析准确性<sup>[12]</sup>。推理环节兼顾实时性与准确性,其中优化算法是关键。具体可采用去中心化架构,通过分布式设备实现资源池化;在传输过程中,以激活值替代参数频繁加载/卸载提升速度,同时借助双注意力缓存机制增强可靠性<sup>[13]</sup>。

### 3.2 大模型驱动的网络自配置挑战及关键技术

大模型因其强大的语义理解、知识推理和代码生成能力,被广泛用于构建面向业务意图的网络“零接触”(Zero-Touch)配置体系。这种技术路线有望彻底改变传统依赖人工命令行接口或静态策略模板的网络配置模式,实现从自然语言需求到设备配置的端到端自动化生成。然而,网络自配置作为典型的任务与环境强绑定的场景,将大模型引入该体系仍面临多重挑战。

#### 挑战1: 意图理解与配置映射精度提升

网络管理人员通常以自然语言或结构化意图表达业务需求。大模型需要准确解析这些复杂且高度专业的表达,自动转换为符合具体设备语法和语义规范的配置命令。然而,由于意图表达的多样性和语义模糊性,加之网络设备厂商和协议标准的多样化,直接使用通用大模型会导致配置命令语法错误或逻辑不匹配的情况。

#### 挑战2: 拓扑感知与全局一致性配置

实际网络配置是一个全局联动行为,涉及设备间的依赖关系与策略协同。缺乏拓扑结构感知,可能导致大模型在生成配置指令时出现上下游设备不一致、路径重叠、策略冲突等问题。

#### 挑战3: 配置验证与事务安全保障

错误配置不仅会影响网络性能,更可能引发严重的安全漏洞和服务中断。因此,在大模型自动生成配置后,必须进行严格的语义校验和行为仿真。

面对上述挑战,研究者们已经探索了许多针对性的解决方案。

针对挑战1: RAG结合网络知识图谱、YANG模型等结构化网络信息,作为大模型的外部知识库,辅助模型在生成配置时调用相关上下文和约束条件,从而显著增强配置语义的一致性和结构化准确率。此外,还需要对大模型进行网络领域指令微调(如对配置数据进行监督微调、强化学习训练),显著提高特定协议语法的支持能力。

针对挑战2: 目前主要通过图结构建模将网络拓扑编码为结构化图数据,结合图神经网络或结构化注意力机制嵌入到大模型的输入表示中,增强模型对设备间依赖关系的理解和推理能力。这种融合方式使得模型不仅基于业务意图,还能根据当前网络拓扑动态生成协调一致的配置方案。

针对挑战3: 当前普遍采用仿真沙箱环境对配置网络行为进行预演,评估配置生成结果的影响,提前发现潜在风险。此外,“人机协同审查”模式也被广泛使用,即配置由模型生成初稿,再由运维人员结合多模型投票或规则引擎审核后才下发。同时,配置下发过程需支持事务化操作与快速回滚机制,若新配置方案出现异常,系统能自动回退至稳定版本,保障网络运行连续性。

大模型驱动的网络自配置是一项涉及语义理解、拓扑感知、安全验证与资源优化的复杂系统工程。尽管当前在意图解析、拓扑建模和安全审计等方面已有初步突破,但要实现真正高效、可靠且具备广泛适用性的“零接触”自动配置系统,仍需在模型可解释性、跨域协调和多层安全保障上持续深入研究。

### 3.3 大模型驱动的网络自优化挑战及关键技术

大模型驱动的网络管理自优化是B5G和未来6G网络的核心方向,旨在通过AI实现网络状态的实时感知、决策与调优。然而,其发展面临两大挑战:

1) 在训练方面,电信网络涉及大量复杂概念,如网络协议、路由算法、网络拓扑等。因此,要使大模型能够理解和推理这些概念,需要借助强大的训练策略。未来研究应着力开发降低幻觉并提升模型输出实际准确性的方法。

2) 在部署方面,网络自优化任务对实时性与资源要求极高。例如,在工业控制或车联网等确定性场景中,需在毫秒级内动态调整资源与路由以保障超低时延和高可靠性。在这种场景下,依赖云端大模型会因链路时延过高而无法满



需求,若将大模型直接部署于边缘,则会面临算力与存储的约束。为此,研究应聚焦云边端协同架构,并结合模型压缩与知识蒸馏等技术,有助于实现低时延、高性能与资源开销间的平衡。

对应于上述挑战,网络自优化所涉及的关键技术研究也分为两个方面。

针对训练挑战,需要构建高质量的通信领域大规模数据集。充分的通信领域相关数据集是训练通信大模型的先决条件。与可以利用互联网上大规模文本语料库的通用领域大模型不同,获得专门针对通信网络的相当大的数据集具有挑战性。现有研究通常专注于一个特定任务,然后构建相应的数据集。一个全面的大规模数据集应该包括网络相关的文档、标准规范、协议、教科书、研究论文和其他相关来源等。

针对部署挑战,边缘计算与模型轻量化成为关键技术路径。具体而言,可以采用云-边-端协同的混合部署架构,将通用或全局性的大模型部署在云端,负责长周期、非实时的全局优化策略生成;同时,在靠近网络设备的边缘节点部署经过模型压缩或知识蒸馏的轻量级模型。这些轻量级模型虽然规模较小,但继承了云端大模型的关键知识,能够基于本地数据进行快速、实时的推理和决策,从而在满足毫秒级时延要求的同时,实现精准的网络自优化。

最后,大模型丰富的现实世界知识将有助于网络优化算法建模和设计,降低基于机器学习(ML)的网络优化的训练和微调难度。具体而言,我们可以使用大模型进行强化学习的奖励函数设计,或者将大模型视为代理,与环境进行交互以探索最优策略;也可以使用大模型帮助凸优化问题建模,放松或去除一些不可行的约束。此外,大模型还可以为他们的决策提供依据和解释,这种能力对于理解电信网络等复杂系统至关重要。

### 3.4 大模型驱动的网络自治愈挑战及关键技术

网络自治愈具体包括故障自诊断和网络自恢复两个方面:

#### 1) 大模型驱动的网络故障自诊断挑战和关键技术

在自智网络的发展背景下,网络自诊断能力作为实现高等级自治的关键能力,正经历由传统方法向智能范式的深度演进。以往依赖规则库和人工经验的诊断方式在处理复杂、多变、跨层级的问题时已暴露出显著局限,难以满足网络规模扩展、业务动态调整与运行状态多维演化的诊断需求。随着大模型技术的快速发展,其在语义理解、知识整合与跨模态推理等方面展现出强大能力,为网络自诊断任务提供了新的技术路径。然而,将大模型应用于网络自诊断并非通过直

接替代传统方案就能实现。

#### (1) 数据质量待提升

当前网络数据呈现出显著的异构性和非结构性特征,覆盖告警、日志、关键绩效指标(KPI)、配置文件等多种类型,且多数数据缺乏标注,存在不平衡分布和高噪声问题。这使得模型训练存在数据质量难以保障的问题,严重影响其泛化能力和稳定性。

#### (2) 可解释性缺失

大模型的“黑盒”属性在诊断任务中带来障碍。网络运营管理系统往往对推理过程的逻辑链条具有强需求,以确保诊断结果的可验证性与操作可控性。而当前大模型生成的诊断建议往往缺乏清晰的因果推理依据,难以直接支撑高风险场景下的闭环控制。

针对上述数据挑战,亟需构建多模态、高质量的训练语料体系,并探索小样本学习、自监督预训练等机制来提升模型的适应性和诊断准确性。此外,针对可解释性缺失的挑战,需要引入可解释人工智能技术,如图因果图建模、图神经网络推理等,可为诊断过程提供结构化支撑,增强模型输出的透明度与可信度。

#### 2) 大模型驱动的网络自恢复挑战及关键技术

在大模型帮助下,网络故障恢复范式有望从“感知响应”迈向“主动免疫”,但同样面临两大挑战。

#### (1) 数据质量与隐私约束

大模型需要高质量、大规模的多源数据,而实际网络中的数据常常存在噪声干扰、信息缺失或异构性等问题,造成语义解析偏差。在联邦云或隐私保护场景,集中式数据访问难以实现,易形成数据孤岛,难以实现跨域的故障诊断和恢复。

#### (2) 策略泛化性与复杂性存在矛盾

网络故障非线性传播增加恢复策略的设计难度。大模型抽象的动作空间若过于简化,则无法覆盖复杂场景;若过于细化,则会导致强化学习探索空间爆炸。此外,大模型与深度强化学习的结合可能带来过高计算开销,不利于在资源受限环境中部署。

针对数据问题,需要研究多源数据语义解析与统一表征技术。借助大模型跨模态嵌入能力,将非结构化日志、时序资源指标和多维告警信息转化为统一语义向量;通过深层注意力机制提取关键特征,捕捉故障语义关联性,提升故障模式识别精度。

针对恢复策略的制定问题,需要研究大模型与深度强化学习融合的策略优化技术,构建“语义解析-策略优化”两阶段架构。大模型负责故障语义理解,深度强化学习通过分



层动作空间建模，学习故障类型与恢复动作的动态匹配，优化恢复策略效率。此外，记忆增强与持续学习机制也值得探索，可引入记忆引导的元控制器，通过存储高价值故障轨迹并基于时序差分（TD）误差采样，增强对罕见故障的学习；同时结合大模型提示微调策略，引导模型聚焦特定语义模式，实现对新型故障的快速适配，进而避免灾难性遗忘。

#### 4 大模型在网络智能运营管理中的应用和实例分析

为展示大模型在网络运营管理中的具体应用，本章将聚焦于第3.4节所探讨的“网络自愈”功能，通过一个完整的故障智能运维实例，验证其可行性与效果。该实例完整覆盖了从故障监测、智能诊断到自动恢复的全过程。

图5展示了一个大模型驱动的网络自愈实例，其技术框架以Ryu控制器为中枢，协同网络仿真环境（Mininet）与大模型智能体（LLM Agent），构建了智能运维闭环。下方的流量曲线直观地验证了该架构的有效性：在“故障注入”导致网络性能骤降后，系统能够自动完成诊断与恢复，在约10 s内使业务流量恢复至稳定状态，其具体的工作流程如下：

第一步：自监测环节。我们首先利用Mininet搭建网络仿真实验环境，并借助基于iperf的流量生成器在网络中注入业务流量。在此期间，系统的拓扑感知和流量监测模块会持续采集KPI，如链路时延、吞吐量等，并将网络状态与操作信息实时同步记录到日志中。这一环节实现了对网络运行状态的动态洞察，是后续诊断与恢复的基础。

第二步：自诊断环节。当监测到网络异常时（例如，我们手动注入一个大流量模拟链路拥塞故障），自诊断流程被

激活。部署在系统中的大模型智能体作为“智能中枢”，开始分析从自监测环节获取的实时数据和历史日志。通过对比正常与异常状态下的流延迟分布等特征，大模型能够运用其强大的推理能力，精准定位故障节点，并判断故障类型，例如是拓扑连接中断还是流量拥塞。这一步如同为网络故障进行“精准画像”，为后续的恢复提供了明确指引。

第三步：自恢复环节。在大模型智能体完成诊断后，系统会基于诊断结论自动执行恢复策略。例如，智能体判断为拓扑故障后，会调用网络控制器（如Ryu），下发指令调整网络路由或隔离故障设备，使业务流量绕开故障点。从实验结果来看，在故障注入后，网络流速会瞬间下降，但自恢复机制能够迅速介入，在约20 s内将网络性能恢复至正常水平并保持稳定。

综上，该实例通过“自监测-自诊断-自恢复”的无缝衔接，验证了大模型驱动下网络智能运维闭环的可行性与高效性，展示了其在提升网络运营自动化水平、降低故障处置时限方面的巨大潜力。

#### 5 网络运营管理的未来发展展望

随着运营商对网络智能化升级需求的持续深化，大模型技术凭借其多维度优势，正加速融入各通信专业域的运营管理与维护场景，成为推动网络智能化向L4+高阶自治演进的关键驱动力。当前，产业界聚焦三大核心突破方向：技术标准引领，针对大模型应用架构、场景需求定义、部署方案设计、智能体及多智能体协同等关键环节，建立统一技术规范以达成业界共识；高价值场景落地，优先布局人工依赖度高、智能化瓶颈显著、智能化需求迫切的高价值场景，从实

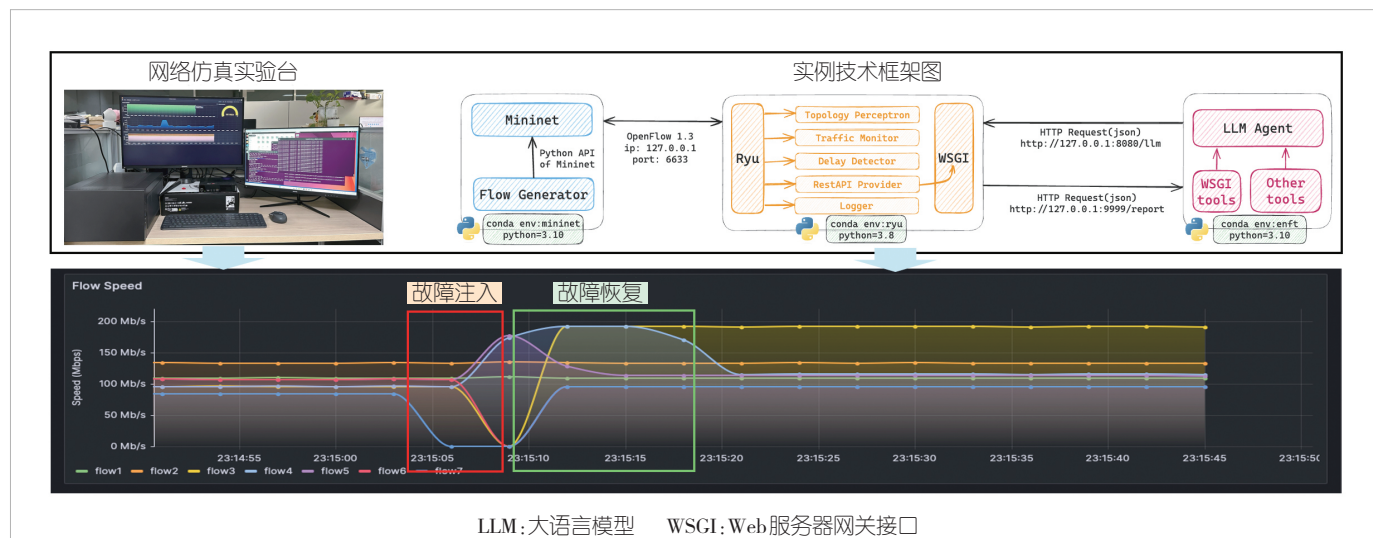


图5 大模型在网络运营管理智能化中的应用

践中发现问题解决问题；能力演进挑战，需突破多模型/多智能体协作机制、场景化评估体系等技术盲区，通过构建动态反馈闭环持续优化模型性能。

未来网络智能化的竞争本质，将取决于大模型在跨域知识融合与自主决策闭环上的突破深度，推动运营管理体系从“人工干预”向“全域自主”跃迁，实现从L3（有限自治）向L4（高级自治）以及L5（完全自治）的高阶自智网络演进。

## 致谢

感谢北京邮电大学在读博士研究生刘新秀、谭灿对本研究工作的支持！

## 参考文献

- [1] 黄韬, 刘江, 霍如, 等. 未来网络体系架构研究综述 [J]. 通信学报, 2014, 35(8): 184–197. DOI: 10.3969/j.issn.1000-436x.2014.08.023
- [2] SHEN Y F, ZHANG J, SONG S H, et al. Graph neural networks for wireless communications: from theory to practice [J]. IEEE transactions on wireless communications, 2022, 22(5): 3554–3569. DOI: 10.1109/TWC.2022.3219840
- [3] MAO H, ALIZADEH M, MENACHE I, et al. Resource management with deep reinforcement learning [EB/OL]. [2025-07-15]. <https://people.csail.mit.edu/alizadeh/papers/deepm-hotnets16.pdf>
- [4] 黄韬, 刘江, 汪硕, 等. 未来网络技术与发展趋势综述 [J]. 通信学报, 2021, 42(1): 130–150. DOI: 10.11959/j.issn.1000-436x.2021006
- [5] MAATOUK A, PIOVESAN N, AYED F, et al. Large language models for telecom: forthcoming impact on the industry [J]. IEEE communications magazine, 2025, 63(1): 62–68. DOI: 10.1109/MCOM.001.2300473
- [6] 孙方平, 钱铮铁. 高阶自智网络关键技术及应用 [J]. 中兴通讯技术, 2024, 30(4): 77–82. DOI: 10.12142/ZTETJ.202404012
- [7] TM Forum. 自智网络白皮书6.0 [R]. 2024
- [8] 宋航, 才建, 袁运栋, 等. 意图驱动的物联网双时间尺度资源分配方法 [EB/OL]. (2025-07-10) [2025-07-15]. <https://kns.cnki.net/kcms/detail/10.1491.tp.20250709.1503.002.html>
- [9] 王晓云, 韩双锋, 刘志明, 等. AI驱动的6G空口: 技术应用场景与均衡设计方法 [J]. 中国科学: 信息科学, 2025, 55(6): 1522–1533
- [10] KOUGIOUMTZIDIS G, POULKOV V K, LAZARIDIS P I, et al. Mobile network traffic prediction using temporal fusion

transformer [J]. IEEE transactions on artificial intelligence, 2025 (99): 1–15. DOI: 10.1109/TAI.2025.3556627

- [11] LIU H I, GALINDO M, XIE H X, et al. Lightweight deep learning for resource-constrained environments: a survey [J]. ACM computing surveys, 2024, 56(10): 1–42
- [12] GUO Z R, ZOU J, XIN P Z, et al. Root cause analysis of power grid 5G network faults based on large language model [C]// Proceedings of 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2025: 624–629. DOI: 10.1109/CSCWD64889.2025.11033346
- [13] BORZUNOV A, RYABININ M, CHUMACHENKO A, et al. Distributed inference and fine-tuning of large language models over the internet [J]. Advances in neural information processing systems, 2023, 36: 12312–12331

## 作者简介



**李文璟**, 北京邮电大学教授、博士生导师, 中国通信学会高级会员; 主要研究领域为无线网络智能管理、B5G/6G 网络架构; 先后主持国家“863”计划课题、国家科技重大专项项目、国家重点研发计划项目、国家自然科学基金重点项目及面上项目等 10 余项; 出版论著 2 本, 以第一起草人身份起草通信行业标准 20 余项。



**方宏林**, 北京邮电大学在读博士研究生; 研究方向为边缘网络故障容忍机制、生成式人工智能。



**喻鹏**, 北京邮电大学未来学院副院长、副教授、博士生导师, IEEE/EAI 高级会员、中国通信学会高级会员; 主要研究方向为 B5G/6G 网络管理与优化; 荣获科技奖励 5 次、国际期刊/会议最佳论文奖 5 次, 近年来主持/参与国家级项目 10 余项, 参与起草了国际行业/企业标准 10 余项。

# 大语言模型赋能智能网络的应用与挑战



## Applications and Challenges of Intelligent Networks Empowered by Large Language Models

牛嘉林/NIU Jialin<sup>1,2</sup>, 邢铭哲/XING Mingzhe<sup>1</sup>,  
张蕾/ZHANG Lei<sup>1</sup>

(1. 中关村实验室, 中国 北京 100194;  
2. 北京邮电大学, 中国 北京 100876)  
(1. Zhongguancun Laboratory, Beijing 100194, China;  
2. Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTETJ.202505003

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20250926.1148.004>

网络出版日期: 2025-09-26

收稿日期: 2025-07-25

**摘要:** 大语言模型 (LLM) 正逐步融入网络的智能规划建设、智能维护、智能优化与智能网络运营等关键环节, 在提升自动化与智能化水平方面展现出显著潜力。基于大模型与智能网络融合的背景, 梳理了大模型在智能网络各关键领域的应用路径, 总结其在提升决策效率、增强服务适配性、降低运维成本等方面的优势。深入探讨了智能网络环境下大模型面临的解空间组合爆炸与 NP 难 (NP-hard) 问题、多维度不确定性、实时性约束、数据异构性、人机协同与成本效益平衡等技术挑战, 并归纳了现有应对思路。未来, 随着多模态融合、在线学习与人机协同等技术的持续进步, 大语言模型有望在推动网络从规则驱动向知识驱动转型的过程中发挥重要作用, 为智能网络的发展提供新思路。

**关键词:** 大语言模型; 智能网络; 人工智能; 应用与挑战

**Abstract:** Large language models (LLMs) are gradually being integrated into key stages of intelligent network development, including network planning and construction, intelligent maintenance, optimization, and operations, demonstrating significant potential in enhancing automation and intelligence. This paper, grounded in the context of the convergence between LLMs and intelligent networks, reviews the application pathways of LLMs across critical areas of intelligent networks. It summarizes their advantages in improving decision-making efficiency, enhancing service adaptability, and reducing operational and maintenance costs. Furthermore, it explores the major technical challenges faced by LLMs in intelligent network environments, including the combinatorial explosion of solution spaces and NP-hard problems, multidimensional uncertainties, real-time constraints, data heterogeneity, human-machine collaboration, and cost-benefit trade-offs, and outlines current strategies for addressing these issues. Looking ahead, with the continued advancement of technologies such as multimodal integration, online learning, and human-machine collaboration, LLMs are expected to play an increasingly important role in facilitating the transition of networks from rule-driven to knowledge-driven paradigms, offering new perspectives for the development of intelligent networks.

**Keywords:** large language model; intelligent network; artificial intelligence; application and challenge

**引用格式:** 牛嘉林, 邢铭哲, 张蕾. 大语言模型赋能智能网络的应用与挑战 [J]. 中兴通讯技术, 2025, 31(5): 11-18. DOI: 10.12142/ZTETJ.202505003

**Citation:** NIU J L, XING M Z, ZHANG L. Applications and challenges of intelligent networks empowered by large language models [J]. ZTE technology journal, 2025, 31(5): 11-18. DOI: 10.12142/ZTETJ.202505003

人工智能 (AI) 技术的发展, 尤其是以大语言模型 (LLM) 为代表的模型范式, 正在重塑网络智能化的技术范式与演进方向。

### 1) AI与网络的融合

随着网络规模与业务复杂度的持续提升, AI 技术与网

络的深度融合已成为推动网络智能化转型的核心路径。AI 通过将网络管理问题的输入特征、输出目标与优化策略进行有效映射, 显著增强了网络系统的自学习能力与动态自治水平。AI 算法在网络中的应用不断深化, 覆盖从资源调度、流量预测到安全防御等多个方面。其中, 强化学习适用于动态优化任务, 而联邦学习等分布式 AI 框架则在保障数据隐私的前提下, 实现了跨站点模型的协同训练与优化<sup>[1]</sup>。

基金项目: 中国工程院战略研究与咨询项目 (2023-JB-13)



大模型的迅速发展为AI与网络深度融合带来了新机遇。这类模型具备强大的语义理解、知识抽取和上下文推理能力,已被应用于网络配置自动生成、异常行为检测和安全事件响应等场景,展现出跨层次、跨域调控网络状态的潜力。当前, AI算法的适配性研究逐步深化: 监督学习在流量预测、路径选择等有标签任务中表现稳健, 而无监督学习通过模式挖掘在恶意行为检测领域实现高精度识别<sup>[2]</sup>。随着模型能力的持续迭代, AI正驱动网络从“人工可控”向“全域自治”加速演进, 成为新一代智能网络的核心引擎。

## 2) 网络智能化需求的提升

现代网络正朝着虚拟化、编程化和弹性化方向快速演进, 新型网络架构如软件定义网络(SDN)与网络功能虚拟化(NFV)的推广, 使得网络结构更加动态复杂, 策略调整的实时性要求显著提升<sup>[3]</sup>。与此同时, 5G与6G网络以更高的速率、更低的时延和更强的连接能力为目标, 对网络资源调度、故障恢复、服务保障等方面提出了更高的智能化要求。

在此背景下, 传统的人工配置与静态策略已难以应对高速发展的业务需求, AI技术成为满足网络智能化需求的核心驱动力。网络数据的激增, 特别是加密流量比例不断上升, 进一步加剧了对自动化分析和决策系统的依赖。基于图神经网络的流量识别模型可实现复杂通信模式的分类判断, 而大模型驱动的意图识别系统则能够将用户自然语言需求高效转换为可执行的网络策略<sup>[4]</sup>。此外, 边缘计算与AI技术的结合也推动了“算力下沉”, 使得边缘节点具备一定的自主决策能力, 从而构建更加高效、协同和智能的分布式网络体系<sup>[5]</sup>。

## 3) 大模型带来的网络变革

大模型正推动网络从规则驱动走向认知驱动、从模块自动化迈向系统智能化。大模型的引入不仅提升了网络系统的性能与效率, 更在架构理念、交互方式与运行机制上带来了深刻变革。

### (1) 网络性能优化

大模型通过统一感知、理解与决策流程, 打破了传统网络中配置、监测与优化的割裂壁垒。模型可自主解析网络状态与业务需求, 实现资源按需分配与动态调度, 支撑网络从被动调整向前瞻自适应转变。

### (2) 网络安全防护

传统安全系统多依赖固定规则与特征匹配, 而大模型具备语义理解与上下文建模能力, 能够识别未知攻击路径与复杂行为链条。大模型生成式机制也推动了“以攻促防”的新范式, 重塑网络安全的响应机制与攻防博弈逻辑。

### (3) 智能网络运维

在大模型驱动下, 网络运维从脚本执行升级为知识交互。模型可解析自然语言指令, 结合上下文生成修复策略, 实现从“事后应对”到“实时响应”的跃迁, 并通过持续学习与知识图谱支撑经验迁移。

### (4) 网络资源调度

相较于传统调度策略依赖静态规则与离线优化, 大模型能够实时感知任务优先级与计算资源分布, 主动做出跨域协同调度决策, 具备“即看即调、即调即优”的动态自演化能力。

### (5) 网络多模态融合

大模型打破了网络中结构化与非结构化数据间的壁垒, 将日志、指令、拓扑、配置等异构信息转化为统一语义空间。这为网络提供了“理解自己”的能力, 实现从数据堆积到知识生成的跃升。

### (6) 网络用户体验优化

传统网络服务依赖用户适应系统, 而大模型实现了系统适应用户, 其对自然语言意图的理解与反馈能力, 使网络响应机制转向按语义驱动配置资源, 推动了从“功能对接”到“体验协同”的转变。

大模型在性能、安全、运维、调度、多模态融合和用户体验等方面对网络的全面重塑, 彰显了其从单一工具向网络认知中枢演进的路径。随着模型在计算效率、跨层协同与实时推理等方面的进一步提升, 其在推动网络向更高自治化和智能化阶段迈进中将发挥越来越关键的作用。

## 1 大模型与网络融合的技术基础与应用

### 1.1 大模型的发展与优化技术

大模型在智能网络应用中的落地, 源于其基础架构与能力体系的演进。早期依赖规则和小规模神经网络的系统, 因数据和算力受限而难以满足复杂场景需求。2017年, Transformer将自注意力机制引入序列建模, 以并行化和长距离依赖建模能力显著提升了语言理解效果<sup>[6]</sup>。当前, 代表性系统如GPT、Gemini、LLaMA等, 依托数十亿至万亿级参数及网页、书籍和代码等跨域语料, 实现了语义理解、逻辑推理与跨模态感知, 为网络流量预测、智能调度和服务感知等场景提供了坚实架构。

为了高效释放大模型能力并适配网络环境, 还需配套完善的训练与优化体系。当前主流做法是先进进行自监督预训练以获取通用语义表征, 再通过有监督微调增强特定任务性能。为降低大模型在网络部署中的资源消耗, 提出了混合精



度训练、模型并行与数据并行等提升算力利用率的方法，以及 LoRA、Adapter 等参数高效微调技术，显著减少了显存占用和边缘部署成本<sup>[7]</sup>。此外，诸如对抗训练、指令微调与人类反馈强化学习等技术被用于提升模型在动态网络环境中的稳定性与安全性，确保其在实际运营中保持高可用性和鲁棒性。

## 1.2 大模型赋能网络的应用

大模型与网络系统的融合正在加速推动网络从数据驱动的“感知型”体系向知识驱动的“认知型”体系转变。在传统网络中，策略配置主要依赖静态规则和有限模型，难以应对动态环境与复杂需求。大模型通过上下文理解与语义建模能力，使网络能够动态感知用户需求、环境变化与服务状态，支撑实时优化与智能决策<sup>[8]</sup>。代表性通用大模型和网络大模型的发展历程如图1所示。

在实际应用中，大模型已广泛辅助网络完成智能规划建设、智能维护、智能优化与智能网络运营，显著提升了网络的自动化与智能化水平。其上下文建模与语义理解能力，尤

其适用于移动通信、边缘计算与物联网（IoT）等高动态环境，能够实现多源异构信息的协同融合与联合优化<sup>[9]</sup>。同时，大模型在自然语言理解与指令生成方面展现出显著优势，使得用户可通过对话式交互精准表达需求，推动网络控制从传统程序式调用向语义驱动配置转型。随着大模型应用逐步渗透至网络管理核心环节，模型上下文协议（MCP）<sup>[10]</sup>与智能体间通信协议（A2A）<sup>[11]</sup>等新兴协议机制也被用于增强上下文共享与多智能体协同推理能力，进一步推动网络智能化水平的提升<sup>[12]</sup>。当前基于大模型的智能网络管理框架及其应用如图2所示。

### 1.2.1 智能规划建设

大模型在智能网络布局规划建设的多个核心环节中展现出实际价值，尤其在资源规划、智能选址与自动化验收方面实现了流程的优化与效率提升<sup>[13]</sup>。

在资源规划方面，运营商已借助大模型分析历史流量日志、用户活跃区域与业务类型，生成动态资源配置策略。通过时间序列建模与场景推理，模型能精准预测不同区域的流

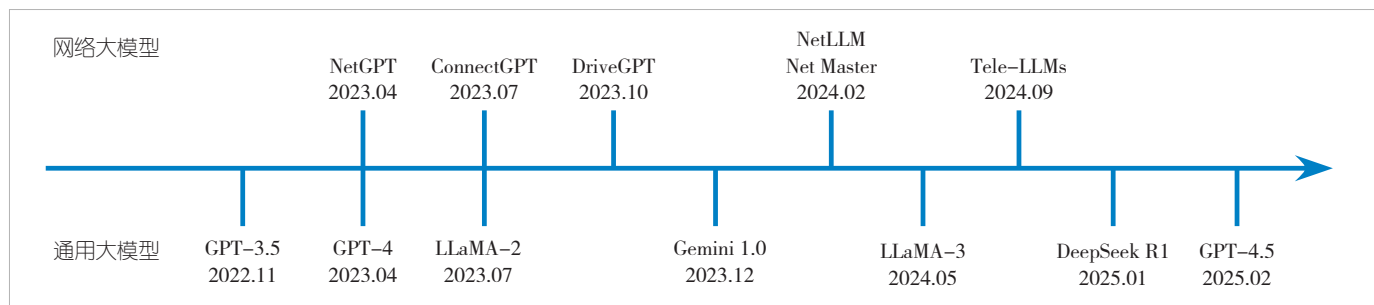


图1 代表性通用大模型和网络大模型的发展历程

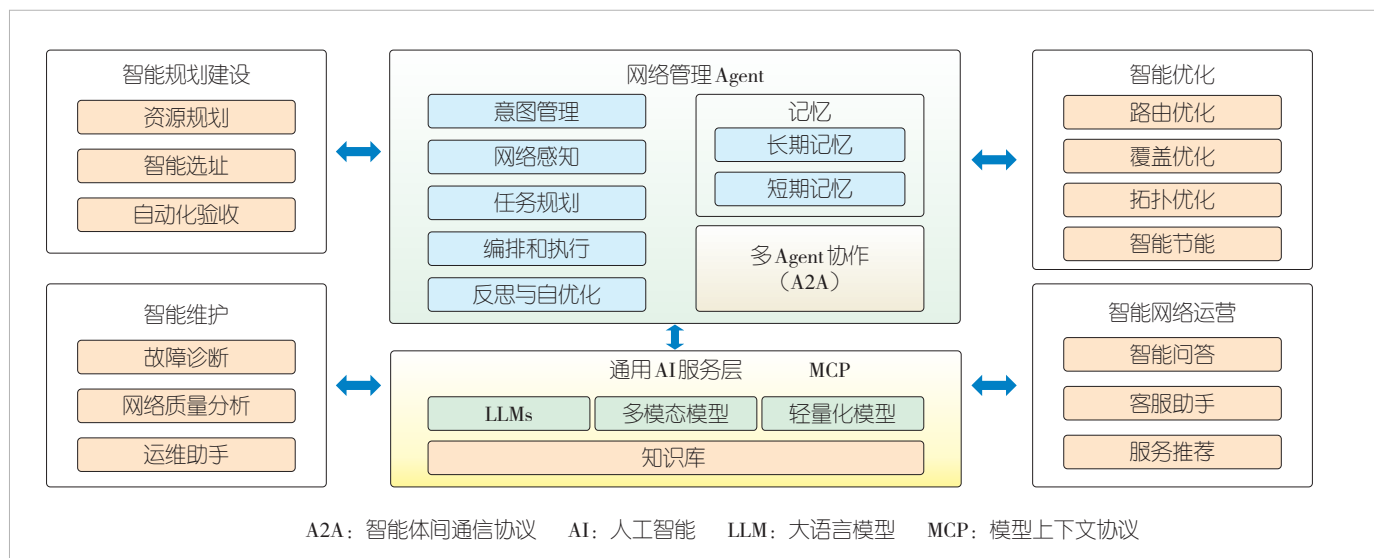


图2 基于大模型的智能网络管理框架及其应用

量峰值,辅助优化基站部署与带宽分配,避免了传统静态规划中常见的资源冗余与服务盲区问题。

在智能选址方面,大模型通过融合地理信息、建筑结构、人口密度、业务需求及部署成本等多源异构数据,自动推荐站点部署方案,减少人工勘测与方案比选的工作量<sup>[14]</sup>。当前,正在逐步为城区高密度区与乡村低覆盖区提供差异化选址支持,以提升网络部署的成本效益与服务弹性。

自动化验收环节同样得到了大模型的重塑。模型通过自然语言接口解析网络验收标准,并对设备配置、接口参数、运行状态进行自动比对和校验,生成结构化验收报告。相比传统人工检查流程,该方法显著降低了配置错误率,缩短了验收周期,提升了系统上线的一致性与规范性。

### 1.2.2 智能维护

在网络智能化转型中,大模型已广泛应用于故障诊断、网络质量分析和运维助手等维护核心场景,推动运维流程向自动化、精准化方向演进<sup>[15]</sup>。

在故障诊断方面,运营商通过接入大模型,对多源告警日志、配置项和历史工单进行语义解析与模式匹配,实现分钟级的故障定位与根因分析。在云网环境中,大模型已被用于对链路震荡、中央处理器(CPU)异常等问题进行快速溯源,并生成修复建议,替代了依赖人工检索的低效流程。

在网络质量分析上,大模型被部署于实时性能监控与异常趋势识别。通过分析丢包率、时延、吞吐等关键指标,模型能够识别视频业务卡顿、物联网(IoT)节点失联等服务劣化现象,并提前发出风险预警。目前已有运营商借助大模型部署了基于性能预测的预防性维护机制,将部分告警量压缩超过50%,显著减轻了一线运维负担<sup>[16]</sup>。

此外,运维助手类大模型被集成至网络管理平台,服务于日常运维场景。工程师可通过自然语言交互调用配置模板、生成脚本或获取操作指引,简化复杂命令的记忆与执行过程<sup>[17]</sup>。在网络割接、参数变更等高风险场景中,大模型还可辅助制定操作步骤并生成回滚预案,显著降低人为错误率<sup>[18]</sup>。同时,通过与工单系统联动,系统可实现从故障感知、策略生成到执行验证的自动闭环处理,加速故障闭环时效。

### 1.2.3 智能优化

在网络性能优化领域,大模型通过多维度智能决策显著提升网络效率与可靠性,其核心应用涵盖路由优化、覆盖优化、拓扑优化与智能节能等场景<sup>[19]</sup>。

在路由优化方面,大模型基于实时流量模式与网络拓扑状态,动态调整数据传输路径以降低时延并提升可靠性。在

SDN中,模型通过融合流量矩阵与拓扑信息,自主优化路径选择策略,有效缓解网络拥塞并增强吞吐能力。结合强化学习算法,模型可实时感知流量波动并自适应调整路由规则,实现网络性能的动态均衡<sup>[20]</sup>。

在覆盖优化方面,大模型被应用于基站参数自适应配置与边缘信号质量提升。通过综合分析基站负载、用户密度分布与环境因素,模型自动生成功率调整与天线优化方案<sup>[21]</sup>。针对移动性强的车联网应用场景,模型基于车对万物通信数据动态调整发射功率,保障关键区域的连续覆盖与服务稳定性。

拓扑优化致力于适应动态业务需求,大模型通过分析节点负载与链路状态,自主优化网络节点布局与资源分配策略<sup>[22]</sup>。在移动边缘计算场景中,模型基于业务需求预测动态重构拓扑连接,缩短了服务响应时间。在卫星-地面融合网络中,通过轨道与链路状态预测实现全球覆盖的连续性保障。

在智能节能方面,大模型被部署用于基站能耗建模与动态休眠调度。结合流量预测与设备运行监测,模型自动制定节能策略,在保障服务质量的前提下降低网络整体能耗。

### 1.2.4 智能网络运营

在网络运营层面,大模型的应用已成为连接用户与网络的智能中枢,广泛应用于智能问答、客服助手与服务推荐中,推动了智能化的客户服务与运营管理<sup>[23]</sup>。

智能问答系统利用自然语言理解技术,快速响应用户关于网络业务和故障排查的咨询。当用户用日常语言描述问题时,系统结合知识库和实时网络数据,提供精准解决方案,自动生成诊断报告和修复建议。

客服助手集成多轮对话引擎与知识图谱,自动处理用户投诉和工单。系统能自动解析投诉并关联网络告警,动态分配优先级,从而提升工单处理效率。开放问答系统的应用使得运营商能够支持多语种服务,减少人工干预并提升服务的全球覆盖性。用户投诉自动分类通过语义分析与模式识别技术,将非结构化投诉准确映射到具体网络问题类别。结合图神经网络和大模型框架,系统精准区分“覆盖问题”“速率问题”等,提升分类准确性。模型还能通过历史投诉数据预测高频故障区域,提前部署维护资源,从而实现主动预防。

在客户留存与服务推荐方面,大模型通过分析用户行为特征、服务使用模式及历史交互记录,构建动态用户画像,基于此生成定制化套餐优化建议与增值服务推荐策略,提升用户满意度与长期留存率。模型进一步结合网络资源供给状态与用户需求趋势,确保推荐方案可行性<sup>[24]</sup>。

## 2 大模型赋能网络应用的技术挑战与解决方案

随着大模型在网络中的广泛应用，其模型复杂性与系统融合深度不断提升，也带来多项关键性技术挑战<sup>[25]</sup>。针对这些挑战，需进一步研究优化相应解决方案<sup>[26]</sup>。当前大模型赋能智能网络面临的技术挑战与解决方案如图3所示。

### 2.1 资源调度场景的解空间组合爆炸与NP难

在智能网络管理任务中，虚拟网络功能部署（VNF）、服务功能链设计（SFC）等问题因涉及带宽、计算、存储等多资源联合优化及拓扑约束，已被证明属于NP难（NP-hard）问题<sup>[27]</sup>。随着网络向分布式云和多接入边缘计算（MEC）扩展，物理节点数量和链路复杂度呈指数级增长，使得传统的穷举搜索和经验式启发式算法在涉及千万级节点的动态编排场景中，难以兼顾毫秒级时延敏感型业务的实时调度，与跨域负载均衡的质量保障，其解空间维度已远超多项式时间求解能力的边界。

针对上述挑战，分布式智能与机器学习深度融合的架构被提出。一方面，云边协同的异构计算架构将整体优化任务分解为多层次子问题。云端负责集中式训练以生成全局策略，边缘节点则应用轻量化模型进行实时推理，从而显著减轻单节点搜索全量解空间的压力<sup>[28]</sup>。另一方面，混合整数线性规划与启发式松弛技术相结合的方法，在服务功能链嵌入等应用中展现显著优势<sup>[29]</sup>。通过数学规划与规则引擎的协同，将计算耗时控制在业务可接受范围内，同时保证了解的次优质量。此外，将深度强化学习（DRL）引入动态网络切片部署的研究也取得了突破：DRL智能体通过与网络环境的持续交互，自主适应流量波动和节点故障等不确定性因素，在非稳态条件下展现出较强的策略泛化能力。

现有DRL方案多基于静态环境假设，面对真实网络中的突发流量峰值和拓扑剧变时，仍会遭遇策略震荡和收敛速度下降的问题<sup>[30]</sup>。因此，如何通过分层强化学习架构实现离线策略预训练与在线微调的深度耦合，以及借助联邦学习机制在多域环境中同步更新模型，平衡解空间探索效率与策略稳定性，是突破NP-hard问题传统求解范式的关键课题。

### 2.2 网络状态的多维度不确定性

网络系统在实际运行过程中常面临多维度不确定性因素，主要包括用户行为剧烈波动、链路状态不稳定变化以及外部策略的频繁调整<sup>[31]</sup>。这些因素导致网络状态难以精确感知，策略决策难以稳定执行，影响服务质量保障和资源调度效率。在动态场景下，如移动边缘计算或多租户环境中，不确定性可能引发预测偏差、资源冲突和策略漂移，使得传统依赖静态配置的方案难以满足实时性和可靠性要求。

为应对上述挑战，智能系统需具备动态感知与自适应调整能力。一种有效路径是构建基于概率图模型与在线学习机制的策略框架，通过建模网络状态与环境变量的依赖关系，实时更新参数以修正策略偏差。实践表明，该方法可有效降低流量突变和策略冲突引发的服务中断风险。在此基础上，结合时序建模与隐空间表示技术可进一步增强系统的异常识别能力。长短期记忆网络用于提取流量变化规律，变分自编码器则通过重构机制捕捉潜在异常特征，两者协同可将异常检测和响应延迟控制在毫秒级范围内<sup>[32]</sup>。同时，为解决跨域数据孤岛与隐私保护问题，联邦学习通过分布式训练方式实现多域信息的安全聚合，在保障数据合规的同时提升预测准



图3 当前大模型赋能智能网络应用的技术挑战与解决方案



确性与系统稳健性。上述技术协同构建了从“动态感知-特征提取-协同决策”的闭环优化链路，为应对多维不确定性提供了系统性解决方案。

### 2.3 实时场景的时延约束

在工业控制、自动驾驶等对时效性要求极高的应用场景中，网络管理任务需在毫秒级甚至亚毫秒级内完成感知、决策与响应，任何延迟都可能导致系统异常或安全风险。以SDN为例，当网络拓扑发生动态变化时，控制器必须在极短时间内完成规则更新与故障恢复，确保数据流连贯性。传统集中式架构依赖全局同步，受限于信息传递延迟和中央处理瓶颈，难以满足极致实时性的需求<sup>[33]</sup>。分布式架构虽具备一定的响应优势，但局部感知信息的不完备又容易引发决策偏差，造成“低延迟高误差”或“高精度慢响应”的两难局面。

为有效应对上述挑战，边缘推理与分层协同决策成为关键策略。高实时性任务（如故障检测、负载异常响应）被下沉至边缘节点，借助轻量级神经网络模型（Tiny-YOLO、MobileNet）进行本地快速推理<sup>[34]</sup>。而复杂的全局优化决策（如多域资源调度）则由云端集中处理，形成云-边-端分层协作体系。同时通过知识蒸馏等技术，将云端大模型压缩迁移至边缘节点，实现不同计算层级间的策略同步与模型适配，保障全局一致性与推理高效性。

此外，为适配边缘设备资源受限的特点，自适应精度调整与渐进式推理机制被引入，通过优先输出近似结果并逐步迭代精化，进一步降低决策延迟<sup>[35]</sup>。在实际应用中，这一分层推理与动态协同框架显著降低了关键路径延迟，支撑了工业网络、车联网等领域对高可靠、低时延智能控制的需求。

### 2.4 异构数据的融合难问题

在智能网络管理过程中，系统需依赖大规模、多源数据支持决策制定。这些数据涵盖设备日志、性能指标、流量报文等多个维度，存在显著的结构、粒度与更新频率差异，形成高度异构的信息环境<sup>[36]</sup>。同时，由于设备故障、链路中断等因素，部分数据存在缺失、损坏或延迟上报现象，进一步削弱了数据完整性与时效性。尤其在5G及未来网络架构中，实时流量激增带来的高数据速度，加剧了数据处理的复杂性，对系统的吞吐能力与数据质量保障提出更高要求。传统集中式数据处理方案受限于全局同步延迟与边缘节点资源受限，难以在保证实时响应的同时兼顾特征提取精度，易导致模型训练偏差和推理决策滞后。

为应对上述挑战，可通过分层数据处理与自适应特征抽

象的协同机制加以优化。一方面，可在边缘节点引入轻量化预处理机制，对原始数据进行采样、冗余字段过滤及时间窗聚合，显著降低上行数据量与通信开销，缓解边缘计算资源压力。云端则集中处理高价值异构数据，通过跨域日志关联、异常模式挖掘等方式提升分析深度与准确率。另一方面，利用机器学习模型对多源数据特征进行动态抽象与自适应建模，可根据实时环境变化调整规则阈值，避免传统静态规则引擎在动态场景下失效的问题<sup>[37]</sup>。此外，基于Trans-former的统一编码架构可有效整合异构输入，结合掩码自动编码器对缺失特征进行恢复，提高数据一致性与特征完整性<sup>[38]</sup>。为进一步提升实时处理能力，可在数据采集层引入滑动窗口机制与流式标准化处理技术，稳定特征分布，保障模型输入质量。

### 2.5 运维管理下的人机协同障碍

在智能网络管理体系中，实现自动化流程与人工操作的高效融合，是突破大规模智能部署瓶颈的关键。虽然大模型在故障诊断、资源调度、性能优化等任务中展现出卓越的自动推理与决策能力，但在复杂场景下，如策略冲突仲裁、未知故障应对等，人类专家仍不可或缺。当前人机融合机制主要面临3个方面的挑战：一是语义鸿沟，即人工指令与机器执行语义之间存在映射偏差，易导致意图理解失真；二是信任壁垒，由于AI模型多为黑盒结构，专家难以快速理解决策依据，降低了系统可控性与可信度；三是应急断层，一旦AI系统推理失效，人工接管往往因上下文同步不足而响应滞后，易导致故障进一步扩散。

为解决上述问题，当前研究提出了双向可解释、动态可干预的人机协同框架。一方面，采用SHAP（SHapley Additive exPlanations）、反事实推理等可解释性技术，自动生成决策因果链，并通过可视化方式直观展现模型推理过程，帮助专家快速理解系统行为<sup>[39]</sup>。另一方面，通过模仿学习采集专家操作序列，并结合知识图谱进行领域知识建模，将人类经验转化为模型可用的策略约束，提升系统的策略合规性与决策稳健性<sup>[40]</sup>。此外，为提升人机协同的韧性，设计基于置信度的分级接管机制：当模型输出的置信水平低于预设阈值或推理后果超出安全边界时，系统自动推送决策上下文摘要，触发人工干预流程。该机制不仅缩短了故障响应时间，也降低了人工介入频率，在大规模网络运维与智能控制系统中初步验证了其实用性与有效性。

### 2.6 边缘部署的成本效益平衡

在边缘计算、IoT终端等资源受限的环境中，部署大模



型面临算力-精度-时延的“不可能三角”约束。由于大模型通常具备极高的参数规模和复杂的计算图,其推理延迟和内存占用常常超出边缘设备的硬件承载能力,导致实时性能下降、功耗飙升,甚至系统失效。此外,跨多节点协同推理时,频繁模型同步操作带来了巨大的通信带宽压力和能耗开销,进一步制约了智能网络在大规模场景下的可扩展性<sup>[41]</sup>。

为实现部署成本与性能效益的平衡,当前主要采用模型轻量化、分布式推理与硬件-软件协同优化等多维技术路径。模型轻量化方面,剪枝、量化、知识蒸馏等方法被广泛应用,通过移除冗余参数、降低计算精度或转移知识表征,在尽可能保留准确率的前提下,压缩模型体积与推理复杂度<sup>[42]</sup>。分布式推理系统通过动态负载感知机制,合理划分推理子任务,使边缘节点根据实时资源状况自适应选择推理路径,减少无效计算与通信冗余,从而提升系统整体吞吐量<sup>[43]</sup>。硬件-软件协同优化则依托定制化加速器与高效推理引擎,通过算子融合、内存共享等手段进一步压缩延迟与功耗,确保智能网络在低资源环境下仍具备稳定可靠的运行能力。这些技术体系的协同应用,已在工业边缘、智能交通、智能制造等场景中展现出显著的成本效益提升潜力。

### 3 结束语

大模型作为新一代智能引擎,已在网络领域的多个核心环节实现初步落地。其在网络规划、运维管理、性能优化及用户服务中的应用,展现出良好的可扩展性与智能化潜力。随着模型训练技术与算力平台的持续演进,大模型将进一步推动网络从规则驱动向知识驱动转型。

在未来网络建设进程中,大模型赋能的智能网络将逐步成为主流技术路径之一。通过融合多模态数据、在线学习、人机协同等技术手段,网络运营将朝着更自动化、更个性化的方向发展。随着大模型在通信行业生态中的不断融合与优化,其产业应用前景将更加广阔,将为运营商和终端用户带来更高效、便捷的服务体验。

### 参考文献

- [1] WU J, FANG X. Collaborative optimization of wireless communication and computing resource allocation based on multi-agent federated weighting deep reinforcement learning [EB/OL]. (2024-04-02)[2025-08-10]. <https://arxiv.org/abs/2404.01638>
- [2] TANG F X, MAO B M, KATO N, et al. Comprehensive survey on machine learning in vehicular network: technology, applications and challenges [J]. IEEE communications surveys & tutorials, 2021, 23(3): 2027-2057. DOI: 10.1109/COMST.2021.3089688
- [3] 崔勇, 张蕾, 马川. 面向多目标的一体化融合网络体系结构 [J]. 电子学报, 2023, 51(9): 2277-2288
- [4] ZHOU H, HU C M, YUAN Y, et al. Large language model (LLM) for telecommunications: a comprehensive survey on principles, key techniques, and opportunities [J]. IEEE communications surveys & tutorials, 2025, 27(3): 1955-2005. DOI: 10.1109/COMST.2024.3465447
- [5] CHEN Y, LI R, ZHAO Z, et al. NetGPT: a native-AI network architecture beyond provisioning personalized generative services [EB/OL]. (2023-07-12) [2025-08-10]. <https://arxiv.org/abs/2307.06148>
- [6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2025-08-10]. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
- [7] WU D, WANG X D, QIAO Y Q, et al. NetLLM: adapting large language models for networking [C]//Proceedings of the ACM SIGCOMM 2024 Conference. ACM, 2024: 661-678. DOI: 10.1145/3651890.3672268
- [8] LEE W, PARK J. LLM-empowered resource allocation in wireless communications systems [EB/OL]. (2024-08-06) [2025-08-10]. <https://arxiv.org/abs/2408.02944>
- [9] ZHOU H, HU C, YUAN D, et al. Large language model (LLM)-enabled in-context learning for wireless network optimization: a case study of power control [EB/OL]. (2024-08-01)[2025-08-10]. <https://arxiv.org/abs/2408.00214>
- [10] VENTUREBEAT. Anthropic releases model context protocol to standardize AI-data integration [EB/OL]. [2025-08-10]. <https://venturebeat.com/data-infrastructure/anthropic-releases-model-context-protocol-to-standardize-ai-data-integration/>
- [11] Google. Announcing the agent2agent protocol (A2A) [EB/OL]. [2024-05-15]. <https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/>
- [12] ZHAO X, WANG M, CECCARELLI D, et al. AI based network management agent (NMA): concepts and architecture [EB/OL]. [2025-04-16]. <https://datatracker.ietf.org/doc/draft-zhao-nmop-network-management-agent/>
- [13] BOATENG G O, SAMI H, ALAGHA A, et al. A survey on large language models for communication, network, and service management: application insights, challenges, and future directions [EB/OL]. (2024-12-16) [2025-08-10]. <https://arxiv.org/abs/2412.19823>
- [14] LI Z, XU J, WANG S, et al. StreetviewLLM: extracting geographic information using a chain-of-thought multimodal large language model [EB/OL]. (2024-11-19) [2025-08-10]. <https://arxiv.org/abs/2411.14476>
- [15] YAO K, CHEN D, JEONG J, et al. Use cases and practices for intent-based networking [EB/OL]. [2025-04-16]. <https://datatracker.ietf.org/doc/draft-irtf-nmr-g-ibn-usecases/>
- [16] YU Z Y, MA M H, ZHANG C Y, et al. MonitorAssistant: simplifying cloud service monitoring via large language models [C]//Proceedings of Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering. ACM, 2024: 38-49. DOI: 10.1145/3663529.3663826
- [17] LIN L, JIN Y, ZHOU Y, et al. MAO: a framework for process model generation with multi-agent orchestration [EB/OL]. (2024-08-04) [2025-08-10]. <https://arxiv.org/abs/2408.01916>
- [18] GOEL D, HUSAIN F, SINGH A, et al. X-lifecycle learning for cloud incident management using LLMs [C]//Proceedings of Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering. ACM, 2024: 417-428. DOI: 10.1145/3663529.3663861
- [19] HUANG Y D, DU H Y, ZHANG X Y, et al. Large language models for networking: applications, enabling techniques, and challenges [J]. IEEE network, 2025, 39(1): 235-242. DOI: 10.1109/MNET.2024.3435752

- [20] SONG Y, QIAN X S, ZHANG N, et al. QoS routing optimization based on deep reinforcement learning in SDN [J]. Computers, materials & continua, 2024, 79(2): 3007–3021. DOI: 10.32604/cmc.2024.051217
- [21] QUAN H Y, NI W L, ZHANG T, et al. Large language model agents for radio map generation and wireless network planning [J]. IEEE networking letters, 2025, 7(3): 1. DOI: 10.1109/LNET.2025.3539829
- [22] YE M, HUANG L Q, WANG X L, et al. A new intelligent cross-domain routing method in SDN based on a proposed multiagent reinforcement learning algorithm [J]. International journal of intelligent computing and cybernetics, 2024, 17(2): 330–362. DOI: 10.1108/ijicc-09-2023-0269
- [23] XU Z T, CRUZ M J, GUEVARA M, et al. Retrieval-augmented generation with knowledge graphs for customer service question answering [C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2024: 2905–2909. DOI: 10.1145/3626772.3661370
- [24] LU H, CHAI Z, ZHENG Y, et al. Large memory network for recommendation [EB/OL]. (2025-02-08) [2025-08-10]. <https://arxiv.org/abs/2502.05558>
- [25] Datatracker. Research challenges in coupling artificial intelligence and network management [EB/OL]. [2025-04-16]. <https://datatracker.ietf.org/doc/draft-irtf-nmrg-ai-challenges/>
- [26] Datatracker. Considerations of network/system for AI services [EB/OL]. [2025-04-16]. <https://datatracker.ietf.org/doc/draft-hong-nmrg-ai-deploy/>
- [27] ATTAOUI W, SABIR E, ELBIAZE H, et al. VNF and CNF placement in 5G: recent advances and future trends [J]. IEEE transactions on network and service management, 2023, 20(4): 4698–4733
- [28] GOLKARIFARD M, CHIASERINI C F, MALANDRINO F, et al. Dynamic VNF placement, resource allocation and traffic routing in 5G [J]. Computer networks, 2021, 188: 107830. DOI: 10.1016/j.comnet.2021.107830
- [29] BEHAVESH R, HARUTYUNYAN D, CORONADO E, et al. Time-sensitive mobile user association and SFC placement in MEC-enabled 5G networks [J]. IEEE transactions on network and service management, 2021, 18(3): 3006–3020. DOI: 10.1109/TNSM.2021.3078814
- [30] YAN Z X, GE J G, WU Y L, et al. Automatic virtual network embedding: a deep reinforcement learning approach with graph convolutional networks [J]. IEEE journal on selected areas in communications, 2020, 38(6): 1040–1057. DOI: 10.1109/JSAC.2020.2986662
- [31] Datatracker. Artificial intelligence framework for network management [EB/OL]. [2025-04-16]. <https://datatracker.ietf.org/doc/draft-pedro-nmrg-ai-framework/>
- [32] GAO Z, ZHANG Z, ZHANG Y, et al. Online client scheduling and resource allocation for efficient federated edge learning [EB/OL]. (2024-09-29) [2025-08-10]. <https://arxiv.org/abs/2410.10833>
- [33] LÓPEZ J, LABONNE M, POLETTI C, et al. Priority flow admission and routing in SDN: exact and heuristic approaches [C]//Proceedings of IEEE 19th International Symposium on Network Computing and Applications (NCA). IEEE, 2020: 1–10. DOI: 10.1109/NCA51143.2020.9306725
- [34] HENG L, YIN G F, ZHAO X F. Energy aware cloud-edge service placement approaches in the Internet of Things communications [J]. International journal of communication systems, 2022, 35: e4899. DOI: 10.1002/dac.4899
- [35] ZHAN H, ZHANG X, TAN H, et al. PICE: a semantic-driven progressive inference system for LLM serving in cloud-edge networks [EB/OL]. (2025-01-16) [2025-08-10]. <https://arxiv.org/abs/2501.09367>
- [36] SUN P, ZHAO B, LI X. Research on multi-source heterogeneous data fusion method of substation based on cloud edge collaboration and AI technology [J]. Discover applied sciences, 2025, 7(4): 262. DOI: 10.1007/s42452-025-06725-8
- [37] JIN B W, ZHANG Y, ZHU Q, et al. Heterformer: transformer-based deep node representation learning on heterogeneous text-rich networks [C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2023: 1020–1031. DOI: 10.1145/3580305.3599376
- [38] JEONG J, KU T Y, PARK W K. Denoising masked autoencoder-based missing imputation within constrained environments for electric load data [J]. Energies, 2023, 16(24): 7933. DOI: 10.3390/en16247933
- [39] GILPIN L H, BAU D, YUAN B Z, et al. Explaining explanations: an overview of interpretability of machine learning [C]//Proceedings of IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018: 80–89. DOI: 10.1109/DSAA.2018.00018
- [40] CHRISTIANO P, LEIKE J, Brown T, et al. Deep reinforcement learning from human preferences [EB/OL]. (2017-06-12) [2025-08-10]. <https://arxiv.org/abs/1706.03741>
- [41] SUN Q, YIN Z, LI X, et al. Corex: pushing the boundaries of complex reasoning through multi-model collaboration [EB/OL]. (2023-09-30) [2025-08-10]. <https://arxiv.org/abs/2310.00280>
- [42] MIRZADEH S I, FARAJTABAR M, LI A, et al. Improved knowledge distillation via teacher assistant [J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(4): 5191–5198. DOI: 10.1609/aaai.v34i04.5963
- [43] ZHOU H, EROL-KANTARCI M, POOR H V. Knowledge transfer and reuse: a case study of ai-enabled resource management in RAN slicing [J]. IEEE wireless communications, 2023, 30(5): 160–169. DOI: 10.1109/MWC.004.2200025

## 作者简介



牛嘉林，中关村实验室与北京邮电大学联培在读博士研究生；主要研究方向为大语言模型在网络中的应用、基于知识图谱的安全威胁建模与推理。



邢铭哲，中关村实验室助理研究员；主要研究方向为AI智能体和AI赋能网络；发表论文10余篇。



张蕾，中关村实验室副研究员；主要研究方向为网络安全和网络系统；发表论文10余篇。

# AI智能体赋能网络运营的研究与应用



## Research and Application of AI Agent Empowering Network Operations

郑雨婷/ZHENG Yuting, 程新洲/CHENG Xinzhou,  
王静云/WANG Jingyun

(中国联合网络通信有限公司研究院, 中国 北京 100048)  
(China Unicom Research Institute, Beijing 100048, China)

DOI: 10.12142/ZTETJ.202505004

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20251011.1143.002>

网络出版日期: 2025-10-11

收稿日期: 2025-08-14

**摘要:** 智能体深度重塑网络运营体系, 推动网络架构从分层解耦向全栈智能驱动的垂直整合演进。围绕智能体的发展现状与技术研究, 结合通信网络运营管理的实际需求与现网落地情况, 重点阐述了智能体在网络运营管理中的赋能作用, 包括共享网络融合规划及基于意图的网络运营服务等实际案例。结合6G通感算一体化与天地一体网络等前沿趋势, 进一步剖析了智能体赋能未来网络运营管理的发展方向, 旨在构建意图驱动、闭环自优的智能化网络新范式。

**关键词:** 智能体; 网络运营; 智能网络

**Abstract:** AI agent is profoundly reshaping network operation systems, driving the evolution of network architecture from hierarchical decoupling to vertically integrated, full-stack intelligence. Centered on the current development and technological research of agents, and in light of practical requirements and real-world deployment scenarios in communication network operations management, the enabling role of agents is elaborated, with practical cases examined in areas such as shared network convergence planning and intent-based network operation services. Furthermore, by incorporating cutting-edge trends like 6G integrated sensing, communication, and computing, as well as space-air-ground integrated networks, the future direction of agent-empowered network operations management is analyzed, aiming to establish an intent-driven, closed-loop, and self-optimizing intelligent network paradigm.

**Keywords:** AI agent; network operation; intelligent network

**引用格式:** 郑雨婷, 程新洲, 王静云. AI智能体赋能网络运营的研究与应用[J]. 中兴通讯技术, 2025, 31(5): 19-24. DOI: 10.12142/ZTETJ.202505004

**Citation:** ZHENG Y T, CHENG X Z, WANG J Y. Research and application of AI agent empowering network operations [J]. ZTE technology journal, 2025, 31(5): 19-24. DOI: 10.12142/ZTETJ.202505004

### 1 AI智能体概述及发展现状

近年来, 人工智能(AI)技术持续推动社会各领域的深刻变革。作为“AI热潮的终极形态”, 智能体被视为实现通用人工智能(AGI)的重要路径, 推动人工智能从特定任务的弱AI向“类人”的通用AI跨越, 成为引领新一轮科技革命与产业变革的关键驱动力<sup>[1]</sup>。

智能体基于大语言模型, 尝试对人类解决问题的行为进行模拟。人类在应对复杂任务时通常采用思维链(Chain of Thought)方式, 即将大目标拆解为子任务, 并根据执行反馈动态调整策略, 逐步达成目标。智能体通过模仿这一过程, 被定义为能够超越基础模型能力边界的应用系统, 能够感知环境并调用工具执行动作, 以完成预设目标<sup>[2]</sup>。与弱AI

不同, 在智能体框架下, AI可自主完成任务分解、工具选择与进度控制等环节, 直至任务结束, 人类仅需设定目标、提供资源并监督结果。

当前, 智能体研究进入爆发阶段。在企业应用层面, 领先企业已大规模部署智能体, 如京东云上线超过7 000个Agents应用于客服、供应链管理与数据分析等场景; 微软Microsoft 365 Copilot全面开放, 覆盖Word、Excel、Teams等办公全场景。随着智能体C端应用的涌现, 智能体用户规模迅速扩大。基于技术先进性、商业化成熟度与社会影响力等维度评估, 表现突出的应用包括腾讯元宝助手(融合混元大模型与微信生态)、支持金融法律专用混合专家模型(MoE)的DeepSeek-R1, 以及具备多方言识别能力的讯飞政务大脑



等。此外，DeepSeek-R1的低成本优势与Coze等低代码平台结合，显著降低了智能体的开发门槛，使普通用户无需编程即可快速构建定制化智能体，推动技术普惠。

智能体正加速向各行业渗透与场景深化。在金融领域，智能体应用于基于用户风险偏好的资产自动配置、毫秒级市场分析决策及异常交易实时监测；医疗行业则借助智能体实现辅助诊断、药物研发与健康管理升级；制造业通过预测性维护、智能质检与柔性生产等应用提升能源利用效率；零售与电商领域借助个性化推荐提升转化率，人工智能生成内容（AIGC）虚拟试衣降低拍摄成本，智能仓储优化供应链周转效率。此外，在教育、公共服务、法律、能源与环保等行业，智能体也在不断拓展创新场景。通信网络行业正从“管道提供商”向“智能服务商”转型，智能体已成为其降本增效与业务创新的核心引擎。

## 2 AI智能体的关键技术

智能体能够在无人工干预的情况下，独立与环境或其他智能体交互，通过感知与推理、规划与协调等一系列认知和操作过程来自主执行任务<sup>[3]</sup>。这一自主能力由多项关键技术所支撑。

### 2.1 多模态交互技术

强大的感知与交互能力是智能体的核心基础。作为人机协同迈向人智协同的代表性技术<sup>[4]</sup>，多模态人智交互技术融合了多模态交互与大模型能力。该技术通过整合文本、图像、音频、视频等多源信息，模拟人类的协同认知机制，有效突破了单一模态的信息局限，提升了环境理解的完整性与准确性。在此基础上，它能实现不同模态间的信息互补与跨模态关联学习，从而增强模型的泛化与适应能力，并最终通过支持多通道自然交互带来交互方式的革新。

### 2.2 记忆技术

记忆技术是智能体实现长期交互、个性化服务与情境理解的核心。当前主流技术包括情节记忆、语义记忆、程序性记忆与工作记忆等。情节记忆不同于简单的持久化数据存储，其核心在于对过去交互的理解与回溯。它通过记录并利用历史交互事件或动作序列，使智能体能够在相似情境中快速复用经验。配合多样化的记忆更新机制，情节记忆能有效提升智能体的决策效率与任务执行的准确性。

### 2.3 规划与推理技术

该能力相当于为智能体“大脑”拓展了推理功能，使其能够解析任务、制定计划并执行操作。ReAct（Reasoning

and Acting）便是一种先进的智能体推理框架，它将推理与行动相结合，通过“感知—推理—执行—优化”的迭代循环，持续调整行动策略，实现智能体与环境的动态协同<sup>[5]</sup>。

ReAct框架在处理多步推理任务方面表现优异，具备更强的复杂问题处理能力。它能够实时观察工具执行结果，验证推理逻辑，从而避免基于错误假设的连续失误，减少幻觉现象。此外，ReAct在决策过程中留下的“思考痕迹”使智能体的推理逻辑可追溯，显著增强了模型的可解释性。

## 2.4 工具使用能力

工具的使用标志着人类智慧的飞跃；与之类似，智能体调用外部工具的能力也代表了人工智能领域的重要突破，使其能够突破自身固有能力的限制。检索增强生成（RAG）技术便赋予智能体调用“外部知识库”的能力。该技术使智能体能够在回答问题前先检索相关背景信息，其工作流程包含检索、增强与生成3个核心步骤：通过从外部知识源获取信息，有效提升生成答案的质量与准确性，并显著减轻模型的“幻觉”问题。

## 3 AI智能体赋能网络运营的应用

智能体技术凭借其迈向通用人工智能（AGI）的颠覆性潜力，正驱动各行各业实现跨场景渗透与规模化价值释放。在通信网络领域，5G技术以其高数据传输速率、低延迟和大容量的特性，为物联网、智慧城市及自动驾驶等应用提供了坚实基础，成为现代生产生活的重要基石<sup>[6]</sup>。与此同时，智能体技术的深度融合为通信网络运营注入了全新的智能化能力，推动通信行业从“网络管道商”向“智能服务商”转型。在网络规划建设与运维管理这两大核心任务中，智能体已逐步实现相关应用落地，展现出显著的实践价值。

### 3.1 智能体赋能共享网络规建

中国作为5G发展的引领者，其在5G网络共建共享方面的成功实践为全球树立了标杆。共享网络模式在提升投资效率、构建绿色智能网络方面展现出显著优势<sup>[7]</sup>，但同时也大幅增加了网络规划与建设的复杂度。

网络规划与建设是构建网络基础设施的基础环节，直接决定网络服务质量与成本效益。由于其重要性，规划过程中需综合考量多方面因素，包括地理与人文空间分析、政策与产业空间分析、竞争格局、网络短板识别、发展趋势研判、业务需求评估以及规划策略制定等。对于共享网络这一特殊形态，规划建设面临更为复杂的网络现状，需协调多方策略、平衡各方用户体验、统筹经济成本，并应对更高的安全风险。

如图1所示，共享网络融合规划建设智能体框架集成了多项智能化能力，涵盖5G共享网络多指标协同动态判定、4G一张网智能化站点拆除规划及后效预测、基于隐私计算的业务量精准预测，以及5G共享网络扩容策略智能洞察等关键场景。该框架通过大小模型协同与应用程序编程接口（API）工具调用，最终实现站点级与栅格级的智能化网络规划，并支持规划方案的合理性自动评估。

5G共享网络感知对等多指标协同动态智能判定，基于共建共享网络中区分公共陆地移动网络（PLMN）统计的存量小区多维指标数据，利用智能化方法协同判定小区对等情况，识别不对等问题小区，确保共享双方网络质量与用户感知的一致性，并推动相关生产单位实现规建类网络问题的分析解决闭环。4G一张网智能拆站规划及关拆后评估预测，采用智能化方案预测4G网络中待关停基站，在统筹共享双方覆盖与业务质量的基础上，结合基础数据、用户感知、网络质量与基站负荷等多维信息，科学制定站点关停方案。同时，为保障关停后网络质量与用户感知不下降，系统从接入性能、保持性能、感知体验与负荷变化等维度开展后评估预测，形成从规划到评估的完整闭环。基于隐私计算的共享网络业务量精准预测，融合共享双方B域（业务流量、用户数等）与O域（基础数据、小区流量、用户数等）数据，构建智能预测方案。针对部分B域数据涉及用户敏感信息的问题，遵循“数据可用不可见”原则，通过联邦学习与隐私计算技术进行联合分析，实现整网级别的业务量精准预测。随着5G业务类型多样化与用户规模扩大，网络负荷显著提升，5G共享网络扩容规划智能精准洞察从小区级与整网级两个

维度出发，实现扩容需求的智能预测，在保障用户体验的同时避免过度投资。

5G共享网络融合规建智能体集成上述多项智能化能力，实现站点级与栅格级的精准规划。该智能体融合多维分析能力，灵活协调共享双方策略，全面综合考虑各类要素，最终实现共享网络规划建设的自动化与精准化。

### 3.2 智能体赋能基于意图的网络运营服务

传统的网络管理方式往往需要大量的人力与物力。当管理人员或业务部门提出具体需求时，往往涉及多个子公司或省级分公司之间的流程流转以及人工处理，导致耗时长、成本高。而基于意图的网络运营服务智能体则通过自动化与智能化手段，自动生成并执行相应网络策略，高效处理日常运维任务，从而大幅提升网络灵活性，显著降低运维成本。如图2所示，该智能体以多源数据为底座，构建富含网络知识的网络数据知识图谱，并依托意图网络框架执行任务流程，最终实现全方位的智能化网络管理。

为提高网络运营服务的精准度，基于意图的网络运营服务智能体以网络知识图谱和网络全息洞察作为底层能力支撑，如图3所示。网络全息洞察基于跨域融合数据，构建用户时空综合定位、行为偏好标签及人际关系图谱等核心算法模型；而接入的网络知识图谱则包括网络数据关联关系知识图谱与网络性能指标影响关系知识图谱。网络数据关联关系知识图谱对多类网络数据进行关联梳理与图谱建模并予以存储。网络数据涵盖基础数据、配置参数、性能指标、告警信息等多种类型，每种类型又可下钻至物理实体、逻辑实体等

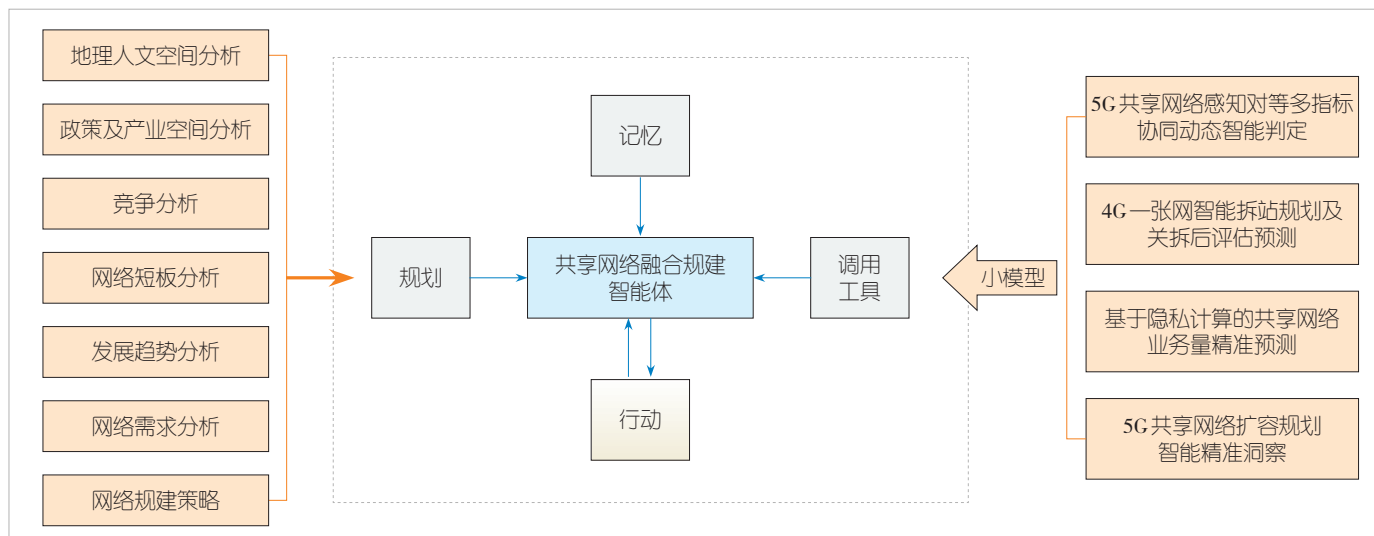


图1 共享网络融合规建智能体框架示意图

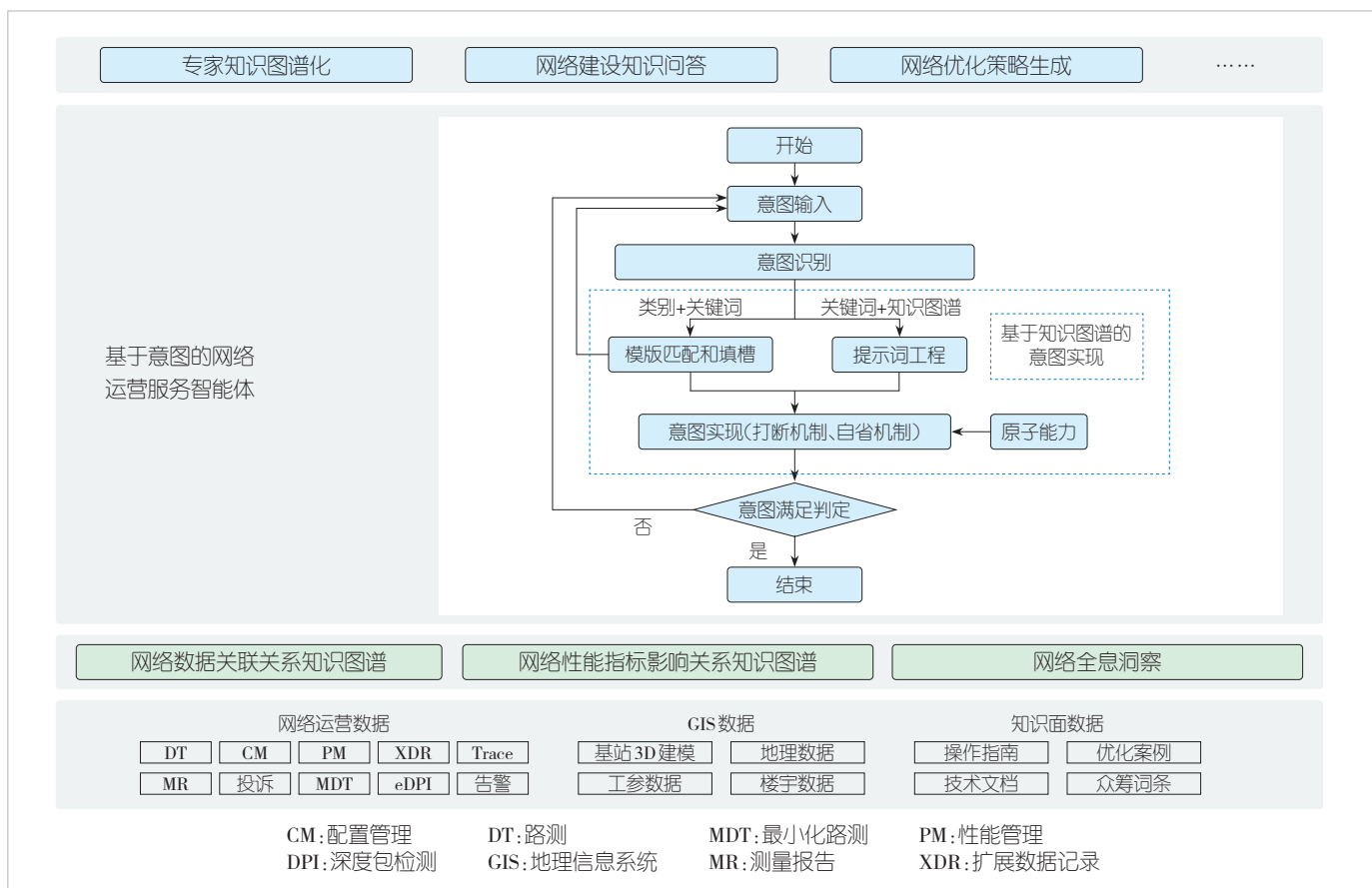


图2 基于意图的网络运营服务智能体框架示意图

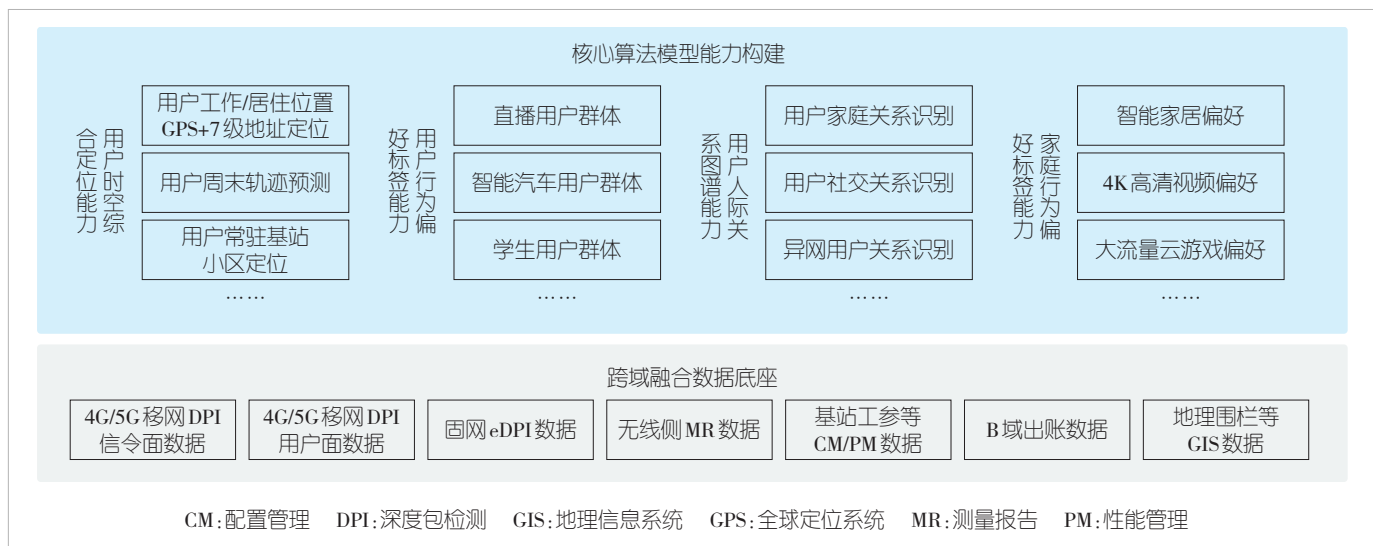


图3 网络全息洞察框架示意图

细粒度维度。据统计，仅基础数据一项，4G网络涉及7张表共157个字段，5G网络则涉及20张表共208个字段。面对如此复杂的数据资源，知识图谱化的关系构建能够更直观地展示和理解各类网络数据及其关联，显著提升网络管理的效率

与准确性，使复杂的网络数据更易于查询与分析。网络性能指标影响关系知识图谱则通过智能化方法挖掘性能指标间的影响规律，将性能指标作为实体、指标间的影响关系作为实体间关系进行建模，构建以性能指标影响关系为核心的网络



运营管理知识图谱，为后续网络规划与优化工作奠定坚实的数据基础和隐性关系知识支撑。图4为网络性能指标影响关系知识图谱的简单示例。

意图驱动管理是提升移动网络自智能力与运营效率的关键。网络运营服务智能体基于意图网络框架执行任务，其整体架构涵盖意图输入、意图识别、意图实现、意图满足判定及结果输出等一系列流程。在意图实现阶段，智能体依托专业构建的网络底层能力，包括基于知识图谱的模板匹配与填槽、嵌入向量模糊匹配形成的提示词工程等。结合打断机制与自省机制等技术支撑，最终实现基于意图的网络运营服务功能。

网络运营服务智能体已在现网投入运行，提供专家知识图谱化网络管理、网络建设知识问答及网络优化策略生成等服务。典型实施案例如根据用户X的网络使用习惯实现个性化配置：智能体接收到该意图后，首先从网络全息洞察中获取用户X的使用数据，经分析生成用户画像；随后从网络知识图谱中提取相关配置信息，结合用户画像生成个性化配置方案，最终自动完成配置部署。相比之下，传统人工配置方法难以实现单用户粒度的个性化服务，人力物力成本高昂；同时缺乏基于大数据分析的用户画像能力，无法从海量使用数据中精准挖掘个性化规律。

基于意图的网络运营服务智能体通过规划、记忆与工具调用，模拟“网络专家”的决策过程，有效整合了通信网络专业知识构建的网络数据知识图谱与网络全息洞察等底层能力。该架构显著提升了网络管理的灵活性，大幅降低运维成本，突破了传统人工管理的能力局限，并推动网络服务架构由分层解耦向垂直整合演进。

## 4 AI智能体赋能网络运营发展趋势

智能体在网络运营中的应用已超越技术探索，进入规模化落地阶段，正成为数字化转型的核心驱动力。未来，它将继续深化与网络智能运营的融合，为其发展注入强劲动能。

### 4.1 AI智能体驱动6G“通感算”一体化网络演进

6G通感算网络通过通信、感知与计算能力的深度融合，为信息的高效获取、处理与应用提供技术支持，推动复杂场

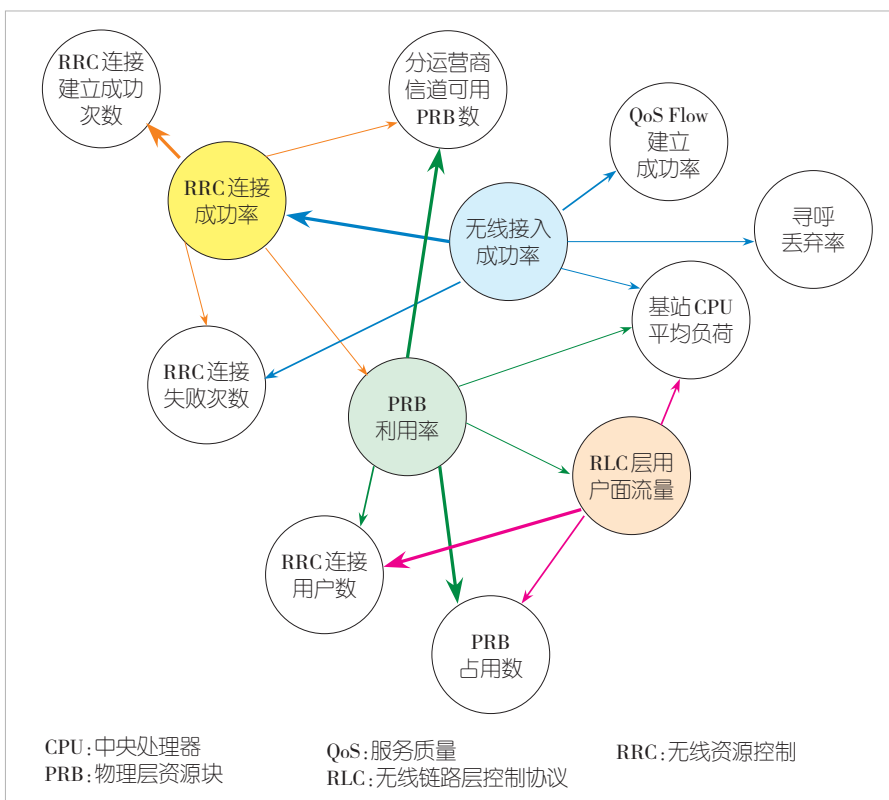


图4 网络性能指标影响关系知识图谱示意图

景的智能化发展<sup>[8]</sup>。智能体凭借其意图驱动能力，实现对通信、感知与计算资源的动态协同，从而赋能6G网络范式的重构。典型应用包括：基于多输入多输出（MIMO）波束成形反演技术实现车辆实时感知通信；利用联邦学习补偿多径效应，提升无线信号成像精度；控制无人机发射定向波束，通过回波时延差构建空中障碍物地图等。

### 4.2 AI智能体引领天地一体网络实现高度自智

面向6G的天地一体化网络架构融合天基网络与地面网络，通过卫星（包括高轨、中低轨卫星等）与地面系统直接连接，构建覆盖广泛、通信高效的多层异构网络体系，以支撑未来全球通信需求<sup>[9]</sup>。智能体通过统一智能化管理空、天、地、海全域资源，赋能天地一体网络实现自智化，突破星地资源割裂的限制，达成全域无缝覆盖。目前，智能体已在此方向取得初步成果，例如：通过星载轻量化智能体实现就地决策，缓解星地时延差异问题；借助区块链联合智能体完成跨域零信任验证，解决星地安全策略不兼容等挑战。

## 5 结束语

智能体正在全球范围内驱动产业格局的深刻变革，其凭借强大的算法与自适应技术框架，为金融、制造、医疗等关

键领域注入新动能。在通信行业，智能体通过智能规划建设、智能运营服务等全栈能力，深度重构网络运营管理体系，推动架构向垂直整合演进；并深度融合6G通感算一体化、天地一体网络等前沿方向，构建起意图驱动、闭环自优的下一代智能化网络范式。

#### 参考文献

- [1] 杨红梅, 赵勋. 人工智能赋能网络安全的挑战与应用 [J]. 中兴通讯技术, 2025, 31(3): 39-43. DOI: 10.12142/ZTETJ.202503007
- [2] Google. 智能体技术白皮书 [R]. 2025
- [3] 陈永伟. 智能体经济的崛起: AI 智能体对商业世界的重塑 [EB/OL]. [2025-08-23]. <http://cjwtyj.chinabidingnews.com/lunwen/itemid-327811.shtml>
- [4] 王镇远, 田东, 董禹, 等. 多模态交互: 从人机协同迈向人智协同 [J]. 数据与计算发展前沿, 2025, 7(3): 81-93
- [5] 孙蒙鸽, 付芸, 刘细文. 智能体赋能科研知识服务的路径解析 [J]. 智库理论与实践, 2025, 10(1): 3-18. DOI: 10.19318/j.cnki.issn.2096-1634.2025.01.01
- [6] 程新洲, 成晨, 刘红杰. 5G+智慧教育 [M]. 北京: 机械工业出版社, 2025
- [7] 李福昌, 贺琳, 周瑶, 等. 5G 共建共享网络发展总结及趋势分析 [J]. 信息通信技术, 2022, 16(3): 51-56
- [8] 吴子君, 张海君, 马旭, 等. 6G 通感算一体化体系架构与关键技术 [J]. 电子与信息学报, 2025, 47(4): 876-887
- [9] 章谦骅. 面向6G的天地一体去中心化网络架构 [J]. 天地一体化信息网络, 2025, 6(2): 57-64

#### 作者简介



郑雨婷, 中国联通研究院网络数据研究工程师; 主要研究方向为5G及AI算法创新及应用; 发表论文40余篇, 获授权国家专利38项。



程新洲, 中国联通研究院网络智能运营研究中心总监、联通集团大数据领域首席专家、教授级高级工程师; 主要研究方向为5G及大数据创新及应用; 已发表论文170余篇, 获授权发明专利100余项。



王静云, 中国联通研究院网络数据研究工程师; 主要研究方向为5G/6G网络管理; 发表论文3篇, 申请专利5项。

## 综合信息

### 中兴通讯技术杂志社第30次编委会议暨2025通信热点技术研讨会隆重召开

2025年8月16日—17日, “中兴通讯技术杂志社第30次编委会议暨2025通信热点技术研讨会”在深圳市召开。100多位来自高校、运营商、研究院所及企业的ICT专家学者参会。资深编委及专家钟义信教授、谈振辉教授、蒋林涛教授、高文院士、张宏科院士、丁文华院士出席会议, 中兴通讯创始人侯为贵、前任董事长李自学、现任董事长方榕、执行副总裁王喜瑜等公司领导莅临参会。

方榕董事长在欢迎辞中指出, 中兴通讯学术刊物作为公司与行业的重要纽带, 见证了公司成长, 助力了公司的发展, 刊物与产学研的融合发展必将为行业创造新的价值。侯为贵先生在致辞中肯定了刊物“科技向善”的公益定位, 并对其进一步发挥产学研纽带作用寄予厚望。

杂志社副主编卢丹回顾了刊物30年发展历程, 汇报了过去一年期刊取得的成果, 提出了产学研融合发展的新思路。

在杂志社成立30周年之际, 为表彰做出突出贡献的团队和个人, 公司领导分别给20位编委颁发了“启航奖”“领航奖”



和“扬帆奖”。方榕董事长为杂志社颁发了“中兴通讯产学研合作先锋奖”, 侯为贵先生为执行主编黄新明颁发了“中兴通讯技术杂志社终身荣誉奖”。

在技术研讨环节, 高文院士介绍了中国算力网计划与鹏城脑海大模型, 另有14位海内外专家围绕人工智能、6G、大模型安全等热点议题分享了最新成果, 获得代表们的一致好评。

此次会议是杂志社30年学术深耕的里程碑, 既是对过往不懈努力深情礼赞, 也是对未来更好发展的深切期盼。

# 检索增强的网络流量预测方法



## Retrieval-Augmented Network Traffic Prediction Method

常远/CHANG Yuan<sup>1</sup>, 吴春鹏/WU Chunpeng<sup>2</sup>,  
王峰/WANG Feng<sup>1</sup>

(1. 中国电信研究院, 中国 北京 102209;  
2. 中国电力科学研究院, 中国 北京 100192)  
(1. Research Institute of China Telecom, Beijing 102209, China;  
2. China Electric Power Research Institute, Beijing 100192, China)

DOI: 10.12142/ZTETJ.202505005

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20250926.1430.006>

网络出版日期: 2025-09-26

收稿日期: 2025-07-25

**摘要:** 网络流量预测是保障网络服务质量的关键技术, 现有时序模型难以融合文本描述的变更事件信息。提出一种融合时序大模型与语言大模型的协同预测框架, 实现变更事件驱动的网络流量动态预测。针对变更事件稀疏性及专业语义理解难题, 设计基于检索增强生成 (RAG) 的变更影响知识库, 通过检索历史相似变更的流量影响特征, 构建可解释的上下文提示。模型采用双阶段架构: 首先使用时序大模型生成基础流量预测, 继而由语言大模型结合检索的变更案例及当前变更描述, 对预测结果进行语义推理修正。实验表明, 在真实网络运维数据集上, 模型在变更事件场景下的预测误差相较于仅通过时序预测的方法有明显下降。

**关键词:** 流量预测; 检索增强生成; 时序预测

**Abstract:** Network traffic prediction is a critical technology for ensuring network service quality, yet existing time series models struggle to incorporate textual descriptions of change events. This paper proposes a collaborative prediction framework that integrates large-scale time series models and large language models to achieve change-driven dynamic network traffic forecasting. To address the sparsity of change events and challenges in professional semantic understanding, we design a retrieval-augmented generation (RAG)-based change impact knowledge base. This retrieves traffic impact characteristics from historically similar changes to construct interpretable contextual prompts. The model adopts a two-stage architecture: First, a large time series model generates baseline traffic predictions; subsequently, a large language model performs semantic reasoning-based refinement of these predictions by incorporating both the retrieved change cases and the current change description. Experiments on real-world network operation datasets demonstrate that our framework significantly reduces prediction errors in change event scenarios compared to time series-only approaches.

**Keywords:** network traffic prediction; retrieval-augmented generation; time series prediction

**引用格式:** 常远, 吴春鹏, 王峰. 检索增强的网络流量预测方法 [J]. 中兴通讯技术, 2025, 31(5): 25-29. DOI: 10.12142/ZTETJ.202505005

**Citation:** CHANG Y, WU C P, WANG F. Retrieval-augmented network traffic prediction method [J]. ZTE technology journal, 2025, 31(5): 25-29. DOI: 10.12142/ZTETJ.202505005

随着网络规模的不断扩大以及业务需求的日益复杂, 网络流量预测作为保障网络服务质量、优化资源调度以及提升运维效率的关键技术, 正受到越来越多的关注。准确预测网络流量的变化趋势, 不仅有助于提前识别潜在的拥塞风险, 还可以为网络容量规划、故障恢复以及安全防护提供决策支持。然而, 传统的流量预测方法<sup>[1-4]</sup>主要依赖于历史流量数据, 通常采用统计模型或深度学习模型对时间序列进行建模, 以捕捉流量的周期性、趋势性和突发性特征。尽管这些方法在常规场景下表现良好, 但在面对网络中发生的计划性变更事件 (如端口关闭、服务迁移、带宽调整等) 时,

往往难以准确反映变更对流量模式的潜在影响。

近年来, 随着大语言模型<sup>[5-8]</sup>在自然语言处理领域的广泛应用, 研究者开始尝试将语言模型的能力引入时间序列预测任务中, 以处理与文本信息相关的预测问题。例如, 在经济预测、天气预报等领域, 已有工作尝试将新闻事件、政策文本等非结构化信息与时间序列预测模型相结合, 提升预测结果的语义解释能力和准确性。然而, 将这一思路应用于网络运维场景仍面临诸多挑战。一方面, 网络环境中的变更事件具有显著的稀疏性, 即在长时间的流量数据积累中, 真正发生且对流量产生显著影响的变更操作相对较少。这使得基于监督学习的方法难以获得足够的训练样本, 从而限制了对语言模型进行微调或端到端训练的有效性。另一方面, 网络

基金项目: 国家电网有限公司总部管理科技项目 (5700-202358842A-4-3-WL)



运维领域具有高度的专业性，涉及复杂的协议、拓扑结构及服务依赖关系。通用语言模型在缺乏领域知识的情况下，往往难以准确理解变更描述的语义，并据此推理其对网络流量的具体影响。

为应对上述挑战，本文提出了一种融合时序大模型与语言大模型的协同预测框架，旨在实现变更事件驱动的网络流量动态预测。该方法的核心思想在于将变更事件的语义信息与流量预测任务进行有机结合，通过构建基于检索增强生成的变更影响知识库，解决历史变更数据稀疏和语义理解受限的问题。具体而言，本文首先将历史变更记录及其对应的流量变化特征组织为结构化的知识条目，并构建高效的检索机制。当面对新的变更描述时，系统将通过语义相似度匹配，从知识库中检索出若干历史相似的变更案例，并将其描述及影响特征作为上下文提示输入给语言大模型。随后，基于历史流量数据，使用时序大模型生成基础预测结果。最终，语言大模型将结合检索到的上下文信息以及当前变更描述，对基础预测结果进行语义推理与修正，从而输出考虑了变更影响的流量预测值。

本文提出的双阶段预测架构不仅有效融合了时序建模与语言理解的优势，还在不依赖大规模标注数据的前提下，提升了模型对网络变更事件的响应能力与泛化性能。通过引入可解释的上下文提示机制，模型在提升预测精度的同时，也为运维人员提供了可追溯的决策依据，增强了系统的透明度与可信度。实验结果表明，在真实网络运维数据集上，该方法相较于仅依赖时序模型的预测方法，在包含变更事件的预测场景中具有显著的误差降低效果，验证了其在复杂网络环境下的实用性与有效性。

## 1 双阶段协同预测框架

### 1.1 架构概述

如图1所示，本文提出的网络流量预测方法基于一种融合时序建模与语言理解能力的双阶段协同预测框架，旨在实

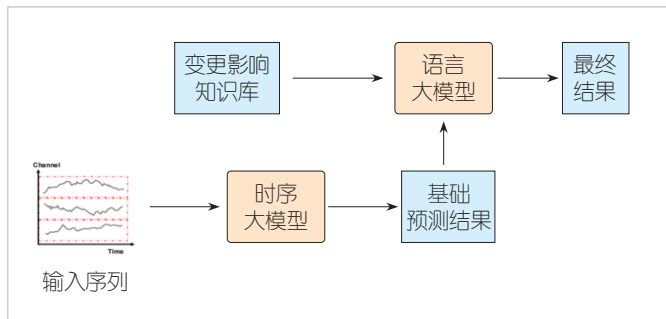


图1 检索增强的网络流量预测架构

现对计划性网络变更事件影响的动态预测。该方法的核心设计目标是解决传统时序预测模型难以处理文本描述性变更信息的问题，同时克服变更事件稀疏、难以通过端到端方式训练语言模型的挑战。系统整体架构由3个关键模块构成：变更影响知识库、时序预测模型以及语言大模型推理模块。其中，变更影响知识库作为外部知识源，负责存储与组织历史变更事件及其对应的流量变化特征，并通过检索增强生成机制为语言模型提供可解释的上下文支持；时序预测模型负责基于历史流量数据生成基础预测结果，捕捉流量的时间演化规律；语言大模型则承担语义理解与推理修正功能，结合当前变更描述与检索到的历史相似案例，对基础预测结果进行语义驱动的修正，从而输出考虑了变更影响的流量预测值。

系统运行流程分为两个主要阶段：第一阶段为基于时序模型的基础预测阶段，输入为当前时刻之前一段时间内的历史流量数据，输出为不考虑任何变更操作影响的未来流量预测结果；第二阶段为语言大模型主导的修正阶段，该阶段不仅接收来自第一阶段的预测结果，还接收当前计划变更的自然语言描述，并通过检索增强生成（RAG）机制从变更影响知识库中检索出若干历史相似变更事件，将其描述与对应的流量影响模式作为上下文提示注入至语言模型中，以辅助其理解当前变更的潜在影响。语言大模型在此基础上对基础预测结果进行语义推理与调整，最终输出融合了变更语义信息的预测值。这种双阶段架构不仅有效分离了时序建模与语义推理的功能边界，也避免了直接对语言模型进行大规模微调的需求，从而提升了方法的实用性与泛化能力。

### 1.2 变更影响知识库的构建

为有效解决网络变更事件语义信息难以建模、历史变更样本稀疏以及大语言模型对网络运维领域理解能力受限的问题，本文构建了一个结构化、语义增强的变更影响知识库。该知识库旨在为语言大模型提供可检索、可解释的历史变更上下文信息，从而在不依赖大规模标注数据与模型微调的前提下，增强其对变更事件语义的理解与推理能力。知识库的构建主要包括3个关键环节：数据收集与预处理、知识条目组织以及检索机制设计，它们分别从数据来源、知识表示与信息检索3个维度构建支持系统。总体的构建流程如图2所示。

在数据收集与预处理阶段，知识库的数据源主要来自网络运维日志系统中记录的历史变更事件及其对应的网络流量数据。变更事件通常以自然语言文本的形式记录，包含变更时间、操作类型（如端口关闭、服务迁移、配置更新等），涉及设备或服务、变更原因、变更可能造成的影响等信息。

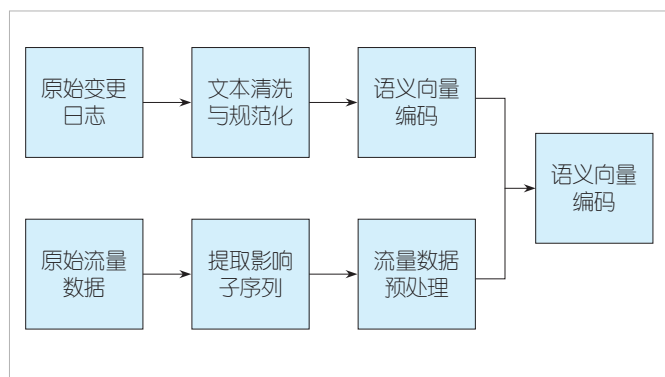


图2 变更影响知识库的构建流程

为了保证知识库中变更条目的质量与可解释性，需对原始变更文本进行清洗与规范化处理。具体而言，首先去除冗余信息与非结构化表达，保留与网络行为直接相关的操作描述；其次，对变更描述中的专业术语进行标准化，例如将“关闭端口 443”与“停用 HTTPS 服务”统一为一致表述，以提升后续语义匹配的准确性。与此同时，针对每条变更记录，提取其发生前后一段时间内的网络流量数据，用于刻画该变更对网络状态量的潜在影响。流量数据通过平滑滤波等手段进行预处理，以消除噪声干扰，保留具有代表性的流量变化模式。

在完成数据预处理之后，下一步是将变更描述与流量序列信息组织为结构化的知识条目，形成可供后续检索与推理使用的知识单元。每个知识条目由两个核心部分构成：变更语义描述与流量影响序列。其中，变更语义描述是对原始变更文本进行语义编码后的向量化表示，本文采用通用文本嵌入模型（BGE 模型<sup>[9]</sup>）对其进行嵌入编码，以保留其语义信息；流量影响序列则直接记录了该变更发生前后一段时间内的原始流量时间序列数据，以数值序列形式保存。通过上述结构化组织方式，每个知识条目不仅保留了变更事件的语义信息，也完整刻画了其对于网络流量的实际影响过程，从而为后续的语义检索与上下文提示生成提供了基础支撑。为确保 RAG 知识库提供准确的修正依据，本系统还引入了人工审核环节，仅收录对流量波动影响显著且语义描述完整的变更事件。

在知识条目的组织完成后，需构建高效的检索机制以保障系统响应效率与预测准确性。本文采用基于语义相似度的检索策略，利用向量空间中的相似性度量方法，从知识库中快速定位与当前变更描述最相似的历史变更条目。具体而言，首先将输入的当前变更描述通过相同的语义编码模型转化为向量表示，然后在知识库中计算其与所有条目中变更语义描述向量的余弦相似度，并依据相似度排序选取 Top-K 个

最相似的历史变更案例作为上下文提示。检索到的 Top-K 变更条目将与其对应的流量影响序列一同作为上下文信息注入语言大模型，辅助其理解当前变更可能带来的网络流量变化趋势。

### 1.3 基于时序大模型的基础预测

整个预测过程的第一阶段为基于时序大模型的基础预测。系统接收当前时刻之前一段时间内的历史流量数据作为输入，目标是生成一个不考虑任何变更操作影响的基础流量预测序列。该阶段的核心任务是捕捉网络流量的固有时序演化规律，包括周期性、趋势性、突发性等特征。为实现对复杂流量模式的高效建模，本文选用 Moirai 作为基础预测模型<sup>[10]</sup>。Moirai 是一种基于 Transformer 架构的通用时序大模型，具有强大的多变量建模能力与长序列预测性能。该模型无需手工设计特征，能够直接接受原始时间序列作为输入，并通过自注意力机制自动学习变量间的复杂依赖关系。此外，Moirai 支持零样本预测，在未见过的目标变量上也能保持良好的泛化能力，这使其特别适用于网络流量预测中可能出现的多维度、多设备、多指标的复杂场景。为进一步提升该模型与网络流量任务的适配性，提高预测的准确率，本文以开源的 Moirai 模型作为基座模型，利用 1 年周期的历史流量数据对模型进行微调。在本系统中，Moirai 的输入为历史流量序列，输出为未来一段时间内的预测值。模型的输入为历史流量序列：

$$\mathbf{X} = [x_{t-T+1}, \dots, x_t] \in \mathbb{R}^T \quad (1),$$

其中， $x_t$  表示  $t$  时刻的流量值， $T$  表示输入序列的长度，例如以 5 min 粒度采集 1 d 的序列长度为  $T = 288$ 。模型的输出则为未来一段时间的流量序列：

$$\mathbf{Y} = [y_{t+1}, \dots, y_{t+M}] \in \mathbb{R}^M \quad (2)。$$

### 1.4 基于语言大模型的语义修正

该阶段的目标是将用户输入的计划性变更描述有效地融入预测结果中，从而生成考虑了变更影响的最终预测流量序列。由于变更事件通常以自然语言形式描述（例如“将于明日 10:00 关闭服务器 A 的端口 443”），传统的时序模型无法直接理解此类信息，因此需要借助语言大模型来完成语义理解与推理任务。在本阶段，系统首先将当前变更描述输入至预训练语言大模型中，获取其语义表示；随后，通过前文所述的 RAG 机制，从变更影响知识库中检索出若干历史相似变更事件，并将其变更描述与对应的流量影响序列作为上下文信息注入语言模型。这种上下文提示机制不仅为语言模型

提供了可解释的推理依据,也有效缓解了变更事件稀疏、语言模型缺乏领域知识所带来的语义理解偏差问题。

语言大模型在接收到基础预测结果、当前变更描述以及检索到的历史上下文信息后,通过设计的提示模板引导其对基础预测进行语义驱动的修正。具体而言,提示模板会明确指示模型:基于当前变更的语义描述,并参考历史相似变更的流量影响模式,对基础预测结果进行调整,输出修正后的未来流量预测序列。模型在生成过程中不仅考虑当前变更的直接语义含义,还结合检索到的历史案例的流量变化趋势,进行类比推理与语义泛化,从而实现对变更影响的动态建模。值得注意的是,这一阶段并不依赖于对语言大模型的微调,而是完全基于其基础能力与上下文学习机制,从而提升了方法的部署灵活性与泛化能力。

## 2 实验验证

### 2.1 数据集

为验证本文所提出方法在面向计划性变更事件的网络流量预测任务中的有效性,实验聚焦于云池出口链路的网络流量监测数据及对应的变更操作记录。该数据集涵盖了连续8个月的流量观测序列与运维变更日志,时间跨度覆盖了典型的业务高峰期与低谷期,具有较强的代表性与现实意义。其中,网络流量数据以5 min为粒度进行采样,记录了云池出口的总流速(单位为bit/s),经过预处理后形成连续、对齐的时间序列数据。该流量序列整体呈现出较强的周期性特征,如每日早晚的访问高峰、周末与工作日之间的流量差异等,同时也受到突发事件(如服务发布、故障恢复、网络攻击等)的影响,表现出一定的非线性与不确定性。

在流量数据之外,数据集还包含同期记录的网络变更日志,每条变更记录均包含变更发生时间、变更标题、变更描述以及变更影响等字段。其中,变更时间精确到分钟级别,可与流量序列进行时间对齐,变更标题为简要概括变更内容的短语,如“某个区出口切流变更”等,变更描述则以自然语言形式详细说明了变更的背景、具体操作内容等,变更影响字段由运维人员记录了该变更可能对网络服务状态、流量分布等方面造成的实际影响情况。需要指出的是,尽管变更

事件在整个时间跨度中相对稀疏,但其对网络流量的影响往往具有显著性和可观察性,因此构成了本文方法中变更影响知识库的核心数据来源。

为了保证数据质量与建模效果,本文在数据预处理阶段进行了多项标准化处理。首先,对流量数据进行了缺失值插补与异常值检测,采用滑动窗口中位数滤波与线性插值相结合的方法填补缺失点。其次,对变更描述文本进行了清洗与规范化,去除冗余信息与非结构化表达,统一术语表述,并对部分描述不完整或语义模糊的变更记录进行了人工补充与标注,以提升后续语义检索与语言模型理解的准确性。最后,将流量数据与变更记录按时间戳进行对齐,构建出每条变更事件对应的变更前后流量时间窗口,用于构建变更影响知识库中的知识条目。

### 2.2 实验对比

为验证本文所提出方法在考虑计划性变更事件影响下的网络流量预测能力,实验选取了预测时间段内存在明确变更事件的历史样本作为测试集(测试集中的变更事件未参与知识库的构建),重点考察本文方法相较于仅依赖时序大模型的预测方法在变更影响建模方面的优势。具体而言,测试样本均来自数据集中变更发生前后的时间窗口,确保每条测试样本均包含一次计划性变更操作及其对网络流量产生的可观测影响。实验对比的两个方法分别为:1)仅使用Moirai时序大模型的预测方法,即不引入任何变更语义信息,仅基于历史流量数据进行未来流量预测;2)本文提出的双阶段协同预测方法,在Moirai预测基础上,结合变更描述文本与检索增强生成机制,通过语言大模型对预测结果进行语义修正。为直观说明两种方法在变更影响建模上的差异,图3展

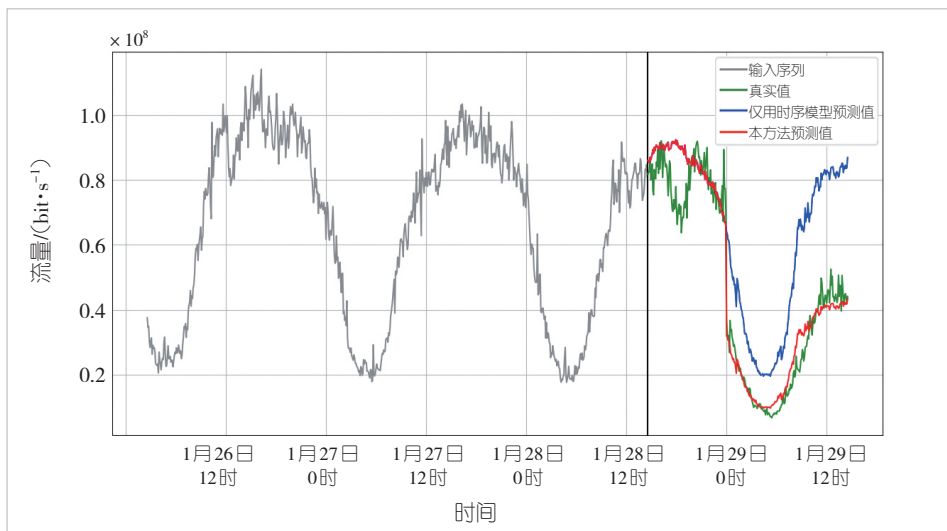


图3 使用Moirai时序大模型预测与双阶段协同预测的效果对比



示了一组测试样本中真实流量曲线、Moirai 基础预测曲线与本文方法预测曲线的对比情况。

从图 2 可以看出, Moirai 模型虽然能够较好地捕捉流量的整体趋势与周期性变化, 但在面对计划性变更事件所引发的流量突增或突降时, 其预测结果仍存在明显偏差。例如, 在某次计划性服务迁移操作发生后, 真实流量在短时间内出现了显著下降, 而 Moirai 的预测结果则延续了历史趋势, 未能准确反映该变更所带来的影响。相比之下, 本文方法在引入变更描述与历史相似案例的基础上, 能够有效识别出变更事件可能引发的流量变化模式, 并对基础预测结果进行合理修正, 使其更贴近真实流量变化趋势。

为进一步量化评估两种方法在变更事件影响预测中的性能差异, 本文采用均方误差 (MSE) 和平均绝对误差 (MAE) 两项指标对测试集上的预测结果进行评价。表 1 展示了两种方法在包含变更事件的测试样本上的性能对比。

从表 1 可以看出, 本文所提方法在两项评价指标上均显著优于仅使用 Moirai 的基础预测方法, 表明本文方法在预测误差方面具有明显优势。这些结果充分说明, 本文所提出的双阶段预测框架能够有效融合变更事件的语义信息, 并结合历史相似案例进行上下文推理, 从而显著提升在变更驱动场景下的流量预测精度。

表 1 使用 Moirai 时序大模型预测与双阶段协同预测的量化对比

测试方法	MSE	MAE
仅使用时序预测模型	0.057 0	0.176 6
本文双阶段预测框架	0.007 4	0.061 1

MAE: 均方误差 MSE: 平均绝对误差

### 3 结束语

本文的研究不仅为网络流量预测提供了一种新的建模思路, 也为时序预测与语言理解的跨模态融合提供了可借鉴的范式。未来的工作将进一步探索该方法在多类型网络场景中的适应性, 并尝试引入更多上下文信息 (如网络拓扑、服务依赖关系等), 以提升预测模型的语义表达能力与推理深度。

#### 参考文献

[1] AQUEDI O, LE V A, PIAMRAT K, et al. Deep learning on network traffic prediction: recent advances, analysis, and future directions [J]. ACM computing surveys, 2025, 57(6): 1–37. DOI: 10.1145/3703447

[2] LIM B, ZOHREN S. Time-series forecasting with deep learning: a survey [J]. Philosophical transactions of the royal society of London series A, 2021, 379(2194): 20200209. DOI: 10.1098/rsta.2020.0209

[3] MASINI R P, MEDEIROS M C, MENDES E F. Machine learning advances for time series forecasting [J]. Journal of economic surveys, 2023, 37(1): 76–111. DOI: 10.1111/joes.12429

[4] BENIDIS K, RANGAPURAM S S, FLUNKERT V, et al. Deep learning for time series forecasting: tutorial and literature survey [J]. ACM computing surveys, 2023, 55(6): 1–36. DOI: 10.1145/3533382

[5] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report [EB/OL]. (2023-03-15) [2025-08-16]. <https://arxiv.org/abs/2303.08774>

[6] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large language models [J]. ACM transactions on intelligent systems and technology, 2025, 16(5): 1–72. DOI: 10.1145/3744746

[7] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models [EB/OL]. (2023-03-31) [2025-08-16]. <https://arxiv.org/abs/2303.18223>

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2018-10-11) [2025-08-16]. <https://arxiv.org/abs/1810.04805>

[9] XIAO S T, LIU Z, ZHANG P T, et al. C-pack: packed resources for general Chinese embeddings [C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2024: 641–649. DOI: 10.1145/3626772.3657878

[10] WOO G, LIU C, KUMAR A, et al. Unified training of universal time series forecasting transformers [EB/OL]. (2024-02-04) [2025-08-16]. <https://arxiv.org/abs/2402.02592>

#### 作者简介



**常远**, 中国电信研究院大数据与人工智能研究所工程师; 主要研究领域为大模型、智能体技术等; 发表论文 10 余篇。



**吴春鹏**, 中国电力科学研究院人工智能所副主任; 主要研究领域为机器学习、边缘计算、生物启发视觉技术; 发表论文 50 余篇。



**王峰**, 中国电信研究院大数据与人工智能研究所副所长; 主要研究领域为云计算、人工智能技术等; 发表论文 20 余篇。

# 原生AI融合网络数字孪生赋能下一代无线网络自治



## Native AI-Integrated Network Digital Twin for Empowering Next-Generation Wireless Network Autonomy

王首峰/WANG Shoufeng, 郭建超/GUO Jianchao,  
边森/BIAN Sen

(亚信科技(中国)有限公司, 中国 北京 100193)  
(AsiaInfo Technologies (China), Inc., Beijing 100193, China)

DOI: 10.12142/ZTETJ.202505006

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20251011.1230.004>

网络出版日期: 2025-10-13

收稿日期: 2025-08-02

**摘要:** 面向下一代无线网络的高水平自治需求, 提出一种融合原生人工智能(AI)与网络数字孪生的一体化架构。该架构涵盖无线网络数字孪生中的数据采集、模型构建与管理等关键环节, 以及原生AI驱动的网络性能预测、AI用例自生成与网络策略自定义等核心能力。通过将原生AI深度融入数字孪生系统, 构建了“数据-模型-决策-验证”的内生智能闭环, 为应对6G网络高复杂度与高动态性环境下的自治挑战, 提供了系统化的架构设计与理论支撑。

**关键词:** 6G; 数字孪生网络; 原生AI

**Abstract:** To address the demand for high-level autonomy in next-generation wireless networks, an integrated architecture that converges native artificial intelligence (AI) with network digital twin (NDT) is proposed. The framework encompasses key processes within the wireless NDT, including data acquisition, model construction, and management, alongside core native AI-driven capabilities such as network performance prediction, automated AI use case generation, and self-customized network strategy formulation. By deeply embedding native AI into the digital twin system, an endogenous intelligent closed loop of "data-model-decision-verification" is established. This provides a systematic architectural design and theoretical foundation for tackling the autonomy challenges posed by the high complexity and dynamic nature of 6G network environments.

**Keywords:** 6G; digital twin network; native AI

引用格式: 王首峰, 郭建超, 边森. 原生AI融合网络数字孪生赋能下一代无线网络自治 [J]. 中兴通讯技术, 2025, 31(5): 30-36. DOI: 10.12142/ZTETJ.202505006

Citation: WANG S F, GUO J C, BIAN S. Native AI-integrated network digital twin for empowering next-generation wireless network autonomy [J]. ZTE technology journal, 2025, 31(5): 30-36. DOI: 10.12142/ZTETJ.202505006

## 1 网络数字孪生融合网络智能化研究现状

### 1.1 数字孪生网络标准现状

全球顶尖咨询公司 Forrester 的研究表明, 数字孪生技术目前主要应用于工业制造、能源、交通等领域, 而在通信网络等基础设施管理方面的应用仍处于萌芽阶段。另据 Gartner 发布的 2023 年十大战略技术趋势预测, 到 2027 年, 全球超过 40% 的大型企业将在其元宇宙相关项目中采用数字孪生技术。然而网络数字孪生技术的实际应用尚处于起步阶段<sup>[1]</sup>。

在标准化与学术研究层面, 国际电信联盟电信标准化部

门 (ITU-T) 在其面向未来网络的 Network 2030 焦点组技术报告中, 将数字孪生列为未来网络 12 个代表性用例之一<sup>[2]</sup>, 并发布了《数字孪生网络需求与架构》标准研究<sup>[3]</sup>, 同时启动了“数字孪生网络能力评级”等项目的立项研究工作。国际电信管理论坛 (TM Forum) 围绕数字孪生网络服务发起了催化剂项目 “Digital network twin for data-driven 5G services”, 并已完成端到端服务等级协议 (SLA) 闭环保障的概念验证 (PoC)<sup>[4]</sup>。互联网工程任务组 (IETF) 也立项了“数字孪生网络: 概念与参考架构”项目, 相关架构文稿已被纳入网络管理研究组 (NMRG) 的组稿范畴<sup>[5]</sup>。此外, 国际知名通信测试厂商 Spirent 指出, 由于 5G 技术复杂且可用

性仍面临挑战,需引入数字孪生技术以构建灵活高效的5G试验平台及沙箱环境。Spirent专注于5G信道仿真和移动定位技术,通过实现高精度无线信道仿真,持续评估与预测网络状态,从而助力运营商提升生产效率、降低运营风险、优化决策并增强网络可靠性。

未来6G网络将突破传统方法的局限,转向以人工智能(AI)为核心驱动力的网络规划、建设、运维与优化方式。该类人工智能方法可规避传统数学建模中的抽象转换问题,更贴合实际网络需求,并能够借助大规模神经网络计算快速生成合理决策。实现该目标需满足两个关键条件:首先,系统需构建与现网高度一致的数字模型,为AI算法提供准确输入数据;其次,针对人工智能决策存在的不可解释性,需引入具备仿真能力的平台以验证其决策有效性。人工智能、数字模型与动态仿真能力的有机融合,构成了智慧原生型的数字孪生系统。通过在数字空间中构建无线网络的数字孪生体,并结合网络可视化、数据开放、动态仿真与智能分析决策等技术,可有效提升未来6G网络的自主性与自治能力。

## 1.2 网络AI标准及技术发展现状

自Release 8起,第3代合作伙伴计划(3GPP)在4G中引入了自组织网络(SON),初步实现了管理面的自动化;Release 16在5G核心网中新增网络数据分析功能(NWDAF),支持在网络内部部署智能化应用;Release 18确立“AI/ML for NG-RAN”与“AI/ML for NR Air Interface”项目,推动支持人工智能的5G系统架构与空口设计<sup>[6-8]</sup>,这表明移动通信标准正持续朝着与人工智能逐步融合的方向演进。近年来,国际标准组织相继启动原生智能相关立项研究,其核心目标均聚焦于构建更适合AI应用的新一代6G网络架构、协议与机制。这一技术趋势已在全球范围内形成共识。国际电信联盟(ITU)发布5G+AI国际标准《机器学习应用于未来网络(含5G)中的架构框架》,明确指出机器学习技术将显著改变网络的运营与优化模式<sup>[9]</sup>。

6G旗舰研究项目Hexa-X致力于构建全新的6G智能网络架构,以实现多项关键使能技术的智能融合。2022年2月,由美国电信行业解决方案联盟(ATIS)成立的Next G Alliance发布报告《Next G Alliance Report: Roadmap to 6G》,将“人工智能原生网络”列为北美6G发展的六大目标之一<sup>[10]</sup>。国际上多家运营商与设备厂商也已陆续展开原生智能及网络自治相关研究。例如,思科(Cisco)推出基于意图的网络方案,将网络转变为可编程平台,通过策略驱动实现全网自动化交互,以适应不断变化的业务需求。Apstra则发出一套供应商无关、基于意图的闭环指挥控制系统,可实

现网络的自动化配置、故障修复与安全防护。

IMT-2030(6G)推进组在《6G网络架构愿景与关键技术展望白皮书》中明确提出,“AI构建网络、网络赋能AI”将成为6G网络的关键能力。2022年1月,6G网络AI联盟(6GANA)发布《6G网络原生AI技术需求白皮书》,指出原生智能(Native AI)将成为未来6G移动通信系统的核心特征之一,并统一了相关术语,提出原生AI架构应具备的十大关键技术特征。中国运营商与设备厂商也相继开展原生智能研究:中国电信研发了共建共享云网智慧运营系统,主导完成了基于共建共享网络的5G人工智能运营技术攻关、系统开发与产业应用;华为推出了意图驱动的智能网络解决方案,致力于在用户与应用程序之间构建可预测、可自愈的网络系统。原生智能网络已成为中国通信领域的研究热点。未来,各企业及行业组织将持续推进6G网络AI相关技术在标准制定、政策监管与产业应用等方面的探索,通过信息与通信技术(ICT)产业、垂直行业、AI服务与解决方案提供商、学术界等多方协同,形成行业共识,共同推动AI成为6G网络全新的基础能力与服务<sup>[11-12]</sup>。

在原生AI与网络融合的技术演进中,现有研究虽已探索人工智能在网络优化中的应用,但仍存在两点局限:一是AI模型大多依赖离线训练,难以实时适应6G网络的高动态环境;二是AI决策与网络控制之间耦合不足,缺乏端到端的闭环验证机制。这些不足为本文提出的“原生AI用例自生成”与“策略自定制”技术提供了创新空间。通过将AI深度融入网络孪生体的全生命周期管理,可在动态场景中实现智能决策与实时验证。

## 1.3 网络自治相关标准及产业进展

在6G研发于全球范围内大规模推进的同时,以TM Forum为代表的国际组织正在持续深化网络自治相关的标准研究。自智网络理念的核心目标,是通过构建体系化方法、技术框架与解决方案,助力运营商实现网络的数智化转型与自治化运营,从而更有效地应对5G/6G时代对网络提出更高要求的复杂商业场景,进一步提升运维效率,实现降本增效,并改善用户体验。为此,TM Forum专门设立了自智网络协作项目(ANP)工作组,聚焦于“通用与专业协同”下的五大标准方向:分级标准、评价体系、架构、接口与关键技术。近年来,该组织已发布多份自智网络最佳实践指南及相关标准,图1为TM Forum定义的自智网络架构<sup>[13]</sup>。

在2023年9月发布的TM Forum第5版自智网络白皮书,对未来技术演进进行了深入分析,明确指出网络自治的关键演进技术具体包括:网络AI基础模型、网络可信性技术、



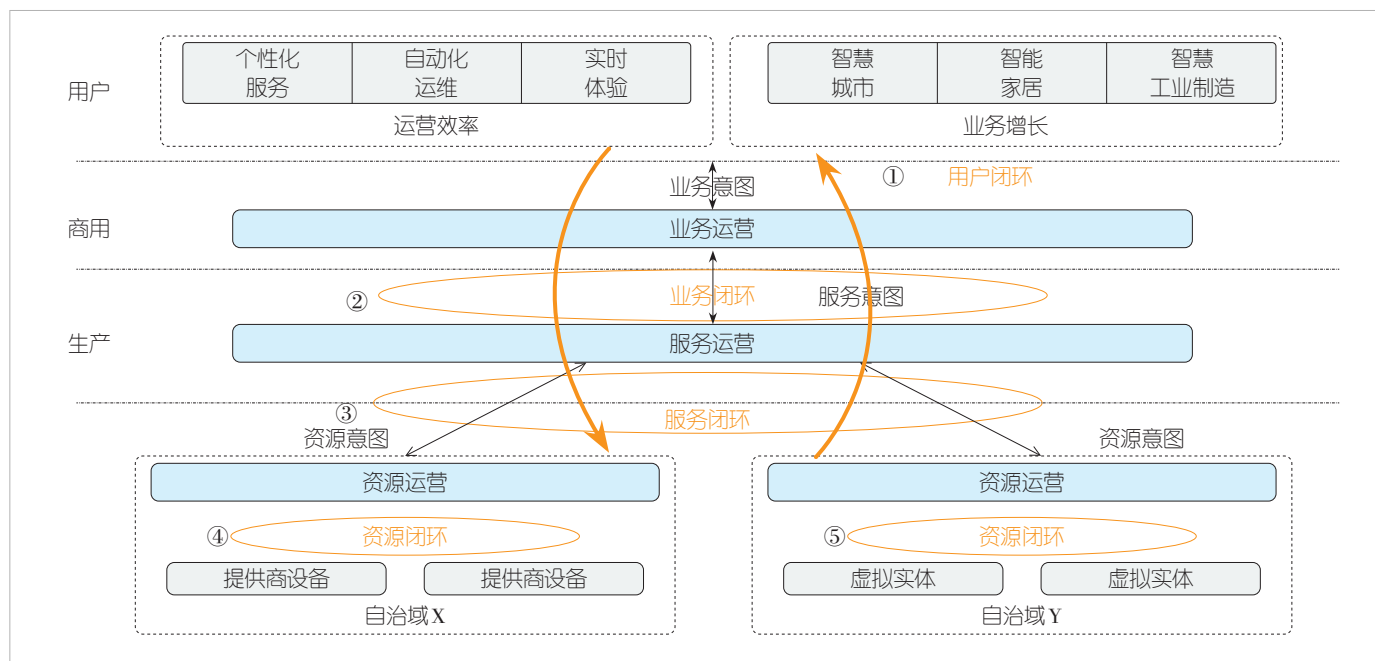


图1 电信管理论坛(TMF)定义的自智网络架构

网络数字孪生、在线网络仿真、网络认知智能及网络人机共生。

与此同时，中国通信标准化协会（CCSA）、3GPP、ITU-T、欧洲电信标准协会（ETSI）等全球标准组织也陆续启动了自智网络相关的标准化工作，积极推动该领域的标准布局。CCSA已成立以TC7（技术工作委员会）和TC610（标准推进委员会）为核心的联合协同工作组，带动多个技术委员会深入合作，立项了40余项自智网络相关标准与研究课题，加速自智网络标准体系的构建。ITU-T SG13则设立了自智网络焦点组FG-AN，并发布了涵盖用例、机制框架和可信评估等方面的标准建议。下一步，ITU-T SG13将持续推进数字孪生网络、意图驱动网络及自智网络成效评估指标等相关标准的制定工作。

当前，自智网络的标准研究主要聚焦于架构框架与分级体系，然而在技术落地层面仍存在明显瓶颈：数字孪生与人工智能的融合机制尚未明晰，跨域场景下的动态映射与策略协同也缺乏有效解决方案。针对这一问题，本研究旨在构建原生AI与数字孪生深度融合的新型架构，重点解决高动态网络环境中的实时映射构建、自治策略的闭环验证等关键挑战，以填补现有标准在技术实现路径方面的空白。

## 2 下一代无线网络自治的挑战

6G作为新一代智能化综合性数字信息基础设施，凭借其空天地海立体覆盖、泛在互联与普惠智能等核心特征，使

得网络系统呈现出高度复杂性、动态性，并对管理能力提出更高要求，从而为实现高阶自治目标带来一系列重大挑战。

### 2.1 网络动态性与复杂性带来的精准映射与实时响应挑战

高动态性与高复杂性是6G网络的核心特征，为实现精准网络映射与实时响应带来了巨大挑战。6G网络将突破传统地面通信的局限，构建空天地海一体化的立体覆盖体系，其应用场景广泛涵盖城市、海洋、空中和太空等多样化环境。在此背景下，网络拓扑随接入智能体的移动及业务切换而动态变化，信道特性也显著受到多路径衰落、多普勒效应等时变因素的复杂影响。与此同时，6G网络需支持从传统人与人、人与物通信向智能体高效互联的范式跃迁，业务类型从语音与数据进一步扩展至沉浸式扩展现实（XR）、工业控制、自动驾驶等全新领域。各类业务在时延、带宽和可靠性等方面呈现出极大差异性，导致网络状态表现出高度的动态性和复杂性。传统网络管理方法依赖于人工预设规则与静态建模机制，难以实现对全场景、多维度网络数据（如泛在终端运行状态、多维环境感知参数等）的高效采集，更无法构建与物理网络实时同步的动态数字映射，因而难以及时响应网络拥塞、故障等动态问题，无法满足6G高阶自治对实时精准管控的根本要求。

### 2.2 网络管理的高需求与人工主导模式的适配性挑战

高管理需求是6G网络规模与复杂性指数级增长的必然

结果,传统依赖人工主导的管理模式已难以适应。6G网络的高阶自治目标要求实现涵盖规划、部署、运维与优化等全生命周期的自动化管理。然而,6G网络规模呈指数扩张,接入设备数量从百亿级向千亿级跃升,网络功能的虚拟化与切片化进一步增加了其逻辑架构的复杂性,导致传统以人为核心的管理模式面临三重困境:

1) 效率低下:依赖人工配置参数与排查故障,响应速度无法满足6G业务对实时性的严苛要求,例如工业控制中毫秒级时延的应用场景;

2) 成本高昂:为维护超大规模网络需投入大量人力资源,致使运维成本随网络扩展急剧上升;

3) 准确性不足:基于人工经验制定的策略难以适应多样化业务场景,容易因策略误判导致网络性能下降,例如切片资源分配不合理引发业务中断。

### 2.3 网络自治的“自智能能力”闭环验证挑战

实现6G的高水平自治,要求网络具备“自感知、自决策、自执行、自修复”的完整闭环能力。然而,网络的高动态性与高复杂性使得传统技术难以构建有效的验证机制:

1) 在物理网络中,策略执行后的效果评估往往伴随较高风险,例如测试新的资源调度策略可能引发实际业务中断;

2) 缺乏高保真虚拟环境的支撑,难以在极端场景(如突发自然灾害导致基站密集故障)下验证自修复与自愈策略的有效性。

这种“决策-执行”环节中验证机制的缺失,可能导致网络在高动态复杂环境中因策略失效而制约其自治的稳定性与可靠性。

### 2.4 原生AI决策的可解释性与网络可信性挑战

原生AI作为6G网络自治的核心驱动力,其决策过程的“黑箱特性”与网络的泛在联接性共同引入了可信性风险:

1) 可解释性不足:基于深度学习等AI算法的决策机制缺乏可追溯性。若因算法偏差导致策略错误(例如误判用户业务类型引发服务质量降级),管理人员难以定位问题根源,无法实施有效干预与修正;

2) 安全风险:6G网络广泛接入智慧交通、工业控制等关键基础设施。原生AI系统一旦遭受恶意攻击(如通过数据投毒篡改训练样本),可能输出错误决策(例如误导自动驾驶车辆的通信资源分配),进而引发严重安全事故。

上述问题阻碍了原生AI在关键任务场景中的可靠应用,构成了实现网络高阶自治的重要信任壁垒。

### 2.5 跨域场景下的网络自治协同挑战

6G网络旨在实现空地海一体化立体覆盖,其协同自治必然涉及地面移动通信、卫星通信与低空通信等多类异构网络,然而当前跨域协同仍面临技术挑战:

1) 跨域数据融合困难:不同域网络在技术特性上存在本质差异,如地面网络侧重高带宽与低时延,而卫星网络受限于链路稳定性与传播时延,导致各域数据在采集维度、格式与更新频率方面难以统一,无法构建支持全局决策的完整数据视图。例如,地面基站的信道状态信息、卫星的星历参数以及无人机的实时轨迹数据分属不同体系,传统数据处理技术难以实现高效融合与协同利用。

2) 跨域策略适配性不足:跨域业务(如远洋船舶经由卫星与地面指挥中心实现实时通信、无人机集群与地面控制系统的协同操作)对资源调度与故障恢复等机制提出差异化要求。依赖人工预设的静态策略难以适应复杂动态环境,如在卫星链路突发拥塞时,地面网络无法迅速做出响应以重新调配资源,从而导致业务连续性面临高风险。

## 3 基于原生AI的网络数字孪生架构

本文提出将网络数字孪生与原生人工智能深度融合,为下一代无线网络的自治能力提供关键支撑。数字孪生通过构建物理网络的精准数字化映射,建立了一个虚拟仿真环境,不仅实现了全域实时网络状态的可视化,也为各类管理策略提供了安全可靠的验证平台。原生AI则依托其用例自生成、策略自定制与智能决策能力,驱动网络实现从规划到运维的全生命周期自动化与智能化管理。二者的协同融合可动态适应复杂多变的网络环境与多样化业务需求,有效克服传统依赖人工管理在效率与准确性方面的局限,显著提升网络响应速度与决策精度。同时,依托双闭环验证机制,该架构能够保障自治策略的有效性与可靠性,增强AI决策的可解释性与系统整体可信度,并有助于强化跨域协同能力,从而为构建支持6G泛在互联与普惠智能愿景的高水平自治网络奠定坚实基础。

图2给出了基于原生AI的网络数字孪生赋能下一代无线网络自治架构,研究创新点主要体现在以下3个方面:

1) 架构创新:提出了一种原生AI与网络数字孪生深度融合的新型自治网络架构,通过构建“数据-模型-决策-验证”的敏捷闭环,实现了网络智能从外部赋能向内生融合的体系性突破。

2) 机制创新:设计了“AI用例自生成”与“网络策略自定制”两大核心机制,并引入智能服务质量(QoS)量化评估与双闭环验证框架,有效解决了传统自治网络中AI

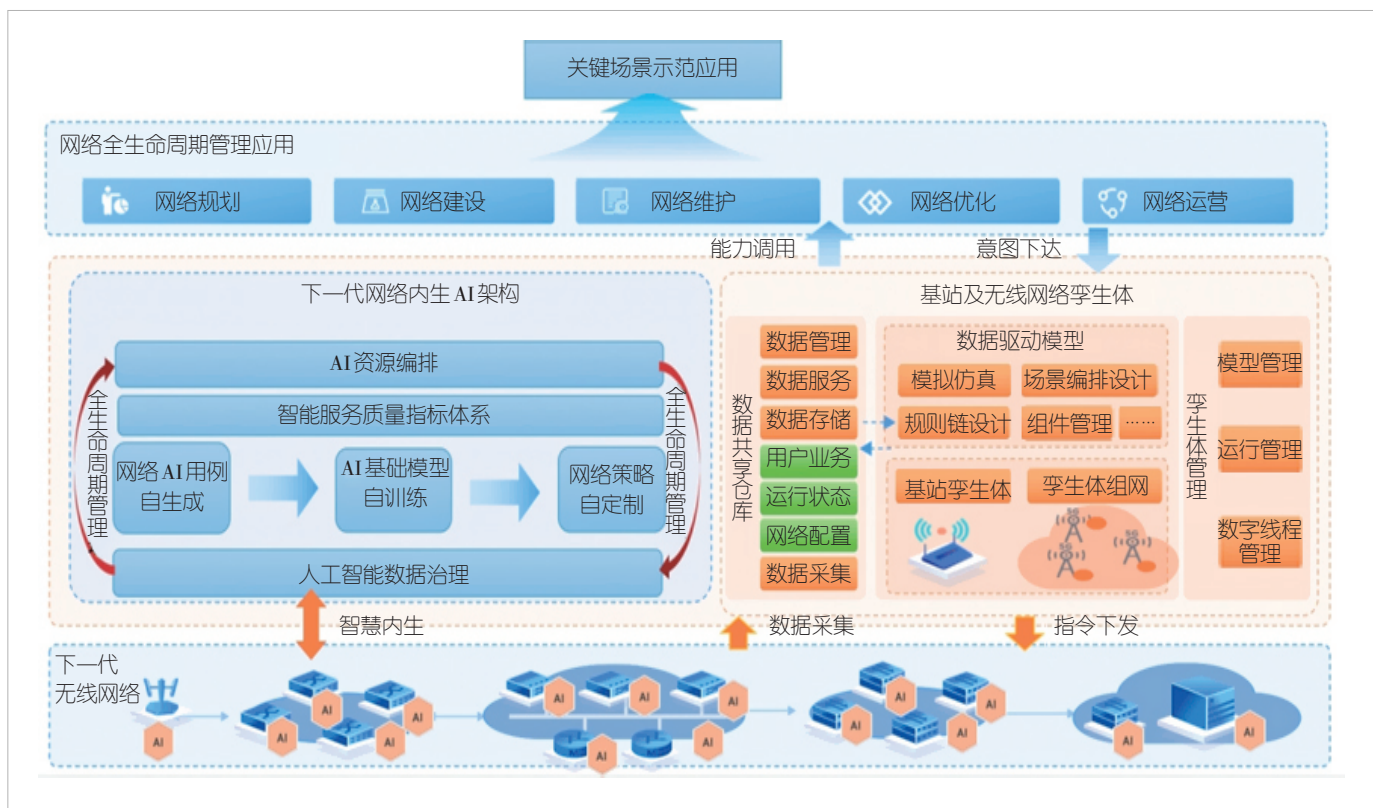


图2 基于原生AI的网络数字孪生赋能下一代无线网络自治框架

决策可解释性差与策略执行风险高的关键问题。

3) 范式创新：推动了网络运维从“被动响应”到“主动生成”的模式转变，使网络具备应对复杂动态环境的自适应与自演进能力，为6G网络实现高阶自治提供了切实可行的技术路径。

### 3.1 基于原生AI的网络数字孪生关键能力

本架构依托按需、多维、高速的无线网络数据采集、传输与存储技术，构建数字孪生基础。通过建立孪生体本体模型，构建统一表征的融合网络孪生数据库，并基于网元模型与拓扑模型的按需组合实现设备孪生体建模；借助数字线程与场景编排等技术，完成网络全生命周期——包括规划、建设、维护、优化与运营的场景建模，实现数字孪生体的全域闭环生命周期管理。通过AI用例自生成、基础模型自训练与网络策略自定制，系统实现从服务需求到网络执行的智能映射；依托AI服务质量评估与保障机制，全面量化AI服务质量需求。基于原生人工智能与数字孪生两项关键技术，分别构建原生AI无线网元与数字孪生系统，并实现二者深度融合。通过形成“数据-模型-决策-验证”的闭环体系，有效应对下一代无线网络在自治化过程中所面临的动态性、复

杂性与可信性等核心挑战。

### 3.2 基于原生AI的网络数字孪生关键技术

#### 1) 高通量的数据采集技术研究

为应对网络动态性与复杂性所带来的精准映射与实时响应挑战，本文采用按需采集与异构数据融合技术，实现空、天、地、海多场景网络数据的实时同步，为构建精准映射提供基础支撑。由于物理网络规模庞大、设备形态多样、流量信息复杂多变，全量数据采集难度极高，因此提出以按需采集为核心策略，其采集类型、频率与方法均以支撑数字孪生网络应用为目标，在保障应用需求的同时兼顾全面性与效率。在对特定网络应用进行建模时，所需数据均可从网络孪生层的数据共享仓库中高效获取。该仓库作为数字孪生网络的“单一事实源”，不仅存储海量的历史与实时网络数据，更通过统一的数据模型与标准化接口，实现多源异构数据的语义对齐与深度融合。例如，将地面基站的信道状态信息、卫星星历参数以及无人机移动轨迹等数据进行融合与标准化处理。此外，依托分布式数据同步机制与实时流处理技术，系统确保数据的一致性与高时效性，为跨场景下的全局决策提供精准、完整的数据视图。



## 2) 高精度的数字孪生建模技术研究

本文通过本体论、知识图谱与图神经网络构成的三级建模方法,构建与物理网络实时同步的动态模型,以提升映射精度。无线网络数字孪生建模主要包括设备网元单体建模和应用场景功能建模两部分。

单体建模是实现大规模网络数据一致性表征的基础。基于本体理论,对实体进行统一表征,构建统一的元数据模型,以确保不同类型、不同厂商设备之间的语义一致性。针对通信网络设备与逻辑网元,融合物理设备信息、环境信息、拓扑节点信息、网络链路信息、容器/虚拟机信息以及网元配置信息,建立无线网络数字孪生体单体模型。在复杂动态场景中,可进一步引入知识图谱与图神经网络(GNN)等技术,以更有效地捕捉网络中实体间的复杂关联与动态变化。功能建模则面向实际网络功能需求,通过整合全生命周期中的多种功能模块,实现动态演进的网络推理与决策。该模型能够根据各类网络应用的具体需求,在多维度上进行构建与扩展。

## 3) 高拟真度的数字孪生呈现

为满足自治闭环能力的验证需求,本文中我们采用高拟真可视化与场景编排技术,构建贴近真实的虚拟试验场,以支持网络策略的有效验证。数字孪生网络可视化需具备多精度网元单体与场景模型构建能力,并能够依据具体业务需求选择适当的可视化层级,从而实现对不同环境的高精度拟真。可视化范围涵盖从宏观的网络拓扑与区域覆盖,到微观的基站内部组件与信号传播路径的精细呈现,支持从小微场景至城市级规模网络的全生命周期管理可视化。该技术能够高度还原真实网络状态,并借助数字线程驱动可视化场景的运行与动态更新,通过线程对可视化内容的实时调整,实现动态可视能力。网络可视化不仅有助于用户理解网络内部结构,还可用于挖掘网络中潜藏的有价值信息。数字孪生网络的可视化面临孪生体规模庞大、虚实映射实时性要求高等挑战。为应对这些挑战,可探索基于游戏引擎的高效渲染技术及分布式可视化架构,以支持大规模、高并发环境的实时数据呈现。

## 4) 网络原生AI用例自生成

为应对网络管理的高需求与人工主导模式之间的适配性挑战,本文中我们提出通过“意图到用例”的自动转换机制替代传统人工配置,以提升网络管理效率。原生AI用例自生成技术重点研究AI用例的生成技术框架及其形式化描述。网络可基于自身数据分析或外部输入(如高层业务意图或网络态势感知结果),自动识别并生成潜在的AI用例描述。该过程可引入自然语言处理、知识推理与强化学习等技术,实

现从业务需求到可执行AI任务的自动化转换。

## 5) 基于QoAIS的原生AI网络策略自定制技术

为应对原生AI决策在可解释性与网络可信性方面面临的挑战,本文引入QoAIS量化指标体系,以实现AI策略的可追溯与可验证。QoAIS是用于评估网络原生AI服务质量的指标体系。每个AI服务对应一套QoAIS指标,而一个AI用例所对应的QoAIS则由其包含的所有AI服务的QoAIS组合构成。在网络接收到AI用例后,需明确其QoAIS要求,以便将其分解为对资源编排、调度与控制的具体指令。QoAIS要求的获取主要有两种途径:一是外部导入,即在引入AI用例描述时一并提供相应的QoAIS要求;二是内部生成,例如当网络根据上层意图信息自动生成AI用例时,可同步基于该意图生成对应的QoAIS指标要求。

## 6) 融合数字孪生网络的AI模型训练与验证技术

在原生AI场景模型的生命周期管理过程中,数字孪生网络通过高通量技术采集多维度网络状态数据,经整合处理后为AI提供输入。原生AI根据场景需求自动生成模型选择、训练、验证及决策实施策略。数字孪生凭借其高精度建模与仿真能力,在虚拟环境中对模型及策略效果进行验证,并将反馈结果用于AI迭代优化,最终形成适配方案,从而支撑无线网络规划、建设、维护、优化全流程的自动化,提升网络自治水平。数字孪生作为“安全可靠的试验场”,能够模拟各类极端场景(如大规模故障、突发流量洪峰、网络攻击等),支持故障注入与异常行为验证,从而在不干扰物理网络正常运行的前提下,全面评估AI策略的鲁棒性与有效性,显著降低新策略的引入风险。为进一步增强原生AI决策的可解释性与网络可信性,集成可解释人工智能(XAI)技术,通过决策过程可视化、特征重要性分析及反事实解释等方法,使AI的“黑箱决策”过程变得可理解、可追溯,便于管理人员定位问题并实施干预修正。

## 4 结束语

本文聚焦6G网络自治需求,提出了一种融合原生AI与网络数字孪生的新型架构。面对6G网络的高复杂性、高动态性以及高管理需求,传统人工主导的管理模式已难以适应。为应对这些挑战,本文所提架构将网络数字孪生与原生AI技术进行深度融合。该架构涵盖无线网络数字孪生的高通量数据采集、高精度建模与高保真呈现等关键技术,并依托原生AI实现网络性能预测、AI用例自生成与网络策略自定制等能力,从而完成从服务需求到网络执行的智能映射。通过构建“数据-模型-决策-验证”的闭环体系,该架构能够有效应对下一代无线网络在动态性、复杂性与可信性等方

面的核心挑战。其关键创新在于，不仅实现了多项技术的系统整合，更构建了一套内生智能、闭环自治的一体化体系，为6G网络实现真正的高阶自治提供了清晰的技术路径与理论支撑。数字孪生作为“安全可靠的试验场”，可模拟极端场景并验证AI策略的鲁棒性及有效性。本研究可为面向下一代无线网络自治的系统设计与原型开发提供理论指导。

## 致谢

感谢亚信科技(中国)有限公司的任志东、王希栋、李赞，以及中国联合网络通信集团有限公司的李凡、陈璇对本研究工作的的大力支持！

## 参考文献

- [1] Gartner. Gartner top 10 strategic technology trends for 2023 [R]. 2022
- [2] ITU-R. Future technology trends of terrestrial international mobile telecommunications systems towards 2030 and beyond [R]. 2022
- [3] ITU-T. Digital twin network - requirements and architecture: Y. 3090 [S]. 2022
- [4] TM Forum. Digital network twin for data - driven 5G services [EB/OL]. [2025-09-10]. <https://www.tmforum.org/catalysts/projects/C21.0.175/digital-network-twin-for-datadriven-5g-services>
- [5] ZHOU C, YANG H W, DUAN X D, et al. Digital twin network: concepts and reference architecture [EB/OL].[2025-08-20]. <https://datatracker.ietf.org/doc/draft-irtf-nmrg-network-digital-twin-arch/>
- [6] 3GPP. Self-organizing networks (SON), concepts and requirements (Release 8): 3GPP TS 32.500 [S]. 2008
- [7] 3GPP. Architecture enhancements for 5G system (5GS) to support network data analytics services: 3GPP TS 23.288 [S]. 2011
- [8] 3GPP. Study on enhancements for Artificial Intelligence (AI)/ Machine Learning (ML) for NG-RAN: 3GPP TR 38.743 [S]. 2024
- [9] ITU-T. Architectural framework for machine learning in future networks including IMT - 2020: ITU - T Y.3172 [S]. 2019

- [10] Next G Alliance. Next G alliance report: roadmap to 6G [R]. 2022
- [11] IMT - 2030 (6G) Promotion group. 6G network architecture vision and key technology outlook white paper [R]. 2021
- [12] 6GANA. 6G Network native AI technology requirements white paper [R]. 2022
- [13] TM Forum. IG1251 autonomous networks - reference architecture v1.0.0 [R]. 2021

## 作者简介



**王首峰**，亚信科技(中国)有限公司创新中心总经理、通信人工智能实验室主任；主要研究方向为将人工智能、大数据和IT技术融入通信网络。



**郭建超**，亚信科技(中国)有限公司标准化工程师；主要研究领域包括数字孪生网络、算力网络、联邦学习等；发表论文10余篇，获授权专利7项。



**边森**，亚信科技(中国)有限公司研发中心网络产品规划部总监；主要研究方向为面向6G的无线网络智能化技术；发表论文10余篇，获授权专利7项。

# 光纤通信技术演进与发展展望： 从基础突破到融合创新



## Evolution and Development Prospects of Optical Fiber Communication Technology: From Foundational Breakthroughs to Convergent Innovation

张海懿/ZHANG Haiyi

(中国信息通信研究院, 中国 北京 100191)

(China Academy of Information and Communications Technology, Beijing 100191, China)

DOI: 10.12142/ZTETJ.202505007

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20251016.1357.002>

网络出版日期: 2025-10-16

收稿日期: 2025-08-18

**摘要:** 总结了光纤通信技术从基础突破到融合创新的发展历程, 分析了光纤通信从光层基础到多域组网的关键技术进展, 并展望了其未来发展趋势, 同时指出中国面临的发展机遇和挑战。针对人工智能、6G 等未来发展新需求, 建议业界继续协同聚力, 推进光纤通信关键技术的基础突破和融合创新, 支撑信息基础设施高质量发展。

**关键词:** 光纤通信技术; 技术突破; 融合创新; 发展展望

**Abstract:** A review is provided of the development of optical fiber communication technology, from fundamental breakthroughs to integrated innovation. The analysis covers key technological advances spanning from optical-layer fundamentals to multi-domain networking, followed by a discussion of future trends and the corresponding opportunities and challenges for China. In response to emerging requirements such as artificial intelligence and 6G, it is recommended that the industry continue collaborative efforts to advance basic breakthroughs and integrated innovation in key optical fiber communication technologies, thereby supporting the high-quality development of information infrastructure.

**Keywords:** optical fiber communication technology; technical breakthrough; convergent innovation; development prospect

**引用格式:** 张海懿. 光纤通信技术演进与发展展望：从基础突破到融合创新 [J]. 中兴通讯技术, 2025, 31(5): 37-49. DOI: 10.12142/ZTETJ.202505007

**Citation:** ZHANG H Y. Evolution and development prospects of optical fiber communication technology: from foundational breakthroughs to convergent innovation [J]. ZTE technology journal, 2025, 31(5): 37-49. DOI: 10.12142/ZTETJ.202505007

## 1 光纤通信技术发展历程

### 1.1 光纤通信基础技术突破阶段

光通信的起源可追溯至古代的“烽火台”，这是一种基于目视的简易信息传递方式。1880年，贝尔与其助手发明了“光电话”，通过光作为载波来调制并传输声音，奠定了现代光通信的基础。然而，受限于光源质量与传输介质的可靠性等关键技术瓶颈，光通信技术在此后数十年始终未能实现实际应用。尽管期间不乏新的尝试（例如利用玻璃等介质进行光信号传输的实验），但均未能取得决定性突破。

自20世纪60年代起，光通信在光源与传输介质领域陆续取得根本性突破。1960年，美国科学家梅曼发明红宝石激光器，为通信所用半导体激光器的出现与发展奠定了重要

基础。1962年，美国通用电气公司工程师成功研制出首台基于砷化镓材料的同质结半导体激光器。1966年，英国标准电信研究所华裔科学家高锟博士提出，光导纤维的高损耗主要源于材料中所含杂质，通过降低杂质含量并改进制备工艺，可将光纤损耗降至20 dB/km。1970年，美国、日本等相继研制出可在室温下连续工作的双异质结半导体激光器；同时，美国康宁公司拉制出世界上第一根损耗为20 dB/km的石英光纤。光源与光纤介质取得的重大突破，开启了光通信技术发展的新纪元，光纤通信由此逐步成为信息传输的主要方式之一。在光源、光纤等基础理论及关键器件实现重大突破之后，光纤通信技术在后续数十年中持续革新并实现规模化商用。中国在该领域也经历了从跟随、追赶、并跑到部分领先的发展历程。



## 1.2 中国光纤通信技术发展

### 1.2.1 技术跟随与追赶阶段

中国光纤通信技术从20世纪70年代至21世纪初,整体处于持续跟进和追赶其他国家先进技术的发展阶段。该时期的技术突破主要体现在从多模光纤传输系统的实验与商用,逐步过渡到单模光纤传输系统的引入与规模部署,以及从准同步数字体系(PDH)、同步数字体系(SDH)等传输体制,演进至波分复用(WDM)大容量传输技术的初步应用与商业化。

1976年,美国部署了首条45 Mbit/s的多模光纤实验链路。随后,欧美日等地区或国家逐步推进光纤通信的商业化,传输速率从数百Mbit/s不断提升至10 Gbit/s,传输距离也从10 km级扩展至200 km级。期间出现的关键技术突破包括:单模光纤的研制与成功应用、光放大器技术的发明,以及半导体激光源的持续革新——如采用量子阱等新型增益材料结构,实现850 nm、1 310 nm和1 550 nm等多波段的覆盖,并显著提升激光器的能量转换效率、输出功率和稳定性。WDM技术的引入进一步推动了系统容量的增长。

1979年,武汉邮科院副总工赵梓森团队拉制出中国第一根实用化光纤,同时上海冶金所、武汉邮科院等已研制出通信用光源发光二极管(LED),标志着中国光纤通信逐步进入实用化阶段。1982年,中国第一条实用化光纤通信线路在武汉建成,贯通武昌、汉阳、汉口3镇,全长13.3 km,速率为8.448 Mbit/s。1986年,中国第一条长途光纤通信线路在武汉至荆州之间建成,速率达34 Mbit/s<sup>[1]</sup>。从1988年开始,中国启动了为期10年的“八纵八横”骨干光纤网络建设,为后续国家信息通信网络发展奠定了坚实基础。在SDH和WDM系统初步部署方面,1997年中国首个采用国际电信联盟电信标准分局(ITU-T) SDH标准速率为622 Mbit/s光纤通信线路在攀枝花建成。1999年,中国首个8×2.5 Gbit/s WDM光纤通信线路在青岛至济南之间建成。

### 1.2.2 并跑和局部领先阶段

自2000年至今,中国逐步进入与国际先进水平并跑、在部分领域实现领先的发展阶段。光纤通信系统传输速率逐步从10 Gbit/s向40 Gbit/s、100 Gbit/s乃至400 Gbit/s升级演进,与欧美国家基本同步推进技术商用。在数字传输与组网技术方面,中国在多业务传送平台(MSTP)、分组传送网(PTN)、切片分组网(SPN)、光传送网(OTN)和光分插复用设备(ROADM)等技术方面,已实现全球技术应用引领或部署规模领先。在面向800 Gbit/s及以上速率、波段扩展、

空分复用、新型光纤等持续扩容的新型传输技术方面,除部分高端芯片制备工艺受限之外,中国的研发节奏基本与其他国家保持一致。另外,随着数据流量的爆发式增长和光纤通信技术的持续革新,宽带光纤接入技术逐步兴起。在国家政策引导和技术产业界的共同推动下,光纤到户(FTTH)成为中国宽带接入网络的重点发展方向。在此期间,中国光纤接入网建设持续加速,吉比特无源光网络(GPON)/以太网无源光网络(EPON)、10G无源光网络(10G PON)等关键技术以约十年一代的节奏稳步推进,实现了代际规模的有序部署。近年来,中国更率先提出光纤到房间(FTTR)技术,并引领其全球规模化部署进程。截至2025年,50G PON技术已进入全国试点应用阶段。无论是在技术选型还是在实际部署层面,中国均已引领全球光纤接入网络的发展潮流。

### 1.2.3 网络部署最新进展

经过多年发展,光纤通信网络已成为新型信息基础设施的关键承载底座,对5G、人工智能(AI)、大数据与算力等新兴业务发展起到关键支撑作用。在此过程中,中国逐步建立起坚实的产业发展基础,在通信设备、光纤光缆和光模块器件等领域已涌现出多家位列全球前十的企业。中国政府高度重视光纤通信网络的建设与发展,近年来已将千兆光网纳入国家顶层规划。2021年,工信部发布《“双千兆”网络协同发展行动计划(2021—2023年)》,明确提出推进千兆光网和5G协同发展。2024年9月,工业和信息化部(后简称工信部)联合多部委发布《关于推动新型信息基础设施协调发展有关事项的通知》,强调要深入开展“双千兆”网络建设,协同推进5G与千兆光网,推动IP承载和光传输融合发展,促进接入网、城域网和骨干网同步扩容升级。2025年1月,工信部办公厅发布关于开展万兆光网试点工作的通知。在政府积极引导及产学研协同推进下,中国光纤通信网络建设与应用成效显著,网络规模与应用水平已居全球领先地位。截至2025年3月<sup>[2]</sup>,全国支持千兆网络服务的10G PON端口数达2 925万个,千兆宽带用户数突破2.18亿个,光缆线路总长度达7 454 × 10<sup>7</sup> km。中国基于ROADM的全光网络规模全球领先,自2023年下半年起启动400 Gbit/s干线传输网络部署,多家运营商持续推进建设。据不完全统计,截至2024年12月底,全网部署的400G长距接口已超过1.29万个。在新技术方面,近两年已逐步开展50G PON、800G速率智算拉远、空心光纤等试点验证工作。当前,中国光纤通信网络正处于千兆普及、万兆启航的关键发展阶段,未来将持续依托5G/6G、大数据、算力与人工智能等新型需求驱动,不断推动关键技术的创新与演进。

2 光纤通信基础技术融合演进

2.1 信号调制和检测

随着各类业务传输需求的持续驱动及光纤通信技术的不断革新，光纤通信系统传输速率已从最初的数十 Mbit/s 提升至 Tbit/s 量级。相应地，光信号的发送与接收技术即调制与检测等基础光层处理技术也融合了技术方案、制备工艺与新型材料等多维度创新，持续推动其向前演进。

1) 强度调制直接检测 (IM-DD)

在光纤通信技术发展初期至 2000 年前后，商用系统中的信号调制与检测基本采用 IM-DD 机制。激光光源外延工艺从扩散法逐步演进为分子束外延 (MBE)、金属有机化合物气相淀积 (MOCVD) 等或多种结合的方式<sup>[3]</sup>。光源类型也经历了从 LED 发展到法布里-珀罗 (FP-LD)、分布式反馈 (DFB-LD)、电吸收调制激光器 (EML) 和垂直腔面发射激光器 (VCSEL) 等多种形态并存。通常情况下，短距离传输采用内调制方式，长距离传输采用外调制方式，相应的工作波长也从 850 nm (短距离) 逐步扩展到 1 310 nm (中短距离) 和 1 550 nm (长距离)。在此期间，得益于相干接收显著提升灵敏度等优势，光通信领域研究人员曾积极开展相干光通信技术的研究和探索。然而，由于该技术对光信号频率与相位处理性能要求过高，同时掺铒光纤放大器 (EDFA)、WDM 等新技术引入有效解决了传输容量和传输距离瓶颈问题，基于相干光通信的应用探索暂时搁置。

自 2000 年起，随着互联网的大规模部署与应用，特别是云计算、大数据等需求推动数据中心基础设施的规模建设，以以太网接口为代表的短距离传输速率持续提升。2010 年，随着 40 GE/100 GE 标准 (IEEE 802.3ba) 的发布，如何选择新的调制方式以满足高带宽、低成本的应用需求，成为业界关注的焦点。经过多轮的热评估与讨论，2017 年发布的 200 GE/400 GE 标准 (IEEE 802.3bs) 正式引入了基于四电平脉冲幅度调制 (PAM-4) 的强度调制方案，并继续沿用波分复用与光纤复用等技术。随后，在 2024 年发布的基于单通道 100 Gbit/s 的 800GE 标准 (IEEE 802.3df)，以及制定中的基于单通道 200 Gbit/s 的 800 GE 与 1.6 TE 标准 (IEEE 802.3dj) 中，这一调制方案继续被采纳。

随着信号处理速率持续提高 (当前商用产品普遍处于 50 ~ 100 Gbaud 水平，并逐步向 200 Gbaud 以上演进)，采用 PAM-4 调制的信号传输距离进一步缩短。业界正在持续评估是否需在 10 km 乃至更短距离中引入低成本相干检测技术，以应对此类传输挑战。

2) 多阶调制相干检测

自 2000 年起，随着 10 Gbit/s 干线传输网络的逐步商用部署，下一代传输速率的研究和选择也逐步提上日程，其中 SDH 和 WDM 技术均将 40 Gbit/s 作为重点探索方向。在该过程中，为提升传输性能，40 Gbit/s WDM 系统逐步引入了光双二进制码 (ODB)、差分相移键控 (DPSK)、差分四相相移键控 (DQPSK) 等多阶调制传输码型，接收侧仍采用直接检测，相关技术在实际网络中实现了小规模商用。2005 年，数字载波相位估计技术在相干接收机中成功实现演示，重新激发了业界对相干光通信的广泛关注。2008 年，北电网络推出了首个基于 DP-QPSK 传输码型、采用 DSP 技术的商用 40 Gbit/s 相干系统<sup>[4]</sup>。随后阿尔卡特朗讯于 2009 年推出 100 Gbit/s 单波长光转发器并在 Verizon 网络启动商用。同期，中国设备商逐步推出了 100 Gbit/s 相干设备，运营商也于 2011 年启动了 100 Gbit/s 速率的干线网络建设，这标志着相干光通信正式进入大规模商用阶段。2023 年，中国启动了 400 Gbit/s 相干系统规模商用。近几年，业界也持续开展 800 Gbit/s ~ 1.2 Tbit/s 速率的技术试验与验证，信号处理速率超过 200 Gbaud 的 1.6 Tbit/s 系统在全球已有商用案例。目前业界公开的典型数字信号处理 (DSP) 芯片处理能力见表 1。

3) 调制和检测用光电材料

随着光通信信号速率持续以 4 倍或 10 倍为单位代际提升，为支撑更大带宽、更高集成度与更低能耗等高性能需求，新型光电子材料不断涌现。在调制和检测环节，激光器与探测器主要采用 III-V 族化合物磷化铟 (InP) 和砷化镓 (GaAs) 作为芯片衬底材料。其中，InP 衬底主要应用于 FP、DFB、EML 等边发射激光器芯片，以及 PIN、APD 等探测器芯片；GaAs 衬底则主要应用于 VCSEL 面发射激光器芯片。近 20 年来，随着硅光技术的发展。基于硅基 (Si) 外延锗材料的探测器芯片也取得明显进展，目前探测器带宽实验室水平达到 100 GHz 量级<sup>[5]</sup>，与 InP 基材料的 200 GHz 水平仍存在一定差距。在调制器材料方面，除最早应用的 LiNbO<sub>3</sub> 体材料和 III-V 族材料 (如 InP) 外，基于硅光和 LiNbO<sub>3</sub> 薄膜的调制技术发展迅速。目前，硅光调制器、LiNbO<sub>3</sub> 薄膜调制器的

表1 业界公开的数字信号处理芯片处理能力进展

厂家	波特率/ Gbaud	支持速率	工艺/nm	发布时间/ 年
思科	136	800 Gbit/s ~ 1.2 Tbit/s	5	2023
NEL	140	800 Gbit/s ~ 1.2 Tbit/s	5	2023
富士通	140	800 Gbit/s ~ 1.2 Tbit/s	5	2023
诺基亚	130+	800 Gbit/s ~ 1.2 Tbit/s	5	2023
英飞朗	148	800 Gbit/s ~ 1.2 Tbit/s	5	2024
Ciena	200	800 Gbit/s ~ 1.2 Tbit/s	3	2024

带宽已超过 110 GHz<sup>[6-7]</sup>, 其中 LiNbO<sub>3</sub> 薄膜调制器的带宽提升潜力要更大。此外, 业界也在探索其他多种材料的调制技术, 如基于铁电薄膜绝缘体 (PLZT) 调制器已有 70 GHz 带宽实验报道<sup>[8]</sup>。为应对未来 200 Gbaud 及以上速率对高集成、低功耗与高性能信号处理的要求, 除了上述材料外, 业界正在积极拓展如钛酸钡薄膜、等离子体、二维材料、有机材料等新兴材料的应用。如等离子体器件带宽可达 110 GHz<sup>[9-10]</sup>, 有望超过 500 GHz<sup>[11]</sup>, 这将共同推动 200 Gbaud 以上的光电子芯片器件的进一步发展。

在光纤通信系统中, 光信号调制和检测的功能主要由光模块来实现。近 10 年来, 光模块及芯片器件的发展情况及趋势如图 1 所示。

## 2.2 光信号放大

在光纤通信技术的发展进程中, 20 世纪 80 年代后期发明的 EDFA 是一个重大里程碑, 彻底解决了信号在长距离传输中的全光中继问题。随着光信号速率不断提高、传输波段逐渐扩展以及传输距离持续增加, 基于不同原理与材料的多种宽谱光放大器相继涌现, 并呈现出加速融合发展与持续演进的趋势。

### 1) 放大器基本类型

目前光纤通信系统中光放大器根据工作机制和介质等特点主要分为 3 类: 第 1 类是稀土掺杂类纤放大器, 以掺杂光纤作为增益介质, 使用特定波长 (如 980 nm 或者 1 480 nm 等) 激光作为泵浦源, 通过稀土离子 (如铒、铥、镨等) 的亚稳态能级将泵浦能量转移至信号光以实现放大。该类放大器主要包括 EDFA、掺铋光纤放大器 (BDFA)、掺铥光纤放大器 (TDFA)、掺镨光纤放大器 (PDFA) 等, 其中 EDFA 在 SDH、WDM 等系统中得到广泛商用, 是目前光纤通信系统中最主要应用的放大器; 第 2 类是半导体光放大器 (SOA), 目前已在光模块内部信号放大等场景实现局部商用; 第 3 类为基于非线性效应的参量放大器, 利用增益介质 (光纤、硅基等) 的非线性效应实现信号放大, 包括已局部商用的光纤拉曼放大器 (FRA), 以及正在研究的硅基拉曼放大器 (SRA)、光纤布里渊放大器 (FBA)、光纤参量放大器 (FOPA) 和硅基参量放大器 (SOPA) 等。上述各类光放大器所支持的增益谱宽如图 2 所示<sup>[12]</sup>。

### 2) 宽谱信号放大

综合考虑光源与光纤等技术特性, 光纤通信系统的长距离传输波段主要集中于 C 波段 (1 550 nm 窗口), 该波段也是 EDFA 的主要应用范围。在部分传输距离要求更高的应用场景中, 常结合使用 FRA 与 EDFA 协同放大光信号。随着信

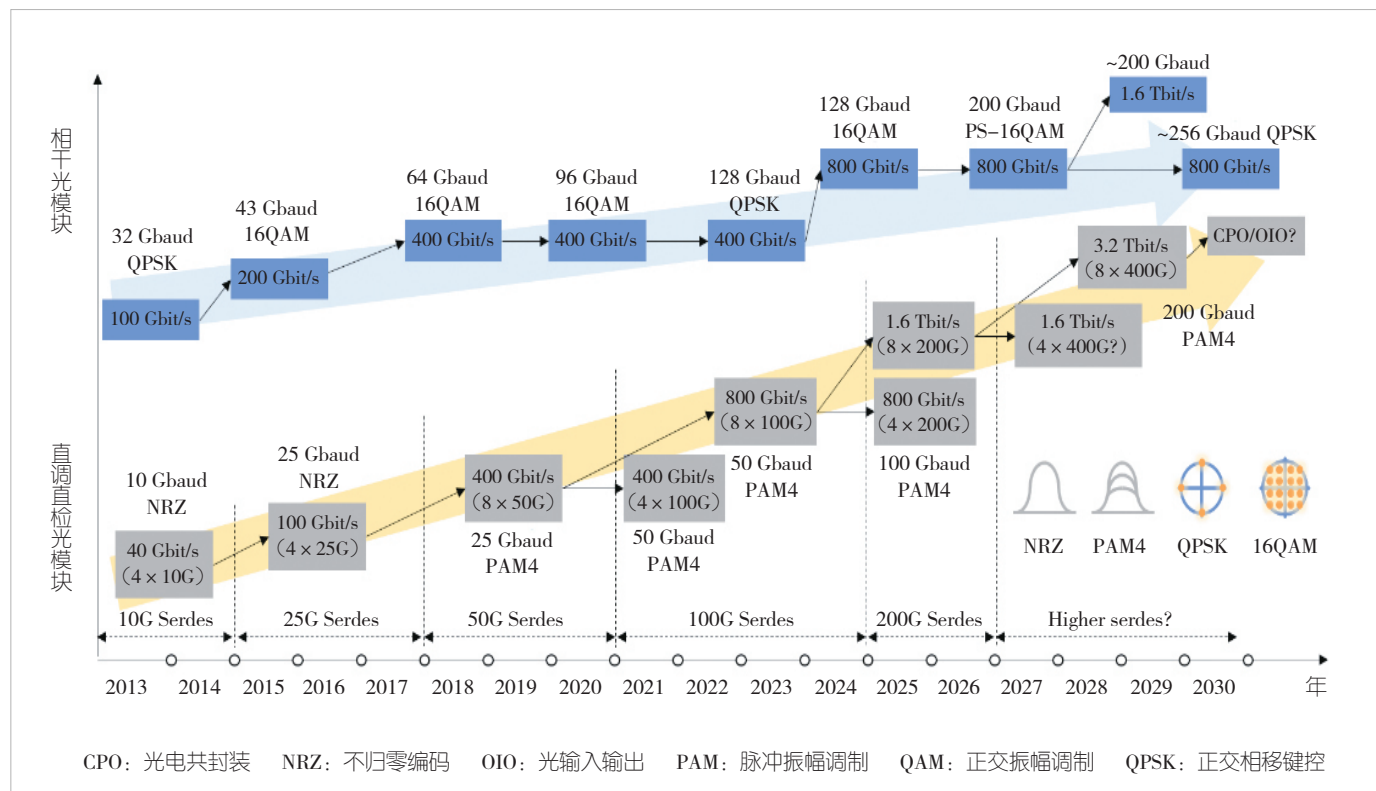


图1 光模块及收发芯片器件发展趋势



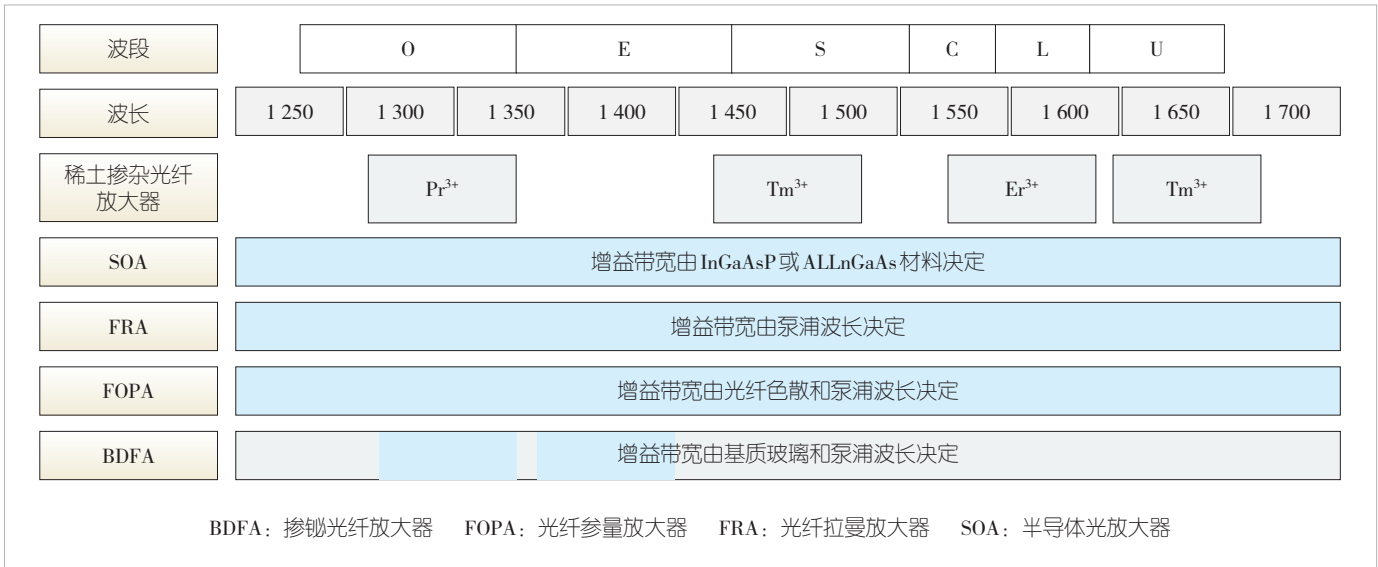


图2 各类光放大技术的增益谱宽

息传输需求的持续增长，光纤通信的传输波段不断扩展。例如，在部署 400 Gbit/s 系统时，传输窗口的带宽已从常规 C 波段的 4.8 THz 扩展至 C 波段与 L 波段各 6 THz。然而，目前能够同时覆盖 C+L 波段的一体化光纤放大器尚未形成成熟的技术方案。传统并联式结构存在增益带隙死区的问题；采用铋铈共掺光纤的光放大器虽可覆盖 C+L 增益带宽，但存在增益均衡性较差的问题。如何实现高增益、低噪声的宽带放大仍是当前研究的重点。

为支持未来 O、E、S、U 等更多波段的信号放大应用，业界正在积极研究其他类型的稀土掺杂光纤放大器。目前该类研究整体仍处于实验室探索阶段，尚未成熟<sup>[13]</sup>。此外，面向更宽传输波段的应用需求，SOA、FRA 及参量放大器等也是具有竞争力的技术方案。然而，除 SOA 在 1 310 nm 窗口、FRA 在 C+L 波段已有一定应用外，大多数宽带放大技术仍处于研究阶段。例如，2017 年诺基亚贝尔实验室采用 SOA 在 100 km 光纤链路上实现了超过 100 nm 的多波段信号放大<sup>[14]</sup>；2020 年阿斯顿大学采用了两级分布式拉曼放大（DRA），在 70 km 单模光纤上实现了 1 475~1 625 nm 的连续谱放大<sup>[15]</sup>；2022 年 NTT 利用基于周期极化铌酸锂（PPLN）的 FOPA 结合分布式拉曼放大，实现了两个 6.25 THz 波段（S+C 或 C+L）的传输实验<sup>[16]</sup>等。面向空分复用（SDM）等多维复用场景的未来需求，SDM 光放大器仍需在增益光纤结构设计、泵浦方式创新及制备工艺优化等方面开展深入研究。

2.3 传输光纤

自传输光纤实现商业化应用以来，其已逐步替代铜缆成

为固定通信网络最主要的传输介质。传输光纤的应用重塑了通信网络架构、产业格局与应用生态。在更高传输性能与技术持续革新的需求推动下，传输光纤正不断融合多维新兴技术，持续向前演进。

1) 基本类型

按照光纤支持信号模式传输能力的差异，通信用光纤主要分为多模光纤和单模光纤。多模光纤主要适用于局域网以及近年来加速发展的数据中心内部短距互联的应用场景。随着信号传输速率从数十 Mbit/s 到 1 Gbit/s、10 Gbit/s，再到 100 Gbit/s、1 Tbit/s 及以上的逐步提升，多模光纤也在不断演进。目前多模光纤主要标准包括 OM1 到 OM5 等多个系列，其中 OM1 和 OM2 主要采用 LED 光源，而 OM3、OM4 和 OM5 采用垂直腔面发射激光器（VCSEL）光源，后者已成为当前主流或即将广泛使用的多模光纤类型。与此同时，业界也正在开展结合 VCSEL 光源优化性能后的波长（如 1 310 nm 等）多模光纤研制<sup>[18]</sup>。

单模光纤通常应用于干线网、城域网、接入网以及传输距离较长的局域网和数据中心等应用场景。随着激光器性能提升和传输技术的演进等，自 20 世纪 80 年代以来，基于光纤色散、损耗、抗弯能力等关键特性，单模光纤逐步形成 G.652、G.653、G.654、G.655、G.656、G.657 等六大类型，如表 2 所示。每种类型又可根据损耗、色度色散、偏振模色散等特性细分为不同的子类。G.652 是目前应用最为广泛的单模光纤，其超低损耗变种于 2007 年前后研制成功，并与 G.652 标准兼容，目前已实现规模商用。相比于 G.652D，G.657 光纤具有更优的抗弯性能，逐渐应用于光纤到房间等

表2 单模光纤基本类型及特性

ITU-T 标准号	单模光纤标准名称	子分类	性能特点
G.652	非色散位移单模 光纤	G.652A	适用的波段范围为O波段和C波段,在1 550 nm波长的最大损耗为0.4 dB/km
		G.652B	适用的波段范围为O波段、C波段和L波段,在1 550 nm波长的最大损耗为0.35 dB/km
		G.652C	在G.652.A基础上消除了1 383 nm附近的水峰,在1 550 nm波长的最大损耗为0.3 dB/km
		G.652D	在G.652.B基础上消除了1 383 nm附近的水峰,在1 530~1 565 nm波长范围的最大损耗为0.3 dB/km
G.653	色散移位单模光纤	G.653A	在1 550 nm波长附近具有标称零色散,在1 550 nm波长的最大损耗为0.35 dB/km
		G.653B	在G.653A基础上定义了1 460~1 625 nm波长范围内的色散系数与波长对应的关系
G.654	截止波长移位单模 光纤	G.654A	截止位移光纤的一个基础分类,模场直径大、损耗低
		G.654B	模场直径比G.654.A更大,适用于长距、大容量的WDM传输系统,如海底传输系统
		G.654C	光纤规格特性跟G.654A类似,PMD的要求更严格
		G.654D	光纤规格特性与G.654B类似,但降低了衰减,在1 550 nm波长的最大损耗为0.2 dB/km
		G.654E	规格特性与G.654B类似,更适用于陆缆传输系统
G.655	非零色散移位 光纤	G.655A	在1 550 nm波长附近衰耗最小、色散较小且不为0,可以用于WDM系统
		G.655B	光纤规格特性与G.654A类似,色散斜率更小
		G.655C	光纤规格特性与G.654B类似,PMD的要求更严格
		G.655D	色散特性在C波段和L波段都有较好的表现,更适用于L波段的波分复用系统
		G.655E	光纤规格特性与G.654A类似,对有效面积和非线性效应的控制更严格
G.656	宽带光传输用 非零色散光纤	——	相比G.655支持的波长范围更广,定义了1 460~1 625 nm波长范围内色散大于某个非零值的单模光纤
G.657	弯曲损耗不敏感 光纤	G.657A	部署在接入网中,可用于1 260~1 625 nm波长范围
		G.657B	通常用于接入网末端的短距离(小于1 000 m)连接,特别是在建筑物内部或建筑物附近,可以用于1 260~1 625 nm波长范围

ITU-T: 国际电信联盟电信标准分局    PMD: 偏振模色散    WDM: 波分复用

场景。

2) 新型光纤

随着信号传输速率的持续提升和单模光纤非线性传输逐渐接近香农极限<sup>[19]</sup>,如何研制出传输性能更优、支撑更大传输容量的新型光纤,已成为业界关注的重点。

目前备受关注的新型光纤主要包括超低损耗大有效面积光纤(G.654E)、空分复用光纤和空芯光纤等。自2016年其标准正式发布以来,G.654E光纤已逐步成为干线网络建设的主要选择。据不完全统计,在2024年建设的光缆中,G.654E光纤占比超过80%。

自2010年非线性传输香农极限提出后,空分复用技术已成为未来传输扩容的重要研究方向。根据复用维度差异,空分复用光纤主要包括多芯光纤、少模光纤和少模多芯光纤3种。目前除多芯光纤在海缆系统中有初步应用外,其他两类主要仍处于实验室探索阶段。例如,NICT于2024年报道了基于15种强耦合模式、在1 001 km传输距离上实现的273.6 Tbit/s数据速率与273.9 Pbit/s·km容量距离乘积,创下了多模传输的最高纪录<sup>[20]</sup>。2024年,北大与长飞联合,基于7芯×10模光纤完成了55 km的弱耦合SDM传输系统搭建,

传输容量为5.27 Pbit/s<sup>[21]</sup>;2025年,NICT通过单模纤芯分布实现了光频梳再生技术,以PDM QPSK调制方式在13 km的39芯少模多芯光纤中完成了传输实验,这验证了12.7 Pbit/s的传输容量<sup>[22]</sup>。面对单模光纤传输容量瓶颈与信息流量快速增长之间的矛盾,空分复用光纤的研究与应用探索预计将持续推进。

自20世纪80年代光子晶体概念提出以来,业界对空芯光纤的研究持续不断。2010—2017年,随着反谐振包层结构的提出,空芯光纤的损耗从早期的约1 000 dB/km显著降低至7.7 dB/km左右。2018年,英国南安普顿大学提出的嵌套反谐振无节点光纤(NANF)将损耗降至1.3 dB/km。通过后续结构与工艺的持续优化,2024年基于双嵌套反谐振无节点结构(DNANF)的空芯光纤损耗已降至0.11 dB/km,低于普通实芯单模光纤在1 566 nm波长处的理论损耗极限(0.139 7 dB/km)<sup>[23]</sup>。2025年,中国长飞公司在OFC会议上报道了0.05 dB/km的最新记录。鉴于空芯光纤具备超宽带、超低时延、超低传输损耗和超低非线性等本征优势,近年来其在损耗性能上的突破引起了业界的广泛关注。例如,微软收购了在空芯光纤技术方面具有优势的英国Lumenisity公

司,以推动其应用探索<sup>[24]</sup>;中国电信、中国移动和中国联通三大运营商也联合企业及高校团队,积极开展相关技术与试点应用<sup>[25-27]</sup>。展望未来,空芯光纤仍处于发展初期,面临工艺尚未成熟、应用特性需深入验证,以及光缆建设与替换周期较长等挑战。预计空芯光纤将持续完善,并有望在延时敏感的数据中心互联等场景中优先开展试点应用。

## 2.4 光子集成与光电融合

随着光纤通信技术的持续发展,业界不断探索如何将多种光芯片或光电分立器件进一步集成,以实现更高集成度、更低能耗和更优性能的光子集成技术。与此同时,基于光、电技术各自优势进行功能协同与融合的光电融合技术,已成为近年的研究热点。

### 1) 光子集成 (PIC)

PIC通过将相同或不同功能的分立光器件集成在一起,其概念最早可追溯至20世纪60年代末<sup>[28]</sup>。到20世纪80年代,在光通信快速发展的推动下,光子集成技术取得了显著进展,阵列波导光栅等重要无源器件相继被研制出来。1986年,首个商用PIC——电吸收调制激光器(EML)问世<sup>[29]</sup>。2000年以后,随着硅基光电子技术的逐步兴起,PIC的集成度进一步提升,进入新的发展阶段。从集成规模来看,中小规模PIC技术已较为成熟并实现广泛商用;大规模PIC在实验室中的集成度可达 $10^4 \sim 10^5$ 量级<sup>[30]</sup>。在材料方面,早期PIC平台以III-V族材料(如InP)为主;近年来,硅基光电子凭借其在更高集成度、互补金属氧化物半导体(CMOS)兼容性与低成本制造方面的优势日益突出。未来,以硅光为基础平台,结合不同材料体系的异质异构集成技术,具有广阔的发展前景。

从集成方式来看,异构/混合集成是当前重要的发展方向,涵盖光芯片与电芯片之间的集成以及光芯片与光芯片之间的集成。在光电芯片集成中,连接方式已从传统的引线键合逐步演进至2.5D芯片倒装和3D集成技术,如硅通孔(TSV)和氧化物通孔(TOV)。3D集成通过在垂直方向上堆叠光电芯片,可实现更短的互连长度、更高的集成密度和更优的高频性能,然而受限于散热问题尚未完全解决,目前仍以2.5D集成为主,台积电、英特尔等企业已较早布局该领域。在光芯片之间的集成方面,异构/混合集成主要用于实现III-V族激光器与硅光芯片的连接,集成结构也从平面走向立体,涵盖多种2.5D/3D封装方式,其中3D倒装是目前的主流技术方案。

从光子集成技术的长远发展来看,硅基单片集成被认为是未来的终极目标。异质集成能够融合多种材料的优势,实

现系统性能的最优化,其主要实现方式包括晶圆级键合与异质外延生长两种。晶圆级键合通过化学或物理作用以及光刻对准等技术将两片同质或异质晶圆紧密结合,随后再制备芯片结构。该技术目前面临晶圆尺寸匹配、表面处理精度高、无法预先筛选已知良品芯片等挑战,预计仍将在中短期内作为主流技术路线。异质外延生长则是在已制作好的晶圆上,选区外延生长其他材料,进而构建芯片结构。若能在光耦合和异质选区外延生长等关键环节实现突破,硅基单片集成有望成为最接近传统CMOS工艺的异质集成方案,并将作为长期发展的重要方向。

### 2) 光电融合

光电融合基于光域与电域的技术特性,旨在实现信息处理功能的高效协同或融合。其概念与内涵较为宽泛,本文主要聚焦于面向连接场景的典型光电融合技术——光电共封装(CPO)与光输入输出(OIO)。这些技术推动光连接从板级、设备级向芯片级演进,逐步实现芯片级光互联。

CPO已成为绿色数据中心领域的热点候选技术。该技术将光引擎与专用集成电路(ASIC)芯片共同封装在同一高速基板上,可显著降低电学损耗,从而减少信号衰减,降低系统功耗和成本,并实现更高的集成度。自2020年起,多家企业相继发布CPO样机,其传输容量持续提升。同年,在美国光纤通信展上,英特尔发布了首款25.6 Tbit/s的CPO样机。2022年,美满公司展示了1.6 Tbit/s的CPO光引擎,未来计划支持其51.2 Tbit/s交换机。博通于2024年展示了51.2 Tbit/s CPO交换机及6.4 Tbit/s FR4光引擎,并于2025年6月发布了102.4 Tbit/s交换机。2025年,英伟达宣布即将推出分别支持InfiniBand和以太网的CPO交换机,其带宽分别达到115.2 Tbit/s和409.6 Tbit/s。

目前,CPO被视为提升单通道速率、实现光芯片与交换芯片间单通道速率超过400 Gbit/s的关键技术方案之一,预计将迎来市场的显著增长。据预测,至2029年,3.2 Tbit/s CPO端口数量将超过1 000万个<sup>[31]</sup>。

OIO技术是目前业界持续探索的研究方向。该技术面向存算网络,通过将计算/存储芯片与光芯片进行封装集成,实现与外部其他芯片之间的高速光互连。Ayar Labs在该领域开展了持续研究,于2023年展示了与英特尔现场可编程门阵列(FPGA)集成的OIO解决方案,可实现双向4 Tbit/s的数据传输。Nvidia、AMD、Intel、NTT等多家行业巨头已对Ayar Labs进行投资,旨在借助其OIO技术突破人工智能数据传输的瓶颈。

2024年,英特尔发布了与中央处理器实施三维共封装的OIO芯粒,其双向带宽达到4 Tbit/s;曦智科技于2023年



推出光互连产品 Photowave, 并在 2024 年将 OIO 技术应用于新华三的 CXL-O 光互连解决方案中。目前, 台积电正在开发面向 OIO 应用的 COUPE 硅光平台, 该平台将扩展处理器 (XPU) 芯片与 CPO 光引擎集成于同一中介层上, 预计可实现功耗降低 90% 以上、延迟减少 95% 以上的显著效益<sup>[32]</sup>。

整体而言, 光子集成技术正沿两个主要方向演进: 一方面, 器件集成度持续提高, 实现了从分立器件到异质异构集成, 并进一步向单片集成发展; 另一方面, 功能复杂度不断提升, 从无源器件扩展到有源/无源混合集成, 并朝着光电融合集成的方向推进, 最终目标为实现“光-电-连-算”的一体化融合。

### 3 光纤通信组网技术融合演进

经过数十年发展, 光纤通信网络已成为信息基础设施的核心承载平台, 在干线网络、城域网、数据中心互联及接入网等多个领域实现了多技术融合与协同演进。面对 AI 与算力需求提升以及 5G/6G 等新型业务承载挑战, 当前光纤通信网络正持续向更大容量、更远距离、更高智能灵活性及更优能效的方向演进。

#### 3.1 干线网络大容量传输技术

干线网络组网技术的演进始终以持续提升传输容量与距离、增强网络可靠性与灵活性为目标, 以适配不断发展的业务需求。PDH、SDH、WDM、OTN/ROADM 等技术逐步成为主流组网方式。20 世纪 80 年代, PDH 技术开始应用于干线网络, 但随着网络规模扩大和业务需求增长, 其标准化程度低、信号复接复杂、管理能力弱及仅支持点对点传输等局限性逐渐凸显。SDH 技术针对上述问题进行了改进, 自 90 年代初起逐步成为干线网络的主要组网技术, 并将最高传输速率从 PDH 的 565 Mbit/s 提升至 SDH 的 40 Gbit/s (实际规模商用的速率为 10 Gbit/s)。

光放大与 WDM 技术的出现显著增强了干线网络的传输能力。自 20 世纪 90 年代中后期起, SDH 数字组网与 WDM 大容量传输相结合, 逐渐成为主流组网方式。至 2000 年左右, 互联网数据业务取代语音业务, 成为带宽需求的主要驱动力, 面向传统语音业务构建的 SDH 体系逐渐显现出承载瓶颈。为解决该问题, 融合了 SDH 与 WDM 技术优势的 OTN 技术完成标准化, 并于 2005 年后逐步在干线网络中规模部署, 至今仍是主流组网技术。OTN 技术传输速率也实现了从 10 Gbit/s 到 40 Gbit/s (2008 年)、100 Gbit/s (2011 年) 乃至 400 Gbit/s (2023 年) 的跨越。

此外, 全光组网始终是业界重点发展方向。随着光层器

件技术的进步及大带宽业务对灵活调度需求的提升, 基于 ROADM 技术的干线全光网络自 2010 年起逐步实现规模部署。目前, 中国已建成全球规模最大的全光网络。

为满足未来干线网络对大传输带宽的需求, 业界正围绕更高单通路速率、更宽传输波段以及空分复用等技术方向持续开展研究。在单通路速率方面, 800 Gbit/s 及以上速率的技术攻关与产品研发已广泛启动, 未来需结合多波段扩展、先进均衡算法及高波特率新型光电器件等多种手段以实现长距离高性能传输。然而, 为支持 200 Gbaud 及以下的信号处理能力, 仍需开发更高带宽的光电芯片, 包括引入如薄膜铌酸锂等新型调制器材料。目前, 全球的主流设备商已可提供 800 Gbit/s 设备样品, 部分厂商更已研制出支持 1.6 Tbit/s 速率的数字信号处理 (DSP) 芯片及设备, 并逐步开展试点商用。

在波段扩展方面, 除 400 Gbit/s 系统已引入 C+L 波段外, 进一步扩展至 E、S、U 等波段成为新的候选方案。全球实验室内多波段系统 (涵盖 O、E、S、C、L 和 U 6 个波段) 的单纤传输容量已达 402 Tbit/s<sup>[33]</sup>, 传输距离为 50 km, 但距离实际商用仍面临诸多挑战, 包括多波段放大器设计, 以及配套器件如多波段光源、相干光模块、合分波器和波长选择开关 (WSS) 等的关键技术攻关。

在空分复用方面, 已报道多项实验与试验结果, 在传输容量与系统性能方面均取得显著进展。部分基于多芯光纤的空分复用技术已在海缆系统中尝试商用, 但距离规模化部署仍存在较大距离。弱耦合/强耦合光纤制备、空分集成放大、多通道 DSP 处理及高性能连接器件等关键技术尚待进一步突破。部分重要实验结果已在本文“新型光纤”章节中予以介绍。

#### 3.2 城域网络多样化传送技术

城域网络在引入干线网络相关技术进行构建的同时, 也根据所承载的主流业务从时分复用 (TDM) 语音向分组数据转变, 持续增强分组化承载能力。截至目前, 伴随承载业务需求的变迁与组网技术的革新, 城域传送网的分组化演进大致经历了 3 个阶段。

2000—2006 年为第 1 阶段, 以基于 SDH 的多业务传送平台 (MSTP) 为代表技术。该技术在继续承载 TDM 语音业务的基础上, 引入通用成帧规程 (GFP) 等数据封装协议, 实现对分组类业务的透明传送与尽力而为的承载。

2006—2015 年为第 2 阶段, 为适应固定网络和移动网络全面 IP 化发展及其电信级传送需求, 先后出现了多协议标签交换-流量工程 (MPLS-TE)、电信级以太网、分组增强

型 OTN，以及基于 MPLS-TP 的分组传送网（PTN）等代表性技术方案，分组承载能力显著增强。

2015 年至今为第 3 阶段，为满足云计算/数据中心互联及 5G 网络切片等发展需求，应对网络扩展性、多业务融合承载与确定性保障等挑战，业界推动了灵活光传送（FlexO）、灵活以太网（FlexE）和 SPN 等创新技术的落地。网络建设逐步演进为综合业务传送网和数据中心互联（DCI）等不同形态，其中 DCI 主要依托 WDM/OTN 技术承载，相关设备形态也根据数据中心具体需求进行了针对性优化。

展望未来，在 AI 和 5G-A/6G 等应用驱动下，城域传送网的发展重点将从固定与移动承载的持续演进，转向智算中心互联等新兴场景。随着 AI 大模型训练与推理能力的不断发展，针对智算中心内高性能服务器图形处理器（GPU）间后端互联（Scale up）及服务器间前端互联（Scale out）等方面日益增长的传输需求，大带宽、低时延、高通量、可靠拥塞控制与无损传输等功能的实现，亟需城域传送网与城域数据网协同演进予以支撑。

支持智算中心分布式部署的 800 Gbit/s、1.6 Tbit/s 及更高速率传输技术，适用于智算中心内部互联的 800 Gbit/s、1.6 Tbit/s、3.2 Tbit/s 等多类型高速光模块技术，旨在提升调度效率与降低能耗的多端口光电路交换（OCS）技术，以及在组网与连接层面多维度发展的光电融合技术等，预计将成为未来几年的关键技术关注点。

3.3 接入网络全光接入技术

作为宽带接入的主流技术，无源光网络（PON）在速率与技术融合上持续演进，目前已进入千兆光网规模部署阶段，并正向万兆光网试点过渡。基于其关键技术特征的演进，PON 技术的发展大致可划分为 4 个主要阶段，各阶段相

应技术标准的演进情况如表 3 所示。

一是 PON 概念兴起与初步应用阶段（约 1995—2005 年）。20 世纪 90 年代，随着万维网（WWW）的兴起，数字用户线（DSL）技术、Cable modem 技术及 PON 技术等新型宽带接入技术成为标准化与产业竞争的焦点。全业务接入网联盟（FSAN）于 1995 年成立，推动 ITU-T 于 1998 年发布首个 PON 国际标准 G.983.1。该标准采用基于异步传输模式（ATM）的链路层封装技术。然而，由于当时从端局到用户侧主要基于电缆连接，基于铜缆的 xDSL 技术得到广泛应用，而基于 ATM 无源光网络（APON）因光缆基础设施缺乏与成本较高等原因，未获得大规模部署。

二是 1 Gbit/s 速率 PON 的规模应用启动阶段（约 2005—2015 年）。随着 IP 技术快速取代 ATM，ITU-T 与 IEEE 分别于 2003 年和 2004 年发布了第 2 代 PON 系统标准——GPON 与 EPON。GPON 采用通用封装方法（GEM）协议替代了 APON 中的 ATM 封装，而 EPON 直接采用以太网帧传输协议。与此同时，用户对带宽需求的持续增长使 DSL 技术面临明显瓶颈，仅能通过牺牲传输距离以提升速率。因此，铜光混合的 FTTx 接入网架构被多数运营商所采纳。中国电信于 2006 年逐步停止长距离铜缆建设，2007 年启动 EPON 设备集采；中国网通也于同年提出“光进铜退”计划，并在北京等城市开展 FTTN 试点。至此，PON 技术正式步入规模商用阶段。

三是 10 Gbit/s 速率 PON 的规模应用的启动阶段（约 2015—2025 年）。自 2007 年起，在中国运营商的推动下，GPON 与 EPON 进入大规模部署阶段。同期，ITU-T 与 IEEE 分别推进 10 Gbit/s 速率 PON 标准的制定：ITU-T 于 2010 年发布 G.987 系列标准，确立 XG-PON 的技术要求；IEEE 于 2009 发布 IEEE 802.3av 标准，定义 10G-EPON 规范；ITU-T 又于 2016 年发布 G.9807 标准，明确对称型 XGS-PON 的规

表 3 TDM-PON 技术标准体系演化

技术体系	ITU-T	速率	IEEE	速率
第 1 代	APON (ITU-T G.983x)	上行: 155 Mbit/s 下行: 622 Mbit/s	—	—
第 2 代	GPON (ITU-T G.984x)	上行: 1.25 Gbit/s 下行: 2.5 Gbit/s	EPON (IEEE 802.3ah)	上行: 1 Gbit/s 下行: 1 Gbit/s
第 3 代	XG-PON/XGS-PON (ITU-T G.987x/G.9807)	上行: 2.5 Gbit/s、10 Gbit/s 下行: 10 Gbit/s	10G-EPON (IEEE 802.3av)	上行: 10 Gbit/s 下行: 10 Gbit/s
第 4 代	50G-PON (ITU-T G.9804x)	上行: 25 Gbit/s、50 Gbit/s 下行: 50 Gbit/s	25G/50G-EPON (IEEE 802.3ca)	上行: 25 Gbit/s、50 Gbit/s 下行: 25 Gbit/s、50 Gbit/s
APON: 基于无源光网络 EPON: 以太网无源光网络		GPON: 千兆无源光网络 IEEE: 美国电气电子工程师学会		ITU-T: 国际电联电信标准化局 XG-PON: 10G 无源光网络
				XGS-PON: 10G 对称无源光网络

范。XG-PON与10G-EPON的带宽较前代提升4~10倍,并可实现平滑演进。2021年,工信部提出推进10 Gbit/s PON规模部署的建设目标,进一步加速了其商用进程。

四是50 Gbit/s速率PON的应用试点启动阶段(自2025年起)。2018年,ITU-T启动了单波长50G-PON的标准制定工作,命名为“G.HSP(Higher Speed PON)”,并于2021年9月发布第1版50G-PON标准G.989x系列。2022年9月,ITU-T批准了该标准的首个修订版本;2023年2月又发布增补标准,新增对称50G-PON光接口技术规格,并支持GPON、XG(S)-PON与50G-PON 3代系统共存。目前,50G-PON已基本完成技术标准化工作。其带宽能力为XG-PON的5倍,并在时延、抖动和可靠性等方面实现技术优化,具备提供确定性业务体验的能力,可支撑更丰富的应用场景,成为“万兆光网”主要技术选择,目前已启动试点部署。

另外,随着视频直播、在线教育、虚拟现实(VR)、超高清视频等千兆应用的广泛发展,用户对家庭网络、企业驻地网等驻地网络的体验需求持续提升。当前,驻地网络仍存在上网质量不佳、Wi-Fi设备能力受限、回传网络质量不稳定及运营商运维手段不足等问题。FTTR作为一种全新的驻地网内组网技术,通过与Wi-Fi协同优化组网,为上述问题提供了有效的解决方案,目前已实现初步规模商用。截至2024年底,中国FTTR用户规模已突破3 000万户。

面向未来更多应用需求,更高速率的100/200G-PON目前已进入技术与标准预研阶段,预计仍需5~10年时间实现成熟。同时,业界对WDM PON的发展也保持开放研究态度。ITU-T于2022年9月启动下一代更高速PON预研项目G.SuP.VHSP,重点关注每波长50 Gbit/s以上光接入物理层所面临的挑战与候选技术,涉及信号调制、多址接入(如共享子载波、TDM、WDM等)、光发射/接收机设计及波长方案等领域。当前光信号处理的调制解调技术主要包括IM-DD与多阶调制-相干检测两类方案。IM-DD在单波长50 Gbit/s以上速率接入应用中面临色散、功率预算、非线性等光层限制,以及波长选择与低成本高功率实现等挑战;而基于相干检测的100G/200G PON具有更优的信号传输能力和更灵活的组网配置潜力。具体技术路线选择仍有待业界持续研究。2024年OFC会议上,加拿大麦吉尔大学首次报道了O波段基于硅光集成的200G相干PON实验<sup>[34]</sup>;2025年OFC会议中,诺基亚贝尔实验室展示了基于硅光、支持突发波长测量与快速本振调谐的100G相干PON实验系统<sup>[35]</sup>。

### 3.4 管理控制

光网络管控系统的发展结合了组网技术演进与新型业务

应用需求的变化,经历了传统网络管理、分布式控制与转发分离、集中式软件定义网络(SDN)架构,以及引入人工智能实现智能化管控等不同阶段。

在传统网络管理架构中,光通信网络及设备主要依赖厂商管理系统及运营商跨厂商管理系统进行控制。该系统通过定义北向接口,实现厂商管理系统与跨厂商系统之间的互通,以解决跨域网络的统一管控问题。

21世纪初,自动交换光网络(ASON)概念被提出。ASON通过引入控制平面,实现了控制与转发的分离,将控制平面功能分布于各传送网设备中,并依托信令协议完成网络控制。2008年前后,ITU-T发布G.8080 ASON体系架构,IETF也基本完成基于通用多协议标签交换(GMPLS)扩展的RFC规范制定,这标志着ASON国际标准化工作趋于成熟。该技术随后在巴西电信、西班牙电信等海外运营商网络中实现规模部署。中国自2004年起开展省内干线试验网建设,2006年中国电信首次在省际干线网络中采用ASON技术,用于承载大客户专线和高等级数据业务。中国电信与中国联通还相继启动了基于ROADM网络的波长交换光网络(WSON)建设,采用分布式计算策略,由首节点负责业务路径计算与端到端连接的建立<sup>[36]</sup>。

在光接入网领域,传统网络管理协议操作复杂,海量固定终端设备为接入网运维带来巨大挑战。2004年,宽带论坛(BBF)发布TR-069技术报告,提出基于IP的终端远程管理协议。该协议采用互联网C/S架构构建管理体系,并定义灵活可扩展的数据模型,有效解决了海量终端的管理效率问题。目前在中国光接入网中,无论是FTTH家庭智能网关还是FTTR主网关,均采用传统网管协议与基于IP的管理协议并存的双栈管理架构。

2012年4月,ONF发布SDN白皮书,获得业界广泛认同,SDN在光传送网中的应用随之成为研究热点。ITU-T、IETF、ONF等多个标准化组织分别从架构、YANG模型和信息模型等方面展开研究。ITU-T扩展了原有ASON控制组件架构,结合SDN技术特点,形成G.7701(通用管控)、G.7702(SDN管控)等系列规范,构建了管控一体化(MCC)架构。IETF基于传送网的抽象与控制(ACTN)架构,形成涵盖体系架构、接口及层协议扩展等一系列标准。ONF定义了传送网管控新型模型TAPI 2.0,适用于L0~L2层的技术建模。软件定义光网络(SDON)在中国的标准化工作已基本完成,三大运营商也已部署基于SDN架构的传送网管控系统。

随着AI技术的引入,基于开放接口架构,融合大模型、机器学习与数字孪生等技术,实现网络管控与运维的智能化



已成为当前光网络管控的研究热点。中国已基本完成网络智能化应用场景及分级评估等相关标准的研究。2025年成为网络智能体应用元年,光通信网络智能体相关研究加速推进,传送网故障管理智能体、接入网家宽装维智能体等标准研究相继立项,相应的测试评估与验证工作也在逐步展开。

图3为光网络智能管控与人工智能结合示意。在设备智能层,通过引入端侧大模型提升设备智能化水平,实现多维度网络态势感知,并基于感知结果开展预测性、预防性的分布式网络控制。在管控智能层,依托统一的采集能力,实现对网络资源及哑资源的全面采集,结合光网络数字孪生与AI技术,形成网络仿真分析等能力;通过智能体实现感知、分析、决策与执行的闭环,增强网络自智能能力,并协同传统网管流程,以开放接口支持运维人员参与管控。在运维智能层,借助大模型赋能智能体应用,增强人机交互能力,实现端到端自动化编排,提升运维效率与水平。

#### 4 光纤通信技术未来发展展望

光纤通信网络作为新型信息基础设施的关键组成部分与承载底座,在5G/6G、人工智能、算力及数据中心等多种新型应用与业务的推动下,其基础技术与组网技术正持续创新与演进,支撑千兆光网向万兆光网的平滑演进,未来发展前景广阔。

在基础技术发展方面,调制、探测、放大和光纤等技术将依托新型光电/电光材料,持续向高速率、宽频谱、集成

化与低能耗方向演进。在光通信调制与探测方面,直调直检光模块速率预计至2030年将达到3.2 Tbit/s,超长距相干光收发模块速率预计同期可达800 Gbit/s甚至1.6 Tbit/s量级。硅光平台通过异质异构集成与光电融合技术,显著提升集成规模与器件性能,发展路径预计从2.5D/3D集成、中等规模异质异构集成,演进至2035年的大规模异质异构集成,并逐步实现中等规模光电单片集成,推动高密度低能耗光互连、光计算与光传感等领域的进步。同时,CPO将逐步投入应用,OIO的应用场景亦持续扩展。

在光信号放大方面,支持多波段拓展的超宽谱光放大器有望逐步突破,尤其是SOA在O波段的应用将进一步扩展,面向SDM等场景的新型光放大技术也将持续研究。在新型光纤方面,除海缆领域继续应用空分复用技术外,也有可能短距互联等场景发掘新需求,但整体仍处于实验室研究阶段。G.654E光纤将逐步成为干线新建线路的主流选择,空芯光纤在制备工艺与应用探索方面预计取得更多进展,其在数据中心互联等实用化场景中的规模应用时间节点目前仍不确定。

在组网技术发展方面,干线网络、城域网络和接入网络整体呈现向高速化、融合化与智能化演进的趋势。在干线网络中,400 Gbit/s OTN/ROADM持续规模化应用,下一代主流速率技术将更多聚焦于400 Gbit/s+,并逐步向更高速率演进。根据历史发展经验推断,1.6 Tbit/s有望成为未来干线网络的主要速率选择。同时,基于多波段扩展、空分复用和空芯光纤等技术的超大容量传输,预计在未来3~5年内仍以实验与试验探索为主。在城域网络方面,OTN/ROADM/SPN等多样化组网技术持续部署与演进,并积极面向6G承载、智算中心分布式互联等新需求,推动新型光通信技术的发展,进一步深化分组技术与光层技术的融合。在接入网络中,10G PON继续规模部署,50G PON应用试点规模逐步扩大,预计需至少3~5年实现规模商用。下一代更高速率B100G PON的预研工作持续推进,其信号传输速率与技术制式将逐步明确。FTTR结合Wi-Fi技术的应用部署范围持续扩展,为未来新业务构建高质量接入环境。以PON为代表的光接入网络正不断拓展至工业PON、工业光总线、车载光通信等更多应用场景。在管理控制方面,光网络将逐步实现设

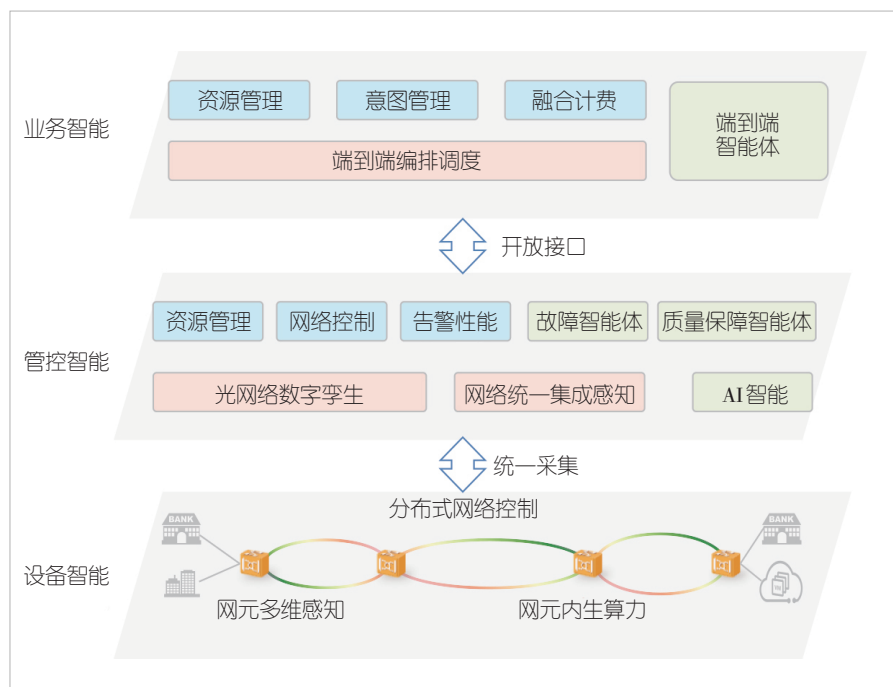


图3 光网络智能管控与人工智能结合示意图

备层智能、网络管控系统智能化及业务运营智能化,进一步提升整体智能化水平。此外,基于光纤介质的通信与感知融合技术日益受到业界关注,目前已在地震波探测、管道检测等领域开展探索性应用,未来应用范围将进一步扩大。

随着产业数字化与数字产业化进程的不断深入,高速光纤通信技术面临强劲的发展需求。然而,中国在该领域仍存在多方面的短板与挑战,包括原始理论创新不足、新型材料研发滞后、超高波特率光电器件关键技术尚未突破、高端平台制备工艺及核心仪器仪表依赖外部供应等问题。此外,中国产业同质化竞争加剧以及国际技术环境的持续压制,也进一步制约了行业的高质量发展。面对这一态势,亟需整合产学研用各方资源,充分发挥协同优势,围绕核心产品与工艺装备攻关、国际标准制定与产业应用推广等重点环节加强协作。应积极利用如ITU-T近期发起的ION-2030等国际项目契机,巩固并提升中国在全球光通信领域的竞争实力,推动光纤通信技术及产业实现健康、有序发展,为制造强国、网络强国和数字中国战略的顺利实施提供坚实支撑。

## 5 结束语

在国家政策引导、新兴业务需求拉动及产业各界的协同努力下,中国光纤通信技术实现了从跟随追赶到部分领跑的跨越,在技术研究、产品开发与规模部署等方面取得显著进展,目前已步入千兆网络广泛普及、万兆光网初步启动的新阶段。回顾其发展路径,光纤通信技术始终以基础技术突破与融合创新为主要演进模式。面向未来AI、6G与算力应用等高需求场景,亟需重点突破超高速光电处理、新型材料、光子集成与光电融合、新型光纤及放大等基础技术,并推动干线网、城域网(含智算中心互联)与接入网等多维组网技术的深度融合。与此同时,人工智能与光通信技术的双向赋能趋势日益明显。展望中国高速光纤通信技术的发展,机遇与挑战并存。呼吁业界围绕关键技术与产业瓶颈,协同攻坚、有序推进创新,持续强化新型信息基础设施的承载能力,从而支撑中国新质生产力加快形成,为数字经济高质量发展提供坚实根基。

## 参考文献

- [1] 赵梓森. 光纤通信的过去、现在和未来 [J]. 光学学报, 2011, 31(9): 1-3. DOI:10.3788/AOS201131.0900109
- [2] 工业和信息化部. 2025年一季度通信业经济运行情况 [R]. 2025
- [3] 王德, 李学干. 半导体激光器的最新进展及其应用现状 [J]. 光学精密工程, 2001, (3): 279-283. DOI: 10.3321/j.issn: 1004-924X. 2001.03.018
- [4] SUN H, WU K T, ROBERTS K. Real-time measurements of a 40 Gb/s coherent system [J]. Optics express, 2008, 16(2): 873-879. DOI:

- 10.1364/oe.16.000873
- [5] SHI Y, LI X, ZOU M J, et al. 103 GHz germanium-on-silicon photodiode enabled by an optimized U-shaped electrode [J]. Photonics research, 2024, 12(1): 1-6. DOI: 10.1364/prj.495958
- [6] HAN C H, JIN M, TAO Y S, et al. Ultra-compact silicon modulator with 110 GHz bandwidth [C]//Proceedings of Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2022: 1-3
- [7] HE Y T, LIU H, SUN C Z, et al. Dual-band 390 gbps high coupling efficiency thin film lithium niobate modulator with 3-dB bandwidth exceeding 110 GHz [C]//Proceedings of Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2025: 1-3
- [8] YOKOYAMA S, YIN Y X, YAZDANI S A, et al. 200 GBd electro-optic PLZT modulator for O-band transmission [C]//Proceedings of Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2025: 1-3
- [9] BLATTER T, KULMER L, XU C R, et al. Plasmonic ring resonator modulator demonstrating IM/DD >400G per lane [C]//Proceedings of ECOC 2024; 50th European Conference on Optical Communication. VDE, 2024: 418-421
- [10] HENI W, BAEUERLE B, LEUTHOLD J, et al. Plasmonic photonic integrated circuits: technology, performance, applications, and future prospects [C]//Proceedings of ECOC 2024; 50th European Conference on Optical Communication. VDE, 2024: 636-639
- [11] LEUTHOLD J, SMAJIC J, FEDORYSHIN Y, et al. Plasmonic-based devices with >500 GHz bandwidth [C]//Proceedings of ECOC 2024; 50th European Conference on Optical Communication. VDE, 2024: 1968-1971
- [12] AGRELL E, KARLSSON M, POLETTI F, et al. Roadmap on optical communications [J]. Journal of optics, 2024, 26(9): 093001. DOI: 10.1088/2040-8986/ad261f
- [13] 文建湘, 庞拂飞, 杨媛媛, 等. 超宽带光纤放大器研究进展与发展瓶颈(特邀) [J]. 光学学报, 2024, 26(9): 1-16. DOI: 10.3788/AOS250901
- [14] RENAUDIER J, MESEGUER A C, GHAZISAEIDI A, et al. First 100-nm continuous-band WDM transmission system with 115Tb/s transport over 100km using novel ultra-wideband semiconductor optical amplifiers [C]//Proceedings of European Conference on Optical Communication (ECOC). IEEE, 2017: 1-3. DOI: 10.1109/ECOC.2017.8346084
- [15] IQBAL M A, KRZCZANOWICZ L, PHILLIPS I, et al. 150nm SCL-band transmission through 70km SMF using ultra-wideband dual-stage discrete Raman amplifier [C]//Proceedings of Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2020: 1-3
- [16] KOBAYASHI T, SHIMIZU S, NAKAMURA M, et al. 50-Tb/s (1 tb/s × 50 ch) WDM transmission on two 6.25-THz bands using hybrid inline repeater of PPLN-based OPAs and incoherent-forward-pumped dra [C]// 2022 Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2022
- [17] CCSA. 空分复用光器件技术研究 [R]. 2025
- [18] 陈皓. 数据中心用多模光纤技术及发展趋势 [J]. 现代传输, 2019, (6): 11-14
- [19] ESSIAMBRE R J, KRAMER G, WINZER P J, et al. Capacity limits of optical fiber networks [J]. Journal of lightwave technology, 2010, 28(4): 662-701. DOI: 10.1109/JLT.2009.2039464
- [20] VAN DEN HOUT M, DI SCIULLO G, LUÍS R S, et al. Transmission of 273.6 Tb/s over 1001 km of 15-mode multi-mode fiber using C-band only 16-QAM signals [J]. Journal of lightwave technology, 2024, 42(3): 1136-1142
- [21] QIAO G, YANG Y, JI H L, et al. 5.27 peta-bit/s weakly-coupled SDM-WDM transmission over 55-km 10-mode 7-core fiber for SDM-priority scheme [C]//Proceedings of Optical Fiber

- Communication Conference (OFC). IEEE, 2024
- [22] ORSUTI D, PUTTNAM B J, LUÍS R S, et al. S/C/L-band transmission in few-mode MCF with optical frequency comb regeneration via single-mode core seed distribution [J]. Journal of lightwave technology, 2025, 43(4): 1786–1793
- [23] SATO S, KAWAGUCHI Y, SAKUMA H, et al. Record Low Loss Optical Fiber with 0.1397 dB/km [C]// 2024 Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2024
- [24] SAKR H, BRADLEY T D, JASION G T, et al. Hollow core NANFs with five nested tubes and record low loss at 850, 1060, 1300 and 1625nm [C]//Proceedings of Optical Fiber Communication Conference (OFC). IEEE, 2021
- [25] 中国移动. 中国移动开通全球首个800G空芯光纤传输技术试验网 [EB/OL]. [2025-06-28]. [https://www.10086.cn/aboutus/news/groupnews/index\\_detail\\_50009.html](https://www.10086.cn/aboutus/news/groupnews/index_detail_50009.html)
- [26] 中国通信网. 中国电信联合业界发布全球首个单波1.2Tbit/s、单向超100Tbit/s空芯光缆传输系统现网示范工程 [EB/OL]. [2025-07-08]. <https://www.c114.com.cn/news/117/a1265976.html>
- [27] 通信世界网. 中国联通携手北理工、上海诺基亚贝尔及长飞突破空芯光纤单波传输速率记录 [EB/OL]. [2025-07-08]. <http://www.cww.net.cn/article?id=589787>
- [28] MILLER S E. Integrated optics: an introduction [J]. The bell system technical journal, 1969, 48(7): 2059–2069
- [29] KAWAMURA Y, WAKITA K, ITAYA Y, et al. Monolithic integration of InGaAs/InP DFB lasers and InGaAs/InAlAs MQW optical modulators [J]. Electronics letters, 1986, 22(5): 242–243. DOI: 10.1049/el:19860166
- [30] PEREZ-LOPEZ D, TORRIJOS-MORAN L. Large-scale photonic processors and their applications [EB/OL]. [2025-06-28]. <https://www.nature.com/articles/s44310-025-00075-4>
- [31] Lightcounting. A resurgence in CPO development [EB/OL]. [2025-06-28]. <https://www.lightcounting.com/newsletter/en/december-2024-aocs-dacs-linear-drive-pluggable-and-co-packaged-optics-303>
- [32] ANAND. Forum list [EB/OL]. [2025-06-25]. <https://www.anandtech.com/show/21373/tsmc-adds-silicon-photonics-coupe-roadmap-128tbps-on-package>
- [33] PUTTNAM B, LUÍS R, PHILLIPS I, et al. 402 Tb/s GMI data-rate OESCLU-band transmission [EB/OL]. [2025-06-25]. <https://opg.optica.org/abstract.cfm?uri=ofc-2024-Th4A.3>
- [34] ZHANG J S, WEI Z X, MISAK S, et al. First demonstration of 200-G coherent PON at O-band with heterogeneously-integrated SiP tx and rx with lasers [C]//Proceedings of Optical Fiber Communication Conference (OFC) 2024. IEEE, 2024
- [35] BORKOWSKI R, STERN B, VIJAYAN K, et al. Burst-mode coherent PON upstream with rapid wavelength measurement and fast local oscillator tuning over 20 nm [C]// 2025 Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2025
- [36] 吕凯, 齐斌, 钟胜前, 等. ROADM全光交换网络关键技术发展与应用展望 [J]. 电信科学, 2022, 38(7): 37–42

## 作者简介



张海懿, 正高级工程师, 中国信息通信研究院技术与标准研究所所长; 主要从事高速光通信、量子信息和人工智能等领域的技术研究、产业咨询与标准制定等工作; 曾获国家科技进步奖二等奖3次, 中国通信标准化协会科学技术奖一等奖3次、二等奖2次, 中国通信学会科学技术奖二等奖2次, 中国标准创新贡献奖三等奖1次, 2013年享受政府特殊津贴; 已发表论文10余篇, 提交国际标准文稿10余篇, 出版专著2部。



# 下一代AI大模型计算范式洞察



## Insights into Computational Paradigm of Next-Generation AI Large Model

熊先奎/XIONG Xiankui, 王程晨/WANG Chengchen,  
蔡文豪/CAI Wenhao

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

DOI: 10.12142/ZTETJ.202505008

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20250926.1148.002>

网络出版日期: 2025-09-26

收稿日期: 2025-08-10

**摘要:** 现代大模型规模随扩展定律持续扩大, 近万亿的模型参数量带来了算法、硬件、工程领域的一系列困境。Transformer架构固有的计算效率低下问题愈发凸显, 引发了研究人员对通用人工智能 (AGI) 实现路径的深入思考。一方面, 针对现有自回归Transformer架构, 已形成注意力机制、低精度量化、参数共享等算法改进方向, 以及集群系统优化、硬件系统升级等工程改进方向; 另一方面, 下一代AI大模型计算范式正朝着不以next token prediction为核心的方向演进, 具体包括两类路径: 一是从更高抽象层次进行预测的扩散和联合嵌入预测架构, 二是从物理第一性原理和计算基材特性出发构建的动力学模型、热力学模型和能量模型。同时, 新型计算范式与新型计算基材相结合, 有望从根本上改变传统AI算法软件与硬件割裂的局面, 成为迈向AGI的高效路径。

**关键词:** 大语言模型; 计算范式; 人工智能

**Abstract:** The continuous expansion of modern large-scale models, guided by scaling laws, has led to a series of challenges in algorithms, hardware, and engineering due to model parameters approaching the trillion-scale mark. The inherent computational inefficiency of the Transformer architecture has become increasingly evident, prompting in-depth reflection among researchers regarding the path to achieving artificial general intelligence (AGI). On one hand, improvements to the existing autoregressive Transformer architecture are being pursued along two main avenues: algorithmic enhancements such as refined attention mechanisms, low-precision quantization, and parameter sharing, as well as engineering advancements including cluster system optimization and hardware upgrades. On the other hand, the next-generation computational paradigms for AI models are evolving away from the core framework of next token prediction. This shift includes two distinct pathways: first, architectures that operate at higher levels of abstraction, such as diffusion and joint embedding prediction models; and second, approaches grounded in first principles of physics and the characteristics of computational substrates, including dynamic, thermodynamic, and energy-based models. Concurrently, the integration of novel computational paradigms with new computational substrates holds the potential to fundamentally alter the traditional disconnect between AI software and hardware, constituting an efficient pathway toward AGI.

**Keywords:** large language model; computational paradigm; artificial intelligence

**引用格式:** 熊先奎, 王程晨, 蔡文豪. 下一代AI大模型计算范式洞察 [J]. 中兴通讯技术, 2025, 31(5): 50-56. DOI: 10.12142/ZTETJ.202505008

**Citation:** XIONG X K, WANG C C, CAI W H. Insights into computational paradigm of next-generation AI large model [J]. ZTE technology journal, 2025, 31(5): 50-56. DOI: 10.12142/ZTETJ.202505008

## 1 LLM现状及瓶颈

### 1.1 LLM架构相对固化的现状

2020年, OpenAI揭示了大模型规模扩展定律 (Scaling Laws)<sup>[1]</sup>: 大语言模型 (LLM) 的最终性能取决于计算量、参数量和训练数据量的堆叠扩展<sup>[2-8]</sup>。拥有175B参数量的GPT-3模型<sup>[2]</sup>在自然语言理解、知识问答等多项任务中, 获得了远超同期模型的性能。近年来, 以DeepSeek-V3、GPT-4o、Llama4、Qwen3、Grok4为代表的大模型无不在证

明这个定律。

构建一款先进的基础大模型, 需要堆叠数十万卡算力、收集数百太字节海量语料, 基于自回归 (AR) Transformer架构, 采用预训练 (Pre-training) 和后训练 (Post-training) 等手段, 完成其内部近万亿参数量的训练。整个训练过程的沉没成本极为高昂, 如X.AI的Grok4模型, 在2个150 MW功率的数据中心构建的20万卡分布式集群里, 耗时半年才完成预训练。因此, LLM的预训练探索和实践主要在工业界

完成，而学术界只能集中在理论层面的研究和较小规模（参数量<7B）的实践。然而，尽管当前架构仍有一系列算法、硬件、工程、成本等瓶颈问题，为实现通用人工智能（AGI）的愿景并验证 Scaling Law 的有效性，产业界不断增大投入。模型规模持续增加的趋势短期内难以改变。

本文尝试从企业视角，剖析大模型架构的关键因素、发展契机和潜在技术路径。

## 1.2 LLM 架构的关键瓶颈

Transformer 架构的计算效率低，访存需求大<sup>[9]</sup>。特别是基于 Decode-only 的自回归结构算术强度仅为 2，即每读取 1 字节数据只能完成 2 次计算。卷积神经网络（CNN）高达数百的算术强度，其高数据复用率可充分满足图形处理器（GPU）/特定领域架构（DSA）的矩阵乘加单元需求；而 Transformer 架构因数据搬移开销较大，导致模型算力利用率（MFU）较低。同时，当前硬件难以并行运算 Transformer 架构中的 Softmax、Layer-norm、Swish 等特殊非线性算子。总之，LLM 架构对先进工艺和高带宽存储器（HBM）的依赖大、工程成本高，这是阻碍其规模应用、性能进一步提升的关键瓶颈。

未来，随着基础模型参数量的持续增加、推理模型长思维链输出上下文长度的飙升，以及以生物制药为代表的 AI for Science 等新型高性能计算应用的普及，Transformer 架构瓶颈将愈发突出，这与摩尔定律放缓的趋势相矛盾。依赖先进工艺提升算力和能效的技术路径将遭遇“功耗墙”“内存墙”等问题。计算和存储分离的冯·诺依曼架构在大模型规模 and 算力不断增长的需求下将面临严峻挑战。

## 1.3 迈向 AGI 的 LLM 发展路线

当前 LLM 在实践过程中或多或少存在幻觉、可解释性差等问题，这些问题在 Scaling Law 不断提升模型能力的过程中被掩盖。但 Transformer 自回归架构的核心是“next token prediction”，导致部分 AI 科学家如 LECUN 等认为，从稀疏编码和等价映射原理看，现有 LLM 难以真正理解物理世界。因此，关于物理世界映射、世界模型构建的路线，在学术界仍有很大争议。

从工业界角度看，Scaling Law 路线仍然需要进一步探索，因为平台期过后可能存在指数上升的拐点。这种路线的核心是商业闭环下的工程优化能力，同时需探索非 AR 模式乃至非 Transformer 模式的全新计算范式和算法<sup>[10]</sup>。未来 AGI 的发展路线，大概是开发能“感知”、能“物理思考”、能“实践”的认知大模型与具身大模型，这类模型需直接对齐

可解释组件，并能通过实践反馈机制形成所谓的自主意识<sup>[11]</sup>。因此，高能效端侧硬件、高效率算法将成为探索具身大模型工程化的关键。

## 2 LLM 自回归模式的工程改进和优化

针对前文所述问题，学术界和工业界基于自回归 LLM 开展了一系列算法、系统、硬件的改进和优化工作。

### 2.1 算法改进和优化

#### 2.1.1 注意力机制优化

文档理解、代码分析、检索增强生成（RAG）等应用场景要求模型支持长上下文输入，而以 DeepSeek-R1 为代表的推理模型又要求模型支持长思维链输出。序列长度增加会导致自注意力机制计算复杂度呈  $O(N^2)$  上升。因此，分组查询注意力（GQA）、多头潜在注意力（MLA）等注意力机制的改进，以及以 Flash-Attention 为代表的算子优化，已被广泛采用，Linear-attention、RWKV、Mamba 等线性注意力机制展现出巨大应用潜力。此外，旋转位置编码（RoPE）插值方案被进一步优化，部分注意力机制如原生稀疏注意力（NSA）、混合块注意力（MoBA），以及针对多卡场景的长上下文推理框架（如 Ring-attention、Tree-attention）也被用来降低计算量。

#### 2.1.2 低精度量化

Decode-Only 架构中典型的运算过程是矩阵向量乘法（GEMV），该运算数据搬移频繁、计算效率低，既消耗算力，又占用带宽。

利用硬件原生 FP8、FP4、MXFP 等低精度数据类型进行模型量化，既能够有效减少内存带宽需求，又可以等效增加芯片算力利用率。现有研究证明，4 bit 量化拥有相对最优扩展率<sup>[12]</sup>，在推理场景中已得到实际应用。然而，量化引入的误差，难免导致模型能力下降，同时非线性层的量化/反量化操作也有额外开销。因此，量化技术只能缓解计算和带宽瓶颈。

#### 2.1.3 循环递归参数复用

循环式 Transformer 架构<sup>[13]</sup>，例如 Universal Transformer、混合专家 Universal Transformer（MoEUT）等，通过跨层共享参数实现深度递归。这类架构引入循环神经网络的递归表达能力后，通过参数共享使权重可支持多次计算，从而有效提升算术强度，在内存带宽受限时提升系统性能。然而，当前这种架构的实验规模较小，其扩展后的表达能力和稳定性尚不明确。

## 2.2 集群系统改进

传统CNN（如ResNet、Yolo）的网络参数量和计算量只在MB和GOPS（10亿次每秒）量级，在当前百TOPS级别算力（能效比2TOPS/W）的算力单元中，通常单卡/单机即可工作。而现代LLM由于巨大的参数量和计算量，会不可避免地引入多卡/多机的集群系统，通过张量并行（TP）、数据并行（DP）、流水线并行（PP）和专家并行（EP）等并行计算范式，加速训练和推理过程。

基于MoE的分布式计算范式可以降低超大参数规模模型的训练强度。其核心原理是每次前向计算时仅激活top-K个专家，从而降低算力需求。例如，Deepseek V3便通过这种方式将前馈神经网络（FFN）的计算量缩减为原来的 $1/32^{[14]}$ 。

P/D分离的部署可以利用Prefill/Decode在计算和带宽需求上的差异：Prefill阶段是计算密集型，追求首个token生成时间（TTFT）；Decode阶段是访存密集型，追求单个token生成时间（TPOT）。二者分离部署，不仅互不影响<sup>[15]</sup>，还能充分利用硬件利用率。

云端AI系统能够协同解决端侧算力资源受限情况下的大模型部署问题。端侧部署参数量较小的模型，可实现本地实时推理。对于复杂任务的拆解和深度思考任务，可通过云端部署参数量较大的模型来完成。分析结果将被反馈至端侧，从而通过端云AI协同搭建“快慢思考”系统<sup>[16]</sup>。

## 2.3 硬件工程优化

LLM集群借用了传统高性能计算（HPC）集群工程经验来优化当前计算范式，具有以下工程化技术创新：

1) 微架构DSA化：在通用图形处理器（GPGPU）中，引入了更多DSA领域采用的专用架构设计。如Nvidia GPU Tensor Core引入异步数据搬运模式以及混合精度训练，借鉴数据流计算范式的相关经验。

2) 互联优化：通过将集群划分为Scale Up和Scale Out域，引入匹配计算范式的互联技术。Scale Up作为高带宽域，使用总线类技术（如Nvlink），提供200 ns超低延迟、数千节点高并行度、原生内存语义的超节点连接，以摆脱Amdahl's law扩展率的约束。而Scale Out则借用远程直接内存访问（RDMA）类技术支持通用扩展，复用HPC集合通信原语（如NCCL），建立并行计算软件模型。

3) 光电混合集群：在当前国产化算力能力受限情况下，基于硅光工艺以及晶圆级扩展的“小电算、大光联”软硬件架构有望成为构建万卡、10万卡以上集群的关键技术。

4) 新型计算范式：在解决带宽问题的过程中，“存算一

体”“Data-Centric”等突破冯氏架构“内存墙”“功耗墙”限制的一些新型计算范式也得到了高度关注。

5) 算网存仿真平台：万卡以上超大规模集群部署的寻优问题，需要通过仿真平台对算、网、存系统进行算力部署和工作流的优化。构建高准确率、高时效性的仿真架构是亟待研究的问题。

当前，有两个前瞻性硬件工程技术至关重要：

1) 基于光IO技术重构先进计算体系结构，是优化LLM计算范式的关键技术。可助力Scale Up百纳秒级超低延迟的超节点连接、内存池化和拉远等架构级创新。

2) 基于3D动态随机存取存储器（DRAM）和无电容DRAM提供大容量、高带宽的内存，并结合LLM计算范式“读多写少”“顺序多于随机”等访存特点，采取异构介质（如高带宽闪存）、层次化缓存、压缩计算、存算一体等架构设计，构建超越高带宽内存（HBM）的新型内存体系。

## 3 下一代AI大模型计算范式演进和展望

通过Scaling Laws持续扩展超大参数模型实现AGI的路线，受到算力、带宽、能耗、语料多方面的限制。AGI的实现需要进行根本性变革，如将基于物理第一性原理的算法模型与计算基材硬件工程相结合。

### 3.1 下一代AI大模型发展趋势

产业界正在探索不以next token prediction为核心的下一代AI大模型范式。基于能量、动力学等第一性原理的模型由于能有效表述各种分布并在物理系统中自然演化，有望成为下一代AI大模型的核心架构。例如，由Hinton提出的玻尔兹曼机，受统计物理中伊辛模型和玻尔兹曼分布的启发，引入了随机、递归的神经网络，能够学习数据的潜在分布，解决复杂组合优化问题。后续的受限玻尔兹曼机和深度置信网络，促进了人工智能技术的快速发展，并促进了生成式模型在图像生成、自然语言处理和强化学习等领域中的广泛应用。

然而，这些基于能量、动力学原理的模型在现有冯·诺依曼计算机上运行时，其能耗和计算效率仍面临显著挑战。这是因为，基于布尔逻辑的确定性计算架构，在处理基于统计和概率的生成式模型时面临以下两个关键问题：其一，互补金属氧化物半导体（CMOS）器件的物理特性限制了其在随机过程模拟方面的硬件实现能力；其二，在面对自然语言处理中的语义模糊性、动态环境下的实时决策等非确定性需求时，现有计算范式效率显著下降。这一瓶颈催生了面向统计和概率等新型计算范式的需求：通过算法和硬件联合设



计,打破存储器与运算器分离的传统流程。这有望大幅提升能效比和计算性能,为突破当前AI算力瓶颈提供全新思路。

### 3.2 未来模型发展方向

目前,针对下一代AI模型设计主要有以下两种思路:

其一,可能仍是Transformer,但不再是next token prediction自回归。从更高抽象空间、更强表达能力、长期学习能力的目标出发,设计新一代模型结构,代表工作包括:1) Diffusion LLM 架构<sup>[17]</sup>,代表模型包括 LLaDA、Mercury 等,通过扩散方法将自回归模型串行化生成过程,改进为从粗粒度到细粒度的并行化生成过程。在相同计算资源和模型规模下,这种架构能够提升10倍以上的推理吞吐量,将计算能耗减少到原架构的1/10,同时提升模型的逆向推理能力和上下文关注长度等指标性能;2) 联合嵌入预测架构<sup>[18]</sup>,代表模型包括联合嵌入预测模型(JEPA)、大型概念模型(LCM)等,通过将语言、图像、视频等数据编码到高层潜空间中,学习世界模型级别的抽象表示,并在表示空间中通过基于能量的模型替代基于概率的模型进行预测,从而有效提升模型的表达效果与规划能力。

其二,基于物理第一性原理,从计算基材特性出发,根据物理过程的动力学特性、能量变化趋势设计模型架构和数据流,代表工作包括:1) 液态神经模型(LFM)<sup>[19]</sup>,代表模型包括液态结构状态空间模型(LSSM),其核心原理是液态时间常数(LTCN)模型:

$$\frac{d\mathbf{x}(t)}{dt} = -\frac{\mathbf{x}(t)}{\tau} + f(\mathbf{x}(t), I(t), t, \theta) \times (A - \mathbf{x}(t)) \quad (1).$$

LFM是一种由小型生物神经动力学模型启发的新型时间连续循环神经网络(RNN),可以通过反向传播进行训练,并在时间序列预测任务中表现出良好的边界和稳定的动态特性、卓越的表达能力和较高的内存效率<sup>[20]</sup>。2) 以Hopfield网络、受限玻尔兹曼机(RBM)、深度置信网络(DBN)等为代表的基于能量的模型(EBM),为概率密度估计和表示学习提供了一种统一的框架。这类模型的理论基础都可追溯到统计物理中的自旋玻璃模型。EBM通过定义能量函数来表示所希望学习的概率分布,因而也可作为生成模型来学习数据分布并生成与训练数据类似的新样本。与显式定义概率分布的模型相比,EBM具有更大的灵活性,能够建模更加复杂的依赖关系。近年来,基于能量的模型理论仍在不断发展,同时也面临不少挑战。其中,配分函数的计算和采样效率问题仍是制约模型应用的主要瓶颈。此外,能量函数的设计缺乏系统的指导原则,往往需要依赖经验和启发式方法。同时,模型的表达能力、泛化性能等仍缺乏更深入的研究。

### 3.3 下一代计算范式展望

在未来AI计算中,相较于算力,能耗将成为更为根本的限制。现有AI计算低效的根本原因是,神经网络的实现依赖于传统冯·诺依曼计算架构通过二进制操作“模拟”神经网络的计算。这种方法实质上是使用高精度的逻辑计算来处理仅需低精度的人工智能任务,大量能量被用于数据搬移和纠错,导致资源的低效利用。为了在进一步提高计算性能的同时降低计算能耗,研究者们探索了多种新型计算范式,其主要思想是采用非冯·诺依曼计算结构和存算一体。目前比较重要和热点的研究包括如下路线:

#### 3.3.1 物理原理启发的计算架构

物理神经网络(PNN)是利用物理第一性原理构建人工智能的技术路径。现有技术路线包括光计算、量子计算、电磁计算等。

光计算是一种利用光子作为信息载体进行计算和传输的计算模式,具有超高速、超宽带、低延迟、高并行等优势。光计算利用光干涉、衍射、强度/相位调制等物理特性直接在模拟域执行特定的计算任务,尤其在AI计算中展现出颠覆性潜力。例如,清华研究团队推出了太极系列光计算系统,利用空间对称和互易特性实现了训推一体的光神经网络(ONN)<sup>[21]</sup>。但光计算目前仍面临集成度、器件性能、系统复杂度、精度、软件生态等多重严峻挑战,成熟度仍然较低。

量子计算是一种遵循量子力学规律调控量子信息单元进行计算的新型计算模式。现有的量子算法和量子神经网络框架需在有限的量子比特和较大的计算错误率约束条件下运行。例如,使用量子加权张量混合网络(QWTHN)实现大模型微调<sup>[22]</sup>,将FFN训练转化为二次无约束二次规划问题(QUBO)并通过量子Ising机求解,利用量子位构建储层并实现储备池计算等。然而,量子计算目前由于技术路线未收敛、量子比特位数量有限、工作环境苛刻等问题,暂时难以实现广泛应用。

电磁计算直接利用电磁波(微波、毫米波、太赫兹波)的特性进行信息处理,而非依赖传统的电子开关状态。其核心优势包括超高速操作、高并行性、低传输损耗等。计算实现形式主要分为微波/毫米波模拟计算、可编程电磁处理<sup>[23]</sup>以及电磁存内计算。电磁计算通过物理定律直接映射数学运算,在特定领域(线性变换、实时处理)展现出应用潜力,当前仍处于实验室阶段。

#### 3.3.2 基于材料特性的模拟计算架构

研究者们正探索多种神经形态器件,这些器件利用材料

的本征物理现象模拟生物系统的复杂行为，通过特定的连接方式，构建单元间相互耦合的系统，能够利用系统自身演化特性替代传统计算过程。因此，利用材料的本征特性，推动算法、软件与硬件的联合设计，有望根本性地改变传统AI算法软件与硬件割裂的局面，从而实现软硬件的协同优化。现有技术路线包括概率计算、吸引子网络、热力学计算等。

概率计算系统依赖具有真随机特性的概率比特单元(p-bit)，它是位于量子计算和数字逻辑之间的中间计算范式，能够比传统计算机更好地利用自然和概率的潜在属性，在组合优化、因式分解、密钥生成、马尔可夫链蒙特卡洛(MCMC)采样等应用场景中均有较大优势。此外，概率计算系统还能够训练随机神经网络和深度生成模型，例如深度玻尔兹曼机<sup>[24]</sup>。

吸引子是动力系统中不同初始条件下趋向的一组数值，可以在动力学系统中实现记忆功能。2024年，LI等利用可变电阻式存储器(RRAM)器件的双向阻变特性实现回滞型神经元<sup>[25]</sup>，并据此构建了一种双极性忆阻器电路涌现的循环神经网络，相比于传统Hopfield网络具有硬件高效、记忆容量大等优势。

热力学计算基于热力学原理，利用自然界固有的计算能力，开发新的信息处理网络的设计原则，应用于未来计算系统。Normal Computing通过构建具有精确表达的状态空间、表现力丰富的非线性函数以及可扩展能力的硬件单元，从而高效地从复杂分布中进行采样，解决物理仿真和机器学习任

务中的计算瓶颈问题。

### 3.3.3 生物启发的计算架构

生物启发计算通过模拟自然系统的信息处理机制重构计算架构，突破传统冯·诺依曼瓶颈。目前主流的研究方向包括类脑计算和DNA计算等。

类脑计算泛指一类受脑启发的新型信息处理架构，这类架构依托大规模并行计算平台，有望突破存储与计算分离的冯·诺依曼架构瓶颈，为通用智能问题提供高效解决方案。

DNA计算是一种利用分子的生化特性进行信息存储与处理的新型计算范式，具有高存储密度、低功耗等优势。未来DNA计算将通过硅基和生物混合计算，赋能AI时代数据处理。

生物启发计算架构正从专用加速器向通用计算范式跃迁。短期看，类脑计算芯片在边缘智能领域将率先爆发；中长期则将形成“硅基+生物群体协同”的融合架构，最终实现生物级能效的智能计算系统。

## 4 中兴通讯面向下一代AI大模型计算范式的探索与实践

### 4.1 存内计算架构

中兴通讯利用8T SRAM数字存内计算技术实现了12.31 TOPS/W@INT8的高能效AI加速器，同时也在进行xpu-pim异构架构探索，如图1所示。该架构基于压缩和量化实现端

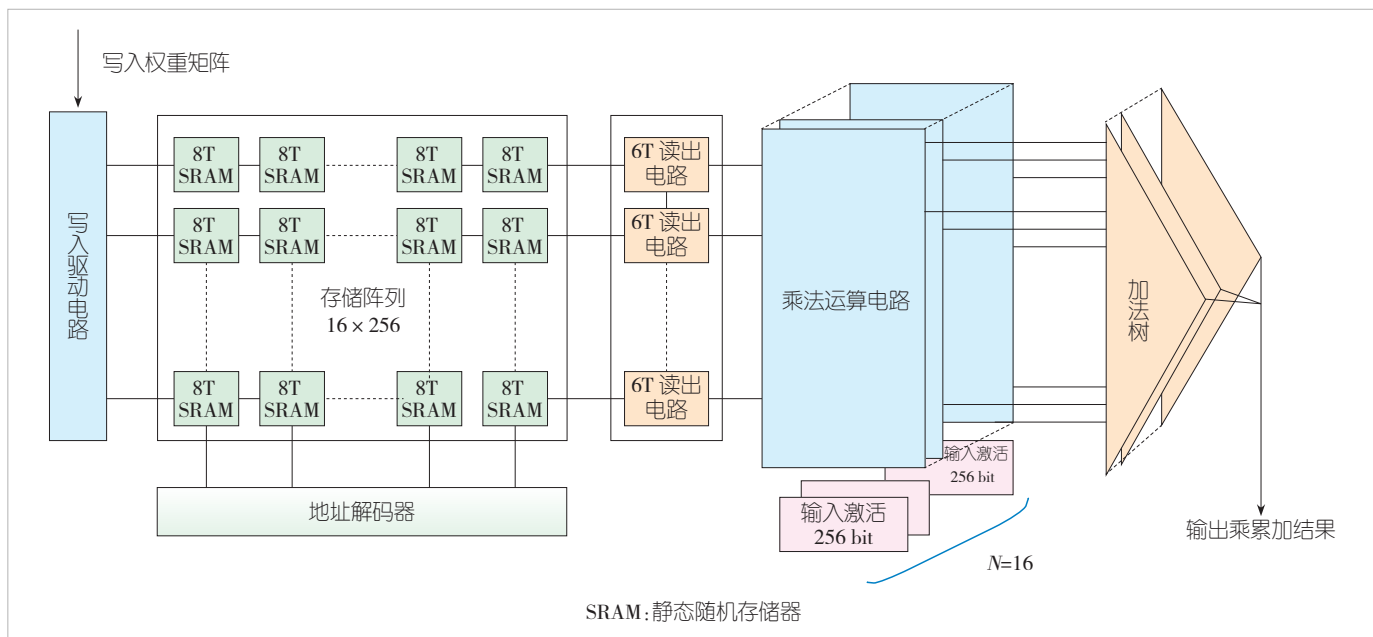


图1 存内计算架构示意图

侧大模型加速，在能效和吞吐量上具有数量级提升，近存架构将在端侧场景下发挥显著能效优势。

4.2 新型 AI 算法和硬件实现

中兴通讯在新型 AI 算法和硬件实现方面，探索了从物理第一性原理出发的新型技术路线。例如，基于循环式 Transformer 架构的高效参数共享特性，中兴通讯探索了其在替代多层 Transformer 架构上的能力。使用 GPT-2 small 的单个 Transformer 层作为模型“基块”，可以在减少超过 50% 参数量的同时保持模型的表达能力不下降。随着基块结构的改进，基块层数和循环次数可以进一步降低，如表 1 所示。

同时，稀疏玻尔兹曼机（DBM）架构由于其稀疏特性和基于最小化能量的推理目标，特别适合利用非易失性存储器执行端侧低功耗任务。DBM 的能量方程可以描述为：

$$E = -\sum_{i < j} J_{ij} m_i m_j + \sum h_i m_i \tag{2},$$

其中， $J_{ij}$  是耦合矩阵， $h_i$  是偏置向量， $m_i$  表示每个神经元的状态<sup>[18]</sup>。各个神经元串行更新并达到玻尔兹曼平衡的过程可以表示为：

$$m_i(t) = \text{sgn}(\tanh[-\beta I_i(t)] - \text{rand}_{U_i[-1,1]}) \tag{3},$$

$$I_i(t + \Delta t) = \sum J_{ij} m_j(t) + h_i \tag{4}.$$

在数千神经元的规模下，利用 GPU 完成单 batch 训练需要超过 10 h。而基于 FPGA 的 DBM 的快速计算单元，采用概率计算范式，通过例化数千个神经元及它们之间的稀疏连接，从而将单 batch 的训练时间缩短至 5 min，实现了超过 2 个数量级的加速效果。未来，使用 RRAM、MRAM 等非易失性存储器件，能够进一步降低计算开销，提升推理速度，以满足 DBM 在端侧推理场景的广泛应用需求。

此外，在光连接、新型内存等支撑性工程技术，以及计算存储分离的数据池化系统、内存语义互联系统、大规模仿真平台等架构技术方面，中兴通讯也展开了一系列前瞻性研究。

5 结束语

现代 LLM 基于 Scaling Laws 持续扩展，参数量接近万亿。巨大的模型规模引发了跨越算法设计、硬件架构和系统工程的多方面挑战。基于二次注意力复杂度的 Transformer 架构的内在计算效率瓶颈越来越显著，而这也推动了对通用人工智能可行途径的思考。

一方面，对当前流行的自回归变压器范式的增量改进，集中在算法改进（稀疏注意力机制、低精度量化、参数共

表1 基于 GPT-2、Qwen3 基块构建循环神经网络的训练结果

基块类型	GPT-2	GPT-2	GPT-2	GPT-2	GPT-2	Qwen3
基块层数	12	1	6	3	2	1
循环次数	1	12	6	12	18	6
最佳损失	3.35	3.65	3.30	3.35	3.45	3.41

享）和工程优化（集群系统、硬件工程）上。另一方面，越来越多的研究正在探索超越 next token prediction 的计算范式，代表性方向包括：1）基于扩散和联合嵌入预测架构的更高抽象层次模型；2）从物理学和底层计算基础得出的第一性原理模型，具体包括动力学模型、热力学模型和能量模型。

至关重要的是，这些新兴的算法范式与新型计算基材（如神经形态、光子和模拟内存加速器）的融合，为统一的硬件-算法协同设计框架提供了前景，有望成为通往 AGI 的高效路径。

参考文献

[1] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [EB/OL]. (2020-01-23) [2025-08-15]. <https://arxiv.org/abs/2001.08361>

[2] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. (2020-05-28) [2025-08-15]. <https://arxiv.org/abs/2005.14165>

[3] 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

[4] 何斯琪, 穆琛, 陈迟晓. 基于存算一体集成芯片的大语言模型专用硬件架构 [J]. 中兴通讯技术, 2024, 30(2): 37-42. DOI: 10.12142/ZTETJ.202402006

[5] 冯文佼, 李宗航, 虞红芳. 低资源集中的大语言模型分布式推理技术 [J]. 中兴通讯技术, 2024, 30(2): 43-49. DOI: 10.12142/ZTETJ.202402007

[6] REN T Q, LI R P, ZHAO M M, et al. Separate source channel coding is still what you need: an LLM-based rethinking [J]. ZTE communications, 2025, 23(1): 30-44. DOI: 10.12142/ZTECOM.202501005

[7] 裴丹, 张圣林, 孙永谦, 等. 大语言模型时代的智能运维 [J]. 中兴通讯技术, 2024, 30(2): 56-62. DOI: 10.12142/ZTETJ.202402009

[8] 韩炳涛, 刘涛. 大模型关键技术与应用 [J]. 中兴通讯技术, 2024, 30(2): 76-88. DOI: 10.12142/ZTETJ.202402012

[9] 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. DOI: 10.12142/ZTETJ.202402003

[10] ZHAO H, WU H Q, YANG D J, et al. BriLLM: brain-inspired large language model [EB/OL]. (2025-03-14) [2025-08-15]. <https://arxiv.org/abs/2503.11299>

[11] PILOTO L S, WEINSTEIN A, BATTAGLIA P, et al. Intuitive physics learning in a deep-learning model inspired by developmental psychology [J]. Nature human behaviour, 2022, 6(9): 1257-1267. DOI: 10.1038/s41562-022-01394-8

[12] OUYANG X, GE T, HARTVIGSEN T, et al. Low-bit quantization favors undertrained LLMs: scaling laws for quantized LLMs with 100T training tokens [EB/OL]. (2024-11-26) [2025-08-15]. <https://arxiv.org/abs/2411.17691>

[13] DEGHANI M, GOUWS S, VINYALS O, et al. Universal



- transformers [EB/OL]. (2025-03-17)[2025-08-15]. <https://arxiv.org/abs/1807.03819v3>
- [14] DAI D D, DENG C Q, ZHAO C G, et al. DeepSeekMoE: towards ultimate expert specialization in mixture-of-experts language models [EB/OL]. (2024-01-11)[2025-08-15]. <https://arxiv.org/abs/2401.06066>
- [15] ZHONG Y M, LIU S Y, CHEN J D, etc. DistServe: disaggregating prefill and decoding for goodput-optimized large language model serving [EB/OL]. (2024-06-06)[2025-08-15]. <https://arxiv.org/pdf/2401.09670>
- [16] TIAN X Y, GU J R, LI B L, etc. DriveVLM: the convergence of autonomous driving and large vision-language models [EB/OL]. (2024-06-25)[2025-08-15]. <https://arxiv.org/pdf/2402.12289>
- [17] NIE S, ZHU F Q, YOU Z B, etc. Large language diffusion models [EB/OL]. (2025-02-14)[2025-08-15]. <https://arxiv.org/abs/2502.09992>
- [18] ASSRAN M, DUVAL Q, MISRA I, et al. Self-supervised learning from images with a joint-embedding predictive architecture [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 15619-15629. DOI: 10.1109/CVPR52729.2023.01499
- [19] HASANI R, LECHNER M, AMINI A, et al. Liquid time-constant networks [J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(9): 7657-7666. DOI: 10.1609/aaai.v35i9.16936
- [20] HASANI R M, LECHNER M, WANG T H, et al. Liquid structural state-space models [EB/OL]. (2022-09-26)[2025-08-15]. <https://arxiv.org/abs/2209.12951>
- [21] XUE Z W, ZHOU T K, XU Z H, et al. Fully forward mode training for optical neural networks [J]. Nature, 2024, 632: 280-286. DOI: 10.1038/s41586-024-07687-4
- [22] KONG X F, LI L, DOU M H, etc. Quantum-enhanced LLM efficient fine tuning [EB/OL]. (2025-03-17)[2025-08-15]. <https://arxiv.org/abs/2503.12790v1>
- [23] LIU C, MA Q, LUO Z J, et al. A programmable diffractive deep neural network based on a digital-coding metasurface array [J]. Nature electronics, 2022, 5: 113-122. DOI: 10.1038/s41928-022-00719-9
- [24] NIAZI S, CHOWDHURY S, AADIT N A, et al. Training deep Boltzmann networks with sparse Ising machines [J]. Nature electronics, 2024, 7: 610-619. DOI: 10.1038/s41928-024-01182-4

- [25] LI Y X, WANG S Q, YANG K, et al. An emergent attractor network in a passive resistive switching circuit [J]. Nature communications, 2024, 15(1): 7683. DOI: 10.1038/s41467-024-52132-9

## 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。



**王程晨**，中兴通讯股份有限公司技术预研工程师；主要研究方向为大模型软硬件协同设计、先进计算范式等。



**蔡文豪**，中兴通讯股份有限公司技术预研工程师；主要研究方向包括深度学习算法、大语言模型、模拟计算、无线通信系统等。

# 星地一体化语义通信网络: 探索与展望



## Semantic Communication for Satellite-Terrestrial Integrated Networks: Exploration and Prospects

李东博/LI Dongbo<sup>1,2</sup>, 王新宇/WANG Xinyu<sup>1,2</sup>,  
尹志胜/YIN Zhisheng<sup>3,4</sup>, 承楠/CHENG Nan<sup>3,4</sup>, 刘劼/LIU Jie<sup>1,2</sup>

(1. 哈尔滨工业大学, 中国 哈尔滨 150001;  
2. 智慧农场技术与系统全国重点实验室, 中国 哈尔滨 150001;  
3. 西安电子科技大学, 中国 西安 710071;  
4. ISN 全国重点实验室, 中国 西安 710071)  
(1. Harbin Institute of Technology, Harbin 150001, China;  
2. State Key Laboratory of Smart Farm Technologies and Systems, Harbin 150001, China;  
3. Xidian University, Xi'an 710071, China;  
4. State Key Laboratory of ISN, Xi'an 710071, China)

DOI: 10.12142/ZTETJ.202505009

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20251016.1403.004>

网络出版日期: 2025-10-16

收稿日期: 2025-08-10

**摘要:** 围绕星地一体化网络中的语义通信, 系统综述了其架构设计、关键技术与建模框架, 并针对路由与资源管理等问题展开探讨。面向星地链路长时延、强多普勒频移及星上算力受限等约束条件, 提出了任务驱动的语义传输机制与鲁棒联合源信道编码策略, 同时探讨了多模态语义处理及知识库在线更新等关键问题。通过构建以语义意图与重要性感知为核心的跨层协同机制, 形成了面向星地场景的统一通信架构与可演进技术路线, 为该领域未来发展提供了理论支撑与系统指引。

**关键词:** 星地一体化网络; 语义通信; 智能化网络; 联合源信道编码; 大模型

**Abstract:** A systematic survey of semantic communication within integrated satellite-terrestrial networks is presented, focusing on architecture design, key technologies, and modeling frameworks, while also addressing issues related to routing and resource management. In response to constraints such as long round-trip delays, severe Doppler effects, and limited onboard computing capabilities in satellite-ground links, a task-driven semantic transmission mechanism and a robust joint source-channel coding strategy are proposed. Key challenges including multimodal semantic processing and real-time knowledge base updates are also examined. By establishing a cross-layer coordination mechanism centered on semantic intent and importance awareness, a unified communication architecture and an evolvable technical pathway tailored for satellite-ground scenarios are developed. This work offers theoretical foundations and systematic guidance for future advancements in the field.

**Keywords:** satellite-terrestrial integrated network; semantic communication; intelligent network; joint source-channel coding; large model

**引用格式:** 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

**Citation:** LI D B, WANG X Y, YIN Z S, et al. Semantic communication for satellite-terrestrial integrated networks: exploration and prospects [J]. ZTE technology journal, 2025, 31(5): 57-65. DOI: 10.12142/ZTETJ.202505009

**国**际电信联盟 (ITU) 在《IMT 面向 2030 及未来发展的框架和总体目标建议书》<sup>[1]</sup>中明确了 6G 的 6 个典型应用场景与 15 项关键性能指标, 标志着通信技术向多维度能

力融合演进, 如图 1 所示。其核心特征体现为通信、感知、计算、人工智能 (AI) 与安全等要素的一体化集成, 以及星地融合的泛在连接。当前, 通信系统性能逐渐逼近理论极限, 现有通信体制难以匹配人工智能的设计范式, 且缺乏对复杂应用场景的灵活适配能力, 已成为制约发展的主要瓶颈<sup>[2]</sup>。在通信领域, 无损高效传输是核心需求, 而多模态语义的提取与处理属于人工智能的关键任务。更高层次的语义

**基金项目:** 国家自然科学基金重点项目 (62350710797); 中国博士后科学基金特别资助项目 (2022TQ0091); 中国博士后科学基金面上资助项目 (2022M720958); 黑龙江省重点研发计划项目 (2022ZX01A23); 全国重点实验室课题 (JD2023GJ01)

信息则为通信与智能的深度融合提供了途径<sup>[3]</sup>。在此背景下,语义通信作为一种面向应用的新兴通信范式,通过与星地一体化网络深度融合,有望在保障高效可靠传输的同时,增强星地泛在连接能力,为突破理论极限、实现场景自适应通信等挑战提供新的解决思路。

当前,语义通信在星地一体化网络中的应用不断拓展<sup>[4]</sup>。借助人工智能技术,其在资源调度、抗干扰能力、行为建模与信道建模等方面展现出显著优势。现有研究主要聚焦于语义通信对网络能力的增强与关键问题的解决,例如优化分布式边缘学习、压缩联邦学习数据以及加强隐私保护。在路由与资源优化方面,语义通信通过引入新型寻址与路由机制,有效提升了卫星网络的传输效率与可扩展性,并推动了网络仿真技术的发展。结合传统比特通信方式,语义通信进一步增强了星地网络在复杂环境下的鲁棒性与资源利用效率,在多模态数据处理、安全性保障及资源分配等方面表现出明显潜力。总体而言,语义通信为星地一体化网络提供了高效、智能、安全的解决方案,是推动网络智能化与自适应演进的关键技术。

尽管语义通信在星地一体化网络中展现出广阔前景,其进一步发展仍面临诸多挑战。在技术层面,挑战主要集中于

多模态数据中语义信息的高效提取、异构语义的有效融合以及计算复杂度的有效控制。在网络架构方面,要集成应用语义通信,就需对现有网络进行深度改造,具体涉及通信协议更新、硬件设备升级,并需构建新的安全机制以保障数据传输与处理的安全可靠。此外,语义通信目前缺乏统一的技术标准,导致系统间互操作性不足。因此,亟需推动行业标准的制定以支撑其规模化部署与应用。

## 1 星地一体化语义通信网络架构

语义通信作为一种区别于传统通信的新兴范式,其核心在于解决“传输符号如何精准传达含义”这一根本问题。近年来,随着人工智能技术的快速发展,语义通信已逐渐成为通信领域的重要研究方向<sup>[5]</sup>。星地一体化网络,是一种融合了不同轨道卫星与地面蜂窝移动系统的多层次、多连接、多接入新型网络架构<sup>[6]</sup>。二者均被视为6G网络的关键组成部分,因此在星地一体化网络环境下开展语义通信研究具有显著必要性。

语义通信与星地一体化网络的结合,源于其能够有效应对空间通信环境中的独特挑战。相较于传统地面网络,星地一体化网络中的语义通信不仅需优化带宽利用与降低时延,

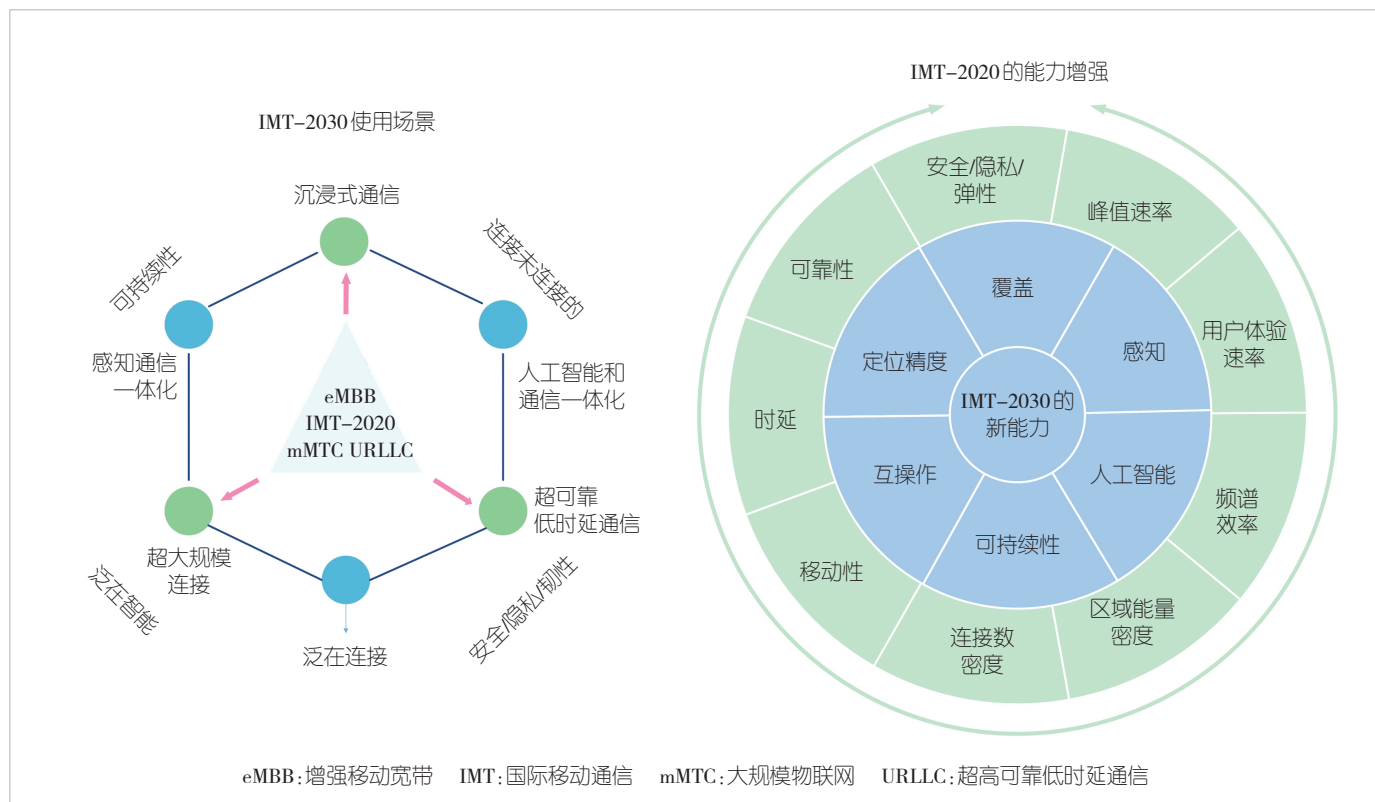


图1 6G的6个典型场景与15个性能指标



还需解决由长传输时延、资源严格受限及显著多普勒效应所引起的语义传递困难。在地面网络中,语义通信主要致力于提升带宽效率、降低传输时延,并利用上下文信息减少冗余数据传输。而在星地一体化网络中,语义优化更具重要意义,因其直接关系到有限能量与带宽资源的高效分配。同时,该系统还需确保信息在多样化传输条件下的语义保真度,从而提升关键数据传输的准确性与可靠性,以支撑广泛的应用场景需求。

## 1.1 语义通信的原理与特性

语义通信是一种基于信息内容理解的通信范式，其核心思想超越了传统的比特级传输模式。在语义通信中，信息的传输不仅是对原始数据流的简单复制，更侧重于对信息内涵的理解与语义层面的表达。其基本原理在于从原始信息中提取关键语义要素，通过组织、归纳与特征编码，实现语义的高效压缩与表示，从而在源头减少传输的数据量<sup>[7]</sup>。在接收端，通过相应的语义解码与重构过程，保障信息含义的准确传递。图2展示了一种典型的语义通信系统模型，其中语义编码与解码对应于B级语义通信，而语法编译码部分属于A级技术通信范畴。

主要模块的定义及功能如下:

1) 语义知识库。语义知识库作为语义通信所特有的关键组成部分,其功能与传统通信系统有明显区别。该模块主要负责从信源或信宿中准确提取语义上下文信息,并收集信道传播环境中的语义相关特征。这些信息作为先验知识,为

语义编码与解码过程提供关键指导,从而保障信息传递的准确性。语义知识库的内容形式多样,可包括知识图谱、语义标签或环境特征等。

2) 语义编码器。语义编码器在语义知识库的支持下,与传统信源编码器相比,更注重从信源消息中提取与语义相关的特征,并传输任务所需的关键信息。它不仅可以根据语义属性和信道状态指导编码过程以实现精确解码,还具备一定的抗干扰能力,能够有效抑制传输过程中的噪声与扰动,从而确保编码结果既保持语义完整性,又具备信道适应性。

3) 语义译码器。语义译码器根据信宿需求, 将接收到的语义信息重构为符合人类情感认知的具体消息, 或转换为智能物联网终端可执行的任务指令。

相较于传统通信, 语义通信具有以下明显特征: 首先是高效性, 由于传输的是信息的语义而非原始数据, 因此可以明显减少传输数据量, 提高通信效率<sup>[8]</sup>; 其次是抗干扰性, 语义通信对信号的形式不敏感, 因此在一定程度上能够抵抗信道噪声和干扰<sup>[9]</sup>; 然后是自适应性, 语义通信可以根据信道条件和传输需求自适应地调整编码策略, 以优化通信性能<sup>[10]</sup>; 最后是智能化, 语义通信涉及对信息内容的理解 and 处理, 因此需要借助人工智能技术来实现高效的语义提取和表示<sup>[8]</sup>。但是, 我们仍需要看到, 尽管语义通信相较于传统通信存在巨大优势, 但其通信系统的设计也更为复杂。换句话说, 语义通信就是以系统的复杂性为代价来提高通信效率。

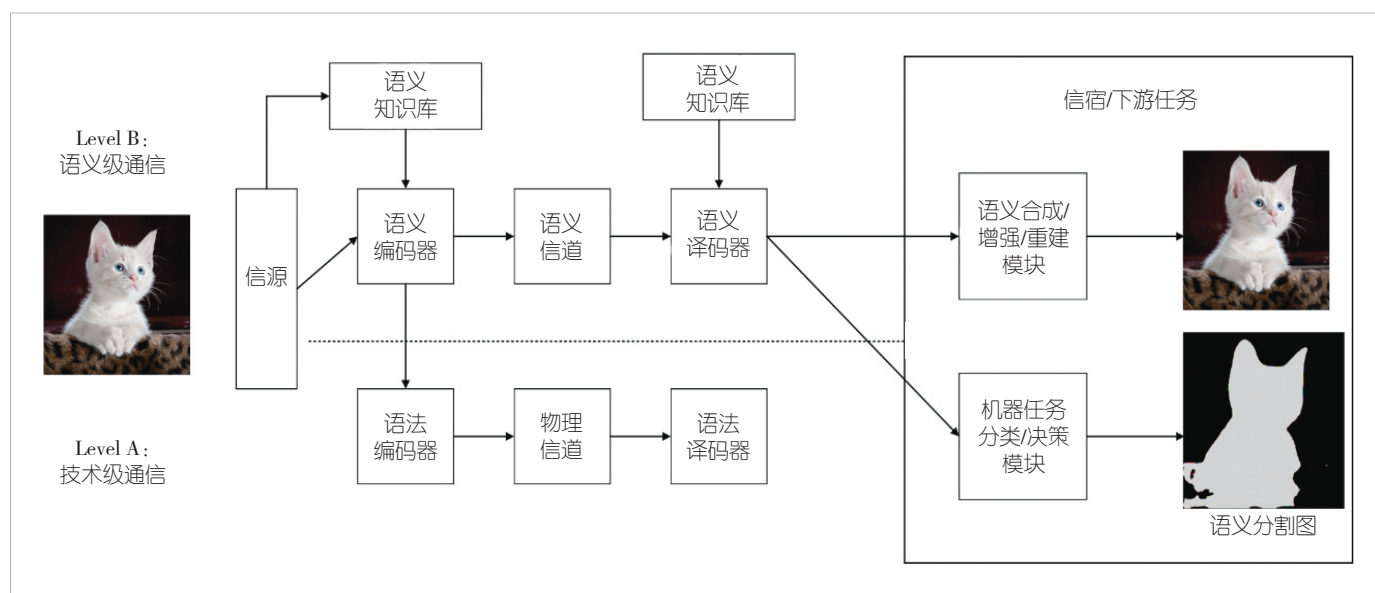


图2 语义通信系统的经典架构

### 1.2 星地一体化语义通信网络架构

星地一体化网络是一种将卫星通信网络与地面通信网络紧密融合的新型通信体系。它通过卫星和地面通信设施的无缝连接,实现全球范围内的宽带通信覆盖,旨在为用户提供无缝、高效、可靠的通信服务<sup>[11]</sup>。星地一体化网络充分利用了卫星通信的广覆盖优势和地面通信的高容量特性,有效弥补了单一通信网络的不足,为全球通信提供了全新的解决方案<sup>[12]</sup>。

随着星地一体化网络的日益发展,卫星网络与地面基础设施正加速融合,这将形成高度分区的网络拓扑,进而对路由协议的设计提出了新的要求,以实现高效的网络操作。此外,由于融合网络需同时支持多样化服务并跨越隔离的网络域,实现内容感知路由已成为一个亟须考量的目标<sup>[13]</sup>。这也为语义通信的应用提供了潜在空间。

未来,物联网设备的大规模接入与卫星端有限的频谱资源之间将构成显著矛盾。语义通信因能有效压缩传输信息量,成为提升信道与频谱资源利用效能的关键途径。在星地一体化网络中,卫星凭借其广域覆盖优势,可为地面特别是传统网络盲区提供连接服务;而语义通信则能进一步增强该网络保障可靠连接的能力,其应用场景如图3所示。

在星地一体化语义通信网络架构中,语义知识库、语义信道、语义噪声及联合源信道编码等关键要素,共同构成了高效智能通信系统在语义层实现信息传递的基础。该架构中的语义知识库,依托知识表示、推理与检索等功能,并借助长短期记忆网络与大型语言模型,增强了其对通信内容的理解与处理能力,从而能够支撑个性化与情境感知服务,同时提升对多模态及非结构化数据的处理效能。

语义噪声是语义通信中的关键挑战,源于外部干扰或系统缺陷引起的信息失真,具体表现为信道干扰与解码偏差。为抑制其影响,需采用增强推理、优化编码及冗余抑制等技术来提升可靠性。自适应语义解码则通过动态优化解码过程来适应环境变化,从而减少失真。

联合源信道编码对语义通信至关重要,通过融合源编码与信道编码来优化通信链路。其中,源编码借助深度学习消除信息冗余,信道编码则通过引入冗余与纠错机制保障信息完整。该技术通过全局优化降低传输开销,并利用语义知识库与机器学习动态调整编码参数,从而显著提升通信效率与系统适应性。该技术根据网络状态的自适应策略能有效增强网络灵活性,是复杂环境下提升系统性能的方案。

### 1.3 星地一体化语义通信网络中的信息传输需求

在星地一体化网络中部署语义通信虽会引入额外计算开销,但由于其能大幅节约通信资源,该代价通常被视为可接受的权衡<sup>[14]</sup>。然而,将语义通信融入星地一体化网络面临挑战:星地链路的长距离和高时延对语义信息的实时提取与编码要求更高;同时,系统必须在保障信息准确性的前提下,进一步降低通信开销,以适应卫星平台在资源与能量方面的

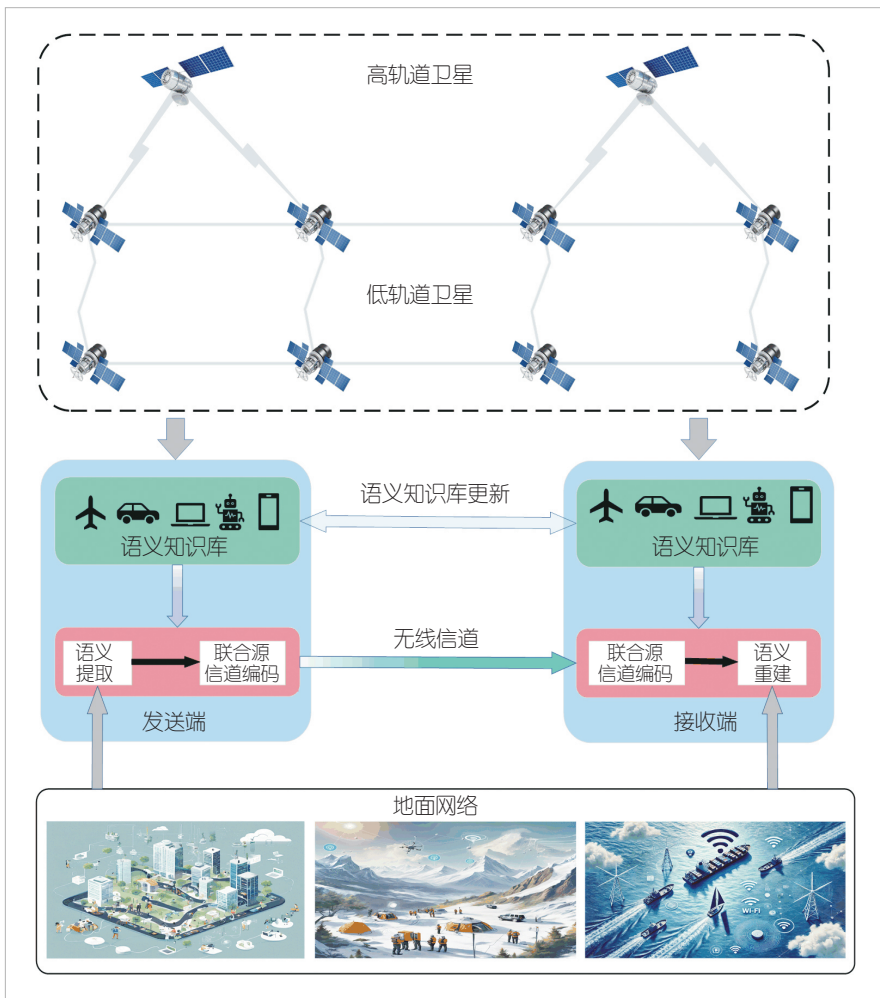


图3 星地一体化语义通信网络架构

严格限制。这些问题的解决,是充分发挥语义通信潜力的必要前提<sup>[15]</sup>。

大型卫星网络的寻址与路由问题是星地一体化网络面临的关键挑战之一。由于星座拓扑的动态时变特性,传统IPv6协议的性能易受卫星星座类型与路由算法的制约。然而,若利用已知轨道参数这一卫星固有特性,语义寻址能够通过有限的索引键实现卫星标识,再与源路由技术相结合,可进一步构建新型语义路由方案<sup>[16]</sup>。该方法对于硬件性能、功耗、链路带宽受限且系统复杂度高的卫星网络而言尤为重要。

尽管语义通信显著增强了星地一体化网络的性能,但现有通用语义框架难以直接适用于卫星部分,因此需研究新的语义技术。与此同时,未来6G网络明确以提升链路与信息容量、开拓新频谱并优化无线通信频谱效率为核心目标<sup>[17]</sup>。此外,星地链路存在的严重路径损耗与遮挡问题,也持续威胁着通信的稳定性。面对卫星端严苛的环境与复杂的链路条件,创新性的语义通信解决方案至关重要。引入无线能量收集、零能耗无线电等技术,可有效降低任务成本,并为人工智能提供无能耗连接的新方案<sup>[17]</sup>。

在星地一体化语义通信中,安全传输是另一个关键挑战。通过结合多点共生安全策略与数字孪生技术,可以显著提升空间-空中-地面集成网络的安全传输性能<sup>[18]</sup>。引入绿色干扰技术,在卫星-地面一体化通信中实现共生安全,能够在提升系统整体安全性的同时兼顾能源效率<sup>[19]</sup>。多域资源复用方案也为卫星辅助物联网的物理层安全提供了有效途径,具体方法包括利用相邻节点产生的自诱导同信道干扰,在主信道与窃听信道之间制造不均匀信号破坏,从而保障卫星与物联网节点间的安全传输<sup>[20]</sup>。此外,无人机(UAV)的引入为多波束卫星支持的车辆通信提供了额外的物理层安全保障<sup>[21]</sup>。其灵活部署能力可用于优化上行链路的安全策略,在确保信息机密性的同时,维持网络公平性<sup>[22]</sup>。上述创新技术表明,星地一体化网络在应对复杂安全威胁时,能够有效整合多样化的空间与空中资源,实现更高效、更安全的数据传输,为未来智能化空间通信网络的构建提供了理论基础与技术支撑。综上所述,语义通信不能直接适配到星地一体化网络当中,需要根据星地网络这一全新场景的要求进行能耗、算法、编解码效率、安全等方面的优化。

## 2 星地一体化语义通信网络关键技术

语义通信的发展离不开AI技术的支撑,而AI技术本身也早已被应用于星地一体化网络。例如,文献[23]在总结AI解决该网络所面临的问题时,提出了利用AI进行行为与信道建模以实现资源调度与抗干扰的初步构想,这被视为语义

通信在星地一体化网络中的早期雏形。

### 2.1 语义赋能的星地一体化网络

当前,星地一体化网络中语义通信的研究部分聚焦于利用语义能力为其他网络需求赋能。例如,文献[24]从边缘学习视角,探讨了分布式边缘学习与语义通信设计的相互作用;文献[14]研究了概率语义通信的能效,并在无人机中继场景下,提出了最小化总通信与计算能耗的迭代算法;文献[25]针对联邦学习中的通信瓶颈,利用语义通信将模型尺寸从5 MB压缩至5 kB,大幅降低了数据量;文献[26]提出了一种语义通信辅助的星载边缘云框架,通过剪枝与分割学习,在节省40.50%通信资源的同时降低51.43%的隐私风险;文献[27]则引入了“语义信息年龄”这一新指标,以评估应用层信息的时效价值,有效降低了卫星指令与传感器数据的平均年龄。

基于以上分析,语义通信在星地一体化网络中展现出多方面的优势,包括通信效率的提升、能耗的降低、资源管理效能的优化以及隐私安全保障的增强。通过与边缘学习、概率通信及联邦学习等技术结合,语义通信能够实现更高的能效、更低的传输时延,并有效支持通信资源的优化配置与智能计算卸载。此类技术融合显著减少了数据传输量,节约了系统资源,同时降低了隐私泄露风险。此外,语义新鲜度等新型度量指标的引入,进一步提升了通信的时效性与准确性。语义通信不仅推动了星地融合通信技术的发展,也为构建未来高效、智能、安全的集成网络提供了重要路径。

### 2.2 星地一体化语义通信网络中的路由与资源优化

语义通信在星地一体化网络中的另一重要应用方向聚焦于路由与资源优化。具体而言:文献[13]提出了语义赋能的算法,以实现多链路有效数据分发,并驱动网络管理、路由与资源分配,同时阐明了关键技术特征,构建了一个高效可扩展的多连通性模型;文献[16]设计了一种基于语义协议的全新寻址与路由方案,相较于传统IPv6协议,该方案能显著减轻卫星内的工作量,更适合太空环境下的大规模卫星网络;文献[28]进一步提出一种面向大规模星座的寻址路由方法,其利用卫星轨道参数进行语义寻址,支持通过星间链路实现数据路由,为巨型低轨星座提供IP连接能力;文献[29]则面向未来发展,描述了语义网络方法并提供了相应的星地通信系统仿真工具;文献[30]论证了本体工程在卫星网络仿真中提供语义支持的可行性,明确了其应用目标与内容。文献[31]将语义技术与软件定义网络(SDN)相结合,提出了



异构控制器本体映射算法,该算法能生成一致的全局网络视图,并具备高精度与低计算延迟的优势。

上述研究表明,语义通信在星地一体化网络中发挥着关键作用。它有效提升了多链路数据分发、网络管理与资源分配的效能,并通过创新的寻址与路由方案,显著增强了卫星网络的效率与可扩展性。面对大规模卫星星座的挑战,语义寻址策略保障了星间链路的高效连接与数据路由,为低轨卫星网络提供了可靠的IP互联基础。此外,语义通信有力推动了星地网络仿真工具的完善与验证能力的提升,并与软件定义网络深度融合,有效增强了网络灵活性与智能水平,为实现全局知识的一致性及高效管理提供了关键技术路径。语义通信在星地一体化网络中的核心价值将有力推动其向更高效、智能的方向演进。

### 2.3 星地一体化语义通信网络的模型

在星地一体化网络背景下,语义通信模型的研究已取得多方面进展。文献[32]提出了一种面向大规模物联网服务的语义授权星地一体化网络架构,通过语义通信与传统比特通信的协同,提升了系统在高资源效率与恶劣信道条件下的鲁棒性,并展现出显著增强的连接能力。文献[33]则构建了一种基于语义通信的星地一体化网络框架,系统阐释了在融合过程中所面临的多模态数据处理、安全机制与资源分配等关键挑战,并展望了语义通信对未来移动网络的潜在影响。文献[34]深入分析了星地一体化网络在大规模场景、高动态信道及受限设备能力方面的基本局限,基于香农信息论论证了语义通信替代传统比特通信的必然性,并详细阐述了相关性能指标体系与关键技术。其他研究也提出了各有侧重的语义通信模型:文献[35]聚焦于实时频谱感知,通过灵活部署无人机检测偏远地区频谱可用性;文献[4]提出基于生成式基础模型的语义卫星通信,借助分割与重建技术显著降低带宽需求,并在高噪声干扰下实现语义特征的准确恢复;文献[36]将语义通信与正交时频空调制相结合,有效抑制多普勒效应,提升传输速率与频谱效率;文献[37]则提出了语义驱动的卫星接入网切片方案,在考虑能源约束与资源稀缺的条件下,验证了其在节能与多任务支持方面的有效性。

综上所述,语义通信显著增强了星地一体化网络的鲁棒性与资源效率,其与比特通信的协同工作进一步提升了系统的连接能力与环境适应性。面对大规模物联网、高动态信道及受限设备能力带来的挑战,语义通信正逐步取代传统通信方式,成为支撑未来网络的关键技术。该技术通过传递信息意图,有效降低了带宽需求,提升了频谱利用

效率,并在高噪声干扰下保持了良好的传输准确性与可靠性。同时,语义通信也在多模态数据处理、安全保障与资源动态优化等方面推动着技术进步,为移动网络的演进开辟了新的可能性。

### 2.4 语义通信的关键技术创新

为实现语义通信与星地一体化网络的有效融合,仅依靠网络架构本身的创新仍显不足,语义通信本身也面临若干亟待突破的技术瓶颈。

在计算卸载与语义寻址方面,如何在实现高效语义信息压缩的同时保持较低的语义失真率,是当前语义通信面临的关键挑战。随着人工智能技术的持续演进,更多先进模型的开发与应用有望推动语义通信性能的进一步提升。大模型技术<sup>[38]</sup>的快速发展也将为星地一体化网络与语义通信的深度融合提供重要支撑。

在星地一体化语义通信网络中,语义信息提取与压缩技术是实现高效传输的核心。通过引入自然语言处理与计算机视觉方法,可从海量原始数据中提取关键语义信息并进行有效压缩,从而显著降低传输负载,在有限带宽条件下提升通信效率。星地链路的干扰处理技术是保障网络稳定性的关键,结合信道估计与智能干扰消除算法,可有效抑制由大气扰动与电磁干扰导致的信号衰减,提升传输质量与系统鲁棒性<sup>[39]</sup>。

多模态数据融合技术通过整合音频、视频与文本等异构数据,并依托深度学习算法实现跨模态语义提取,增强了系统在复杂数据环境下的理解与推理能力,为构建准确语义场景提供了支持。动态拓扑管理技术则基于自适应路由与网络优化机制,实时调整通信路径,在节点失效或网络拥塞时维持数据传输稳定性,并实现资源的高效调配。在此基础上,联合源信道编码技术结合机器学习对编码参数进行优化,进一步提升了复杂信道条件下的传输可靠性与带宽利用效率<sup>[40]</sup>。

## 3 挑战与展望

2024年7月,北京邮电大学张平院士团队率先建成国际首个通信与智能融合的6G外场试验网,实现了6G典型应用场景下通信性能的全面提升,标志着以“通信与智能融合”为特征的6G关键技术取得新突破。该成果为星地一体化语义通信网络的研究奠定了理论基础,两者的深度融合已在学术界与产业界形成共识<sup>[33]</sup>。尽管星地一体化语义通信网络展现出显著潜力与优势,但在实际部署中仍面临诸多挑战,主要体现在星地链路复杂干扰、网络动态拓扑、多模态

数据传输以及语义知识库构建等方面。

### 3.1 星地一体化语义通信网络面临的挑战

1) 星地链路面临复杂的干扰环境: 通信信号在穿越大气层过程中易受电磁干扰、雨衰、信道衰落及多径效应等因素影响, 引起信号衰减与随机噪声, 从而降低通信的可靠性与稳定性。语义通信以信息含义为核心, 可在一定程度上缓解干扰造成的影响, 但在极端条件下, 其抗干扰能力仍有待提升。此外, 卫星高速运动所导致的链路时延与动态变化, 也为实时数据传输带来挑战。因此, 发展能够适应动态干扰条件的鲁棒语义编码与解码技术, 成为提升星地链路通信质量的关键<sup>[41]</sup>。

2) 星地一体化网络具有动态变化的拓扑特性: 受卫星节点高速运动的影响, 网络拓扑结构持续发生变化, 传统静态路由协议难以有效适应这一特点。因此, 如何设计能够适应动态拓扑的路由算法, 从而实现高效的语义路由, 已成为当前研究的关键难题。动态拓扑不仅影响路由决策的准确性, 还可能引发通信中断或传输延迟。为应对这一问题, 网络需引入具备自适应能力的新型路由协议, 借助机器学习与预测技术动态调整路由策略。同时, 系统还应实现快速的重新配置与恢复机制, 以保障通信的连续性与可靠性<sup>[42]</sup>。

3) 多模态数据传输: 多模态数据传输在星地一体化网络中具有关键作用, 其目标在于整合视频、音频、图像与文本等异构数据, 实现跨模态的语义统一理解。传统系统通常对不同数据实施分离处理, 而语义通信则将其纳入统一框架中进行协同传输。在此过程中, 多模态数据的实时性与同步性至关重要。为此, 需在语义编码中引入更为智能的调度与优化机制, 从而有效提升数据传输的效率与准确性<sup>[43-44]</sup>。

4) 语义知识库的构建: 语义知识库作为星地语义通信网络中智能信息处理的核心组成部分, 其构建面临多方面的挑战。当前基于深度神经网络与知识图谱的方法, 常受限于知识表示的有限性以及高昂的数据标注成本。同时, 在动态网络环境下, 频繁的知识更新需消耗大量计算与能源资源。为提升更新效率, 已有研究探索引入在线学习与迁移学习等机制, 以减少模型重新训练的需求, 并增强系统对动态环境的响应能力。此外, 知识共享过程中涉及的安全与隐私问题, 尤其在传输敏感信息时, 需借助分布式架构及差分隐私、联邦学习等隐私保护技术以降低相关风险。上述关键技术的有效结合, 将为星地一体化网络智能化通信的发展提供更高效支撑。

### 3.2 星地一体化语义通信网络的发展趋势

展望未来, 星地一体化语义通信网络技术的快速发展将带来更高效、更智能的通信能力, 主要体现在以下几个方面:

1) 基于现有基础设施的语义通信: 北京邮电大学张平院士团队的6G外场试验为星地一体化语义通信网络提供了基于4G/5G基础设施的可行演进路径。通过引入语义层技术, 现有网络可在无需大规模硬件更新的情况下, 提升智能化处理能力, 优化资源分配与数据传输效率, 从而改善通信质量与用户体验, 为6G系统的平滑演进奠定基础。

2) 更高效的语义提取与重建算法: 未来研究将致力于开发更高效的语义提取方法, 以在复杂干扰环境中准确捕获关键信息。此类算法将融合深度学习的最新进展, 增强对语义特征的识别与重构能力, 并促进多模态数据的协调传输, 实现无缝跨模态融合。

3) 高性能联合源信道编码算法: 联合源信道编码在语义通信中的进一步应用, 有助于降低冗余、增强鲁棒性, 从而提升通信效率与传输可靠性。该方向的进展将推动网络在受限带宽与复杂信道条件下实现高质量的语义传输, 有效节约频谱与功率资源。

3) 基于大模型的星地一体化语义通信架构: 大模型凭借其强大的语义理解与推理能力, 可显著提升星地语义通信网络的智能水平, 弥补私有知识库的表示局限。同时, 其在资源分配、网络管理以及语义知识库动态更新方面的潜力, 也有助于提升系统传输准确性、灵活性与安全性。

4) 星地语义通信中的非正交多址接入(NOMA)技术: NOMA在星地一体化网络中具有重要潜力, 可在同一频谱资源下支持多用户并发接入, 显著提升频谱效率。未来研究将进一步探索NOMA与语义通信的融合机制, 通过语义信息优先级调度实现智能资源分配, 缓解网络拥塞, 提升高密度用户环境下的整体通信性能。

## 4 结束语

语义通信技术在星地一体化网络中的应用, 彰显了其作为未来通信关键使能技术的重要地位。随着人工智能、边缘计算、软件定义网络等技术的快速发展, 语义通信持续取得创新突破, 为星地一体化网络的演进提供了新的理论支撑与实现路径。本文系统综述了星地一体化语义通信网络的体系架构与关键技术, 剖析了当前面临的主要挑战, 并对未来发展趋势进行了展望。通过深入探讨语义通信在星地网络中的融合机制与关键技术路径, 本文论证了其在提升通信效率、降低系统能耗、优化资源分配以及增强隐私保护等方面的核

心价值, 以期为语义通信与星地一体化网络的深度融合与开展提供参考。

#### 参考文献

- [1] ITU-R. Framework and overall objectives of the future development of IMT for 2030 and beyond [R]. 2023
- [2] 牛凯, 张平. 语义通信的数学理论 [J]. 通信学报, 2024, 45(6): 7-59
- [3] ZHANG P, LIU Y M, SONG Y L, et al. Advances and challenges in semantic communications: a systematic review [EB/OL]. (2023-12-05) [2025-08-05]. <https://www.sciengine.com/NSO/doi/10.1360/nso/20230029>
- [4] JIANG P, WEN C-K, LI X, et al. Semantic satellite communications based on generative foundation model [J]. IEEE journal on selected areas in communications, 2025, 43(7): 2431-2445. DOI: 10.1109/JSAC.2025.3559113
- [5] 张亦弛, 张平, 魏急波, 等. 面向智能体的语义通信: 架构与范例 [J]. 中国科学: 信息科学, 2022, 52(5): 907-921
- [6] 刁兆坤, 杨丽, 王振章. 6G 空天地一体化网络架构及其构建 [J]. 通信世界, 2024(4): 36-39
- [7] LIU Y T, WANG X J, NING Z L, et al. A survey on semantic communications: technologies, solutions, applications and challenges [J]. Digital communications and networks, 2024, 10(3): 528-545. DOI: 10.1016/j.dcan.2023.05.010
- [8] 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
- [9] ZHANG Y M, WANG F Y, XU W J, et al. Semantic communications: a new paradigm for networked intelligence [C]// Proceedings of IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2022: 1-6. DOI: 10.1109/mlsp55214.2022.9943480
- [10] YANG W, DU H, LIEW Z, et al. Semantic communications for future internet: fundamentals, applications, and challenges [J]. IEEE communications surveys & tutorials, 2023, 1(25): 213-250. DOI: 10.1109/COMST.2022.3223224
- [11] 徐益平. 天地一体化网络发展趋势与挑战 [J]. 现代雷达, 2017, 39(7): 12-16
- [12] 杨昕, 孙智立, 刘华峰, 等. 新一代低轨卫星网络和地面无线自组网网络融合技术的探讨 [J]. 中兴通讯技术, 2016, 22(4): 58-63. DOI: 10.3969/j.issn.1009-6868.2016.04.012
- [13] COLA T D. Enabling effective multi-link data distribution in NTN-based 6G networks [C]// WSA & SCC 2023; 26th International ITG Workshop on Smart Antennas and 13th Conference on Systems, Communications, and Coding. IEEE, 2023: 1-5
- [14] ZHAO Z X, YANG Z H, CHEN M Z, et al. Energy-efficient probabilistic semantic communication over space-air-ground integrated networks [J]. IEEE transactions on wireless communications, 2025: 1. DOI: 10.1109/twc.2025.3569102
- [15] DE GAUDENZI R, OTTERSTEN B, PEREZ-NEIRA A, et al. Guest editorial space communications new frontiers: from near earth to deep space [J]. IEEE journal on selected areas in communications, 2024, 42(5): 1023-1028. DOI: 10.1109/jsac.2024.3378585
- [16] HAN L, RETANA A, WESTPHAL C, et al. Large scale LEO satellite networks for the future Internet: challenges and solutions to addressing and routing [EB/OL]. [2025-08-08]. [https://www.researchgate.net/publication/373665888\\_Large-Scale\\_LEO\\_Satellite\\_Networks\\_for\\_the\\_Future\\_Internet\\_Challen](https://www.researchgate.net/publication/373665888_Large-Scale_LEO_Satellite_Networks_for_the_Future_Internet_Challen)
- ges\_and\_Solutions\_to\_Addressing\_and\_Routing
- [17] STRINATI E C, BELOT D, FALEMPIN A, et al. Toward 6G: from new hardware design to wireless semantic and goal-oriented communication paradigms [C]// Proceedings of ESSCIRC 2021 - IEEE 47th European Solid State Circuits Conference (ESSCIRC). IEEE, 2021: 275-282. DOI: 10.1109/esscirc53450.2021.9567793
- [18] YIN Z S, CHENG N, LUAN T H, et al. DT-assisted multi-point symbiotic security in space-air-ground integrated networks [J]. IEEE transactions on information forensics and security, 2023, 18: 5721-5734. DOI: 10.1109/tifs.2023.3313326
- [19] YIN Z S, CHENG N, LUAN T H, et al. Green interference based symbiotic security in integrated satellite-terrestrial communications [J]. IEEE transactions on wireless communications, 2022, 21(11): 9962-9973. DOI: 10.1109/twc.2022.3181277
- [20] YIN Z S, CHENG N, HUI Y L, et al. Multi-domain resource multiplexing based secure transmission for satellite-assisted IoT: AO-SCA approach [J]. IEEE transactions on wireless communications, 2023, 22(11): 7319-7330. DOI: 10.1109/twc.2023.3250227
- [21] YIN Z S, JIA M, CHENG N, et al. UAV-assisted physical layer security in multi-beam satellite-enabled vehicle communications [J]. IEEE transactions on intelligent transportation systems, 2022, 23(3): 2739-2751. DOI: 10.1109/tits.2021.3090017
- [22] YIN Z S, CHENG N, SONG Y C, et al. UAV-assisted secure uplink communications in satellite-supported IoT: secrecy fairness approach [J]. IEEE Internet of Things journal, 2024, 11(4): 6904-6915. DOI: 10.1109/ijot.2023.3313197
- [23] FOURATI F, ALOUINI M S. Artificial intelligence for satellite communication: a review [J]. Intelligent and converged networks, 2021, 2(3): 213-243. DOI: 10.23919/icn.2021.0015
- [24] XU W, YANG Z H, NG D W K, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing [J]. IEEE journal of selected topics in signal processing, 2023, 17(1): 9-39. DOI: 10.1109/jstsp.2023.3239189
- [25] MU J S, CUI Y H, OUYANG W J, et al. Federated learning in 6G non-terrestrial network for IoT services: from the perspective of perceptive mobile network [J]. IEEE network, 2024, 38(4): 72-79. DOI: 10.1109/mnet.2024.3380647
- [26] ZHENG G H, NI Q, NAVAIE K, et al. Semantic communication in satellite-borne edge cloud network for computation offloading [J]. IEEE journal on selected areas in communications, 2024, 42(5): 1145-1158. DOI: 10.1109/jsac.2024.3365879
- [27] GAO R H, LI Y, XU Y L, et al. Semantic LTP: an age-optimal bundle delivery mechanism in space disruption-tolerant networks [J]. IEEE journal on selected areas in communications, 2024, 42(5): 1159-1174. DOI: 10.1109/jsac.2024.3365872
- [28] HAN L, RETANA A, WESTPHAL C, et al. New IP based semantic addressing and routing for LEO satellite networks [C]// Proceedings of IEEE 30th International Conference on Network Protocols (ICNP). IEEE, 2022: 1-6. DOI: 10.1109/icnp55882.2022.9940332
- [29] DĄBROWSKA-KUBIK K. Semantic network of ground station-satellite communication system [M]// Knowledge-Based and Intelligent Information and Engineering Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010: 369-378. DOI: 10.1007/978-3-642-15393-8\_42
- [30] LIN Q, XIONG Z, LI Z. Semantic modeling in satellite network simulation [C]// Proceedings of Asia Simulation Conference - 7th International Conference on System Simulation and Scientific Computing. IEEE, 2008: 877-881. DOI: 10.1109/asc-icsc.2008.4675486



- [31] IEEE. Message from the SmartCity 2021 general chairs [C]// Proceedings of IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). IEEE, 2021: 571 - 578. DOI: 10.1109/hpcc-dss-smartcity-dependsys53884.2021.00014
- [32] DENG D H, WANG C W, XU L X, et al. Semantic communication empowered NTN for IoT: benefits and challenges [J]. IEEE network, 2024, 38(4): 32 - 39. DOI: 10.1109/mnet.2024.3383604
- [33] PENG H X, ZHANG Z H, LIU Y L, et al. Semantic communication in non-terrestrial networks: A future-ready paradigm [J]. IEEE network, 2024, 38(4): 119 - 127. DOI: 10.1109/mnet.2024.3380817
- [34] MENG S Q, WU S H, ZHANG J M, et al. Semantics-empowered space-air-ground-sea integrated network: new paradigm, frameworks, and challenges [J]. IEEE communications surveys & tutorials, 2024, 27(1): 140 - 183. DOI: 10.1109/comst.2024.3416309
- [35] YI P, CAO Y, XU J R, et al. Semantic communication for remote spectrum sensing in non-terrestrial networks [C]//Proceedings of IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2022: 1 - 6. DOI: 10.1109/mlsp55214.2022.9943492
- [36] CHEN W B, LI S Y, JU C, et al. Semantic-enhanced downlink LEO satellite communication system with OTFS modulation [J]. IEEE communications letters, 2024, 28(6): 1377 - 1381. DOI: 10.1109/lcomm.2024.3386748
- [37] YOU C Q, HE X Q, ZHANG Y J, et al. SemSAN: semantic satellite access network slicing for NextG non-terrestrial networks [C]//Proceedings of ICC 2024 - IEEE International Conference on Communications. IEEE, 2024. DOI: 10.1109/icc51166.2024.10622974
- [38] RONG B, RUTAGEMWA H. Leveraging large language models for intelligent control of 6G integrated TN-NTN with IoT service [J]. IEEE network, 2024, 38(4): 136 - 142. DOI: 10.1109/mnet.2024.3384013
- [39] 吕守晔, 戴金晟, 张平. 信源信道联合的新范式: 语义通信 [J]. 中兴通讯技术, 2023, 27(2): 2-8. DOI: 10.12142/ZTETJ.202302002
- [40] 辛港涛, 樊平毅. 语义信息论的回顾与展望 [J]. 中兴通讯技术, 2023, 27(2): 9-12. DOI: 10.12142/ZTETJ.202302003
- [41] 徐晖, 陈山枝, 艾明. 面向6G的星地融合网络架构 [J]. 中兴通讯技术, 2023, 27(2): 9-15. DOI: 10.12142/ZTETJ.202305003
- [42] 缪德山, 邓凌越, 孙建成, 等. 6G星地融合无线网络及关键技术 [J]. 中兴通讯技术, 2024, 30(4): 42-49. DOI: 10.12142/ZTETJ.202404007
- [43] 瞿重希, 毛浩斌, 许憧, 等. 面向6G的星地融合网络频谱共享技术 [J]. 中兴通讯技术, 2024, 30(4): 50-56. DOI: 10.12142/ZTETJ.202404008
- [44] 张可, 林文超, 王野. 面向星地通信的低复杂度通用编译码技术 [J]. 中兴通讯技术, 2024, 30(5): 30-40. DOI: 10.12142/ZTETJ.202405006

## 作者简介



**李东博**, 哈尔滨工业大学计算学部副研究员、智慧农场技术与系统全国重点实验室通信与计算系统研究室负责人; 研究方向为空天智能系统、星地融合网络、大模型与多模态融合、语义通信等; 主持/参与国家级/省部级项目 20 余项; 发表学术论文 20 余篇。



**王新宇**, 哈尔滨工业大学在读博士研究生; 主要研究方向为卫星网络、语义通信等。



**尹志胜**, 西安电子科技大学、ISN 全国重点实验室副教授; 主要从事空天地一体化网络与传输方面的研究, 具体方向为空天安全通信、电磁对抗、智能传输等; 主持/参与国家级/省部级项目 20 余项; 发表学术论文 100 余篇。



**承楠**, 西安电子科技大学、ISN 全国重点实验室教授, 博士生导师, 国家高层次青年人才; 长期从事空天地海一体化智能网络与通信、电磁空间孪生及其应用、大模型与智能体通信相关研究; 主持多个科研项目; 发表学术论文 100 余篇。



**刘劼**, 哈尔滨工业大学讲席教授、人工智能研究院院长, 智慧农场技术与系统全国重点实验室主任, 物联网智能技术工信部重点实验室主任, 农业农村部东北规模化智慧农业重点实验室主任, 国家高层次人才, IEEE Fellow, ACM 杰出科学家, ACM SIGBED CHINA 发起人和主席; 主持科技创新人工智能重大专项、国家自然科学基金重点等项目; 研究方向为智能感知、智能物联网系统和人机物融合系统; 发表学术论文 140 余篇。

# 5G 基站节能面临的 关键问题和解决方案



## Key Problems and Solutions of Energy-Saving for 5G Base Stations

王小锋/WANG Xiaofeng, 韩茜/HAN Qian

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

DOI: 10.12142/ZTETJ.202505010

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20240726.1651.004>

网络出版日期: 2024-07-29

收稿日期: 2024-05-25

**摘要:** 在全球“双碳”战略推动下, 5G基站高能耗问题亟待解决。剖析了5G基站商用节能存在的四大关键问题: 凝露引发的硬件风险、多制式共用射频限制节能效果、性能与感知下降, 以及复杂场景节能策略难定义。针对性地提出软硬件协同、频谱规划优化和感知-能耗动态平衡等方案, 基于人工智能和大模型构建“大模型+大数据”驱动的逆向节能决策体系, 为5G网络实现高效、智能、可靠节能提供系统性解决路径。

**关键词:** 5G; 节能; 神经网络; 大模型

**Abstract:** Driven by the global "dual-carbon" strategy, the high energy consumption of 5G base stations has become an urgent issue to address. This paper analyzes four key challenges in the commercial deployment of 5G energy-saving technologies: hardware risks induced by condensation, limited energy-saving effectiveness due to multi-mode shared radio frequency units, degradation in network performance and user perception, and difficulties in defining energy-saving strategies for complex scenarios. To address these challenges, this paper proposes targeted solutions, including hardware-software co-design, spectrum planning optimization, and dynamic balancing between user perception and energy efficiency. Furthermore, leveraging artificial intelligence and large models, this paper establishes a "large model + big data"-driven inverse energy-saving decision framework, providing a systematic approach to achieve efficient, intelligent, and reliable energy savings for 5G networks.

**Keywords:** 5G; energy saving; neural network; large model

**引用格式:** 王小锋, 韩茜. 5G基站节能面临的关键问题和解决方案 [J]. 中兴通讯技术, 2025, 31(5): 66-70. DOI: 10.12142/ZTETJ.202505010

**Citation:** WANG X F, HAN Q. Key problems and solutions of energy-saving for 5G base stations [J]. ZTE technology journal, 2025, 31(5): 66-70. DOI: 10.12142/ZTETJ.202505010

气候与环境变化已成为21世纪人类社会面临的最严峻挑战之一。为应对气候变化, 中国明确提出2030年前实现碳达峰、2060年前实现碳中和的“双碳”战略目标, 推动经济社会全面绿色转型。在此背景下, 作为能源消耗和碳排放的重要领域, 信息通信技术 (ICT) 产业的绿色低碳发展备受关注。5G网络作为新型基础设施的核心, 其建设规模持续扩大。截至2025年4月, 全国5G基站总数已突破374万个, 占全球总量的六成以上。然而, 5G基站单站功耗较4G显著提升, 整体能耗呈指数级增长, 导致运营商面临巨额电费支出和碳排放考核的双重压力。基站能耗问题已成为制约5G可持续发展的关键瓶颈。为响应国家“双碳”战略, 中国移动提出“C<sup>2</sup>三能一碳达峰碳中和行动计划”, 构建“三能六绿”绿色发展模式; 中国电信发布“天翼零碳计划”, 推进网络节能降耗; 中国联通则实施“碳达峰、碳中和”专

项行动, 强调绿色网络建设。三大运营商均将5G网络节能作为重点任务, 积极推动节能技术的研发与商用部署。然而, 在实际商用环境中, 节能技术的落地仍面临诸多挑战, 如节能效果受限、用户体验下降、硬件安全风险等问题。

### 1 节能功能

5G基站的基本节能功能包括符号关断、增强型符号关断、射频通道关断、载波关断、深度休眠、自动启停 (也称为极致节能), 以及调压。表1简要介绍了5G基站的基本节能功能的原理和应用建议。

图1给出了4发射通道射频拉远单元和64发射通道有源天线单元两款典型设备, 在空载条件下各种节能功能的节能收益。相比于中、高负荷, 空载或者低负荷时的节能收益更明显。当开启自动启停时, 设备的功耗可以降到10 W以内,

表1 5G基站基本节能功能原理、应用建议

节能功能	原理简述	应用建议
符号关断	关闭空闲符号的功放	无特殊限制,可用于大多数场景
增强型符号关断	通过调度汇聚,增加空闲的符号个数	不适合时延敏感类业务的传输
射频通道关断	低负荷时关闭部分发射通道	对于网络覆盖和多天线性能有影响,小区内边缘用户时慎用
载波关断	关闭网络容量层的载波及使用的硬件器件,达到节能目的	适用于多载波共覆盖的组网场景,恢复时间在30~60 s之间
深度休眠	在网络轻载或者空载情况下,休眠射频模块,减少设备能耗	适用于多载波共覆盖的组网场景,恢复时间为分钟级,一般不超过3~4 min
自动启停	关闭射频模块的绝大部分硬件设备,减少设备能耗	除了考虑网络的需求之外,还需要关注周围环境的温度/湿度条件是否适合开启本功能
调压	通过软件对功放的漏压进行控制,达到节能目的	开启后会对业务传输产生一定的影响,高负荷或者重要数据传输时慎用

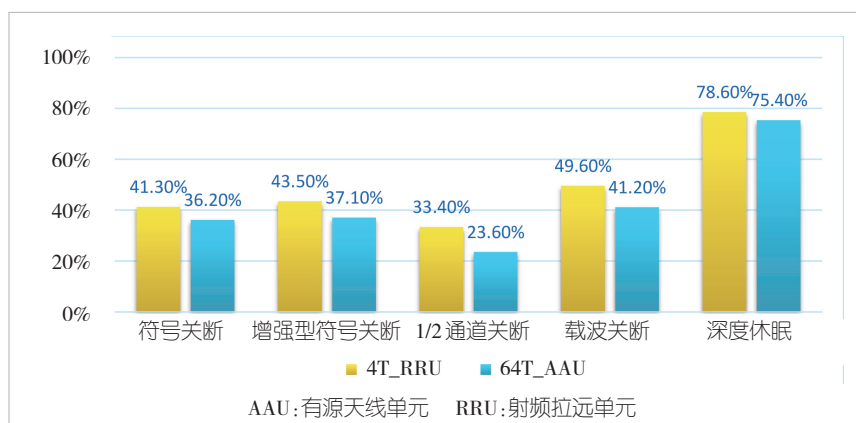


图1 典型AAU/RRU射频设备空载时部分节能功能的节能收益对比

有一些设备甚至可以降到5 W以内。调压可以单独开启,也可以和其他节能功能叠加开启。在商用网络中开启调压功能,可以在原有基础上增加5%~13%的节能收益。

## 2 关键问题

在5G基站的商用过程中,硬件器件的可靠性、无线配置差异、组网配置差异,都会对设备的节能收益产生影响<sup>[1-4]</sup>。

### 1) 节能可能导致硬件设备损坏

深度休眠、自动启停等节能功能,在一定条件下会导致设备表面产生“凝露”<sup>[5]</sup>,进而导致设备损坏。凝露是指当一定温度和湿度的空气遇到温度较低的固态表面时,空气中的水蒸气析出并以液体形式在器件表面形成露滴。凝露和设备内的灰尘混合后,会在器件表面形成导电通道,容易产生短路,损坏电子元器件。

### 2) 节能受设备特征影响

随着射频器件硬件集成度和无线制式复用度的提升,设备体积、重量和成本都在不断降低。这对节能产生了重要影响。这种影响主要体现在:不同无线制式载波复用相同的射

频通道和功放器件。当出现部分无线制式的载波可关闭而部分无线制式的载波工作时,被复用的功放硬件将无法关闭获取节能收益。例如,与全球移动通信系统(GSM)/通用移动通信系统(UMTS)共享硬件的4G/5G设备节能收益普遍偏低。

### 3) 节能导致网络性能和用户感知下降

通道关断通过关闭部分发射通道实现节能,但是关闭部分发射通道会削弱无线系统的赋形能力;载波关断通过关闭节能频层载波实现节能,而关闭载波带来的小区退服会在网络中形成节能频层覆盖空洞,

使得空洞区域的其他频层小区的负荷上升。这些后果会影响网络的基础指标和用户感知指标,严重的会导致网络性能和用户感知下降。

### 4) 节能策略复杂而难以制定

在复杂的商用网络环境中,如何制定网络的节能策略,才能在保障网络性能指标和用户体验的前提下,尽可能降低能耗?该问题的求解属于受大量因素影响的非线性寻优,难以确定最优解。而且随着节能区域的扩大,寻优、求解涉及的参数将成倍增长,求解的复杂程度也随之增加<sup>[6]</sup>。

## 3 解决方案

硬件可靠性、射频设备特征、节能与网络指标及用户体验的平衡关系,以及最优节能策略的求解,是影响商用环境下节能功能使用的4个主要因素。本节针对这些关键问题,逐一提出对应的解决方案。

### 3.1 软硬件协同提升器件可靠性

针对设备凝露的问题,核心解决思路是阻断设备的过快降温。相关方案可以结合外部的环境温度和湿度特点,采用



“递进式节能”的方法，逐步降低设备的温度。

在低温、潮湿的环境下，实施节能时若直接关闭射频设备，容易因设备温度快速下降产生凝露。对此可以通过分阶段逐步关闭硬件设备的方法，来减缓射频设备温度下降的速度。比如，有些较新的射频设备安装了传感器，在监测环境的温度和湿度以及设备自身的温度时，可自动控制硬件的关闭节奏，避免凝露的产生。

### 3.2 基于射频器件特征最大化节能收益

针对设备特征对节能的影响，可以从网络节能、硬件设计两个方面来构建解决方案。

#### 1) 网络节能

当前主流运营商掌握多个商用频段的频谱资源。以中国移动为例，4G的商用频段包含1 800 MHz（Band3）、900 MHz（Band8）等，5G商用频段包含2.6 GHz（n41）、4.9 GHz（n79），并与中国广电共享700 MHz（n28）频段。中国电信和中国联通也在850 MHz、900 MHz等频段部署了4G/5G网络。

（1）隔离基础覆盖层和补热层频谱资源：在网络频谱的规划和使用上，把承担基础覆盖而不能关闭的网络层部署在低频段、小带宽的频谱资源上面；把承担补热的4G/5G容量层部署在高频段、大带宽的频谱资源上面。比如，把GSM、UMTS和承担基础覆盖的长期演进（LTE）网络，部署在低频段、小带宽的频谱资源上；把5G、补热的4G网络部署在高频段、大带宽的频谱资源上。这样便于低负荷时段独立关闭5G、4G的补热频层的设备硬件，从网络频层规划上避免补热层和基础覆盖层因为硬件的共享带来关闭时的冲突，保障节能效果。

（2）迁移低负荷时段小区用户：针对多频层覆盖的区域，低负荷时段可以关闭补热频层的硬件设备。在低负荷时段到来时，将驻留在补热频层小区的用户迁移到基础覆盖频层，为补热频层的设备关闭创造条件。比如，在条件允许的情况下，将驻留在5G容量层的小区用户迁移到4G基础覆盖层，甚至迁移到2G、3G的频层上，可为5G频层设备的关闭创造条件。

（3）避免容量频层和基础覆盖频层复用硬件设备：在网络规划时，尽量避免将GSM/UMTS/LTE基础覆盖频层和容量频层的4G、5G频层部署在同一个射频设备上。在无法避免的情况下，尽量规划其他的基础覆盖层，并在需要节能的时候，将驻留在共用设备上的基础覆盖层用户迁移到独立覆盖层，从而创造关闭容量频层设备的条件。

#### 2) 硬件设计

硬件的设计不仅要考虑如何提高复用度、降低设备体积和重量，还要考虑如何避免影响硬件的可关闭性。具体而

言，在硬件设计中应提升节能能力相近的无线制式的硬件复用度，避免节能能力差异较大的无线制式复用同样的硬件器件。将节能设计理念和约束引入到硬件设计中，有助于在成本、体积、重量和节能之间找到平衡点。

### 3.3 平衡节能与网络性能、用户感知

节能会导致用户迁移、容量频层关闭，这在一定程度上会影响网络性能和用户感知。但是从更宏观的视角做深层逻辑分析，节能未必导致网络性能下降和用户感知下降。

为了保障覆盖、性能和感知，无线网络常通过“做加法”（如提升发射功率、增加信道控制等）来实现优化。这些优化在一定程度上提升了覆盖、性能和用户感知，但也会带来一些不利影响。比如，功率提高会增加干扰，带宽扩容会加剧系统间的冲突，差异化业务处理会导致低优先级用户业务感知受损。而节能在低负荷时段关闭容量频层小区，可以降低区域内的干扰，进而能提升区域内用户的感知。因此，从更大范围、更长时间的观测，节能未必导致用户感知下降。

此外，节能所引发的用户迁移，未必会导致移动性指标、连接类指标和用户感知类指标恶化。移动性指标、连接类指标和用户感知类指标，与用户的行为、业务变化有关。无线网络中，每天会产生大量的用户小区切换和用户业务变化。如果这种小区切换与业务变化导致移动性指标、连接类指标和感知类指标恶化，首先需要从网络的规划和优化上找原因和解决方法，而不应该关闭节能，损害节能收益。

### 3.4 AI助力节能策略寻优

首先，明确无线网络节能的目标：降低能耗是核心目标之一；保持网络关键KPI稳定、维持良好用户感知，同样是重要目标。综合来看，商用无线网络节能的核心目标是：在维持网络覆盖性能、保障关键性能指标（KPI）与用户感知达标的基础上，实现能耗最小化。商用网络的节能策略选择，本质上是多因素影响的非线性寻优问题。如何做到能耗、覆盖、网络指标和用户感知的平衡，是节能策略选择需要达到的目标。同一片无线网络在不同的时段和负荷条件下，网络性能或者用户感知的要求可以不同；同样，同一片网络在不同的时段和负荷条件下，为保证网络覆盖和用户感知，需要的网络资源也不同。比如，针对城市热点区域，在深夜节能时间段，要把降低能耗作为优先考虑目标；在白天低负荷的节能时段，要平衡能耗和指标；在白天繁忙时段，要保证网络指标和用户感知优先。

其次，节能策略部署的区域越大，策略的复杂程度就越高。这是因为，目标区域越大，包含的站点越多，要考虑和

分析的输入参数就越多,节能策略就越复杂。

最后,节能策略属于多因素共同作用的非线性多目标寻优。对于多因素共同作用的非线性多目标寻优,传统的寻优方法显得力不从心。

基于神经网络的深度学习被证明是可行的解决方法。图2是节能策略智能决策系统的结构框图,它主要由数据维护模块、模型训练模块、策略制定模块、策略执行模块组成。

- 数据维护模块:主要功能包括学习数据维护/提供、获取并维护基站的反馈数据、向模型训练模块提供训练数据、向策略制定模块提供输入数据<sup>[7]</sup>;

- 模型训练模块:主要功能包括基于训练数据进行神经网络的学习、将学习后的结果部署/更新到策略制定模块、接收策略制定模块的预期结果、将预期结果和网络反馈的实际结果进行对比、优化神经网络模型性能;

- 策略制定模块:主要负责接收模型训练模块的模型部署和更新、接收输入数据并进行节能策略的制定、将节能策略下发给策略执行模块;

- 策略执行模块:主要负责接收策略制定模块下发的节能策略、将节能策略下发给各个目标网元。

在节能策略智能决策系统中,模型训练模块和策略制定模块需要使用神经网络来实现节能策略的寻优和决策<sup>[9]</sup>。

以卷积神经网络(CNN)方法为例,在工程实践中,特征过滤器可以基于深度学习学习方法学习得到;卷积层、池化层的子层数量及卷积层深度,则需依据具体任务场景灵活配置;输出结果也需要根据实际的计算目标进行针对性设定。依托神经网络算法的自优化能力,可在系统能耗与网络性能(如精度、推理速度)之间建立动态权衡机制,进而生成适配具体应用场景的最优节能策略。

基于神经网络算法的智能节能,通过对网络负荷和网络关键指标的分析,对可节能时间段进行动态调整。通过延长可节能时间段、在指定时段内使用节能效果更优的节能方法,可获取更高的节能收益。图3是在4G/5G混合、多频段网络中,应用AI算法对21:00(前一天)—6:45(第二天)时段内的载波关断、深度休眠和自动启停节能时段进行调整的结果。调整后的结果表明,自动启停的节能时段比例明显增加,

载波关断和深度休眠的节能时段比例存在下降。这种改变使得验证区域内的节能收益增加了15%,能效提升了11%。

#### 4 基于“大模型+大数据”的逆向方案

AI助力节能策略寻优,使用神经网络来解决多入参的非线性多目标寻优问题。这种方法存在难于收敛、计算复杂等问题。大模型的出现为这种非线性多目标寻优问题的求解带来了新的解决方案。基于大模型,我们可以通过从结果向前追溯的逆向求解方案来解决该问题。

图4展示了基于大模型+大数据的逆向求解方案,该方案主要包含下面5个部分:

- 1) 训练大模型:从商用网络采集海量节能数据,训练并构建大模型;

- 2) 构建优秀节能案例库:从商用网络采集大量的优秀节能数据,构建优秀节能案例知识库;

- 3) 为目标网络匹配适用的节能数据集:针对需要做节能优化的目标网络,提取其特征数据并“喂给”大模型。大模型依据目标网络中的典型特征,在优秀节能案例知识库中匹配最合适的案例集合;

- 4) 节能数据集预处理:将选择到的案例样本数据(其中包含节能策略和参数配置),传递到策略执行单元,并对案例

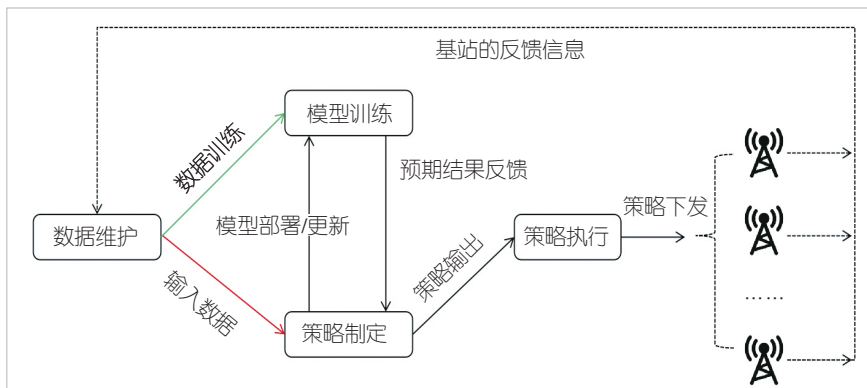


图2 节能策略智能决策系统框图<sup>[8]</sup>

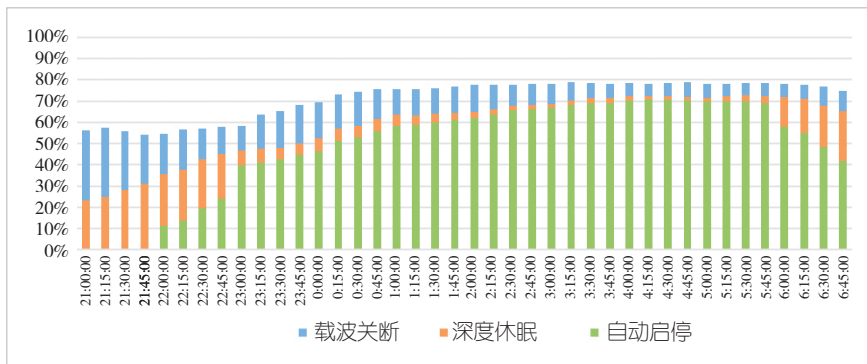


图3 应用AI算法优化后的节能功能时长占比

样本数据集进行预处理；

5) 向目标网络下发节能数据集：策略执行单元将预处理后的节能数据集，下发给目标网络中的各个节点。

通过策略选择和多次尝试，从中选择最适合目标区域网络的节能策略。

大模型+大数据的逆向方案具有如下3个突出优点：

1) 优秀节能案例知识库中的案例集合具有良好的可迁移性：其数据集合均是从商用网络中采集的有效节能策略/配置数据，且已从中筛选出有效的策略/配置样本。只要是特征类似的网络或者基站，都可以直接复用；

2) 避免神经网络算法的“不收敛”

问题：筛选到的节能案例数据集，是经过商用实践验证的有效数据集合，将其用于目标网络的节能优化，可有效避免神经网络计算的不收敛、收敛慢的问题；

3) 系统支持持续迭代：数据越丰富，模型的精度越高，特征匹配和节能数据选择的精度也会同步提升。

因此，基于大模型+大数据的逆向求解方案，在降低节能策略寻优算法复杂度的同时，维持网络覆盖性能、保障关键KPI指标和用户感知达标，并进一步实现能耗最小化、能效最优化。

## 5 结束语

本文介绍了当前主流的无线设备节能方法，对无线基站节能的主要问题进行分析，针对神经网络算法的弊端，提出了基于大模型+大数据的逆向求解方案。随着全球气候挑战的不断加剧，绿色低碳已是大势所趋。如何结合网络、基站特征选择有效的节能策略，在保持网络覆盖质量、保障用户感知的基础上，尽可能降低设备能耗，是5G及未来无线网络的重要研究方向。

### 参考文献

- [1] 郭诚, 陈梦竹. 面向5G-A的无线网络节能关键技术 [J]. 中兴通讯技术, 2022, 27(2): 11-15. DOI:10.12142/ZTETJ.202306003
- [2] 周均翼, 周琳, 张舜卿. 面向节能减排的跨制式融合感知通信系统 [J]. 中兴通讯技术, 2022, 27(2): 16-22. DOI: 10.12142/ZTETJ.202306004
- [3] SUN S Y, WEN G Y. Optimal design of wireless power transmission systems using antenna arrays [J]. ZTE communications, 2022, 20(2): 19 - 27. DOI: 10.12142/

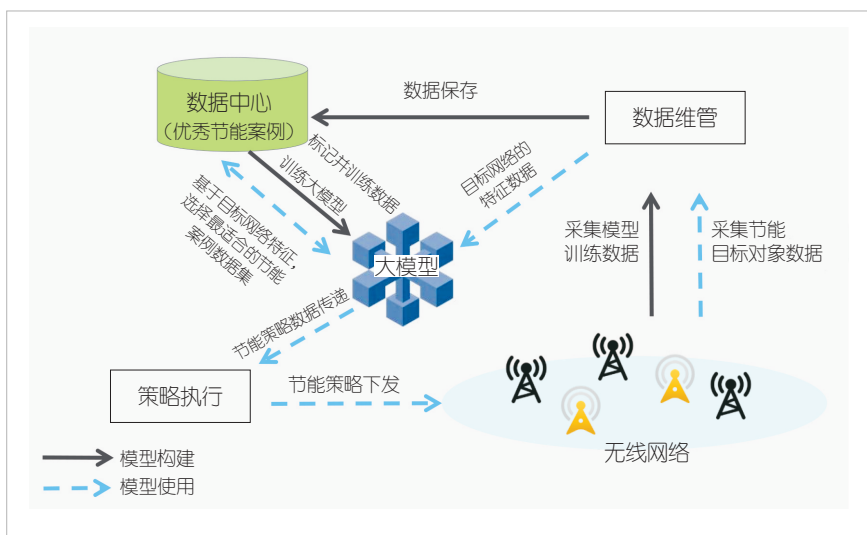


图4 基于“大模型+大数据”的节能策略逆向求解方案

ZTECOM.202202004

- [4] 陈天贝, 李娜, 陶小峰. 低开销智能反射面辅助无线通信研究综述 [J]. 中兴通讯技术, 2022, 27(2): 29-38. DOI: 10.12142/ZTETJ.202306006
- [5] 中兴通讯股份有限公司. AAU自动启停凝露分析技术白皮书 [R]. 2021
- [6] 王映民, 孙韶辉. 5G移动通信系统设计与标准详解 [M]. 北京: 人民邮电出版社, 2020
- [7] 3GPP. Study on enhancement for data collection for NR and EN-DC (Release 17): 3GPP TR 37.817 [S]. 2022
- [8] 3GPP. Study on system and functional aspects of energy efficiency in 5G networks (Release 17): 3GPP TR 32.972 [S]. 2024
- [9] 伊恩·古德费洛, 约书亚·本吉奥, 亚伦·库维尔. 深度学习 [M]. 赵申剑, 黎或君, 符天凡, 等, 译. 北京: 人民邮电出版社, 2017

### 作者简介



**王小锋**，中兴通讯股份有限公司RAN产品规划总工；主要研究方向为5G无线产品、大规模MIMO、无线网络节能等；先后参与WiMax、LTE和NR系统的研发与规划工作；申请发明专利4项。



**韩茜**，中兴通讯股份有限公司产品经理；主要研究方向为LTE、5G测试终端物理层和系统设计与开发；先后参与过WiMax系统研发，LTE、NR测试终端的研发和项目交付工作；拥有专利2项。



# 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

陈 力 中国科学技术大学

陈 为 北京交通大学

陈琪美 武汉大学

陈舒怡 哈尔滨工业大学

陈思衡 上海交通大学

官 科 北京交通大学

韩凯峰 中国信息通信研究院

何 姿 南京理工大学

侯天为 北京交通大学

胡 杰 电子科技大学

黄 晨 紫金山实验室

李 昂 西安交通大学

刘 凡 东南大学

刘春森 复旦大学

刘俊宇 西安电子科技大学

卢 丹 中兴通讯股份有限公司

陆游游 清华大学

宁兆龙 重庆邮电大学

祁 亮 上海交通大学

秦晓琦 北京邮电大学

秦志金 清华大学

史颖欢 南京大学

唐万恺 东南大学

王景璟 北京航空航天大学

王兴刚 华中科技大学

王勇强 天津大学

温森文 华南理工大学

吴泳澎 上海交通大学

武庆庆 上海交通大学

夏文超 南京邮电大学

徐梦炜 北京邮电大学

徐天衡 中国科学院上海高等研究院

杨川川 北京大学

尹海帆 华中科技大学

于季弘 北京理工大学

张 娇 北京邮电大学

张宇超 北京邮电大学

章嘉懿 北京交通大学

赵昱达 浙江大学

赵中原 北京邮电大学

周 伊 西南交通大学

朱秉诚 东南大学

### 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、  
陆 平、史伟强、屠要峰、王会涛、熊先奎、  
赵亚军、赵志勇、朱晓光

双月刊 1995 年创刊  
第 31 卷 总第 185 期  
2025 年 10 月 第 5 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: 王利  
执行主编: 黄新明  
副主编: 卢丹  
编辑部主任: 王萍萍  
责任编辑: 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: 2025 年 10 月 25 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