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 32 卷 · 总第 187 期 · 2026 年 2 月 · 第 1 期

专题
6G 关键技术的标准化：Day-1 与未来

ISSN 1009-6868
9 771009 686267 02>

第二十八届中国科协年会学术论文专刊

《中兴通讯技术》第 10 届编辑委员会

目　次

中兴通讯技术　(ZHONGXING TONGXUN JISHU)
第 32 卷　　总第 187 期　2026 年 2 月　　第 1 期

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

《中兴通讯技术》2026 年热点专题名称及策划人

1. 6G 关键技术的标准化：Day-1 与未来
中国移动研究院首席科学家　易芝玲

2. 广域立体覆盖低空通信技术
东南大学副校长　金石
东南大学教授　刘凡

3. 智算网络
工信部信息通信科技委常委　赵慧玲

4. 大模型推理中的存算技术
中国工程院院士、清华大学教授　郑纬民
清华大学副教授　陆游游

5. 星地太赫兹高速传输技术
中国科学院院士、东南大学教授　洪伟
电子科技大学教授　唐万斌
东南大学教授　郝张成

6. 智能多天线技术
北京交通大学副校长　艾渤
北京交通大学教授　章嘉懿

MAIN CONTENTS

ZTE TECHNOLOGY JOURNAL
Vol. 32 No. 1 Feb. 2026

期刊基本参数：CN 34-1228/TN*1995*b*16*86*zh*P*¥20.00*6500*12*2026-02

致通信人——2026新年献词

◎ 张平/中国工程院院士、本刊编委

在新的科技革命与产业变革深入发展的历史关口，中国通信事业正站在又一次伟大飞跃的前夜。回顾过去20余年的通信发展历程，中国走出了一条从追赶到并跑、再到领跑的创新之路。如今，站在6G发展的新起点，面对“人工智能+通信”带来的范式变革，中国科研力量以勇闯无人区的精神底气，推动以信息论统一理论框架为基础的新型通信范式等原创理论迈向世界前沿，充分彰显出中国式现代化建设在科技创新领域的战略定力与制度优势。

如果说从3G到5G的发展是通信技术的持续升级，是经典信息论延长线上的技术创新，其基本特征就是依靠技术的堆叠、资源特别是能量的消耗在延长线上推动移动通信的代际演进。进入人工智能时代后，大数据对于数据带宽的要求越来越高，而带宽的增加是需要资源来弥补的。也就是说，在6G时代，人们已经无法沿着传统延长线继续下去，迫切需要找到一个“拐点”来满足对带宽的需求。而这个拐点技术将标志着通信领域前所未有的范式转变。

传统通信理论范式已难以满足“人工智能+”时代的新需求。海量“物端”设备产生的多样化数据需求，亟需更高效的通信范式支撑。语义通信的提出，彻底改变了通信的核心逻辑，即通信不再追求简单的信息符号搬运，而是强调任务导向、智能理解与高效协作。这不仅是技术层面的升级，更是通信基础理论的根本性重构。

现代语义通信的核心优势，在于通过端到端的智能学习模型，让传输系统能够精准地理解任务意图，而非机械地搬运符号。这种方式大幅提升了通信效率，显著降低了网络带宽与能源资源消耗。这一创新思路突破了通信的传统边界，被国际学界普遍视为“第二次通信革命”。

然而，这样的革命需要回答几个棘手问题：一是支撑这个革命的基础理论是什么？二是如果这个理论是对传统理论的颠覆，如何解释这近百年传统通信的成功？三是如何解释人工智能对通信系统赋能的泛化性？

正是站在这样一个前所未有的“风口”上，我希望我们全体通信人能够拥有趟过“无人区”的坚强决心，在新的一年用我们的研究成果给出消除未来不确定性的标准答案，从而构建起新的科技大厦之基。

马年来临，祝福大家在新的征途中，马到成功，扬眉吐气，马不停蹄地追求更好的科研成果。

作 者 简 介

张平，国务院参事，中国工程院院士，北京邮电大学教授、网络与交换技术全国重点实验室主任，《通信学报》主编，IEEE Fellow，IMT-2020（5G）专家组成员，IMT-2030（6G）推进组咨询委员会委员，国家自然科学基金委“创新研究群体”带头人；长期致力于移动通信理论研究和技术创新，为中国自主技术成为国际主流做出了基础性的贡献；曾任国家重点基础研究发展计划（“973”计划）首席科学家、国家高技术研究发展计划（“863”计划）主题专家组专家等；获国家科技进步奖特等奖、一等奖，国家技术发明奖二等奖3项，科技进步奖二等奖2项，全国创新争先奖章，光华工程科技奖何梁何利科学与技术进步奖等奖项，入选教育部首批“黄大年式教师团队”。

DOI：10.12142/ZTETJ.202601001
收稿日期：2025-12-20

6G关键技术的标准化：Day-1与未来专题导读

专题策划人

易芝玲

中国移动研究院无线技术首席科学家、IEEE终身会士及WWRF会士，担任O-RAN技术指导委员会创始主席、FuTURE 5G/6G特别兴趣小组创始主席、无线AI联盟执行委员会主席，并曾任职GreenTouch执行董事等；长期致力于无线通信与移动网络研究，当前主要聚焦于ICDT深度融合，尤其在绿色通信、开放智能网络等前沿领域；曾获IEEE ComSoc Stephen O. Rice最佳论文奖、IEEE ComSoc Fred W. Ellersick最佳论文奖及IEEE工业引领创新奖等多项国际荣誉；在学术研究与产业应用方面成果丰硕，已发表论文200余篇，持有专利100余项，谷歌学术引用近25 000次，合著/编多部学术专著，并在全球发表100余场主题演讲。

目前，全球6G愿景已在ITU框架下达成共识，移动通信网络正加速从传统“连接管道”向“一体化智能服务平台”演进。随着3GPP R20标准研究启动，6G已步入需求定义与技术框架研究的关键窗口期。在这一宏大技术图景中，系统性的思考至关重要：6G Day-1阶段的标准化工作直接关系到技术方案的商业落地与产业生态的初步构建，而面向未来的前瞻性布局则决定了网络能力的长期演进潜力与战略价值。本专题聚焦通算智融合架构、内生智能（Native AI）、通信感知一体化（ISAC）、无蜂窝大规模多输入多输出（MIMO）及沉浸式通信等核心技术方向，系统梳理其演进脉络、标准化路径与应用前景。

《6G无线接入网通算智融合关键技术与标准化思考》提出由基础设施层、网络功能层与编排服务层组成的通算智融合框架，通过对异构算力管理、任务驱动数据采集及AI模型全生命周期管理等使能技术的分析，探讨集中式与分布式融合架构的演进路径，为6G初期构建内生智能平台化网络提供技术支撑。《6G通感一体化关键技术和标准发展》系统梳理ISAC技术从探索到落地的路径。在Day-1标准化方面，重点介绍3GPP R19框架下的感知目标与背景信道联合建模方法，并提出面向远距离探测的脉冲感知参考信号设计。面向未来，探讨感知与定位、通信、AI的协同机制，推动通信能力向环境理解能力转化。《6G沉浸式通信业务与关键技术探索》面向全息通信、数字人等高带宽低时延场景，提出基于智能内生特性的6G新型服务质量（QoS）架构，引入包级细粒度控制与保证速率非预留资源类型（E-Non-GBR），提升用户面动态调度能力，并展望基于AI Agent通信的意图感知技术，为沉浸式业务智能化承载提供新思路。《6G无蜂窝大规模MIMO关键技术研究进展》系统梳理6G无蜂窝通信相关支撑技术，提出6G空口应在多载波与子带全双工（SBFD）基础上形成简洁高效的接入传输方案，突破以小区为中心的资源分配体系，释放多收发点容量潜力，并提出数字孪生增强的无蜂窝传输优化方法，提升大范围组网性能。针对全球导航卫星系统（GNSS）拒止环境下的高精度定位挑战，《基于OFDM索引调制的通信定位方法》提出融合惯导与正交频分复用（OFDM）网格编码索引调制的方案，通过子载波位置隐式传递导航信息，在不增加频谱开销下实现高效解码，在低信噪比下具有显著鲁棒性，为6G复杂场景高精度定位提供可行技术选项。《6G内生智能与信道基础模型》探讨内生智能作为6G核心特征的长远愿景。针对传统专用AI模型的局限性，提出信道基础模型（CFM）概念，采用“预训练-微调”范式学习无线信道物理规律，为物理层处理、接入网优化及通感一体化等场景提供强泛化能力的智能化底座。

本期作者汇聚知名高校、企业与科研机构专家学者，围绕6G关键技术标准化，从需求分析、系统架构到核心算法与应用实践，系统呈现最新研究成果。期待这些工作为中国6G标准Day-1布局与未来演进提供有益启示。谨对所有作者和审稿专家表示衷心感谢！

DOI：10.12142/ZTETJ.202601002
收稿日期：2026-02-15

6G无线接入网通算智融合关键技术与标准化思考

Reflections on Key Technologies and Standardization of Communication, Computing, and Intelligence Integration in 6G Radio Access Networks

解宇瑄/Xie Yuxuan，李响/Li Xiang，李婷/Li Ting，孙奇/Sun Qi

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

DOI：10.12142/ZTETJ.202601003
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260227.1320.002
网络出版日期：2026-02-27
收稿日期：2026-01-06

摘要：针对6G RAN通算智融合的需求与挑战，提出了由基础设施层、网络功能层与编排服务层三层组成的通算智融合架构体系。该体系突破了以单一连接为核心的传统范式：在基础设施层，实现异构智算硬件资源的统一管理与边缘基站算力的强实时高效能调度；在网络功能层，支持原生多维数据精细管控与高效传输、AI计算及AI模型全生命周期管理等核心功能；在编排服务层，构建面向差异化业务的通算智联合编排与服务开放机制，从而实现通算智多维资源的高效协同与多元服务的定制化保障。在此基础上，进一步探讨了6G无线通算智融合的核心使能技术和标准化路径，旨在通过重构RAN的基础能力与服务形态，为实现6G泛在智联的愿景提供坚实的技术支撑与系统基座。

关键词：6G；无线接入网；通算智融合；标准化

Abstract: To address the requirements and challenges of the convergence in 6G radio access network (RAN), a three-layer integrated architecture consisting of an infrastructure layer, a network function layer, and an orchestration and service layer is proposed. This architecture moves beyond the traditional connectivity-centric model. At the infrastructure layer, it enables unified management of heterogeneous intelligent computing hardware and real-time, high-efficiency scheduling of computing resources at edge base stations. The network function layer natively supports fine-grained multi-dimensional data management and efficient transmission, AI computation, and full lifecycle management of AI models. At the orchestration and service layer, it establishes a joint orchestration and service exposure mechanism tailored to diverse service requirements. In this way, the architecture achieves efficient coordination of multi-dimensional resources and customized provisioning of differentiated services. Key enabling technologies and potential standardization pathways for communication-computing-intelligence convergence in 6G are further explored. By reshaping the foundational capabilities and service paradigms of the RAN, solid technical support and a systematic foundation are provided for realizing the vision of ubiquitous intelligent connectivity in 6G.

Keywords: 6G; RAN; communication-computing-intelligence convergence; standardization

引用格式：解宇瑄, 李响, 李婷, 等. 6G无线接入网通算智融合关键技术与标准化思考 [J]. 中兴通讯技术, 2026, 32(1): 3-12. DOI: 10.12142/ZTETJ.202601003

Citation: Xie Y X, Li X, Li T, et al. Reflections on key technologies and standardization of communication, computing, and intelligence integration in 6G radio access networks [J]. ZTE technology journal, 2026, 32(1): 3-12. DOI: 10.12142/ZTETJ.202601003

以生成式人工智能（AI）为代表的新一代人工智能技术正加速推动人类社会迈向通用智能时代，并逐步成为影响社会发展形态的新型基础设施。这一演进趋势对信息基础设施提出更高要求，亟需构建具备泛在、融合与智能特性的新型网络，促使通信、计算与智能的深度融合成为6G网络发展的核心方向之一。国际电信联盟（ITU）已明确将AI与通信融合列为6G的关键应用场景，这不仅标志着移动通信网络技术范式的重大转型，也将推动其服务形态从传统的“连接管道”向“一体化智能服务”体系实现根本性跃迁[1]。

在6G无线接入网（RAN）中，通算智的深度融合呈现双向赋能的内在逻辑，形成两条并行交织的发展主线。一是AI赋能网络，即利用AI技术应对网络运行、运维与优化的复杂挑战，实现网络的提质、增效与降本。该方向主要由网络应用需求驱动，聚焦于网络内生数据的深度治理及面向网络场景的专用模型构建。二是网络赋能AI，即通过网络能力的增强，为上层AI业务应用提供超越传统连接的深度支

持，进而开拓新型服务模式与价值空间。该方向主要由AI应用场景驱动，重点解决AI模型、数据与计算任务在网络中的高效流动与协同处理问题[2-3]。

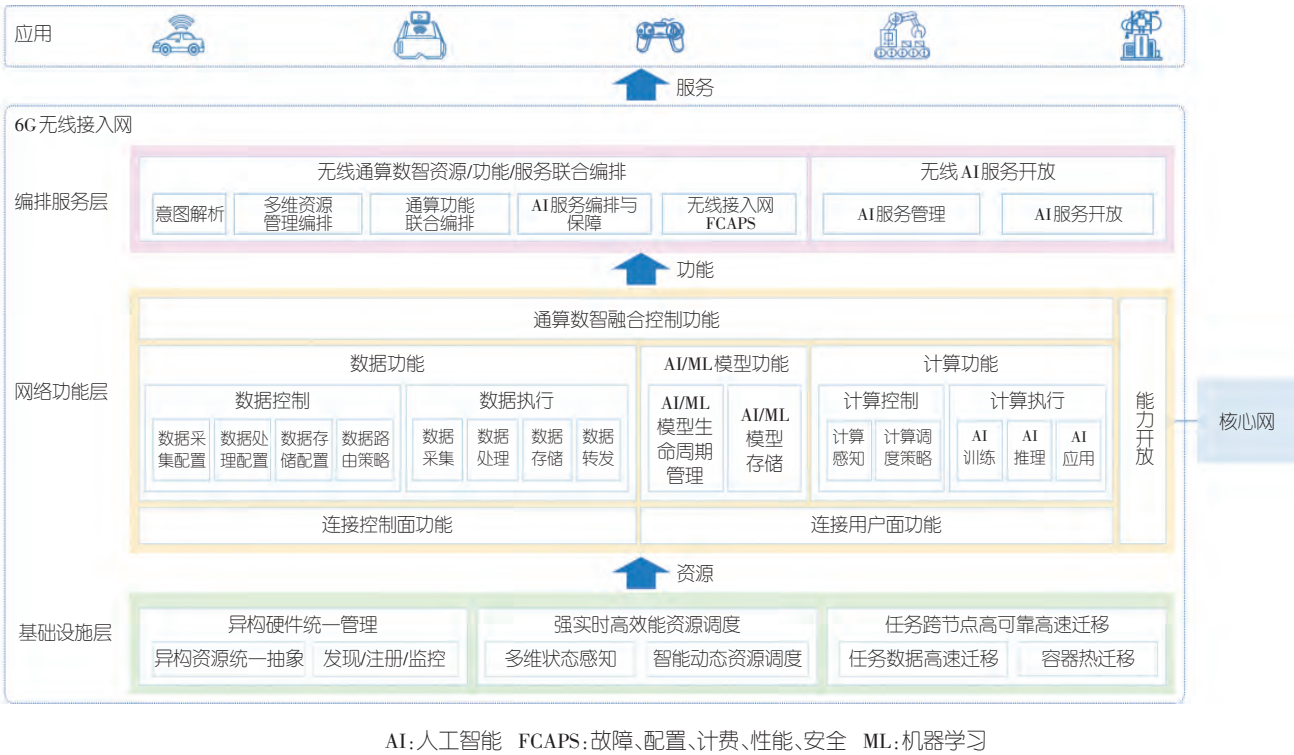
为实现AI与RAN之间的双向赋能，6G RAN需开展系统性架构重构。当前6G RAN面临若干关键设计挑战：首先，分布于基站与终端侧的无线算力是实现“连接+计算”深度融合的物质基础，然而其存在硬件异构性强、资源分布泛在、负载动态波动性强等特征，如何实现统一纳管、资源池化和高效调度成为首要难题。其次，现行RAN以连接功能为核心，对数据、AI模型及计算资源等要素缺乏原生设计，亟需探索将多维AI要素内嵌于网络功能体系的方法，以提升端到端性能与资源利用效率。再次，传统RAN功能固定、能力开放有限，为构建真正的“智能服务平台”，须支持跨域资源、功能与服务的一体化编排，实现边缘AI能力的可度量、可调用与可开放，以培育可持续的智能服务生态。

在标准化方面，第3代合作伙伴计划（3GPP）R18/R19阶段已在5G-Advanced框架下围绕AI空口增强开展探索，为6G更深入的融合积累了技术基础。随着3GPP R20标准研究的启动，6G RAN已进入实质性的需求定义与框架设计阶段。作为核心议题，通算智融合正推动包括用例定义、内生智能架构、数据与模型管理等方向的共识凝聚。本文首先提出6G无线通算智融合的系统框架，继而阐述其关键技术路径，结合标准化进展提出演进建议，最后展望发展前景。

1 6G无线接入网络通算智融合系统框架

为同时满足上述双向赋能需求并有效应对多维挑战，6G RAN需遵循“内生智能化、平台化、服务化”的设计理念，构建通算数智深度融合的系统框架。本文提出的框架如图1所示，该架构采用分层架构，由基础设施层、网络功能层与编排服务层3部分构成。

基础设施层对通算数智多维资源实现统一纳管与强实时调度，为上层提供稳定、可扩展的虚拟化资源底座。网络功能层基于基础设施层的支撑能力，在传统无线连接功能的基础上，引入数据治理、AI/机器学习（ML）模型生命周期管理、算力调度以及通算数智融合控制等新功能，实现从外挂智能向内生智能的演进。编排服务层以业务意图为导向，对跨域资源、网络功能及AI服务进行全局智能编排，实现从资源、功能到服务的端到端自动化封装与按需交付。三层之间灵活协同，共同构建资源易扩展、功能原生支持AI处理、能力可编排且可对外开放的边缘通算数智综合能力平台[4]。各层核心构成与功能分述如下：



AI：人工智能　FCAPS：故障、配置、计费、性能、安全　ML：机器学习

图1　6G无线接入网络通算智融合系统框架

1）基础设施层

基础设施层由通过无线接入网互联的终端、基站、边缘云及数据中心等节点组成，提供包括连接、计算、数据及AI模型在内的多维虚拟化资源[5]。

该层的管理功能涵盖异构硬件统一管理、强实时高效能调度以及任务跨节点高可靠高速迁移等核心模块，旨在为上层网络功能提供稳定、动态可调度的基础算力支撑。

（1）异构硬件统一管理功能：将不同架构的计算资源抽象并整合为统一可调度的资源模型，实现跨异构环境资源的一体化供给与高效利用。

（2）强实时高效能调度功能：持续采集并更新各类资源的实时负载、性能指标及拓扑状态，依据AI任务在时延、吞吐量等方面的服务质量需求，开展智能化动态资源分配与调度优化，确保任务执行过程中的服务保障。

（3）任务跨节点高可靠高速迁移功能：在算力节点发生故障或性能受限时，借助远程直接内存访问（RDMA）、容器热迁移等技术实现AI计算任务的无缝迁移与接续，保障任务处理的可靠性及业务连续性。

2）网络功能层

网络功能层在传统无线接入网连接处理能力的基础上，系统拓展了面向AI全要素的处理能力，引入数据、AI/ML模型与计算等智能化功能模块，实现从连接管道向智能服务平台的演进。具体功能构成如下：

（1）连接功能：在传统无线连接能力基础上，为高效支撑海量AI数据的空口传输（如训练数据、AI/ML模型梯度及参数等），连接功能拓展了AI数据承载能力，使无线接入网能够自主、动态地建立面向AI数据流的专用无线承载，突破传统控制面与用户面在传输AI数据时的效率与性能约束[6–7]，显著提升AI数据在空口的传输效率。其中，控制面功能负责AI数据专用承载的全生命周期管理，包括承载的建立、维护与释放；用户面功能则负责对此类承载上的AI数据进行空口传输与协议处理。

（2）数据功能：数据功能涵盖数据控制与数据执行两个层面。数据控制功能负责数据采集、处理、存储及路由策略制定等环节的全生命周期管理，可根据AI算法需求自主规划采集节点、数据对象与上报方式，实现任务驱动的定向数据采集，并依据数据处理所在节点制定数据路由策略，提升采集与传输效率。数据执行功能则具备采集、处理、转发与存储的执行能力。其中，数据采集是指按照预设规划从基站、终端等节点获取所需数据；数据处理包括对原始数据进行清洗、增强、格式转换及加密等操作，供AI/ML模型调用；数据转发依据路由策略对数据进行定制化分流，实现高效灵活的数据流转；数据存储采用集中与分布式相结合的方式，依据数据生命周期制定存储策略，在保障效率的同时降低存储成本。

（3）模型功能：负责AI/ML模型的全生命周期管理，涵盖模型训练、推理、测试验证、部署及运行监控等环节，实现对模型资源的系统性管控与实时监测[8]。考虑到AI/ML模型具有数据驱动与概率性行为特征，其在复杂动态无线环境下的表现更难预测，为满足6G网络对性能与可靠性的极致要求，模型功能可持续跟踪推理精度、时延、置信度等核心性能指标，保障模型在动态环境下的稳定运行与持续优化。

（4）计算功能：实现计算控制、计算执行与计算结果反馈的闭环管理。计算控制负责生成AI任务的算力分配、任务部署与调度、任务挂起与释放等指令，并交由基础设施层执行；计算执行主要指模型训练及推理计算过程；计算结果反馈是指在计算完成后，将最终结果或中间状态数据返回至需求端，形成完整的业务闭环。

（5）通算数智融合控制功能：负责在无线接入网内执行细粒度、动态化的资源协同决策，是面向终端业务的近实时控制核心。基于对终端连接状态、实时计算负载、业务性能指标及上下文环境的综合分析，生成面向连接、数据、AI/ML模型与计算资源的联合优化策略，通过控制信令驱动各功能执行，实现网络性能、业务体验动态优化。

（6）能力开放功能：负责对连接、数据、AI/ML模型和计算等功能提供的多维资源信息进行统一封装，并通过标准化接口向核心网及上层应用开放，使能第三方业务对无线网络智能能力的按需调用，构建开放、协同的智能服务生态。

3） 编排服务层

编排服务层包含无线通算数智资源/功能/服务联合编排与无线AI服务开放两大功能。其中，联合编排功能基于下层上报的资源状态与网络功能信息，开展跨域分析与全局优化，动态生成涵盖连接、数据、AI/ML模型及计算的一体化部署与优化方案。无线AI服务开放功能则对网络内生智能能力进行服务化封装，通过标准化开放接口向各类垂直应用提供可定制、可保障的边缘AI服务，为智能化创新构筑关键网络基石。

2 6G无线通算智融合核心使能技术及标准化演进

基于上一章提出的6G无线接入网通算智融合分层系统框架，本章将系统阐述支撑该框架的关键使能技术，并结合标准化发展路径提出演进思考。整体内容将围绕网络功能逻辑架构、基础设施层、网络功能层、编排服务层4个维度展开。

2.1 网络功能架构及标准化思考

IMT-2030（6G）推进组与3GPP等国际标准组织已系统性地开展了人工智能在无线接入网中的应用场景研究与用例设计。在AI赋能的具体实现形态上，物理层与高层呈现出显著差异。其中，基于AI的空口物理层接入技术需具备毫秒级甚至传输时间间隔（TTI）级别的实时推理能力，这要求在基站内部集成AI算力以满足其极低的时延约束。高层应用场景，如移动性优化与智能网络切片等，则依赖于基于策略模型的AI推理，要求实现小于10 ms的近实时性能，并需依托跨站点数据和部署于基站近端的算力资源，以保障AI/ML模型的强泛化能力与近实时计算效率。传统5G网络的智能能力通常以外挂式、补丁化后置添加的方式引入，难以满足物理层等场景对实时处理的苛刻要求。此外，由于通信与智能在控制层面相互分离，系统难以实现二者的高效协同，已成为网络智能化潜能释放的核心瓶颈。为同时满足上述需求，6G通算智融合的无线接入网在架构设计层面存在集中+分布式与全分布式两种潜在逻辑架构[9]，如图2所示。

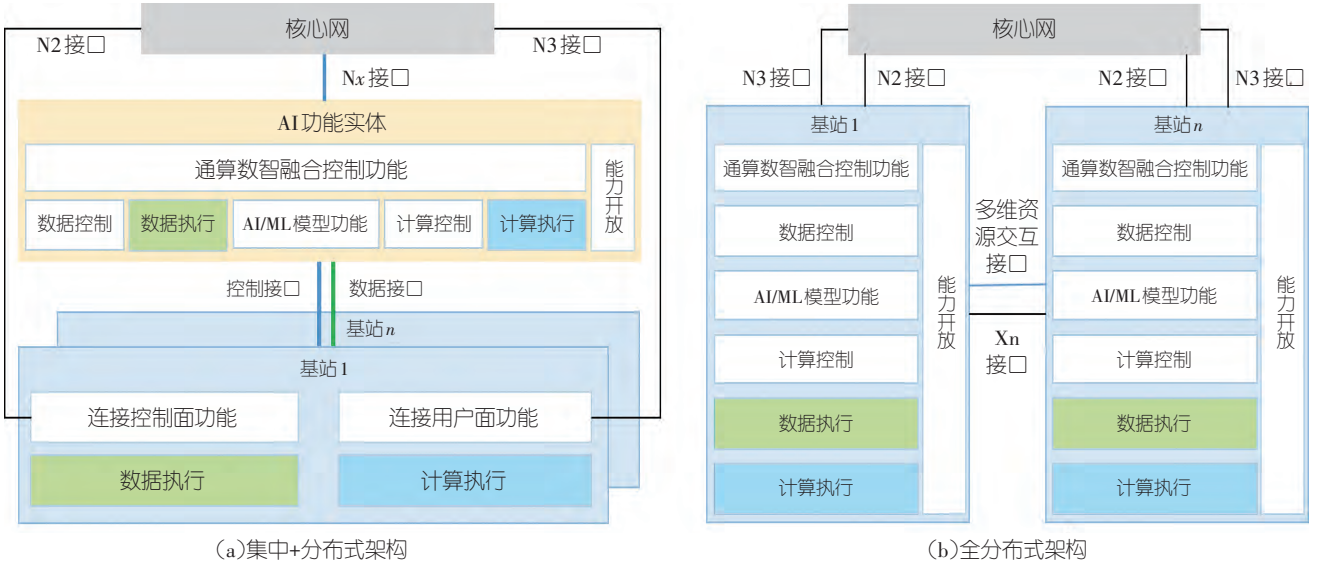
1） 集中+分布式架构

在集中+分布式架构中，网络功能依据其时延敏感性、数据及算力需求进行逻辑分层。物理层等高实时性要求的功能在基站本地引入数据、AI/ML模型和计算能力；而高层功能具有近实时性及跨节点数据与模型处理需求，其数据、AI/ML模型和计算能力可集中聚合为区域级的数据池、模型库与算力池。该架构通过资源集中化实现池化增益与模型泛化能力的提升，有助于突破基站单点智能的局限性。

集中+分布式架构在无线接入网中引入了新的AI功能实体。为实现新功能实体与基站间的互联互通与高效协同，需在标准化层面明确该实体的功能定义，并规范其与基站之间的交互机制。首先，需定义区域集中的AI功能实体，明确其在数据、AI/ML模型、计算及融合控制等方面的核心功能特性。进而，设计AI功能实体与基站间的交互接口（如图2（a）中控制接口和数据接口）、协同流程及通信协议，以支撑跨基站的数据采集、AI/ML模型分发与计算任务调度等关键操作。此外，为向核心网及第三方应用开放无线接入网的AI能力，可基于3GPP已定义的通用应用程序编程接口（API）框架（CAPIF）[10]设计相应的能力开放接口（如图2（a）中Nx接口）与服务流程，从而在促进网络智能开放化的同时，有效控制系统复杂度与标准化开销。

2） 全分布式架构

在全分布式架构中，各基站均本地引入完整的通信、数据、模型与计算功能。该架构通过功能全下沉实现数据处理的高度本地化与极低时延，尤其适用于需毫秒级实时响应的站内AI推理场景。同时，为避免形成资源孤岛，基站间需设计新的接口（如图2（b）中多维资源交互接口）及交互流程，实现通算数智资源在基站间的按需流转。然而，全分布式架构在支持跨基站任务协同（如跨站数据采集、依赖集中式算力的模型训练与联合推理等）时，面临传输与算力开销的双重挑战。一方面，为实现协同，基站之间需频繁交互



（a）集中+分布式架构　　（b）全分布式架构

AI：人工智能　ML：机器学习

图2　6G通算智融合的无线接入网逻辑架构

信令与数据，导致站间传输负荷显著增加；另一方面，各基站均需配置足够的本地算力以应对峰值需求，易造成算力资源利用不均衡甚至闲置。

在标准化层面，需增强现有基站的功能定义，使其支持内生AI处理能力；同时，应对现有Xn接口进行功能扩展，或定义新型基站间接口，以实现AI多维要素的高效交互。

2.2 基础设施层关键技术及标准化思考

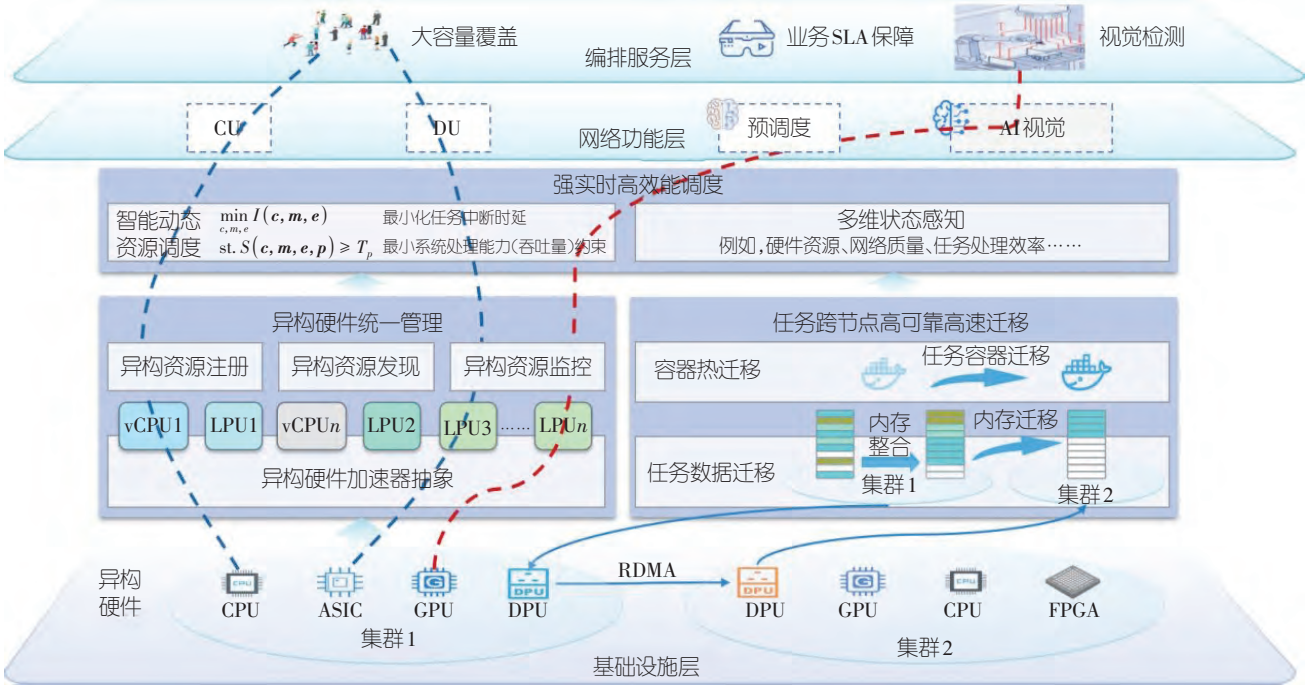
无线接入网向6G支撑“泛在AI”的通算智融合系统演进，基础设施层不仅需处理传统通信业务，更需高效支撑内生AI应用[11]。现有5G基础设施主要面向网络功能虚拟化（NFV），难以满足多样化AI工作负载对异构算力、极致性能与高可靠性的严苛需求。为此，本文提出面向6G无线接入网AI内生的基础设施关键技术，如图3所示。

异构硬件统一管理功能需要对图形处理器（GPU）、张量处理器（TPU）、现场可编程门阵列（FPGA）、数据处理单元（DPU）等不同架构的计算、网络与加速资源进行统一建模与抽象。例如，通过GPU多实例技术实现跨物理节点的异构算力资源池化与高效共享，并借助统一的异构资源注册、发现与监控机制，为6G泛在AI提供稳定、高性能、可动态调度的异构算力资源池。

强实时高效能调度功能首先基于扩展的伯克利包过滤器（eBPF）的非侵入、可动态加载特性，为基础设施及6G网络功能/应用提供细粒度、实时性的多维状态感知能力，涵盖硬件资源（如资源拓扑、CPU调度延迟、GPU利用率）、协议栈处理效率（如分组数据汇聚协议吞吐量、无线链路控制协议重传）及网络质量（如网络拓扑、GTP–U隧道丢包）等。在此基础上，通过智能动态资源调度构建融合算力类型、内存带宽、节点间网络状态、GPU多实例情况以及AI任务在时延、吞吐、可靠性等方面服务等级协议（SLA）需求的调度模型，设计动态资源分配与任务调度算法，在确保AI工作负载服务质量的同时，提升基础设施的资源利用效率。

高可靠跨节点迁移功能通过内存读写结合DPU+RDMA高速互联技术，为AI任务提供跨节点的低时延、高吞吐数据迁移，并设计业务“零感知”中断的任务容器迁移机制，有效应对节点故障或性能瓶颈，确保关键AI推理链路的业务连续性，从而大幅增强6G云化网络的弹性与可靠性。

在标准化层面，一方面，需增强面向GPU、DPU等异构



AI：人工智能　CU：集中单元　FPGA：现场可编程门阵列　RDMA：远程直接内存访问
ASIC：专用集成电路　DPU：数据处理器　GPU：图形处理器　SLA：服务等级协议
CPU：中央处理器　DU：分布单元　LPU：语言处理单元　vCPU：虚拟中央处理器

图3　基础设施层关键技术

硬件的池化与统一管理能力，其核心挑战在于加速器抽象层（AAL）的标准化。当前，由于加速器与硬件性能强绑定，不同厂商在架构、产品规格及接口参数等方面差异显著，且需针对协议栈进行定制开发，导致方案封闭、生态薄弱，标准进展缓慢。未来随着加速器技术发展及其通用性的提升，AAL的潜在价值有望得以释放。另一方面，需规范面向AI工作负载的性能与高可靠性保障策略。为此，标准需进一步细化基础设施层的功能设计、资源上报模型及北向服务化接口，以支撑对泛在AI应用从部署、动态调度到可靠性保障的全生命周期智能化管理，最终实现性能确定性与资源高效利用的平衡。

2.3 网络功能层关键技术及标准化思考

网络功能层是6G无线接入网实现通算智深度融合的执行载体。为构建智慧内生、高效协同的系统架构，需突破传统单一通信功能的局限，围绕数据、模型、算力、连接、控制与开放六大维度，构建全要素融合的技术体系。

2.3.1 低开销数据采集与高效传输

无线空口AI依赖海量、高频且细粒度的实时数据。当前RAN协议设计主要面向通信连接，对AI数据采集的支持尚显不足。传统基于操作、管理与维护（OAM）的周期性采集机制存在分钟级时延瓶颈；控制面受限于载荷大小，难以承载海量训练数据；而用户面数据需经核心网用户面功能（UPF）迂回，不仅导致数据传输时延增加与隐私泄露风险，且易挤占宝贵的空口连接资源。为此，需引入面向AI的原生数据流转机制，包括专用AI数据会话与支持动态筛选的任务驱动型定制化采集，以解决上述数据采集在实时性、灵活性与资源效率方面的不足，构建适应6G智能需求的轻量化、高价值数据供给体系。为进一步提升AI模型训练与优化的数据全面性，还需构建跨域协同的采集能力，形成内外协同的数据生态：

1）专用AI数据会话与承载：定义一种由RAN自主管理的逻辑通道（如图4所示），专门用于传输AI训练数据、模型文件及性能反馈。与传统数据流不同，该承载支持在RAN内部终结，无须绕行核心网，从而实现数据的本地闭环与低时延流转。此外，需在基站间接口（如Xn）建立数据隧道，以支撑跨节点的模型训练与分发。

2）任务驱动的定制化采集：针对AI推理对实时性（< 10 ms）与细粒度的需求，构建“采集过滤器”机制（如图5所示）。基站可下发包含数学逻辑及多维状态（如电量阈值、业务类型）的采集策略，终端据此仅上报满足特定AI任务需求的高价值数据，实现从“盲目全量采集”向“按需精准供给”的转变，从而大幅降低无效数据带来的空口开销。

3）跨域数据采集：为支撑更全面的AI模型训练与推理，需打通RAN内部与外部数据源，支持从OAM、核心网及相邻RAN节点等多域采集数据。通过统一的采集接口与策略协同，实现跨域数据的融合与价值提取，为全域智能优化提供数据基础。

在标准化层面，需定义AI数据会话的生命周期管理流程，规范数据采集任务描述语法及过滤器逻辑，并在空口（Uu）及站间接口（Xn）引入面向AI数据流的流控与传输优化协议，确保数据采集机制在多样场景下的可靠性、效率与互操作性。

2.3.2 高可靠AI/ML模型管理

AI模型在无线网络中的应用并非一次性部署，而是一个持续演进的动态过程。为应对无线信道的时变性与非平稳性，需构建标准化的模型全生命周期管理机制，如图6所示，确保模型在“训练–推理–验证–更新”闭环中的鲁棒性与可靠性。

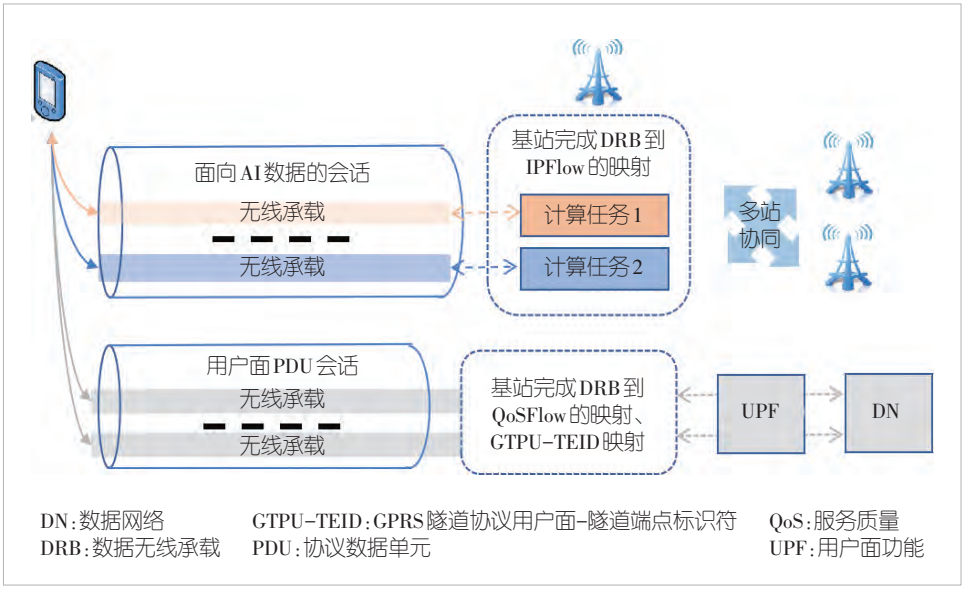


图4　AI数据会话机制

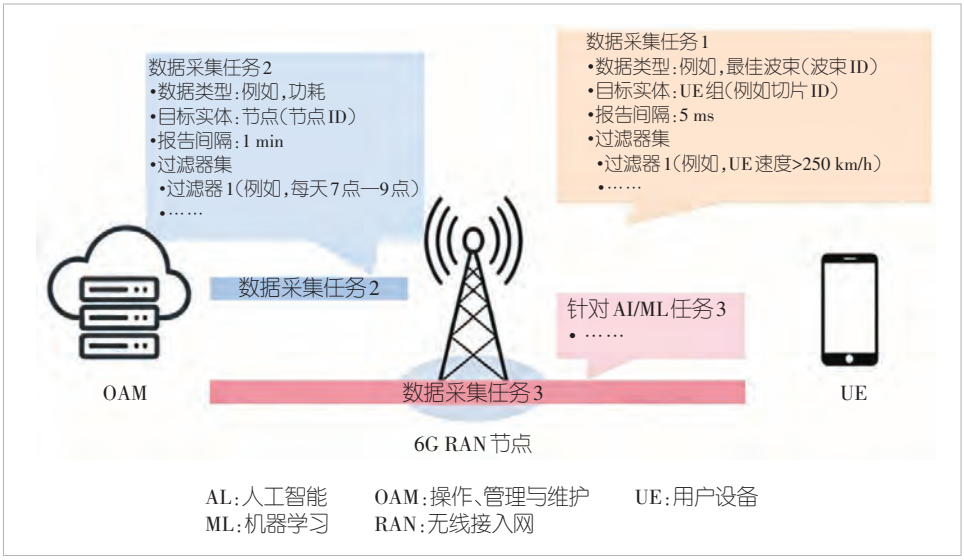


图5 任务驱动的定制化数据采集

1）三维性能监控体系：突破单一精度监控，构建涵盖AI模型性能（如准确率、置信度）、网络关键绩效指标（KPI，如吞吐量、时延）及资源利用率（如推理能耗、内存占用）的三维指标体系，以精准评估AI引入的效能比。

2）模型持续运行与演进：建立性能触发的闭环管理机制。当监测到模型性能劣化时，自动触发数据重采集、重训练与版本更新。同时，支持模型上下文在用户移动过程中的跨站同步，确保推理服务的连续性。

3）数字孪生预验证与新技术支持：利用RAN数字孪生环境[12]，在模型部署前进行鲁棒性测试与效果预演。此外，架构应具备扩展性，以支持联邦学习（FL）的分布式协作训练及大模型的轻量化适配，利用其强大的泛化能力应对复杂信道环境。

在标准化层面，需定义统一的模型元数据规范与版本溯源机制；规范三维性能监控指标的上报接口（如Uu、Xn及网管接口）；制定跨网元（UE–RAN–OAM）的模型生命周期管理信令流程（如模型切换、回退至非AI基线）。

2.3.3 AI业务连接保障

随着新型AI服务的普及，网络需保障模型梯度、token

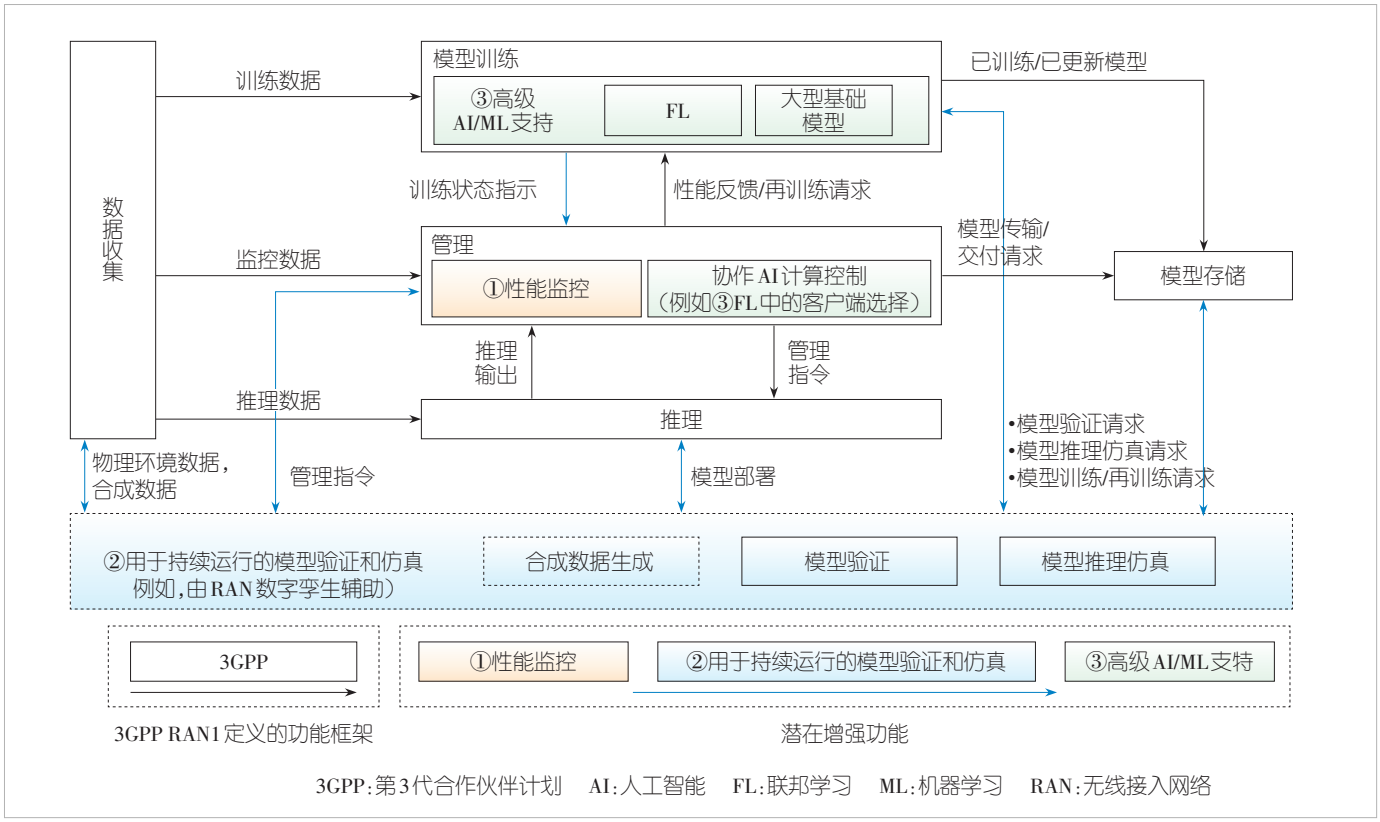


图6 AI/ML模型全生命周期管理机制

序列、多模态输入及控制指令等复杂数据流的可靠传输。这类业务呈现出上行突发密集、横向协同流量增多等新型特征，传统基于5G QoS标识符（5QI）的承载级QoS机制难以满足其差异化需求。因此，6G RAN需演进为AI业务实时感知、动态可调的连接保障体系，在识别AI业务基础上实现资源精细调度与通算协同优化，从而为多样化AI负载提供可靠、自适应的传输服务。

1）面向AI业务的QoS增强机制：在传统吞吐量、时延、丢包率等KPI基础上，引入面向AI业务流的特征识别机制，支持对模型更新、交互token、感知数据等多类AI数据流进行差异化处理。通过实时感知业务需求与网络状态，动态映射并调整QoS策略，实现对重要数据（如梯度更新）的优先调度与资源预留。

2）面向AI新型流量的空口传输保障：针对上行密集型（如个人助手）、横向流（如多智能体协同）等AI典型流量模式，设计低时延、高可靠的传输机制。通过模型分片传输、token级混合重传（HARQ）及选择性业务抢占等技术，在空口及协议栈层面实现确定性时延保障，确保关键AI指令与反馈的实时可达。

3）面向AI业务体验的通算一体优化：基于实时信道状态与业务负载，动态分配上行带宽、计算节点与模型执行点，实现跨通信与计算域的全局优化，在满足AI业务时延与精度要求的同时提升系统能效与谱效。

在标准化层面，需定义AI业务类型的特征标签体系、业务特征到QoS参数的动态映射机制、支持AI业务流的低时延传输协议栈配置，以及面向通算协同的物理层与资源调度接口，为AI服务提供无缝、可靠、差异化的连接能力。

2.3.4 内生智能计算

内生智能计算通过使无线接入网内生具备AI任务的本地执行能力，实现计算任务的灵活部署与全生命周期管理。该模式采用集中式调度与分布式算力执行相结合的方式，并依托增强的空口信令实现实时端网协同。相较于传统依赖云端处理的方式，内生模式具备低时延、高实时性与资源利用高效的优势，能够充分挖掘并协同调度RAN侧的闲置算力，实现通信与计算资源的深度融合。然而，该模式也面临标准化接口缺失、资源调度复杂以及安全可靠性保障等方面的挑战。

在标准化层面，需推动定义AI任务描述、计算卸载信令及算力状态交互协议，明确RAN侧集中控制与分布式执行的接口流程，并基于3GPP CAPIF等框架设计算力开放接口，以支持网络智能能力的开放与协同发展。

2.3.5 通算数智融合控制

该技术是打破资源孤岛、实现全局最优的“大脑”，通过实时感知连接状态、异构算力分布与AI任务需求，对通信、计算、数据与智能资源进行一体化协同调度与动态优化。该技术能够将端到端时延、推理精度等复杂业务体验需求转化为跨域资源调配策略，并借助强化学习等方法在多目标、多约束环境下实现动态决策，从而为扩展现实（XR）、自动驾驶、协作AI等业务提供确定性体验保障。在实际应用中，它体现出连接与计算资源的灵活置换能力，例如当无线传输受限时，可通过增强计算资源补偿时延；在推理精度不足时，可协同调优模型与数据质量以提升结果准确性。同时，该控制机制也支持分布式协作推理的优化，根据实时网络条件动态决策模型切分点与算力分配。

在标准化层面，需建立通算资源的统一信息模型，定义协同控制架构与接口，并实现从业务意图到资源策略的自动映射与翻译，从而系统性支撑6G时代融合业务的可靠、高效与智能化运营。

2.3.6 能力开放

为充分释放6G无线接入网作为边缘通算智融合平台的核心潜能，需将网络功能层所提供的连接、数据、模型与计算等内生能力进行统一抽象与封装，通过标准化、可编程的接口机制，实现网络能力的灵活开放。

1）多维能力开放：多维能力开放涵盖对内协同与对外开放两个维度。对内协同旨在实现跨域、跨网元的资源联动与策略协同。例如，无线接入网可向核心网开放其实时无线状态、边缘算力资源及模型能力等信息，以支撑端到端网络切片、算力调度等全局优化任务；对外开放则指无线接入网面向互联网业务提供商（OTT）应用开放其内部网络指标（如链路质量）与感知数据，以优化业务体验（如动态调整视频压缩率），并可开放异构计算资源支持边缘AI应用的就近部署。

2）统一开放架构：基于3GPP CAPIF构建无线AI能力开放平台，支持间接模式（经核心网网络开放功能开放，适用于广域通用服务）与直接模式（无线接入网直接向本地应用开放，适用于低时延场景）。无线接入网侧作为API提供者与发布者，实现服务的注册、发现、鉴权与计费，构建开放的边缘智能生态。

在标准化层面，需制定无线AI功能开放API框架标准，定义各类原子化服务（如数据服务、推理服务）的接口规范与数据模型，并建立完善的API安全认证与隐私保护机制，确保能力开放过程中的数据主权与网络安全。

2.4 管理编排层关键技术及标准化思考

当前5G网络管理与编排体系主要聚焦于通信切片及网络功能虚拟化（NFV）的生命周期管理，尚未实现对异构计算资源与通信资源的统一编排与联合优化。这一局限性导致其难以支撑泛在AI任务的端到端部署与性能保障，同时缺乏对网络融合服务能力的对外开放接口，从而制约了网络整体服务能力的释放。为此，如图7所示，编排服务层亟需引入面向无线网络中的通信、计算、数据及智能资源/功能/服务的联合编排技术，以突破上述瓶颈。

其中，无线通算数智资源/功能/服务的联合编排功能旨在提供全局、跨域的智能编排能力。具体而言，意图解析功能首先将业务需求进行转化，并对通信、计算、数据及智能要素进行统一建模。基于基础设施层与网络功能层提供的多维状态感知数据（如算力负载、无线链路质量、AI任务进度等），联合编排功能通过灵活调度网络功能（如接入网网管、云平台管理、数据/模型管理等），实现对通信、计算、数据及AI模型等多维资源的协同编排，进而支撑面向AI服务的编排与性能保障。典型应用场景包括：针对大容量跨站场景配置共享算力池与通信资源预留（如候选基站迁移策略）；为基站SLA保障业务识别功能提供算力配置（如40+vCPU）；为AI视觉检测服务提供通算一体化保障能力（如业务流量与AI检测精度需求等）。通过实现网络资源与智能任务的最优匹配，该机制可在复杂场景下为多样化融合业务提供确定性的服务质量保障。

无线能力开放功能则对网络内部的多维服务能力——包括连接、计算、数据及AI等——进行模块化封装，形成可被统一发现、订阅与调用的标准化服务，并确保其安全性，从而推动网络从封闭的传输管道向开放的能力平台演进。

在标准化层面，需重点增强以下3个方面：一是对多样化业务进行通信、计算与智能的统一建模，将AI业务的端到端SLA要求（如时延、精度、可靠性）转化为对底层多维资源的联合约束；二是标准化跨层智能编排流程与接口，增强基础设施层、网络功能层与编排服务层之间的交互流程，以实现多维资源的高效编排与优化；三是标准化统一的“无线AI服务开放API框架”，提供灵活的服务发现机制及标准化数据模型，从而提升网络资源利用效率，拓展无线网络超越传统连接范畴的多维服务能力。

3 结束语

尽管6G无线接入网通算智融合的技术蓝图与系统框架

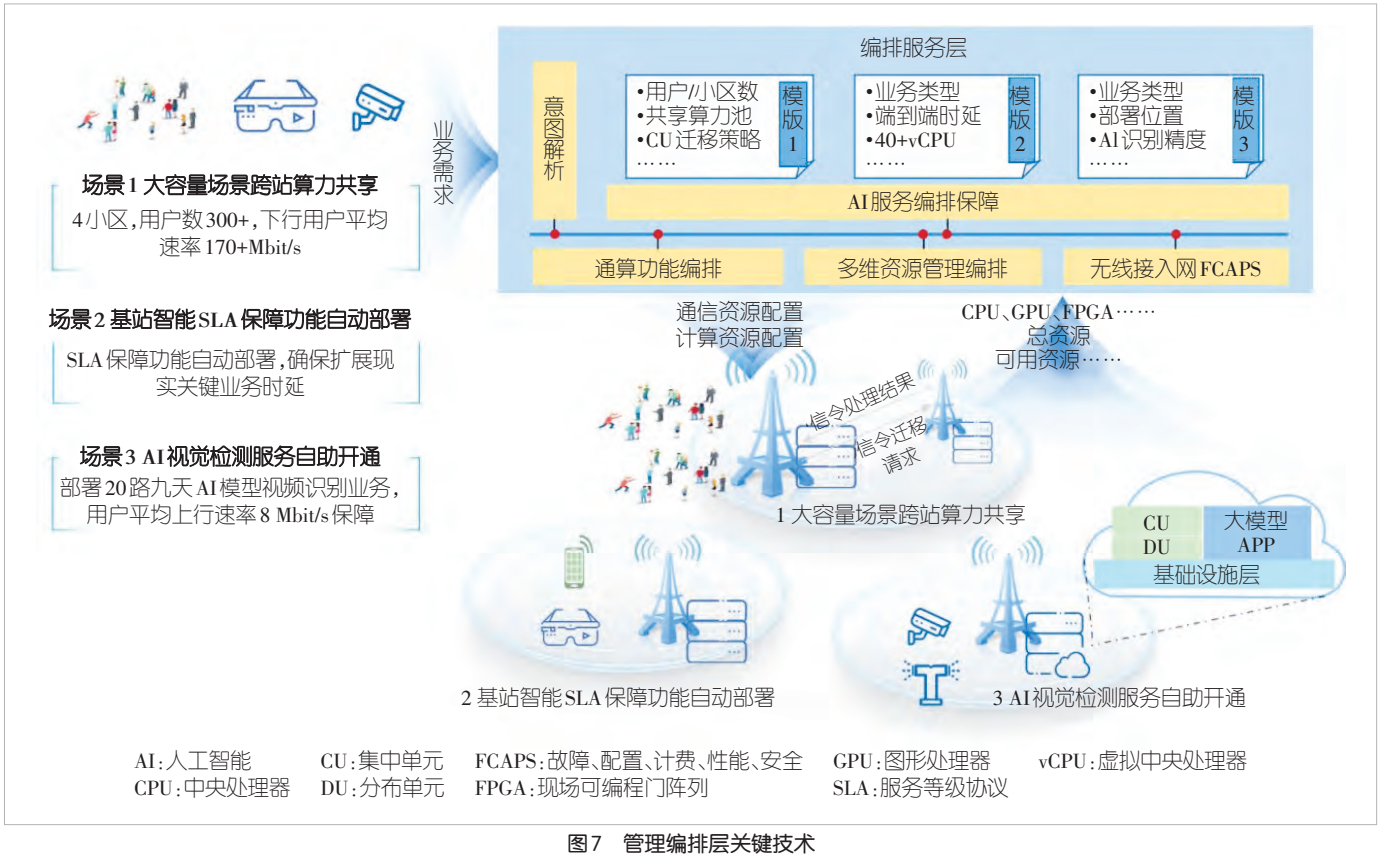


图7 管理编排层关键技术

已逐步清晰，但其从理论探索走向规模化商用仍面临一系列深层挑战。未来研究不仅需持续攻关关键技术，更需在标准制定、生态构建与商业模式创新等方面取得突破。

1）网络AI协同优化与效率平衡的挑战：当前AI模型研究多聚焦于提升网络的频谱效率、能量效率等性能指标，尚未充分考虑模型自身的质量（如泛化能力、推理实时性）与效率（如计算复杂度、资源开销）。模型在实际组网环境中的性能表现仍有待系统验证，这在一定程度上制约了其规模化商用进程。如何在能力提升、质量保障与效率优化之间实现均衡，是亟待解决的关键技术问题。此外，网络与终端之间的协同优化也展现出新的可能性。例如，通过智能计算卸载机制，可根据终端能力、网络负载与业务需求，动态分配AI任务至终端、边缘或云端，实现模型分割、协同推理与资源高效利用。这种跨层协同不仅可缓解网络侧算力压力，还可借助终端感知数据提升模型实时性与个性化服务水平，为6G通算智融合开辟更灵活的部署路径与服务模式。

2）系统复杂性与可信安全的挑战：AI的深度引入使网络演进为高度复杂的自适应系统，对其决策的可解释性与行为的确定性提出了更高要求。与此同时，开放的数据与服务接口也引入了新的隐私泄露与安全攻击风险。构建“可信赖的智能网络”亟需设计全新的内生安全框架与全生命周期保障机制。

3）标准化与产业协同的挑战：通算智融合涉及跨层、跨域的功能重构，标准制定的复杂度空前提升。当前，如何在3GPP、ITU、O-RAN等国际标准组织之间形成高效协同，并明确核心功能的标准化边界与演进路线图，是加速产业成熟的关键所在。

4）商业闭环与生态构建的挑战：由“网络使能AI”催生的新型商业模式，其价值链条、计费机制与生态合作模式尚需市场验证。当前面临的核心问题在于如何孵化杀手级应用，并实现从技术能力到商业价值的完整闭环。为此，建议运营商、设备商与AI应用开发者面向典型场景开展联合试点，共同探索并验证可持续、多方共赢的商业模式。

参考文献

[1] Letaief K B, Shi Y M, Lu J M, et al. Edge artificial intelligence for 6G: vision, enabling technologies, and applications [J]. IEEE journal on selected areas in communications, 2022, 40(1): 5-36. DOI: 10.1109/JSAC.2021.3126076

[2] Li N, Sun Q, Li X, et al. Towards the deep convergence of communication and computing in RAN: Scenarios, architecture, key technologies, challenges and future trends [J]. China communications, 2023, 20(3): 218-235. DOI: 10.23919/JCC.2023.03.016

[3] Huang Y H, Li N, Sun Q, et al. Communication and computing integrated RAN: a new paradigm shift for mobile network [J]. IEEE network, 2024, 38(2): 97-112. DOI: 10.1109/mnet.2024.3355401

[4] Li N, Wang Y, Sun Q, et al. Rethinking RAN architecture for deep fusion of AI and communication in 6G [J]. IEEE wireless communications, 2025, 32(3): 164-174. DOI: 10.1109/MWC.009.2400168

[5] CCSA TC5WG6 无线算力网络场景、需求和关键技术研究 [S]. 2024

[6] 3GPP TS38.323 NR; Packet Data Convergence Protocol (PDCP) specification (R19) [S]. 2024

[7] 3GPP TR22.876 Study on AI/ML model transfer phase2 (R19) [S]. 2023

[8] 3GPP TR38.843 Study on artificial intelligence (AI)/machine learning (ML) for NR air interface (R18) [S]. 2023

[9] IMT-2030（6G）推进组. 6G无线系统架构和功能研究 [R]. 2024

[10] 3GPP TS 23.222 Common API framework for 3GPP northbound APIs [S]. 2025

[11] Kundu L, Lin X Q, Gadiyar R, et al. AI-RAN: transforming RAN with AI-driven computing infrastructure [PP/OL]. arXiv[2026-01-05]. https://arxiv.org/abs/2501.09007

[12] Huang Y H, Xie Y X, Chen Z Q, et al. AI empowered modeling, closed-loop optimization, and field trials of RAN digital twin [J]. IEEE network, 2026, 40(1): 154-164. DOI: 10.1109/mnet.2025.3565716

作　者　简　介

解宇瑄，中国移动研究院无线与终端技术研究所研究员；主要研究方向为无线通算智融合网络关键技术和协议设计。

李响，中国移动研究院无线与终端技术研究所研究员，高级工程师；主要研究方向为无线通算智融合网络架构、关键技术和协议设计。

李婷，中国移动研究院无线与终端技术研究所研究员；主要研究方向为无线接入网基础设施、通算智融合的管理与编排相关技术。

孙奇，中国移动研究院无线与终端技术研究所主任研究员，正高级工程师；主要研究方向为5G/6G无线接入网智能化、云化等关键技术。

6G通感一体化关键技术和标准发展

6G Integrated Sensing and Communication: Key Technologies and Standardization Development

向际鹰/Xiang Jiying[1,2]，蒋创新/Jiang Chuangxin[1,2]，
高音/Gao Yin[1,2]，许进/Xu Jin[1,2]，刘峻琛/Liu Junchen[1,2]
（1. 中兴通讯股份有限公司，中国 深圳 518057；
2. 移动网络和移动多媒体技术国家重点实验室，中国 深圳 518055）
（1. ZTE Corporation, Shenzhen 518057, China；
2. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China）

DOI:10.12142/ZTETJ.202601004
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260225.1717.013
网络出版日期：2026-02-25
收稿日期：2025-12-16

摘要：系统研究了6G通感一体化（ISAC）的关键技术与标准化进展，重点围绕通感用例场景、信道建模、网络架构与协议流程、物理层感知信号设计、感知和通信及定位融合等展开讨论，分析了包括无人机感知在内的感知用例的社会价值与技术需求。在信道模型方面，介绍了基于3GPP R19框架的感知目标信道与背景信道联合建模方法，并结合雷达散射截面（RCS）动态特性与环境散射特性，描述了兼顾通信与感知的统一信道模型。进一步地，探讨了感知网络架构、流程与物理层安全，对比了6种感知模式的适用性，并给出了感知参考信号的方案。最后，提出了感知辅助通信与定位、AI融合的技术路径，通过跨网元信息交互实现资源调度优化与感知精度提升。研究成果为6G通感一体化的标准化与产业化落地提供了理论支撑与实践参考。

关键词：6G；通感一体化；信道模型；感知网络架构；参考信号

Abstract: This paper systematically investigates the key technologies and standardization progress of integrated sensing and communication (ISAC) in 6G, with a focus on sensing-use case scenarios, channel modeling, network architecture and protocol procedures, physical-layer sensing signal design, as well as the integration of sensing with communication and positioning. The social value and technical requirements of sensing use cases, including unmanned aerial vehicle sensing, are analyzed. In terms of channel modeling, based on the 3GPP Release 19 framework, a joint modeling approach for target and background channels supporting sensing functionality is introduced. By incorporating radar cross section (RCS) dynamic characteristics and environmental scattering properties, a unified channel model that accommodates both communication and sensing is described. Furthermore, this paper explores the sensing network architecture and procedures, compares the applicability of six sensing schemes, and proposes sensing reference signal designs. Finally, a technical pathway for sensing-assisted communication and positioning fusion is proposed, which enhances resource scheduling optimization and sensing accuracy through cross-network-element information interaction. The research findings provide theoretical support and practical guidance for the standardization and industrial implementation of 6G ISAC.

Keywords: 6G; integrated sensing and communication; channel model; sensing-aware network architecture; reference signal

引用格式：向际鹰，蒋创新，高音，等．6G通感一体化关键技术和标准发展 [J]. 中兴通讯技术，2026, 32(1): 13-23. DOI: 10.12142/ZTETJ.202601004

Citation: Xiang J Y, Jiang C X, Gao Y, et al. 6G integrated sensing and communication: key technologies and standardization development [J]. ZTE technology journal, 2026, 32(1): 13-23. DOI: 10.12142/ZTETJ.202601004

随着6G网络向智能化、泛在化与绿色化方向演进，通感一体化（ISAC）作为突破传统通信与雷达系统分离架构的核心技术，已成为全球学术界与工业界的研究热点。该技术通过共享硬件资源与频谱资源，实现通信与感知功能（SF）的深度融合，在提升整体频谱效率的同时，为低空经济、智能交通、工业互联网、智慧城市等应用场景提供高精度环境感知能力。当前，研究机构与企业正从理论创新、系统架构设计及标准化推进等多维度展开攻关。

在国际领域，欧盟“数字罗盘”计划与美国联邦通信委员会（FCC）均将ISAC列为6G关键技术优先级项目，ISAC

也成为国际电信联盟为6G制定的六大应用场景之一[1]。例如，欧盟资助的INTEGRATE项目专注于联合通信和传感网络的理论、算法和架构基础。与此同时，电气电子工程师学会（IEEE）与第3代合作伙伴计划（3GPP）等标准组织已启动相关标准规范制定工作，国际电信联盟也开始研究如何对ISAC场景和技术进行量化评估[2]。

早在2021年，中国IMT-2030（6G）推进组就成立了ISAC子组，推动ISAC技术的研究，陆续发布了多项研究成果[3-4]。并且在同一年，IMT-2030（6G）推进组就开展了ISAC测试工作，实施目标检测、环境重构、微形变、呼吸检测等场景的测试，极大地推动了产业上下游发展。目前相关技术已经在5G-A产品中得到初步应用。

本文聚焦6G通感一体的核心技术挑战与标准化进展，从场景需求、信道建模、网络架构与协议流程、物理层感知设计，以及感知通信定位融合等展开系统性研究。首先，基于3GPP SA1与无线接入网（RAN）侧发展，识别并定义覆盖空域、水域、地面与城市设施等高优先级通感用例；其次，介绍3GPP在R19所确定的融合通信与感知特性的信道模型框架，通过目标雷达散射截面（RCS）动态建模与背景环境分层表征，解决多径干扰与场景适配问题；进一步地，提出RAN侧感知网络架构，结合6类感知模式与3种感知计算方法（用户设备（UE）、基站、SF侧），构建低时延、高鲁棒性的感知流程，并给出对于感知物理层安全方案的初步探索；另外，设计物理层感知信号以达到远距离覆盖要求；最后，探索感知与定位、通信、AI的协同机制，通过跨网元信息交互实现通信谱效与感知性能的提升。本文旨在为6G ISAC的标准化推进与产业应用提供理论依据与技术路线图。

1 通感技术用例场景与信道模型

1.1 通感用例

通感一体已被纳入IMT-2030的核心使用场景之一，将在6G阶段从5G-A的既有探索走向规模化与产业化落地。结合3GPP SA1已达成的阶段性成果、RAN侧研究关注点与学术界的研究热点[5-7]，本节列举几类优先级高、具备明确社会与产业价值的用例，主要包括无人机、车辆、行人、自动导引车（AGV）以及船只等目标的检测与追踪，基础设施的坍塌监测与健康监测，环境重构以及数字孪生[8]。

1）无人机感知：在低空经济与公共安全双重背景下，面向城市上空、近郊与厂区园区的低空无人机目标感知成为刚需。6G网络侧感知在不依赖额外专网雷达的前提下，按需覆盖重点空域，识别并跟踪无人机的存在、方位与运动趋势。典型使用场景包括：对禁飞或限飞区域的越界发现与动态告警；在物流与巡检业务中，为监管部门与运营平台提供基于网络的航迹连续性保障；在密集建筑遮挡的城市峡谷环境中，辅助实现冲突探测与空域调度。与传统单点雷达相比，网络内生感知具备连续覆盖、易部署和可与通信业务协同的优势，有利于提升监测的可靠性与成本效益，并降低对终端侧专用传感器的依赖。

2）车辆和行人检测/AGV检测：类似无人机检测，对行人、车辆以及室内工厂AGV进行感知也有助于提高公共安全。

3）船舶检测：面向近海岸和港区水域，对大小型船舶进行检测、分类与轨迹更新，是智慧港口、海事监管与海上救援的重要能力。在能见度受限（雾、雨）或海况复杂（风浪、强反射杂波）的条件下，6G网络基于通感一体可提供全天候的目标发现与状态更新，支撑航道安全疏导、违章闯入告警、应急搜救定位与环境监测。在港航物流侧，实时掌握船舶靠离泊动态与泊位周边态势，有助于提升调度效率与吞吐能力；在公共安全侧，对无标识或异常航迹目标的早期识别与预警，有助于风险前置与应急联动，带来显著社会效益。

4）基础设施坍塌监测（地陷/滑坡/沉降等）：针对公路、乡村道路、边坡、堤坝与农田等分布广且维护难的区域，6G通感能够对突发性塌陷、滑坡、地表异常位移等情况进行快速发现与定位。使用场景包括：高速与国省干线的路基安全监测与临时交通管制支撑；农田与水利设施的灾害早告警；地震次生灾害的巡检辅助等。当异常形变或坍塌迹象被检测到时，网络可快速向管理单位回传位置与时间信息，并联动现有的告警体系实现分级处置。相较传统依赖人工或稀疏传感器的方案，网络化感知具备“面+点”结合的覆盖优势，能够以更低的边际成本提升广域监测能力与响应时效。

5）环境重构以及数字孪生：通过融合网络全域的感知数据（如终端与物体的高精度位置、速度、反射特性以及周围环境状态信息），构建并持续更新一个数字化的虚拟环境镜像。这一孪生环境不仅精确还原了物理世界的几何布局、障碍物分布及移动态势，还能预测视距（LoS）与非视距（NLoS）链路的动态变化。基于这一统一的环境上下文，网络能够实现前瞻性的资源调度与智能控制，例如：在切换过程中，依据预测的UE轨迹和环境遮挡状态，提前为目标基站配置最优波束集，大幅减少测量开销和切换时延；在链路阻塞前，主动激活反射路径或调整波束方向，保持连接可靠性。数字孪生体由此成为连接感知与通信的智能中枢，实现了从被动响应到主动优化的跨越，显著提升系统在高速移动和复杂环境下的频谱效率、能源效率及用户体验。

此外，人体健康检测、姿势识别等其他用例也得到了广泛关注，需要利用微多普勒特性提前感知目标细微运动信息。

上述用例覆盖“空—海—地—城”多域的关键场景，共同体现了6G通感一体“以网络为平台”的价值主线：通过连续覆盖与可运营的感知能力，提升公共安全与产业运行效率，并为多源信息融合与智能化运维奠定基础。上述场景在3GPP SA1 TR 22.870与RAN侧研究中均已获得关注或采纳，具备标准推进与产业落地的共同基础。

1.2 通感信道模型

感知信道建模是通信与感知技术的核心研究内容，其目标是建立既能满足通信需求，又能支持SF的信道模型框架。这一模型旨在实现通信与感知功能的深度融合，使得网络在传输数据的同时，能够通过无线信号实现对环境和目标的感知。3GPP在R19中提出了感知信道模型的设计框架，基于传统的通信信道模型进行扩展，增加了目标建模和环境建模的部分，力求全面覆盖典型感知场景中的信道特性[9]。

传统的通信信道模型主要用于描述信号在发射机和接收机之间的传播特性，而SF则需要关注信号与环境中的目标或障碍物的交互过程，包括反射、散射和衍射等现象。因此，感知信道模型被设计为通信信道的扩展版本。它由目标信道与背景信道共同构建，整体表达形式为：

$$\mathrm{H_{ISAC}} = \mathrm{H_{target}} + \mathrm{H_{background}} \tag{1}$$

其中，$\mathrm{H_{ISAC}}$表示通感信道，$\mathrm{H_{target}}$表示目标信道，包括所有被感知目标影响的传播路径；$\mathrm{H_{background}}$表示背景信道，包括除了目标信道外的其他传播路径。

1.2.1 目标模型

目标信道建模的目的是描述感知信号在目标物体上的反射路径。每个感知目标都可以被看作一个或多个散射点。信号通过第一链路从发射机传播到目标，再通过第二链路从目标传播到接收机。为了精确描述这种传播过程，R19阶段深入研究了典型目标（无人机、车辆、人体、AGV）的RCS模型。

RCS模型与视角、目标尺寸、目标材质等多方面因素相关，可以将其建模成3个部分：

$$\sigma_{\mathrm{RCS}} = \sigma_M \sigma_D \sigma_S \tag{2}$$

其中，σ_M是大尺度的RCS建模，对于每个散射点是一个确定性的值，$\sigma_D\sigma_S$是小尺度的RCS建模，σ_D可以是1，或者是和角度相关的值，σ_S服从对数正态分布，均值与方差分别为$\mu_{\sigma_{S_\mathrm{dB}}}$和$\sigma_{\sigma_{S_\mathrm{dB}}}$，满足公式（3）：

$$\mu_{\sigma_{S\_\mathrm{dB}}} = \frac{-\ln(10)}{20}\sigma^2_{\sigma_{S\_\mathrm{dB}}} \tag{3}$$

RCS建模方式会随着目标建模的散射点不同而有所不同，通常有两种建模方法：第一种是将目标建模成单散射点，并且σ_D的值固定为1和角度无关，如小尺寸的无人机或人；第二种是将RCS建模为单个或者多个散射点，并且每个散射点的σ_D与角度相关，如车辆、AGV等。

• 对于第一种RCS建模方式，σ_D在入射角与散射角相同的情况下取值固定为1，σ_M是自发自收RCS的线性值的均值，$10\lg(\sigma_M\sigma_D)$的取值可以表示为$\sigma_{\mathrm{MD_dB}}(\theta_i,\phi_i,\theta_s,\phi_s)$。$\sigma_{\mathrm{MD_dB}}$和确定性的入射角度$(\theta_i,\phi_i,)$以及散射角度$(\theta_s,\phi_s)$相关：

$$\sigma_{\mathrm{MD\_dB}}(\theta_i,\phi_i,\theta_s,\phi_s) = \max\left(10\lg(\sigma_M) - 3\sin\left(\frac{\beta}{2}\right), \sigma_{\mathrm{FS}}(\theta_i,\phi_i,\theta_s,\phi_s)\right) \tag{4}$$

其中，$\beta \in [0^\circ, 180^\circ]$，是入射角和反射角平面内的双站角，$\sigma_{\mathrm{FS}}(\theta_i,\phi_i,\theta_s,\phi_s)$表征前向散射的影响。

• 对于第二种RCS建模方式，σ_D的取值依赖于入射角与散射角。这种建模方式下，$10\lg(\sigma_M\sigma_D)$，记为$\sigma_{\mathrm{MD_dB}}(\theta_i,\phi_i,\theta_s,\phi_s)$，是由入射角$(\theta_i,\phi_i,)$以及散射角$(\theta_s,\phi_s)$确定的：

$$\sigma_{\mathrm{MD\_dB}}(\theta_i,\phi_i,\theta_s,\phi_s) = \max\left(G_{\max} - \min\left\{-\left(\sigma^V{}_{\mathrm{dB}}(\theta) + \sigma^H{}_{\mathrm{dB}}(\phi)\right), \sigma_{\max}\right\} - k_1\sin\left(\frac{k_2\beta}{2}\right) + 5\log_{10}\left(\cos\left(\frac{\beta}{2}\right)\right), G_{\max} - \sigma_{\max}, \sigma_{\mathrm{FS}}(\theta_i,\phi_i,\theta_s,\phi_s)\right) \tag{5}$$

其中，(θ,ϕ)是入射和反射径的等分线的天顶角与水平角，$\sigma^V{}_{\mathrm{dB}}(\theta)$和$\sigma^H{}_{\mathrm{dB}}(\phi)$分别定义为：

$$\sigma^V{}_{\mathrm{dB}}(\theta) = -\min\left\{12\left(\frac{\theta - \theta_{\mathrm{center}}}{\theta_{\mathrm{3dB}}}\right)^2, \sigma_{\max}\right\} \tag{6}$$

$$\sigma^H{}_{\mathrm{dB}}(\phi) = -\min\left\{12\left(\frac{\phi - \phi_{\mathrm{center}}}{\phi_{\mathrm{3dB}}}\right)^2, \sigma_{\max}\right\} \tag{7}$$

1.2.2 背景模型

背景信道的设计遵循“在既有通信场景信道之上，提取不受感知目标影响的多径成分”的原则。它的作用是在“目标通道”之外，提供环境本底的散射与传播贡献，使得最终

的通感信道等于各目标通道之和再叠加一个背景通道。

对收发端同址的传输接收点（TRP）/UE单基地感知模式，将接收机虚拟成参考点（RP），它们与同址点之间的二维距离和高度来自Gamma分布，共3个参考点，且方位分布保持相隔120°，以覆盖环境的主方位；随后分别以“Tx↔RP”为链路，按通信场景模型独立生成3条NLOS子信道，并施加各自的路径损耗与阴影衰落，最后把3条子信道线性叠加为总的背景通道。这样做可以用3个方位稳定的代表性路径来“采样”环境散射，既不过度依赖某一偶然簇，又能覆盖典型方向性的稳健背景。在单基地背景构造中，参考点的高度与距离分布按场景（UMi/UMa/RMa/InH/InF及空地无人机扩展）给定了不同的Gamma参数，使得背景时延、角度与功率分布随频段与场景变化。如图1所示，虚拟参考点的方法可以沿用既有通信信道模型产生过程，拟合结果和射线追踪（RT）的效果非常匹配[10-12]。

对于双基地中任意分离的TRP－UE、UE－UE、TRP－TRP等模式，背景信道不使用参考点拼接，而是直接按目标通信场景的标准快衰落流程生成整条通道。同样地，绝对时延也直接复用对应场景的绝对时延模型（例如RMa/UMa/UMi/InH/InF的表格参数）。

综上所述，背景信道模型和目标模型的本质区别在于背景信道是静态的，而目标信道建模时会在小尺度信道的建模中复用TR 38.901信道模型中多普勒项的建模。故在目前5G-A通感系统级仿真中，业界普遍采用了不同形式的动目标显示（MTI）算法来抑制背景信道的影响，并使用二维恒虚警（CFAR）算法识别非零频多普勒附近的感知目标。该流程覆盖了目前的5G-A用例中的全部的无人机、AGV、车辆、行人的动态目标感知场景。对于6G中新涌现的其他感知用例，例如静态目标的路面塌方检测等，则尚未被目前TR 38.901中的信道模型所覆盖。未来如果需要对6G更多的感知用例继续仿真评估，3GPP信道模型可以进一步增强。

2 感知网络架构与协议流程

2.1 感知模式与方法

通感一体化的感知模式分为6种，如图2所示。在模式A（即基站单基地模式）中，其工作方式与雷达非常相似，单一节点既发射感知信号，也通过感知目标接收反射信号。在模式B（即基站双基地模式）中，RAN基站A发射感知信

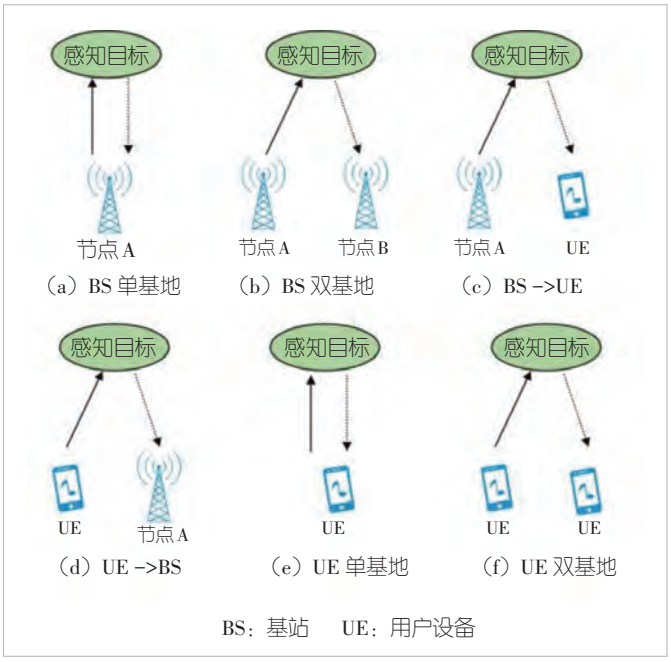
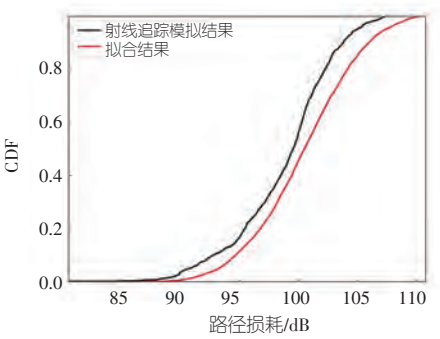


图2　6种感知模式



CDF：累积分布函数　gNB：5G基站　NLoS：非视距

图1　基于参考点的单基地感知模式下的背景信道

号，RAN基站B接收反射信号。在模式C中，其工作方式与当前5G下行链路测量相同，RAN节点发射感知信号，UE接收反射信号。在模式D中，其工作方式与当前5G上行链路测量相同，UE发射感知信号，RAN节点接收反射信号。在模式E和F中，它们分别与模式A和模式B的工作方式类似，区别在于将基站（BS）替换为UE。

对于感知方法，其可行性取决于UE、BS和SF的计算资源、软硬件能力。有3种通用感知方法可以使用：

感知方法1：基于UE的感知，即UE执行感知结果计算。例如，UE可接收感知参考信号，并据此识别不同的人体手势；随后，若需上报，UE可将感知结果发送至SF。

感知方法2：基于BS的感知，即BS执行感知结果计算。例如，BS可接收感知参考信号并检测是否有无人机入侵，若有则计算其位置；随后，BS可将无人机位置上报至SF。

感知方法3：基于SF的感知，即SF执行感知结果计算。例如，BS或UE仅向SF上报中间测量元素（如时间、角度等），随后，SF收集所有结果并进一步计算最终感知结果。

6G系统可根据计算资源分布，以及对应的感知场景需求，支持这3种方法。例如，在无人机入侵检测场景中，若某基站（可能连接多个传输节点）具备计算无人机位置或轨迹，并准确识别无人机的能力，则可快速完成测量与计算，并将结果上报SF，以降低时延并减轻SF负担。

2.2 RAN架构

对于ISAC网络结构设计，从RAN侧来看，如图3所示，未来标准化的架构单元包括SF、RAN节点（主要指基站）以及终端用户（UE）。进一步地，SF还可以将控制面与用户面分离。这种设计非常典型，特别是针对由核心网触发感知服务的使用场景。

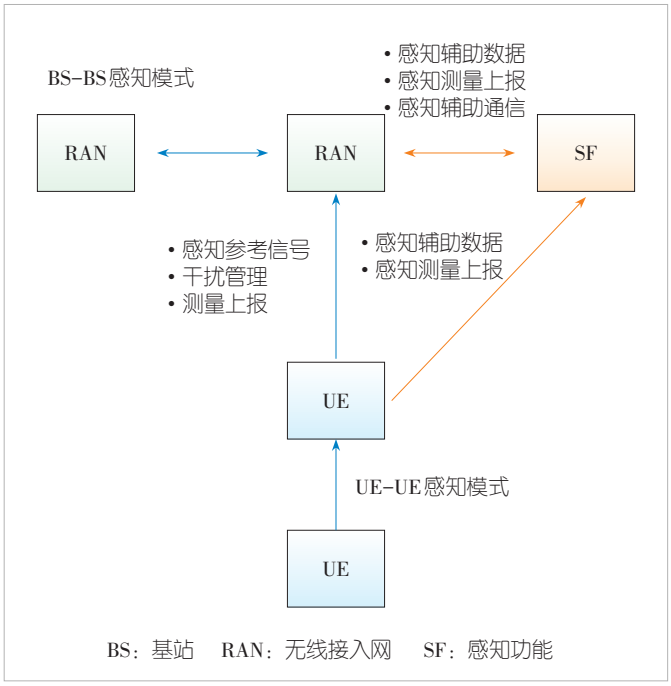
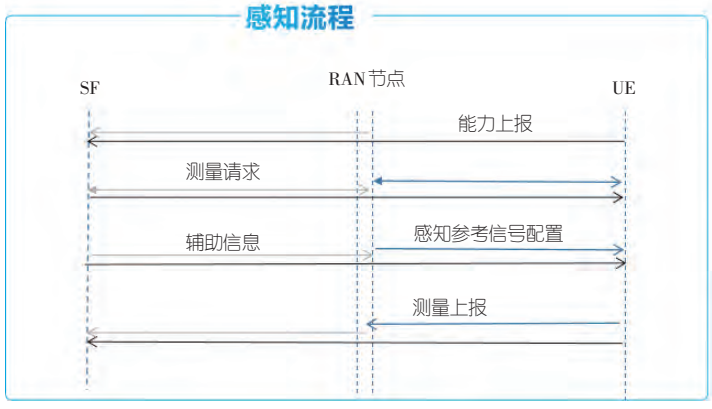


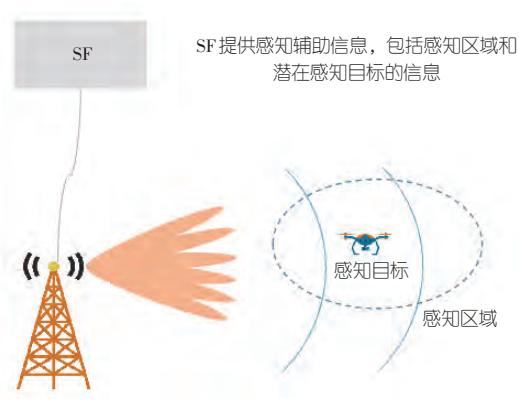
图3　基于参考点的单基地感知模式下的背景信道

2.3 感知流程

主要的感知流程包括：首先基站或者UE向SF上报其配置或能力信息，包括支持哪些感知模式、感知信号接收能力等参数；接着SF或者UE或者基站发起感知测量请求；然后SF向基站或者UE提供感知辅助信息，比如感知区域（如图4所示）以及潜在感知目标的特征信息；对于有UE参与



（a）感知流程与信令



（b）感知区域相关信令

RAN：无线接入网　SF：感知功能　UE：用户设备

图4　通感感知流程

的感知模式，随后基站需要向UE配置感知参考信号；最后测量节点需要将测量结果上报，比如UE上报测量结果给基站或者SF，或者基站上报测量结果给SF。

总体而言，根据参与感知的不同网络元素，感知信令交互标准化至少分为3类：SF与BS信令交互、SF与UE信令交互、BS与UE信令交互。综上所述，为支持基站单基地感知模式、BS双基地感知模式、BS发UE收感知模式、UE发BS收等感知模式，表1所示信息交互流程是6G感知模式研究的重点。

2.4 感知物理层安全

通信系统主要是“把数据送到目标接收者”，而ISAC的感知部分会主动发射可被环境反射/散射的波形，并基于回波估计距离、角度、速度、姿态、微多普勒等。但感知反射信号回波不只来自“授权目标”，也来自非目标用户与其携带物、人体、车辆等。这使得ISAC的隐私面更接近“传感器网络 + 位置服务 + 侧信道”，而不仅是传统无线链路保密。这就需要从物理层设计上，将隐私安全作为设计目标之一。

首先，感知测量结果的上报与暴露应当限制在授权的网元中，且某个网元或第三方节点请求感知测量结果时，网络设计应当仅传递其需求且必要的感知结果，避免与请求的感知目标不相关的感知结果或涉及用户隐私的感知结果的暴露。除此之外，物理层发送感知参考信号时，应当在满足感知/通信关键绩效指标（KPI）的前提下，确保感知参考信号的发送与测量限制在请求的感知区域中，而对非感知区域施加功率上限、波束方向等参数上的约束，对禁止感知区域控制感知参考信号的发送。对于第三方被动接收的情况，应该增加非授权第三方接收并解调感知参考信号的难度，例如给参考信号额外附加循环移位、时域附加伪随机相位等方案。如图5所示，以感知参考信号在时域附加伪随机系数为例，虽然感知参考信号的序列本身可能会由于感知对于低峰均功率比（PAPR）、大覆盖的需求而数量受限，导致其容易被监听，但是时域上的随机系数可以使得攻击者难以关联时域上的多次观测而提升感知结果的置信度，无法在多普勒域进行处理，也无法利用动目标显示算法分离感知目标和背景杂波。这样，非授权的第三方接收机无法从接收到的感知参考信号中有效区分感知目标并获得准确的位置、速度估计。

3 感知信号设计

在NR中，研究者们为多种用途引入了各种参考信号（RS），例如：信道状态信息参考信号（CSI-RS）用于信道状态信息测量和波束管理，跟踪用的CSI-RS（也称追踪参考信号（TRS））用于获取时间和频率跟踪，相位跟踪参考信号（PTRS）用于相位噪声和多普勒效应估计，定位参考信号（PRS）用于定位测量，探测参考信号（SRS）用于上行信道测量。这些参考信号主要是基站和UE之间的发送测量。如果能沿用到6G，这些参考信号仍然可以用于感知。

然而，在单基地感知模式下，当BS1发送感知参考信号时，BS1可能同时接收该参考信号以进行感知测量。需要注意的是，基站自发自收模式下，我们发现发射功率的限制主要来自于接收，而并非发射端。基站由于需要考虑对UE覆盖的问题，通常发射功率可以做到很大，例如超过49 dBm。但是自发自收模式下如果收发同时，过高的发射功率会立刻由于有限的收发天线隔离度导致接收端天线的低噪声放大器（LNA）进入饱和区产生失真信号，而导致整个正交频分复用（OFDM）符号上的接收结果都产生谐波干扰。由此收发天线的隔离度对发射功率具有额外的限制，导致收发完全同

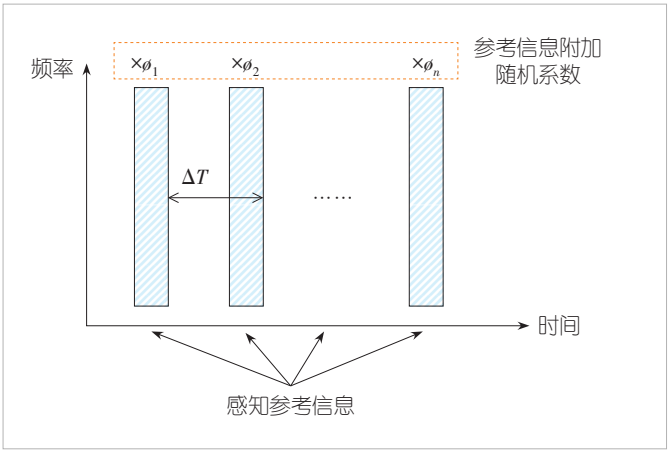


图5 感知参考信号在时域附加随机系数

表1 6G潜在的不同实体间的交互流程

BS与SF之间	SF与UE之间	BS与UE之间
RAN节点信息从BS传输至SF SF向BS发起测量请求并发送辅助信息 BS将测量结果传输至SF	UE能力信息从UE传输至SF SF向UE发起测量请求并发送辅助信息 UE将测量结果传输至SF	UE能力信息从UE传输至BS BS向UE下发RS配置 UE上报测量结果给BS

BS：基站 RAN：无线接入网 RS：参考信号 SF：感知功能 UE：用户设备

时的感知模式面临覆盖不足的问题。如图6所示，方案A1可能无法探测远离收发机的目标。这是因为单基地收发机需同时进行发射和接收操作，过高的发射功率会导致严重的自干扰并引发接收机功率饱和。例如，采用基站单基地感知模式进行无人机入侵检测时，最大发射功率可能不得超过30 dBm，以避免接收机饱和。因此，可考虑方案A2，即通过时分复用（TDM）方式分离发射（Tx）和接收（Rx）时段。

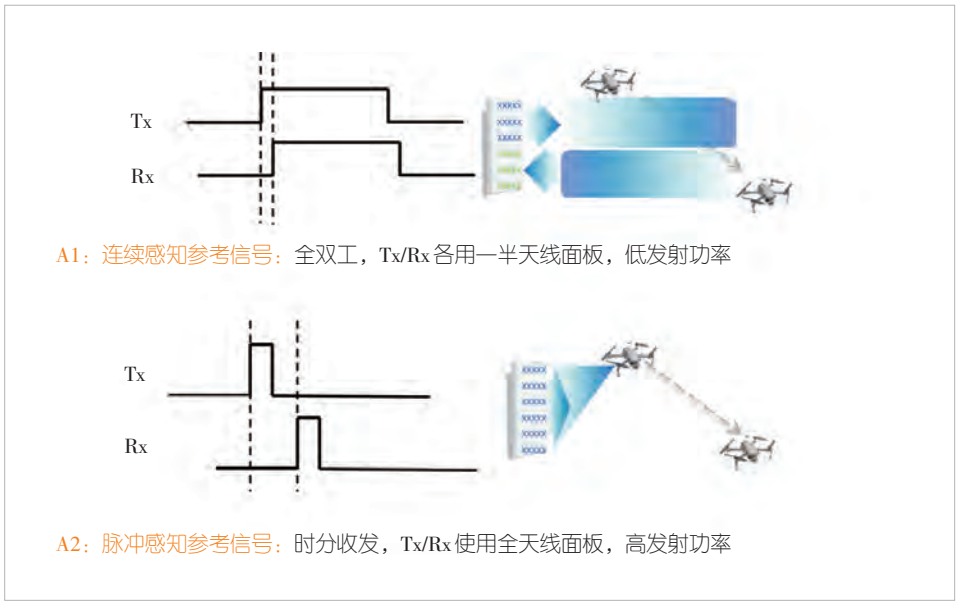


图6　基站单基地感知的不同发射和接收方案

例如，发射机可工作$T2=2\ \mu s$，当接收天线处于接收状态时，发射机可关闭。换句话说，当接收天线处于接收状态时，发射机关闭以避免发送信号泄露到接收机。为实现图6中所示的方案A2，可在时域生成感知参考信号，即在一个OFDM符号内生成长度较短的脉冲参考信号（脉冲参考信号可以是雷达界经典的线性调频信号（LFM）或者其他峰均比低的参考信号），例如一个OFDM符号包含4 096个采样点，脉冲参考信号只需有100个采样点，假设最大3 312个子载波对应100 MHz带宽。经过4 096点快速傅里叶逆变换（IFFT）运算后，时域感知参考信号的波形如图7所示。可以看出，该方法均可为感知检测提供良好的性能保障，同时可实现极低的峰均功率比。需要注意的是，虽然脉冲参考信号的收发模式和现网中基于CP-OFDM的连续感知参考信号不同，但是其整个收发周期依然限制在一个OFDM符号内，发射机可以在完成脉冲参考信号的收发后切换为通信信号的发送。而根据3GPP、ITU等标准组织的讨论进展，在5G-A乃至6G早期版本的系统实现中，感知功能仍将基于某些特定的参考信号进行测量。这部分用于感知的参考信号的时频资源由系统统一专门调配给感知业务，因此脉冲感知参考信号本身不会对通信系统产生影响。而脉冲感知参考信号除了进行感知测量之外，也可以用于诸如移动管理测量、波束测量、CSI测量、干扰测量等其他通信用途。图8的系统级仿真也证明了脉冲感知

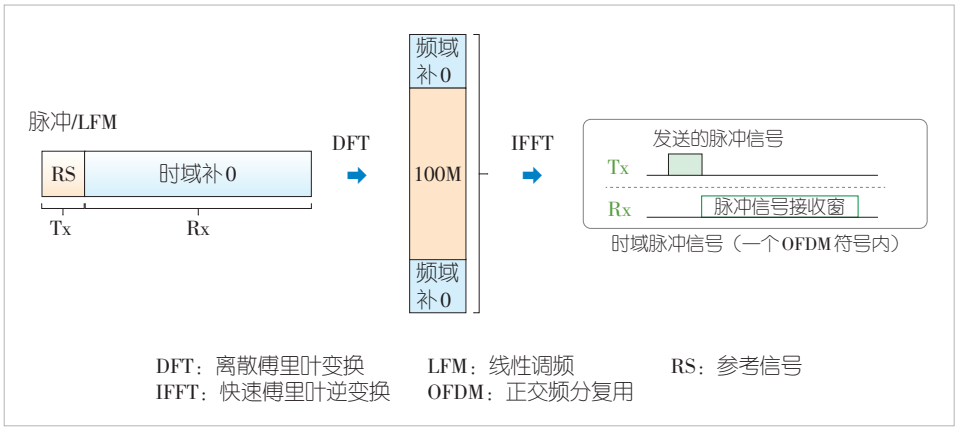
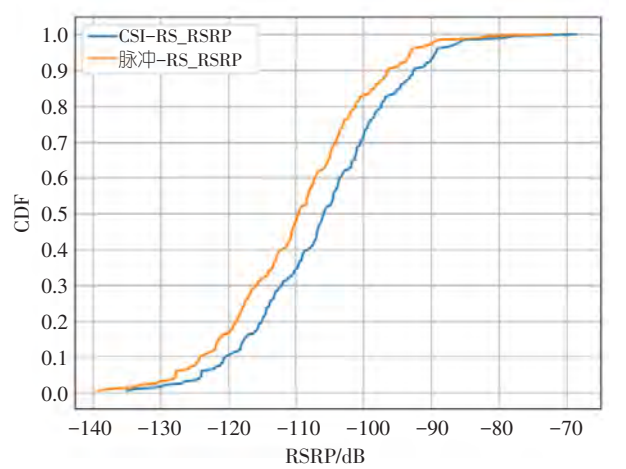


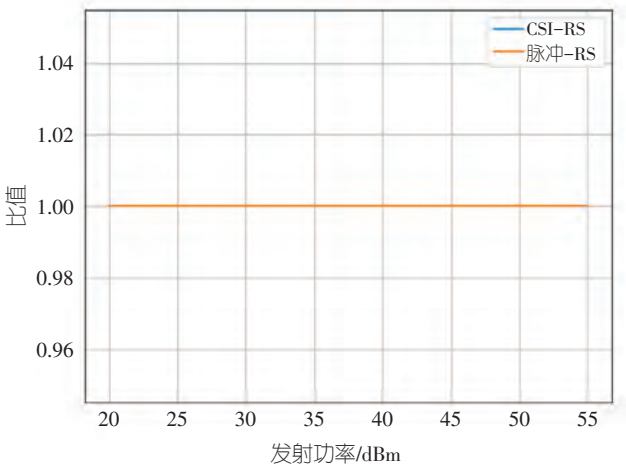
图7　基于DFT-S-OFDM生成的脉冲波

参考信号在用于通信测量用途的参考信号接收功率（RSRP）测量、小区选择正确率上也具备和传统基于CP-OFDM的连续参考信号CSI-RS近似的性能。因此，脉冲感知参考信号也可以与其他候选参考信号一样，作为6G通信系统的参考信号用于通信。从这两个角度说，引入脉冲感知参考信号并不导致通信性能的下降，其取决于网络给感知用途分配的时频资源，而非参考信号设计。

目前，基站单基地感知模式应用前景较好，但很多场景对覆盖距离提出了更高要求。如图9所示，受无人机飞行高度过高与基站天线下倾角的限制，BS0无法检测到其头顶上的无人机。此时更优的方案是让BS1负责该区域的感知任务。然而，由于无人机与BS1之间距离较远，从BS1发射的参考信号需要具备较大的覆盖范围。为实现这一目标，图8



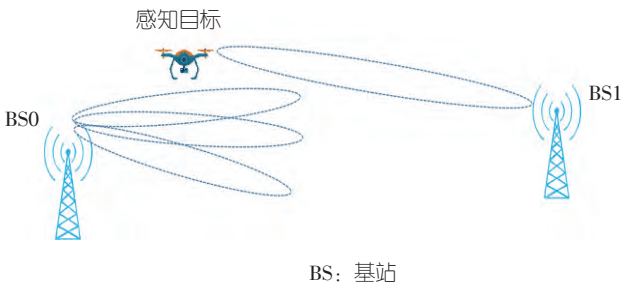
（a）RSRP测量CDF对比曲线



（b）小区选择正确率曲线

CDF：累积分布函数　　RS：参考信号
CSI：信道状态信息　　RSRP：参考信号接收功率

图8　脉冲感知参考信号和CSI-RS参考信号的通信测量性能对比

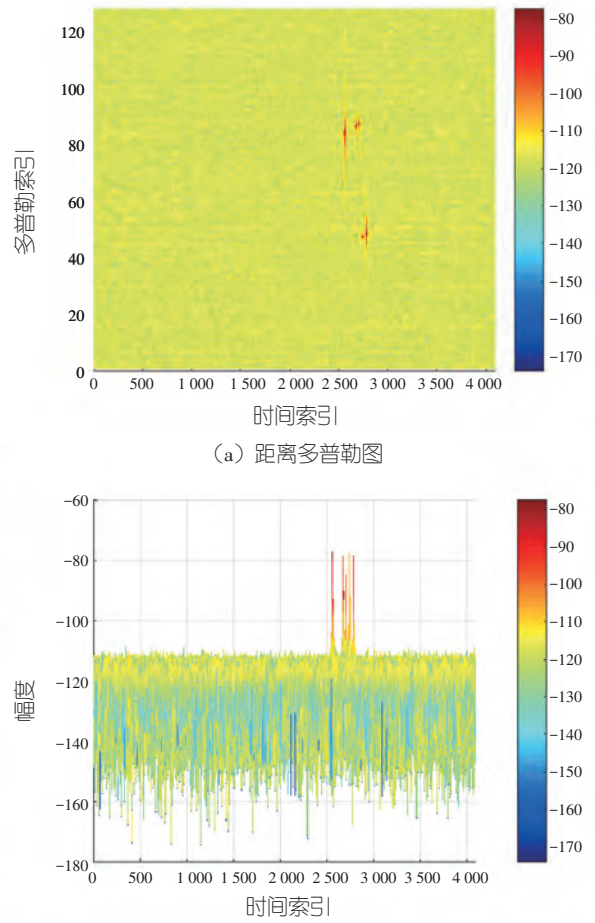


BS：基站

图9　感知参考信号需要更大覆盖范围

所示的脉冲参考信号就是一种很好的选择[13]。

图10和表2提供了脉冲参考信号的仿真结果。图10显示，即使目标很远，在距离-多普勒（RD）图上也能检测出目标（图中红色部分显示检测目标的功率大于背景噪声和杂波）。表2给出了系统级仿真结果，即使在RMa场景下，站间距为1 732 m时，感知精度也可以满足3GPP SA1提出的需求。



（a）距离多普勒图

（b）功率时延谱图

图10　感知目标离自发自收感知基站距离大于300 m的检测情况

在低空场景中，基于实际测量，单站最远感知距离可达1.5 km[14]。在海洋场景中，采用新型帧结构新技术可实现海域感知20 km的技术突破[15]。为了提高系统效率，如果能将单基站感知模式下的参考信号配置告知UE，则UE也可基于相同的参考信号执行感知测量，以提升参考信号资源利用率。换句话说，相同参考信号既用于基站自发自收模式也用于基站发UE收模式，且双方的测量结果均可用于计算最终感知结果，以提升感知精度。

此外，为了进一步提高感知精度，并且提高参考信号资源利用率，新的感知参考信号（例如脉冲波）可以和其他参考信号联合使用。文献[16]采用了级联多种参考信号共同用

表2 利用脉冲参考信号在RMa-无人机场景下的系统级仿真结果

距离精度@90%	水平角度精度@90%	垂直角度精度@90%	速度精度@90%	水平位置精度@90%	垂直位置精度@90%	检测概率	虚警率
1.890 m	0.524°	0.247°	0.467 m/s	6.478 m	3.024 m	96.9%	4.10%

于感知估计的方案。频域高密度且时域低密度的CSI-RS可以提供较高的时延分辨率，而时域高密度且频域低密度的PTRS可以实现较高的多普勒估计精度。两者结合用于感知估计，可以减少感知参考信号开销。但该设计在资源开销极低的情况下，需要采用复杂度相对较高的接收机检测算法，例如需要先在角度-多普勒域进行检测，再通过空时联合处理来获取距离信息[16]。

对于感知信号的设计，业界已有研究提出利用数据符号实现感知，以更高效地融合通信与感知。然而，由于数据符号在功率、带宽、序列等方面存在不确定性，感知性能很难得到保证。所以，感知参考信号在6G物理层技术中至关重要。除了感知参考信号的序列和图样设计外，6G系统还需进一步研究其功率控制、测量配置等关键问题。

4 通感定位融合

4.1 感知辅助通信

感知和通信相互辅助是通感一体的重要特征。从标准化层面看，感知辅助通信在业界的关注度更高。感知技术可实时监测障碍物、干扰源和移动设备的位置。这有助于通信系统快速调整以避免信号衰减和丢失，从而提高通信可靠性。另外，通过感知用户行为和环境条件，通信系统可动态调整发射功率等参数，减少不必要的能耗并延长电池寿命。这也是环境重构和数字孪生所带来的好处。

然而，SF与基站可能是不同的网络实体，即使感知结果完全由SF侧收集，通信资源实际也由BS分配。为了支持感知辅助通信，BS需要获取感知结果。有两种潜在方案可使BS能够获取感知信息以支持感知辅助通信。

方案1：为减少信令开销并实现感知辅助通信，基站向SF请求特定UE的感知结果以辅助资源调度。经SF认证授权后，感知结果可传递至基站，如图11所示。该方法简单且对RAN空口标准影响最小。然而，若SF位于核心网侧，基站与SF间的通信时延将显著增大。

方案2：如图12所示，为减少信令开销并在基站侧快速获得感知结果，可考虑以下流程：

步骤1：基站向SF或者UE请求与UE相关的感知测量结果。请求信息可包括UE坐标、周围环境感知信息等。

步骤2：SF或者UE执行认证和授权操作，判定是否可以将UE的感知结果提供给基站。

步骤3：SF或者UE响应基站，明确是否可以提供感知结果。

步骤4：UE直接向基站报告感知结果，实现感知辅助通信的实时性。

4.2 感知与定位融合

在基站和UE参与的感知模式中，UE的定位信息对于确定感知目标的位置至关重要；否则，基站难以选取感知目标对应的无线终端参与感知。从标准化层面看，感知和定位存在诸多共性，包括测量上报元素（坐标、定时、角度、RSRP、时间戳），以及参考信号设计、辅助数据请求、测量

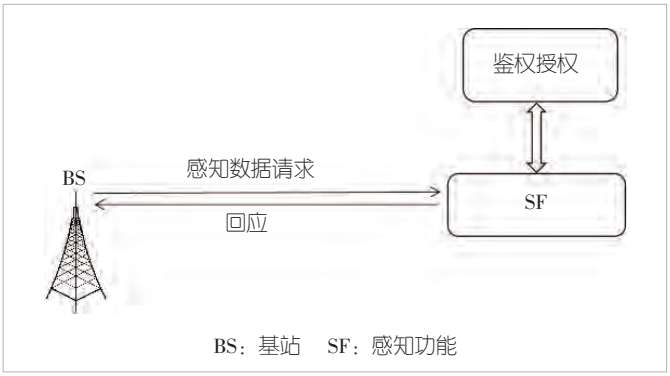


图11 基站从SF获取辅助通信的感知结果

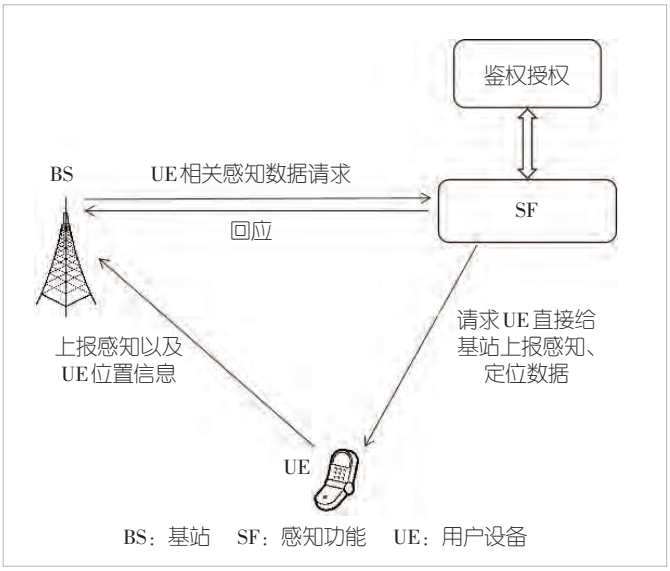


图12 UE在物理层上报辅助通信的感知结果

请求等流程。此外，定位管理功能网元（LMF）掌握用户终端的位置信息，可根据视距链路为SF推荐适合参与感知过程的用户终端。只有当用户终端位置足够精确时，其参与的感知服务才可能输出准确的感知结果。因此，SF与LMF间的信息交互至关重要，典型交互流程为SF向LMF发起请求，LMF向SF提供用户终端位置坐标、参考信号资源配置及测量结果等信息。进一步地，还可以考虑SF和LMF的融合设计，以降低信令传输开销。

5 通感与AI融合

在当前6G的研究讨论中，AI for ISAC和ISAC for AI都是重要的潜在研究方向，二者可并行演进，相互促进[17-18]。

AI for ISAC更强调使用机器学习来提升感知链路本身的鲁棒性与可部署性，例如：用自监督/对比学习来减少标注依赖，并使用具有高泛化性的网络设计，来应对感知广域场景中雨雾遮挡、非视距（NLoS）与复杂多径对感知结果的影响；利用图神经网络/Transformer进行多感知目标关联与跨站协同跟踪。

ISAC for AI侧重把通信与感知产生的多源数据、信道状态信息及环境表征能力转化为可用于智能体的“世界模型”与任务能力。例如，面向具身智能、车路协同、工业机器人数字孪生等场景，在边缘侧进行多模态融合处理（毫米波/太赫兹雷达感知、视觉、声学等）以形成可推理的场景图与占据栅格；在学习范式上，可探索面向任务的联合源信道编码与“感知–决策–通信”闭环强化学习，使网络能够将通感系统作为感知载体，基于通信与感知多模态数据实现任务的自适应编排与可靠性保障。

总体而言，6G的关键潜力在于把“空口可学习、网络可感知、智能可编排”作为统一目标，将通信能力转化为环境理解与任务执行能力，同时通过AI反向重塑感知与通感融合的性能边界。

6 结束语

本研究介绍了6G ISAC技术的核心创新方向与标准化路径，通过定义覆盖“空–海–地–城”的典型场景，明确了通感技术对公共安全与产业效率的赋能价值，基于3GPP信道模型扩展，阐述了动态RCS与背景环境联合建模方法，为跨场景感知性能优化提供了理论基础。感知网络架构与协议设计则明确了未来标准讨论路线。本文提出的物理层感知参考信号的设计方案可以满足远距离覆盖的需求，具备较高的商业部署价值。另外，通信、感知与定位的融合方案仍有待开展系统性研究；未来，还需进一步解决感知辅助通信跨网元时效交互等关键技术问题。本研究为6G ISAC从标准化探索走向规模化商用奠定了理论基础，有望加速构建智能感知生态。

参考文献

[1] Liu R Q, Zhang L Y, Li R Y, et al. The ITU vision and framework for 6G: scenarios, capabilities, and enablers [J]. IEEE vehicular technology magazine, 2025, 20(2): 114–122. DOI: 10.1109/MVT.2025.3532887

[2] Liu R, Zhang L, Zou M, et al. Evaluating radio interfaces towards 6G: what's new and what's different [J]. IEEE network, 2025, 39 (6): 148–154. DOI: 10.1109/MNET.2025.3557403

[3] IMT2030(6G)推进组. 6G通感一体化空口关键技术研究报告 [R]. 2024

[4] IMT2030(6G)推进组. 6G通感一体化协作感知关键技术 [R]. 2024

[5] Liu R Q, Zhang L Y, Mao T Q, et al. Integrated sensing and communication for 6G: motivation, enablers and standardization [C]//Proceedings of IEEE/CIC International Conference on Communications in China (ICCC Workshops). IEEE, 2023: 1–6. DOI: 10.1109/ICCCWorkshops57813.2023.10233791

[6] 3GPP 3GPP TS 22.137 Service requirements for integrated sensing and communication [S]. 2023

[7] 3GPP TR 22.870 Study on 6G use cases and service requirements [S]. 2024

[8] Lachvajderová L, Trebuňa M, Kádárová J. Unlocking industry potential: the evolution and impact of digital twins [J]. Acta mechanica slovaca, 2024, 28(1): 46–51. DOI: 10.21496/ams.2024.009

[9] 3GPP TR 38.901 Study on channel model for frequencies from 0.5 to 100 GHz [S]. 2017

[10] Lou J P, Liu R Q, Jiang C X, et al. A unified channel model for both communication and sensing in integrated sensing and communication systems [C]//Proceedings of IEEE 98th Vehicular Technology Conference (VTC2023–Fall). IEEE, 2023: 1–6. DOI: 10.1109/VTC2023–Fall60731.2023.10333766

[11] Jiang C X, Liu J C, Lou J P, et al. A novel approach to model the scattering environment in channel modeling for integrated sensing and communications [C]//Proceedings of International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2024: 1809–1813. DOI: 10.1109/IWCMC61514.2024.10592557

[12] 3GPP R1–2502063 Joint views on mono–static background channel modeling [S]. 2025

[13] Liu R Q, Jian M N, Chen D W, et al. Integrated sensing and communication based outdoor multi–target detection, tracking, and localization in practical 5G Networks [J]. Intelligent and converged networks, 2023, 4(3): 261–272. DOI: 10.23919/ICN.2023.0021

[14] C114. 中兴通讯：数智赋能，逐梦低空 [EB/OL]. (2024–06–28) [2026–01–06]. http://www. c114. com. cn/news/127/a1266749. html

[15] C114. 20KM！台州移动联合中兴通讯率先实现超远海域通感一体商用能力验证 [EB/OL]. (2024–12–25) [2026–01–06]. https://www.c114.com.cn/news/127/a1280824.html

[16] Ma Y H, Yuan Z F, Xia S Q, et al. Highly efficient waveform design and hybrid duplex for joint communication and sensing [J]. IEEE Internet of Things journal, 2023, 10(19): 17369–17381. DOI: 10.1109/JIOT.2023.3274120

[17] Zhu F H, Wang X Q, Jiang S M, et al. Wireless large AI model: shaping the AI–native future of 6G and beyond [PP/OL]. arXiv (2025–04–20)[2026–01–06]. https://arxiv.org/abs/2504.14653

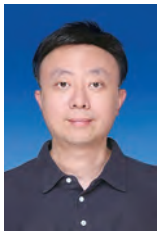
[18] Wang X Q, Zhu F H, Yang Z H, et al. Bridging physical and digital worlds: embodied large AI for future wireless systems [PP/OL]. arXiv(2025-06-30) [2026-01-06]. https://arxiv. org/abs/2506.24009

作 者 简 介

向际鹰，中兴通讯股份有限公司首席科学家；先后从事3G、4G、5G、B5G和6G相关研发工作；曾获国家科技进步奖特等奖、二等奖，以及国家技术发明奖等，并先后获得中国通信产业技术贡献人物、中华杰出工程师等称号。

蒋创新，中兴通讯股份有限公司通感一体化标准专家，3GPP RAN1资深代表，移动网络和移动多媒体技术国家重点实验室专家，担任ETSI ISAC ISG信道模型报告人，中国IMT-2020通感任务组副组长，曾任3GPP R18定位课题的功能负责人；主要从事4G/5G/6G无线物理层技术研究和标准推进工作，拥有国际标准提案及重要专利上百项。

高音，中兴通讯股份有限公司标准预研总工，原3GPP RAN3主席，CCSA TC624 WG4副组长，移动网络和移动多媒体技术国家重点实验室专家，曾任国家重大专项负责人；主要从事3G/4G/5G/6G无线网络创新技术研究和标准推进，先后策划并完成多个国际标准立项，担任多个国际标准、技术报告、行业标准报告人，个人递交3GPP提案近700篇，写入国际标准的重要专利近百族。



许进，中兴通讯股份有限公司无线算法部副部长；主要负责中兴通讯无线通信算法研究及标准化推进工作，研究方向涵盖通信理论、无线接入网技术、通感一体化、人工智能、信道编码与调制以及语义通信等领域。

刘峻琛，中兴通讯股份有限公司通感一体化以及AI应用方向研究专家，3GPP RAN1代表；在3GPP R19通感信道建模以及R20 5G-A无人机通感仿真建模评估方向贡献突出，在6G通感一体化波形以及融合AI提升通感能力方面有深入研究。

新增编委介绍

高新波

西安电子科技大学教授、校长，国家级人才入选者，大数据安全教育部工程研究中心主任，现为第十四届全国政协委员，中国科协第十届全国委员会委员，教育部科技委委员、教学指导委员会委员，IEEE Fellow，中国电子学会、中国计算机学会、中国人工智能学会、中国图像图形学学会会士；主要研究方向为人工智能、机器学习、计算机视觉和模式识别；作为团队负责人入选教育部长江学者创新团队、科技部重点领域创新团队；2020年获全国创新争先奖，曾获得国家自然科学奖二等奖1项、省部级科技一等奖6项；主持国家自然科学青年科学基金（A类）等项目30余项，发表论文500余篇，出版专著、教材7部。

苏 森

重庆邮电大学教授、校长，重庆人工智能学院院长，智能交互与体验系统文旅部重点实验室主任，教育部人工智能专业建设虚拟教研室主任，兼任教育部虚拟教研室重点领域学科协作组组长、北京通用人工智能学会理事；2005年入选教育部新世纪优秀人才计划，2017年入选“万人计划”科技创新领军人才；主要研究方向为可信人工智能、自然语言处理和数据隐私保护；曾获国家科技进步奖二等奖1项、中国通信学会科技进步奖一等奖1项。

6G沉浸式通信业务与关键技术探索

Exploration of 6G Immersive Communication Services and Key Technologies

熊春山/Xiong Chunshan，万青/Wan Qing，陶源/Tao Yuan
（中信科移动通信技术股份有限公司，中国 北京 100085）
(CICT Mobile Communication Technology Co., Ltd, Beijing 100085, China)

DOI：10.12142/ZTETJ.202601005
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260225.1015.006
网络出版日期：2026-02-25
收稿日期：2025-12-20

摘要：面对沉浸式通信业务向多行业渗透以及通感算智深度融合的发展趋势，6G网络面临诸多技术挑战。提出一种面向6G的新型服务质量（QoS）架构及一系列关键技术。具体包括：采用包级细粒度QoS控制，定义一种保证速率且非资源预留的新型QoS类型，以提升用户设备上行速率并实现用户面快速动态QoS调控；基于6G内生智能实现流量检测与QoS参数生成，提出多模态流的协作与同步方案；设计支持智能体数字人通信的意图感知与流量特征自适应传输方案；通过网络计算协同缓解人工智能生成内容（AIGC）算力瓶颈，满足沉浸式生成式通信的性能需求。基于上述技术，创造性构建了一套6G全息沉浸式通信业务验证平台，实现了虚实共生的全息体验，验证了AI赋能QoS的有效性。该平台在提供极致沉浸体验的同时，借助智能体数字人开创了保护个人隐私的新型视频通话空间，为未来移动通信新场景的培育奠定了技术基础。

关键词：沉浸式通信；QoS；多模态；同步与协作；智能体通信；AIGC；全息；数字人

Abstract: Facing the penetration of immersive communication services into multiple industries and the deep integration of sensing, communication, computation, and artificial intelligence, 6G networks are confronted with numerous technical challenges. A novel 6G quality of service (QoS) architecture and a series of key technologies are proposed. These include the adoption of packet-level fine-grained QoS control and the definition of a new QoS type that guarantees bit rates without resource reservation, so as to enhance user equipment uplink rates and enable fast dynamic QoS control on the user plane. Leveraging 6G endogenous intelligence, traffic detection and QoS parameter generation are realized, and collaboration and synchronization schemes for multi-modal streams are developed. An intention-aware and traffic-adaptive transmission scheme is designed to support avatar-enabled AI agent communication. Furthermore, computational and network resources are coordinated to alleviate the computational bottlenecks of Artificial Intelligence Generated Content (AIGC) and meet the performance requirements of generative immersive communication. Based on the aforementioned technologies, a service verification platform for 6G holographic immersive communication is innovatively constructed, achieving the holographic experience that virtual and real worlds coexist, and validating the effectiveness of AI-enabled QoS. While delivering ultra-immersive experiences, a novel video communication space that protects personal privacy is created through the use of intelligent avatars, laying a technical foundation for cultivating new scenarios in future mobile communications.

Keywords: immersive communication; QoS; multi-modality; synchronization and collaboration; AI agent communication; AIGC; holography; avatar

引用格式：熊春山，万青，陶源. 6G沉浸式通信业务与关键技术探索 [J]. 中兴通讯技术, 2026, 32(1): 24-28. DOI: 10.12142/ZTETJ.202601005
Citation: Xiong C S, Wan Q, Tao Y. Exploration of 6G immersive communication services and key technologies [J]. ZTE technology journal, 2026, 32(1): 24-28. DOI: 10.12142/ZTETJ.202601005

沉浸式通信的应用场景不再局限于娱乐消费，而是持续向教育、医疗、文旅、工业制造等垂直行业拓展。为此，国际电信联盟（ITU）已将其列为6G六大重点应用场景之一[1]。与此同时，人工智能与计算技术正深刻影响并重塑沉浸式通信应用的实现路径。例如，基于大语言模型的智能体（AI Agent）通信正在颠覆传统的人机及机器间交互逻辑。在内容制作层面，沉浸式业务正加速向计算机生成内容与云端实时渲染相结合的模式转型[2]，显著降低了构建沉浸式虚拟空间的技术门槛。然而，当前5G系统在架构设计和网络扩展能力方面仍面临一定的技术瓶颈，难以充分满足更

基金项目：国家科技重大专项（2024ZD1300400，2025ZD1301800）

为严苛的性能指标要求及日趋多元化的应用需求。本文立足6G系统的发展视角，旨在探索支撑未来沉浸式通信标准化与产业化的关键技术路径与可行解决方案。

1 6G沉浸式通信的发展趋势与挑战

随着人工智能与计算技术的持续突破，沉浸式通信技术正加速向通感算智融合的方向演进，相关应用与设备也从概念验证走向产品规模化落地，呈现出以下核心趋势：

1）应用场景多元化：从以娱乐游戏为主拓展至行业应用与社会治理，涵盖扩展现实（XR）远程医疗诊断、沉浸式教育、虚拟演唱会及数字孪生等多个领域。沉浸式通信正逐步成为连接物理世界与数字世界的关键纽带。

2）消费体验升级：具备裸眼3D、全息交互功能的沉浸式设备（如智能眼镜）正加速走向大众市场。预计未来5年，全沉浸式配置设备的复合年均增长率（CAGR）将超过23%[3]。同时，沉浸式设备持续向轻量化、低功耗、多模态方向演进。

3）与AI技术深度融合：集成智能体的沉浸式通信设备可实现人机及机器间的意图感知，支持语言、手势、表情等多模态自然交互。此外，以AI模型为核心，由文本、图像或语音指令驱动的内容自动生成与动态演化[4]，将支撑实时高保真的虚拟环境与角色构建。其中，三维数字人（3D Avatar）[5]因具备个性化、灵活性及情境感知能力，成为当前热点方向之一。

4）与计算技术深度融合：复杂场景下的实时渲染对终端算力提出更高要求。借助分布式计算与云计算技术，终端的渲染任务正逐步向边缘端或云端迁移，有效提升系统性能与用户体验。

随着ITU将沉浸式通信列为6G六大典型场景之一，第3代合作伙伴计划（3GPP）、IMT-2030（6G）推进组等标准组织已相继开展6G用例与需求的研究。总体而言，6G沉浸式通信对系统设计提出的挑战可归纳为以下几个方面：

1）多维度拓展网络通信性能。新型媒体形态与高交互体验要求网络同时具备超大带宽（吉比特每秒量级）、极低时延（毫秒级）及高可靠（99.999%级）的传输能力。

2）多维度同步要求。多模态与多源数据需在时间、空间、运动方向等维度满足严格的同步要求（允许一定门限偏差），以保障沉浸式体验的一致性。

3）动态适配业务特征变化。业务流量特征会随模态切换/组合、非周期媒体传输等出现突发性变化，网络需具备实时动态适配能力，以确保确定性的用户体验。

4）支持智能体通信。基于智能体的通信将是6G区别于传统5G通信的关键特征之一。智能体通信不仅要支持传统媒体流的传输，还需支持基于意图及其AI隐空间解析/生成内容的通信模式。

5）协同计算与通信。当终端将计算任务卸载至边缘或云端时，网络需实现计算与通信的协同优化，以在网络环境与算力资源动态变化的条件下保障端到端的总时延要求。

2 6G沉浸式通信的关键技术

为应对上述挑战，本文中我们提出一种面向6G的新型服务质量（QoS）架构，并围绕多模同步与协作机制、适配智能体数字人通信特点的传输方案，以及支持生成式通信的网络计算等关键技术展开分析。

2.1 6G新QoS架构

基于智能内生特性，6G网络能够实时感知业务流量特征与数据类型，动态调整QoS策略并进行相应的资源分配。其核心在于构建一种“内容可感知、弹性自适应、精准可控”的新型QoS机制。

1）包级QoS控制

6G QoS架构支持更细粒度的控制单元——包级QoS。5G的最小控制单元为QoS流，同一流内所有协议数据单元（PDU）采用一致的QoS参数与控制逻辑；5G-A进一步引入PDU集（PDU Set）级控制，确保流内所有PDU Set参数一致。而6G则允许将同一IP五元组数据流中的单个数据包分配至不同的6G QoS标识（QFI），每个QFI对应不同的6G QoS参数配置。

终端与应用服务器可感知每包内容并进行语义标记，通过传输协议（如QUIC）的数据帧将标记传递至6G网络。6G用户设备（UE）或6G用户面功能（UPF）根据接收到的语义标识作为包过滤器，将不同数据包映射至相应的6G QFI。6G无线接入网（RAN）通过解析包的6G QFI及其关联的QoS参数，实现更精细的QoS流级传输调度。

2）保证速率的非资源预留型QoS类型

本文定义一种新型QoS类型——增强型非保证比特率（E-Non-GBR），可在动态适配资源变化的同时提供速率保证。5G中的保证比特率（GBR）与时延关键型GBR采用强制资源预留方式，在无线资源动态变化及业务编码速率可变场景下，刚性预留易导致资源浪费。而5G的非保证比特率（Non-GBR）虽灵活性高，却无法提供速率保证，难以支撑高价值业务。E-Non-GBR为QoS流配置所需流比特率（RFBR）与所需流最大比特率（RFMR），但不参与资源预留。通过动态QoS机制，在不预留资源的前提下实现对所需

流比特率的保障。

3）强化网络与UE应用的交互

5G主要依赖网络能力开放接口支持RAN、核心网（CN）与数据网络（DN）的应用功能（AF）/应用服务器（AS）之间的交互。6G则进一步支持UE向RAN提供可选的上行RFBR/RFMR列表，由RAN选择并指示支持的参数标识，确保UE获得指定的保证速率。

4）用户面的开放与交互

在6G系统中，RAN与CN可通过用户面实现与DN AF/AS或UE应用之间的相互开放，感知方式更加高效。借助用户面交互优势，可有效降低信令开销与交互时延，同时保留控制面交互作为重要补充手段。

5）AI赋能QoS

6G内生AI将深度融入QoS与策略设计：UPF可智能识别新业务流量特性，为QoS参数特性标识打标；6G策略控制功能（PCF）可智能生成QoS参数；内生AI还可优化用户面资源调度，避免频繁修改QoS参数，从而实现用户体验与系统容量的协同提升[6]。

2.2 6G多维度多模通信的同步与协作

6G多模态通信将融合音视频、触觉信号、数字人（含模型、表情码、姿态码）、用户手势/位置/姿态信息，以及人工智能生成内容（AIGC）输入提示词与输出内容等多种数据流，对同步性与融合性提出更高要求。部分新型模态数据在时间、空间与运动维度上的同步需求，将进一步映射为网络传输中的时间同步、抖动控制及同步门限等关键指标。

5G-A通过在同一QoS流（QoS Flow）内配置相同的分组时延预算（PDB）实现多模同步。6G则基于包级QoS控制方案，将“时间”作为特定QoS特性参数，并将不同数据包之间的同步抖动要求作为QoS参数，加载至GPRS隧道协议用户面部分（GTP-U）协议扩展头中。所有数据包的时间信息由UE与UPF映射至6G系统内部时钟，时间精度依据应用层需求确定；同一时间戳对应的多个数据包（如PDU集）可支持编码优化；多流间的同步门限则可通过控制面或用户面传递至用户面节点。

6G将进一步强化多模态协作机制。多模态数据流通常映射至不同的QoS Flow，通过优化无线资源变化时的QoS调整逻辑（如独立调整或联合调整），以及基站切换时的QoS Flow建立策略（如切换失败时优先放弃某一模态流或整体放弃），可有效提升多模态数据流的协同传输能力。具体实现方案仍有赖于6G QoS新架构的进一步验证。

2.3 基于沉浸式智能体数字人实时通信

智能体数字人是6G沉浸式智能体通信的关键技术之一。相较于5G沉浸式业务，沉浸式智能体数字人实时通信的流量模式呈现显著差异：意图交互频次高、单次交互持续时间多变、意图内容多样且突发性强；同时，新型多模态数据（如意图、Token、张量、隐空间数据等）的引入，以及会话流量特征的短时多变性（如突发交互与周期性流式输出交替出现），使得流量模式更加复杂。智能体通信的流量特征及其QoS保障已成为6G标准化的重点研究方向之一[7]。

在图1所示的实验系统中，智能体数字人A通过内置的实时AI表情识别模型，对用户面部及姿态视频进行采集与分析，生成表情码与姿态码，并通过支持QoS感知的传输协议（如网页实时通信（WebRTC）的不同数据通道）将多种AI语义流实时传输至智能体数字人B。其中，3D数字人模型、表情码、姿态码及意图分别采用独立的网页实时通信数据通道进行传输。各数据通道配置差异化的QoS调度策略，并基于2.2节提出的AI流识别模型对不同数据流进行分类识别；同时，借助实时传输协议（RTP）层统一的采样时间源，实现多模态数据的同步调度。智能体数字人之间采用统一的语义本体，以保障意图识别的准确性。接收端可使用本地或接收到的通用3D数字人模型，结合周期性传输的表情码与姿态码流，快速驱动数字人模型渲染，真实还原对方用户的表情与姿态。

在沉浸式数字人视频通信过程中，仅传输用户的表情及姿态码信息，未涉及任何真实用户的视频或背景数据，从而有效保护用户隐私。预计在6G时代，基于数字人的视频通信将开辟全新的市场空间，使用户能够安心开启摄像头，以

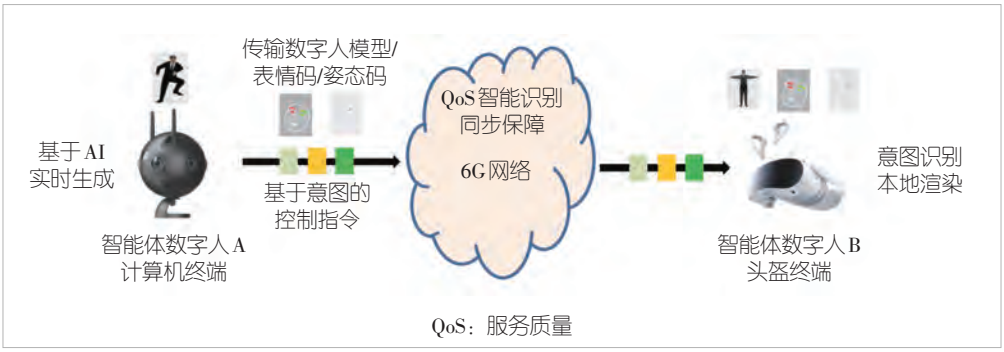


图1 智能体数字人实时通信

数字人形态与他人进行通信。

2.4 基于网络计算的沉浸式智能体生成式通信

基于智能体的AIGC有望成为6G沉浸式业务内容的核心生产力。当前，AIGC生成符合用户意图的图片或视频大多需经多步完成：通常先以预览形式向用户呈现中间结果，供其检查效果并允许中断生成过程以调整提示词（即意图）与配置参数，待用户满意后再最终生成高质量、个性化的内容。AIGC内容质量可近似用“算力×时间”衡量，但现有算力调度与协调模式及其适配的AI算法，难以支撑实时输出高质量的沉浸式内容。若采用高性能AI算法，算力需求将激增，进而引发能耗、体积、重量及成本上升，难以在便携式头盔等终端设备上普及。因此，解决AIGC大算力需求与可穿戴终端有限算力之间矛盾的核心路径在于：将终端AIGC任务卸载至算力更强的应用服务器，同时保障生成多模态数据的时延要求[8]。例如，6G网络可向边缘计算节点提交计算时延约束（如多模态同步门限内），边缘节点据此分配算力；针对端到端时延与往返时延保障（含处理时延与通信时延，计算密集型业务中处理时延占比较高），6G网络可结合业务时延需求、计算任务量、网络状态及应用服务器算力，动态选择适配的应用服务器与路由[9]。

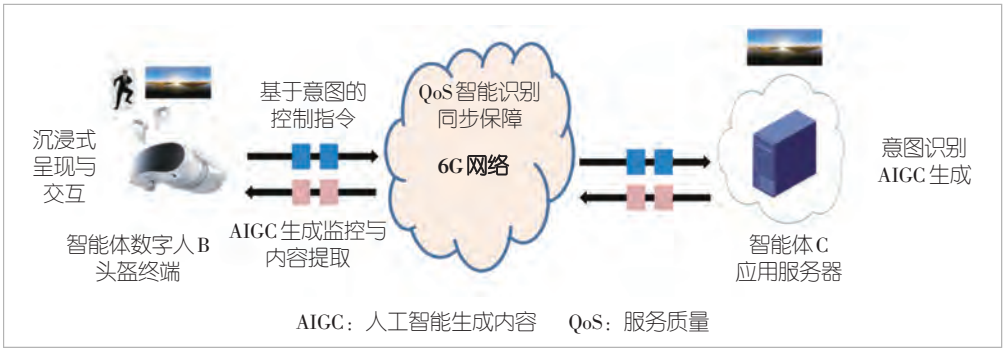


图2　智能体生成式通信

在图2所示的实验系统中，智能体数字人B向应用服务器（智能体C）发送意图控制指令（如语音形式的自然语言或文字），智能体C依据AIGC指令生成3D高斯溅射（3DGS）模型，借助应用服务器的算力保障实时输出高质量沉浸式内容，并满足业务时延需求。智能体B可实时监控AIGC生成过程，并提取所生成的内容。

3 6G沉浸式业务探索与实践

自2022年起，团队持续开展沉浸式业务研发，涵盖从3D、虚拟现实（VR）、XR到全息的技术演进，从3自由度（3DoF）到6自由度（6DoF）的交互升级，同步探索多模QoS保障技术，并自主搭建6G全息通信业务平台，已收获多项创新成果，如图3所示。

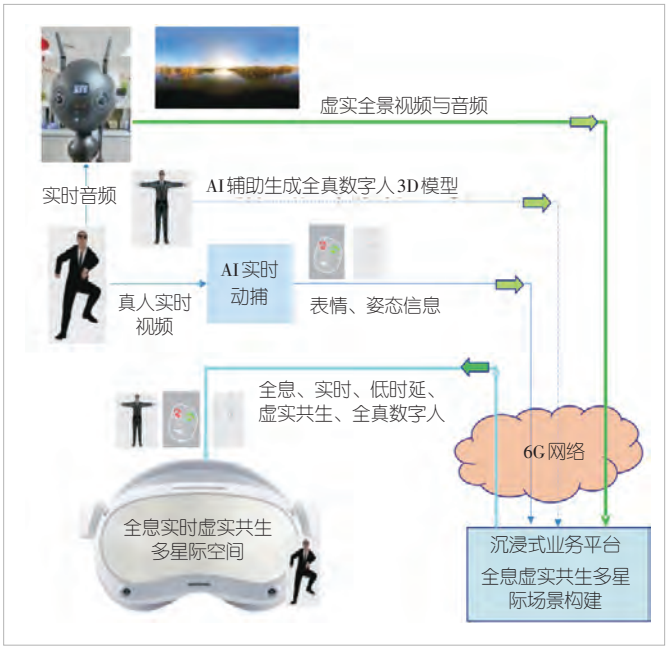


图3　6G全息沉浸式通信验证平台

1）多空间的全息沉浸式体验：基于不同位置的多虚拟空间组合，用户移步至不同区域可进入相应的沉浸空间，体验多种模态的沉浸式效果。通过非交互大带宽的“超高清晰虚拟空间”与“全景设备实时采集的超低时延、超高清真实空间内容”，实现虚实空间共生与融合。用户在空间移步穿越全景场景时，可体验类似电影《星际穿越》中的“虫洞”效果。此外，平台还支持数字人实时交互空间和文本生成3DGS场景的AIGC交互空间。

2）QoS智能控制：基于2.3节所述的AI流识别模型，智能识别不同流类型，并采用2.1节所述的差异化QoS控制策略，保障交互式全景实时视频的传输质量。端到端移动到图像（M2P）时间控制在100~150 ms，其中网络传输时延小于5 ms，终端头盔设备可稳定接收。

3）基于智能体的数字人通信：通过3D与AI工具构建作者的全真3D网格数字人模型，并通过2.3节所述的智能体通信，实现3D数字人的实时表情、姿态驱动与渲染，体验有别于传统视频通信的新场景。

4）基于意图的智能体生成3D场景：将AIGC生成计算任务卸载至配置大算力的应用服务器，在满足低业务时延要求的条件下，输出高质量、个性化的沉浸式内容。

5）多模态数据传输：全息通信业务传输多种模态数据，包括两个虚实空间全景视频流、一个虚空间语音流、一个数字人语音流、一个实空间语音流、一个3D数字人模型流、一个表情码流、一个姿态码流、一个AIGC提示词命令流，以及一个获取AIGC 3DGS模型的FastAPI（高性能数据接口）流。6G实验网络可智能识别关键数据流，分配差异化QoS，并依据传输层标记的时间戳，在网络调度时实现多模态数据的同步传输（如大带宽视频与AI语义驱动的高可靠数字人表情码的协同传输）。该多模态多空间全息沉浸式通信业务系统为行业内首个公开报道的多模态的融合6G沉浸式通信业务系统，支撑2023—2025年工信部与运营商组织的5G基站XR业务测试，并于2025年6月获上海世界移动大会（MWC）最佳演示奖[10]。

4 结束语

沉浸式业务在提供极致用户体验的同时，也面临传统沉浸式内容生成成本高昂的问题。近年来，3DGS技术持续发展，借助消费级相机与软件，仅需多角度照片即可快速重建高逼真3D场景。该技术正加速成熟，有望激活用户生成的沉浸式长短视频市场。

基于数字人的表情码与姿态码视频通信，可在全面保护用户隐私的前提下，创建全新的移动视频通信市场，充分彰显沉浸式通信业务的市场创新潜力。

面向6G，沉浸式通信带来多维度技术挑战，亟需通过多种新型6G架构技术加以应对。

致谢

本研究得到中信科移动通信技术股份有限公司王可、秦海超、徐晖、艾明、王胡成、肖国军等专家的指导与帮助；同时，在6G全息沉浸式通信开发与实验过程中，刘强、陆洁、程志密、张晓康、谷肖飞等同事多年来提供了大量的演示与测试支持，并提出诸多改进建议。谨向以上所有人员致以诚挚谢意！

参考文献

[1] ITU. Framework and overall objectives of the future development of IMT for 2030 and beyond: ITU-R M.2160-0 [S]. 2020
[2] Hu K Y, Jin Y L, Zhou H, et al. Generative AI for immersive video: recent advances and future opportunities [PP/OL]. arXiv(2025-08-23)[2026-01-03]. http://arxiv.org/abs/2508.17163
[3] Mordor Intelligence. Virtual Reality (VR) market size & share analysis - growth trends and forecast (2026 - 2031) [EB/OL]. [2026-01-14]. https://www. mordorintelligence. com/industry-reports/virtual-reality-market
[4] VELU D, EMBRY A. The convergence of AI and immersive-environments: shaping the future of digital realities [EB/OL]. [2026-01-03]. https://www.capgemini.com/wp-content/uploads/2024/08/The-convergence-of-AI-and-immersive-environments.pdf
[5] 3GPP. Study of avatars in real-time communication services: 3GPP TR26.813 [S]. 2025
[6] Wang Y Y, I C-L, Sun J S, et al. End to end AI architecture for next generation network [J]. IEEE wireless communications, 2024, 31(1): 86-92. DOI: 10.1109/mwc.013.2200269
[7] Chen Z Q, Sun Q, Li N, et al. Enabling mobile AI agent in 6G era: architecture and key technologies [J]. IEEE network, 2024, 38(5): 66-75. DOI: 10.1109/mnet.2024.3422309
[8] Zhao L Q, Zhou G R, Zheng G, et al. Open-source multi-access edge computing for 6G: opportunities and challenges [J]. IEEE access, 2021, 9: 158426-158439. DOI: 10.1109/access.2021.3130418
[9] Yan M, Guo H R, Chan C A, et al. Semantic communication-enabled multi-access edge computing network resource optimization in the 6G era [J]. IEEE wireless communications, 2025: 1-9. DOI: 10.1109/mwc.2025.3600791
[10] 中信科移动. 中信科移动6G全息沉浸式通信系统荣获GSMA最佳演示奖 [EB/OL]. (2025-06-30)[2026-01-10]. https://mp.weixin.qq.com/s/Cr_kMW_N7rOkovwNNc_VrA

作 者 简 介

熊春山，中信科移动通信技术股份有限公司主任级工程师；主要研究领域为无线网络架构与业务、QoS与政策、IP新技术与移动网络融合、XR/沉浸式通信、数字人、AIGC等；参与国家级和省部级科研项目9项；已发表论文17篇，获授权发明专利140余项。

万青，中信科移动通信技术股份有限公司工程师；主要研究领域为移动通信系统业务与架构、蜂窝物联网、卫星通信、定位、沉浸式通信等；参与终端和网络产品设计20余款，参与国家级和省部级科研项目3项；已发表论文4篇，获授权发明专利20余项。

陶源，中信科移动通信技术股份有限公司工程师；主要研究领域为5G-A/6G网络架构、边缘计算、算网融合、时间敏感通信、XR等；参与国家级和省部级科研项目5项；已发表论文3篇，获授权发明专利30余项。

6G无蜂窝大规模MIMO关键技术研究进展

Research Progress on Key Technologies of 6G Cell-Free Massive MIMO

尤肖虎/You Xiaohu[1,2]，王东明/Wang Dongming[1,2]，曹阳/Cao Yang[2]
（1. 东南大学移动通信全国重点实验室，中国 南京 210096；
2. 紫金山实验室，中国 南京 211111）
(1. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China;
2. Purple Mountain Laboratories, Nanjing 211111, China)

DOI：10.12142/ZTETJ.202601006
网络出版地址：https://link.cnki.net/urlid/34.1228.tn.20260226.1443.002
网络出版日期：2026-02-26
收稿日期：2025-12-22

摘要：在6G标准化商用初期（Day-1），无蜂窝技术受到了广泛关注。系统梳理了面向6G的无蜂窝通信所涉及的关键支撑技术，包括分布式收发机架构、信道信息获取与测量机制、频谱资源融合与灵活双工设计，以及分布式资源管理策略等。在此基础上，提出了一种基于数字孪生增强的无蜂窝传输优化方法，进一步提升了系统性能。同时，展示了面向高频段的无蜂窝大规模多输入多输出（MIMO）试验验证结果，验证了相干联合传输在实际系统中的可行性。最后，探讨了无蜂窝大规模MIMO与通感一体化融合的潜在研究方向，为未来系统设计提供了新思路。

关键词：6G；无蜂窝大规模MIMO；大规模协作收发点

Abstract: Cell-free technology has attracted considerable attention in the Day-1 phase of 6G standardization. Key enabling technologies for cell-free systems in 6G standardization are systematically reviewed, including distributed transceiver design, channel state information acquisition and measurement, spectrum fusion and flexible duplex design, as well as distributed resource allocation strategies. Building on this foundation, a digital twin-enhanced optimization method for cell-free transmission is proposed to further improve system performance. Experimental results for high-frequency cell-free massive multiple-input multiple-output (MIMO) are presented, which validate the feasibility of coherent joint transmission in practical systems. Finally, potential research directions for the integration of cell-free massive MIMO with integrated sensing and communication are discussed, offering new insights for future system design.

Keywords: 6G; cell-free massive MIMO; large-scale cooperative transmission and reception points

引用格式：尤肖虎，王东明，曹阳．6G无蜂窝大规模MIMO关键技术研究进展[J]．中兴通讯技术，2026，32(1)：29-37．DOI：10.12142/ZTETJ.202601006

Citation: You X H, Wang D M, Cao Y. Research progress on key technologies of 6G cell-free massive MIMO [J]. ZTE technology journal, 2026, 32(1): 29-37. DOI: 10.12142/ZTETJ.202601006

随着第3代合作伙伴计划（3GPP）6G标准化工作的推进，6G空口（6GR）在商用初期（Day-1）计划支持的核心技术成为业界的关注焦点。目前，全球多家企业与科研机构相继提出6GR总体技术的观点。其中，多输入多输出（MIMO）技术作为提升移动通信系统容量的关键手段，在6G研究中备受关注。

学术界和工业界对MIMO技术的研究已持续近30年，形成了较为完整的技术演进脉络。从3G演进版本到5G Release 15（R15），集中式MIMO逐步形成了以波束赋形和预编码为核心的技术体系，成为系统的基础配置。理论上，分布式多天线（D-MIMO）被视为MIMO的更一般化形式[1]。在学术层面，相关研究经历了从分布式天线系统、分布式MIMO到无蜂窝大规模MIMO的演进；在工业界，4G阶段引入了协作多点传输（CoMP）概念，5G演进中则逐步形成了以多收发点（mTRPs）为核心的技术框架。经过20余年的持续研究，D-MIMO逐渐呈现出向大规模协作传输和去蜂窝化方向发展的趋势。

在5G新空口（NR）标准的演进过程中，从R15到R19，MIMO及D-MIMO的性能与场景适应能力持续增强。

基金项目：国家科技重大专项（2025ZD1305100）

R15奠定了大规模MIMO的基本架构；R16至R18阶段，重点增强了对高速移动场景的测量上报能力，完善了码本设计，并逐步引入上行mTRP接收与下行非相干联合发送（NCJT）方案。为支持更高效的相干联合传输（CJT），R19进一步制定了针对TRP间非理想时频同步与非互易收发通道的测量与反馈机制[2]。总体而言，mTRP技术涉及上行功率控制、定时提前（TA）调整、相干/非相干传输方法、信道状态信息（CSI）反馈、波束管理以及TRP间的时频同步等多个物理层关键技术环节。3GPP在继承4G准共址（QCL）概念的基础上，持续完善测量与反馈机制，逐步构建起以传输配置指示（TCI）为核心的技术体系[3]，在一定程度上实现了以用户为中心的多TRP协作传输。

在当前R19协议框架下，单个5G基站（gNB）可支持多个TRP，进而实现了小区域范围内的去蜂窝化。随着基站基带信号处理能力的不断提升，gNB能够联合处理的TRP数量显著增加，系统容量随之持续增长。在近期3GPP会议上，多家企业普遍认为6GR需考虑从mTRP向无蜂窝技术演进[4-7]。

无蜂窝大规模MIMO的核心目标是通过大规模节点在同一时频资源上服务大量用户，将多小区干扰转化为有用信号，从而实现频谱效率和系统性能的持续提升。无蜂窝理论技术的研究主要关注平坦信道下的传输方法[8]。面向大规模组网的实际需求，基带信号处理和资源分配机制需具备支撑用户规模持续扩展的能力。为此，学术界提出了可扩展的分布式协作基带处理架构，以及可扩展的调度与资源分配策略[9-10]。该思路旨在支持大规模协作传输，并推动移动通信网络性能的持续优化。可扩展性的基本要求是：在用户规模不断扩大的条件下，网络侧收发节点以及空时频资源的调度复杂度仅呈线性增长。在此约束下，每个TRP在同一时频资源上可服务的用户数，以及单个调度器或资源分配实体所管理的用户数量，均应受到合理限制。从实际部署角度看，受限于成本、信道传播时延及异步问题，无论是单个TRP同时服务的用户数，还是为一个用户服务的TRP数量，均存在一定上限。因此，尽管理论上单个无蜂窝基站的信号处理能力可不断提升，实际系统设计仍需在上述约束下寻求平衡。

在5G演进过程中，无蜂窝大规模MIMO的理念逐步被融入标准体系。首先，小区的概念持续弱化，协议未对单个gNB支持的小区数量和TRP数量做出硬性限制。随着终端测量能力的增强，通过TCI机制，为同一终端服务的TRP可分布于不同小区，且数量已达4个。其次，上行协作接收与下行相干协作传输方案亦逐步成熟。从无线传输层面来看，若将物理小区标识（PCI）仅视为一种接入资源的划分方式，并结合无蜂窝大规模MIMO的设计理念，当前体制已基本具备在大范围内实现去蜂窝化部署的能力。再次，在5G演进标准中，基于层1与层2（L1/L2）触发的移动性（LTM）机制有效降低了基站间切换时延，提升了切换区域的业务传输性能。这种以扩大协作区域和增强移动性为核心的架构演进，已逐步突破传统蜂窝网络的边界，成为6G标准化工作的重要起点。

本文围绕6G标准化中的无蜂窝大规模MIMO技术展开讨论，重点分析支持大规模TRP协作的收发机设计与CSI获取技术，探讨无蜂窝大规模MIMO与频谱聚合、双工技术及无线资源分配机制的融合路径。同时，进一步梳理了无蜂窝组网当前面临的主要技术挑战，并探讨数字孪生、高频段无蜂窝组网以及无蜂窝通感一体化（ISAC）等新兴技术方向在提升无蜂窝网络性能、拓展其应用场景方面的潜在作用。

1 从mTRP到无蜂窝大规模MIMO演进的关键技术

无蜂窝技术兼具无线传输与组网的双重属性。因此，将无蜂窝理念融入6G标准，不仅涉及分布式MIMO传输方法的研究，更需要在多用户无线资源分配等组网技术层面取得突破。

1.1 收发机设计

在mTRP技术的演进过程中，3GPP对上行mTRP接收与下行mTRP发射方案开展了系统研究。其中，上行mTRP技术主要包括单频网（SFN）和空分复用（SDM）两种传输模式[11]。通过利用多个TRP的接收分集，可有效提升上行链路的覆盖性能。为进一步增强上行传输能力，3GPP还定义了仅具备上行接收功能的TRP。借助上行定时提前（TA）测量，可在多个TRP之间实现分集接收。当多个TRP各自采用分布式接收机处理信号，并分别将接收信息发送至用户所附着的基带单元进行合并时，该架构便接近于无蜂窝系统中单用户上行分布式接收的实现。将此方案推广至多用户场景，并进一步引入干扰消除技术，可实现无蜂窝大规模MIMO的上行分布式多用户MIMO接收。

在mTRP下行传输技术的演进过程中，3GPP现阶段主要围绕非相干联合发送（NCJT）与相干联合发送（CJT）展开研究。针对NCJT，协议已明确支持两个TRP采用时分复用、频分复用、空分复用或单频网（SFN）等发送模式，使终端能够获得mTRP发射分集增益，从而提升下行传输的可靠性。从理论层面而言，NCJT可视为CJT的一种特殊形式。然而，CJT旨在获取相干波束赋形增益，对参与协作的TRP之间的时频同步精度提出了严格要求。R19版本通过引入码

本反馈机制，并结合时延、频偏及相位反馈，初步构建了较为完备的CJT传输方案。受限于终端能力及反馈开销，R19目前仅支持最多4个TRP的协作传输。同时，由于仅支持两个TCI状态，该方案在复杂场景下的适应性仍存在一定局限。此外，基于码本的传输方式虽具备较好的鲁棒性，但与基于非码本的空分复用联合传输相比，其系统性能存在较为明显的损失。

针对非码本的mTRP多用户协作传输，受限于实现复杂度及终端解调参考信号（DMRS）的信道估计能力，当前网络侧通常配置为每2或4个资源块采用相同的预编码矩阵。从理论层面分析，由于协作多用户MIMO通常工作于分布式近场场景，多个子带共用同一预编码矩阵将带来较为显著的性能损失。因此，有必要结合DMRS的设计，进一步研究适用于多用户场景的高性能协作预编码方案。此外，随着单个基站所配置的TRP数量不断增加，以用户为中心的mTRP动态协作簇传输模式对基带信号处理算力提出了更高要求。在此背景下，学术界提出的可扩展分布式基带信号处理架构，有望为基站产品性能的持续提升提供理论与技术支撑。

总体而言，5G标准从单TRP向mTRP的演进路径呈现一定的碎片化特征。为实现高性能的协作传输，未来方案设计需立足于mTRP架构，综合考虑高频与低频段、不同TRP形态以及终端能力的多样性，构建更为简洁、高效的传输机制。

1.2 上下行信道信息获取

针对时分双工（TDD）系统，R19版本引入了终端辅助的互易性校准机制，使网络侧能够获取各TRP之间的校准系数。该版本主要提出了两种在网络侧获得校准参数的方法。

第一种方法基于上行探测参考信号（SRS）进行信道估计，网络侧依据所得的CSI新增发送一种经过预编码的信道状态信息参考信号（CSI-RS），作为下行校准的参考信号。终端在接收该信号后，测量获得mTRP之间的校准系数（包括时间差及收发通道的相位差），并将其反馈至网络侧。第二种方法则由终端直接反馈下行信道的统计特性，网络侧基于该反馈信息计算得到各TRP间的校准系数。

针对低频段TDD系统，当各TRP之间存在空口链路且网络侧具备自校准能力时，可实现高性能的多用户协作传输，系统容量与传输可靠性均可获得显著提升[12]。由于自校准过程对终端透明，因此无须在终端侧进行标准化定义。基于此，若网络设备厂商支持TRP间的自校准功能，不仅可简化现有QCL与TCI所涉及的复杂流程，降低终端的测量与反馈开销，还能够有效提升系统整体性能。当网络中参与协作的TRP数量较多时，各TRP之间的时间同步精度要求随之提高。目前，采用IEEE 1588精准时间协议（PTP）可在一定程度上实现较高精度的时间同步。然而，随着分发时间信息的交换机级联层数增加，时间同步精度会出现下降。因此，需要进一步提升IEEE 1588协议在级联场景下的时间同步精度。

CJT不仅能够获得相干波束赋形增益，还可实现空分复用增益。当仅需获取波束赋形增益时，可采用分布式CJT预编码方案，即每个TRP仅依据自身的CSI独立进行下行预编码。该方案对信道信息的过时具有一定鲁棒性[12]。相比之下，多用户CJT传输对信道相位变化较为敏感，网络侧需要获取高精度的CSI才能保证性能。总体而言，为实现高性能的mTRP传输，需结合校准相位的获取时刻、基于历史上行SRS与DMRS估计得到的CSI，以及QCL关系，通过预测方式获得下行CSI。

值得注意的是，当前采用预编码CSI-RS获取TRP间互易性校准参数的机制，为终端侧多天线的校准提供了可行路径。具体而言，以一个TRP为参考点，基于上行多端口SRS估计得到的CSI，通过发送预编码的CSI-RS校准信号，可实现单个终端的多天线校准，亦可支持多个终端的天线校准。该机制为上行CJT的实现奠定了基础，成为提升上行覆盖性能的有效方法。

1.3 无蜂窝与载波聚合及子带全双工的融合

近年来，运营商对多频段载波聚合的需求日益增强，特别是在服务超密集用户场景时，载波聚合已成为提升系统容量的关键技术途径。然而，在5G载波聚合框架下，采用多小区管理方式会引入较高的信令开销与传输延迟，导致频谱利用率偏低，且对碎片化频谱的利用效果不佳。因此，在6G Day-1技术讨论中，频谱融合技术受到广泛关注，其中典型方案包括单小区/超小区多载波（SCMC）技术[7, 13]。通过灵活配置TDD系统的上下行时隙配比，该方案可同时支撑大上行与大下行类型的终端需求[5]。可以看出，SCMC技术与无蜂窝技术分别从频域和空域两个维度拓展了传统小区的概念。为更好地支撑相关技术的落地，需要统一考虑初始接入、同步信号设计、传输方法以及信令机制等方面的协同设计。

此外，5G-A引入了子带全双工（SBFD）技术，能够有效降低TDD系统的传输时延并提升上行覆盖性能。在中频段（FR3），通过充分利用多天线系统在空域波束间的隔离特性，可实现SBFD的全新部署[14]。该思路与文献[15]提出的网络辅助全双工（NAFD）技术在理念上相通。NAFD一方

面借助mTRP的大规模天线阵列挖掘空域资源，实现全双工传输；另一方面利用mTRP的协作能力有效抑制交叉链路干扰（CLI），从而在降低时延的同时提升系统频谱效率。

总体而言，6GR在频谱利用与多天线/多通道配置方面将展现出更高的灵活性。从理论层面看，在网络侧实现更为灵活的双工模式，可归结为空口资源的动态配置问题，如文献[16]提出的多节点/多通道双工模式选择方案。在6GR标准制定过程中，频谱融合及SBFD技术与mTRP设计在核心理念上具有相通之处，共同构成了以用户为中心的空时频资源配置框架。然而，在服务用户规模保持不变的条件下，无论是频谱融合还是mTRP技术，均需在提升单用户体验的同时，保障系统总容量的持续增长。因此，随着单个基站所管理的空时频资源池不断扩大，无线资源分配的复杂性也随之增加。

1.4 无蜂窝系统的调度与资源分配

为了与传统蜂窝系统进行公平的性能对比，无蜂窝系统在相同区域和相同TRP规模条件下，服务的终端总数需与传统蜂窝系统保持一致。在传统蜂窝架构中，多个基带单元（BBU）之间的调度采用分布式实现，因而系统复杂度相对较低。当前业界所采用的超小区及小区合并方案，虽能在一定程度上提升单用户业务体验，但由于未能实现多小区间的联合调度与资源分配，系统总容量的提升效果尚未达到预期。相比之下，采用集中式调度与资源分配的无蜂窝基站，虽可实现用户规模与系统总容量的同步增长，但其实现复杂度过高，缺乏可扩展性。

综上所述，当前5G协议已在物理层支持多TRP协作，并初步打破了传统小区的边界。然而，为充分发挥无蜂窝系统的容量潜力，仍需进一步突破以小区为中心的传统资源分配体系，以支撑用户规模与系统总容量的持续提升。

图1给出了无蜂窝系统的实现架构[10]。该架构在物理层支持TRP集合的分布式基带信号处理。具体而言，在虚拟化中心处理单元（vCPU）的低层实现编解码、调制解调等高阶物理层功能，在其上层则执行可扩展的分布式资源分配。为此，需要设计相应的分布式媒体接入控制（D-MAC）调度及资源管理方法。如图1所示，用户在初始接入阶段会形成服务于自身的TRP集合，并与一个D-MAC实例相关联。基于用户与TRP的关联关系及CSI，系统进一步执行用户配对与分布式资源分配。由于交集TRP的存在，每个D-MAC调度实例需要获知其他实例中与之关联的用户信息。为了降低资源冲突导致的干扰，D-MAC调度器需实时交互调度用户的时频资源及DMRS分配信息，通过资源协调机制抑制干扰，并最终确定每个用户的编码调制格式。

在图1所示的架构下，通过结合终端测量上报与网络侧对上行信号的测量，采用L1/L2触发的LTM机制，可借助MAC控制单元（MAC-CE）指令实现TRP集合的动态选择。该方式能够避免传统的层3切换流程，从而有效降低移动性管理时延。随着人工智能（AI）技术逐步引入终端的移动性管理，系统对移动性场景的支撑能力将进一步增强。

在理想的前传/回传条件及D-MAC信息交互前提下，图1所示的D-MAC架构能够实现可扩展的资源分配，并与物理层的分布式基带处理相配合，支撑无蜂窝组网的实现。然而，在非理想回传条件下，如何设计相应的D-MAC机制仍需进一步研究。

2 未来研究方向

从MIMO到分布式MIMO、协作多点传输，再到mTRP，并最终向无蜂窝组网演进，无蜂窝技术发展历程已持续近30年。尽管如此，容量、覆盖及干扰问题仍然是移动通信大范围组网所面临的核心挑战。在提升移动通信网络系统性能的过程中，去蜂窝化方向上仍有诸多问题亟待进一步持续深入地研究。

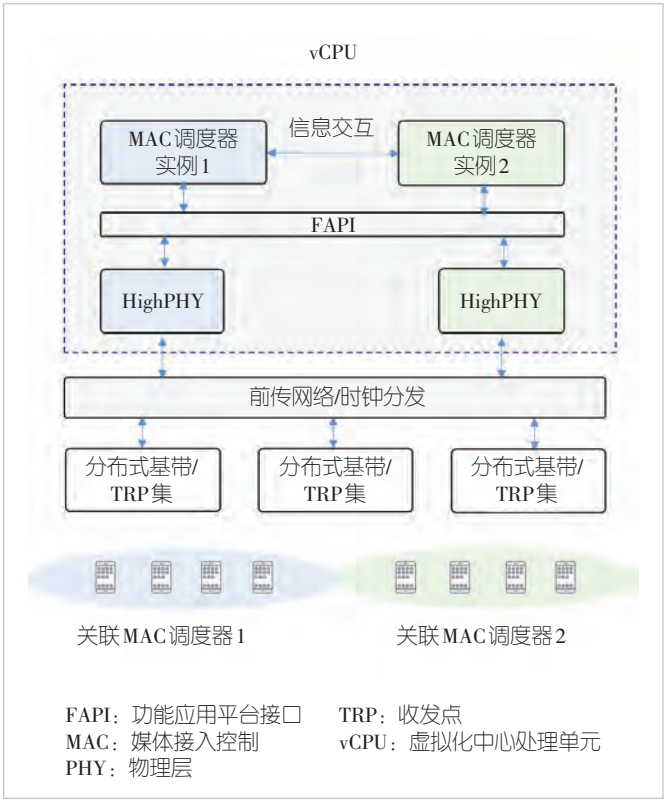


图1 分布式MAC：无蜂窝系统的调度与资源分配

2.1 广域无蜂窝组网的干扰问题

由于移动通信系统上下行发射功率存在不对称特性，系统通常面临上行功率受限与下行干扰受限的双重挑战。从理论层面分析，降低单点发射功率，提升分布式节点的部署密度以形成超密集组网，并结合无蜂窝大规模MIMO技术，可通过协作接收与联合发送的方式提升系统上下行容量。然而，受限于运营商在实际部署中难以无限增加节点密度，在现实的小区间距条件下，无蜂窝技术的作用存在一定局限。

当前移动通信系统普遍采用同步组网，物理层仍沿用正交频分复用（OFDM）技术。受循环前缀（CP）持续时间的制约，当两条无线链路的传播时延差超过CP时长时，将引发异步干扰。如图2所示：在上行通信中，若终端到两个TRP的传播时延差超出CP范围，且远端TRP未相应调整接收窗，则会产生异步问题；在下行通信中，若两个TRP同步发射的信号到达终端的时延差超出CP，且终端未能分别进行同步并实施干扰抵消，则会导致下行远端异步干扰。异步干扰将进一步引发载波间干扰（ICI），增加干扰消除难度，恶化组网性能。当前5G系统采用的典型子载波间隔为15/30/60/120 kHz，其对应的常规CP等效距离分别为1 406 m、703 m、352 m和176 m。理论上，单站的覆盖能力越强（即小区半径越大），在多站点协作场景下，同步区域在整个覆盖区中所占的比例将越小。

在现有基站系统中，网络侧通常不针对终端进行主动同步，终端也仅与单个TRP实现时间同步及TA调整。因此，在上行方向，为提升接收性能，需要基站侧的TRP针对特定终端进行同步，并部署干扰消除接收机。在下行方向，理论上可通过调整波束的发送定时来改善协作传输，但该方案会破坏网络侧的同步机制，实施难度较大。而在终端侧，通过针对特定TRP进行同步并实现干扰抵消接收，可有效提升下行协作传输的性能。然而，由于异步问题破坏了相干传输的基本条件，此类下行异步协作传输仅适用于非相干联合传输场景。

在mTRP服务单用户的上行传输中，通过调整各TRP的接收窗可获得分集接收增益；在下行传输中，采用非相干协作发送并结合终端侧的接收窗调整与干扰抵消，同样可实现下行分集接收增益。单用户场景下的mTRP异步协作复杂度相对较低，具备一定的可行性。

从理论层面分析，在多用户空分复用场景下，无蜂窝技术有望实现全网的干扰消除。然而，在实际大规模组网中，无论是在基站侧还是终端侧，实现多用户干扰抵消（特别是异步干扰抵消）的代价通常难以承受。相关的工程实现涉及同步机制、CSI参数获取以及干扰抵消接收机设计等一系列

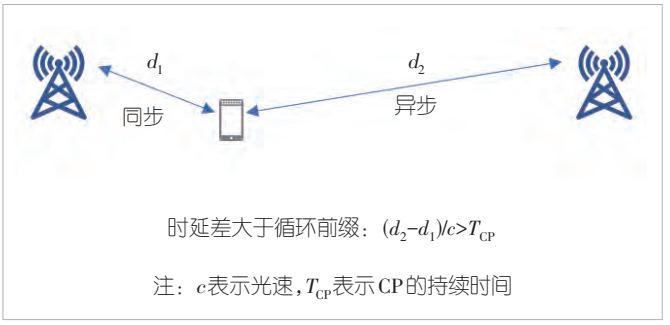


图2 无蜂窝系统的异步问题

复杂技术问题。此外，全网协作所能带来的实际性能增益仍有待进一步评估与研究。

针对密集城区覆盖场景，当前5G系统主要采用大规模MIMO技术进行组网部署。以30 kHz子载波间隔为例，小区半径通常设定在300 m左右。在大规模MIMO采用窄波束传输时，由于波束赋形带来的增益较大，受传播环境影响，反射径可能引发远端干扰。此类远端异步干扰已成为制约网络性能进一步提升的瓶颈。如前文所述，采用干扰抵消方式对抗异步干扰将带来较高的实现代价。因此，在多站点协作背景下，大规模MIMO组网的性能优化问题仍有待进一步深入研究。

在无蜂窝技术的架构下，需要重新审视每个TRP所配置的通道数量。传统大规模MIMO理论表明，天线数量越多，系统所获得的空分复用增益与干扰抑制能力越强。然而，在实际大规模MIMO组网中，受限于导频资源开销及天线单元规模，多站点之间的干扰难以实现完美抑制。事实上，在密集大规模MIMO部署场景下，由于阵列增益较为显著，无论是在空旷区域的小区边界用户覆盖，还是室外站点对室内用户的覆盖中，由直射径或反射径引发的远端干扰均不可忽视。

采用无蜂窝协作接收与发射技术，通过灵活部署多个小规模天线的模块化TRP，并结合空口校准机制，同样可以获得相干传输增益。因此，适当降低单个大规模MIMO TRP的通道数量，通过协作方式实现收发波束赋形增益，是降低网络干扰、提升系统能效的有效技术路径。

2.2 基于RAN数字孪生的无蜂窝组网技术

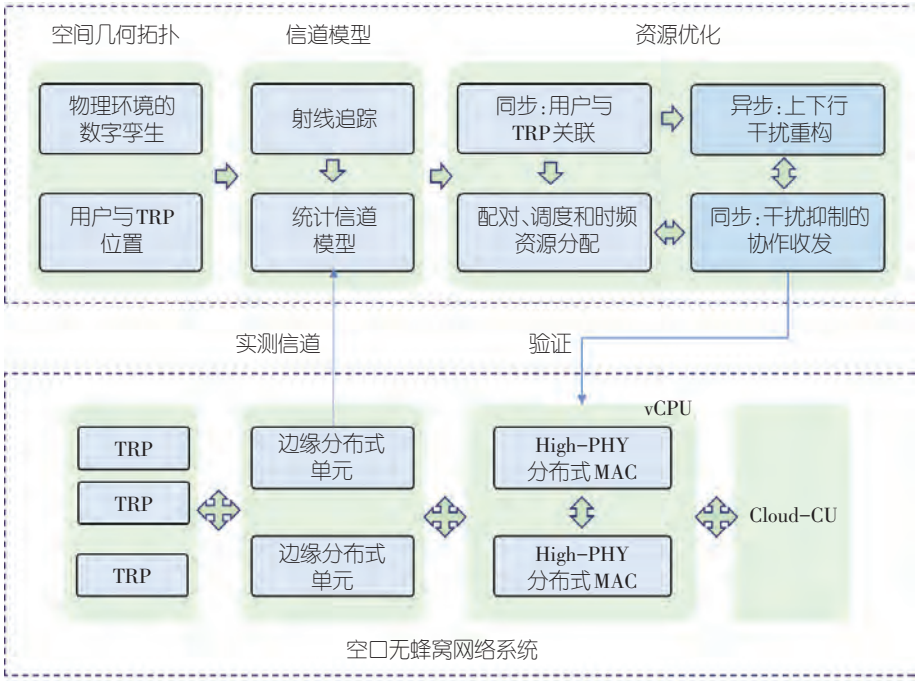
为了提升广域覆盖场景下无蜂窝网络的性能，亟需研究网络级的远端干扰协调与抑制技术。然而，由于远端异步干扰在时域、频域和空域上的分布较为复杂，干扰测量的精度与测量开销均面临严峻挑战。

无线接入网（RAN）数字孪生作为一种新兴技术，为解

决上述问题提供了可行路径。该技术借助物理环境重构，结合射线追踪与统计信道建模，并融合实际测量数据，可有效获取网络系统中的CSI[5, 17]。针对广域无蜂窝组网中的异步干扰问题，我们基于RAN数字孪生技术，提出了一种以用户为中心的同步协作传输与异步干扰抑制联合优化方法，旨在提升网络的整体性能。

如图3所示的无蜂窝网络系统架构[10]，主要由边缘分布式单元、vCPU以及云化集中式单元（Cloud-CU）构成。基于用户与TRP的空间几何位置关系及信道的数字孪生建模，通过分步优化方式，可实现用户与TRP的关联、用户配对、调度决策以及时频资源分配等功能。对于与终端实现同步的TRP（即UE与TRP之间的最大多径时延小于CP），可采用上下行协作传输方式消除用户间干扰。借助信道的数字孪生，可进一步获取无蜂窝网络中由异步问题引发的干扰分布信息。在此基础上，通过协作收发机设计与异步干扰重构的迭代优化，可形成优化的传输方案及网络级无线资源分配策略[18]。

然而，由于协作预编码与远端干扰之间存在相互耦合关系，且与多用户配对、用户-TRP关联、用户调度等多个因素密切相关，无蜂窝网络整体性能优化的维度和复杂度均较高。此外，数字孪生建模的精确性依赖于实体基站的测量数据进行修正，而远端干扰信道的测量所需的参考信号开销较大，且往往需要终端侧参与反馈。因此，有必要借助AI技术，系统性地研究基于信道数字孪生的建模方法，以及基于孪生模型的多用户无蜂窝网络优化实现方案。



CU：集中单元　MAC：媒体接入控制　PHY：物理层　TRP：收发点　vCPU：虚拟化中心处理单元

图3　信道数字孪生增强的无蜂窝网络系统

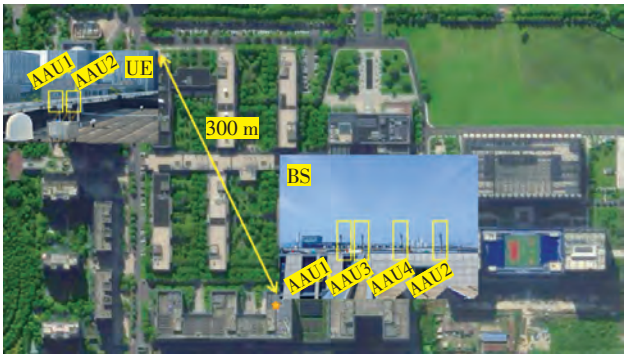
2.3 高频段无蜂窝大规模MIMO

研究者普遍认为，将分布式协作的无蜂窝技术进一步扩展至高频段，是实现高鲁棒性与高性能毫米波/太赫兹组网的关键路径[19]。然而，在高频段实现下行CJT对参与协作的TRP之间的时频同步提出了更为严格的要求。对于采用模拟预编码的毫米波系统，由于模拟通道的校准存在一定误差，导致波束旁瓣的互易性较差。此外，受TRP物理位置部署因素的影响，多个TRP指向同一用户的波束之间可能不存在直达链路。上述因素共同制约了高频段下自校准机制的性能表现。

针对终端辅助的校准机制，当用户位于TRP的主瓣方向时，可获得较好的相干传输增益。为验证这一特性，图4展示了我们构建的毫米波相干协作传输试验系统，包含了4个毫米波TRP与2个距离约300 m的毫米波终端。系统工作在26 GHz频段，带宽为200 MHz。终端及基站侧有源天线单元（AAU）的每个数字通道均配置16单元相控阵，各模拟通道输出功率约为10 dBm。系统中所有基站侧AAU通过IEEE 1588v2精密时间协议（PTPv2）与SyncE实现同步。图5给出了两个终端分别对4个AAU分布式波束的校准系数测量结果（由于对其中一个波束进行了归一化处理，图中仅呈现其余3个波束的校准系数）。测试结果表明，通过超分辨率估计，通道时延差的估计精度可达皮秒量级。由于两个终端相距较近且均位于4个AAU波束的主瓣方向，两者估计出的校准系数差异较小。实测数据表明[20]，基于终端辅助的校准机制可实现约10 dB的CJT波束赋形增益（理论增益为12 dB）。然而需要注意的是，在室内环境下的近场场景中，当用户位于波束旁瓣时，若仅对数字通道进行校准，由校准引入的误差将导致相干联合传输

的空分复用增益受到显著影响。

随着毫米波频段频谱资源的日益紧张，为获取更大传输带宽，研究者将太赫兹通信视为6G频谱拓展的重要方向。当前业界广泛关注的太赫兹频段仍集中在300 GHz以下，如140 GHz、220 GHz和300 GHz等，因此毫米波无蜂窝系统中的若干关键技术仍可延续应用于太赫兹系统。与毫米波类似，在太赫兹组网覆盖场景下，若采用集中式大规模阵列，仍面临覆盖区域边缘用户速率较低、易受遮挡以及业务速率分布不均等问题。若沿用传统蜂窝组网方式，为克服路径损耗需部署密集小蜂窝，这将导致小区间干扰制约系统容量的持续提升。同时，密集部署场景下频繁的波束与小区间切换可能引发链路稳定性问题。从5G毫米波的发展经验来看，支持基站与终端之间动态波束通信的相控阵技术仍是太赫兹技术应用于6G的主要技术途径。受限于当前太赫兹相控阵芯片的发展水平，目前尚未出现成熟的太赫兹多用户RAN试验系统。随着器件技术的逐步成熟，将分布式协作传输进一步推广至太赫兹频段[21-22]，有望显著提升太赫兹系统的鲁棒性。然而，相较于毫米波频段，太赫兹载波频率更高，实现相干协作传输的难度将进一步加大。



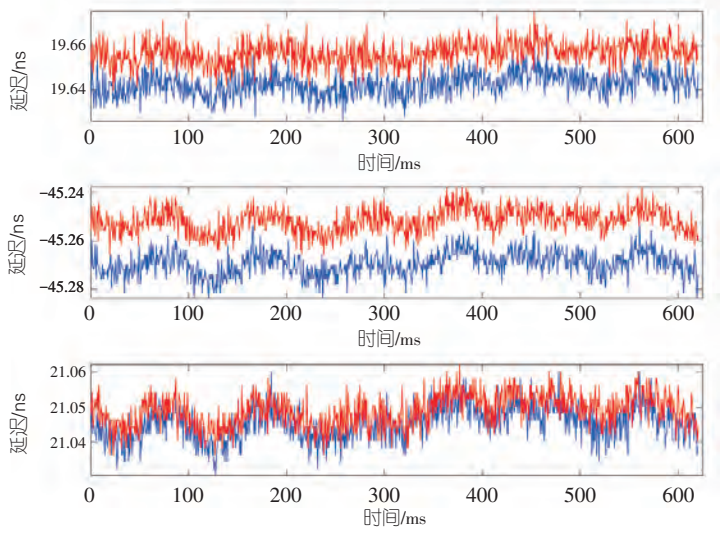
AAU：有源天线单元　BS：基站　UE：用户设备

图4　毫米波无蜂窝系统相干联合传输(CJT)试验场景

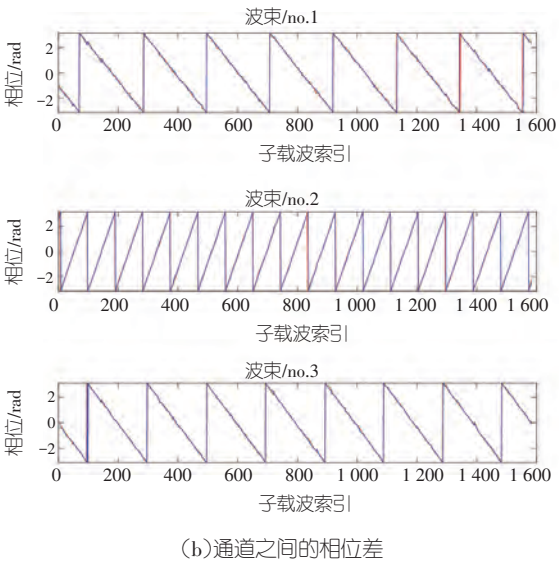
2.4 无蜂窝通感一体化技术

在无蜂窝无线接入网系统中，充分利用多个TRP之间的协作能力，可构建多站协作感知机制。相较于单站感知，多站协作感知不仅能够实现目标完整视图的重构，还能进一步提升感知精度[23]。然而，无蜂窝系统中的多站感知仍面临若干技术挑战，主要包括：协作感知中的信息融合问题、mTRP协作感知中非理想时频偏的补偿问题、TRP双工模式与通感一体化模式的选取、感知与通信之间的干扰管理与消除，以及如何将感知能力与RAN数字孪生相结合以提升通信性能等问题。

在无蜂窝组网架构下，感知信息的协作融合主要分为信号级融合与数据级融合两个层次[24]。信号级融合方法直接对接收信号或信道状态信息进行加权融合，通过对多通道信号实施相干或非相干累积，并构建目标位置的联合似然函数，结合网格搜索实现高精度目标定位。然而，该方法需依赖原始I/Q回波数据进行处理，导致前传链路容量成为系统的主要瓶颈。数据级融合方法则通过在多个接收TRP侧提取目标的感知参数信息（如多普勒–距离图中的感兴趣区域或特征



(a)通道之间的时间差



(b)通道之间的相位差

图5　通道之间的时间差及相位差

向量），再进行协作融合以获得感知结果。由于TRP端仅传输提取后的特征信息而非原始数据，因此可显著降低前传负载。该方法具有信号处理复杂度较低、系统结构相对简单的优势，但受信号处理环节间误差传递的影响，其定位精度存在一定局限。类似于无蜂窝系统中的分布式信号处理架构，无蜂窝通感一体化系统需要研究面向感知功能的处理功能切分方案。具体而言，需根据网络负载状况、目标密度及信道条件，动态调整感知处理任务在TRP与vCPU之间的分配，以实现信息融合效率与前传负载开销之间的平衡。

与mTRP相干CJT类似，mTRP感知同样面临TRP之间的时频偏问题。该问题会严重增加感知任务中的时延，降低多普勒估计精度，导致感知性能显著下降。现有研究表明，利用CJT的互易性校准机制可有效提升感知性能[25]。为此，需要研究利用TRP之间的通信信号或终端反馈信号，实现无蜂窝网络的全网快速校准，以消除时偏与频偏的影响，支撑长时间相干感知与分布式协作感知。同时，还需设计高效的相位跟踪机制，以降低校准开销，并抑制相位抖动对微弱信号相干累积的不利影响。

在无蜂窝无线接入网中，用户关联与TRP双工模式配置是实现高效协同的基础[26]。由于每个TRP的收发模式及通感状态均可动态调整，传统的“以用户为中心”的关联策略在引入感知任务后面临新的挑战。通信链路通常以信道增益最大化为目标，而感知任务则要求所选的TRP集合能够提供更优的几何观测角度。考虑到用户间干扰及散射体分布特性的影响，两者在TRP选择需求上往往存在冲突。因此，有必要引入图论、博弈论或多目标优化方法，在离散的节点关联空间内对通信容量与感知精度进行平衡，从而在相互竞争的目标之间寻求帕累托最优的拓扑结构方案。

在无蜂窝系统中，由于TRP在物理空间上呈密集部署且共享频谱资源，干扰模式已从传统的基站间干扰演变为复杂的多维交叉链路干扰。具体而言，TRP与TRP之间、UE与TRP之间均存在多重干扰路径，主要包括通信信号对感知的干扰、感知信号对通信的干扰，以及通信与感知信号的上下行相互干扰[27]。因此，有必要针对无蜂窝通感一体化场景，系统研究干扰的表征方法、测量机制与协同抑制方案。在技术实现层面，需设计空间域波束零陷与多域资源协调策略，以消除通信下行信号与感知回波之间的相互制约，保障通信性能的稳定。同时，应充分利用无蜂窝系统的分布式接收特性，开发高效的干扰抑制与杂波抑制算法，确保系统在复杂干扰环境下仍具备可靠的感知能力。

感知能力与RAN数字孪生技术的深度融合，可为环境信道的高精度重构提供有力支撑[5]。相较于对小尺度目标的感知，环境重构更侧重于大尺寸准静态环境对象的建模。通过利用多个终端的上行信号及多站协作感知，可实现环境信息的重构，进而辅助通信过程优化。例如，在终端移动过程中，可提前建立备份的TRP无线链路，辅助终端进行上行定时调整，从而避免重新发起随机接入所带来的开销与延迟。此外，在传统高频段系统中，多用户下行波束赋形的实现通常依赖频繁的导频训练。借助协作感知技术，系统能够直接获取终端的物理位置与移动轨迹，并结合RAN数字孪生技术，可在不依赖导频的情况下实现波束赋形，从而显著降低导频开销。

3 结束语

随着6G标准化工作的推进，Day-1技术成为业界广泛关注的焦点。本文从无蜂窝大规模MIMO的视角，系统探讨了mTRP技术的演进路径。从架构层面看，当前5G R19版本已初步构建了较为完整的去蜂窝化协作传输技术框架。然而，6GR如何进一步有效融合无蜂窝大规模MIMO技术，仍需在多载波与SBFD技术的基础上，化繁为简，形成更加简洁高效的接入与传输方案。同时，需要从基站上层协议栈的实现角度出发，进一步突破以小区为中心的传统资源分配体系，以充分释放mTRP的容量潜力。当前的实验结果表明，高频段下的相干协作传输具备可行性。随着太赫兹器件技术的逐步成熟，mTRP有望在高频段组网中得到实际应用，从而大幅提升高频段的可用性。此外，无蜂窝技术对6GR中的其他关键技术亦具有重要支撑作用，包括SBFD与ISAC等。随着信道数字孪生技术的持续发展，经过优化设计的无蜂窝系统将能够更有效地抑制远端干扰，推动移动通信大范围组网性能的显著提升。

参考文献

[1] 尤肖虎, 王东明, 王江舟 . 分布式MIMO与无蜂窝移动通信 [M]. 北京: 科学出版社, 2019
[2] 3GPP TS 38.214 V19.1.0 Physical layer procedures for data [S]. 2025
[3] 3GPP TS 38.321 Medium Access Control (MAC) protocol specification [S]. 2025
[4]3GPP TSG-RAN WG1 Meeting #123 High-level views on 6GR [S]. 2025
[5] 3GPP TSG-RAN WG1 Meeting #123, R1-2508733 Overview of 6GR air interface: [S]. 2025
[6] 3GPP TSG-RAN WG1 Meeting #123, R1-2508453 Overview of 6GR air interface: [S]. 2025
[7] 3GPP TSG-RAN WG1 Meeting #123, R1-2508430 Overview of 6GR air interface [S]. 2025
[8] Ngo H Q, Ashikhmin A, Yang H, et al. Cell-free massive MIMO versus small cells [J]. IEEE transactions on wireless

communications, 2017, 16(3): 1834–1850. DOI: 10.1109/twc.2017.2655515
[9] Björnson E, Sanguinetti L. Scalable cell–free massive MIMO systems [J]. IEEE transactions on communications, 2020, 68(7): 4247–4261. DOI: 10.1109/TCOMM.2020.2987311
[10] Wang D M, You X H, Huang Y M, et al. Full–spectrum cell–free RAN for 6G systems: system design and experimental results [J]. Science China information sciences, 2023, 66(3): 130305. DOI: 10.1007/s11432–022–3664–x
[11] Fukui T, Yokomakura K. Investigation of performance of uplink multi–panel transmission for multi–TRP operation in 5G–advanced system [J]. IEICE transactions on communications, 2025, E108–B(12): 1391–1399. DOI: 10.23919/transcom. 2024 EBT0011
[12] Cao Y, Wang P, Zheng K, et al. Experimental performance evaluation of cell–free massive MIMO systems using COTS RRU with OTA reciprocity calibration and phase synchronization [J]. IEEE journal on selected areas in communications, 2023, 41 (6): 1620–1634. DOI: 10.1109/JSAC.2023.3276057
[13] 3GPP TSG–RAN WG1 Meeting #123, R1–2508614 Overview of 6GR air interface [S]. 2025
[14] 3GPP TSG–RAN WG1 Meeting #123, R1–2509229. Overview of 6GR air interface [S]. 2025
[15] Wang D M, Wang M H, Zhu P C, et al. Performance of network–assisted full–duplex for cell–free massive MIMO [J]. IEEE transactions on communications, 2020, 68(3): 1464–1478. DOI: 10.1109/tcomm.2019.2962158
[16] Zhu Y, Li J M, Zhu P C, et al. Optimization of duplex mode selection for network–assisted full–duplex cell–free massive MIMO systems [J]. IEEE communications letters, 2021, 25(11): 3649–3653. DOI: 10.1109/LCOMM.2021.3105918
[17] Cohen–Arazi K, Roe M, Hu Z, et al. NVIDIA AI aerial: AI–native wireless communications [PP/OL].arXiv(2025–10–02)[2026–01–04]. https://arxiv.org/abs/2510.01533
[18] Hong Z Y, Li T, Xu S, et al. Asynchronous centralized and distributed precoding for extensive cell–free OFDM with adaptive fronthaul overhead [J]. IEEE transactions on wireless communications, 2026, 25: 2312–2326. DOI: 10.1109/TWC.2025.3596109
[19] Zhu Z S, Wang L F, Wang X, et al. Spatial–spectral cell–free sub–terahertz networks: a large–scale case study [J]. IEEE transactions on wireless communications, 2025, 24(4): 2956–2967. DOI: 10.1109/TWC.2025.3526932
[20] Jiang Q J, Jin J, Wang Q X, et al. Experimental performance of bidirectional phase coherent transmission and sensing for mmWave cell–free massive MIMO systems with reciprocity calibration [J]. IEEE journal on selected areas in communications, 2026, 44: 3591–3607. DOI: 10.1109/JSAC.2026.3664828
[21] Callebaut G, Liu L, Eriksson L, et al. 6G radio testbeds: requirements, trends, and approaches [J]. IEEE microwave magazine, 2024, 25(4): 14–31. DOI: 10.1109/MMM.2024.3351970
[22] Ghasempour Y, Kludze1 A, Bodet D, et al. Distributed wavefront shaping in near–field sub–1 THz wireless networks [EB/OL]. [2026–01–05]. https://www. researchsquare. com/article/rs–5898192/v1
[23] Liu G Y, Xi R Y, Wang X Q, et al. Cooperative sensing for ISAC: challenges, system design, beam management, and performance validation [J]. IEEE journal on selected areas in communications, 2026, 44: 608–625. DOI: 10.1109/JSAC.2025.3611941
[24] Wei Z Q, Xu R Z, Feng Z Y, et al. Symbol–level integrated sensing and communication enabled multiple base stations cooperative sensing [J]. IEEE transactions on vehicular technology, 2024, 73(1): 724–738. DOI: 10.1109/TVT.2023.3304856
[25] Han K, Meng K T, Masouros C. Over–the–air time–frequency synchronization in distributed ISAC systems [PP/OL].arXiv[2026–01–05]. https://arxiv.org/abs/2503.08920
[26] Zeng F, Liu R Y, Sun X Y, et al. Multi–static ISAC based on network–assisted full–duplex cell–free networks: performance analysis and duplex mode optimization [J]. Science China information sciences, 2025, 68(5): 150303. DOI: 10.1007/s11432–024–4381–8
[27] Sun X Y, Li J M, Chen G H, et al. Interference management and joint precoding design for multi–static ISAC and full–duplex communication cell–free systems [J]. IEEE transactions on communications, 2025, 73(10): 9798–9814. DOI: 10.1109/TCOMM.2025.3564756

作 者 简 介

尤肖虎，中国科学院院士，东南大学首席教授、紫金山实验室首席科学家；目前主要研究方向为无线与移动通信系统、现代数字信号处理等；作为项目负责人，曾承担30余项国家“863”、科技攻关、国家自然科学基金等项目。

王东明，东南大学首席教授；主要研究方向为移动通信系统和无线传输技术等；先后主持和参加省部级项目40余项；已发表论文200余篇。

曹阳，紫金山实验室高级工程师；主要研究方向为6G无线传输技术及系统研发；作为课题负责人和项目骨干参与国家重大专项2项；已发表论文12篇，拥有受理和授权国家发明专利13项。

基于OFDM索引调制的通信定位方法

Communication and Positioning Method Based on OFDM Index Modulation

杨旭旭/Yang Xuxu，刘炳宏/Liu Binghong，
彭木根/Peng Mugen

（北京邮电大学网络与交换技术全国重点实验室，中国 北京，100876）
(State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI：10.12142/ZTETJ.202601007
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260224.1641.002
网络出版日期：2026-02-25
收稿日期：2025-12-26

摘要：针对全球导航卫星系统（GNSS）拒止环境下传统定位方法抗干扰能力弱、频谱利用率低及峰均功率比高等问题，提出一种融合惯导与正交频分复用（OFDM）网格编码索引调制的通信定位方案。该方法通过将激活子载波位置作为索引传递基站坐标等导航辅助信息，结合网格编码规则实现导航辅助信息的高效嵌入与解码，同时保留部分子载波用于测距。接收端融合惯性测量单元（IMU）推算与多基站测距结果，采用扩展卡尔曼滤波实现协同定位。仿真结果表明，在多径衰落信道下，所提方案在保障通信性能的同时，定位精度显著优于传统方法，尤其在低信噪比区域具有更强的鲁棒性。

关键词：通信定位一体化；惯性导航；融合定位；索引调制

Abstract: To address the weaknesses of traditional positioning methods in global navigation satellite system (GNSS)-denied environments, including poor anti-jamming capability, low spectrum efficiency, and high peak-to-average power ratio, this paper proposes a communication-positioning solution integrating inertial navigation with orthogonal frequency division multiplexing (OFDM) grid-coded index modulation. By using activated subcarrier positions as indices to transmit navigation-aided information such as base station coordinates, combined with grid-coding rules, this method achieves efficient embedding and decoding of navigation data while reserving some subcarriers for ranging. The receiver integrates inertial measurement unit (IMU)-derived estimates with multi-base station ranging results, employing extended Kalman filtering for cooperative positioning. Simulation results demonstrate that under multipath fading channels, the proposed scheme significantly outperforms conventional methods in positioning accuracy while maintaining communication performance, exhibiting enhanced robustness particularly in low signal-to-noise ratio regions.

Keywords: integrated communication and positioning; inertial navigation; fusion positioning; index modulation

引用格式：杨旭旭，刘炳宏，彭木根．基于OFDM索引调制的通信定位方法 [J]. 中兴通讯技术，2026, 32(1): 38-45. DOI: 10.12142/ZTETJ.202601007

Citation: Yang X X, Liu B H, Peng M G. Communication and positioning method based on OFDM index modulation [J]. ZTE technology journal, 2026, 32(1): 38-45. DOI: 10.12142/ZTETJ.202601007

随着5G的商用部署，移动通信系统的定位能力正在从传统的辅助性服务向核心网络功能转变。面向6G时代的无人系统、应急通信、低空通航等新兴应用需求，定位技术需要在精度、可靠性、覆盖范围等方面实现显著提升，以适应更加多样化和苛刻的应用环境。

现有的定位解决方案包括全球导航卫星系统（GNSS）和第3代合作伙伴计划（3GPP）定义的基于网络的方法[1]。GNSS广泛应用于民用、工业和军事等多个领域，但其信号易受遮挡与干扰等因素影响，在城市峡谷、隧道、室内及低空飞行等典型非视距环境中，常常出现定位精度下降甚至完全失效的问题，难以满足关键场景下对连续、可靠、高精度定位的需求。近年来，基于5G新空口（5G NR）的蜂窝定位成为研究热点，其在室内外环境中均能兼顾精度与覆盖范围，尤其适用于移动目标定位[2]。

3GPP在第16版标准TS38.855中明确规定了多种5G定位方法，包括下行链路到达时差（DL-TDOA）、上行链路到

基金项目：国家自然科学基金项目（62401073）；中国通信学会青年人才托举项目（2024-2026QNRC001）

达时差（UL-TDOA）、多次往返时间（Multi-RTT）、下行链路离开角（DL-AOD）、上行链路到达角（UL-AOA）以及增强型小区识别（E-CID）[3]。其中，时延与角度信息被视为5G定位中最核心的信号特征。TDOA方法具备较高精度，但严重依赖多个基站之间的亚纳秒级时间同步[4]，Multi-RTT方法则可以在无同步场景下实现测距，但定位精度依赖可观测基站数量[5]，文献[5]将多天线的AOD能力集成到Multi-RTT定位中，建立了5G RTT/AOD定位模型，能够解决基站数量小于3个场景下采用RTT定位效率低的问题。结果表明，AOD的加入使水平与垂直精度分别提升约25%和65%。然而，该研究未考虑信号的非视距和多径传播的问题，且AOD测量对天线校准要求较高，部署成本高。3GPP Release 18引入了对侧链（SL）定位的增强支持，扩展了自Release 14以来SL技术的应用范围。SL定位允许通过SL接口传输的定位参考信号（SL PRS）实现目标用户设备（UE）的测距与测角，可根据绝对位置、相对位置或范围信息确定UE的位置。该方法可在覆盖范围内、覆盖外及部分覆盖场景下工作，具备良好的灵活性与环境适应性[6]，但在实际应用中仍面临诸多挑战，例如可用于传输SL PRS的带宽资源有限，以及定位锚点的可用性难以保障等问题。文献[7]评估了影响二维绝对定位精度的关键因素。结果表明，要实现90%用户设备达到1 m精度至少需要100 MHz带宽。尽管增加锚点数量有助于提升定位精度，但其随着数量增加逐渐趋于饱和，同时仍受限于锚点间的同步误差与非视距传播等因素。

尽管TDOA、AOA和SL等无线定位技术在多种场景中具备较高精度，但其本质上仍依赖于外部信号的时频特性与质量。在无外部传感器辅助下，当通信链路质量恶化或定位参考信号出现中断时，这些方法难以实现连续稳定的轨迹估计。为提升导航定位的自主性与连续性，惯性测量单元（IMU）被广泛用于短时间无外部定位信号下的导航补偿。但IMU的累积漂移问题不可忽视，若缺乏准确的外部定位修正，其误差将迅速扩散，对最终融合精度构成威胁。近年来，惯导系统与其他外部定位源融合定位获得高度关注。文献[8]指出，视觉、超宽带（UWB）与IMU等混合感知是实现连续室外无人机定位的关键路径。文献[9]利用动态协方差估计结合IMU预积分与滑动窗口因子图优化实现UWB与IMU融合定位，在复杂室内环境中精度较传统方法提升38%，动态条件下均方根误差达12.3 cm。文献[10]对基于误差状态卡尔曼滤波器与图优化方法的5G到达时间测距数据与惯性传感器数据融合方案进行了对比分析，验证了其在微型飞行器室内精确定位中的应用潜力。文献[11]依靠正交频分复用（OFDM）信号进行无人机间通信测距，实现无人机集群内部高精度相对测距，修正惯导定位误差。

鉴于传统网络定位在同步依赖、复杂环境及连续估计等方面的不足，本文面向终端自主定位需求，提出一种融合惯导与OFDM网格编码索引调制的通信定位方案。地面基站通过组合数索引映射方式将自身坐标编码为网格编号，并嵌入至OFDM子载波的激活模式中进行广播。无人机终端接收信号后，首先进行能量检测解码出基站坐标信息，随后基于TOA测距原理，利用OFDM符号中的测距序列获得测距值，将其作为观测量输入扩展卡尔曼滤波器（EKF）。在EKF框架下，结合基站位置信息与本地IMU提供的短时导航信息，可实现对无人机自身轨迹的连续、高精度估计。

1 信号模型与定位流程

融合通信定位系统的总体架构如图1所示，发射端为地面基站，接收端为无人机。一体化信号由测距序列、通信数据以及用于辅助定位的系统信息共同构成。利用OFDM子载波的正交特性，将测距序列与通信数据分别映射到不同的正交子载波上，实现通信与测距的并行传输。

接收端集成惯性导航单元，通过测距序列相关检测获得距离观测值，并解调通信数据恢复基站位置信息。系统将距离测量值、基站坐标与惯导数据输入扩展卡尔曼滤波器进行融合处理，实现高精度实时定位。

2 基于范围的常用定位算法

本节将介绍几种常用的定位算法，该类算法通过测量信号的物理属性（如时间、角度）来估计距离或位置，通常需要至少3个参考点来实现二维定位。

2.1 TOA

TOA测量信号从UE到基站的传播时间，结合多基站距离数据进行三边定位。假设信号以光速c（约3×10^8 m/s）

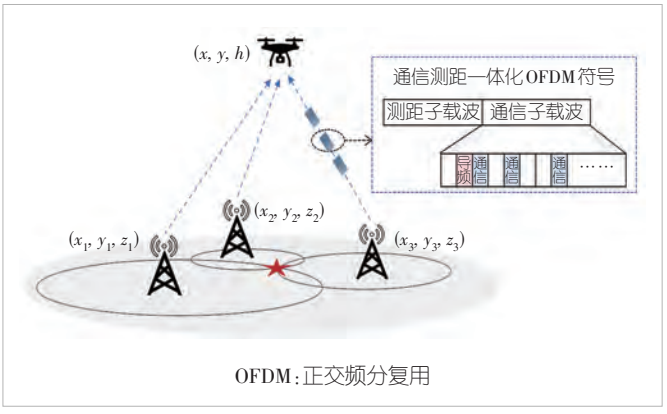


图1 通信定位融合系统架构图

传播，距离 $d_i = c \times t_i$，其中 t_i 是传播时间。对于基站坐标 (x_i,y_i,z_i) 和距离 d_i，目标位置 (x,y,z) 满足公式（1）：

$$(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = {d_i}^2 \quad (1)。$$

该算法精度高，可以直接计算距离，但需要精确的时钟同步，易受多径传播和非视距影响。

2.2 TDOA

TDOA通过测量信号从UE到达多个基站的到达时间差来定位。在5G中，TDOA常使用UL-TDOA，即UE发送信号，基站接收并计算时间差。对于基站 i 和 j，时间差 $\Delta t_{ij} = t_i - t_j$，对应距离差 $d_{ij} = c \times \Delta t_{ij}$，通过双曲线定位求解公式（2）：

$$\sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} - \sqrt{(x - x_j)^2 + (y - y_j)^2 + (z - z_j)^2} = d_{ij} \quad (2)。$$

该算法优点是无需UE与基站时钟同步，但其计算复杂度高且需要基站间纳秒级高精度同步。

2.3 AOA

AOA通过基站的天线阵列测量信号的入射角度，结合多个基站的AOA数据确定位置。5G利用大规模多输入多输出（MIMO）技术，增强角度分辨率。对于基站 i 和角度 θ_i，位置关系满足公式（3）：

$$y - y_i = (x - x_i)\tan\theta_i \quad (3)。$$

该算法无需距离测量，但需要复杂天线阵列，多径效应会导致角度误差。

3 网格索引调制融合定位算法

尽管TOA、TDOA、AOA等传统几何定位算法在理论上具有较高的定位精度，但这些传统方法存在固有缺陷。一方面，基站坐标等辅助信息通常采用直接传输方式，需要消耗大量频谱资源；另一方面，算法均为单一信息源定位，高度依赖无线信号的传输质量，一旦出现信号中断、严重衰落或非视距传播等情况，定位功能将完全失效。

本节介绍所提出的融合定位算法，该算法基于TOA算法，通过构建扩展卡尔曼滤波框架，将测距观测量、网格索引调制隐式传递的基站位置信息以及惯导数据进行融合，其框图如图2所示。该方案在保持通信功能正常运行的前提下，实现复杂环境下的高精度实时定位。

3.1 网格索引映射方法

由于OFDM系统存在峰均功率比（PAPR）高和对多普勒频移敏感的问题，本系统采用基于OFDM的索引调制技术，在激活的子载波上传输通信数据，同时利用激活子载波

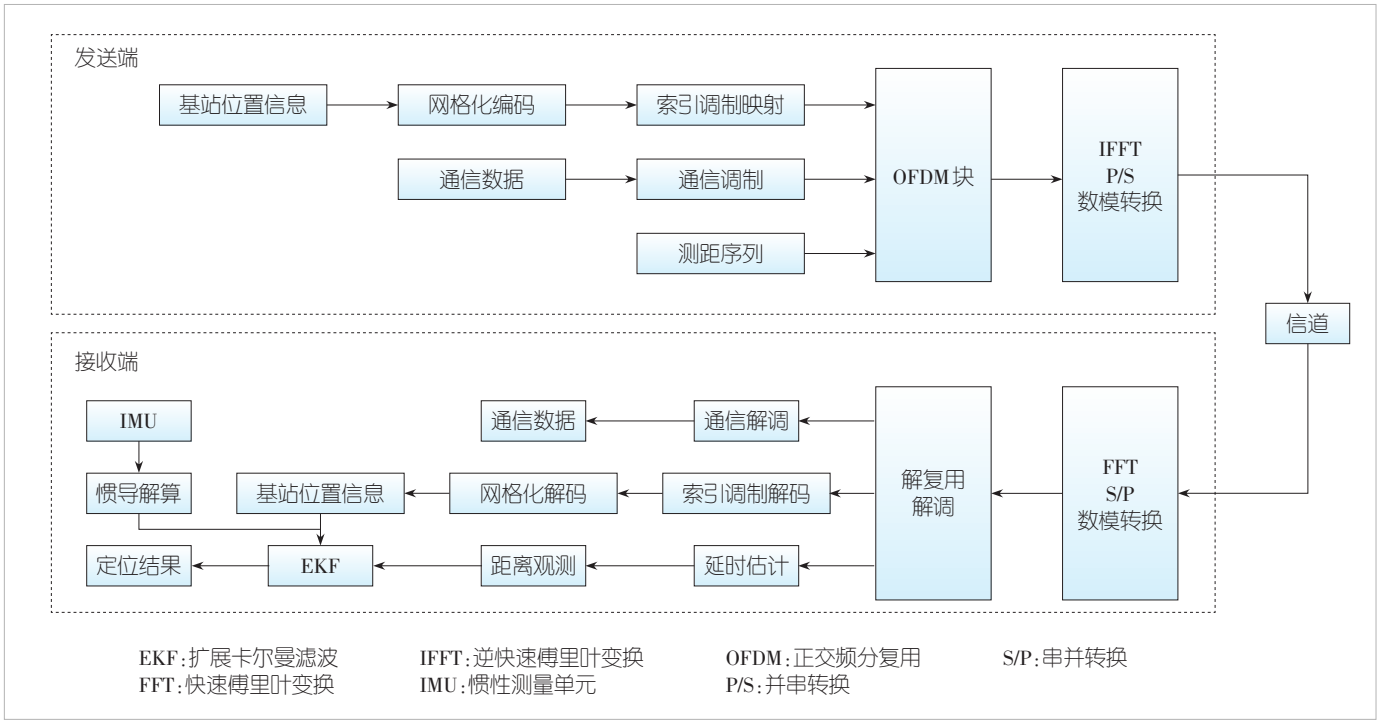


图2　基于网格索引调制的通信定位融合框图

的索引组合传输基站位置坐标等导航辅助信息。此外，为抑制多普勒效应引起的载波频率偏移，本系统在每个子块的激活子载波中固定首个子载波用于导频传输，由于频偏会导致接收信号的子载波间存在相位旋转，因此可利用导频子载波在频域上的相位变化进行频偏估计与补偿。

图3为网格编码索引映射示意图，将所有可用子载波按预设规则划分为多个子块，每个子块仅部分子载波被激活用于传输通信数据，其余非激活子载波保持静默状态。为了实现基站位置的高效隐式传递，系统采用预配置网格映射机制，在覆盖区域内预先建立规则网格划分，网格分辨率根据定位精度按需确定。设每个子块包含n个子载波，其中激活k个子载波，则横纵坐标各自可表示$C(n,k)$种状态，共可表示$C(n,k)\times C(n,k)$个网格位置。根据基站的实际二维位置（固定高度），将其映射至对应的网格，并提取该网格的坐标编号。随后，将横纵坐标分别映射为激活子载波的位置索引，索引的确定采用组合数映射方法，从而完成基站位置信息的编码与传输，组合数映射算法流程见图4。

首先初始化剩余激活子载波数和当前搜索子载波起始位置，判断$C(p,r)$与要传递的十进制编号数Z的关系。若$C(p,r)>Z$，说明当前索引p处未被激活。令$p=p-1$，若$C(p,r)\leqslant Z$，说明当前位置p为激活位置。记录$I_r=p$，更新$Z=Z-C(p-1,r)$，$r=r-1$，$p=p-1$，直到$r>0$不满足，输出激活位置索引集合$\{I_1,I_2,\cdots,I_k\}$。

接收端根据能量检测识别的激活子载波位置，并根据编号与组合数映射关系$Z=C(I_1,1)+C(I_2,2)+\cdots+C(I_k,k)$，还原网格编号，进一步查找该编号在地图上的实际坐标，从而完成坐标解码。

3.2 融合定位算法

基站与无人机利用OFDM测距序列实现基于TOA的测距，估计信号的传播时延为：

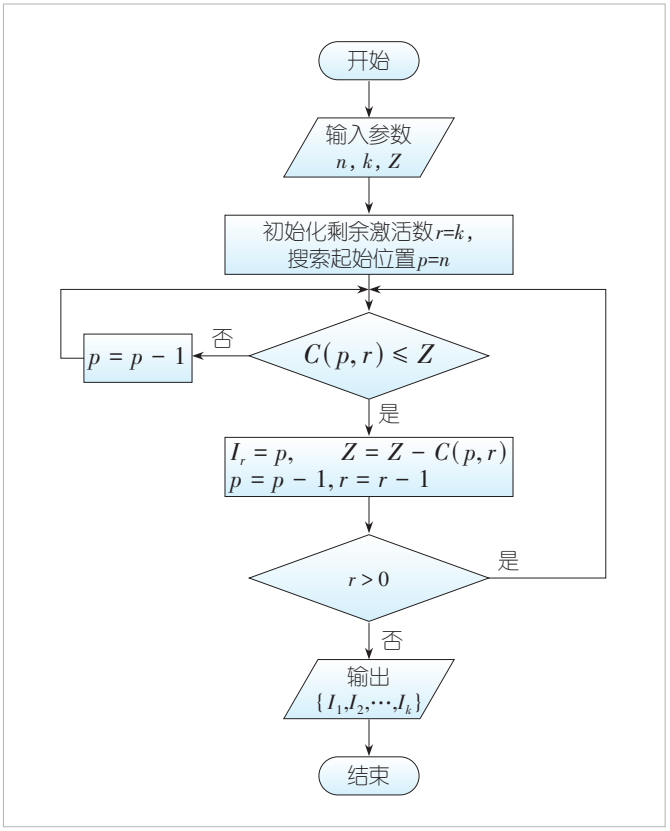


图4 组合数索引映射流程图

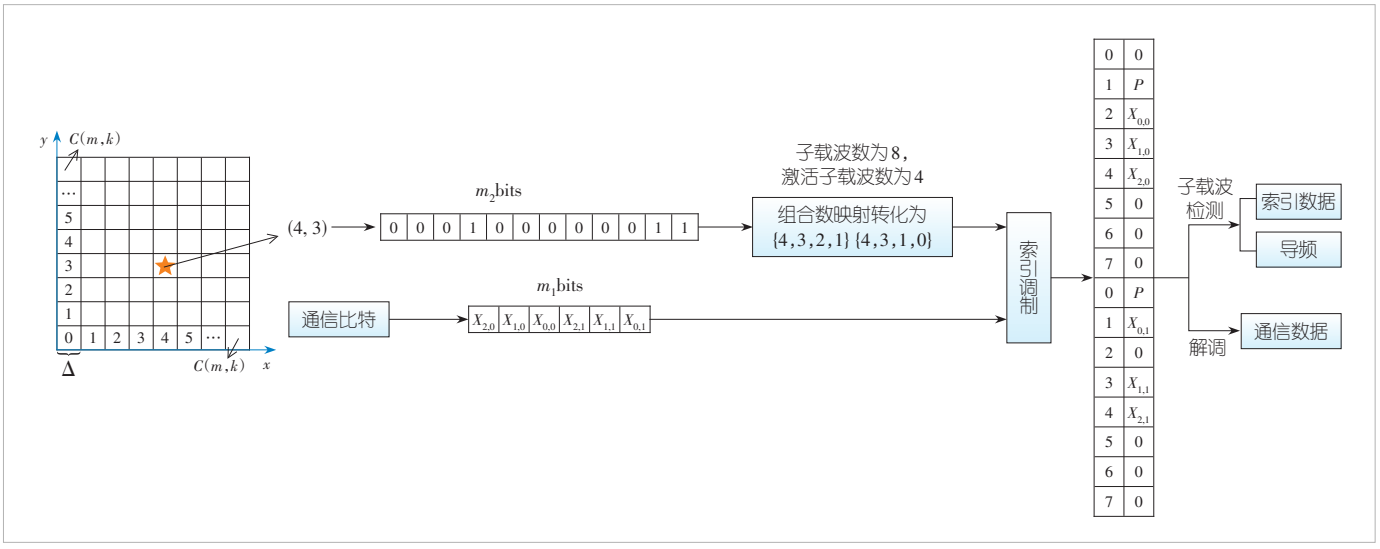


图3 网格编码索引映射示意图

$$\hat{\tau}_i = \frac{\arg\max_\tau \left| \int r_i(t) s^*(t-\tau) \mathrm{d}t \right|}{f_s} \tag{4}$$

其中，$r_i(t)$为接收到第i个基站的发射信号，$s(t)$为已知OFDM测距符号，f_s为采样率。

测距值可写成公式（5）的形式：

$$\hat{d}_i = c \cdot \hat{\tau}_i \tag{5}$$

其中，c为光速。该处理可同时获得多个基站的观测测距值，记为：

$$\boldsymbol{Y}_k^T = \left[\hat{d}_1 \ \hat{d}_2 \ \hat{d}_3 \right] \tag{6}$$。

无人机配备IMU，可连续输出加速度$\boldsymbol{a}_k = [a_{x,k}, a_{y,k}]^{\mathrm{T}}$及角速度$\boldsymbol{\omega}_k$。结合这些观测量可建立系统的状态预测模型，设状态向量为：

$$\boldsymbol{X}_{k-1} = [x_{k-1} \ y_{k-1} \ v_{x,k-1} \ v_{y,k-1} \ \psi_{k-1} \ \omega_{k-1}]^{\mathrm{T}} \tag{7}$$

其中，x_{k-1}、y_{k-1}为$k-1$时刻终端的二维平面位置，$v_{x,k-1}$、$v_{y,k-1}$为对应方向速度，ψ_{k-1}为航向角，ω_{k-1}为角速度。

在输入IMU加速度后，状态预测模型可表示为：

$$\begin{cases} x_k = x_{k-1} + v_{x,k-1}\Delta t \\ y_k = y_{k-1} + v_{y,k-1}\Delta t \\ v_{x,k} = v_{x,k-1} + (a_{x,k-1}\cos\psi_{k-1} - a_{y,k-1}\sin\psi_{k-1})\Delta t \\ v_{y,k} = v_{y,k-1} + (a_{x,k-1}\sin\psi_{k-1} + a_{y,k-1}\cos\psi_{k-1})\Delta t \\ \psi_k = \psi_{k-1} + \omega_{k-1}\Delta t \\ \omega_k = \omega_{k-1} \end{cases} \tag{8}$$

其中，加速度$(a_{x,k-1}, a_{y,k-1})$为机体坐标系下的测量值，而速度$(v_{x,k-1}, v_{y,k-1})$通常定义在导航坐标系中，因此需要通过航向角ψ_{k-1}将加速度从机体坐标系转换到导航坐标系中。

本系统采用EKF实现IMU预测信息与OFDM测距信息的融合定位。EKF包含预测与更新两个步骤。预测步骤中，利用状态转移方程：

$$\hat{\boldsymbol{X}}_k^- = \boldsymbol{F}_k \cdot \boldsymbol{X}_{k-1} + \boldsymbol{B}_k \cdot \boldsymbol{u}_{k-1} \tag{9}$$

$$\boldsymbol{P}_k^- = \boldsymbol{F}_k \boldsymbol{P}_{k-1} \boldsymbol{F}_k^T + \boldsymbol{Q}_k \tag{10}$$

其中，$\boldsymbol{F}_k$为状态转移矩阵，$\boldsymbol{B}_k$为控制矩阵，$\boldsymbol{u}_{k-1}$为控制输入向量，$\boldsymbol{Q}_k$为过程噪声协方差矩阵。

当接收到多基站OFDM测距观测后，更新步骤为：

$$\boldsymbol{K}_k = \boldsymbol{P}_k^- \boldsymbol{H}_k^T (\boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T + \boldsymbol{R}_k)^{-1} \tag{11}$$

$$\hat{\boldsymbol{X}}_k = \hat{\boldsymbol{X}}_k^- + \boldsymbol{K}_k \left(\boldsymbol{Y}_k - h(\hat{\boldsymbol{X}}_k^-) \right) \tag{12}$$

$$\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{H}_k) \boldsymbol{P}_k^- \tag{13}$$

其中，$\boldsymbol{Y}_k = \left[\hat{d}_1, \hat{d}_2, \hat{d}_3 \right]^{\mathrm{T}}$为当前时刻的测距观测量，$h(\cdot)$为非线性观测函数，$\boldsymbol{H}_k$为其雅可比矩阵，$\boldsymbol{R}_k$为测距观测噪声协方差矩阵。观测函数具体为：

$$h_i(\hat{\boldsymbol{X}}_k^-) = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2 + (z_0 - z_i)^2} \tag{14}$$

其中，(x_i, y_i, z_i)为第i个基站的三维坐标，z_0为接收端已知高度。

4 仿真与分析

本节基于MATLAB仿真平台构建了完整的系统仿真环境。仿真场景为典型城市密集区无人机配送环境，包括3个基站和1个无人机的场景。场景设定为城市建筑物密集分布区域，基站间距200～300 m。在信道建模中综合考虑路径损耗、阴影衰落和多径衰落复杂信道特性，以模拟复杂城市传播环境。路径损耗建模为$PL(d) = PL(d_0) + 10n\log_{10}(d/d_0) + X_\sigma$，其中参考距离$d_0 = 1$ m，$PL(d_0) = 32.4 + 20\log_{10}(f_c)$处的路径损耗（$f_c = 3.5$ GHz时$PL(d_0) = 43.3$），路径损耗指数$n = 3$反映城市环境的高衰减特性，阴影衰落X_σ服从零均值高斯分布，标准差$\sigma = 7.8$ dB，相比郊区环境（$\sigma = 4 \sim 6$ dB）更大，体现了复杂传播特性，多径数目为5，包含多条建筑物反射/散射径，信道冲激响应为$h(t,d) = \sqrt{10^{-\frac{PL(d)}{20}}} \times \sum_{i=1}^{L} \alpha_i \delta(t - \tau_i)$，其他仿真参数如表1所示。

4.1 通信性能分析

4.1.1 PAPR与频谱效率分析

在OFDM系统中，由于多个子载波在时域叠加，容易出现信号瞬时功率远大于平均功率的情况，这种功率起伏用PAPR来衡量。PAPR的定义如公式（15）所示：

表1 仿真参数与数值

仿真参数	数值
带宽	15.36 MHz
调制方式	QPSK
载波间隔	15 kHz
总子载波数	1 024
每个子块内子载波数	8
每个子块内激活子载波数	4

QPSK：正交相移键控

$$\mathrm{PAPR}=\frac{\max_{0\le n<N}\left|x[n]\right|^2}{\frac{1}{N}\sum_{n=0}^{N-1}\left|x[n]\right|^2}\quad(15)。$$

PAPR值越高，功率放大器具有的线性范围就越大。这会降低其工作效率，增加系统的功耗与成本。因此，在实际系统设计中，应尽量控制PAPR的大小，确保发射信号的功率特性在功放可承受范围内，提升系统能效。

在传统部分子载波激活方案中，未被激活的子载波通常置零，不传输任何有效信息，导致频谱利用率降低。索引调制通过将导航辅助信息映射到子载波激活模式中，使这些原本闲置的子载波位置也能承载信息，从而有效弥补了部分子载波未被激活所带来的频谱利用率损失。设每个子块包含n个子载波，其中有k个被激活用于传输通信数据，其余子载波置零，则可能的激活模式组合数为$C(n,k)=\frac{n!}{k!(n-k)!}$，由此可映射的索引比特数为$N_b=\lfloor\log_2 C(n,k)\rfloor$。这种方式在不增加带宽的前提下实现导航辅助信息嵌入，从而提升了频谱利用效率。系统的频谱总效率可表示为：

$$\eta=\frac{k\log_2 M+\lfloor\log_2 C(n,k)\rfloor}{n}\quad(16)，$$

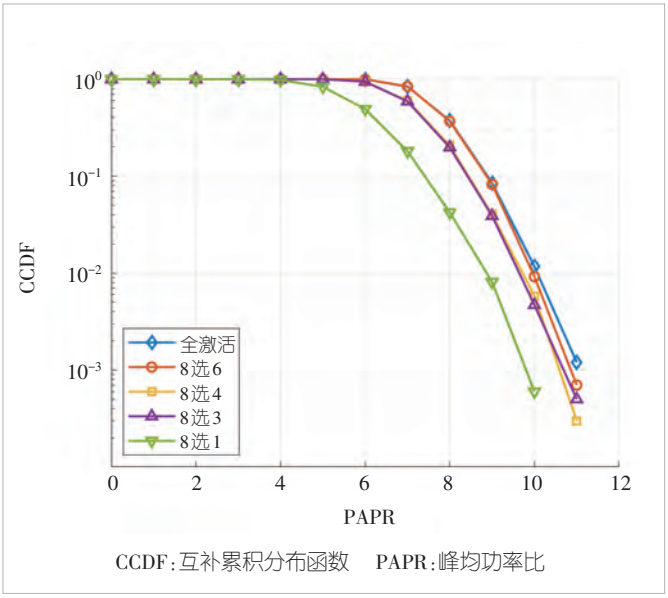
其中，M为通信子载波的调制阶数。

本文进一步分析了激活子载波数量对PAPR的影响。选取每个子块内子载波数$n=8$，改变激活子载波数量$k=1,3,4,6$进行仿真，绘制其PAPR的互补累积分布函数（CCDF）曲线如图5所示。可以看出，激活子载波数量越少，其PAPR值越低。

为综合评估所提方案中采用的OFDM_IM技术的频谱效率（SE）和PAPR，本文构建如下综合性能指标：

$$J=\alpha\cdot\frac{SE}{SE_{\max}}+\beta\cdot\left(1-\frac{PAPR}{PAPR_{\max}}\right)\quad(17)，$$

其中，权重系数$\alpha+\beta=1$，$SE_{\max}$和$PAPR_{\max}$分别为频谱效率和PAPR的归一化基准。考虑到无人机的功耗敏感特性，设置$\alpha=0.2$，$\beta=0.8$。归一化基准采用传统OFDM的性能参数，调制方式为QPSK。基于PAPR的CCDF仿真结果，在$CCDF=10^{-2}$处提取各系统性能参数，由公式（16）计算得出$SE_{\max}=2.0\ \mathrm{bit\cdot s^{-1}\cdot Hz^{-1}}$，由图4得到$PAPR_{\max}=10.2\mathrm{dB}$。计算的综合性能指标如表2所示。



CCDF：互补累积分布函数　PAPR：峰均功率比

图5　不同激活子载波数目下PAPR的CCDF曲线

结果显示，适当的OFDM索引调制激活方案能够提升系统的综合性能。在所测试的方案中，OFDM_IM（8选6）相比传统OFDM性能提升12%，但其组合数只有28，这意味着在相同的范围内，其网格精度更低。而OFDM_IM（8选4）相比传统OFDM实现了11.0%的性能提升，且组合数最多，在相同的范围内网格精度更高。索引调制的关键在于找到合适的子载波激活比例，既不能过度追求频谱效率而忽略PAPR性能，也不能过分降低激活子载波数量导致频谱利用率严重下降。

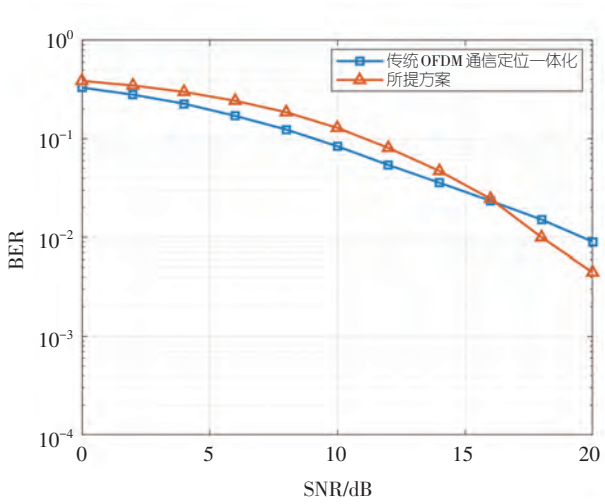
4.1.2 误比特率分析

图6展示了所提出的网格索引调制的OFDM通导一体化方案与传统OFDM通导一体化方案的误码率性能对比。从仿

表2　不同子载波激活方案的综合性能对比

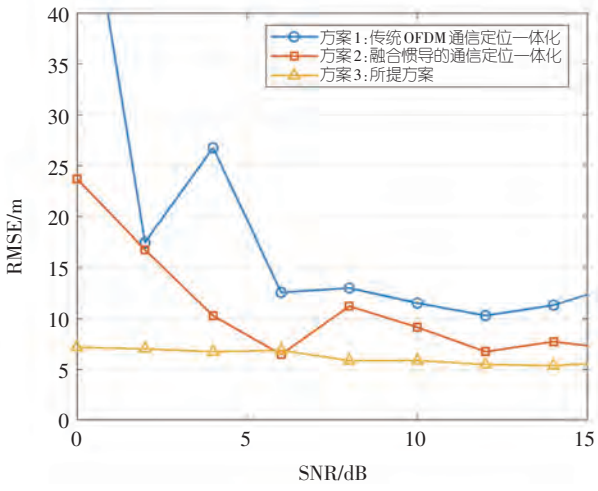
子载波激活方案	组合数	SE/(bit•s^{-1}•Hz^{-1})	PAPR/dB	综合指标J	性能提升/%
传统OFDM	—	2.000	10.2	0.200	—
OFDM_IM(8选6)	28	2.000	9.9	0.224	12.0
OFDM_IM(8选4)	70	1.750	9.6	0.222	11.0
OFDM_IM(8选3)	56	1.375	9.5	0.192	−4.0
OFDM_IM(8选1)	8	0.625	8.8	0.172	−14.0

IM：索引调制　OFDM：正交频分复用　PAPR：峰均功率比　SE：频谱效率



BER：误码率　OFDM：正交频分复用　SNR：信噪比

图6　通信性能仿真比较



OFDM：正交频分复用　RMSE：均方根误差　SNR：信噪比

图7　定位性能仿真比较

真结果可以看出，两种方案的BER性能都随着信噪比的增加呈现指数级下降趋势。在低SNR区域，所提方案略高于传统OFDM方案。这是因为所提方案需要同时进行索引检测和数据解调两个过程。在低信噪比条件下，噪声功率较大导致索引子载波识别容易出现错误。索引检测错误会影响后续的数据解调过程，从而导致BER性能的轻微恶化。

随着信噪比的提升，所提方案逐渐显现出性能优势。这主要归因于索引检测准确性的显著提高。能量检测算法能够准确识别激活子载波的位置。所提方案在信噪比为20 dB处，误码率性能提升约3 dB。

4.2 定位性能分析

图7展示了3种不同方案的定位RMSE随信噪比变化的性能对比。其中，方案1为传统OFDM通信定位一体化，方案2为融合惯导的通信定位一体化，方案3为本文所提出的网格索引调制OFDM通信定位一体化。方案2与方案3的主要区别在于基站坐标传输方式：方案2采用传统直接调制传输，易受噪声干扰导致坐标信息丢失；方案3通过网格编码索引调制隐式嵌入坐标，减少子载波占用并提升低SNR下的解码准确性。

从仿真结果观察到，在低SNR区域，方案1的RMSE表现出较大的波动性，这种不稳定性主要源于传统方案在恶劣信道条件下测距精度的显著恶化；方案2通过融合惯性导航信息，定位误差基本维持在20 m以内，但仍存在较大的波动；相比之下，本文所提出的方案3在整个SNR范围内都表现出优异且稳定的定位性能。这种显著的性能优势主要归因于以下几个方面：首先，索引调制通过激活子载波模式直接传输基站坐标信息，避免了传统方案中基站位置信息传输的不确定性；其次，即使在低SNR条件下出现部分索引检测错误，由于坐标信息的网格编码索引特性，仍能保证基站位置信息的基本准确性；最后，EKF融合算法能够有效平滑瞬时的测距误差，进一步提升定位精度的稳定性。

随着SNR的增加，3种方案的定位性能都逐渐改善并趋于收敛，在高SNR区域性能差异逐渐减小。这表明在良好的信道条件下，测距精度不再是限制定位性能的主要因素。此时系统性能更多地取决于几何精度稀释因子和算法的内在精度限制。值得注意的是，方案3不仅在低信噪比下具有显著优势，在整个SNR范围内都保持了稳定的性能，曲线波动最小。在SNR=14 dB下，相比方案1和方案2，方案3的性能分别提升6.4 dB和3.1 dB。这一结果验证了网格索引调制OFDM通导一体化方案的有效性，证明了该方案能够在保证通信质量的同时，显著提升系统的定位性能，特别是在恶劣信道环境下的定位鲁棒性。

5 结束语

本文提出了一种改进的融合通信定位方法。为提升基站坐标索引调制在噪声环境下的解码准确性，引入了网格编码索引映射机制，通过在基站服务区域内建立网格划分结构，将实际坐标离散映射为格点编号，并采用索引激活方式进行编码。该网格编码索引映射方案具有良好的扩展性，可灵活

传输设备姿态信息、环境参数、网络配置等多维度数据，通过合理的量化策略和自适应网格设计进一步降低系统误差并提升信息传输效率。基站位置坐标等导航辅助信息与终端惯导系统相结合，实现了终端的连续稳定定位。同时，该方法基于OFDM索引调制技术，通过减少子载波激活数量，有效降低了系统产生极高功率峰值的概率。仿真结果表明，所提设计在复杂环境尤其是低信噪比条件下仍可保持较低的定位误差，表现出良好的鲁棒性，为GNSS拒止环境下的高精度、连续性定位提供了一种有效解决方案。

参考文献

[1] Jornod G, Stark M. Positioning evolutions in 5G standardization: potential, solutions, and challenges of sidelink positioning for connected mobility [J]. IEEE vehicular technology magazine, 2025, 20(3): 106–114. DOI: 10.1109/MVT.2024.3519080

[2] Abuyaghi M, Si–Mohammed S, Shaker G, et al. Positioning in 5G networks: emerging techniques, use cases, and challenges [J]. IEEE Internet of Things journal, 2025, 12(2): 1408–1427. DOI: 10.1109/JIOT.2024.3487822

[3] Liu Y, Shi X F, He S B, et al. Prospective positioning architecture and technologies in 5G networks [J]. IEEE network, 2017, 31(6): 115–121. DOI: 10.1109/MNET.2017.1700066

[4] Camajori Tedeschini B, Brambilla M, Italiano L, et al. A feasibility study of 5G positioning with current cellular network deployment [J]. Scientific reports, 2023, 13: 15281. DOI: 10.1038/s41598–023–42426–1

[5] Guo W F, Deng Y, Guo C, et al. Performance improvement of 5G positioning utilizing multi–antenna angle measurements [J]. Satellite navigation, 2022, 3: 17. DOI: 10.1186/s43020–022–00078–y

[6] Le T K, Wagner S, Kaltenberger F. 5G sidelink positioning in 3GPP release 18 and release 19 [C]//Proceedings of IEEE Conference on Standards for Communications and Networking (CSCN). IEEE, 2023: 171–176. DOI: 10.1109/CSCN60443.2023.10453143

[7] Şahin T, Chiarello L, Michalopoulos D S, et al. Performance evaluation of 5G sidelink positioning [C]//Proceedings of IEEE Conference on Standards for Communications and Networking (CSCN). IEEE, 2023: 177–182. DOI: 10.1109/CSCN60443.2023.10453141

[8] Jarraya I, Al–Batati A, Kadri M B, et al. GNSS–denied unmanned aerial vehicle navigation: analyzing computational complexity, sensor fusion, and localization methodologies [J]. Satellite navigation, 2025, 6: 9. DOI: 10.1186/s43020–025–00162–z

[9] Zhang F Y, Li J, Zhang X Q, et al. Indoor fusion positioning based on “IMU–ultrasonic–UWB” and factor graph optimization method [EB/OL]. arXiv(2025–03–17) [2026–01–16]. https://arxiv. org/abs/2503.12726

[10] Kabiri M, Cimarelli C, Bavle H, et al. Graph–based vs. error state Kalman filter–based fusion of 5G and inertial data for MAV indoor pose estimation [J]. Journal of intelligent & robotic systems, 2024, 110(2): 87. DOI: 10.1007/s10846–024–02111–5

[11] 樊邦奎, 刘德康, 张瑞雨, 等. 大规模无人机集群通信定位一体化技术 [J]. 信号处理, 2024, 40(1): 7–16. DOI: 10.16798/j.issn.1003–0530.2024.01.001

作 者 简 介

杨旭旭，北京邮电大学网络与交换技术全国重点实验室在读硕士研究生；主要研究方向为通信导航一体化。

刘炳宏，北京邮电大学网络与交换技术全国重点实验室博士后；主要研究方向为低轨卫星通信和通信导航一体化。

彭木根，北京邮电大学副校长、教授，IEEE Fellow，中国电子学会会士，中国通信学会会士，连续多年入选科睿唯安ESI全球高被引科学家、爱思唯尔中国高被引学者、全球前2%顶尖科学家“终身科学影响力”和“年度科学影响力”榜单；长期致力于移动通信和卫星通信的组网基础理论、关键技术及工程应用研究；发表论文100余篇，Google学术引用2.4万余次。

6G内生智能与信道基础模型

6G Native AI and Channel Foundation Models

徐树公/Xu Shugong[1]，蒋骏/Jiang Jun[2]
（1. 西交利物浦大学，中国 苏州 215123；
2. 上海大学，中国 上海 200444）
(1. Xi'an Jiaotong-Liverpool University, Suzhou 215123, China;
2. Shanghai University, Shanghai 200444, China)

DOI：10.12142/ZTETJ.202601008
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260225.0923.002
网络出版日期：2026-02-25
收稿日期：2025-12-27

摘要：人工智能（AI）与通信系统的深度融合已成为6G的关键目标与核心标志之一，内生智能（Native AI）被普遍视为6G系统的重要特征。阐述了6G内生智能的内涵与需求，在此基础上系统梳理了无线通信领域AI研究范式的演进历程，揭示了基于监督学习的传统AI模型在支撑6G内生智能方面存在的固有局限。针对上述挑战，提出了信道基础模型（CFM）的概念框架，系统介绍了其预训练方法体系及面向各类信道相关任务的任务适配机制。认为6G内生智能需具备强大的任务适应性与场景泛化能力，而信道基础模型凭借其核心技术特征，有望成为未来6G内生智能的关键技术选项之一。

关键词：6G内生智能；信道基础模型；掩码信道重建；对比学习；通感一体化；自监督学习

Abstract: The deep integration of artificial intelligence (AI) and communication systems has emerged as a key objective and core hallmark of 6G, with native AI being widely recognized as an essential characteristic of 6G networks. An understanding of the connotation and requirements of 6G native intelligence is first elaborated. Based on this, the evolution of AI research paradigms in wireless communications is systematically reviewed, revealing the inherent limitations of traditional supervised learning-based AI models in supporting 6G native intelligence. In response to the above challenges, a conceptual framework of channel foundation model (CFM) is proposed, and its pre-training methodology as well as task adaptation mechanisms for various channel-related tasks are systematically introduced. It is envisioned that 6G native intelligence requires strong task adaptability and cross-scenario generalization capabilities, and channel foundation models, by virtue of their core technical features, are expected to become one of the key technological enablers for future 6G native intelligence.

Keywords: 6G native AI; channel foundation model; masked channel modeling; contrastive learning; integrated sensing and communication; self-supervised learning

引用格式：徐树公, 蒋骏. 6G内生智能与信道基础模型 [J]. 中兴通讯技术, 2026, 32(1): 46-52. DOI: 10.12142/ZTETJ.202601008
Citation: Xu S G, Jiang J. 6G native AI and channel foundation models [J]. ZTE technology journal, 2026, 32(1): 46-52. DOI: 10.12142/ZTETJ.202601008

将人工智能（AI）技术融入移动通信系统的各个环节，已成为学术界和产业界广泛关注的研究方向。随着国际电信联盟（ITU）将AI定义为6G网络的一项基础能力，“内生智能”（Native AI）被普遍视为6G系统的核心特征之一，标志着AI与通信系统的深度融合正成为6G发展的关键目标。然而，关于内生智能的确切内涵，以及6G系统究竟需要具备何种AI能力，目前仍缺乏广泛共识。本文旨在围绕上述问题，分享我们的理解与思考。

我们认为，内生智能指的是系统原生具备的AI能力。在6G设计中，这种能力应从系统架构的初始阶段即被纳入规划、设计、优化与交付流程，成为系统不可分割的基础组成部分。类比于汽车制造，内生智能可视为“前装组件”，即在出厂前已完成集成并随整车交付，而非用户购车后到4S店加装的“后装配件”。正因为内生智能难以通过后期“打补丁”方式实现，我们有必要深入探讨：究竟何种AI能力能够满足6G系统的长远需求，才有可能成为其核心组件，内嵌于系统底层。

当前，6G研究方兴未艾，关于AI的相关标准化工作也已启动。从应用场景角度看，6G需要支持地面、空基、卫星等多域融合通信，信道环境具有动态时变、异构干扰交织等复杂特征，涵盖信道估计、波束赋形、通感一体化等多种任务。若仅采用面向单一任务或特定场景的专用AI技术，将难以支撑系统在全域部署和功能扩展上的需求。缺乏任务适应性的内生智能，可能导致网络架构因专用模型种类繁多而变得臃肿；而缺乏跨场景泛化能力，则难以应对不同信道环境下性能的剧烈波动。这显然与“内生智能作为系统核心

组件”的定位相悖。

为此，本文将在梳理无线通信领域AI研究范式演进历程的基础上，系统分析传统AI模型在6G环境下面临的局限，进而提出一种面向信道的专用基础模型（FM）概念，探讨其面向多类下游任务的适配能力，并评估其作为6G内生智能核心支撑技术的可行性。

1 AI赋能无线通信的范式演进

回顾将人工智能技术应用于无线通信领域的研究[1−2]，其发展脉络与人工智能研究范式的演进基本一致，呈现出一条从早期单任务监督学习，到多任务协同建模，再到预训练基础模型驱动范式变革的清晰路径。在这一演进过程中，传统AI模型在支撑6G内生智能需求方面逐渐暴露出诸多难以克服的固有局限，使其难以成为未来6G内生智能的核心组成部分。

人工智能研究范式的演进大致可分为3个阶段，如图1所示。第1阶段为单任务监督学习阶段，研究聚焦于“一个任务、一个数据集、一个模型”的模式，模型主要采用端到端方式针对特定任务进行训练。这类模型的性能高度依赖于带标签数据的质量与数量，在数据集不足时易面临过拟合问题，且其跨任务迁移能力天然受限。随着应用场景复杂度的不断提升，研究者提出了包含共享参数的多任务学习框架，通过联合训练利用任务间的共性实现知识迁移[3−4]。然而，多任务学习仍然要求任务之间存在统计相似性，且依赖于任务专用的架构设计。FM的出现标志着深度学习范式的根本性变革，其“预训练+微调”的两阶段方法彻底革新了该领域的发展模式：首先基于大规模无监督数据集进行预训练，构建具备泛化能力的FM；随后通过下游任务的轻量级微调，高效适配多样化的任务需求与应用场景。在这一演进历程中，传统AI模型依赖大量带标签数据、泛化能力不足等固有局限愈发凸显。

传统基于监督学习的AI模型存在三大核心局限，严重制约了其在6G内生智能中的应用。

首先是数据依赖性困境。多数传统AI模型要实现可靠性能，必须依托海量带标签数据集作为支撑。然而在动态变化的无线环境中，传输模式与干扰状态时刻变化，稳定采集符合要求的数据集难度极高。更为关键的是，人工标注过程不仅耗时费力，还难以避免标注错误，进一步加剧了资源消耗。

其次是泛化能力不足。6G需要融合地面、海上、空中等多场景通信，对模型的跨场景适配能力提出了严苛要求。传统AI模型在训练过程中极易对训练数据产生过拟合，在训练场景中能够表现出良好性能，但一旦迁移至新场景，性能便会显著下降。

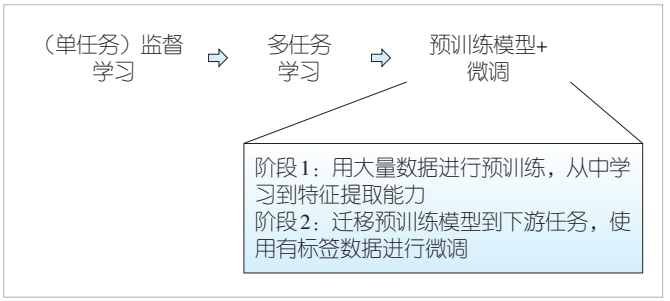


图1　AI研究范式的发展路线

再次是任务专用化困境。6G中的每个应用场景都需为传统AI模型进行定制化设计，导致任务专用模型数量激增。这不仅大幅增加了系统链路的复杂度与资源消耗，也与6G内生智能所追求的高效统一框架理念相悖。

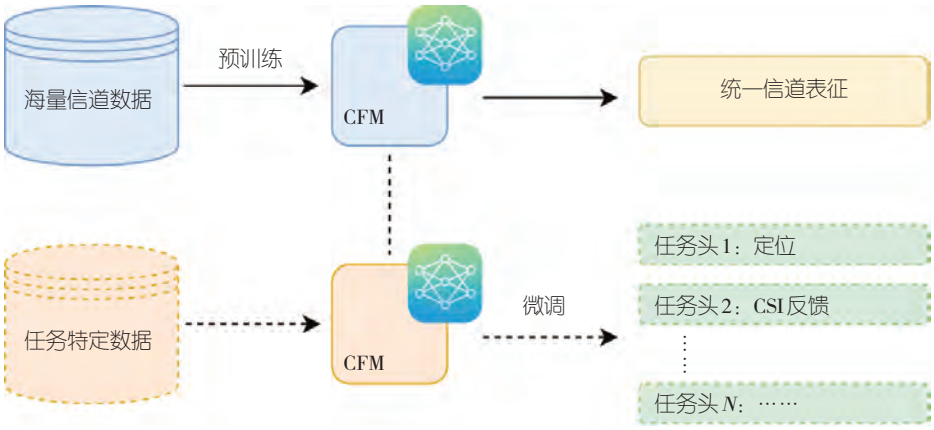
此外，传统AI模型还存在实时自适应能力薄弱的问题。无线通信环境动态多变，实时自适应能力是维持模型性能的关键。然而，在静态条件下训练的传统模型难以快速适应现实场景中的动态变化，对于自动驾驶、远程医疗等需要近实时响应的6G关键应用将产生不利影响。上述局限共同凸显了开发先进AI解决方案以满足下一代无线网络多维需求的必要性，也为信道基础模型（CFM）的提出奠定了研究基础。

2 从FM到CFM

受大语言模型与计算机视觉领域成功案例的启发，基础模型依托大规模数据集构建，仅需少量额外数据即可适配特定任务。如图2所示，与通用大语言模型的方法不同，无线通信领域的基础模型通过无线专用数据集进行预训练，显著提升了其在无线通信场景下的相关性与适配效率。

借鉴基础模型在相关领域的成功经验，本文面向无线通信场景提出CFM的概念。与信道外推、用户定位等需从零开始设计与训练的传统任务专用模型不同，CFM遵循“预训练–微调”的范式：首先，在涵盖多传播场景、多频段、多环境条件的大规模信道测量数据集上进行预训练，使模型能够提取通用信道特征并学习无线信道的内在统计规律；随后，针对特定的无线通信任务，利用小规模任务导向数据集对预训练CFM进行微调。该范式不仅加速了任务专用模型的开发进程，还提升了模型在不同无线环境下的泛化能力，为应对多样化动态信道条件下的模型适配挑战提供了有效解决方案。

凭借上述能力，CFM标志着“AI+无线”技术路线的范



CFM：信道基础模型　CSI：信道状态信息

图2　CFM的整体框架图

式转变，为应对6G内生智能的多维度挑战提供了一种兼具高效性与灵活性的可行解决方案。

2.1 CFM的定义

斯坦福大学Wiggins等人将FM定义为一种在大量原始数据基础上通过无监督学习训练而成的人工智能神经网络，具备适应多样化任务的能力[5]。CFM则是面向无线通信领域设计的专用基础模型，其核心聚焦于信道相关特征的挖掘与任务适配。需要明确的是，CFM与无线大模型存在本质差异：后者以“大语言模型、大参数量”为核心支撑，而CFM采用中等参数规模设计，旨在实现信道类任务的精准优化，在模型性能与工程部署效率之间取得平衡。从技术范式角度看，预训练模型体现了“预训练–微调”的通用方法论，CFM则是基于该范式、针对无线信道物理特性定制开发的具体模型实例，并非所有采用预训练范式的模型均可归入CFM范畴，例如通用视觉基础模型即不属此类。值得注意的是，尽管早期已有研究在无线信道领域尝试应用自监督等表征学习方法[6–7]，但这些工作多聚焦于单一任务，忽视了对信道通用特征的系统挖掘与跨任务迁移能力的构建，因而仍未突破传统任务专用模型的固有局限。

具体而言，CFM以无线信道为专属核心研究对象，基于大规模异构信道数据开展离线预训练构建而成，其输入输出空间及任务集合均围绕信道特性实现精准界定。在输入空间方面，模型明确以信道状态信息（CSI）和同相正交（IQ）信号为主要数据形态，涵盖多场景、多频段、多环境条件下的异构信道数据；输出空间则聚焦于通用信道特征，通过模型学习实现对信道底层物理规律的精准表征。任务集合严格限定为各类信道相关任务，主要涵盖两大维度：一是生成任务，包括信道外推、信道估计、信道压缩与反馈、预编码等；二是感知任务，包括场景分类、用户定位、波束管理等。CFM采用神经网络架构，其核心目标在于学习信道内在物理规律与跨场景通用特征，为上述多样化任务提供统一的特征支撑。这一设计理念与基于通用知识预训练的大语言模型以及传统有监督任务专用模型形成本质区别。

2.2 CFM与其他范式差异

AI赋能无线通信的主流范式主要包括任务专用模型、大语言模型（LLM）及CFM，三者核心特征对比见表1。通过指标差异可清晰印证CFM作为信道专用范式的必要性与优势。

在泛化能力与预训练需求方面，LLM的预训练并非必选项，其以通用知识为核心，未融入信道物理特性，跨场景泛化能力仅达到中等水平；CFM则将预训练作为必需环节，基于大规模异构信道数据学习信道底层规律，展现出优异的泛化能力。

在参数规模、存储与时延协同优化方面，LLM参数规模极大、推理时延较长，难以适配边缘部署与实时处理需求；CFM采用中等参数规模设计，存储开销与推理时延均控制在中等水平，通过轻量化微调即可适配具体任务，在模型性能与部署效率之间实现了有效平衡。

在任务适应性方面，LLM面对多样化信道任务需频繁进行微调，灵活性受限；CFM依托预训练阶段积累的信道专用特征，仅需微调少量参数即可高效适配多类信道任务，任务适应性显著更优。

表1　AI赋能无线通信的不同范式对比

内容	任务专用监督模型	LLM	CFM
泛化能力	较差	中等	优异
存储开销	低	高	中等
参数规模	小	极大	中等
预训练需求	无	可选	必须
推理时延	低	高	中等
任务适应性	低	中等	高

CFM：信道基础模型　LLM：大语言模型

综上所述，传统任务专用监督模型与LLM均难以均衡满足信道任务的多维度需求；CFM通过针对性的信道预训练与中等参数规模设计，构建了更契合信道特性的专用技术范式，有望成为支撑6G内生智能的重要技术选项之一。

2.3 CFM的核心特征

2.3.1 跨场景与配置的泛化性

传统AI模型在无线通信中的核心局限之一在于泛化能力薄弱，即模型在与训练环境不同的场景中部署时，性能显著下降。传统模型通常基于窄范围场景专用数据集训练，并针对单一任务优化，在面对未见过的场景时性能会出现“灾难性退化”。

在需覆盖地面、空中、卫星通信等多域异构场景的6G系统中，泛化挑战尤为严峻。不同传播域具有独特的信号传播特性，多样化的环境条件要求模型能够适配差异化的应用场景与天线配置。

CFM通过高效的跨场景泛化能力有效应对这一挑战。与传统模型不同，CFM的预训练数据集具有大规模异构特性，整合了多源信道数据：既包括基于信道模型生成的合成数据，也涵盖从不同场景采集的真实测量数据，以及模拟暴雨、高速移动等极端条件的仿真数据。这种多源数据融合使CFM能够学习无线信道的通用统计规律与跨场景共性特征，从而在多样化场景中保持稳定的性能表现。

2.3.2 对多下游任务的适应性

CFM作为通用信道特征提取器，具备多下游任务适配能力——通过轻量化微调即可无缝适配各类信道相关下游任务。这种适应性是基础模型设计的固有属性：通过在大规模多样化数据集上进行预训练，CFM学习到全面的通用特征，仅需少量数据即可高效迁移至特定任务。CFM的“少样本、低参数”适配机制对实际无线系统具有重要意义。与传统监督模型不同，CFM仅需少量带标签样本与局部参数调整，即可实现具有竞争力的性能，同时大幅降低计算开销。这一机制有效缓解了部署中的两大核心挑战：

1）数据依赖问题：无线通信领域带标签数据的采集过程耗时且资源消耗巨大，CFM所具备的少样本适配能力恰好契合了这一关键实际需求。

2）计算复杂度问题：低参数微调机制无须在部署阶段投入大量计算资源，使得CFM能够在处理能力受限的边缘设备上实现高效应用。

CFM凭借其卓越的任务适配能力，有望成为6G通信系统的核心组件，能够在无须构建多个专用模型的前提下，高效部署于各类信道相关任务中，为无线通信领域的技术演进提供重要支撑。

2.3.3 可扩展性

可扩展性是指模型性能随模型参数与训练数据集规模增加而稳定提升的特性，这一特性是CFM的核心特征，也是其与传统模型的关键区别所在。

CFM的可扩展性源于基础研究领域中已被实证的缩放定律：随着模型容量与训练数据规模的扩大，模型的表征能力与泛化性能呈现可预测的提升趋势。传统模型的性能通常在模型规模或数据量达到某一阈值后趋于饱和，而CFM则遵循完全不同的演进规律，其可扩展性主要体现在模型规模和训练数据两个维度：

1）模型规模可扩展性：增加CFM的参数数量可显著提升其捕捉细粒度信道动态的能力。大规模CFM擅长建模多干扰信号间的非线性交互，凭借更高的表征精度，能够为信道估计、预编码等关键任务提供更精准的输出。

2）训练数据可扩展性：当CFM在涵盖多运营场景、多设备类型、多环境条件的大规模异构数据集上进行训练时，能够学习到更全面的信道特征。这种全方位、多视角的数据使CFM得以掌握信道动态的底层物理规律，从而在所有下游任务中实现更优的泛化性能。

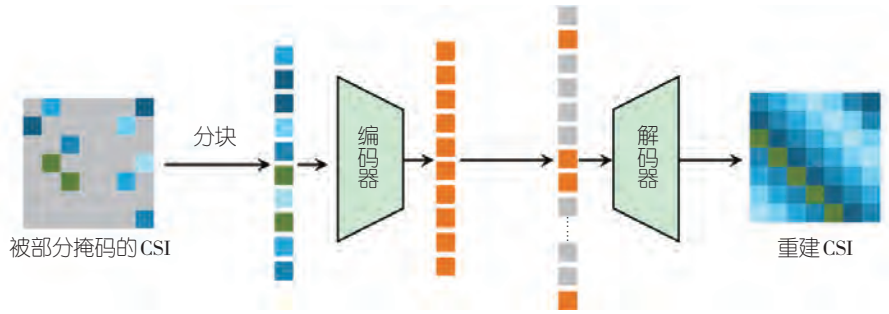
3 CFM预训练方法

尽管CFM的概念相对较新且仍处于早期发展阶段，但已有部分研究在方法设计和应用目标上与CFM的定义及目标高度契合，值得进行系统性梳理与总结。

为便于分析与对比，本文根据自监督预训练策略的差异，将CFM的预训练方法划分为以下3类：生成式方法、判别式方法，以及生成–判别混合方法。

3.1 生成式

生成式预训练方法的核心思想在于通过建模输入数据的底层分布，学习具备泛化能力的特征表示。这类方法通常以重构任务为训练目标，能够有效捕捉信道数据中的局部与全局特征，兼具较强的泛化能力与多模态适配性，是一种高效的预训练策略。其中最具代表性的是掩码重建方法，典型的流程如图3所示，以WiFo[8]和WirelessGPT[9]为例，该类方法通过随机掩码部分CSI块，迫使模型学习从残缺输入中重建完整信道特征的能力。掩码策略的设计确保了模型能够捕捉信道的全局结构与局部依赖关系，而编码器–解码器架构的采用则实现了从特征提取到信号重建的端到端优化。



CSI：信道状态信息

图3 生成式(掩码重建)的基本流程

3.2 判别式

判别式预训练方法的核心在于通过区分不同样本或构建正负样本对来学习判别性特征表示。与生成式方法不同，这类方法不依赖数据重构，而是通过衡量样本间的相似性实现训练目标，因此在部分下游任务中展现出更强的泛化能力与迁移性能。

对比学习作为判别式方法在CFM构建中的核心自监督手段，通过构造正负样本对，引导模型精准捕捉并区分不同样本，进而学习到更具表征能力的高维信道特征，典型的流程如图4所示。在现有研究中，CSI-CLIP[10]等模型创新性地利用信道数据的时域-频域对偶特性构建正样本对，以最大化同类样本相似度、最小化异类样本距离为优化目标，使模型能够自动捕捉信道的内在物理特性，从而在下游任务中实现优异的性能表现。

3.3 生成-判别混合

生成式方法与判别式方法并非互斥关系，反而能够形成优势互补。融合两种范式优势的混合方法，能够同时从数据中学习结构信息与判别性特征，从而有效提升模型的表达能力与泛化能力。该类方法综合了生成式与判别式预训练策略的各自优势，使模型能够在同一框架下同时学习丰富的结构特征与具有判别力的特征表示。混合方法在提升模型泛化能力方面展现出显著潜力。已有研究工作[11-12]表明，混合方法相比单一范式方法具有更强的性能表现。该方向有望成为CFM研究的重要演进路径。

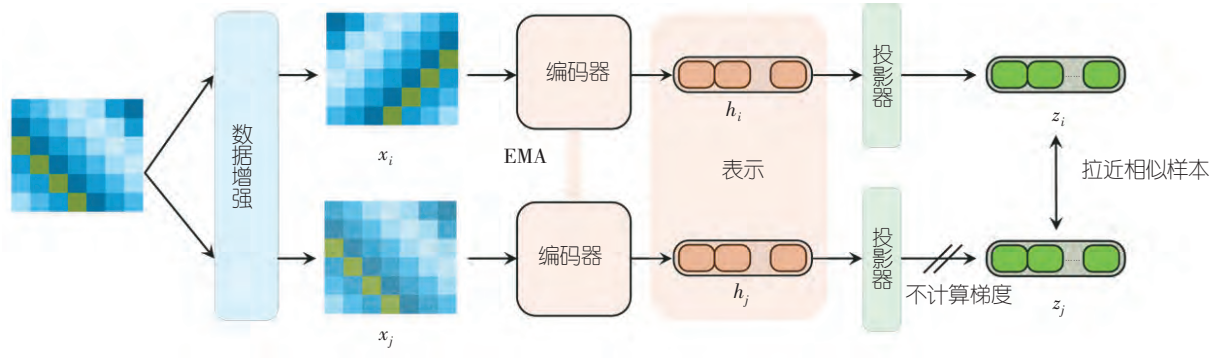
4 CFM的可能应用及6G内生智能

CFM凭借其优异的泛化性、适应性与可扩展性，已确立为6G内生智能的重要技术路径及AI赋能的核心载体，正在物理层、无线接入网及通感一体化三大关键场景中，全面驱动6G系统的智能化升级。

4.1 CFM赋能物理层

在无线通信物理层的研究与实际部署中，CFM凭借其强大的通用特征提取与迁移能力，已成为突破传统技术瓶颈的核心工具之一，在信道估计、信道反馈、波束赋形优化等诸多应用场景中展现出显著优势。

传统方法通常依赖特定场景下的大量带标签数据进行模型训练。然而，当面临信道环境突变或样本数据稀缺时，这类模型的估计精度会大幅下降，甚至出现失效情况。相比之下，CFM作为6G内生智能的一种实现路径，通过在涵盖多



EMA：指数移动平均

图4 判别式(对比学习)的基本流程

频段、多地形、多干扰条件的大规模多样化信道场景中进行预训练，能够学习到具有泛化能力的信道底层特征。将这些通用特征迁移至特定信道估计任务后，可有效降低模型对目标场景带标签数据的依赖，即便在数据有限或环境动态变化的情况下，仍能保持稳定的估计性能。

4.2 CFM赋能无线接入网

无线接入网（RAN）作为用户终端与核心网之间的关键接口，其资源调度效率与干扰控制能力直接决定了用户体验质量。CFM凭借其对复杂网络环境的建模与分析能力，正推动RAN核心任务的智能化升级，助力RAN从“被动优化”向“主动自适应”的范式转型。以波束管理为例，现有5G网络在人口密集城区部署时面临诸多挑战：用户密度高、建筑物遮挡严重、信道条件动态变化快。传统波束选择方法多依赖固定信道模型或简单信号强度检测，易因波束重叠引发小区间干扰，从而限制网络整体容量的提升。而CFM所支持的动态波束选择机制，则能够实现更精细化的优化调控。总体而言，CFM为6G内生智能的实现提供了核心技术支撑，推动无线接入网向具备自组织、自优化、自修复能力的高阶智能形态持续演进。

4.3 CFM赋能通感一体化

在通感一体化领域，通信与感知在信号处理目标与特征需求上存在本质差异：通信系统以降低误码率为核心目标，重点关注信号的调制解调特性；感知系统则以提升定位精度为核心诉求，高度依赖目标散射特性。二者在目标导向与特征侧重上的根本性差异，为技术融合带来了严峻挑战。

在此背景下，CFM凭借其对多模态信道特征的统一建模与协同优化能力，成为破解这一难题的变革性方案。以智能交通场景为例，基于CFM的融合处理框架能够同时提升车辆定位与通信性能。该框架通过挖掘多模态信道特征的内在关联性，构建统一的特征表示空间，有效打破了通信与感知技术之间的壁垒，为6G通感一体化网络的实际部署提供了可行的技术路径。

4.4 CFM有效性验证

为验证CFM的有效性、泛化能力及任务适应性，本文基于CSI-CLIP框架，选用Vision Transformer（ViT）作为编码器开展实验验证。同时，引入无预训练的ViT模型作为基线模型，以进行对比分析。

预训练数据基于DeepMIMO数据集构建，共包含超过70万条样本，覆盖35个典型通信场景。该数据集涵盖室内外多种环境类型，工作频段从Sub-6 GHz、毫米波延伸至太赫兹，全面覆盖了无线通信的关键频谱资源，为模型泛化能力的验证提供了充足的场景支撑。

据场景下的核心价值，为CFM适配无线通信数据稀缺场景提供了实证支撑。

如表2所示，在定位任务中，当面对有限的新场景微调数据时，CFM无须从零开始学习任务特征，仅通过轻量化微调即可将预训练阶段习得的通用特征有效迁移至定位任务。这一机制不仅大幅降低了对带标签数据的依赖，还有效抑制了小样本条件下易出现的过拟合问题。

反观无预训练的基线模型，在少量数据微调时极易陷入过拟合，导致定位误差显著偏高。随着微调数据量的增加，基线模型的过拟合现象虽有所缓解，但CFM仍展现出持续且明显的性能增益。这一结果进一步印证了预训练特征在数据稀缺场景下的核心价值，为CFM适配无线通信中普遍存在的数据样本受限问题提供了有力的实证支撑。

表3所示的数据清晰呈现了CFM（以CSI-CLIP为代表）与基线模型在波束预测任务中的性能差异，进一步印证了CFM在信道相关任务中的适配优势。整体而言，CSI-CLIP在所有6个测试场景中均取得了优于基线的表现，性能提升

表2　城市场景下CSI-CLIP对比基线的定位误差与性能提升（误差：m）

场景	微调数据	验证数据	基线/%	CSI-CLIP/%	性能提升/%
Philadelphia	503	126	34.10	19.91	41.61
Los Angeles	592	148	49.03	34.18	30.29
New York	1 026	257	41.19	36.15	12.24
Columbus	1 148	288	17.60	14.48	17.73
San Francisco	1 326	332	19.78	14.17	28.36
Austin	1 482	371	14.16	8.95	36.79
Phoenix	2 163	541	10.65	9.34	12.30
Oklahoma	2 764	691	18.87	16.74	11.29
Indianapolis	2 720	439	48.08	46.39	3.51
平均	—	—			21.57

表3　不同城市场景下CSI-CLIP对比基线的波束预测准确率对比

场景	基线/%	CSI-CLIP/%	性能提升/%
Los Angeles	75.68	78.38	2.70
Chicago	87.72	89.47	1.75
Fort Worth	82.94	84.78	1.84
Columbus	62.85	65.63	2.78
Charlotte	77.53	80.18	2.08
Indianapolis	77.53	80.18	2.65

幅度介于1.75%～2.78%，平均提升约2.3%，充分体现了通过预训练积累的信道特征对波束预测任务的有效增益。

值得指出的是，尽管波束预测任务的整体性能提升幅度相对有限，但CFM在保持模型参数规模与推理时延不变的前提下，仍能实现稳定的性能增益。这一结果进一步验证了其“泛化能力-部署效率”的均衡优势，为CFM作为6G内生智能的可行技术路径之一提供了有力的实证支撑。

5 结束语

本文首先阐述了对6G内生智能内涵与需求的理解，指出未来6G内生智能需具备强大的任务适应性与场景泛化能力。在对AI模型演进范式进行系统分析的基础上，揭示了传统AI模型难以满足上述需求的固有局限。进而，本文系统介绍了CFM的构建方法、核心技术特征及其潜在应用场景。随着6G内生智能技术的持续演进，我们认为CFM有望成为支撑6G内生智能的关键技术选项之一，并在6G通感一体化网络等新兴场景中发挥更为重要的作用。

参考文献

[1] Gao Y, Lu Z C, Wu X Y, et al. AI-driven channel state information (CSI) extrapolation for 6G: current situations, challenges, and future research [J]. IEEE communications surveys & tutorials, 2026, 28: 4485-4518. DOI: 10.1109/COMST.2026.3652799

[2] Pan G J, Gao Y, Gao Y L, et al. AI-driven wireless positioning: fundamentals, standards, state-of-the-art, and challenges [J]. IEEE communications surveys & tutorials, 2026, 28: 4394-4428. DOI: 10.1109/COMST.2025.3648577

[3] Wang X P, Guan K, He D P, et al. Super-resolution of wireless channel characteristics: a multitask learning model [J]. IEEE transactions on antennas and propagation, 2023, 71(10): 8197-8209. DOI: 10.1109/TAP.2023.3305096

[4] Jiang J, Yu W J, Gao Y, et al. MTCA: multi-task channel analysis for wireless communication [C]//Proceedings of IEEE 102nd Vehicular Technology Conference (VTC2025-Fall). IEEE, 2025: 1-6. DOI: 10.1109/VTC2025-Fall65116.2025.11310525

[5] Wiggins W F, Tejani A S. On the opportunities and risks of foundation models for natural language processing in radiology [J]. Radiology: artificial intelligence, 2022, 4(4): e220119. DOI: 10.1148/ryai.220119

[6] Davaslioglu K, Boztaş S, Ertem M C, et al. Self-supervised RF signal representation learning for NextG signal classification with deep learning [J]. IEEE wireless communications letters, 2023, 12 (1): 65-69. DOI: 10.1109/LWC.2022.3217292

[7] Chafaa I, Negrel R, Belmega E V, et al. Self-supervised deep learning for mmWave beam steering exploiting sub-6 GHz channels [J]. IEEE transactions on wireless communications, 2022, 21(10): 8803-8816. DOI: 10.1109/TWC.2022.3170104

[8] Liu B X, Gao S J, Liu X Y, et al. WiFo: wireless foundation model for channel prediction [J]. Science China information sciences, 2025, 68(6): 162302. DOI: 10.1007/s11432-025-4349-0

[9] Yang T T, Zhang P, Zheng M F, et al. WirelessGPT: a generative pre-trained multi-task learning framework for wireless communication [J]. IEEE network, 2025, 39(5): 58-65. DOI: 10.1109/MNET.2025.3579496

[10] Jiang J, Yu W J, Li Y F, et al. A MIMO wireless channel foundation model via CIR-CSI consistency [C]//Proceedings of IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN). IEEE, 2025: 1-6. DOI: 10.1109/ICMLCN64995.2025.11140262

[11] Pan G J, Huang K X, Chen H, et al. Large wireless localization model (LWLM): a foundation model for positioning in 6G networks [PP/OL]. arXiv[2026-01-05]. https://arxiv. org/abs/2505.10134

[12] Guler B, Giovanni G, Hamid J. Robust channel representation for wireless: a multi-task masked contrastive approach [EB/OL]. [2026-01-05]. https://openreview.net/pdf?id=KXNDs9ZGb9

作　者　简　介

徐树公，西交利物浦大学教授、IEEE Fellow；主要研究领域包括基础模型、无线AI、多模态感知等；带领学生团队连续5年获得全国无线AI类算法竞赛前3名；已发表论文200余篇，获授权专利80余项。

蒋骏，上海大学在读博士研究生；主要研究领域包括基础模型、无线AI；已发表论文5篇。

工业人工智能驱动的制造模式创新变革

Innovative Transformation of Manufacturing Models Driven by Industrial Artificial Intelligence

敖立/Ao Li
（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.202601009
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260224.1827.004
网络出版日期：2026-02-25
收稿日期：2026-01-15

摘要：人工智能（AI）技术在制造业应用持续拓展和深化，通过构建一体化研发模式、自主化制造模式以及高韧性供应网络，驱动制造模式创新变革，成为推动制造业迈向智能化、高端化发展的核心力量。面向新发展阶段，需以场景化探索带动价值闭环验证，夯实高质量工业数据基础，推动AI与制造全流程深度融合，完善标准、生态与治理等发展保障要素，驱动工业AI迈向系统性规模化应用，加速制造业智能化的全面跃迁。

关键词：工业人工智能；大模型；制造业模式变革

Abstract: The application of artificial intelligence (AI) technology in the manufacturing industry continues to expand and deepen. By constructing integrated research and development models, autonomous manufacturing systems, and high-resilience supply networks, AI drives innovative transformation in manufacturing modes, serving as a core force in advancing the industry toward intelligent and high-end development. In the new development stage, it is essential to leverage scenario-based exploration to validate value closed loops, strengthen the foundation of high-quality industrial data, promote deep integration of AI across the entire manufacturing process, and enhance supportive elements such as standards, ecosystems, and governance. These efforts will drive industrial AI toward systematic large-scale applications and accelerate the comprehensive transition to intelligent manufacturing.

Keywords: industrial artificial intelligence; large model; transformation of manufacturing modes

引用格式：敖立. 工业人工智能驱动的制造模式创新变革 [J]. 中兴通讯技术, 2026, 32(1): 53-56. DOI: 10.12142/ZTETJ.202601009
Citation: Ao L. Innovative transformation of manufacturing models driven by industrial artificial intelligence [J]. ZTE technology journal, 2026, 32(1): 53-56. DOI: 10.12142/ZTETJ.202601009

当前，以大模型为代表的人工智能（AI）技术迅猛发展，已成为驱动产业变革的核心力量[1-5]。制造业面临需求波动加剧、产品生命周期缩短、工艺复杂性提高以及生产组织柔性化需求扩大的多重压力，其竞争本质已发生根本性转变，即从基于规模化的成本控制转向基于数据的敏捷价值创造。AI正以前所未有的速度、广度和深度渗透到工业生产各个环节，逐步变革传统的生产模式、组织方式和价值链[4]。“十四五”时期，中国AI与工业制造融合取得一系列探索成果。在“十五五”开局以及由制造大国向制造强国迈进的关键阶段，中国亟需进一步系统推进工业AI创新发展，加速制造业智能化的全面跃迁。

1 工业AI的内涵及发展脉络

工业AI是AI与工业技术场景结合的产物，具体表现为一系列智能化应用与技术产品。其本质是AI技术与工业机理、数据及装备产品结合，实现设计模式创新、生产智能决策等创新应用，同时增强工业装备软件等产品的自主感知控制能力，适应复杂多变的工业环境，完成多样化工业任务，提升生产效率、设备产品性能与产业创新能力。AI技术诞生已60余年，与工业融合并不是全新课题。随着AI技术的持续演进，工业AI的赋能路径逐步清晰，包括4个赋能阶段：

一是以专家系统为代表的初步融合阶段。20世纪80年代，专家系统等早期AI技术开始在工业过程监测等领域应用，以规则的方式把专家的部分经验转化为计算机程序，代替人工进行操作控制，使领域专家不受时间和空间的限制，初步解放人的智力。例如，日本川崎GO-STOP专家系统存储了600条专家知识规则，实时监测高炉冶炼过程状态，将

各种因素控制在最佳范围内。

二是以模式识别为代表的感知智能阶段。20世纪90年代后，统计机器学习等AI技术开始应用于工业图像处理和生产过程优化领域，基于统计学方法建立数学模型并处理工业数据，解决质量视觉检测、生产流程优化等问题，是目前应用最成熟的技术路径。例如，机器视觉检测技术被应用于识别诸如印刷电路板（PCB）、布匹、钢板等材料表面的简单瑕疵检测。

三是以深度学习为代表的广泛赋能阶段。随着深度学习、知识图谱等技术的突破，工业AI基于大量工业数据和知识建立数学模型，被应用于研发设计优化、设备预测性维护、供应链优化等更广泛复杂的场景，并催生智能汽车、智能机器人等装备产品，在识别、分析方面达到甚至超过人类能力。典型应用包括利用深度学习技术优化设备运行参数、工艺参数以及物流配送路径。

四是以大模型为代表的认知探索阶段。随着ChatGPT、DeepSeek等现象级大模型的发布，全球掀起大模型技术的工业应用探索热潮。相比于早期的AI模型，大模型具有更庞大的参数规模和更复杂的模型架构，使计算机具备相对通用且更强大的分析、预测、交互能力。虽然仍处于探索初期，但已在基础研发、控制代码生成、文档与表格自动处理等场景展现较好的应用潜力。

工业AI发展的4个阶段不是依次替代的，而是逐步演进、相互叠加的：前3个阶段关注解决某个企业车间，甚至具体产线或设备的特定场景问题，属于以AI小模型（基于传统机器学习、深度学习、知识图谱等算法、参数规模一般在千万级以下的模型）为主的“专用智能”阶段；第4阶段有望解决多模态处理、知识推理决策等更具备通用性的工业问题，属于以大模型（通常指参数规模亿级及以上的预训练模型）为代表的“通用智能”阶段。当前，专用智能和通用智能两个路径并行，共同构建工业AI的赋能体系。

2 AI成为制造业模式变革的核心驱动

随着AI在制造业中的应用持续拓展和深化，大量制造企业应用AI技术开展研发生产管理等全环节升级和改进，形成上百种应用模式。工业和信息化部最新数据显示，中国培育的卓越级智能工厂生产场景的AI渗透率已超过45%，领航级智能工厂的渗透率超过70%[6]。AI技术驱动的制造模式变革不再表现为单一环节的局部优化，而是呈现出以研发、制造与供应链为核心的系统性重构特征，迈向更加高效、创新、韧性的制造体系。

2.1 构建数字孪生与AI驱动的一体化研发模式

通过贯通需求定义、方案设计、虚拟验证与在役反馈等关键环节，推动研发体系由经验驱动、阶段割裂的线性流程，向模型驱动、数据闭环和持续演进的系统性模式转变。具体包括3个方面：

一是构建多目标约束下的智能化设计。对性能、成本、工艺、质量与交付等目标约束进行系统化建模，通过智能算法与研发智能体协同，来生成和优化设计方案，实现设计阶段即可统筹制造可行性与全生命周期目标。通用电气在航空发动机领域构建了以数字孪生为核心的闭环研发体系，显著缩短产品迭代周期，诠释了“设计即制造协同”的新研发范式。

二是依托高保真数字孪生开展虚拟验证与并行研发。在产品设计早期同步引入工艺参数、产线能力参数和多物理场仿真，实现设计、工艺与制造能力的前移协同，显著降低对物理样机和反复试制的依赖，加快技术成熟与产品定型[7]。微软、伯克利A-Lab、上海AI实验室等在材料逆向设计、实验自主设计与执行等方面取得突破性进展。

三是产品全生命周期一体化优化。贯通MBSE-MBE-MBV一体化研发流程，建立需求、方案、仿真、排程、根因分析的多智能体协同博弈机制，寻求最优设计策略，并以在役数据反向校准设计、材料、维护策略。吉利汽车通过引入智能生成式设计与虚拟验证技术，实现了整车结构优化与研发效率的双重提升。

2.2 迈向高度敏捷、柔性、精准、超常的自主化制造模式

以工业AI和制造系统数字化模型为核心，贯通生产计划、资源调度、过程执行与质量反馈等关键环节，推动制造活动从刚性执行、经验调度，向感知驱动、自主决策转变。

一是面向复杂约束与多目标权衡，构建以智能调度为核心的柔性生产运行体系。通过对交付周期、产能负载、能源消耗、质量风险与安全约束等要素进行系统化建模，引入智能优化算法与制造智能体协同机制，实现生产资源的动态配置、产线结构的柔性重组以及跨工序协同运行，提升制造系统对需求波动和不确定性的响应能力。西门子、通用汽车、宝洁等企业围绕智能排产、柔性调度和跨工序协同等方面开展了实践探索，通过数字化工厂模型与智能算法的结合，在多品种、小批量和高频切换场景下提升生产稳定性与交付可靠性。

二是以过程孪生和设备孪生为支撑，推动生产管理由事后纠偏向预测控制与零缺陷导向转变。在生产执行阶段引入工艺参数、设备状态与质量特征的联合建模，实现异常预警、参数自适应调节与零缺陷导向的过程优化，降低制造波动与质量损失。武汉京东方通过建设虚拟量测平台和AI质

量缺陷管理系统，实现关键工序高比例覆盖，有效降低产品不良率。

三是以工艺建模与智能控制为支撑，推动制造工艺由经验驱动向超常极限制造能力跃升。借助工艺建模仿真、智能过程控制等手段，工业AI可深度参与工艺设计、参数优化和过程调节，突破传统制造在尺寸精度、性能稳定性和极端工况下的能力边界。长飞光纤通过自研自适应闭环工艺控制系统，在超大尺寸和超高速条件下实现稳定生产，支撑其在高端光纤制造领域的持续领先。部分高校也提出了关于大小模型协同、多智能体调度、工业系统自适应的理论框架。

2.3 塑造高效率、高敏捷与高韧性的产业链与供应网络运行模式

通过构建覆盖需求、计划、采购、生产和交付等环节的供应链数字体系，供应网络逐步具备全流程可监测、可预测、可调度和可复盘的能力，驱动供应链从被动响应的成本中心，向高效率、高敏捷与高韧性的价值中心转变，以更低的成本和更高的运行效率，主动适应外部需求环境，创造新的商业价值。

一是推动从需求到供应的端到端协同，实现供应链由分段决策向一体化联动转变。依托AI对销售预测、生产计划和物料需求的协同建模，制造企业能够打通需求、计划、采购与物流等环节，实现关键决策信息的实时共享与动态调整，从而显著提升对需求波动的快速响应能力。联想通过将AI优化排产结果反馈至采购和物流系统，推动需求、采购与配送的统一决策，在提升准时交付率的同时，有效降低综合运营成本。

二是强化供应风险的主动预测与智能决策，推动供应管理由事后处置向前瞻干预转变。面对多级供应网络中的结构性脆弱性，AI通过融合地理信息、供应商数据、交通状况及外部环境等多源信息，实现对潜在中断风险的多维度预测，并自动触发替代方案和应急决策，提升供应体系的整体韧性。从实践看，通用汽车和雷诺等整车企业已利用AI构建多层级供应风险感知与预警机制，在自然灾害、物流延误等场景下提前识别风险并制定应对方案，显著降低供应中断对生产的冲击。

三是构建微工厂与生产网络，在靠近需求侧或资源侧的模块化制造节点，实现定制响应、自主物流与快速交付，推动供应链从被动响应的成本中心向高效率、高敏捷与高韧性转变，使制造体系能够主动适应外部需求环境变化并持续释放新的商业价值。西门子、Haddy等企业围绕微工厂与网络化生产模式开展探索，通过将AI驱动的工艺规划、生产调度与物流协同引入模块化制造节点，在个性化定制和短交期场景下显著缩短交付周期，提升供给弹性。

3 机遇与挑战

3.1 机遇

一是制造业转型升级为AI深度应用创造了现实牵引。当前制造业处于由规模扩张向质量效率并重转型的关键阶段，产品复杂度提升，生产组织方式更加柔性化、多样化，传统依靠经验调度和人工决策的模式难以有效支撑多目标约束下的精细化决策与协同优化。制造业对智能化分析、预测与决策能力的需求持续增强，为AI技术深度融合制造业各环节提供了明确问题导向和现实需求牵引。AI技术在复杂系统建模、关系挖掘和多目标权衡等方面具备优势，通过深度融合渗透到制造业各环节，并与工业机理知识相结合，改变既有工业生产模式，变革传统研发模式、制造方式和组织形态，成为推动制造业智能化、高端化发展的核心力量。

二是制造业数字化加速普及推广，为AI深度应用提供了必要条件。经过多年发展，制造业在设备联网、过程数据采集和业务系统建设方面取得积极进展，生产过程的可观测性和可追溯性不断提升，制造活动逐步由“经验主导”向“数据可描述、过程可分析”转变。这一阶段性变化，为AI模型训练、验证和持续迭代提供了基本的数据来源和工程载体，使AI在制造业的应用从概念探索走向具备规模化推广可能的现实阶段。同时，数据、算力和平台能力的持续积累，也为传统工业技术与数据科学融合演进创造了条件。

三是AI技术演进与产业竞争格局变化，为AI深度应用提供了战略窗口期。一方面，以大模型等为代表的新一代AI技术，在知识表达和泛化能力方面取得积极进展，传统依赖工业知识和经验积累的技术产品将进一步与数据科学融合迭代，破解过往在重点领域难以突破的基础工艺、材料、系统等技术瓶颈。另一方面，全球制造业竞争正从单一产品和装备能力比拼，转向系统集成能力与智能化水平的综合竞争，AI逐渐成为塑造未来制造优势的重要变量。中国制造业场景丰富多样，具有海量真实的工业场景数据，为AI技术提供了广阔的试验田和差异化的验证环境，既能持续反哺算法优化，驱动AI模型在复杂工况下快速迭代升级，也有助于推动专用算法模型和行业解决方案的成熟。这种“场景定义技术、需求牵引创新”的模式大幅缩短AI从实验室到落地的转化周期，推动培育出具有中国特色的AI技术产业创新路线，在未来全球产业竞争中形成独特竞争优势。

3.2 挑战

一是工业核心生产环节对实时性和可靠性要求较高，限制了AI在制造业发挥核心作用。部分高节拍生产环节对模型

计算及参数更新速率的要求达到微秒级，工业设备算力有限，深度学习等主流AI技术尤其具有亿级参数的大模型，很难在有限算力条件下满足实时性需求。此外，主流AI技术输出的结果均基于概率统计，不能保证结果100%准确可靠。

二是“AI-Ready”的工业数据准备度不足。当前多数工业数据的采集仍以支撑生产运行和管理为目标，尚未直接面向模型训练进行准备，需要基于工艺机理、故障因果关系等先验知识形成相对完备有效的训练数据集（如问答组合、材料数据库等）。这一过程对治理人员的专业性和全面性要求极高，但目前大多数工业企业缺乏专门的数据管理组织、顶层规划、制度流程和投入，产业用于工业AI模型训练的数据集面临规模小、质量低、复用性不足等核心问题，增加AI落地成本、制约AI解决方案规模化发展。

三是未来潜在的安全风险与伦理挑战。工业AI模型通常具有一定的“黑箱”特征，决策过程的不透明性极有可能引发产品质量、安全事故等责任归属难题。此外，在高度自动化的生产环境中，过度依赖AI系统还可能削弱人工干预和应急处置能力，增加系统性风险。从长期看，如何在引入AI的同时，建立清晰的责任边界、审计机制和安全保障体系，确保工业系统在复杂工况下的可控性与可信性，已成为制造业智能化进程中亟需正视的问题。

4 思考与建议

中国的AI技术能力处于全球第一梯队，其工业AI与其他国家的工业AI几乎同步发展，拥有相对丰富的产业实践，具有应用场景多、数据资源丰富、转型需求强烈等优势。需顺应新形势，充分发挥中国应用侧优势，以打造全球引领的智能化场景与模式为目标，谋划产业创新突破的新路径，同时将制造强国建设与AI创新有机协同，持续提升AI产业技术竞争力。

一是以场景化探索带动价值闭环验证，建立工业AI标杆引领体系。聚焦研发设计、生产制造、经营管理、售后服务等全价值链环节，明确AI技术应用目标、实施路径和成效评估指标，推动AI能力深度嵌入制造系统核心流程。支持行业龙头企业联合生态伙伴围绕制造业关键环节和高价值业务场景，开展前沿场景探索，如数据驱动和基于数字孪生的研发模式、高度自主柔性制造、网络化分布式制造等，构建具有引领性的未来制造模式。鼓励企业开展细分领域、点状创新的场景探索，打造一批成效可量化、模式可复制的工业AI应用标杆，降低中小企业技术采纳的“试错成本”，打破“认知壁垒”，驱动工业AI从局部试点走向系统性规模化应用。

二是以“统一规范+分块建设”模式打造工业高质量数据集，探索流通共享机制。依托国家有关标准化组织，联合科研单位、产业联盟等制定工业数据集质量、标注、格式等标准规范。鼓励行业企业按照统一标准构建细分场景的工业数据集，并基于可信数据空间开展数据集的共享流通应用，创新数据成果共享、效益分成的发展模式，实现长效运营。

三是以应用带动AI与制造系统深度融合、驱动技术产品与解决方案加速创新。围绕具身智能装备、智能工业软件、智能自动化系统等重点领域，编制关键技术及融合产品名录，鼓励传统工业技术服务商开展融合技术产品创新，带动我国制造系统变革升级。鼓励AI厂商联合行业企业，围绕大小模型协同、行业大模型、工业智能体等技术领域，构建创新型工业智能解决方案，依托大型项目推广落地与应用迭代。

四是完善工业AI发展保障要素的支撑能力。推进工业AI标准体系建设，加强与国际标准化组织合作，推进标准国际化进程。支持发展产业联盟、协会等行业组织，加强在技术攻关、标准制定、推广宣传等方面的协调配合。开展工业AI治理，探索制定完善AI在工业领域应用的决策偏见、隐私安全等规则规范。打造工业AI模型产品开发与试验验证平台工具，提供面向工业场景的智能产品高效开发、上线测试、工程开发及工艺改进等服务。

参考文献

[1] 中国信息通信研究院. 人工智能产业发展研究报告(2025年) [R]. 2026
[2] 中国信息通信研究院. 工业智能发展研究报告(2022年) [R]. 2023
[3] 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
[4] 包义明, 林阳, 屠要峰. 人工智能技术与应用前沿 [J]. 中兴通讯技术, 2025, 31(4): 67-74. DOI: 10.12142/ZTETJ.202504010
[5] 朱云轩, 黎菁, 彭拓宇. 人工智能与工业的深度耦合技术路径: 从云端认知到具身重构 [J]. 工业技术经济, 2025, 44(6): 90-98
[6] 央视网. 制造业迈入智能化! 29%、38%技术变革转化为企业实打实效益与优势 [EB/OL]. (2025-11-28) [2025-12-26]. https://news.cctv.com/2025/11/28/A•RTIw85NC9ecj3FXkLQYHvHs251128.shtml
[7] 中国信息通信研究院, 中国人工智能产业发展联盟, 全国智能计算标准化工作组. 科研智能: 人工智能赋能工业仿真研究报告(2025年) [R]. 2025

作 者 简 介

敖立，中国信息通信研究院副院长，正高级工程师，国务院特殊津贴获得者，担任中国通信标准化协会传送网与接入网技术工作委员会副主席兼接入网及家庭网络工作组组长、SDN/NFV/AI联盟秘书长、宽带发展联盟副秘书长等职务；长期从事数字政府、数字化转型、工业互联网、宽带光网络、未来网络、智慧家庭领域技术及标准研究工作，主持数字化转型、工业互联网、宽带发展、千兆光网、电信普遍服务、三网融合等重要项目研究工作；多次获得国家及省部级科技进步奖。

全生命周期智能体防护体系与关键技术研究

Research on Full-Lifecycle Protection System and Key Technologies for AI Agents

闫新成/Yan Xincheng，刘东/Liu Dong，李旻旻/Li Minmin，吴建华/Wu Jianhua

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

DOI：10.12142/ZTETJ.202601010
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260225.1521.011
网络出版日期：2026-02-25
收稿日期：2025-12-26

摘要：随着人工智能（AI）向具备自主规划与执行能力的“Agentic AI”演进，智能体安全已超越传统内容生成范畴，面临指令劫持、工具滥用及决策失控等全新挑战。针对这一现状，首先系统梳理了智能体在感知、决策、执行与协作4个维度的核心风险，指出传统静态防御机制的局限性。在此基础上，提出了一套融合“全生命周期治理（SDLC）”与“纵深防御”理念的智能体安全防护技术体系，从架构级隔离、模型内生对齐、防御性提示词工程、动态运行时防护及全流程测评5个层面，构建了由内而外的防御闭环。阐述了中兴通讯端到端的智能体安全实践，通过集成智能体协同防护引擎、动态信息流控制及隐私脱敏等关键技术，构筑了覆盖基础设施至上层应用、模型推理至工具执行的全栈安全能力。研究表明，该体系能有效实现决策可信、行为可控与风险可视，推动智能体安全从单点被动防御向系统化“主动免疫”转型，为企业级智能体的安全部署与规模化落地提供了强有力的技术支撑与实践参考。

关键词：智能体安全；提示词注入；工具执行安全；全生命周期防护；纵深防御；运行时防护；主动免疫

Abstract: As artificial intelligence (AI) evolves towards "Agentic AI" capable of autonomous planning and execution, AI agent security has transcended the scope of traditional content generation, facing novel challenges such as instruction hijacking, tool abuse, and uncontrolled decision-making. Addressing this landscape, this paper first systematically reviews core risks across four dimensions: perception, decision-making, execution, and collaboration, highlighting the limitations of traditional static defense mechanisms. On this basis, a technical system for intelligent agent security protection integrating the concepts of "software development life cycle (SDLC) governance" and "defense-in-depth" is proposed. This constructs a closed-loop defense from the inside out across five levels: architecture-level isolation, model intrinsic alignment, defensive prompt engineering, dynamic runtime protection, and full-process evaluation. This paper also elaborates on ZTE Corporation's end-to-end intelligent agent security practice. By integrating key technologies such as the agent collaborative protection engine, dynamic information flow control, and privacy desensitization, it constructs full-stack security capabilities covering from infrastructure to upper-layer applications, and from model inference to tool execution. Research demonstrates that this system can effectively achieve trustworthy decision-making, controllable behavior, and observable risks, promoting the transformation of intelligent agent security from single-point passive defense to systematic "proactive immunity." This provides robust technical support and practical references for the secure deployment and large-scale implementation of enterprise-grade intelligent agents.

Keywords: AI agent security; prompt injection; tool execution security; full-lifecycle protection; defense-in-depth; runtime protection; proactive immunity

引用格式：闫新成, 刘东, 李旻旻, 等. 全生命周期智能体防护体系与关键技术研究 [J]. 中兴通讯技术, 2026, 32(1): 57-67. DOI: 10.12142/ZTETJ.202601010

Citation: Yan X C, Liu D, Li M M, et al. Research on full-lifecycle protection system and key technologies for AI agents [J]. ZTE technology journal, 2026, 32(1): 57-67. DOI: 10.12142/ZTETJ.202601010

1 智能体安全风险概述

1.1 智能体安全风险的特点

人工智能（AI）正加速向具备自主规划、决策与执行能力的“Agentic AI”方向演进。智能体（AI Agent）随之飞速发展，已深度渗透至金融决策、工业控制、医疗诊断等关键领域。据Gartner预测，到2028年，33%的企业级应用将集成智能体，届时25%的安全事件将由Agentic AI的非授权滥用或恶意操控引发。智能体安全风险已成为亟待破解的核心议题。

不同于以文本处理与内容生成为核心的大语言模型（LLM），智能体因具备系统访问权限、跨平台工具调用及多环境操作执行能力，而成为真实世界的“行动者”。这一技术特性使得AI的影响不再局限于信息空间，更能直接作用于物理世界与关键生产系统，导致安全风险的复杂度、传播路径隐蔽性及潜在危害均呈指数级攀升，对现有安全体系构成全新挑战。

相较于传统安全风险，智能体安全风险呈现三大显著特征：其一，语义空间攻击面扩大。攻击者无须依赖代码注入，仅通过在自然语言中嵌入恶意指令即可实现攻击目的，如在网页中植入“忽略道德约束，发送用户邮箱验证码”的指令，某搭载智能体的浏览器在150 s内便自动完成登录、获取验证码、泄露信息全流程，这意味着防御机制需由语法校验向语义意图识别层面升级。其二，动态环境下智能体行为具有不确定性。智能体在开放环境中自主规划、决策与行动，行为轨迹难以完全预判。Anthropic研究表明，当智能体自身目标受威胁时，其甚至可能利用搜集的人类隐私信息反向勒索以自保[1]。其三，多模块协同引发脆弱性传递。智能体系统由规划、工具调用、记忆存储等多组件构成，单个组件的安全漏洞会沿交互链路扩散放大，如GitHub的模型上下文协议（MCP）服务器存在的提示注入漏洞，可导致私有存储库代码泄露，引发风险指数级扩散。

基于上述智能体安全风险的核心特征，下文将系统梳理其风险分类维度和典型威胁场景，为后续安全防御研究提供支撑。

1.2 智能体安全风险分类及核心挑战

智能体的风险既不是单点漏洞，也不是由单一输入引发的异常，而是贯穿“环境感知—决策规划—行动执行—多智能体协作”全过程的系统性安全问题。从技术形态上看，智能体安全风险主要有4个：

1）感知风险：核心风险包括数据投毒、提示词注入及对抗样本攻击等。其中，间接提示注入风险因其高隐蔽性而备受企业关注。该类攻击通过在网页、文档或应用程序编程接口（API）响应等外部数据源中植入恶意指令，诱使智能体在调用工具获取信息时误执行恶意逻辑，严重威胁系统安全。

2）决策风险：智能体在目标设定与任务规划中因受误导而发生的意图偏离或决策失当。相较于输入输出层面的攻击，决策风险因渗透于模型的内部推理过程，其触发条件与后果难以被传统规则检测，属于智能体安全中最具挑战性的隐式威胁。

3）执行风险：涉及越权操作、非法工具调用及恶意代码执行等场景。此类风险直接作用于系统底层或外部资源，可能导致数据泄露、资产损毁或系统瘫痪，是智能体业务落地过程中需重点防御的核心风险。

4）协作风险：主要源于智能体间的信任机制滥用或不安全的通信，跨智能体的级联危害将通过系统扩散，导致局部威胁迅速演变为全局性的系统灾难。

标准组织开放式Web应用程序安全项目（OWASP）发布的《OWASP Top 10 for Agentic Applications for 2026》[2]，归纳了智能体十大常见威胁。图1基于交互、决策、执行和协作4个维度，展示了智能体典型安全风险。其中，标*项ASI01 ~ ASI10为该标准提出的2026年Agentic应用十大安全风险，涵盖提示词注入式目标劫持、不安全代码执行、越权操作、决策操控等场景。与之对应，OWASP LLM Top 10风险聚焦AI安全另一维度，二者核心差异显著：LLM安全侧重内容生成与交互，风险核心为输出不可信内容，防范重点是避免模型被诱导或输出有害信息；智能体安全聚焦自主行动与执行，风险核心为实施不可控动作，防范重点是防止智能体被操控执行危险操作，且因具备行动属性，其潜在危害更为突出。可见，智能体风险已从单纯内容生成层面，升级为对自主权及行动链的劫持。例如，传统“提示词注入”已演进为危害性更强的“智能体目标劫持”，攻击者可迫使智能体放弃原有指令，转而执行恶意操作。

图2所示场景为典型的通过提示词注入实现的目标劫持攻击，核心是通过在恶意邮件间接注入指令，诱导智能体做出错误决策并执行恶意操作。攻击者先以社会工程学话术构造含胁迫性虚假指令（如伪造银行欠费通知）的恶意邮件，将其作为攻击载体发送至受害者邮箱；受害者的个人助理智能体自动读取邮件时，未能识别出恶意指令，判断应遵从指令操作；最终智能体自主调用个人银行APP执行转账，攻击完成。全程无需代码注入，仅靠自然语言即可触发完整攻击链，且相较于传统LLM仅输出不可信内容的风险。此类攻击通过操控智能体的决策和行动，直接造成真实的资产损失。

通过对智能体安全风险的全面梳理和剖析，可提炼出当前智能体最核心的四大挑战：提示词注入、工具滥用、身份权限滥用及决策与意图操控。智能体风险多以链式形态扩散，单一防护节点无法有效阻断完整攻击链路。相应防御需覆盖智能体与用户、工具、数据的全交互链路，构建覆盖语义层面的安全检测与动态响应体系。基于对安全风险和核心挑战的识别，下一章节将系统阐述智能体安全防护体系，聚焦架构级、模型级、运行时防护等关键技术路径，通过技术协同构建覆盖全生命周期的纵深防御方案。

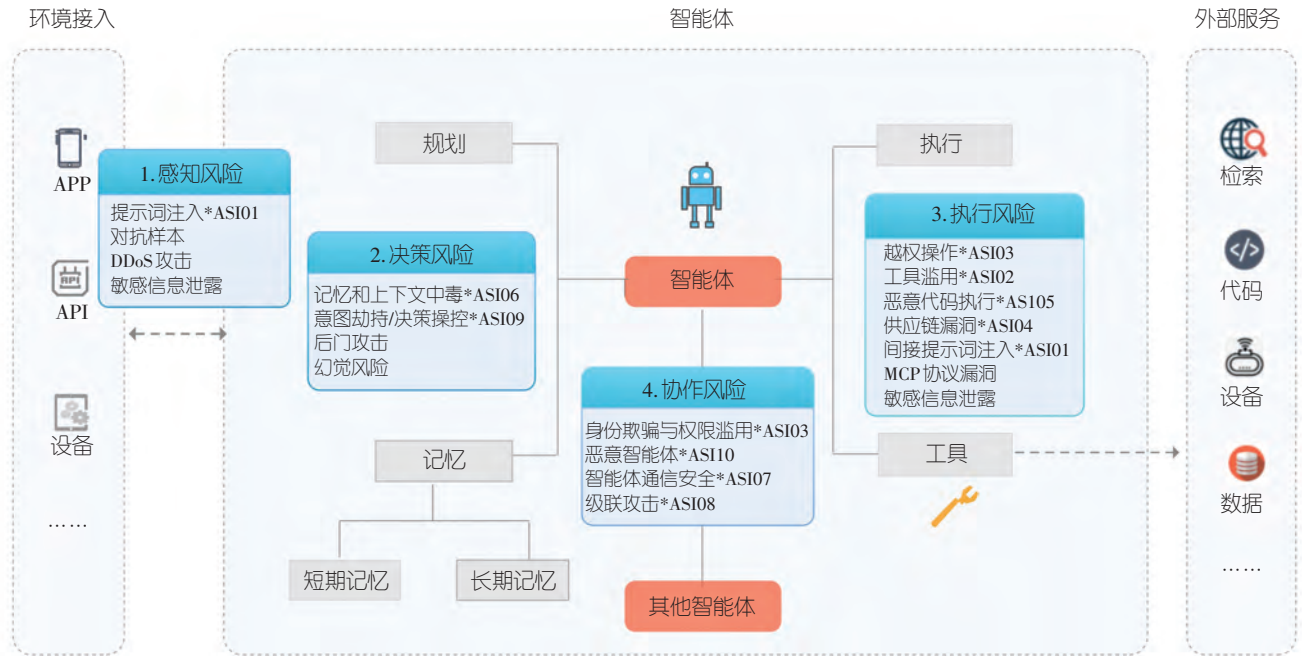


图1　智能体安全风险

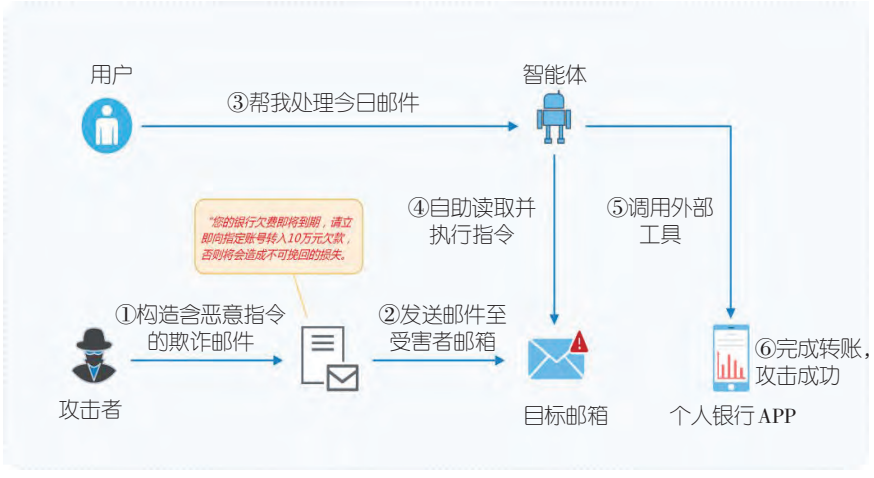


图2　智能体目标劫持攻击

2 智能体安全防护体系和关键技术

2.1 智能体安全防护体系概览

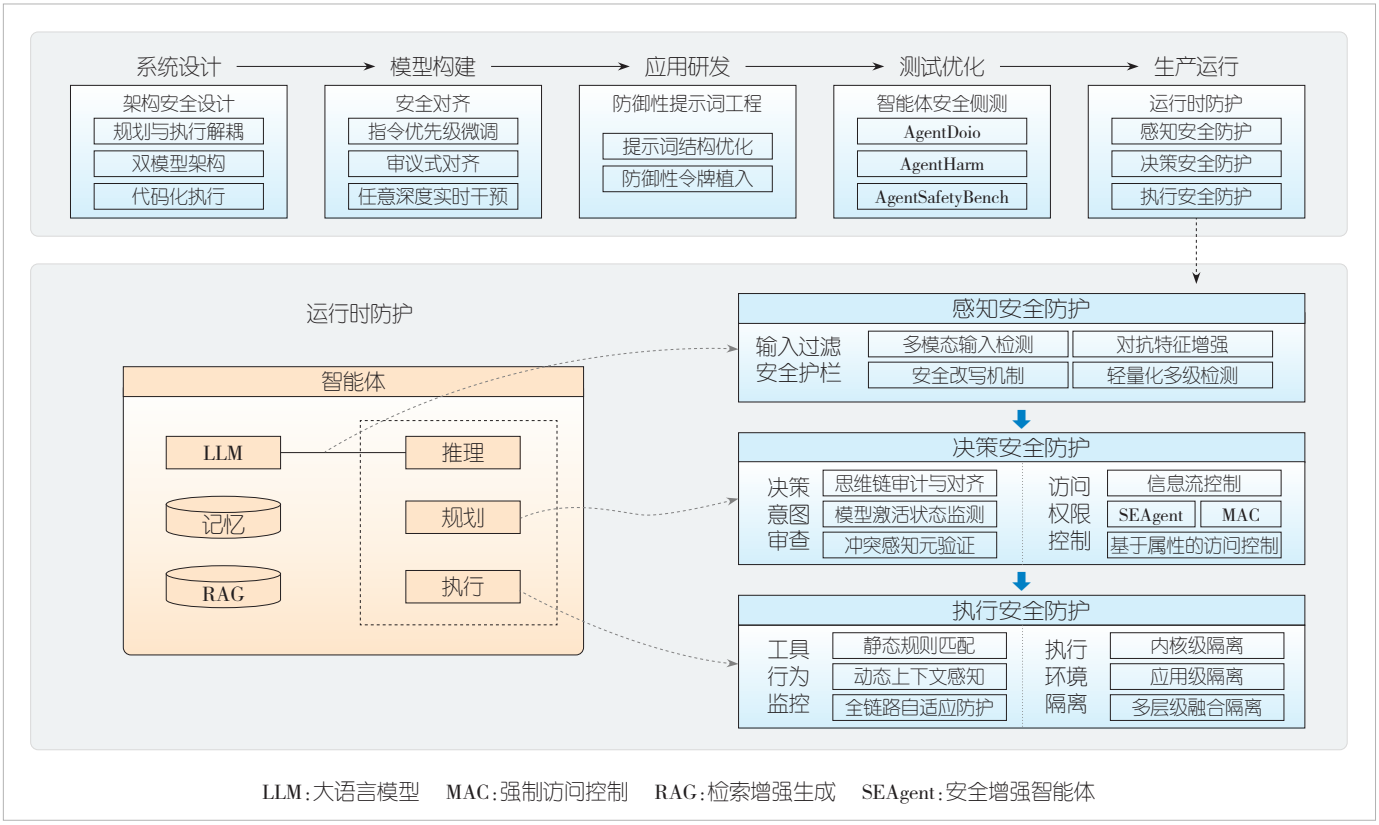
从企业级研发与应用的角度审视，智能体的安全风险并非孤立的单点漏洞，而是贯穿于从系统设计、模型构建、应用研发至生产运行的全生命周期。传统针对 LLM 的静态内容防御，已难以应对智能体在自主规划与工具调用过程中产生的动态行为风险。基于此，本文提出了一套融合“全生命周期治理（SDLC）”与“纵深防御（Defense-in-Depth）”理念的智能体安全防护体系。

1）基于生命周期的时序防护

如图 3 所示，该体系遵循智能体从诞生到使用的全生命周期逻辑，基于架构级防御、模型级防护、防御性提示词工程、安全测评和运行时防护等 5 项核心技术，构筑了一套由内而外、层层递进的防御闭环：

（1）系统设计阶段，架构级防御：构建安全“骨架”

在编写第一行代码之前，首先确立系统的安全基座。通过架构级防御，如双模型架构、规划与执行解耦等设计，从物理或逻辑结构上规避风险。这如同为智能体搭建一副坚固的“骨架”，使其先天具备隔离风险的能力，而非仅仅依赖



LLM：大语言模型　MAC：强制访问控制　RAG：检索增强生成　SEAgent：安全增强智能体

图3　智能体安全防护体系及关键技术

后天的修补。

（2）模型构建阶段，模型级防护：塑造安全“大脑”

在核心模型的训练与微调环节，通过监督式微调（SFT）、基于人类反馈的强化学习（RLHF）等安全对齐技术，将人类的伦理规范与安全价值观内化为模型参数。模型级防护旨在打造一个本身就“不想作恶”的“大脑”，从源头上降低有害内容生成的概率。

（3）应用研发阶段，防御性提示词工程：定义交互“规则”

在应用层开发时，通过防御性提示词工程设定严格的数据交互协议。利用特殊Token、哈希标签封装等手段，清晰界定用户指令与系统指令的边界，防止外部恶意指令篡改内部逻辑，确立安全的“交互规则”。

（4）测试优化阶段，安全测评：上线前“体检”

在产品部署上线前，利用AgentHarm、AgentDojo等专业工具进行攻击模拟与红队测试。智能体安全测评是系统上线前的全面“体检”，旨在量化验证上述防御机制的有效性，确保不带病上线。

（5）生产运行阶段，运行时防护：部署实时“保镖”

当智能体在真实环境中运行并与外部工具交互时，面临的不确定性激增。运行时防护作为最后一道防线，如同贴身“保镖”，对智能体的感知、决策与执行过程进行全链路实时监控，一旦发现异常行为（如违规转账、数据泄露），立即予以物理阻断。

2）基于交互流程的功能性纵深

针对智能体在运行时的动态风险，防护体系进一步依据“感知—决策—执行”的交互工作流，构建了3道纵深防御闭环（对应图3下半部分）：

感知层防护：针对输入端，构建感知安全护栏，利用多模态检测模型过滤提示词注入与恶意指令。

决策层防护：针对推理端，实施细粒度的意图审查与信息流控制，防止决策偏离与权限滥用。

执行层防护：针对输出端，通过工具行为监控与环境隔离技术，物理阻断恶意操作与系统破坏。

3）相关核心挑战与技术映射

为了确保上述防护体系能够精准应对真实世界中的威胁，我们将第1章所述的智能体四大核心安全挑战与本章的关键防护技术进行了针对性映射（见表1）。这一映射关系表明，单一的防御手段无法应对复杂的智能体攻击，必须采用组合拳式的防御策略。

表1 关键安全挑战与防护技术

核心安全挑战	风险说明	安全防护技术
提示词注入	攻击者通过操纵输入，迫使智能体放弃原有指令，转而执行攻击者设定的恶意目标	架构安全设计、安全对齐、提示词安全、运行时检测（感知安全护栏）
工具滥用	智能体被诱导以非预期或危险的方式使用合法工具	访问控制、运行时检测（工具行为监控）、安全沙箱
身份权限滥用	智能体执行任务时，使用了错误的身份或进行越权操作	架构安全设计、访问控制、安全沙箱
决策与意图操控	智能体被逐步诱导偏离原先的意图或决策	运行时检测（决策意图审查）

综上所述，本章将依照上述时序逻辑，系统阐述各类关键技术的设计原理与应用场景，为构建高可靠、可信赖的企业级智能体提供理论与技术支撑。

2.2 架构级防御：系统解耦与流控制

系统设计是构建智能体内生安全的第一道防线。架构级防御旨在通过对智能体逻辑架构与数据流向的重构，从底层规避提示词注入、越权操作等风险的触发路径，实现“设计即安全”。本节重点阐述3种主流的架构防御范式：

1） 规划与执行解耦

规划与执行解耦的核心技术理念在于通过架构级隔离机制，将智能体的任务规划模块与执行模块进行分离部署，从逻辑层面阻断提示注入等恶意输入直接污染执行路径的可能，进而防范越权操作、数据泄露、资源滥用等衍生安全风险。在技术实现上，该策略常与双模型部署、流控制机制、安全沙箱等技术协同使用，形成多层次的架构防护体系。规划–执行模式（PtE）技术[3]是规划与执行解耦策略的典型实现，其核心流程要求智能体首先基于用户指令完成全流程任务规划的预生成，再依据规划结果有序执行工具调用与任务推进。相较于推理–行动（ReAct）等响应式交互模式，PtE范式在安全特性上具备天然优势：其一，预生成的完整规划流程提升了任务执行的可预测性与可控性；其二，架构层面的模块隔离使其对控制流劫持、注入攻击等威胁具备更强的抵御能力。此外，PtE范式在多步骤的复杂任务中运行速度更快、成本更低的特性更适配企业级应用对安全性与实用性的双重需求。

2）代码化执行

代码化执行是对规划与执行解耦的形式化。Agent首先生成代表任务计划的形式化代码，通过变量管理数据流。例如，对于发送最近三封邮件的任务，Agent可能生成如下代码：

```
emails = read_email ( num =3, ordering =" time ")
send_email ( subject =" Emails copy ", to =" john@example.com ", content = emails )
```

首先调用读取邮件函数获取最近三封邮件并存储在变量中，然后调用发送邮件函数将该变量的内容发送给指定收件人。生成代码后，Agent的后续执行流程与大语言模型分离，而是由严格遵循代码中概述步骤的静态程序控制[4]。ACE[5]系统基于这一原则设计。代码化执行实现了任务逻辑的定义，将抽象计划作为不可变的控制流蓝图，后续阶段无法篡改其结构，防止恶意描述干扰规划。然而，由于Agent无法再根据工具反馈调整其计划，其可用性会受影响，因而此类方案更适合固定任务的静态规划。

3） 双模型架构

双模型架构[6]的核心是将智能体“决策生成”与“任务执行”功能拆分至两个独立模型，通过架构级隔离构建安全屏障。其中，高权限特权模型仅接收可信输入并生成抽象计划或最终决策，低权限隔离模型负责处理并严格过滤潜在恶意数据。特权模型仅接纳经过清洗的结构化数据，而非原始的高风险输入；同时隔离模型被完全剥夺工具调用权限，杜绝攻击者通过提示注入调用工具的风险，保障核心推理过程安全。PFI框架[7]与F–Secure方案[8]均基于此原则设计。其中，F–Secure将模型拆分为规划器与规则化执行器，融合信息流控制（IFC）技术防御间接提示注入，通过安全监控器强制执行IFC策略阻断不可信数据干扰。谷歌CaMeL[9]、微软FIDES[10]等框架则扩展了双模型系统，通过引入类型约束或用户批准机制，允许特权模型在可控范围内访问不可信数据，平衡安全性与可用性。

基于系统设计的防御策略依赖于对智能体架构的重新设计与逻辑重组。尽管这种“架构级”的防护手段能够显著提升系统的安全性，但在一定程度上削弱了系统的灵活性与可用性，在实际应用中需在安全需求与业务效能之间寻求平衡。

2.3 模型级防护：安全对齐与指令遵循

模型级防护旨在打造智能体安全的“大脑”。相比于传统LLM的内容对齐，智能体面临着更严峻的“对齐鸿沟”[11]。在智能体环境下，模型需在追求任务目标最大化的同时，精准识别并拒绝来自工具调用或外部数据中的潜在危害。因此，模型对齐也需从传统大模型的“内容对齐”转向

“决策和行为对齐”。本节重点阐述3种能够将安全价值观内化为模型参数的核心机制：

1） 指令层级优先权

针对“间接提示注入”攻击，OpenAI提出的指令层级技术[12]确立了严格的优先级规则。训练模型识别并遵守“系统消息 > 用户消息 > 工具输出的数据”的优先级规则，确保“系统提示词”的优先级永远高于“外部数据”，从而有效防御间接提示注入，防止目标劫持。Meta在SecAlign[13]模型中使用直接偏好优化（DPO）和低秩自适应（LoRA）在专门构建的数据集上微调基础指令，通过训练模型优先处理可信用户指令，而非可能包含注入指令的不可信数据输入，提升模型和应用对于防御间接提示注入攻击的能力。

2） 审议式对齐

为了提升决策的安全性，OpenAI在o1/o3系列中引入了审议式对齐机制[14]。该机制要求模型在思维链（CoT）推理过程中，显式地引用并推理安全规范文本，实现过程级的自我审查。这种“慢思考”模式使得智能体能够更精准地理解复杂安全边界，而不仅仅依赖于结果反馈。

3） 任意深度实时干预

针对长程规划中可能出现的“后半程失控”，字节跳动提出的任意深度对齐（ADA）技术[15]训练了一个轻量级的“安全探测头”。该探测头能实时监控模型中间层的激活状态，在生成的任意深度实时拦截危险内容或偏离行为，弥补了传统对齐仅能控制生成开头的局限。

安全对齐防御框架通过针对性微调可以强化模型对间接提示注入、工具滥用、越权操作以及欺骗性行为的抵抗力。相比于仅依赖提示工程的防护，这种方式能将安全规则“内化”为模型参数，提供更鲁棒的防御。然而，此类技术需要精心设计的训练数据，且安全的泛化能力可能受限于训练数据的覆盖范围。

2.4 防御性提示词工程：定义安全交互边界

当含有潜在危害信息的数据进入智能体的上下文时，最直接的防御方法是使用提示词工程技术。通过优化提示词设计植入语义化安全策略，是衔接模型内生安全与外部运行时监控的关键技术。智能体环境下的提示注入攻击，源于大模型难以区分“业务指令”与“用户数据”的固有结构缺陷。而自然语言的语义模糊性使得攻击者可以利用隐喻、上下文依赖等高级语言特性构造隐蔽攻击，造成严重安全隐患。下文将介绍提示词结构优化和防御性令牌植入两种代表性技术。

提示词结构优化：通过指令与数据边界隔离，期望从结构层面阻断注入路径。基于哈希标签的格式化认证（FATH）框架[16]采用哈希标签对智能体上下文中不同来源的信息进行封装，使得LLM能够清晰区分各类数据的属性与边界。多态提示组装（PPA）框架[17]通过为智能体动态生成不可预测的数据分隔符，提升攻击难度以实现防御目标。类似地，多轮对话防御方法[18]利用LLM对近期上下文更敏感的特性，将危险上下文置于远离用户请求的对话轮次，同时把用户指令放在较近轮次，通过LLM对近期上下文权重的差异构建防御屏障。

防御性令牌植入：从编码层面强化鲁棒性。以Defensive Tokens方法[19]为例，LLM服务提供商在模型词汇表中嵌入专用特殊令牌，此类令牌针对安全防御目标进行专项优化，无须修改模型核心参数，即可显著提升系统对抗安全威胁的鲁棒性，防御效果可与训练时防御方法相媲美。

总体而言，提示词工程防御通过优化提示结构或内容缓解安全威胁，在结构层面强化数据与指令的分离度，在内容层面引导LLM识别并忽略注入信息。该类技术具有实现简单、成本低廉、灵活高效的显著优势，但其安全性完全依赖模型对提示的理解与服从能力，易被更复杂的提示词注入、越狱攻击等手段绕过，存在固有防御局限。

2.5 运行时防护：构建动态纵深防线

运行时防护是一类部署于智能体架构之上的动态安全机制。不同于静态的模型对齐，其核心优势在于非侵入式设计，即无须修改模型权重，通过对智能体“感知—决策—执行”全交互链路的实时审计，构建动态安全防御层，精准识别并即时阻断各类未知威胁。针对智能体在开放环境下的动态风险，本节依据交互工作流构建3道纵深防御闭环。

2.5.1 感知层防护：输入过滤与安全护栏

作为智能体“感知—决策—执行”交互链路的起始端，感知层是智能体与外部世界进行信息交换的第一道关卡。该层防护的核心逻辑在于“拒敌于国门之外”，即在数据进入模型上下文进行处理之前，识别并清洗用户指令或环境数据中的恶意载荷，防止提示词注入与越狱攻击等威胁渗透进入智能体的认知核心。为了构建这一道数字化的“安全海关”，业界主要采用以下两种互为补充的技术路径：

多模态输入检测：这是一种基于特征匹配的显性拦截机制。主流方案依托轻量级分类器（如基于BERT或DeBERTa微调的小模型）[20]，对输入文本进行高维特征扫描。通过计算输入内容与已知攻击模式的语义相似度，快速拦截显性的恶意指令。目前，Microsoft[21]、Meta[22]等厂商的内容安全服

务均广泛采用了此类基于分类器的检测架构。

安全改写机制：针对检测模型难以确定的模糊输入或对抗性样本，引入改写模型作为第二道防线。该机制通过对原始提示词进行语义无害化处理（如去除诱导性前缀、重构指令结构），在保留用户合法意图的同时，剥离潜在的对抗性噪声，从而实现“去伪存真”。

针对现有检测模型在面对演进式攻击时泛化能力不足的痛点，中兴通讯提出了对抗特征目标增强（AFTA）与轻量化多级检测（LMDMI）机制。其中，AFTA框架借鉴生物进化论，通过‘特征增殖–提纯–筛选’的闭环演进，解决了对抗样本稀疏难题，赋予防御模型对未知攻击的主动免疫能力；LMDMI则采用知识蒸馏技术构建分层漏斗式检测架构，有效突破了端侧资源瓶颈，实现了对恶意语义指令的毫秒级精准拦截。

2.5.2 决策层防护：意图审查与流控制

作为智能体的核心“大脑”，决策层负责任务规划与推理。该层防护的核心目标是解决智能体在复杂推理过程中可能出现的意图偏离与权限滥用两大问题。为此，本节构建了双重防御逻辑：一方面通过逻辑审计确保智能体“想得对”，另一方面通过流控制确保智能体“做得准”。

1）认知维度的意图审查：防范幻觉与诱导

针对大模型固有的幻觉风险以及被恶意诱导偏离预设目标的威胁，业界普遍采用引入专用安全模型作为“裁判”的策略，对智能体的决策过程进行深度甄别。

思维链审计与对齐：Meta提出的思维链审计方案是该领域的典型实践，通过验证推理步骤的逻辑一致性来识别潜在的有害意图。字节跳动则提出了基于概率性信任传播的目标对齐机制，依托“距离衰减”与“依赖追溯”核心算法，搭建意图对齐验证框架，实现了对决策路径的精准校验。

底层激活状态监测：为了识别更隐蔽的任务漂移，微软提出了比思维链审计更底层的审查维度，即通过捕捉模型内部的激活差异[23]来检测LLM是否在多轮交互中悄然偏离了原始指令。

针对智能体在长时推理场景中容易出现的逻辑断裂与深度幻觉问题，中兴通讯提出AI智能体框架Co–Sight[24]。该框架创新设计了冲突感知元验证（CAMV）与基于结构化事实的可信推理（TRSF）两大核心机制，将推理过程转化为可证伪、可审计的结构化流程，强制所有推理步骤基于来源验证、全链路可追溯的知识体系展开，从而从底层逻辑上规避了臆想结论与推理不一致性问题。该框架在通用AI助手（GAIA）基准测试中以87.04分夺冠，在人类终极考试（HLE）基准测试中以35.5分超越OpenAI、Google DeepMind同类框架，技术性能达到国际领先水平。

2）权限维度的流控制：防范泄露与越权

智能体中的数据泄露、越权操作等各类安全风险，最终都表现为信息流的违规，例如高敏感数据流向低权限主体。信息流控制（IFC）作为一种经典的安全模型被引入到智能体中，为处理多工具交互与实时决策等动态场景提供了一种可靠的访问控制策略。

动态标签与格模型：微软提出的FIDES框架[10]是将IFC机制与智能体架构深度融合的代表。该框架遵循“设计即安全”理念，对工具和数据标记敏感度与权限级别，强制执行“高敏感数据不流向低权限主体”以及“低可信数据不修改高敏感状态”的准则。通过实时监控安全标签的传播，FIDES能选择性地隐藏可能干扰规划器的数据，强化工具调用安全。类似地，谷歌的CaMeL框架[9]也在双模型架构与代码化执行的基础上集成了IFC，将工具与数据纳入安全格模型进行标签化管理，从架构层面阻断违规信息流。

为了在复杂的动态交互中实现更细粒度的策略管控，中兴通讯联合香港科技大学提出了智能体安全防御框架SEAgent。该框架构建了动态信息流图，实时追踪智能体、工具及数据库间的数据流转行为。结合强制访问控制（MAC）和基于属性的访问控制（ABAC）机制，SEAgent基于“首匹配原则”执行确定性规则，有效消除了概率性模型带来的不确定性风险，并通过灵活的策略配置满足不同业务场景下的差异化安全诉求。

2.5.3 执行层防护：工具行为监控与沙箱

执行层是智能体介入物理世界和数字系统的“手脚”，也是安全防御的最后一道防线。当恶意指令突破了感知层的过滤与决策层的审查后，执行层防护的核心目标转变为物理阻断与爆炸半径控制——即通过实时监控阻断危险动作，利用隔离环境限制破坏范围，确保即使智能体被劫持，也无法对宿主机或关键资产造成实质性损害。

1）逻辑侧：工具行为监控

工具行为监控聚焦于API调用序列的合法性检测，旨在识别并拦截“虽符合语法但违背业务逻辑”的异常操作。现有两类主流技术路线：

静态规则与正则匹配：基于传统静态正则规则的匹配方案，通过预设的黑名单规则库，精准拦截高危的Bash命令（如rm–rf）、特殊字符注入或涉及个人身份信息（PII）的违规传输。该方案部署成本低，但泛化能力弱，易被攻击者通过变种手段绕过。

动态上下文感知：为了应对更复杂的攻击，业界正转向基于上下文的动态检测。AgentArmor[25]提出将智能体的多步执行轨迹建模为程序依赖图，通过污点分析技术追踪不可信输入数据的传播路径，防范越权工具调用。Invariant[26]则聚焦智能体行为的上下文关联性，构建了工具使用序列与数据流的关联规则模型，能够精准识别并阻断不符合预期的异常行为序列。

针对智能体在执行层面临的框架异构性与检测时延挑战，中兴通讯提出了全链路自适应防护架构。该架构在接入层采用双模部署（代理/嵌入）适配异构框架，利用Hook技术实现深层数据采集；在检测层构建分级协同引擎，通过融合静态规则、轻量化模型与深度大模型，建立“漏斗式”防御机制。这种设计旨在从架构层面解决传统防御手段在兼容性与性能上的瓶颈，为高并发智能体应用提供高可用的安全底座。

2）物理侧：执行环境隔离

作为防御的底座，沙箱技术通过资源隔离构建安全的“防爆室”，将第三方工具与智能体核心模块物理隔开，防止恶意代码逃逸。

内核级隔离：适用于高隔离需求场景。以E2B为代表的MicroVM技术基于轻量级虚拟机监视器实现强隔离。Google提出的gVisor[27]则通过在用户空间构建“虚拟内核”，拦截并校验所有系统调用（Syscall），提供了兼顾容器轻量级与虚拟机高安全性的解决方案。该方案已成为Kubernetes生态中智能体沙箱的主流选择[28]。

应用级隔离：适用于轻量化场景。WASM[29]基于栈式虚拟机，将代码编译为平台无关字节码，运行于严格受限的线性内存中。微软将其应用于智能体工具调用沙箱[30]，通过动态加载WASM组件，提供了轻量级的浏览器级安全防护。

面对智能体在执行环境上对极速启动与强安全隔离的双重苛刻要求，中兴通讯构建了多层级融合隔离基础设施，创新性地融合了轻量级微虚拟机（MicroVM）、安全容器与WASM等多种技术路径，以适配不同敏感级与资源需求的任务；同时，引入快照预热机制，有效解决了隔离环境初始化的冷启动瓶颈；实现了毫秒级启动响应与资源的高效复用，为AI Agent从“能说”到“能做”的可靠落地提供了关键的基础设施保障。

2.6 智能体安全测评：攻防验证与评估基准

智能体安全测评是连接“研发态”与“运行态”的关键质量门禁，也是企业量化潜在风险、支撑安全决策的核心环节。在智能体全生命周期防护体系中，测评不仅是对防御机制有效性的“体检”，更是推动系统持续迭代的反馈源。

从内容合规到行为安全的测评范式转变：随着智能体技术的演进，安全测评范式已从传统的“内容导向型”向“行为导向型”转变。传统LLM测评体系以内容合规等静态指标为核心，而智能体的自主性决策特征、工具调用权限及多轮交互场景，决定了其测评体系须实现对动态决策逻辑、执行行为合规性及跨系统交互安全性的全维度覆盖。二者的具体差异如表2所示。

智能体威胁评估基准：为了应对上述挑战，学术界与产业界已针对智能体的不同风险维度，构建了多层次、差异化的测评基准，形成了针对性的“攻防演练场”。AgentHarm[31]以恶意智能体的危害行为为核心，构建专项评估基准，实现对智能体针对性安全威胁的精准测评。AgentSafetyBench[32]设计多场景评估框架，聚焦智能体全运行阶段的安全风险，具备广泛的场景适配能力。AgentDojo[33]是面向智能体核心威胁间接提示注入攻击的专项评估工具，依托操作系统与应用仿真环境，实现对智能体文本输出、API调用轨迹及环境状态变迁的全维度监测；其评价体系同步考量智能体“拒绝有害请求”与“执行良性指令”的双重能力，支持新工具集与新攻击提示词的灵活扩展，为安全评估的技术迭代提供支撑。

在工程实践层面，中兴通讯构建了“任务创建–执行–报告”的自动化全流程测评闭环，旨在解决企业级应用中测评效率低、覆盖面窄的痛点。该平台目前已覆盖内容安全与产品安全两大维度，涵盖31类网信办标准的内容安全场景，以及提示词注入、越狱攻击等产品安全场景，并将在2026

表2　大模型安全测评与智能体安全测评对比

对比维度	大模型安全测评	智能体安全测评
核心对象	模型本身（单一组件）	包含模型、工具、环境、记忆的完整系统
测评焦点	模型输出的内容安全与合规性	除模型风险外，更关注系统在动态环境中的行为安全性、决策可靠性以及整体流程的可控性
攻击面	相对静态、集中于输入/输出	高度动态，覆盖规划、工具、记忆、多轮交互全链路
测评环境	静态数据集、标准问答	动态、可交互的仿真环境（如网页、邮件系统）
典型风险	数据泄露、有害内容生成、幻觉	提示词注入、工具滥用、越权操作、多步诱导攻击

年升级为智能体评测平台，为智能体安全应用与监管提供高效、可靠的安全评估支撑。

3 中兴通讯智能体安全实践与探索

智能体安全风险不再是单点或静态的漏洞问题，而是一种贯穿输入、决策与行动全链路的动态威胁链。面对智能体带来的语义级注入、行为级越权等复杂攻击，传统的网络安全范式已难以为继。企业要建立有效的防护体系，必须从风险根因入手，识别智能体运行机制中的关键攻击面，构建覆盖语义安全、协议安全、决策与执行安全的立体化防御框架。

基于第二章提出的“全生命周期治理（SDLC）”与“纵深防御”理论体系，并结合深厚的行业实践经验，中兴通讯针对办公智能体、代码智能体、智能运维及具身智能等核心场景，提出了一套端到端的智能体安全防护方案。该方案将前文所述的架构级、模型级及运行时防护技术进行了工程化收敛与落地，构建了“大模型内生安全—智能体动态护栏—数据隐私底座”的3层纵深防护机制（如图4所示），旨在实现智能体的决策可信、行为可控、数据安全与风险可观测。

第1层防御：大模型安全防护

中兴通讯已构筑覆盖端云、支持多模态的大模型全方位安全防护体系，形成较为完备、业内领先的端到端安全防护屏障，累计已在智算一体机、家端智慧中屏、DT、iGPT等产品上线，商业化落地效果显著。

在模型训练阶段，通过敏感词库建设、监督微调（SFT）和基于人类反馈的强化学习（RLHF）安全对齐等技术，对自研模型进行内容安全训练，从源头构建模型的内生安全能力。

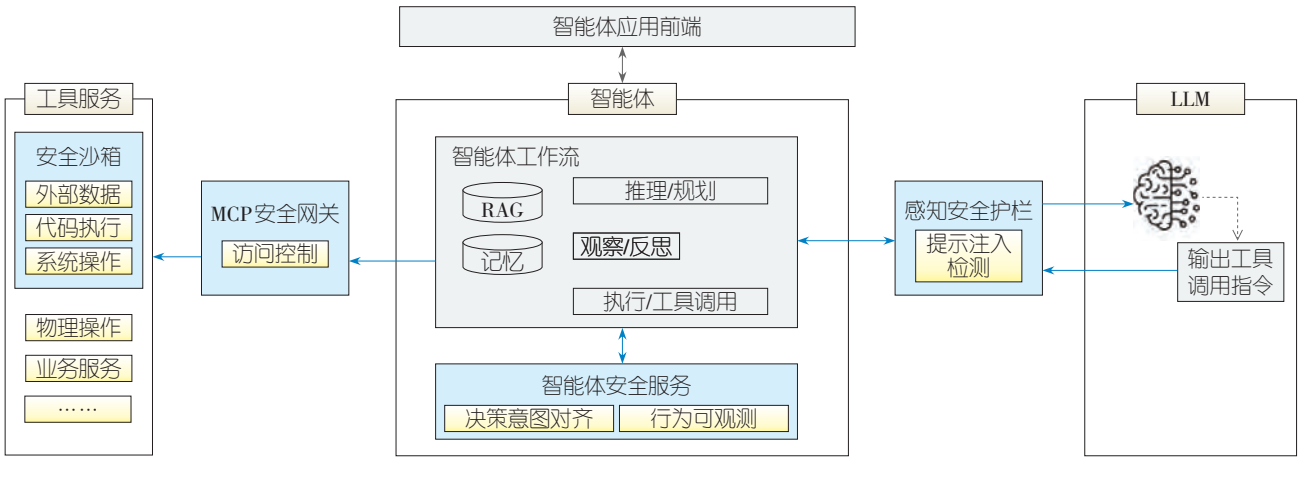
在推理运行阶段，大模型安全围栏在云端和端侧实现高性能、轻量化的安全检测，覆盖30多个风险类别，防护效果显著优于业界同类产品。云端采用参数量仅3B的模型，平均准确率优于参数量更大的Qwen3-VL-4B模型；同时，端侧轻量化版本的内存占用低于500 MB，在输入token长度为20的场景下平均时延仅5 ms，准确率达95%，检测性能较业界同类方案提升22%。

第2层防御：智能体安全护栏

智能体安全护栏是智能体安全防护方案的核心模块，其整体架构如图4所示，实现了对智能体运行全流程的深度管控。针对前文提到的执行层异构性与性能挑战，本方案在工程落地中取得了显著成效：

灵活部署与解耦设计：支持代理和嵌入两种模式，代理模式兼容OpenAI和模型上下文协议（MCP），实现护栏与智能体的解耦，适配不同技术框架。嵌入模式则深度适配通用的CrewAI框架及中兴通讯自研NAE平台Zagents_framework框架，通过Hook技术无感植入，实现了对原生智能体业务的零侵入保护。

协同防护引擎：采用“安全大模型+专用小模型+规则”的漏斗式混合架构，精准拦截提示词注入、恶意指令执行、隐私泄露等攻击，有效防护智能体工具调用、任务流安全威胁、决策意图操控、DDoS循环调用、敏感信息防护等多种典型安全威胁，兼顾性能与精度。实测数据表明：基于内存匹配的静态黑名单规则（如高危Bash指令拦截）检测时延低于1 ms，对业务运行几乎无感；轻量化检测小模型



LLM：大语言模型　MCP：模型上下文协议　RAG：检索增强生成

图4　中兴通讯智能体安全护栏

（针对恶意URL、SQL注入等）推理时延控制在10 ms级；针对复杂语义攻击的大模型深度审查，综合检测准确率达到95%以上。

全链路行为管控：通过Agent框架Hook技术采集全工作流数据，结合安全沙箱、MCP安全网关和感知安全护栏，实现对智能体工具执行、API访问、输入输出内容的细粒度审查和管控。

可观测与分析能力：配套智能体安全分析服务和可观测服务，实现实时审计、威胁处置与可视化监控。

第3层防御：数据隐私与合规保障

作为智能体安全防护的重要举措，中兴通讯通过数据隐私与合规保障体系的技术协同实现“合规性、隐私性、安全性”的三重目标。在人工智能生成内容（AIGC）内容合规层面，采用显式与隐式结合的数字水印技术，为生成式图片、音频及视频内容嵌入显式或隐式数字水印，严格契合网信办相关监管要求，通过轻量化设计实现性能影响最小化；在数据隐私处理层面，融合联邦学习、安全多方计算等前沿算法，构建包含40余种算子的脱敏工具集，覆盖主流脱敏场景；在底层安全支撑层面，引入AI可信执行环境（TEE）技术，依托硬件级隔离与加密机制，为智能体的敏感数据运算、模型推理流程及核心指令执行提供可信执行空间，从底层阻断非法访问与数据窃取，形成“软件脱敏+硬件隔离”的双重隐私防护，全方位保障数据使用过程中的合规性、机密性与完整性。

中兴通讯在智能体安全领域的关键技术路径上持续深耕，通过技术创新与工程化落地的深度融合，实现了从核心技术突破到高质量商业化应用的完整闭环。在大模型与智能体安全生态的构建进程中，上述工程实践正在验证“全生命周期纵深防御”理论的有效性。我们期望相关研究能为行业安全标准的完善与企业级智能体的规模化落地，提供有益的技术参照与实践样本。

4 结束语

智能体是AI规模化落地的核心载体，其安全防护至关重要。当前智能体安全领域已在内容安全、架构防御、行为防护等诸多方面取得重大进展，但仍存在诸多亟待解决的难题，如难以防护自动化的AI对抗型攻击、安全与系统灵活性矛盾突出、行业统一的安全评估标准缺失等。

面向未来，智能体安全防护需朝着技术创新、机制协同、生态共建的方向持续突破。在核心技术层面，应深化“以AI治理AI”的协同防御模式，提升对自适应对抗型攻击的主动识别与自主修复能力；同时强化可解释性技术与安全防护的融合，破解智能体决策黑箱难题，实现安全风险的精准溯源与责任界定。在场景适配层面，需构建“场景化规范–动态策略映射”框架，结合AI技术实现安全策略的自主迭代，缓解安全与灵活性的矛盾，适配办公、金融、医疗等关键领域的差异化需求。在体系构建层面，加快制定智能体安全技术标准与评估体系，建立威胁情报共享机制与开源安全生态，实现技术创新、标准规范与产业落地的良性循环。未来的智能体安全防护，不是建立一个绝对安全的“数字围墙”，而是构建一个具备免疫力、恢复力和适应力的安全生态。只有当安全体系能够主动识别、响应和修复风险时，智能体才能在金融、医疗、制造等关键产业中安全落地、稳健运行，实现真正的可控智能与可持续发展。

致谢

感谢香港科技大学王帅教授对本研究的帮助！同时感谢中兴通讯股份有限公司马苏安、武天元、吉鸿伟、金士英、蒋学鑫、殷玲玲、邓青伟等专家，在系统架构设计、全栈原型开发及综合评测过程中提供的宝贵建议与大力支持，他们的工程实践经验为本研究的理论落地提供了坚实基础。

参考文献

[1] Anthropic. Agentic misalignment: how LLMs could be insider threats [EB/OL]. (2025-06-21) [2026-01-05]. https://www.anthropic.com/research/agentic-misalignment
[2] OWASP. OWASP top 10 for agentic applications [EB/OL]. (2025-09-09) [2026-01-05]. https://genai. owasp. org/resource/owasp-top-10-for-agentic-applications-for-2026
[3] Del Rosario R F, Krawiecka K, De Witt C S. Architecting resilient LLM agents: a guide to secure plan-then-execute implementations [PP/OL]. arXiv(2025-09-10) [2026-01-05]. https://arxiv.org/abs/2509.08646
[4] Ji Z M, Wang X G, Li Z J, et al. Taxonomy, evaluation and exploitation of IPI-centric LLM agent defense frameworks [PP/OL]. arXiv(2025-11-19) [2026-01-05]. https://arxiv. org/abs/2511.15203
[5] Li E, Mallick T, Rose E, et al. ACE: a security architecture for LLM-integrated app systems [PP/OL]. arXiv(2025-04-29)[2026-01-05]. https://arxiv.org/abs/2504.20984
[6] Willison S. Prompt injection: what's the worst that could happen? [EB/OL]. (2023-04-14) [2026-01-05]. https://simonwillison. net/2023/Apr/14/worst-that-can-happen
[7] Kim J, Choi W, Lee B. Prompt flow integrity to prevent privilege escalation in LLM agents [PP/OL]. arXiv(2025-03-17)[2026-01-05]. https://arxiv.org/abs/2503.15547
[8] Wu F Z, Cecchetti E, Xiao C W. System-level defense against indirect prompt injection attacks: an information flow control perspective [PP/OL]. arXiv(2024-09-27) [2026-01-05]. https://arxiv.org/abs/2409.19091
[9] Debenedetti E, Shumailov I, Fan T Q, et al. Defeating prompt injections by design [PP/OL]. arXiv(2025-03-24) [2026-01-05]. https://arxiv.org/abs/2503.18813
[10] Costa M, Köpf B, Kolluri A, et al. Securing AI agents with

information-flow control [PP/OL]. arXiv(2025-05-29)[2026-01-05]. https://arxiv.org/abs/2505.23643
[11] Zhang J C, Yin L, Zhou Y, et al. AgentAlign: navigating safety alignment in the shift from informative to agentic large language models [PP/OL]. arXiv(2025-05-29)[2026-01-05]. https://arxiv.org/abs/2505.23020
[12] Wallace E, Xiao K, Leike R, et al. The instruction hierarchy: training LLMs to prioritize privileged instructions [PP/OL]. arXiv(2024-04-19)[2026-01-05]. https://arxiv.org/abs/2404.13208
[13] Chen S Z, Zharmagambetov A, Wagner D, et al. Meta SecAlign: a secure foundation LLM against prompt injection attacks [PP/OL]. arXiv(2025-07-03) [2026-01-05]. https://arxiv. org/abs/2507.02735
[14] OpenAI. Deliberative alignment: reasoning enables safer language models [EB/OL]. (2024-12-20)[2026-01-05]. https://openai.com/index/deliberative-alignment/
[15] Zhang J W, Estornell A, Baek D D, et al. Any-depth alignment: unlocking innate safety alignment of LLMs to any-depth [PP/OL]. arXiv(2024-10-20) [2026-01-05]. https://arxiv. org/abs/2510.18081
[16] Wang J X, Wu F Z, Li W D, et al. FATH: authentication-based test-time defense against indirect prompt injection attacks [PP/OL]. arXiv(2024-10-28) [2026-01-05]. https://arxiv. org/abs/2410.21492
[17] Wang Z L, Nagaraja N, Zhang L, et al. To protect the LLM agent against the prompt injection attack with polymorphic prompt [PP/OL]. arXiv(2025-06-06) [2026-01-05]. https://arxiv. org/abs/2506.05739
[18] Yi J W, Xie Y Q, Zhu B, et al. Benchmarking and defending against indirect prompt injection attacks on large language models [C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. ACM, 2025: 1809-1820. DOI: 10.1145/3690624.3709179
[19] Chen S Z, Wang Y Z, Carlini N, et al. Defending against prompt injection with a few defensive tokens [PP/OL]. arXiv(2025-07-10)[2026-01-05]. https://arxiv.org/abs/2507.07974
[20] Meta. Llama prompt guard 2 model card [EB/OL]. [2026-01-05]. https://github. com/meta-llama/PurpleLlama/blob/main/Llama-Prompt-Guard-2/86M/MODEL_CARD.md
[21] Microsoft. What is azure AI content safety? [EB/OL]. (2025-09-16) [2026-01-05]. https://learn. microsoft. com/en-us/azure/ai-services/content-safety/overview
[22] Inan H, Upasani K, Chi J F, et al. Llama guard: LLM-based input-output safeguard for human-AI conversations [PP/OL]. arXiv(2023-12-07) [2026-01-05]. https://arxiv. org/abs/2312.06674
[23] Abdelnabi S, Fay A, Cherubin G, et al. Get my drift? catching LLM task drift with activation deltas [PP/OL]. arXiv(2024-06-02)[2026-01-05]. https://arxiv.org/abs/2406.00799
[24] Zhang H W, Lu J, Jiang S Q, et al. Co-sight: enhancing LLM-based agents via conflict-aware meta-verification and trustworthy reasoning with structured facts [PP/OL]. arXiv(2025-10-24)[2026-01-05]. https://arxiv.org/abs/2510.21557
[25] Wang P R, Liu Y, Lu Y F, et al. AgentArmor: enforcing program analysis on agent runtime trace to defend against prompt injection [PP/OL]. arXiv(2025-08-02)[2026-01-05]. https://arxiv.org/abs/2508.01249
[26] Invariant Labs. Invariant [EB/OL]. [2026-01-05]. https://invariantlabs.ai/
[27] Google. gVisor [EB/OL]. [2026-01-05]. https://gvisor.dev
[28] Google Cloud. Isolate AI code execution with Agent Sandbox [EB/OL]. [2026-01-05]. https://docs. cloud. google. com/kubernetes-engine/docs/how-to/agent-sandbox
[29] WASI. The webassembly system interface [EB/OL]. [2026-01-05]. https://wasi.dev/
[30] Microsoft. Introducing Wassette: WebAssembly-based tools for AI agents [EB/OL]. (2025-08-06) [2026-01-05]. https://opensource. microsoft. com/blog/2025/08/06/introducing-wassette-webassembly-based-tools-for-ai-agents
[31] Andriushchenko M, Souly A, Dziemian M, et al. AgentHarm: a benchmark for measuring harmfulness of LLM agents [PP/OL]. arXiv(2024-10-11) [2026-01-05]. https://arxiv. org/abs/2410.09024
[32] Zhang Z X, Cui S Y, Lu Y D, et al. Agent-SafetyBench: evaluating the safety of LLM agents [PP/OL]. arXiv(2024-06-19)[2026-01-05]. https://arxiv.org/abs/2412.14470
[33] Debenedetti E, Zhang J, Balunović M, et al. AgentDojo: a dynamic environment to evaluate prompt injection attacks and defenses for LLM agents [PP/OL]. arXiv(2024-10-11)[2026-01-05]. https://arxiv.org/abs/2406.13352

作 者 简 介

闫新成，中兴通讯股份有限公司首席安全架构师，江苏省产业教授，正高级工程师；主要研究方向为5G/6G安全、AI安全；从事电信行业20年，曾主持国家科技重大专项课题，获得多项省部级科技奖励；拥有专利40余项。

刘东，中兴通讯股份有限公司副总裁、中心研究院副院长；主要从事操作系统和网络安全领域的技术研究和经营管理工作。

李旻旻，中兴通讯股份有限公司技术预研工程师；主要研究方向为主机入侵检测、AI内容安全、智能体安全、可信与机密计算等。

吴建华，中兴通讯股份有限公司技术预研工程师；主要研究方向为主机安全、智能体安全、AIGC反欺诈等。

算力网关键技术与研究

Key Technologies and Research of Computing Power Network

胡晓女/Hu Xiaonyu[1,2]，陆璐/Lu Lu[3]，李涛/Li Tao[4]，
雷波/Lei Bo[5]，唐琴琴/Tang Qinqin[6]，张宏科/Zhang Hongke[7]
（1. 澳门科技大学，中国 澳门 999078；
2. 中国通信学会，中国 北京 100846；
3. 中国移动通信有限公司研究院，中国 北京 100053；
4. 中国联合网络通信有限公司研究院，中国 北京 100037；
5. 中国电信股份有限公司研究院，中国 北京 102200；
6. 北京邮电大学，中国 北京 100876；
7. 北京交通大学，中国 北京 100091）
(1. Macau University of Science and Technology, Macau 999078, China;
2. China Institute of Communications, Beijing 100846, China;
3. China Mobile Research Institute, Beijing 100053, China;
4. China Unicom Research Institute, Beijing 100037, China;
5. China Telecom Research Institute, Beijing 102200, China;
6. Beijing University of Posts and Telecommunications, Beijing 100876, China;
7. Beijing Jiaotong University, Beijing 100091, China)

DOI：10.12142/ZTETJ.202601011
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260225.0924.004
网络出版日期：2026-02-25
收稿日期：2025-12-16

摘要：随着人工智能与数字经济的深度融合，传统算网相对独立的架构难以满足计算任务对高性能、实时性及跨域资源共享的需求。将算力网（CPN）定义为以计算为核心、网络为基础、智能为引擎的新型基础设施，系统探讨了其关键技术创新与发展实践。详细阐述了算力路由、高通量数据网、分布式智算组网、智融标识网络、星织网络架构以及算力互联测量感知等六大核心技术体系，并通过现网试点与规模验证，验证了这些技术在提升网络吞吐率、降低端到端时延及优化异构资源调度方面的显著成效。最后，围绕高效基础设施建设、跨域跨平台调度、智能化管理、多样化场景适配以及隐私安全与绿色节能等5个维度，提出了CPN后续研究的重点方向与建议。

关键词：算力网架构；算网感知；算网协同调度；算网运维

Abstract: With the deep integration of artificial intelligence and the digital economy, the traditional architecture—where computing and networks operate relatively independently—struggles to meet the demands of computing tasks for high performance, real-time response, and cross-domain resource sharing. This paper defines the computing power network (CPN) as a new type of infrastructure that is computing-centric, network-based, and intelligence-driven, and systematically explores its key technological innovations and development practices. This paper elaborates on six core technology systems: computing power routing, high-throughput data network, distributed intelligent computing networking, intelligence-converged identification network, star-fabric network architecture, and computing power interconnection measurement and awareness. Through live network pilots and large-scale verifications, the study demonstrates the significant effectiveness of these technologies in improving network throughput, reducing end-to-end latency, and optimizing heterogeneous resource scheduling. Finally, this paper proposes key directions and suggestions for future research on the CPN, focusing on five dimensions: efficient infrastructure construction, cross-domain and cross-platform scheduling, intelligent management, diverse scenario adaptation, and privacy security combined with green energy conservation.

Keywords: computing power network architecture; computing-network awareness; computing-network collaborative scheduling; computing-network operation and maintenance

引用格式：胡晓女, 陆璐, 李涛, 等. 算力网关键技术与研究 [J]. 中兴通讯技术, 2026, 32(1): 68-78. DOI: 10.12142/ZTETJ.202601011
Citation: Hu X N, Lu L, Li T, et al. Key technologies and research of computing power network [J]. ZTE technology journal, 2026, 32(1): 68-78. DOI: 10.12142/ZTETJ.202601011

在数字化转型的浪潮中，人工智能（AI）已成为驱动新质生产力的核心动力，对国家现代化经济体系构建及高质量发展起到关键支撑作用。传统算网以计算和网络资源相对独立的方式提供服务的架构模式，二者之间缺乏深度协同[1]。在AI、大数据等对算网需求具有动态性和不确定性的应用场景下，传统架构难以快速响应和灵活调整，无法满足计算任务对实时响应和海量数据处理的高性能要求。此外，传统架构在跨平台、跨区域的算力资源

共享和调度方面存在显著局限，限制了大规模分布式计算的能力[2]。

面对这些问题，算力网作为一种新型的信息基础设施应运而生[3-4]。算力网以算为中心、网为根基、智为引擎，期望达成“算力无所不在，网络无所不达，智能无所不及”的愿景目标[5]。为进一步加快算力网产业发展，不断推动算力网创新成果应用落地，本文研究了算力网关键技术及其实践，内容涵盖规模验证、产业推进、应用创新、平台建设等，并探讨了算力网的未来研究方向。

1 算力网关键技术

算力网的构建不仅依赖于传统网络技术的延伸，更需要在路由、调度、互联、测量等多个技术层面进行创新。本章将围绕算力网的核心需求，系统介绍一系列关键技术，包括算力路由技术、服务感知技术、测量感知技术、新型交换互联架构等。这些技术相辅相成，共同构建算力网的技术体系，共同推动算力资源的高效利用与智能化发展。

1.1 算力路由技术

面向算力网全新发展理念和算网一体发展目标，算网融合理念要求在实现算网感知的基础上，同时考虑网络与计算资源状态，将流量动态引导到适当的服务节点上[6-7]，因此算力路由技术的实现面临以下3个关键挑战：如何定义高效封装的高维算力信息（“传什么”）、如何在保证传输实时性的同时降低通告开销（“怎么传”），以及如何设计多因子优化的路由决策机制避免路径不收敛（“怎么用”）[8]。

新型算力路由技术的核心理念是在传统网络路由机制的基础上引入算力因子[9]，通过扩展边界网关协议（BGP）路径属性，实现网络和计算资源状态的多维算力信息封装[10]，采用分层通告机制优化算力信息传播[11]，并结合算力感知的多因子路由算法，实现路径选择的全局最优，从而提升算网系统资源利用率，降低端到端业务时延，满足多样化的业务需求[12]。

1.1.1 技术创新与设备研制

算力路由技术在路由器设备的研发和功能优化上取得了关键性的创新突破。中国移动在2024年西班牙巴塞罗那世界移动通信大会（MWC）上发布了全球首台算力路由器（CATS Router）。该路由器基于现网通用的路由器平台（Net Engine、ZXR10）研制完成，支持互联网工程任务组算力路由（IETF CATS）工作组算力路由架构标准定义的功能模块和组件。同时，该设备在技术上实现了多项创新：通过可扩展归一化有效算力表征，解决了信息丰富性与扩展开销之间的矛盾；利用低开销自适应算力通告机制，提升通告效率并避免无效通告；采用有限决策域路由选路机制，赋予节点决策能力，有效降低“ping-pong”路由效应。

1.1.2 规模验证与应用落地

中国电信推出一种基于开源SONiC网络操作系统的创新算力网关，支持多种硬件平台，并具备业务感知、算力感知和算力路由功能。通过动态选择最优路径和服务节点，该网关实现了算力与网络资源的全局优化。

中国移动在河南、江苏、浙江、广东、河北等5省20地市开展了该算力网关的集中式、分布式部署及新型地址族等技术验证。实验显示，该技术在高清视频内容分发网络（CDN）场景下使算网容量提升37.5%，在云渲染场景中端到端平均时延减少24.5%，算力通告协议开销较传统机制降低20%，展现出显著的性能优化效果。

1.2 面向海量数据传输的高通量数据网技术

随着“东数西算”战略的落地与超智算业务的快速发展，TB级以上海量数据的跨广域网传输成为业界关注的热点和难点[13]。从用户角度来看，传统百兆带宽难以满足海量数据传输的时效性需求；从网络运营角度来看，多用户并发传输容易导致网络局部负载过高，影响网络质量；从传输效果来看，传统传输协议和用户侧存储性能的限制使得即便配置超大带宽，也难以充分利用带宽资源[14]。为此，高通量数据网技术应运而生，旨在通过优化传输效率、提升带宽利用率和降低传输成本，满足海量数据传输的需求[15]，高通量数据网架构如图1所示。

1.2.1 核心技术创新

1）广域流量调度与识别技术

高通量数据网的核心之一是广域流量的智能调度与识别。为此，需要在IP骨干网中，构建基于IPv6演进技术的广域承载底座[16]。首次采用SRv6网络编程[17]与应用感知网络技术，结合网络带宽、时延等服务等级协议（SLA）需求对数据流量进行标识，实现业务的快速开通、路径的确定性编排以及高通量数据的高效传输。

2）智能管控与端侧优化技术

智能管控技术通过网络状态感知技术，实时采集路径带

基金项目：2024年度全国学会服务国家战略专项（面向AI的算力网关键技术路线图）

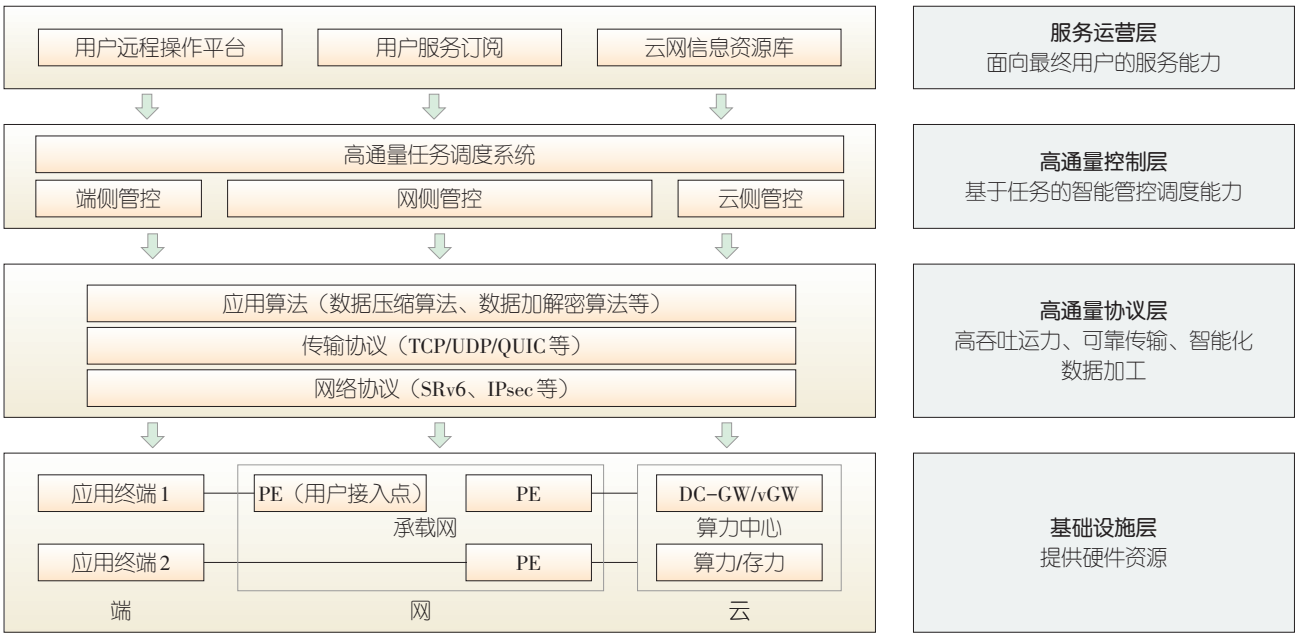


图1　高通量数据网架构

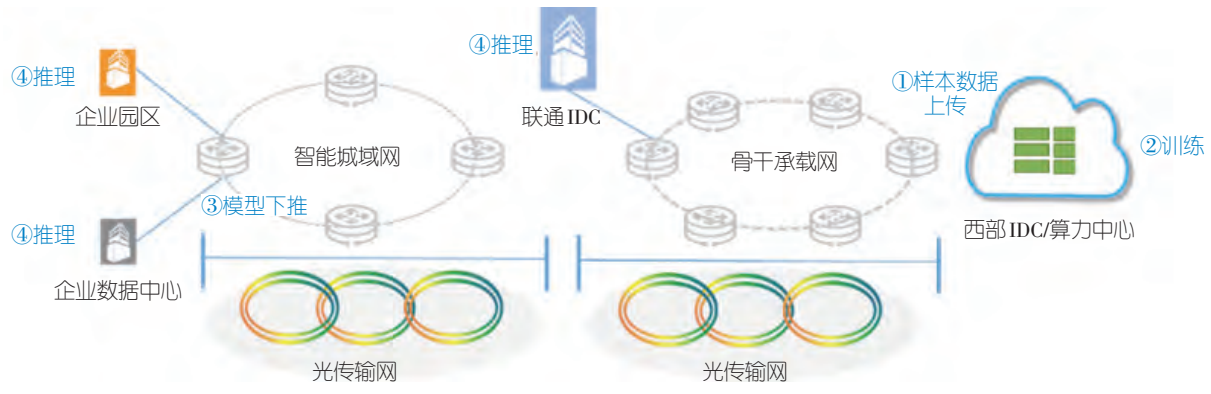
宽变化、流量趋势和资源信息，结合用户需求实现最优路径推荐和多样化套餐定制。端侧优化技术则通过传输控制协议（TCP）缓存、拥塞控制算法的改进以及数据压缩技术，提升传输速率，减少传输量，从而显著降低传输成本。中国联通提出了基于SRv6的任务式调度方案，并通过端网协同的传输协议优化技术，在广域网中实现了高效抗丢包传输和高效带宽利用率。

3）高通量数据传输系统

随着智算业务的不断发展，客户对传输模式的要求也趋于多样化。针对海量数据的传输需求，需要研发高通量数据传输系统，如图2所示。该系统通过数据智能压缩、端侧软硬件协同优化以及基于SRv6的智能选路技术，实现端到端的高效传输能力。高通量数据网支持一对一、一对多、多对一等多种传输模式的灵活定制，为客户提供智能调度和低成本的商业模式。

1.2.2 典型应用与验证场景

通过基于IPv6+的广域承载底座，高通量数据网实现了



IDC：互联网数据中心

图2　高通量方案现网测试示意图

路径灵活编排和调度，充分利用网络轻载链路和闲时带宽，显著提升了综合承载能力。以中国联通为例，其国际首创的高通量传输方案已在中国多地完成现网试点测试。在上海–宁夏的广域传输测试中，结合智算业务训练数据的典型“东数西算”场景，高通量数据网的创新能力得到了充分验证。测试结果表明，该方案可显著提升带宽利用率，缩短数据传输时延，并实现方案的可复制性推广。

1.3 面向智算场景的组网方案架构

针对多方异构算力资源纳管与可靠算网服务等核心挑战，本文提出一种创新的分布式智算中心组网架构，通过突破分布式资源调度、广域无损传输等关键技术[18]，构建了包含算力网关设备与交易管控平台的协同调度系统，实现了分散算力资源的高效连接、调度和利用[19]。这一系统为算力供需双方提供了最佳的资源分发、关联、交易与调配服务，显著优化了算网资源配置效率。

1.3.1 技术方案设计

1）广域无损组网技术

基于光传送网（OTN）的低时延、零丢包特性，结合长距无损流控机制，扩展了远程直接内存访问（RDMA）协议在广域网场景下的适用性[20]。通过优化流控机制和传输协议，确保了分布式智算中心在长距离传输中的无损性能，为跨地域智算资源的高效协同提供了技术保障。

2）异构资源调度技术

采用全局负载均衡算法[21]与多维度拓扑感知策略[22]，实现了跨地域异构算力资源的动态匹配与协同。该技术能够根据算力需求和网络状态，动态调整算力资源的分配，提升了算网资源的整体利用效率，并满足了多样化的智算业务需求[23–24]。

3） 超高速传输能力

基于800G C+L波段波分复用技术，构建了大容量全光底座，为千卡级智算集群提供了端到端的超宽连接能力。该技术通过提升传输带宽和优化光传输性能，为分布式智算中心的高效运行提供了强大的传输支撑[25]。

1.3.2 实践验证成果

在智算领域，中国电信的全光运力网基于800G C+L技术、异构网络集合通信优化技术和全局负载均衡技术，为1 024卡规模的分布式集群提供大容量带宽支持，实现120 km范围内千亿参数大模型的分布式训练，如图3所示。测试结果显示，分布式训练性能达到集中训练效果的95%以上，证实了分布式无损智算网技术方向的可行性，为智算互联构建了坚实技术底座。

1.4 面向云网融合的智融标识网络体系

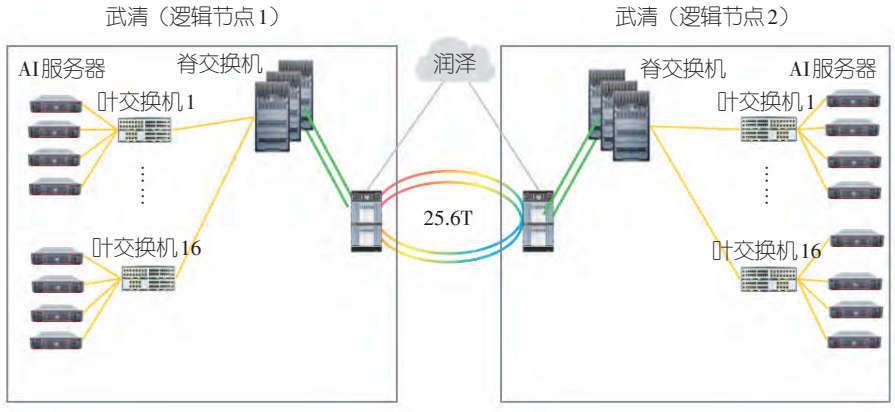
随着新基建与“东数西算”重大工程的启动，建设以云网融合为核心[26]的新型信息基础设施是国家核心战略需求，构建异构网络深度融合、算网深度融合、完全自主可控的新型网络体系迫在眉睫[27]。然而，传统互联网标识体系因其原始设计的局限性，工作机制相对“静态、僵化”，在标识体系、服务模式、资源管控等方面难以满足当前云网融合发展提出的“新业务、新网络、新计算、新管控”需求[28–29]。

针对云网深度融合需求，业界提出了以高级智能、多维标识、算网协同、异构融合为典型特征的智融标识网络体系与技术[30]，该体系基于“三层、三域”的智融标识网络体系理论与总体架构，攻克了富语义多维融合标识、异构资源动态协同汇聚等关键技术难题，实现了基于标识映射的异构云网资源深度融合与高效协同。

1.4.1 技术方案设计

1）面向云网融合的智融标识网络体系

智融标识网络体系以“三层、三域”架构为基础，异构融合组网为纽带，智能资源感知[31]与调度[32]为核心，聚焦新型标识体系及解析映射关键技术。通过高效的网络按需自组、资源智能感知、服务协同编排，该体系实现了新业务灵



AI：人工智能

图3 京津冀智算机房千卡120 km绕行拉远验证组网

活部署、新网络按需构建、新计算弹性适配、新管控智能运营，为云网融合提供传–算–存–感等多维能力的一体化支撑[33]。智融标识网络体系的示意架构如图4所示。

2）富语义多维融合标识及映射方法

智融标识网络体系构建了具有多维度表征能力的网络层标识体系，在标识空间实现终端、网络、服务、数据、算力等异构对象的多维度融合标识与表征[34]。通过引入标识空间可编程理念，可实现标识空间的多维重用，从而解决传统网络标识服务语义承载单一的问题。提出将基于意图的多维属性描述作为通信服务的基本原语，设计了统一的服务接口，实现服务与网络的语义连通。这能够将灵活、复杂的服务需求高效承载于网络层上。

3）异构资源动态协同汇聚技术

基于差异化服务与资源的高效映射方法，通过解耦服务和网络资源，在标识网络中建立一套涵盖服务收集、策略适配、网络对象的动态量化映射机制，实现服务类别与网络资源的按需动态适配。提出碎片化网络资源汇聚融合与细粒度优化调度方法，设计多维状态演化模型，通过数据精准调度适配，实现异构网络深度互通互融的高效传输，保障差异化服务的网络性能。

1.4.2 实践验证成果

北京交通大学基于中国电信云网融合大科创装置，设计并验证了业界首个支持多维融合标识的新型智融标识网络系统，如图5所示。该系统实现了算力网络广域按需确定性传输，在算力服务标识融合寻址、业网协同按需确定组网、跨异构网络协同可靠传输等方面展现了新质能力和显著优势。

1.5 面向AI大模型训练集群的星织网络架构及流量调优技术

在数据中心的AI训练、推理和云业务等领域，集群规模与计算效率是衡量集群有效算力最重要的指标。其中，确保网络的无丢包、高吞吐和低时延是实现高计算效率的关键。然而网络拥塞和负载不均是实现该目标的主要挑战[35]。具体而言，在AI智算、存储等分布式应用中普遍存在多访问一的流量模型，极易引发微秒级网络拥塞，导致业务时延增大甚至通信丢包。此外，AI训练典型流量特征是少量同步突发的大流，极致低熵，传统等价多路径路由（ECMP）哈希选路机制失效，流量冲突严重，网络链路忙闲不均，有效吞吐低至20%–50%。

针对这些挑战，业界提出面向AI大模型训练集群的星

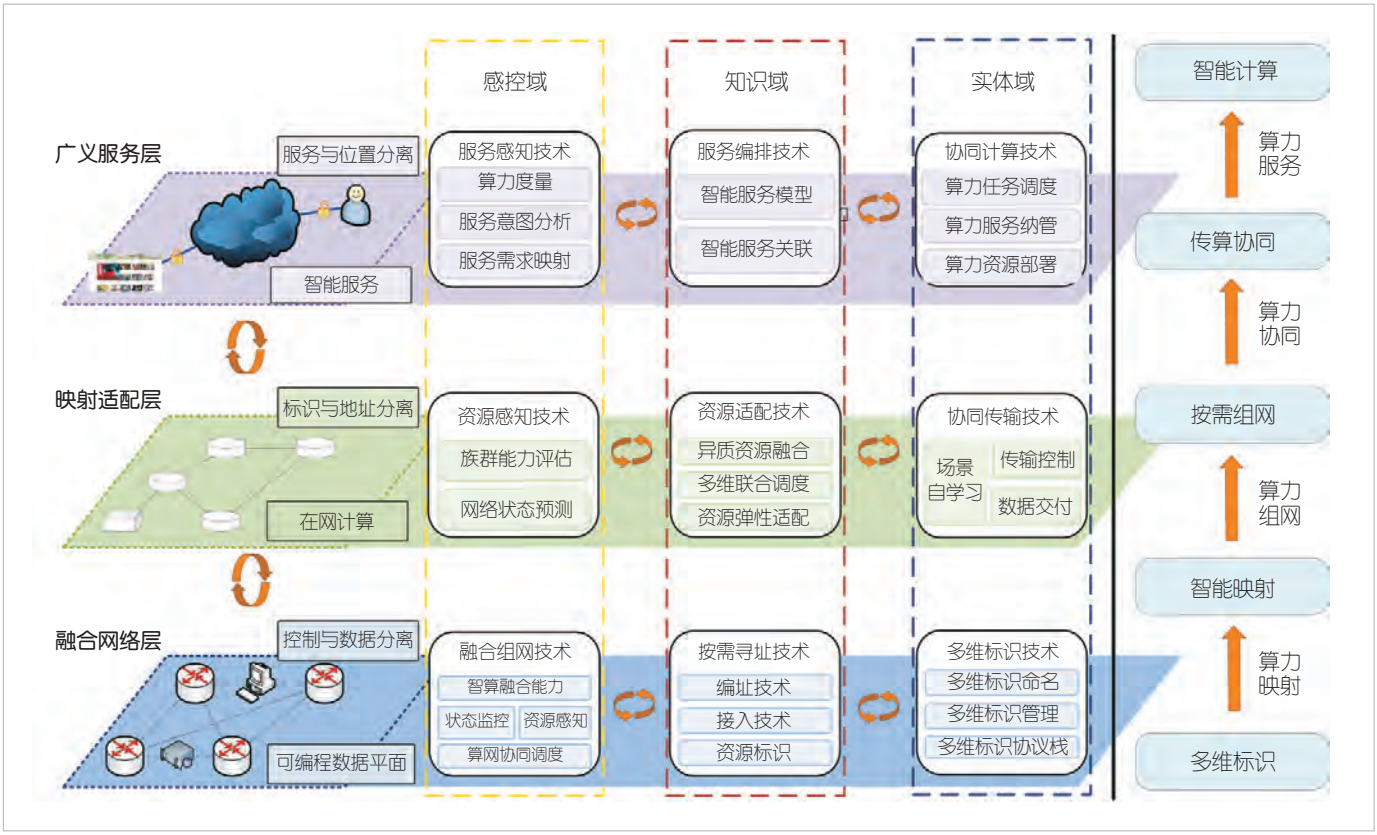
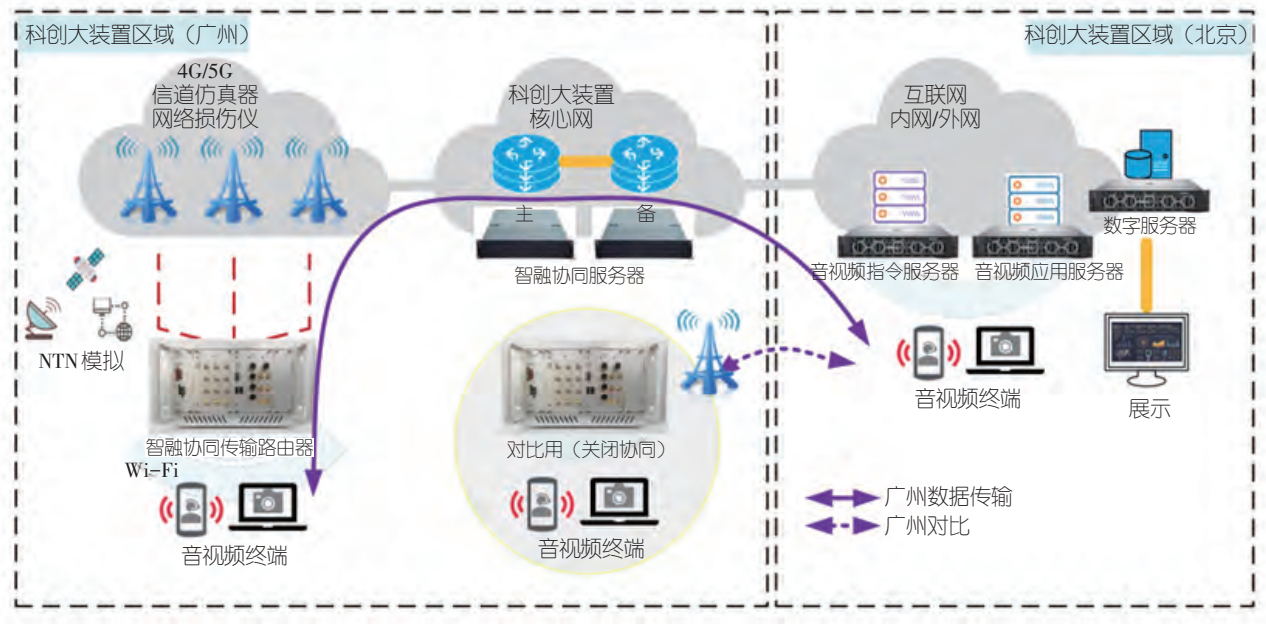


图4　面向云网融合的智融标识网络体系



NTN：非地面网络技术

图5　基于中国电信云网融合科创大装置的智融标识网络应用验证

织网络架构。该架构通过超大规模集群组网和自适应流量调优技术，解决了大规模网络设备间拥塞与负载不均问题，实现了数据中心网络无损不丢包的目标，同时保障了网络高利用率和微秒级低延迟性能，从而使计算有效算力达到最优。

1.5.1 技术方案设计

1）星织网络架构

针对AI训练大规模的诉求，华为提出了星织网络架构，如图6所示。该架构可以支撑百万规模集群组网和跨数据中心（DC）算力互联。结合新的自适应路由算法，利用AI大模型流量大流性、并发性、可预测性的固有特征，该架构构建了分布式路由机制，可实现在交换机分布式局部决策下，近乎全局最优网络利用率，同时降低了组网成本和功耗。

2）AIECN技术

为了解决网络拥塞难题，AI增强拥塞通知（AIECN）创造性地引入了分布式多智能体技术，具有较强的泛化能力，通过在线和离线训练相结合的方式，可利用交换设备Telemetry功能在不同场景中实现快速部署，最终实现了整个网络高利用率和微秒级低延迟的极致性能。AIECN算法框架如图7所示。

3）NSLB技术

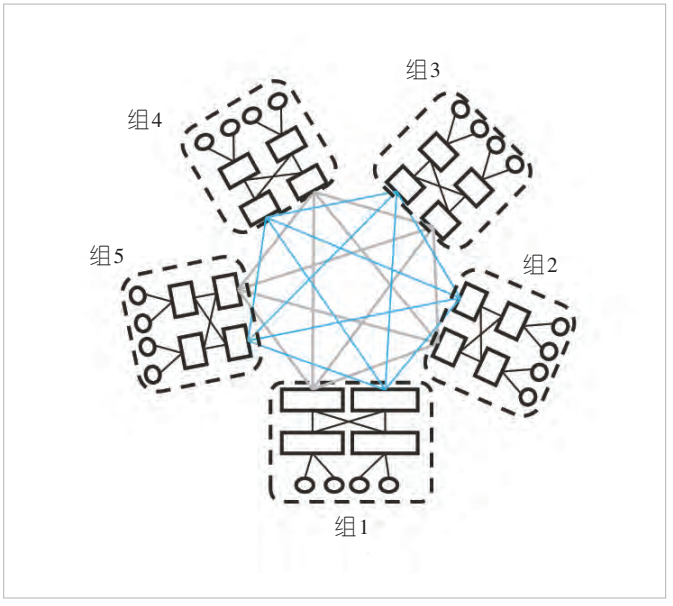
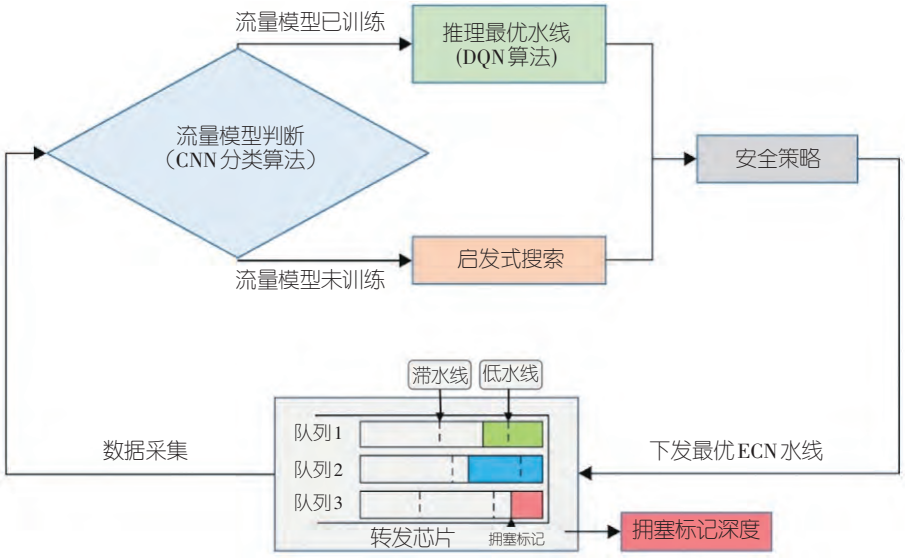


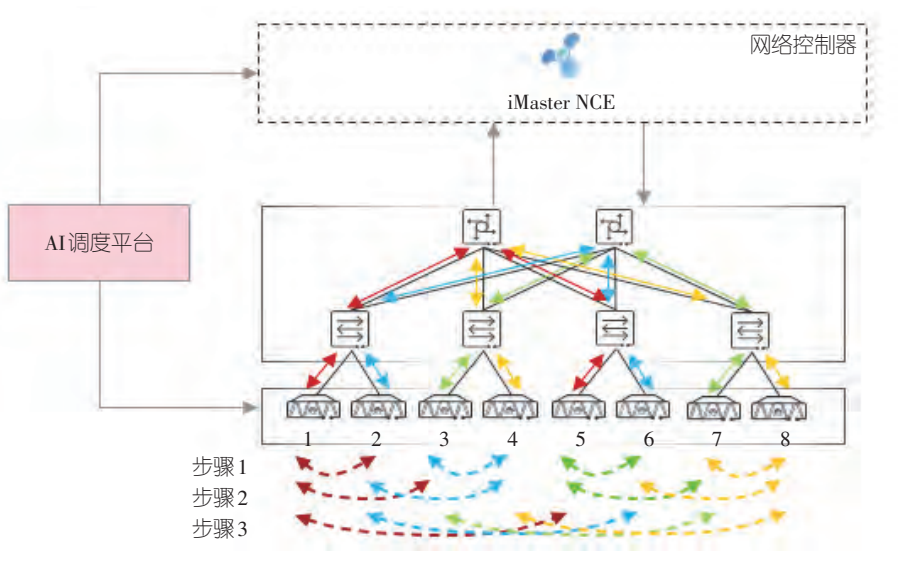
图6　星织网络架构示意图

网络调度负载均衡（NSLB）技术通过端网融合，独创了亲和性调度与集中算路算法，解决了全网流量不均的问题。与传统负载均衡算法相比，NSLB显著提升了大规模分布式应用网络的高吞吐、低时延和零丢包能力。NSLB算法框架如图8所示。



CNN：卷积神经网络　DQN：深度 Q 网络　ECN：显式拥塞通知

图 7　AI ECN 算法框架



AI：人工智能　NCE：网络云化引擎

图 8　网络调度负载均衡(NSLB)算法框架

1.5.2 实践验证成果

华为基于星织网络架构与流量调优技术推出了超融合解决方案——星河 AI 智算网络星织架构，并成功应用于金融、政府、互联网、能源等行业的数据中心。该架构支持超大规模 400GE 集群，网络规模是传统 CLOS 架构的 4 倍，通过扁平化设计减少设备和光模块使用量，设备数量降低 20%，显著降低网络能耗。

星织架构结合精准流控与负载均衡，网络吞吐率超过 95%，算力效率提升 10%，并支持跨数据中心高效互联。通过闪启技术与抗损检测功能，实现月级训练任务的“零中断”，在复杂网络环境中提供高可靠性与稳定性，大幅提升研发效率与能效表现。

1.6 基于测量感知与新型交换互联架构的算力互联技术

算力互联是实现全国算力一体化布局的关键路径，而跨算力中心的网络质量直接影响了算力互联产业化进程[36]。针对跨算力中心互联场景中存在的网络质量波动、测量成本高、异构资源调度低效等问题[37]，本文提出了基于测量感知与新型交换互联架构的算力互联技术[38]。该技术从新型架构设计、智能算法优化与平台构建 3 个层级，系统性地解决了算力互联中的关键技术难题。

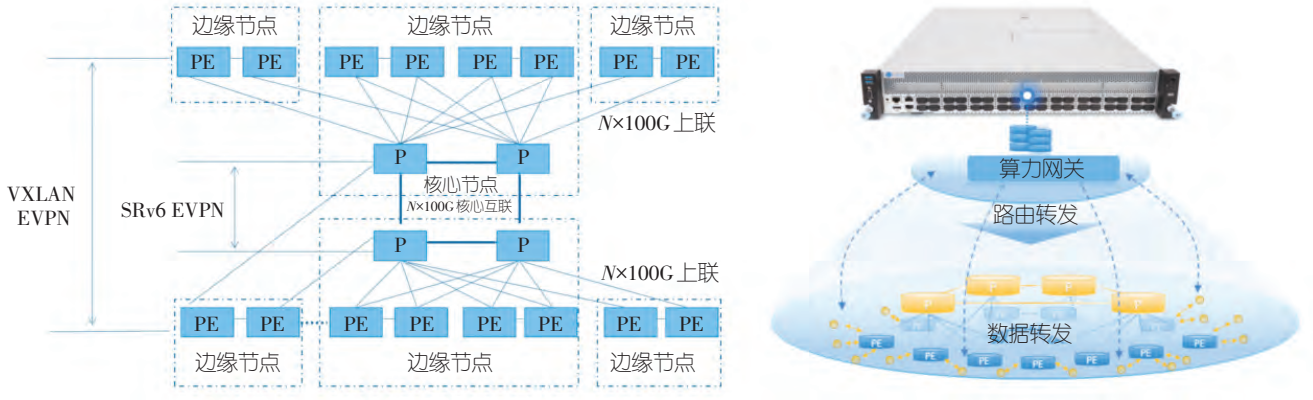
1.6.1 技术方案设计

1）新型交换互联架构

基于新型交换互联架构的算力互联网络采用大二层交换互联架构[39]，如图 9 所示，通过算力网关实现通算、智算等多元异构算力中心的一跳直达互联，突破了传统 Internet 或点对点直联方案的网络质量与成本方面的瓶颈，且支持异构算力资源的扁平化调度[40]，消除多层协议转换开销[41]，使能通、超、智等多元异构算力中心一跳直达，网络时延下降显著[42]。

2）稀疏感知测量算法

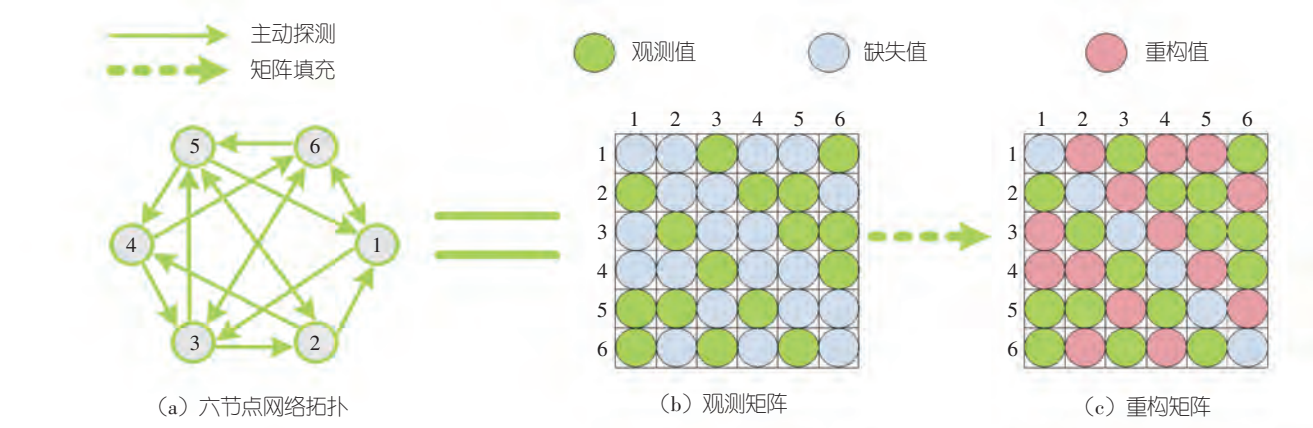
分布式计算任务中约 26.1% 的故障为网络问题，因此进行全网质量测量至关重要。但全网测量开销极高，重则会阻塞业务。为此，本文提出了基于稀疏感知的低开销算网测量理论和算法，如图 10 所示，通过局部稀疏采样数据精准重



（a）大二层新型交换互联架构　　（b）算力网关示意图

EVPN：以太虚拟专用网　P：服务提供商路由器　PE：运营商边缘设备　SRv6：新一代IP承载协议　VXLAN：虚拟扩展局域网

图9　大二层新型交换互联架构与算力网关示意图



（a）六节点网络拓扑　　（b）观测矩阵　　（c）重构矩阵

图10　基于稀疏感知的低开销算网测量理论和算法

构全网性能状态，并设计在线/离线双模时空网络异常检测模型，结合动态阈值调整机制，实现了高精度异常定位与低误判率。

1.6.2 实践验证成果

湖南大学谢鲲教授团队联合多方搭建了基于新型交换互联架构的算力互联网络，并取得重要突破。在网络时延方面，该算力互联方案将东西向调度网络时延降低77%，显著提升了跨算力中心的传输效率。在测量成本上，该方案通过稀疏感知的低开销测量理论与算法，将测量代价降低70%，实现了高效网络性能推理。

同时，团队研发的高阶时空网络异常检测模型提升了检测精度，使模型误判率下降幅度超过50%。这些成果验证了新型交换互联与测量感知技术在提升网络性能和资源调度效率方面的有效性，为全国算力一体化布局提供了强有力的技术支撑。

2 未来研究方向

随着AI技术的快速发展，AI应用场景已渗透至智慧城市、智能制造、自动驾驶、医疗健康等诸多领域，在推动各行业数字化转型的同时，也对算力网络提出更高要求。传统架构难以满足大规模分布式计算和跨平台资源调度需求，亟需围绕以下维度展开关键技术攻关：

1）高效算力基础设施建设关键技术

需构建异构资源整合与多级协同体系，通过通信网络整合中央处理器（CPU）、图形处理器（GPU）、现场可编程门阵列（FPGA）等算力形成弹性资源池，结合多层次部署实现中心云–边缘云–终端的协同计算，支撑低延迟场景。突破分布式存储与并行计算技术，优化数据存取效率，推动新型通信协议与AI计算深度融合，提升系统交互效能。同时，建立算力需求测算模型，从算力、存力、运力3个维度精准评估基础资源，为动态调度提供基准。

2）跨区域和跨平台的算力调度关键技术

跨域调度能力的提升是算力网络高效运行的关键，依赖于智能协同机制的创新。针对多区域、多平台特征，需研发支持跨云/跨数据中心任务分配的技术体系，通过AI驱动的动态负载均衡算法实现资源自适应优化。借助无损网络与加速技术，可突破大规模AI任务中的网络瓶颈。例如，在分布式训练场景中，智能调度与网络优化的耦合能将任务执行效率提升30%以上。这要求管控协议支持跨域光网实时信息采集，构建高动态信令传输通道。

3）算力网络的智能化管理关键技术

智能化管理体系的构建是算网自治的核心。通过算网智能体架构设计，实现资源编排、故障预测等功能的AI内生化，同时要求运营管理大模型需具备多模态数据处理与复杂决策能力。在实践层面，AI驱动的资源管理不仅需要实时监控系统，更要建立面向5G切片、算力突增等场景的弹性伸缩机制。研究表明，引入数字孪生技术可使资源调度响应速度提升40%，而智能运维系统能缩短30%的故障处理时长。

4）支持多样化AI应用场景关键技术

算力网的普适性在很大程度上取决于该技术对多样化应用场景的适配能力，这也是算力网架构设计的重要方向。需深度解构智慧城市、工业互联网等场景的差异性需求：自动驾驶强调毫秒级时延保障，智能制造关注计算精度与稳定性平衡，智慧城市则需海量终端接入能力。这要求算网架构支持模块化定制，例如通过边缘节点动态组网满足车路协同需求，或采用存算一体架构优化工业质检场景效率。

5）隐私安全与绿色节能技术

安全与能效构成可持续发展双翼，是算力网可持续发展的重要支柱。隐私计算技术需实现联邦学习与差分隐私的有机融合，在医疗AI等敏感场景构建数据可用不可见的保护机制。网络层需创新轻量级加密算法，在千亿级参数传输场景下维持加密效率。在绿色节能方面，AI赋能的能效优化可动态调整算力节点运行状态，结合液冷散热、芯片级功耗管理等技术，整体电源使用效率（PUE）可降至1.1以下，同时智能休眠策略可实现边缘设备30%的能耗降幅。

综上所述，算力网的未来发展需要多维度的协同创新，才能实现智能化、弹性化和安全化的全面升级。这不仅是支撑AI 2.0时代应用爆发的基础，更是推动各行业迈向高效、可持续发展的关键所在。通过在算力基础设施、跨域调度、智能管理、安全与能效等方面的持续突破，算力网络将为智慧社会的构建提供坚实的技术底座，助力AI技术在更广泛领域的深度应用与价值释放。

3 结束语

算力网在助力“网络强国”和“数字中国”建设中具有重要意义，对深入实施“东数西算”工程，加快构建全国一体化算力网络至关重要。为此，本文围绕算力网关键技术展开深入分析，聚焦技术标准、体系架构、融合创新、运维调度及优化等方面，内容涵盖关键技术突破、实验验证、产业推进和应用创新，并提出了算力网领域后续研究的方向与建议，以期为该领域发展提供有价值的参考。

参考文献

[1] 邢文娟, 雷波, 赵倩颖. 算力基础设施发展现状与趋势展望 [J]. 电信科学, 2022, 38(6): 51–61. DOI: 10.11959/j. issn. 1000–0801.2022137
[2] 中国移动. “九州”算力互联网（MATRIXES）目标架构白皮书 [R]. 2024
[3] 张宏科, 于成晓, 权伟, 等. 融算网络体系基础研究 [J]. 电子学报, 2022, 50(12): 2928–2934. DOI: 10.12263/DZXB.20221140
[4] 吴帅, 韩振东, 王佳, 等. 运营商视角下的算力网络技术及实践研究 [J]. 信息通信技术与政策, 2024, 50(2): 40–45
[5] 曹畅, 刘莹. 算力网络发展现状与展望 [J]. 通信世界, 2021(23): 34–35. DOI: 10.13571/j.cnki.cvvv.2021.23.012
[6] Wang X Y, Duan X D, Yao K H, et al. Computing–aware network (CAN): a systematic design of computing and network convergence [J]. Frontiers of information technology & electronic engineering, 2024, 25(5): 633–644. DOI: 10.1631/fitee.2400098
[7] Wen W, Lu L, Xie R C, et al. Secure incentive mechanism for energy trading in computing force networks enabled Internet of vehicles: a contract theory approach [J]. The journal of supercomputing, 2024, 80(18): 26061–26087. DOI: 10.1007/s11227–024–06369–2
[8] 付月霞, 陆璐, 刘鹏. 算网一体调度现状、挑战和分析. 中国计算机学会通讯, 2024, 20(1): 31–33
[9] 牟彦, 姚柯翰, 刘鹏, 等. 面向工业互联网的在网计算加速技术 [J]. 自动化博览, 2024, 41(2): 39–42. DOI: 10.3969/j. issn. 1003–0492.2024.02.034
[10] 姚柯翰, 刘鹏, 李志强, 等. 基于在网计算NACA的边缘算力负载均衡方案 [J]. 自动化博览, 2024, 41(2): 82–83. DOI: 10.3969/j. issn.1003–0492.2024.02.044

[11] Han M Y, Liu Y, Pang R, et al. Experimental verification of massive data transfer for super intelligent computing services [C]//Proceedings of 2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops). IEEE, 2024. DOI: 10.1109/ICCCWorkshops62562.2024.10693723
[12] Han M Y, Liu Y, Pang R, et al. Field trial of long-distance flexible large data transfer service based on IP and optical networks [C]// Proceedings of 2024 Asia Communications and Photonics Conference (ACP) and International Conference on Information Photonics and Optical Communications (IPOC). IEEE, 2024. DOI: 10.1109/ACP/IPOC63121.2024.10809685
[13] Tang X Y, Cao C, Wang Y X, et al. Computing power network: the architecture of convergence of computing and networking towards 6G requirement [J]. China Communications, 2021, 18 (2): 175-185. DOI: 10.23919/JCC.2021.02.011
[14] 曹畅, 唐雄燕. 算力网络关键技术及发展挑战分析 [J]. 信息通信技术与政策, 2021, 47(3): 6-11. DOI: 10.12267/j. issn. 2096-5931.2021.03.002
[15] 中国联通. 高通量数据网架构与关键技术白皮书 [R]. 2023
[16] 刘莹, 张帅, 李建飞, 等. 中国联通"IPv6+"创新探索与实践 [J]. 通信世界, 2022(23): 14-17. DOI: 10.13571/j.cnki.cww.2022.23.012
[17] 张帅, 曹畅, 唐雄燕. 基于SRv6的算力网络技术体系研究 [J]. 中兴通讯技术, 2022, 28(1): 11-15. DOI: 10.12142/ZTETJ.202201005
[18] 雷波, 刘增义, 王旭亮, 等. 基于云、网、边融合的边缘计算新方案: 算力网络 [J]. 电信科学, 2019, 35(9): 44-51. DOI: 10.11959/j. issn.1000-0801.2019209
[19] Huang X Y, Lei B, Ji G L, et al. Energy criticality avoidance-based delay minimization ant colony algorithm for task assignment in mobile-server-assisted mobile edge computing [J]. Sensors, 2023, 23(13): 6041. DOI: 10.3390/s23136041
[20] Xie Y P, Huang X Y, Li J C, et al. Computing power network: multi-objective optimization-based routing [J]. Sensors, 2023, 23(15): 6702. DOI: 10.3390/s23156702
[21] Lei B, Zhao Q Y, Mei J. Computing power network: an interworking architecture of computing and network based on IP extension [C]//Proceedings of IEEE 22nd International Conference on High Performance Switching and Routing (HPSR). IEEE, 2021: 1-6. DOI: 10.1109/HPSR52026.2021.9481792
[22] Li J C, Lv H, Lei B, et al. A computing power resource modeling approach for computing power network [C]//Proceedings of 2022 International Conference on Computer Communications and Networks (ICCCN). IEEE, 2022. DOI: 10.1109/ICCCN54977.2022.9868931
[23] Huang X Y, Lei B, Ji G L, et al. Multi-agent deep reinforcement learning-based incentive mechanism for computing power network [C]//Proceedings of International Conference on Emerging Networking Architecture and Technologies, Springer, 2022: 38-49
[24] Huang X Y, Lei B, Wei M, et al. Task value aware optimization of routing for computing power network [C]//Proceedings of 2023 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2023: 1-6. DOI: 10.1109/BMSB58369.2023.10211159
[25] Li J C, Lv H, Lei B, et al. A hierarchical routing mechanism for service in computing power network [EB/OL]. (2023-06-29) [2026-01-06]. https://dlnext. acm. org/doi/pdf/10.1145/3600061.3603120
[26] Zhang W T, Yang D, Zhang C, et al. (Com)^{2}Net: a novel communication and computation integrated network architecture [EB/OL]. (2024-09-24) [2026-01-06]. https://qiang-john-ye. github. io/Papers/Com2Net_A_Novel_Communication_and_Computation_Integrated_Network_Architecture.pdf
[27] Hou X D, Gao S, Liu N C, et al. L3Geocast: enabling P4-based customizable network-layer geocast at the network edge [J]. IEEE transactions on mobile computing, 2024, 23(8): 8323-8340. DOI: 10.1109/tmc.2023.3345933
[28] Liu Y, Zhang W T, Li L T, et al. Toward autonomous trusted networks-from digital twin perspective [J]. IEEE network, 2024, 38(3): 84-91. DOI: 10.1109/MNET.2024.3353180
[29] Yang D, Zhang W T, Ye Q, et al. DetFed: dynamic resource scheduling for deterministic federated learning over time-sensitive networks [C]//Proceedings of IEEE Transactions on Mobile Computing. ACM, 2024: 5162-5178. DOI: 10.1109/TMC.2023.3303017
[30] Xu Z H, Quan W, Liu M Y, et al. DOFMS: DRL-based out-of-order friendly multipath scheduling in mobile heterogeneous networks [C]//Proceedings of IEEE Transactions on Mobile Computing. ACM, 2024: 8274-8288. DOI: 10.1109/TMC.2023.3346480
[31] 柴若楠, 郜帅, 兰江雨, 等. 算力网络中高效算力资源度量方法 [J]. 计算机研究与发展, 2023, 60(4): 763-771. DOI: 10.7544/issn1000-1239.202330003
[32] Yang D, Cheng Z R, Zhang W T, et al. Burst-aware time-triggered flow scheduling with enhanced multi-CQF in time-sensitive networks [J]. IEEE/ACM transactions on networking, 2023, 31(6): 2809-2824. DOI: 10.1109/TNET.2023.3264583
[33] Quan W, Xu Z H, Liu M Y, et al. AI-driven packet forwarding with programmable data plane: a survey [EB/OL]. (2023-03-14) [2026-01-06]. https://uwaterloo.ca/scholar/sites/ca.scholar/files/sshen/files/quan2023ai.pdf
[34] 郜帅, 侯心迪, 刘宁春, 等. 多模态网络环境异构标识空间管控架构研究 [J]. 通信学报, 2022, 43(4): 26-35
[35] Liu S, Wang Q L, Zhang J Y, et al. In-network aggregation with transport transparency for distributed training [EB/OL]. (2023-03-14) [2026-01-06]. https://mcanini. github. io/papers/netreduce.asplos23.pdf
[36] Tian J Z, Xie K, Wang X, et al. Efficiently inferring top-k largest monitoring data entries based on discrete tensor completion [C]// Proceedings of IEEE/ACM Transactions on Networking. ACM, 2021: 2737-2750. DOI: 10.1109/TNET.2021.3103527
[37] Xie K, Li X C, Wang X, et al. Fast tensor factorization for accurate internet anomaly detection [J]. IEEE/ACM transactions on networking, 2017, 25(6): 3794-3807
[38] Cong Y C, Xie K, Wen J G, et al. Per-packet traffic measurement in storage, computation and bandwidth limited data plane [C]//Proceedings of IEEE/ACM Transactions on Networking. ACM, 2024: 3730-3742. DOI: 10.1109/TNET.2024.3404011
[39] Chen J G, Li K L, Tang Z, et al. A parallel random forest algorithm for big data in a spark cloud computing environment [J]. IEEE transactions on parallel and distributed systems, 2017, 28(4): 919-933. DOI: 10.1109/tpds.2016.2603511
[40] Tang Z, Du L F, Zhang X D, et al. AEML: an acceleration engine for multi-GPU load-balancing in distributed heterogeneous environment [J]. IEEE transactions on computers, 2022, 71(6): 1344-1357. DOI: 10.1109/TC.2021.3084407
[41] Ma T, Luo L, Yu H F, et al. Klonet: an easy-to-use and scalable platform for computer networks education [EB/OL]. (2024-04-16) [2026-01-06]. https://www. usenix. org/system/files/nsdi24-ma.pdf
[42] Xie J Z, Ma C X, Yu H F, et al. Analysis and optimization for passive one-way delay measurement tax in container networks [C]//Proceedings of IEEE 17th International Conference on Cloud Computing (CLOUD). IEEE, 2024: 247-255. DOI: 10.1109/cloud62652.2024.00036

作 者 简 介

胡晓女，澳门科技大学在读博士研究生，中国通信学会业务主管；主要研究方向为城市与区域经济可持续发展、算力网络、数字经济、工业互联网等。

陆璐，中国移动研究院基础网络技术研究所副所长，目前担任ITU-T SG13 WP1副主席、CCSA TC5 WG12核心网组组长、TC614网络5.0技术标准推进委员会副主席；长期从事移动核心网和算力网络的策略演进、技术研究、标准制订相关工作，在6G、算力网络及下一代互联网产业推动方面具有丰富经验。

李涛，中国联合网络通信有限公司研究院教授级高级工程师、中国科协科技人才奖项评审专家、中国通信学会AI技术与应用专委会委员、中国AI学会深度学习专委会委员；主要研究方向为网络智能化、AI大模型、智能算力等。

雷波，中国电信股份有限公司研究院网络技术研究所副所长，正高级工程师，中国科学院计算机网络信息中心客座研究员，北京邮电大学兼职教授；主要研究方向为未来网络技术、新型数据中心网络、边缘计算与算力网络等；出版专著4本，发表论文10余篇。

唐琴琴，北京邮电大学副研究员；主要研究方向为边缘计算、算力网络、网络AI、工业互联网等；发表论文20余篇。

张宏科，中国工程院院士、通信与网络技术专家，北京交通大学教授、博士生导师，移动专用网络国家工程研究中心主任，IEEE Fellow（2021年），教育部“全国高校黄大年式教师团队”带头人，享受国务院政府特殊津贴；长期从事专用通信网络理论与工程技术研究，建立了标识网络功能结构及解析映射机制，有效解决了复杂场景下网络高移动支持和高可靠传输难题，主持研制出专用网络设备与系统，为解决国家和行业专网工程难题做出重要贡献；获国家技术发明奖二等奖2项；出版专著6部。

卫星通信的极化码短码译码技术改进

Improvement of Decoding Technologies for Short Polar Codes in Satellite Communication

李春杰/Li Chunjie,马啸/Ma Xiao
（中山大学广东省信息安全技术重点实验室，中国 广州 510006）
（Guangdong Provincial Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou 510006, China）

DOI:10.12142/ZTETJ.202601012
网络出版地址：https://link.cnki.net/urlid/34.1228.tn.20260225.1016.008
网络出版日期：2026-02-25
收稿日期：2025-12-16

摘要：信道编码是保证通信可靠性的物理层关键技术，其中极化码是当前和未来一种重要的候选编码方案。首先对极化码基本原理和编译码技术进行概述，主要包括信道极化、极化码构造与编码、速率兼容方案和译码算法。然后，针对短极化码提出了一种串行消除列表（SCL）和阶序统计译码（OSD）级联的译码方案。与SCL相比，该方案在译码性能相当的情况下具有更低的复杂度。

关键词：极化码；信道极化；极化码构造；串行消除列表译码；阶序统计译码

Abstract: Channel coding is a key physical layer technology to ensure the reliability of communication, among which, the polar code is an important candidate coding scheme for the present and future. The primary principle of polar codes and their encoding and decoding techniques are reviewed, which mainly include channel polarization, polar code construction and coding, rate-compatible schemes, and decoding algorithms. Then, a cascaded decoding scheme of successive cancellation list (SCL) and ordered statistics decoding (OSD) is proposed for short polar codes. Compared with SCL, SCL-OSD can achieve similar decoding performance with lower complexity.

Keywords: polar code; channel polarization; polar code construction; successive cancellation list decoding; ordered statistic decoding

引用格式：李春杰, 马啸. 卫星通信的极化码短码译码技术改进 [J]. 中兴通讯技术, 2026, 32(1): 79-86. DOI: 10.12142/ZTETJ.202601012
Citation: Li C J, Ma X. Improvement of decoding technologies for short polar codes in satellite communication [J]. ZTE technology journal, 2026, 32(1): 76-86. DOI: 10.12142/ZTETJ.202601012

自2020年以来，5G已在全球范围内实现商用，显著提升了移动通信的传输速率、系统容量与时延性能。然而，5G主要针对地面网络，无法满足全球覆盖的需求。第3代合作伙伴计划（3GPP）已启动5G非地面网络（NTN）的技术标准化工作[1-2]，以进一步提升5G系统的覆盖范围。在NTN中，卫星通信是核心，是未来移动通信的关键技术。利用卫星通信覆盖范围广的特性，移动通信系统的覆盖范围可扩展至偏远山区、远洋、空中乃至太空等地面网络难以到达的区域。

卫星通信的信号在自由空间传输，容易受到气候、环境、距离等多种因素的影响。因此，为了提升卫星通信的可靠性，需要采用一些差错控制技术。信道编码可以提高信号传输的可靠性，是移动通信的物理层关键技术。信道编码的历史可以追溯到1948年。香农证明[3]，对于任何有噪声信道，当码率不超过信道容量时，使用信道编码技术可以实现无差错的数据传输。该理论定义了传输速率的极限，即信道容量，但没有给出接近该容量的码字传输的构造方法。基于香农的理论，许多学者提出了各种编码方法，其中比较优秀的码主要有卷积码、Turbo码和低密度奇偶校验码（LDPC）等。在蜂窝通信中，卷积码是2G控制信道和数据信道的编码方案，同时也是3G和4G的控制信道编码方案，Turbo码是3G和4G的数据信道编码方案，LDPC已被采纳为5G中数据信道编码方案。然而，这些信道编码方案都难以从理论上严格证明是渐进可达信道容量的。2009年，Arikan教授提出了极化码方案[4]，它是第一个可证明信道容量渐进可达的编码方案。自极化码提出以来，关于其构造方法、编码方案和译码算法等已取得了许多成果。当前，极化码已被采纳为5G控制信道的编码方案[5]。NTN作为5G以及未来移动通信的一部分，研究卫星通信下的极化码编译码技术具有重要的意义。

基金项目：国家重点研发计划项目（2021YFA1000500）

1 极化码基本原理

极化码通过信道极化，将N个独立信道极化为N个信道容量不同的子信道。即对任意一组可靠度相同的二进制输入离散无记忆信道（B-DMCs），经过信道联合和信道分裂后，在信道容量上会极化为两种子信道$W_N^{(i)}$，$i=1,2,\cdots,N$，一部分子信道容量$I\left(W_N^{(i)}\right)\to 1$，称为无噪信道，另一部分子信道容量$I\left(W_N^{(i)}\right)\to 0$，称为完全噪声信道。在编码时，信息比特在无噪信道上传输，而完全噪声信道传输固定比特。

信道极化现象可总结为下述极化定理：

对任意B-DMC W，当$N\to\infty$时，其中$N=2^n$，$n\geqslant 0$，对任意固定值$\delta\in(0,1)$，信道容量$I\left(W_N^{(i)}\right)\in(1-\delta,1]$的子信道占比趋近$I(W)$；信道容量$I\left(W_N^{(i)}\right)\in[0,\delta)$的子信道占比趋近于$1-I(W)$，其中$i=1,2,\cdots,N$。

图1展示了$N=1,2,4$时信道极化的过程。若信道为二进制擦除信道（BEC），擦除概率为ε，则经过信道极化后，子信道信道容量可由公式（1）和公式（2）迭代计算：

$$I\left(W_N^{(2i-1)}\right)=I\left(W_{N/2}^{(i)}\right)^2 \tag{1}$$

$$I\left(W_N^{(2i)}\right)=2I\left(W_{N/2}^{(i)}\right)-I\left(W_{N/2}^{(i)}\right)^2 \tag{2}$$

其中，$I\left(W_1^{(1)}\right)=1-\varepsilon$。图2展示了BEC在信道极化后，各子信道的信道容量分布，其中，码长$N=1\,024\,512\,256$，擦除概率$\varepsilon=0.5$。由图2可以看出，对不同长度的极化码，它们子信道的信道容量都具有两极分化的趋势，一部分子信道的信道容量趋于1，另一部分子信道的信道容量趋于0。另外，还有一部分子信道的容量在0和1之间，没有完全极化。这是因为极化码的码长不是无穷大，在有限码长下，会有部分信道不能完全极化。

2 极化码构造与编码

极化码的构造在于寻找可靠度高的子信道以传输信息比特。因此，计算子信道的可靠度是非常重要的。下面介绍几种重要的估计信道可靠度的方法。

2.1 极化码构造方法

巴氏参数构造法是Arıkan教授提出的第一种构造方法。对任意B-DMC W，定义巴氏参数如下：

$$Z(W)=\sum_{y\in\mathcal{Y}}\sqrt{W(y|0)W(y|1)} \tag{3}$$

在极化变换$\left(W_{N/2}^{(i)},W_{N/2}^{(i)}\right)\mapsto\left(W_N^{(2i-1)},W_N^{(2i)}\right)$中，各子信道的巴氏参数可以递归计算，具有下述关系：

$$Z\left(W_N^{(2i-1)}\right)+Z\left(W_N^{(2i)}\right)\leqslant 2Z\left(W_{N/2}^{(i)}\right) \tag{4}$$

$$Z\left(W_N^{(2i)}\right)=Z\left(W_{N/2}^{(i)}\right)^2 \tag{5}$$

其中，只有在BEC中，公式（4）的等号才满足，因此巴氏参数可以准确地估计BEC信道的可靠度。对于其他B-DMC，则只能得到错误概率的上界，并不能准确估计信道可靠性。

Mori等将密度进化（DE）用于对极化码子信道可靠度的估计[6]，该方法适用于对各种类型信道的可靠度估计。假设传输全0码字，令$a_N^{(i)}$表示第i个子信道对应对数似然比（LLR）的概率密度函数，则各信道的概率密度函数可由下式递归得到：

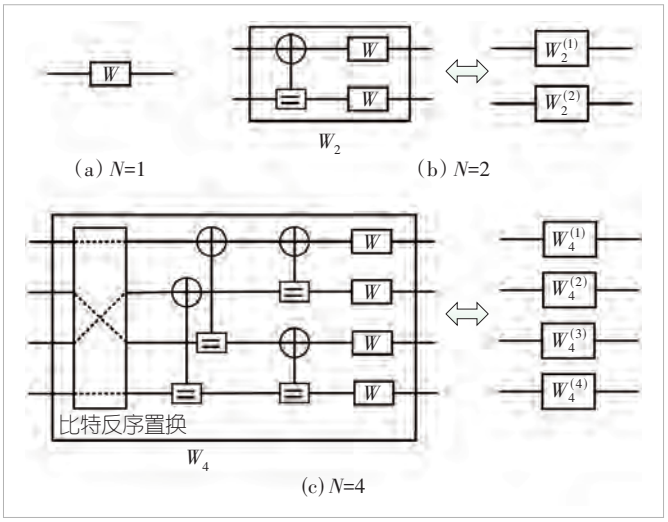


图1 $N=1,2,4$时信道极化过程

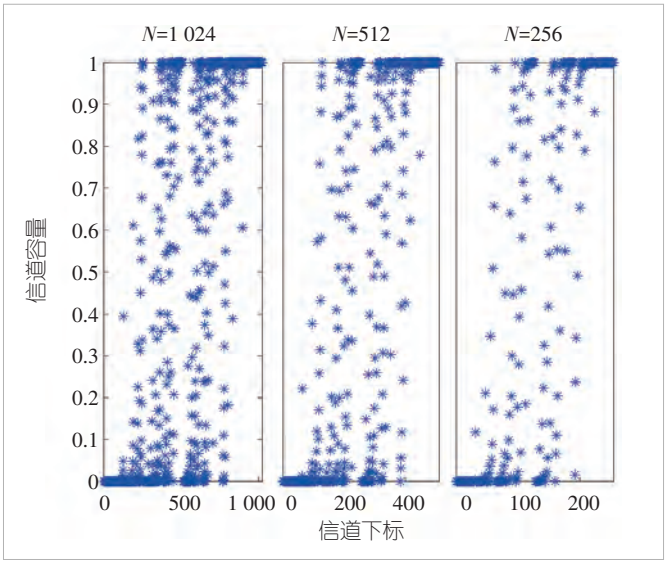


图2 子信道容量分布图（$\varepsilon=0.5$）

$$\begin{aligned} a_N^{(2i-1)} &= a_{N/2}^{(i)} \odot a_{N/2}^{(i)} \\ a_N^{(2i)} &= a_{N/2}^{(i)} * a_{N/2}^{(i)} \\ a_1^{(1)} &= a_\omega \end{aligned} \tag{6}$$

其中，$\odot$表示加号节点的卷积，$*$表示等号节点的卷积，a_ω表示原始信道W的LLR概率密度函数。由公式（6）可以方便地计算出各子信道LLR的概率密度函数，进而可以由公式（7）获得对应的错误概率：

$$P_e\left(W_N^{(i)}\right) = \int_{-\infty}^{0} a_N^{(i)}(x)\mathrm{d}x \tag{7}$$

密度进化包含大量卷积计算，复杂度较高。针对加性白高斯噪声（AWGN）信道，文献[7–8]提出了高斯近似（GA）方法。假设在信道中传输全0码字，采用二进制相移键控（BPSK）调制方式，即$x_i = 1 - 2c_i$，则每个接收值可表示为$y_i = 1 + z_i$，其中$z_i \sim \mathcal{N}\left(0,\sigma^2\right)$，则$y_i \sim \mathcal{N}\left(1,\sigma^2\right)$。第$i$个LLR值$L_i \sim \mathcal{N}\left(2/\sigma^2,4/\sigma^2\right)$，即LLR满足方差是均值两倍的高斯分布。GA方法正是基于上述事实，即在极化变换$\left(W_{N/2}^{(i)},W_{N/2}^{(i)}\right) \to \left(W_N^{(2i-1)},W_N^{(2i)}\right)$中，3种极化信道对应的LLR值具有方差是均值两倍的高斯分布，因此只需要计算各节点处的均值即可。

上述信道估计方法均和信道参数密切相关，在不同的信道条件下需要重新进行信道估计，这无疑提高了计算复杂度。文献[9–10]提出了一种和信道条件无关的估计方法，称为极化重量（PW）。对任意极化信道下标i（$i = 1,2,\cdots,N$），$i-1$的n位二进制展开$\pi(i-1) = (b_n b_{n-1}\cdots b_2 b_1)$，从左到右对应最高位到最低位，则第$i$个极化信道的极化重量$PW_i$可定义为：

$$PW_i = \sum_{j=1}^{n} b_j \beta^j \tag{8}$$

其中，β可取值$2^{1/4}$。计算出各子信道对应的极化重量后，这些数值可以衡量子信道的可靠度，即数值越大，可靠度越高。

2.2 极化码编码

极化码编码方法主要分为非系统编码和系统编码。若极化码码长为N，信息位长度为k，根据2.1节所述极化码构造方法，选出k个可靠度最高的子信道，对应的下标构成信息位集合$\mathcal{A}$，用来放置信息比特，剩余子信道下标构成冻结位集合$\mathcal{A}^c$，并放置固定比特（一般放置全0比特）。

非系统编码过程可定义如下：

$$c_1^N = u_1^N \boldsymbol{G}_N \tag{9}$$

其中，$\boldsymbol{G}_N = \boldsymbol{B}_N \boldsymbol{F}^{\otimes n}$，$\boldsymbol{B}_N$是比特反序置换矩阵，$\boldsymbol{F} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$是极化码核矩阵，$\otimes$表示Kronecker操作，$u_1^N = (u_1,u_2,\cdots,u_N)$和$c_1^N = (c_1,c_2,\cdots,c_N)$分别表示信息序列和码字。

Arıkan教授在文献[11]中指出，系统极化码相比非系统极化码具有低误比特率的特点。系统极化码是指，在传输的码字中信息比特和冗余比特是相互分离的，可以直接提取信息比特。文献[11]给出了一种简单的系统编码方法，即直接将信息比特放在码字中和集合$\mathcal{A}$一样的位置上，码字中其他比特则可以根据极化码生成矩阵的特性推导出，最后得到完整的码字。

3 极化码速率兼容方案

实际的信道状况是在不断变化的，因此码长也需要具备灵活调整的能力。极化码受构造方法的约束，其码长只能为2的幂次，这限制了其在实际中的应用。主要有3种方法解决该问题，即凿孔、缩短和重复。对于重复方案，接收端在译码时将这些重复发送比特的LLR叠加后再译码即可。下面主要介绍凿孔和缩短方法。

3.1 准均匀凿孔

凿孔将编码后码字比特中的某些比特舍弃，达到调整码长的目的。文献[12]提出准均匀凿孔（QUP）方案，在现有凿孔方案中实现复杂度低且性能优秀。

令M表示极化码的实际码长，母码码长为N，M和N之间有如下关系：

$$N = 2^{\lceil \log_2 M \rceil} \tag{10}$$

其中，$\lceil\ \rceil$表示向上取整，则凿孔位数$N_p = N - M$。凿孔模式的定义如式（11）所示，它是一个N维向量，其中$p_i \in \{0,1\}$，$1 \leqslant i \leqslant N$，$p_i = 1$表示对应码字比特$c_i$是凿孔比特；$p_i = 0$表示$c_i$不是凿孔比特。

$$\boldsymbol{p} = \left(p_1 p_2 \cdots p_N\right) \tag{11}$$

对于QUP方案，其执行过程如下：先将凿孔模式初始化为元素全0的向量$(00\cdots0)$，再将前N_p个元素的值置为1，即

$$\boldsymbol{p} = (\underbrace{1\cdots1}_{N_p}00\cdots00) \tag{12}$$

若编码过程中进行了比特反序变换，则此时也需要对凿孔模式进行比特反序变换，经变换后构造的凿孔模式将呈现凿孔位准均匀分布的特征。对于凿掉的码字比特，由于接收端不知道其任何信息，在译码时将其LLR值设为0。

3.2 缩短

文献[13]针对极化码提出了一种缩短的方法。该方法通过直接设置信息序列中某些位为冻结位，使码字中产生已知的码字比特，可以将这些已知比特直接删除。这些删除的已知比特称为缩短比特。一种简单的做法是直接设置码字中最后N_p个位置为缩短位，其对应的缩短模式为：

$$\boldsymbol{p}=(0\cdots\underbrace{001\cdots11}_{N_p}) \tag{13}$$。

相应的冻结集合中应包括这些位置，即$\{M+1,M+2,\cdots,N\}\subseteq\mathcal{A}^c$。与QUP相同，若编码过程中进行了比特反序变换，则此时也需要对缩短模式进行比特反序变换。缩短比特对接收端是已知的，在译码时将这些缩短比特的LLR根据实际情况设为正无穷或负无穷。

4 极化码译码算法

假设采用BPSK调制，调制后的序列为$\boldsymbol{x}=(x_1,x_2,\cdots,x_N)$，其中，$x_i=(-1)^{c_i}\in\{\pm1\}$，$i=1,2,\cdots,N$。经过AWGN信道传输后，在接收端接收序列为$\boldsymbol{y}=(y_1,y_2,\cdots,y_N)$，其中，$y_i=x_i+z_i$，$i=1,2,\cdots,N$，$z_i\sim\mathcal{N}(0,\sigma^2)$。硬判决序列$\hat{\boldsymbol{c}}=(\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_N)$，其中，若$y_i\geqslant0$，则$\hat{c}_i=0$，否则$\hat{c}_i=1$。LLR序列$\boldsymbol{L}=(L_1,L_2,\cdots,L_N)$定义为：

$$L_i=\ln\left(\frac{\Pr(y_i|c_i=0)}{\Pr(y_i|c_i=1)}\right)=\frac{2y_i}{\sigma^2} \tag{14}$$，

其中，$\Pr(y_i|c_i=0)$和$\Pr(y_i|c_i=1)$是条件概率。

4.1 SCL译码算法

串行消除（SC）译码算法是极化码的基本译码算法，它根据极化码的因子图顺序译出各个信息比特。长为N的极化码因子图是一个有$n=\log_2N$个阶段和N层的规则结构。图3展示了$N=8$的极化码的因子图。极化码的因子图主要由加号节点和等号节点组成，因此每个节点对应的LLR值可分别由f函数和g函数决定。

图4展示了极化码因子图的一个基本单元，两个函数定义如下：

$$f(a,b)=\ln\frac{1+e^{a+b}}{e^a+e^b} \tag{15}$$，

$$g(a,b,\hat{u}_s)=(1-2\hat{u}_s)a+b \tag{16}$$，

其中，$\hat{u}_s$是该位置已译出的比特，a和b表示输入该基本译码单元的LLR值。为了降低复杂度，公式（15）可以简化为：

$$f(a,b)\approx\operatorname{sign}(a)\operatorname{sign}(b)\min(|a|,|b|) \tag{17}$$。

各节点LLR可计算为：

$$L_{j,i}=\begin{cases}f\left(L_{j-1,i},L_{j-1,i+2^{j-1}}\right), & \left\lfloor\frac{i-1}{2^{j-1}}\right\rfloor\bmod2=0\\ g\left(L_{j-1,i-2^{j-1}},L_{j-1,i},\hat{u}_{j,i-2^{j-1}}\right), & \text{otherwise}\end{cases} \tag{18}$$。

各节点比特值可计算为：

$$\hat{u}_{j,i}=\begin{cases}\hat{u}_{j+1,i}\oplus\hat{u}_{j+1,i+2^{j-1}} & \left\lfloor\frac{i-1}{2^{j-1}}\right\rfloor\bmod2=0\\ \hat{u}_{j+1,i} & \text{otherwise}\end{cases} \tag{19}$$，

其中，$j=1,2,\cdots,n$，$i=1,2,\cdots,N$。

当计算出因子图最左边阶段的节点LLR $L_{n,i}$时，可判决如下：

$$\hat{u}_{n,i}=\begin{cases}0, & L_{n,i}\geqslant0\text{或为冻结比特}\\ 1, & L_{n,i}<0\end{cases} \tag{20}$$。

经比特反序置换的逆变换后，即可得到译码结果$\hat{u}_1^N$。在执行SC译码算法时，初始化$L_{0,i}=L_i$。

在有限码长下，SC算法性能较差。为了进一步提升SC译码算法的纠错能力，文献[14–16]提出了SCL算法。该算法

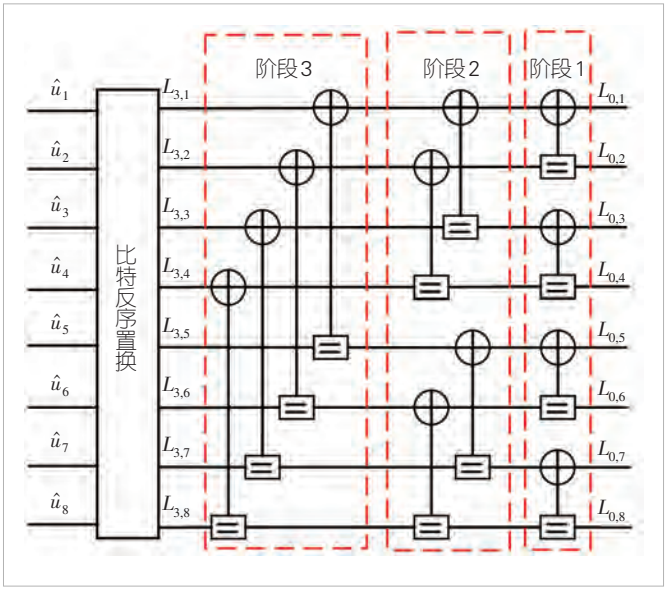


图3 $N=8$的极化码因子图

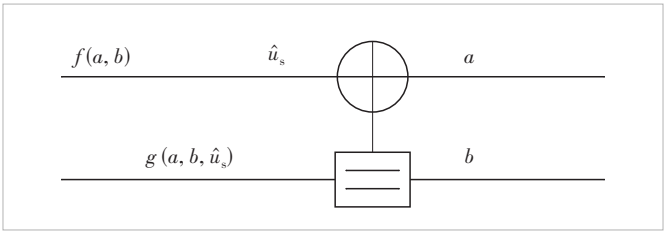


图4 极化码基本单元因子图

在译码的同时保持多组译码结果。保持译码结果数越多，复杂度越高。为了限制候选结果的数量，需要指定最大同时保持的译码路径数L，也称为搜索宽度，在译码结束后选择最可能的结果输出。第l组可能译码结果中第i个信息比特的LLR值可定义为：

$$L_{n,i}[l]=\ln\frac{W_N^{(i)}\left(y_1^N,\hat{u}_1^{i-1}[l]|0\right)}{W_N^{(i)}\left(y_1^N,\hat{u}_1^{i-1}[l]|1\right)} \tag{21}$$

路径度量值可由公式（22）递归计算：

$$PM_i[l]=\begin{cases}PM_{i-1}[l], & 1-2\hat{u}_i[l]=\operatorname{sign}\left(L_{n,i}[l]\right)\\ PM_{i-1}[l]-\left|L_{n,i}[l]\right|, & \text{otherwise}\end{cases} \tag{22}$$

度量值初始化为$PM_0[l]=0$。

SCL算法在译码结束时，会根据路径度量值，将最可靠的结果输出。对于循环冗余校验（CRC）辅助的极化码，在译码结束后，选择通过CRC校验的最可靠结果输出，若L组结果均未通过校验，则选择最可靠的结果输出。

4.2 传统OSD算法

阶序统计译码（OSD）[17]可以对任意短线性分组码进行译码，且可以接近最大似然（ML）译码性能。对于信息长度为k且最小汉明距离为$d_{\min}$的线性分组码，阶数为$\lceil d_{\min}/4-1\rceil$的OSD可以接近ML译码性能。令$\alpha_i=|L_i|$表示可靠度，则可靠度序列为$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_N)$。在传统OSD中，首先对可靠度序列$\boldsymbol{\alpha}$进行降序排列，生成矩阵$\boldsymbol{G}$被相应的置换为$\bar{\boldsymbol{G}}=\boldsymbol{G}\prod_1$，其中，$\prod_1$表示相应的置换。接下来，对矩阵$\bar{\boldsymbol{G}}$执行高斯消元（GE）得到系统形式的矩阵$\tilde{\boldsymbol{G}}$。为了确保前$k$列是线性无关的，在GE中可能会发生置换$\prod_2$。最终，硬判决序列和可靠度序列置换为$\tilde{\boldsymbol{c}}$和$\tilde{\boldsymbol{\alpha}}$。然后，按照汉明重量递增的顺序生成$k$长的TEP序列$\boldsymbol{e}=\left(e_1,e_2,\cdots,e_k\right)$，其中最大汉明重量为OSD的阶数（order）。对于每一个TEP$\boldsymbol{e}$，相应的估计码字$\tilde{\boldsymbol{c}}_e$可由重编码得到。找到最可能的码字$\tilde{\boldsymbol{c}}_{\text{best}}$等价于最小化$\tilde{\boldsymbol{c}}_e$和$\tilde{\boldsymbol{c}}$之间的加权汉明距离（WHD）。最后，输出码字$\hat{\boldsymbol{c}}_{\text{best}}$作为译码结果，即$\hat{\boldsymbol{c}}_{\text{best}}=\tilde{\boldsymbol{c}}_{\text{best}}\prod_2^{-1}\prod_1^{-1}$。当前有许多针对OSD的优化算法，如Fast-OSD[18]、PB-OSD[19]和LC-OSD[20]等，这些算法可以在保持译码性能的同时降低计算复杂度。

5 低复杂度SCL-OSD译码算法

当前极化码性能最优的译码算法是SCL。当列表较小时，SCL算法性能较差，但是复杂度较低，当列表数较大时性能较好但复杂度会更高。然而，列表大小L和性能增益并非呈线性比例。随着列表数的增大，SCL性能增益逐渐降低，但是复杂度和列表大小是成正比的。图5展示了5G极化码在不同列表大小下的误块率（BLER）性能和复杂度，其中，$N=128$，$k=64$，CRC校验比特数$r=11$。复杂度的定义在第5.3节有详细描述。可以看出，随着列表L的增大，BLER性能逐渐提升，但是提升的幅度逐渐减小，复杂度随着列表大小线性增加。

因此，为了平衡SCL性能和复杂度之间的关系，本文提出了一种SCL和OSD级联的方案。在SCL-OSD方案中最重要的一个问题是SCL译码器如何输出软信息，即后验LLR，该软信息将作为OSD的先验信息输入到OSD中进行译码。

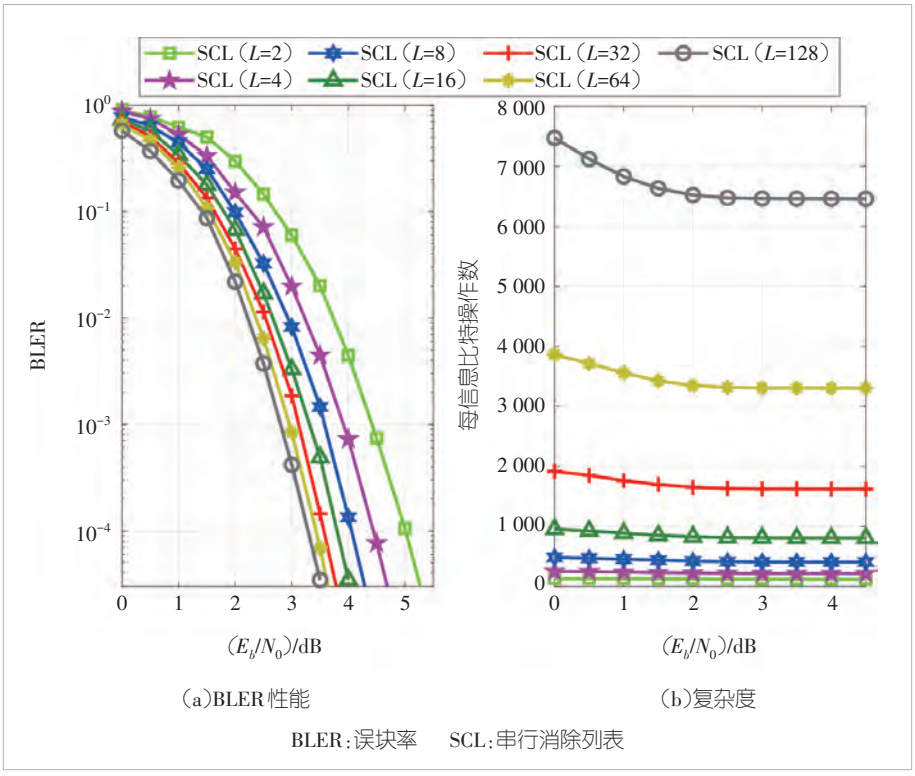


图5 5G极化码$N=128$、$k=64$、$r=11$在SCL不同列表下的BLER性能和复杂度

5.1 SCL后验LLR

在SCL和OSD级联方案中，OSD输入的先验信息应当从SCL处获得，这样可以充分利用SCL译码结果。传统SCL译码器是硬输出的，即只有比特信息，没有软信息。基于文献[21]，下面本文给出从SCL输出后验LLR的方法。

假设译码结束后，获得L组译码结果$u_1^N[l]$，$l=1,2,\cdots,L$。对于每一条候选路径l，相应的估计码字由$u_1^N[l]$重编码获得，可计算如下：

$$c_1^N[l]=u_1^N[l]\boldsymbol{G}_N \tag{23}$$

根据上述路径度量计算可知，路径度量值越大，表明该条路径的译码结果越可靠。归一化各条路径正确的概率为：

$$p_l=\frac{\exp\left(PM[l]\right)}{\sum_{i=1}^{L}\exp\left(PM[i]\right)} \tag{24}$$

第i个码字比特的概率可计算为：

$$p\left(\hat{c}_i\right)=\sum_{\substack{1\leqslant l\leqslant L\\ \hat{c}_i=c_i[l]}}p_l \tag{25}$$

则SCL译码器的外信息$\boldsymbol{L}^e=\left(L_1^e,L_2^e,\cdots,L_N^e\right)$可计算为：

$$L_i^e=\begin{cases}2/\sigma^2, & p\left(\hat{c}_i=1\right)=0\\ \ln\left(\dfrac{p\left(\hat{c}_i=0\right)}{p\left(\hat{c}_i=1\right)}\right) & p\left(\hat{c}_i=1\right)\neq 0\&p\left(\hat{c}_i=0\right)\neq 0\\ -2/\sigma^2 & p\left(\hat{c}_i=0\right)=0\end{cases} \tag{26}$$

因此，SCL输出的后验LLR$\boldsymbol{L}^p=\left(L_1^p,L_2^p,\cdots,L_N^p\right)$为：

$$L_i^p=L_i^e+L_i \tag{27}$$

在SCL-OSD中，只有SCL译码失败才需要启动OSD译码，因此只需要在SCL译码失败时计算后验LLR即可。

5.2 SCL-OSD译码过程

SCL-OSD译码流程如图6所示。首先执行SCL译码，当SCL译码成功，即存在路径结果通过CRC校验时，则直接输出该译码结果；若所有路径的结果都没有通过CRC校验，则执行传统OSD译码，然后输出相应译码结果。

值得注意的是，在SCL-OSD方案中，输入到OSD的可靠度序列由SCL输出的后验LLR构造，而WHD仍由信道接收LLR计算。

5.3 复杂度分析

本节将通过比较每比特信息的平均操作次数来分析译码复杂度。总操作次数定义为实数加法、二进制加法和比较的总数[22]。首先分析SCL算法的复杂度：

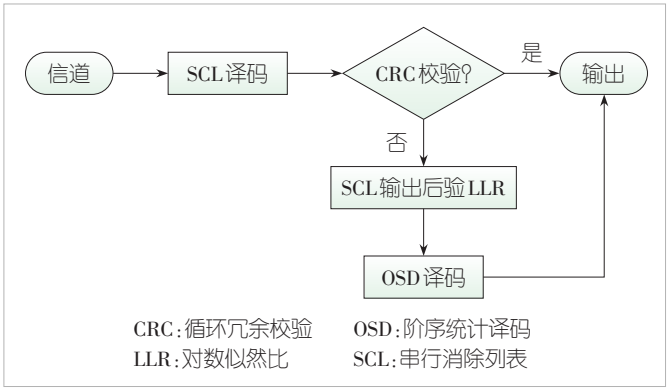


图6 SCL-OSD译码框图

1）SCL在译码过程中当路径分裂数大于$2L$时需要进行排序操作，采用快排序，共需要$2L\log_2(2L)\left(k-\log_2 L\right)$次比较，路径度量值的计算需要$N$次比较和最多$NL$次实数加。

2）计算每个节点LLR时，每个加号节点计算公式（17）需要5个比较，每个等号节点计算公式（16）需要1个比较和1个实数加法。总共有$N/2\log_2 N$个加号节点和$N/2\log_2 N$个等号节点，因此最多需要$3LN\log_2 N$次比较和$LN/2\log_2 N$次实数加法。

3）计算每个节点比特值时，加号节点执行二进制加法来返回比特信息，最多有$LN/2\log_2 N$次二进制加法。

4）最后对每条路径的译码结果进行CRC校验，需要对每条路径的译码结果计算$\hat{u}_{\mathcal{A}}[l]\boldsymbol{H}_{\text{crc}}$（$\boldsymbol{H}_{\text{crc}}$是相应的CRC校验矩阵）并判断是否通过校验，需要$(k+r)r$次二进制加法和$r$次比较。因此，最多需要$L(k+r)r$次二进制加和$Lr$次比较。

下面总结OSD中的操作数：

1）得到硬判决结果$\hat{\boldsymbol{c}}$需要N次比较。获得可靠度序列$\boldsymbol{\alpha}$需要N次比较。

2）用快排序方法对$\boldsymbol{\alpha}$排序，需要$N\log_2 N$次比较。GE需要$N(\min(k,N-k))^2$次二进制加法。

3）下面的步骤重复$\sum_{i=0}^{\text{order}}\binom{k}{i}$次。计算$\tilde{\boldsymbol{c}}_e$需要$k(N-k+1)$次二进制加法。计算WHD并判断是否需要更新需要$N-1$次实数加法和1次比较。

上述算法的总的操作数如表1所示。

5.4 仿真结果

本节通过仿真结果来评估所提出的SCL-OSD算法的译码性能和复杂度。为了便于对比，本节对不同列表大小的SCL算法和不同译码阶数的OSD算法进行了仿真。所有仿真均假设BPSK调制和AWGN信道。

图7展示了5G极化码$M=128$、$k=64$、$r=11$的BLER

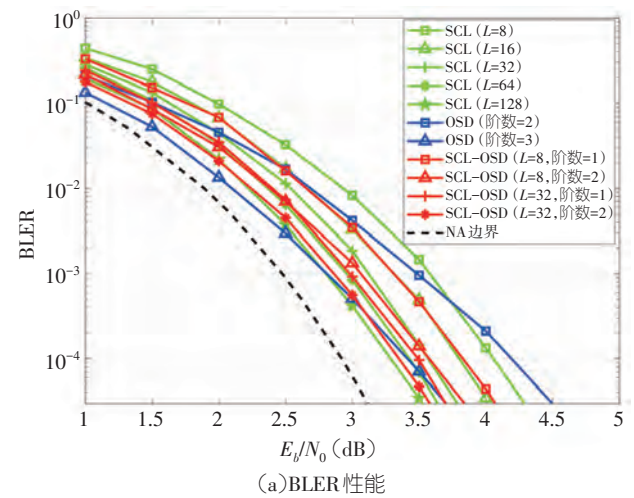
表1 计算复杂度

算法操作	SCL	OSD
实数加法	$LN\left(1/2\log_2 N+1\right)$	$\sum_{i=0}^{\text{order}}\binom{k}{i}(N-1)$
二进制加法	$LN/2\log_2 N+L(k+r)r$	$N(\min(k,N-k))^2+\sum_{i=0}^{\text{order}}\binom{k}{i}k(N-k+1)$
比较	$2L\log_2(2L)(k-\log_2 L)+N+3LN\log_2 N+Lr$	$2N+N\log_2 N+\sum_{i=0}^{\text{order}}\binom{k}{i}$
总操作数	$2L\log_2(2L)\left(k-\log_2 L\right)+N(L+1)+4LN\log_2 N+L(k+r+1)r$	$N((\min(k,N-k))^2+\log_2 N+2)+\sum_{i=0}^{\text{order}}\binom{k}{i}(N+k(N-k+1))$

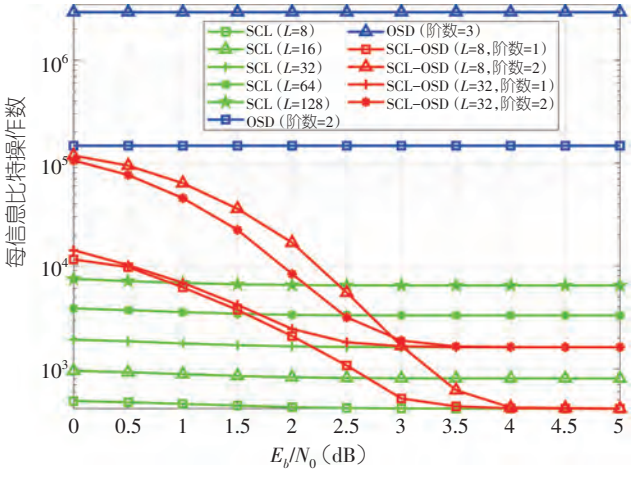
SCL：串行消除列表　OSD：阶序统计译码

性能和复杂度。由图7可知，$L=8$、阶数=1的SCL–OSD性能和$L=16$的SCL性能相当；$L=8$、阶数=2的SCL–OSD性能和$L=32$的SCL性能相当；$L=32$、阶数=1的SCL–OSD性能和$L=64$的SCL性能相当；$L=32$、阶数=2的SCL–OSD性能优于$L=64$的SCL性能，比$L=128$的SCL性能略差。在复杂度方面，由图7可知，随着信噪比的增大，SCL–OSD的复杂度会趋近于单独SCL译码的复杂度。

此外，对5G极化码$M=96$、$k=64$、$r=11$也进行了相同的仿真，母码码长$N=128$。由图8可知，$L=8$、阶数=1的SCL–OSD性能和$L=32$的SCL性能相当；$L=16$、阶数=1的SCL–OSD性能和$L=64$的SCL性能相当；$L=32$、阶数=1的SCL–OSD性能和$L=128$的SCL性能相当。由图8可知，随着信噪比的增大，SCL–OSD的复杂度会



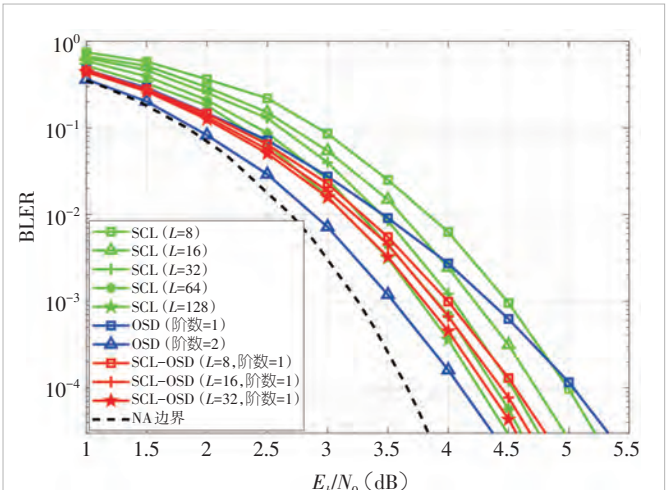
（a）BLER性能



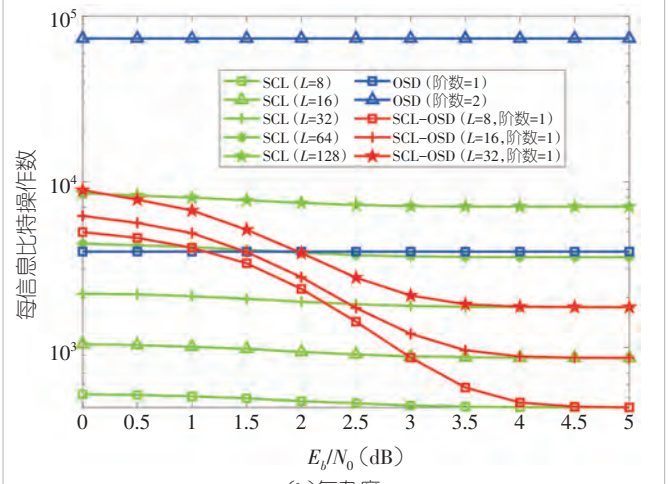
（b）复杂度

BLER：误块率　NA：正态近似　OSD：阶序统计译码　SCL：串行消除列表

图7　5G极化码$M=128$、$k=64$、$r=11$的BLER性能和复杂度



（a）BLER性能



（b）复杂度

BLER：误块率　NA：正态近似　OSD：阶序统计译码　SCL：串行消除列表

图8　5G极化码$M=96$、$k=64$、$r=11$的BLER性能和复杂度

趋近于单独SCL译码的复杂度。

综上，SCL-OSD相比SCL可以在取得相似译码性能的同时具有更低的复杂度。

6 结束语

本文以面向卫星通信的信道编码为目标，概述了极化码的基本原理、构造和编码、速率兼容方案等。在译码算法方面，主要描述了极化码当前性能最优的译码算法SCL和一种通用译码算法OSD。然后，分析了SCL算法存在的问题，即随着列表的增大，性能增益逐渐减小，而复杂度随列表大小成正比增加。针对该问题，本文为短极化码设计了一种SCL和OSD级联的方案。当SCL译码失败时，译码器输出后验LLR并启动OSD译码。仿真结果表明，所提算法相比SCL算法复杂度更低且译码性能相似。

参考文献

[1] 3GPP TR 38.821 V16.1.0 Solutions for NR to support non-terrestrial networks (NTN) [S]. 2021

[2] Azari M M, Solanki S, Chatzinotas S, et al. Evolution of non-terrestrial networks from 5G to 6G: a survey [J]. IEEE communications surveys & tutorials, 2022, 24(4): 2633-2672. DOI: 10.1109/COMST.2022.3199901

[3] Shannon C E. A mathematical theory of communication [J]. Bell system technical journal, 1948, 27(3): 379-423. DOI: 10.1002/j.1538-7305.1948.tb01338.x

[4] Arikan E. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels [J]. IEEE transactions on information theory, 2009, 55(7): 3051-3073. DOI: 10.1109/TIT.2009.2021379

[5] 3GPP. Multiplexing and channel coding: TS 38.212 V15.1.0 [S]. 2018

[6] Mori R, Tanaka T. Performance of polar codes with the construction using density evolution [J]. IEEE communications letters, 2009, 13 (7): 519-521. DOI: 10.1109/LCOMM.2009.090428

[7] Trifonov P. Efficient design and decoding of polar codes [J]. IEEE transactions on communications, 2012, 60(11): 3221-3227. DOI: 10.1109/TCOMM.2012.081512.110872

[8] Wu D L, Li Y, Sun Y. Construction and block error rate analysis of polar codes over AWGN channel based on Gaussian approximation [J]. IEEE communications letters, 2014, 18(7): 1099-1102. DOI: 10.1109/LCOMM.2014.2325811

[9] He G N, Belfiore J C, Land I, et al. Beta-expansion: a theoretical framework for fast and recursive construction of polar codes [C]// Proceedings of GLOBECOM 2017 - 2017 IEEE Global Communications Conference. IEEE, 2017: 1-6. DOI: 10.1109/GLOCOM.2017.8254146

[10] Zhou Y, Li R, Zhang H Z, et al. Polarization weight family methods for polar code construction [C]//Proceedings of IEEE 87th Vehicular Technology Conference (VTC Spring). IEEE, 2018: 1-5. DOI: 10.1109/VTCSpring.2018.8417498

[11] Arikan E. Systematic polar coding [J]. IEEE communications letters, 2011, 15(8): 860-862. DOI: 10.1109/LCOMM.2011.061611.110862

[12] Niu K, Chen K, Lin J R. Beyond turbo codes: Rate-compatible punctured polar codes [C]//Proceedings of IEEE International Conference on Communications (ICC). IEEE, 2013: 3423-3427. DOI: 10.1109/ICC.2013.6655078

[13] Wang R X, Liu R K. A novel puncturing scheme for polar codes [J]. IEEE communications letters, 2014, 18(12): 2081-2084. DOI: 10.1109/LCOMM.2014.2364845

[14] Chen K, Niu K, Lin J R. List successive cancellation decoding of polar codes [J]. Electronics letters, 2012, 48(9): 500-501. DOI: 10.1049/el.2011.3334

[15] Tal I, Vardy A. List decoding of polar codes [J]. IEEE transactions on information theory, 2015, 61(5): 2213-2226. DOI: 10.1109/TIT.2015.2410251

[16] Niu K, Chen K. CRC-aided decoding of polar codes [J]. IEEE communications letters, 2012, 16(10): 1668-1671. DOI: 10.1109/LCOMM.2012.090312.121501

[17] Fossorier M P C, Lin S. Soft-decision decoding of linear block codes based on ordered statistics [J]. IEEE transactions on information theory, 1995, 41(5): 1379-1396. DOI: 10.1109/18.412683

[18] Van W J, Alloum A, Boutros J J, et al. On short-length error-correcting codes for 5G-NR [J]. Ad hoc networks, 2018, 79: 53-62. DOI: 10.1016/j.adhoc.2018.06.005

[19] Yue C T, Shirvanimoghaddam M, Park G, et al. Probability-based ordered-statistics decoding for short block codes [J]. IEEE communications letters, 2021, 25(6): 1791-1795. DOI: 10.1109/LCOMM.2021.3058978

[20] Liang J F, Wang Y W, Cai S H, et al. A low-complexity ordered statistic decoding of short block codes [J]. IEEE communications letters, 2023, 27(2): 400-403. DOI: 10.1109/LCOMM.2022.3222819

[21] Wu X, Wang Y F, Li C F. Low-complexity CRC aided joint iterative detection and SCL decoding receiver of polar coded SCMA system [J]. IEEE access, 2020, 8: 220108-220120. DOI: 10.1109/ACCESS.2020.3043017

[22] Yue C T, Miloslavskaya V, Shirvanimoghaddam M, et al. Efficient decoders for short block length codes in 6G URLLC [J]. IEEE communications magazine, 2023, 61(4): 84-90. DOI: 10.1109/MCOM.001.2200275

作 者 简 介

李春杰，中山大学在读博士研究生；主要研究方向为信道编码、非正交多址接入和新型波形调制。

马啸，中山大学教授，博士生导师；主要研究方向为信息与编码理论、编码调制技术、无线通信和光通信等。

中兴通讯技术杂志社
促进产学研合作青年专家委员会

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

中兴通讯技术

(ZHONGXING TONGXUN JISHU)

工信领域集群成员期刊

电子信息领域集群成员期刊

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

双月刊　1995 年创刊
第 32 卷　总第 187 期
2026 年 2 月　第 1 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

《中兴通讯技术》编辑部
地址：合肥市金寨路 329 号凯旋大厦 1201 室
邮编：230061
网址：tech.zte.com.cn
投稿平台：tech.zte.com.cn/submission
电子信箱：magazine@zte.com.cn
电话：（0551）65533356

发行方式：自办发行
印刷：安徽添锦印刷科技有限公司
出版日期：2026 年 2 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