

AI 驱动的通信网络架构创新与技术演进路径研究

Research on Innovation of AI-Driven Communication Network Architecture and Technology Evolution Path

张 健,张嗣宏(南京中兴新软件有限责任公司,江苏 南京 210012)
Zhang Jian,Zhang Sihong(ZTE Communications Co.,Ltd.,Nanjing 210012,China)

摘 要:

人工智能正在驱动通信网络架构和技术演进升级,并呈现双向赋能的特征。在 Network for AI 方向,需构建支撑人工智能全场景的新型网络架构,通过增加训练网络带宽与降低传输成本提升 AI 算力效能,同时强化推理网络对海量智能体的泛在接入与端边云协同能力;在 AI for Network 方向,需推进算力与运力的硬件协同,实现 AI 能力在网络功能架构中的内生融合。通过算网融合关键技术突破、智能资源编排与分布式协同机制,促进网络系统向高阶自智演进,最终形成 AI 与通信网络双向驱动的良好发展范式。

Abstract:

AI accelerates the evolution of communication architectures and technologies, demonstrating two-way empowerment. To enable Network for AI, developing AI-native network architectures supporting full-scenario applications is critical. Optimizing training network bandwidth and transmission costs enhances AI computational efficiency, while ubiquitous access for intelligent agents and strengthened end-edge-cloud collaboration in inference networks ensure scalable intelligence deployment. To enable AI for Network, Advancing AI-Network integration requires hardware co-optimization of computing and transport capacities, embedding AI natively within network architectures. Breakthroughs in computing-networking convergence, intelligent resource orchestration, and distributed coordination mechanisms will enable self-evolving network intelligence, establishing a sustainable co-evolution cycle between AI and communication systems.

Keywords:

Artificial intelligent; Communication network; AI agent

关键词:

人工智能;通信网络;AI 智能体

doi:10.12045/j.issn.1007-3043.2025.08.005

文章编号:1007-3043(2025)08-0019-05

中图分类号:TN915

文献标识码:A

开放科学(资源服务)标识码(OSID):



引用格式:张健,张嗣宏. AI驱动的通信网络架构创新与技术演进路径研究[J]. 邮电设计技术,2025(8):19-23.

1 概述

近期,以 DeepSeek 为代表的大模型技术取得显著突破,其通过算法创新与架构优化,在保持国际顶尖性能水平的同时展现出突出的成本效益优势。人工智能技术的快速迭代,正在重构通信网络的技术范式与服务边界。AI 驱动的通信网络架构创新主要围绕 2 个方面展开:一方面,Network for AI 致力于构建适应

AI 业务特性的新型网络基础设施,为 AI 训练推理提供全场景网络支撑,使通信网络成为 AI 系统运行的神经中枢;另一方面,AI for Network 聚焦于将 AI 技术深度融入网络体系,推动网络运维、资源调度与安全防护等环节向高阶自智演进。这种双向驱动的技术融合,已成为下一代通信系统研究的核心命题。

当前国内外多个标准化组织、研究机构、运营商等在 AI+通信网络方面进行了积极探索和研究。3GPP 从 AI 重构无线接入网的可行性出发,启动了多项技术研究与标准项目,旨在将 AI 技术与 5G RAN 深度融合,

收稿日期:2025-07-02

重点提升网络频谱效率、构建智能化运维体系、打造极致节能和资源管理等关键能力。3GPP构建了支持多场景的AI统一基本功能架构,囊括了数据收集、模

型训练、管理、推理、传输、识别、监控等模型全生命周期的管理功能(见图1)。

2024年11月,软银(SoftBank)联合英伟达

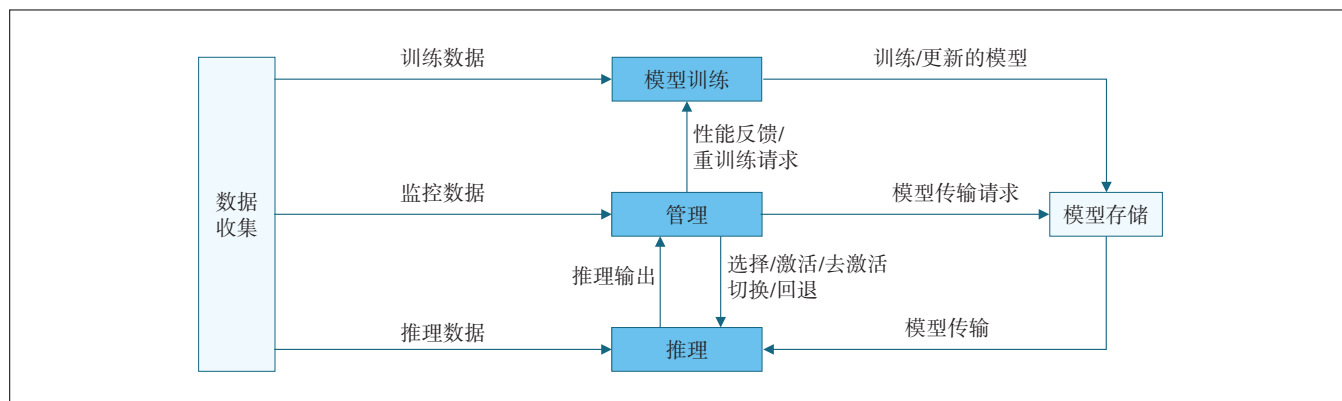


图1 3GPP定义的AI统一基本功能架构

(NVIDIA)、富士通(Fujitsu)、红帽(Red Hat)共同完成了首个AI-RAN“AITRAS”的外场试点。根据英伟达的观点,软银使用其AI Aerial加速计算平台,成功尝试将人工智能和5G电信网络相结合,同时运行人工智能和5G工作负载。

2025年3月,中国移动携手16家合作伙伴发布《智能体通信网络(ACN)白皮书》,并展示了业界首套ACN原型样机。本次发布的ACN白皮书,从目标愿景、显著特征、关键场景、网络能力等方面系统阐述了智能体通信的内涵与整体构想。

此外,SDN/NFV/AI标准与产业推进委员会启动了广域智算网络的关键技术研究,计划在2025—2026年形成广域智算网络关键技术规范和设备规范,并开展关键技术测试验证。中国信息通信研究院则聚焦百公里训练与千公里数据搬移场景,分层分域进行智算互联系统体系研究。

2 AI时代通信网络面临的挑战

通信网络是实现AI泛在普惠的基础连接平台,推动AI成为泛在化的社会级服务。AI是满足通信网络泛在连接的关键路径,赋能网络提升运行效率、降低运维成本、增强用户体验。AI与通信的深度融合已经历了多年的发展,并且取得了重要的进展,但目前仍然存在如下挑战。

a) 带宽与算力失衡。随着大模型参数规模从千亿级提升到万亿级,超大规模GPU集群成为大模型训练的必要条件,机内GPU通信和机外集合通信将产生

大量通信需求。例如,千亿级参数的大模型并行训练所产生的集合通信数据将达到数百GB量级,网络拥塞和丢包会严重影响GPU计算效率。这一矛盾的本质在于算力增速远超带宽增速,进一步加剧“算力孤岛”效应。

b) 数据传输昂贵。多样化泛在算力需求使得数据流通量剧增,数据传输效率直接影响算力供给效率。当用户使用低成本的小带宽专线时,大批量数据传输太慢;当使用高成本的大带宽专线时,对于很多不定期批量数据传输场景,专线利用率不高,当前按月固定带宽收费模式性价比较低。因此,存在“低带宽等不起,高带宽用不起”的困局。

c) 算网功能分离。近年来,越来越多的网元设备不断增强计算能力,服务网络自身规化、建设、维护、优化或者对外提供服务。但从物理形态上看,网元新增的算力资源一般是以相对独立节点的形态存在。从逻辑功能上看,网元内的网络能力和计算能力、跨网元的计算能力之间协同性不足,网络设施缺乏对计算服务和智能应用的感知性,阻碍网络赋能效应的发挥。

d) AI能力外挂。传统通信网络并非从设计初期就引入AI,而是随着技术的不断发展,从引入机器学习再到逐步融入AI和大模型技术。这种外挂模式导致各体系缺乏统一的框架,体系间接口和交互规则不规范,难以实现预验证、在线评估和优化的全自动闭环。此外,外挂模式下现网集中采集标注数据困难,传输及存储开销也大,导致AI模型迭代周期较长,收

敛慢、模型泛化性差。

3 Network for AI 需要从训练和推理 2 个方面综合考量

3.1 训练网络性能为王,机内、机间到算间网络服务于超大规模 GPU 集群

机内网络垂直扩展,即 Scale Up 网络,通过增加处理器、内存、存储等硬件资源,以及优化这些资源之间的互连技术,提升单个节点(如服务器、超节点或主机)内部的处理能力、存储能力和互连能力。Scale Up 网络以张量并行(Tensor parallelism, TP)以及专家并行(Expert parallelism, EP)流量为主,网络带宽和延迟是首要因素。目前的卡间互连技术多为专有协议,业界正在构建基于以太网的 Scale Up 开放标准体系,提升单节点 GPU 数量。

机间网络水平扩展,即 Scale Out 网络,通过把集群横向扩展到大量的 GPU 机柜,增加服务器或机柜数量并优化机间互连技术来扩充系统性能,从而突破单节点的性能极限。Scale Out 网络以流水线并行(Pipeline parallelism, PP)和数据并行(Data parallelism, DP)为主,可扩展性、高可用性、大带宽是首要因素。预计在 2025 年下半年,业界将推出下一代 102.4T 交换芯片,机间交换机的端口速率也将从 400G 提升至 800G 乃至 1.6T。

Scale Up、Scale Out 网络演进如图 2 所示。

算间网络互联多个智算中心并提供低成本数据传输服务:一方面,算间网络需要集成多项技术,包括带宽日历、弹性带宽以及服务等级协议(SLA)的可承诺机制等,充分利用网络资源,应对海量样本数据的低成本迁移需求。另一方面,算间网络需要将多个分

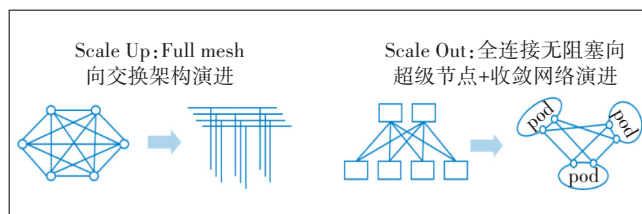


图2 Scale Up、Scale Out 网络演进

散的智算中心互联成一个大型虚拟智算集群,并解决传输时延、丢包对模型训练效率的影响问题,满足用户更大规模的算力需求。

3.2 推理网络泛在接入,使能海量智能体交互

传统通信网络的核心职责在于提供广泛的网络覆盖、高速的连接带宽以及低延迟的传输服务。未来基于大模型的 AI 智能体(AI Agent)正逐步成为智能应用的主流形式(见图 3)。面对这一变革,通信网络的角色与使命也迎来了深刻的拓展与重塑,需要其能够随时随地支持各类智能体之间的互联互通,实现跨厂商、跨生态的智能体间信息交互与任务协作。

为保障用户体验,智能体通常会被部署在网络的不同层级,如在终端部署家庭和个人助理智能体,在边缘位置部署工业制造、人机协作类智能体,在区域中心部署用户界面实时交互类智能体,在中心位置部署文本交互类、语音交互等智能体。网络需要在这些遍布各个层级的资源池和遍布各地的用户之间灵活调度。同时,针对高价值用户,需要构建差异化的入算保障能力。

智能体在行为特征和信息处理上与人存在着显著差异,可能导致网络传统流量模式发生变化。

a) 信息形态。面向人的通信系统,传递的主要是人能够感知和理解的文字、图像、音视频等多种媒体

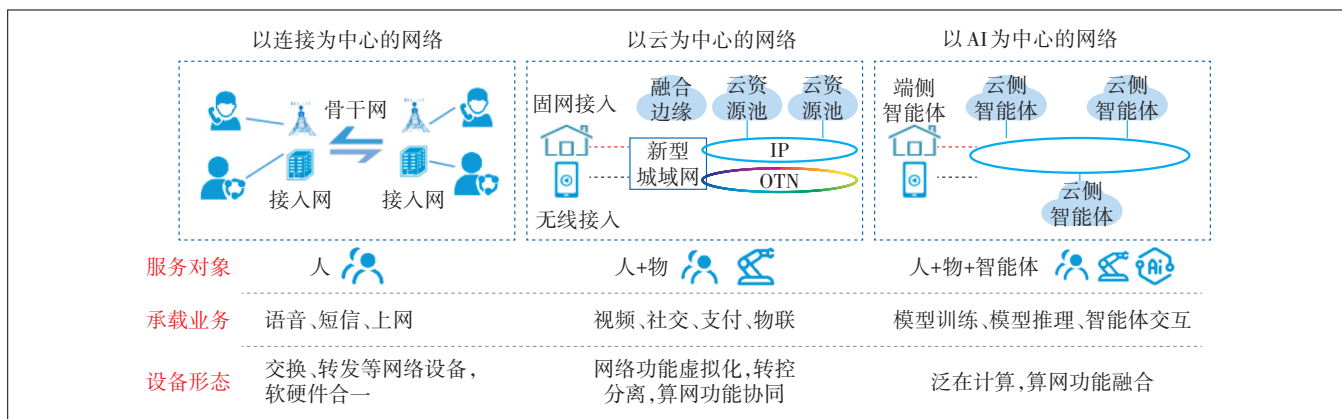


图3 通信网络从以连接为中心到以 AI 为中心的演进趋势

信息。而智能体间的通信,很可能直接传送多种形态的机器媒体信息,如 token、特征向量、隐空间数据、AI 模型等,而不必依赖于听觉和视觉传感器件。

b) 交互频度。人类的通信行为由人的听觉、视觉感受能力和大脑处理速度来决定单位时间内信息流量和带宽的消耗上限。对于具备高端感知和智能能力的高级智能体,其对信息的处理能力有可能超过人类,且可以实现多线程并行通信交互,单位时间内交互频度更高。

c) 数据流向。面向人的通信系统,数据流量多为人通过手机访问外部网络中的应用而产生的南北向流量。但在智能体通信中,不同智能体之间可能频繁直接在网内互访完成信息交互,导致大量流量发生在网络内部。如果多个智能体分布在不同区域,还将导致网络内东西向流量的增加。

未来通信网络还将利用网络内置算力为算力受限的智能终端(如 AI 穿戴设备、家庭终端等)提供端边云协同的混合 AI 推理机制。比如在家庭网络中, FTTR、中屏、Wi-Fi 路由器等终端内置推理算力,与门锁、传感器等弱算力终端协同实现家庭全面智能。在接入网络中, BBU、OLT 等网元设备内置推理算力,一方面可以与终端设备实现智能体协同,另外也可以独

立赋能自动驾驶、工业制造等行业,为运营商网络内生资源和能力变现提供了有效的途径。

4 AI for Network, 硬件走向融合异构, 功能进一步下沉

网络内生智能技术正引领着网络发生深刻变革,促使其从传统意义上单纯的连接管道,逐步转变为能够提供多元化、个性化服务的综合性平台。面向未来日趋复杂的网络功能和场景,现有“外挂式”解决方案和云 AI 服务供应方案存在局限性,难以提供实时高性能 AI 应用和服务,无法满足未来网络实现高阶自智的需求。未来需要从如下几个方面考虑。

a) 硬件形态上,需要高性能、异构化的硬件平台支持。当前部分网元通过内置算力单板的形式已经实现算力和网络在设备层级的融合(见图4)。未来随着异构技术的不断完善,可能会出现专门为融合网络与算力功能而设计的新型芯片架构。这种架构将打破传统的芯片设计思路,将不同类型的芯片(如网络芯片、计算芯片)集成在一个封装内,实现网络处理单元和算力单元的深度融合,以提高整体性能和能效比。

b) 功能架构上,需要将 AI 算法/模型按需编排到

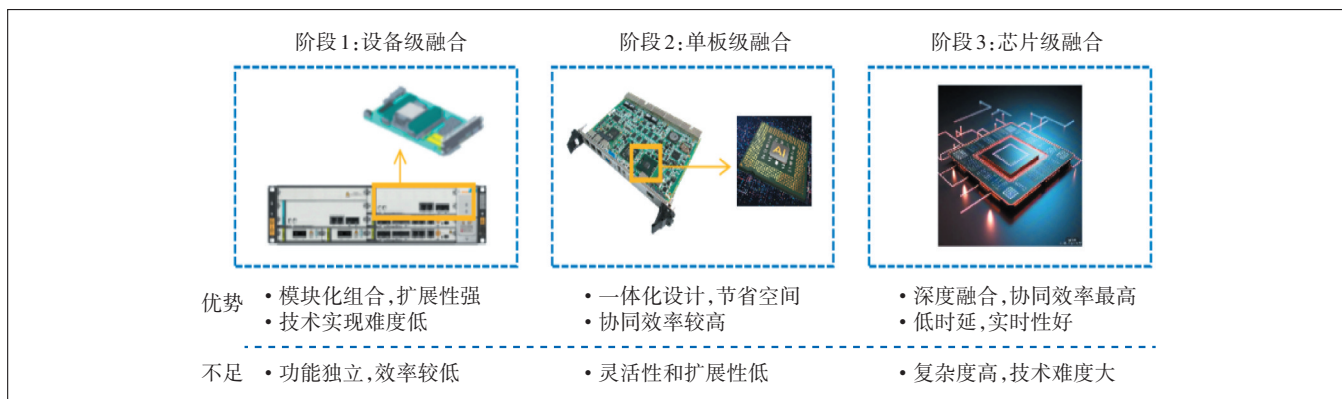


图4 通信网络硬件形态演进趋势

无线、有线或跨域端到端等各领域。通过对 AI 能力的灵活部署和调度,可以为高阶自智提供基础能力,使得网络智能化更高效。AI 对于网络的赋能将主要包括无线网络、有线网络与端到端体验保障等多个方面。

(a) 无线网络。在接入网中,通过 BBU 内置算力进行用户体验感知分析,可有效提升用户的视频体验,并通过智能资源管理实现频谱资源的优化利用和

节能效果提升。在核心网中,3GPP 定义了全新的核心网网元——NWDAF (Network Data Analytics Function),从各核心网网元、应用、运维系统以及运营支持系统收集数据,然后对这些数据进行分析,以提供有关网络性能和健康状况的建议及处理方案,可用于实时流量预测、资源管理和网络安全(见图5)。

(b) 有线网络。在固网接入中,基于网元设备内置算力和 AI 算法,对宽带接入用户的视频、云游戏、云

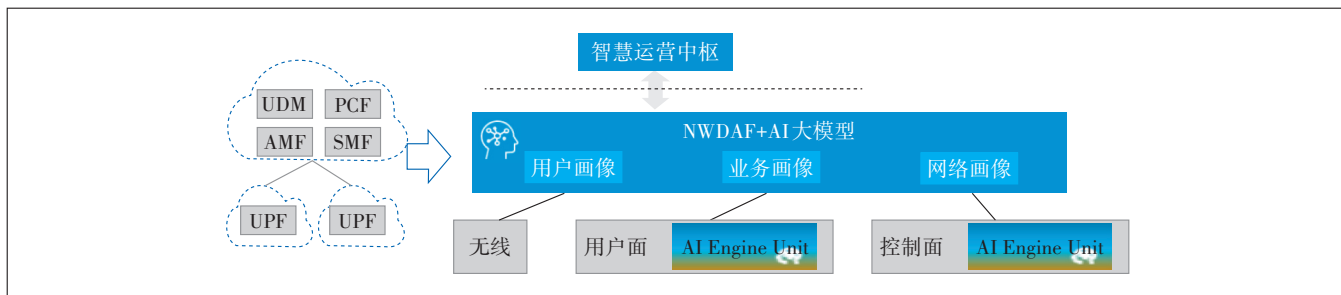


图5 核心网传统架构向智能化架构演进

电脑等业务进行卡顿检测和处理,提升用户体验。在承载网中,基于专家规则和大模型知识图谱,实现对网络各类故障场景(网元间/网元内的光纤中断/劣化故障、单板故障、光模块故障等)进行根因分析和界定定位。

(c) 端到端体验保障。AI还将帮助网络根据用户需求动态调整 QoS 参数,确保跨域跨层次的服务质量满足用户需求,这种端到端的智能优化可带来更流畅的服务体验。

当前网络中 AI 的应用场景以网络运营运维为主,在底层协议和算法方面较少。未来,6G 网络将实现 AI 原生架构,应用场景逐步向通信协议的底层下沉。比如在调制/解调、CSI/Pre-code、预测单模块性能提升以及多模块联合优化等领域,可大幅提升 6G 系统的物理层性能。

c) 在适用场景方面, AI 并非万能普适,需结合具体场景特点灵活选择,发挥最大组合价值。在确定性数学原理和推导过程主导的场景中,例如无线物理层的空口处理,由于其特定的运算需求,采用传统的硬件加速器和优化算法往往能实现更高的效率。

对于存在统计概率或者没有明确数学模型的情况下,比如网络配置、管理、维护及故障定位,无线物理层的信道估计和联合处理等,传统判决式 AI 和大模型可组合协同发挥最大价值。

需要特别说明的是, AI 赋能通信网络不等于大模型+网络,对于实时性要求高的场景,例如状态监控和简单故障诊断,传统判决式模型比大模型更合适。未来通信网络的内生智能硬件层面仍然依靠 CPU+各种传统加速器,以及局部 AI 加速器协同完成。

5 结束语

综上, AI 所引发的变革为通信网络的演进升级带来了新的机遇与挑战。通信基础设施的广泛连接与

持续迭代,如同数字时代的生命线,是 AI 应用与普及的基石,支撑智能世界的运转。而 AI 的快速进步,如同强劲的引擎,又为通信产业注入了前所未有的变革力量,驱动产业迈向新高度。二者双向奔赴,交叉赋能,正在重塑信息通信产业的未来格局,开启智能时代的新篇章。

参考文献:

- [1] 陈斌,裴培,许鹏. 大模型时代通信网络技术及新能力研究[J]. 邮电设计技术, 2024(9):1-6.
- [2] 刘为波,颜彪,沈麟,等. 基于 AI 通信的大规模 MIMO 信道状态信息反馈网络[J]. 电讯技术, 2024, 64(1):29-35.
- [3] 张中虎,杨蕊侨,谢云菁,等. 无线算力网络中面向 AI 任务的分布式通信计算协同研究[J]. 移动通信, 2025, 49(3):37-45.
- [4] 牛萍,黄文楷. AI 在蜂窝网络通信中的融合与创新策略[J]. 通信电源技术, 2025, 42(4):156-158.
- [5] 何世文,戴诗棋,董浩磊,等. 面向 6G 核心网的 AI-Native NWDAR 网元开发架构[J]. 移动通信, 2025, 49(1):81-90.
- [6] 申晓峰. 通信与 AI 融合加速推进内生智能成为未来趋势[J]. 通信世界, 2024(6):48-49.
- [7] 杨光. “AI+”战略创造电信运营商转型新空间[J]. 通信世界, 2024(20):4.
- [8] 孙万飞,左君,徐晖,等. AI 与 6G 网络融合研究与挑战分析[J]. 移动通信, 2024, 48(8):41-45.
- [9] 李贵勇,杨叶茂. 6G 时代的 AI 服务架构与算网技术[J/OL]. [2025-05-22]. <https://d.wanfangdata.com.cn/periodical/txyxxjs20250221001>.
- [10] 段晓东,李婕妤,程伟强,等. 面向智算中心的新型以太网需求与关键技术[J]. 电信科学, 2024, 40(6):146-159.

作者简介:

张健,硕士,主要研究方向为人工智能、算力网络、云网融合、大数据等;
张嗣宏,学士,主要研究方向为算力网络、云网融合、人工智能、大数据、SDN/NFV、5G 网络及业务发展等。

