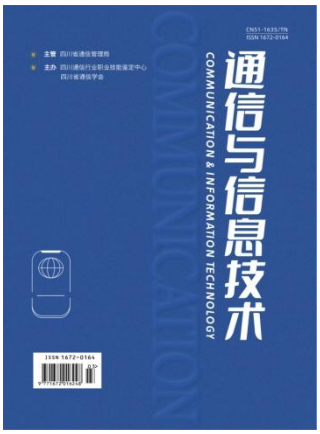

通信与信息技术

Communication & Information Technology

国内统一连续出版物号：CN 51-1635/TN
国际标准出版物号：ISSN 1672-0164
邮发代号：62-166

题目：6G 时代的 AI 服务架构与算网技术

作者：李贵勇，杨叶茂

优先出版日期：2025年2月21日

万方网站：https://w.wanfangdata.com.cn/

《通信与信息技术》为双月刊，逢单月底出刊，是国内外公开出版的自然科学学术期刊，设置了运营一线、热点技术、行业观察、解决方案、专网通信等栏目。

办刊宗旨：面向行业，沟通社会；宣传政策，促进发展；为通信发展服务，为通信企业服务，为通信科技人员和职工服务，为广大通信消费者服务，集信息性、行业性、技术性为一体的综合类通信刊物。

6G 时代的 AI 服务架构与算网技术

李贵勇，杨叶茂

重庆邮电大学 通信与信息工程学院，重庆 400065

摘　要：5G 网络利用人工智能技术加速了移动通信以及各垂直行业的智能化发展。但 5G 网络的对外 AI 服务面临着网络带宽资源、计算资源、AI 模型与数据资源分属不同域，存在跨域调度困难的弊端。除此之外，边缘计算的各节点在 5G 领域应用中也缺乏协同机制。针对以上问题，提出了在 6G 核心网中引入 AI 服务管理网元负责对 AI 任务、算力、带宽、AI 模型进行统一调度，形成一种新的对外 AI 服务架构，并在算力网络的加持下为 AI 算法提供更加稳定高效的服务。首先，总结分析在 5G 网络中的 AI 服务模式及架构，同时对当前边缘计算的特点、缺陷进行了归纳；其次，针对 6G 对外 AI 服务智能化、低时延与实时性的要求，提出一种新型网络架构；最后，总结 6G 内生 AI 与算力网络进行融合的发展前景与可能存在的困境、挑战。

关键词：智慧内生；6G；AI；算力网络

中图分类号：TN929.5　　**文献标志码：**A　　**文章编号：**

1 引言

随着5G的广泛应用，学术领域也逐渐开始对第六代移动通信技术进行深入研究。基于新一代互联网技术和大数据应用的“智”联网时代正在到来，而6G网络作为未来社会和经济发展的关键，将结合算力网络和AI的原生特性，并致力于实现智慧、深度、全息以及泛在这四大核心连接愿景[1]。

面对6G时代对于AI内嵌组网的迫切需求及智慧互联的宏伟蓝图，未来的网络架构将与人工智能达成深度交融，构筑起一个计算资源与人工智能紧密结合的新型网络体系。在5G向智能化迈进的征途中，两大核心策略尤为关键：一是于核心网层面增设诸如网络数据分析功能（NWDAF）等新型逻辑节点；二是利用多样化的模块级AI“附加增强型”功能，对系统性能进行精细优化与强化。这些技术革新均在现有系统架构与协议栈框架的支撑下得以实施，精准聚焦于已明确的特定通信挑战之上[2]。相较于依托5G技术的网络智能化道路，6G时代内生AI的网络将面临更为严苛的标准。6G所展望的智能互联图景，亟需从架构维度促进通信连接、运算力、数据资源与AI算法的紧密交融，力求再度缩减时延并削减通信开销。

面对6G的算力网络组网需求以及泛在连接的远景，我们预见未来的网络将能够在任何时间、任何地点与算力实现无缝的连接。为了满足这一场景对计算能力的要求，边缘计算技术应运而生，成为当前学术界和工业界关注的热点。基于5G技术的传统边缘计算存在两大主要问题：首先，单一的边缘计算节点资源是有限的，这使得它难以有效地处理复杂的任务；其次，边缘计算节点间缺少高效的协作机制，计算资源的分配、调度不完善[3]。相较于5G的边缘计算需求，6G的泛在连接愿景强调在架构设计中实现广泛的算力连通性，充分利用分布式部署的计算资源、数据资源、网络带宽资源，从而提升整体效能。

2 面向 5G 的 AI 服务与边缘计算组网技术

全球移动通信技术已正式迈入5G纪元，商用化的5G技术凭借其卓越的低时延与高可靠性，成功赋能物联网业务，将云端智能无缝融入用户终端，但是当前面向5G的AI服务与边缘计算组网技术也存在诸多挑战和缺陷。

2.1 基于 NWDAF 的对外 AI 服务架构

3GPP在5G标准制定之初，为了将人工智能与网络大数据分析技术融合应用在5G网络，于是在核心网中引入了一个网络数据分析功能模块(NWDAF)。NWDAF在5G网络结构中扮演着关键的功能角色，它的主要任务是对网络中的各种数据进行收集、分析和处理。通过对大量真实场景下的网络拓扑进行挖掘，收集存储不同业务类型间的海量数据流。

这批数据涵盖了用户行为、网络表现、流量模式等多个方面，为AI算法带来了大量的数据支持。NWDAF的设计考虑了与AI技术的融合，它提供了一套标准的接口和协议，使得外部的AI服务能够以图1的“附挂”形式与NWDAF进行交互[4]。得益于此，外部AI服务能够利用NWDAF提供的庞大数据资源和计算分析能力。但这仅仅只是在5G现有的架构

收稿日期：2025 年 1 月 20 日；　**修回日期：**2025 年 2 月 17 日

方案上做增量式AI功能开发，面临着网络带宽资源、计算资源、AI模型与数据资源分属不同域，存在跨域调度困难的弊端，经常会出现从秒级到分钟级的延迟问题，这进一步削弱了服务质量的稳定性，使得实时高性能AI服务变得难以实现，因此亟须在核心网框架内部引入新的对外AI服务架构。

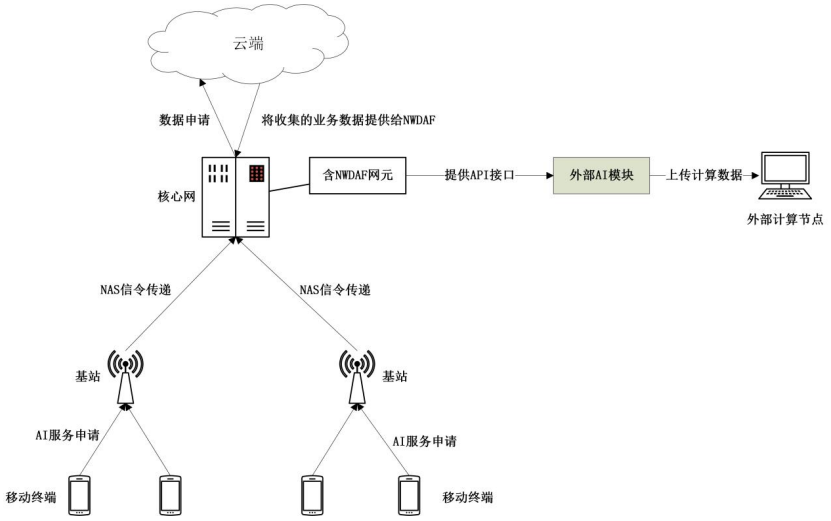

图 1 5G 附挂式 AI 架构

2.2 传统边缘计算的特点和缺陷

随着5G技术和移动互联网的持续进步，新的业务和应用领域如增强现实/虚拟现实（AR/VR）、车联网、环境重构同步定位与地图构建（SLAM）等不断涌现。传统电信运营商已经很难满足这种快速变化的业务发展需要，必须通过各种创新手段来满足用户多样化的服务需求。

这些新兴的应用程序不仅对网络提出了更高的带宽需求，同时也对计算能力提出了更高层次的要求，以保证其稳定运作并提升用户的使用体验。传统的网络中，节点之间的通信都需要依赖固定物理位置进行，这样就导致了网络拓扑结构难以灵活地调整。在这种背景下，边缘计算得以提出。边缘计算是一种将计算资源部署在网络边缘的网络模式，通过将计算资源与处理过程尽可能地放置在数据源附近，从而有效减少时延和通信成本，其网络架构如图2所示。

虽然边缘计算为缩减延迟带来了新颖的解决方案，但在实际部署时仍需跨越重重难关。首要问题在于，独立的边缘计算节点资源有限，尤其面对高计算需求的任务时，边缘节点往往力不从心，难以迅速且高效地完成任务，进而引发节点过载，延长了整个处理流程的时间。再者，尽管边缘计算资源正逐渐遍布更广泛的区域，但目前边缘计算节点间、边缘计算节点与云计算节点间缺失有效的协同调度、通信机制。正是由于这种算力资源的分配与调度体系不够完善，造成了较低的资源利用效率。故而，如何构思出高效的资源协同及任务调度体系，以提升计算资源的整体使用效能，是边缘计算领域亟需攻克的核心难题[5-6]。

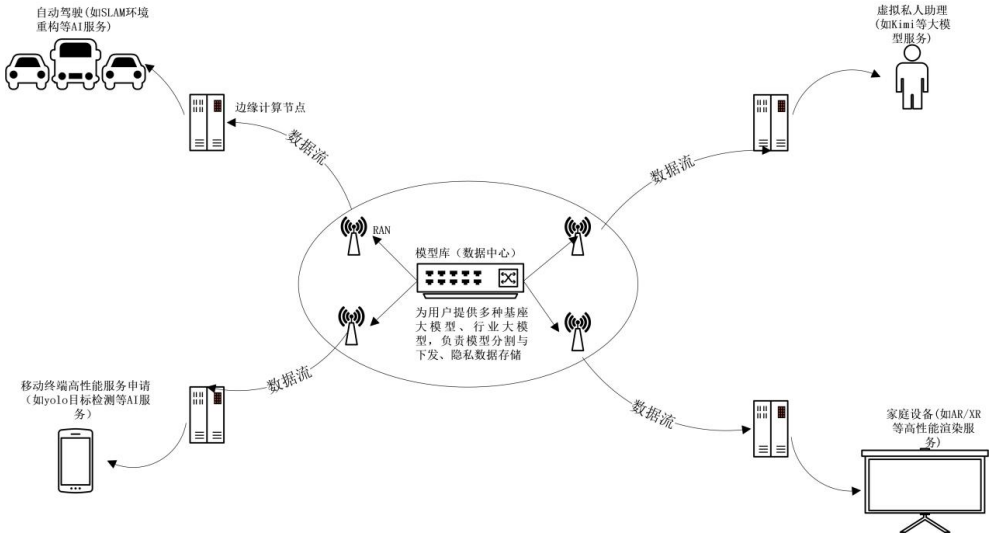

图 2 边缘计算网络架构

3 面向 6G 的 AI 内生与算力网络融合网络架构

移动通信网络和垂直行业智能化发展，6G移动通信系统在设计环节支持AI并以最佳效率实现，实现智能内生、网内各算力资源及网络资源统一调度是一种必然的趋势。

3.1 6G 内生 AI 国内外研究现状

面向6G网络智能化发展以及满足社会各界对新业务新应用的高性能需求，产业界和学术界对智能内生技术进行了广泛的研究和探讨。

华为无线技术实验室的吴建军[7]等提出，将网络中的资源调度从传统的连接资源扩展为连接、计算、数据和算法四要素资源，在任务部署阶段对这些资源进行有效配置，并在任务执行过程中进行实时调度。这种思路有助于将AI嵌入到网络资源管理的全过程，从而提升网络的智能化处理能力。与此类似，IMT-2030(6G)推进组无线AI任务组组长张朝阳[8]等提出的架构，将5G的传统网络架构（即控制面和用户面）扩展为网络层、数据层、智能层和应用层四个层次，力图通过人工智能来实现计算密集型任务的调度与管理，促进网络资源与AI的深度融合。这种设计不仅能够增强网络的自适应能力，还能提高服务的效率和灵活性。

中国移动研究院的刘光毅[9]等则从网络功能、结构和运行三个层面进行深度创新，提出了基于端到端微服务架构的网络设计，并设计了低频、中频和高频协同工作的信令广域覆盖策略，旨在通过智能化组网提升网络的整体效能。同时，还提出了基于云计算和大数据的边缘节点智能化组网技术，推动网络的灵活性和个性化服务发展。

瑞典皇家理工大学的Carbone Paris教授[10]以及法国巴黎索邦大学的Toumi教授[11]分别提出了在网络软件化过程中引入NeuroRAN架构和MLOps管道的概念。NeuroRAN架构通过结合虚拟化资源和去中心化资源管理，使网络能够在大规模网络环境中实现服务的可组合性，进而实现原生AI功能。而MLOps管道的引入，则是将智能集成到决策过程中，把AI/ML模型与网络管理深度交互，推动网络服务管理中的自动化决策过程。

这些研究反映了当前学术界和产业界对AI在6G网络中应用的多样性探索，旨在通过重构网络架构、引入AI智能化调度和管理，提升网络的灵活性、效率和智能水平。

3.2 6G 算力网络国内外研究现状

伴随着数字经济的快速崛起，计算资源已经变成了驱动经济增长的关键因素。作为计算基础的算力是支撑各类数字产业发展的关键要素之一，其在提升数据质量、降低时延等方面具有不可替代的作用。而算力网络通过整合云、边、端的计算资源，实现了资源的高效管理和调度，能够满足多样化的数字应用需求。因此，产业界和学术界对算力网络技术的落地进行了广泛的研究和探讨。

中国工程院院士张宏科[12]等人提出的“三层三域”网络架构为算力网络的发展提供了新的思路。在这一架构中，算力网络被划分为服务层、映射适配层和融合网络层三个层次，同时结合实体域、感控域和知识域三个领域，强调了算力与网络的深度协同。这种架构设计有助于提升算力资源的利用效率，并加快数字产业的创新应用。

Liu[13]等提出的CFN-dyncast技术则针对5G中的多接入边缘计算（MEC）环境，创新性地实现了分布式负载均衡。该技术突破了单个MEC站点的限制，通过动态调度，将客户端请求根据计算站点的负载和网络状况进行优化分配。这不仅提高了算力的整体效率，还有效改善了用户的体验。

在全球范围内，算力网络的研究正朝着协同优化和高效资源调度的方向发展，尤其是在边缘计算、分布式计算框架和智能化网络调度方面取得了显著进展。

3.3 6G 对外 AI 服务架构

6G内生AI指的是在6G网络架构中提供一个完整的AI工作流程运行环境，包括数据采集、数据处理、模型训练、模型推理等环节。在新的对外服务框架中，深度融合了AI服务所需的算力资源、数据资源、AI模型、通信管道以及网络带宽资源，为AI业务提供端到端的支持，将AI服务深度集成到核心网架构中[14-19]。因此，6G所拥有的AI能力不再是附挂式AI，而是一种内生特性，形成了一种新的对外AI服务架构，其基于服务接口的网络架构和流程如图3、图4所示。

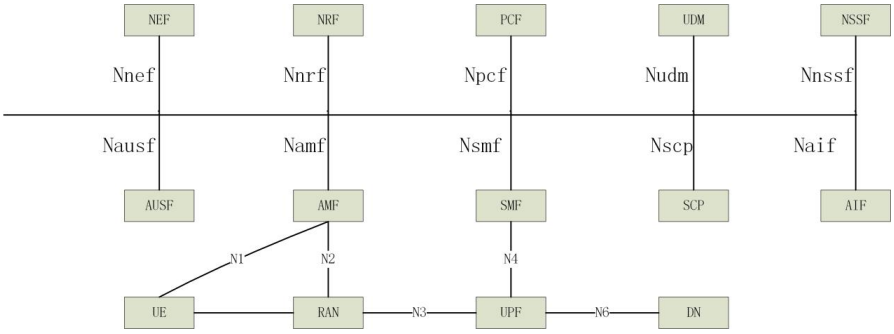

图 3 基于服务接口的网络架构

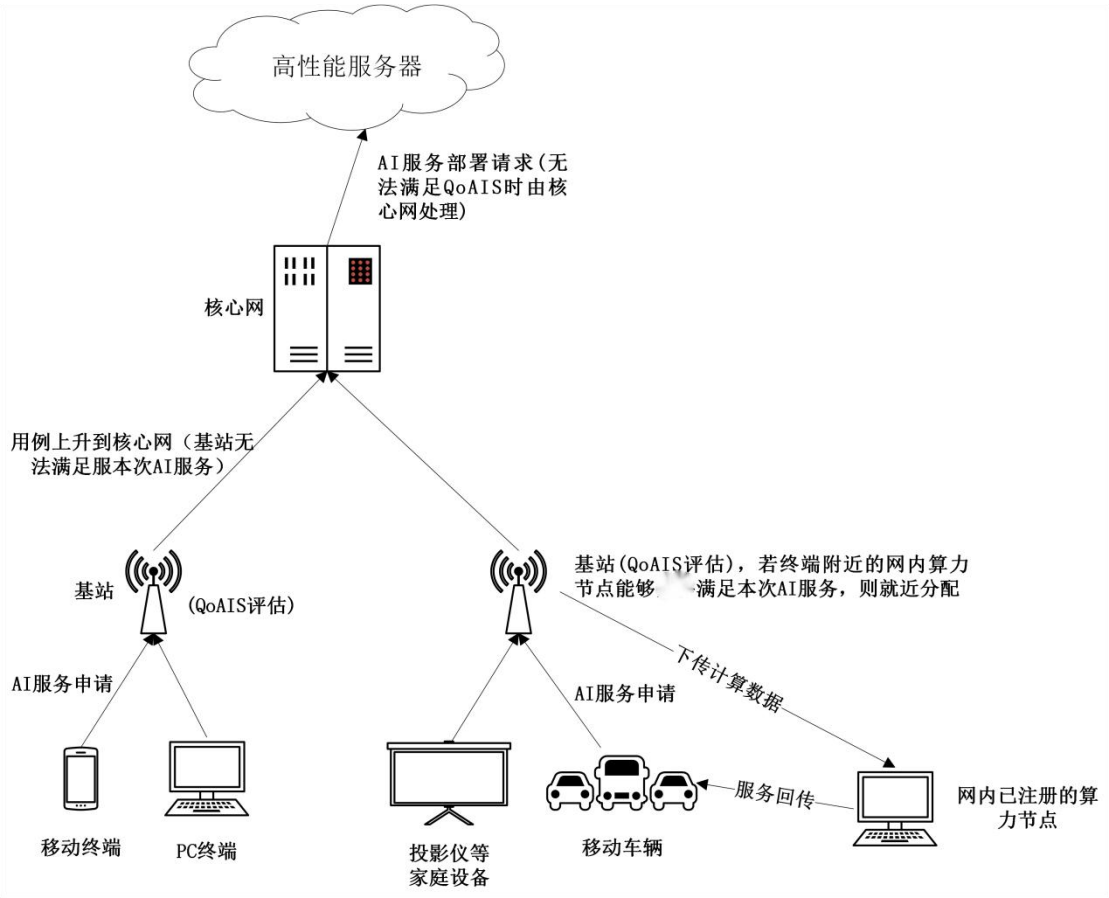

图 4 内生 AI 工作流程

通过在现有的5G核心网网元中，引入一个AI服务管理网元(Artificial Intelligence Function,AIF)，从而形成原生AI能力。新增的网元用于负责AI服务的管理，即创建和删除来自终端UE（User Equipment）的AI服务，并对AI服务资源进行监控。在6G网络架构内对资源进行有效调度分配，从而完成AI服务，其组网内部架构如图5所示。

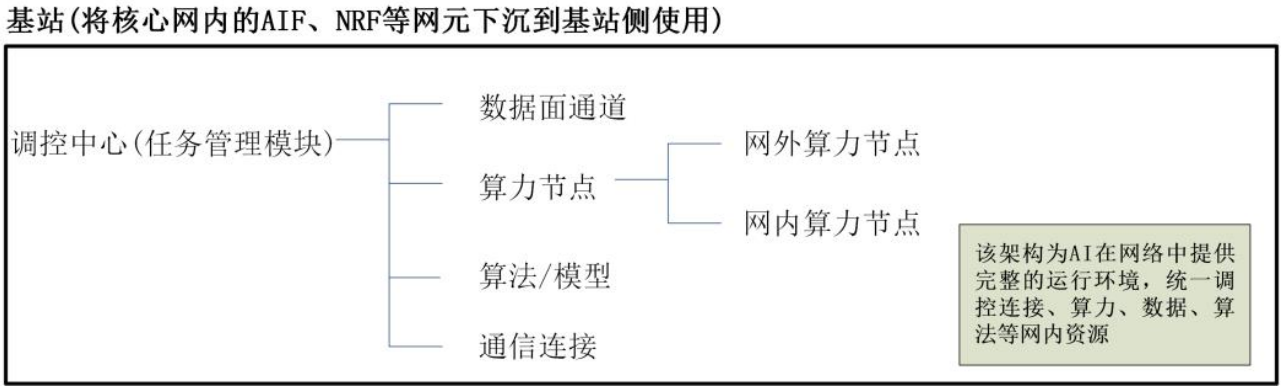

图 5 组网内部架构

AIF网元向任务管理网元提供AI服务模板。UE可根据自己的AI需求填写模板并向任务管理网元申请对应级别的AI服务。AI模板参数包括AI服务种类、时延需求、算力需求、带宽需求等。任务管理网元获取AI服务模板以后对其进行拆解和翻译，并在核心网内获取与AI服务相匹配的算力节点、数据通道等一系列网内资源，并将资源地址交付AI服务管理网元，由其管理本次AI服务。同时，AIF负责向编排系统Kubernetes进行通信、请求在指定边缘计算节点部署AI服务。除此之外，AI服务管理网元还需对本次AI服务的资源进行实时监控，资源包括CPU/GPU/网络带宽/时延等，若当前节点无法持续满足QoAIS(时延过高、算力不足等)，需上报资源阈值告警，经由编排系统评估后，通过算力网络开始进行算力节点迁移，以持续达成高效稳定的AI服务质量。其中编排系统Kubernetes对AI服务中的CPU/GPU/内存等关键资源

进行实时监控，过程如图6、图7所示。

图 6 AI 服务资源监控

图 7 核心网编排

通过引入新的对外AI服务架构，为AI在网络中提供完整的运行环境，统一调度连接、算力、数据、AI算法等，可以有效提升通信质量，降低通信成本和时延。

3.4 6G 算力网络架构

整个6G对外AI服务架构的设计，包括了AI服务请求模板、AI服务管理、AI资源监控等一系列功能。此外，要实现6G内生AI的功能，6G算力网络的支持也是必不可少的。各边缘计算节点将自身能力上报至核心网，由核心网根据算力请求方的需求进行统一调度，真正做到网中有算、算随网动，有效解决各边缘计算节点间、云计算节点与边缘计算节点间无法协同的问题[20]。

3.4.1 基于 Roofline 的算力决策模型

来自加州理工大学的伯利克提出了Roofline模型，该模型的目的是确定当前计算平台在不同计算强度（Operational Intensity）条件下能够达到的理论计算上限。该模型的具体结构可以参考图8。

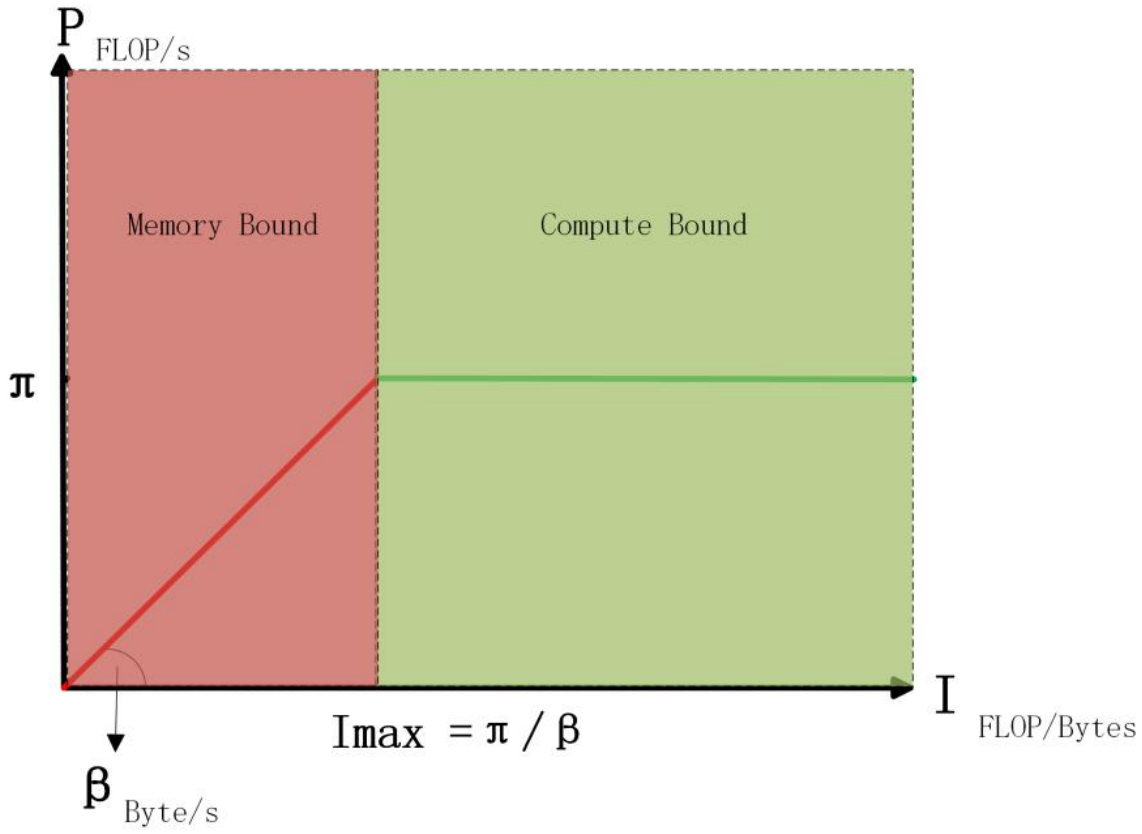

图 8 Roofline 模型

算力π表示计算平台的性能上限，即该平台在最大负载条件下，每秒可以完成的浮点运算次数。带宽β表示计算平台的最大带宽，它描述的是平台在面对最大负载时，每秒可以完成的内存交换能力。计算的强度上限I_{max}是基于算力π与带宽β之间的比率来确定的。

Roof-line划分出两个瓶颈区域：

$$P = \begin{cases} \beta * I, when\ I < I_{max} & Memory-Bound \quad (1) \\ \pi, when\ I \geq I_{max} & Compute-Bound \quad (2) \end{cases}$$

计算瓶颈区域Compute-Bound：无论模型计算强度I多高，其理论性能P始终受限于计算平台的算力π。当I超过平台的最大计算强度I_{max}时，模型进入Compute-Bound模式，此时P与I不再直接相关，而是受限于平台算力。若继续增加计算量，无法在计算效率与精度之间取得最佳平衡。若平台算力无限增长，计算结果将满足甚至超越应用需求。显然，平台算力π提升后，模型的理论性能P将在瓶颈区后继续提高。

带宽瓶颈区域Memory-Bound：当模型的计算强度I低于平台最大计算强度I_{max}时，模型处于“房檐”区域。在这一范围内，模型的理论性能P由平台带宽上限β（即房檐斜率）和计算强度I共同决定，处于Memory-Bound状态。在带宽受限的情况下，若平台带宽β增大（房檐更陡）或计算强度I提高，模型的理论性能P将线性增长。

在算力网络中存在网络算力π^{NET}和本地算力π^{UE}，网络算力一般远大于本地算力，当终端所需的算力π在本地无法满足时，可以选择使用网络算力。当本地算力可以满足终端的算力需求但对计算时延有需求时，由于本地算力的计算强度低，使用本地算力会导致功耗过大或者时延过大，此时网络算力也是更优选择。

时延是决定是否使用网络算力的关键指标。当UE使用本地算力时，此时的时延T_{UE}仅由计算时延T^{UE}_{comp}构成；当UE使用网络算力时，此时的时延T_{CFN}由通信时延T_{comm}和网络侧计算时延T^{NET}_{comp}构成。

UE根据时延决定算力源，当$T_{UE} \leq T_{comm}$时：优先用本地算力；当$T_{UE} > T_{CFN}$时：优先用网络算力；当$T_{comm} < T_{UE} \leq T_{CFN}$时：依据基于Roofline的算力决策模型选择合适的算力提供方。

3.4.2 算力节点的注册与调度

在6G对外AI服务架构中，存在一个算力管理网元负责算力节点和算力服务的管理。算力资源分为网内算力和网外算力，其算力类型、算力状态的上报注册流程也存在差异，网外算力需要通过核心网的NEF模块进行安全校验，网内、网外算力均注册到该网元中，根据算力需求方的需求进行统一调度，其算力网络架构如图9所示。

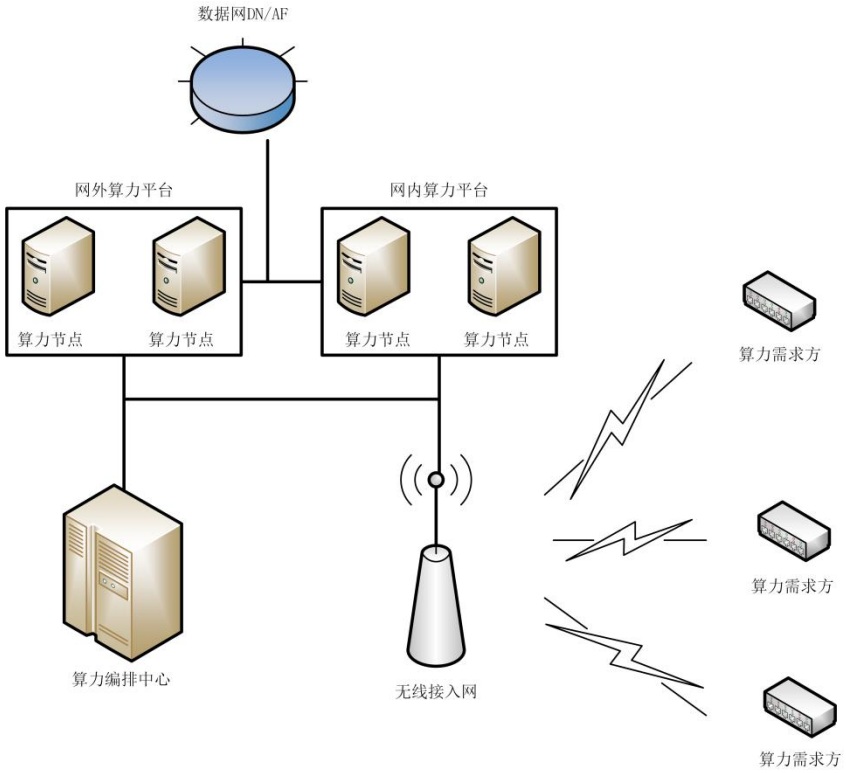

图 9 算力网络架构

当终端UE需要进行计算强度较大、时延要求较高的密集型计算任务时，可发送算力请求到网络侧，由核心网的算力调度模块依据终端的算力请求进行评估，调度合适的算力节点。当算力提供方因各种原因（算力节点不能到达或者时延过长、算力节点负载过重等）不能继续提供所要求的算力服务时，会重新寻找新的算力资源，完成算力迁移流程并继续提供算力服务。其中对于算力节点的选择采用如下方法：如果终端的算力请求中含有时延需求，则说明其属于时延敏感型终端，在满足终端时延需求的算力节点中选择更优节点；否则认为该终端对时延无特殊需求，选择节点时尽量避免算力节点迁移等情况。在上述度量之后，由编排系统Kubernetes对算力节点进行评估，取分值最高的算力节点调度给算力需求方，其中编排系统对算力节点的监控如图10所示。

图 10 算力节点监控

4 面向6G的AI内生与算力网络融合组网技术

当前的通信架构涵盖了核心网络（CN）、无线接入网络即基站端（RAN），以及用户终端UE。步入6G时代，AI服务能力下沉至RAN侧执行，与算力网络融合后，处理枢纽将更加贴近数据源头，能有效削减数据传输负荷，降低通信开销，并大幅缩减响应时间。当用户设备（UE）提出AI服务请求时，无线接入网（RAN）端会迅速作出响应，优先处理该请求，并着手评估此次AI服务的品质指标——即AI服务质量（QoAIS）。通过这一评估过程，系统能够明确该AI服务所需达到的QoAIS标准，进而将这些标准细化为对各类资源（涵盖通信链路资源、计算能力资源、数据储备资源及算法逻辑资源）的编排策略、调度安排及管控措施的具体需求，确保QoAIS得以持续满足。

当RAN侧算力足够满足QoAIS要求时，基站会响应AI服务请求，并可通过基站间协作支持高实时性和高速移动终端的AI应用，如自动驾驶、目标检测和室外定位等。然而，由于基站节点内算力和存储有限，当QoAIS无法在域内达成时，任务将上升到核心网，由核心网调度全局算力和数据资源。在此架构的基础上图11、图12由如下环境所得，实验部署在Ubuntu22.04系统、CPU为i7-10代、GPU为NVIDIA4060、核心网为Open5GS与自研6G Core、基站与终端是以UERANSIM模拟、AI服务则是以SLAM环境重构为例，从图11、图12中不难看出，这种分级部署架构不仅能减轻单一计算节点的性能压力，还能有效降低时延与信令开销。

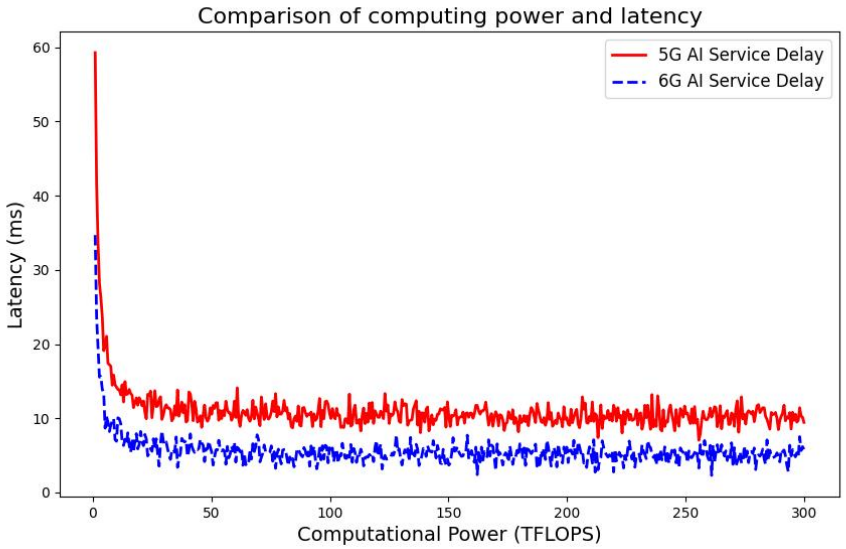

图 11 AI 服务时延对比

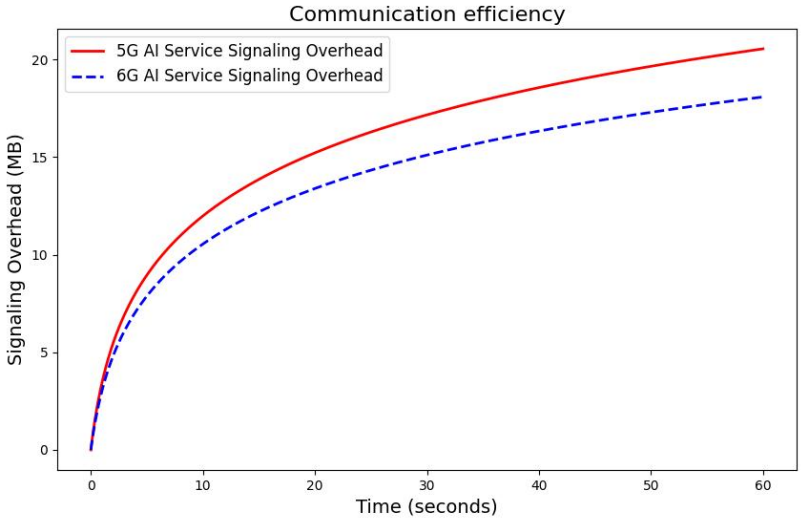

图 12 核心网信令开销

除此之外，可利用如图13所示的支持隐私保护及终端异构的联邦学习解决各基站之间的“数据孤岛”问题，促进多设备间AI模型的分布式协同[21]。对于基站与核心网之间，则可采用知识蒸馏的方法，如图14所示，让规模小、结构简单的RAN侧学生AI模型通过学习来获取核心网教师AI模型的知识进行数据更新[22]。基于上述策略可提升通信效率、计算效率，解决各基站之间、基站与核心网之间的“数据孤岛”困境，从而达成6G内生AI与算力网络真正意义上的融合，实现群智能。

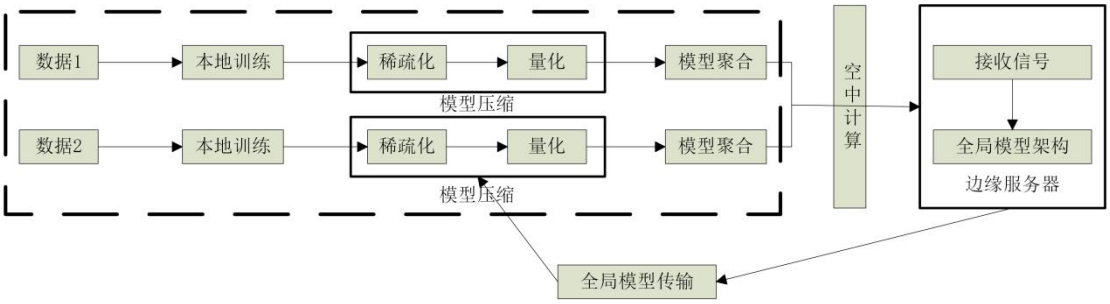

图 13 联邦学习模型

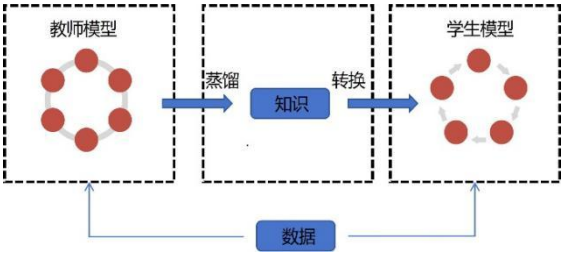

图 14 知识蒸馏框架

5 结束语

通过对5G对外AI服务架构、边缘计算的特点和缺陷进行阐述分析，探讨了6G对外AI服务架构的特点以及6G智能内生与算力网络融合的前景。这一融合不仅提升了网络智能与效率，还加速了数据处理的实时性与灵活性，为构建未来智能社会奠定了坚实基础。展望未来，两者深度融合将深刻影响智能交通、远程医疗等领域，对技术创新、标准制定等提出更高要求，这也将带来众多新问题。面对这些新挑战，如何扩展6G的应用领域并提升服务水平，进而推动社会向更高层次的智能化发展，还需要学术界同仁的共同努力和协作。

参考文献

[1] 李龙.在 MEC-WPT 系统中面向多移动设备剩余能量的资源联合优化[D].吉林长春:吉林大学,2021.

[2] 牛煜霞,赵嵩,贺智敏.NWDAF 网络数据分析功能的标准演进[J].移动通信,2023,47(01):29-33.

[3] 李子姝,谢人超,孙礼,等.移动边缘计算综述[J].电信科学,2018,34(01):87-101.

[4] 夏旭,梅承力.面向智能化切片的服务化等级保障技术增强和研究[J].移动通信,2021,45(01):6-10.

[5] 施巍松,孙辉,曹杰,等.边缘计算:万物互联时代新型计算模型[J].计算机研究与发展,2017,54(05):907-924.

[6] 施巍松,张星洲,王一帆,等.边缘计算:现状与展望[J].计算机研究与发展,2019,56(01):69-89.

[7] 吴建军,邓娟,彭程晖,等.任务为中心的 6G 网络 AI 架构[J].无线电通信技术,2022,48(04):599-613.

[8] 谢雨良,田雨晴,张朝阳.6G 智能内生无线通信网络:现状、挑战、系统设计和架构[J].移动通信,2024,48(08):8-12.

[9] 刘光毅,邓娟,李娜,等.内生智能和端到端服务化的 6G 无线网络架构设计[J].无线电通信技术,2022,48(04):562-573.

[10] Paris C ,Gyorgy D ,James G , et al.NeuroRAN Rethinking Virtualization for AI-native Radio Access Networks in 6G[J].INSIGHT,2023,25(4):74-79.

[11] Toumi N, Dimitrovski T."AI-native architecture for 6G networks and services with model dependencies," 2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Antwerp, Belgium, 2024:901-906.

[12] 张宏科,权伟,刘康.算力网络研究与探索[J].中兴通讯技术,2023,29(01):1-5.

[13] LIU B ,MAO J ,XU L , et al.CFN-dyncast:load balancing the edges via the networkd[C].Proceedings of 2021 IEEE Wireless Communications and Networking Conference Workshops.Nanjing:IEEE,2021:1-6.

[14] KROL M ,MASTORAKIS S ,ORAN D , et al.Compute first networking:distributed computing meets ICN[C].Proceedings of the 6th ACM Conference .Macao,China:ACM,2019:67-77.

[15] 张彤,任奕璟,闫实,等.人工智能驱动的 6G 网络:智慧内生[J].电信科学,2020,36(09):14-22.

[16] 赵亚军,郁光辉,徐汉青.6G 移动通信网络:愿景、挑战与关键技术[J].中国科学:信息科学,2019,49(08):963-987.

[17] 王晴天,刘洋,刘海涛,等.面向 6G 的网络智能化研究[J].电信科学,2022,38(09):151-160.

[18] 李文璟,喻鹏,张平.6G 智能内生网络架构及关键技术分析[J].中兴通讯技术,2023,29(05):2-8.

[19] 乔秀全,黄亚坤.面向 6G 的去中心化的人工智能理论与技术[J].移动通信,2020,44(06):121-125.

[20] 吕廷杰,刘峰.数字经济背景下的算力网络研究[J].北京交通大学学报(社会科学版),2021,20(01):11-18.

[21] 周传鑫,孙奕,汪德刚,等.联邦学习研究综述[J].网络与信息安全学报,2021,7(05):77-92.

[22] 黄震华,杨顺志,林威,等.知识蒸馏研究综述[J].计算机学报,2022,45(03):624-653.

作者简介

李贵勇（1971— ），男，硕士，正高级工程师，研究生导师，研究方向为 4G/5G/6G 移动通信协议、蓝牙/Wi-Fi 通信协议以及 MCU 芯片设计。

第二作者：杨叶茂（2000— ），男，硕士研究生，研究方向为 6G 内生 AI 网络架构及关键技术。

AI service architecture and computing power network in the 6G Era

LI Guiyong,YANG Yemao

School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: 5G networks utilize artificial intelligence (AI) technologies to accelerate the intelligent development of mobile communications and various vertical industries. However, the external AI services of 5G networks face the drawbacks of cross-domain scheduling difficulties, as network bandwidth resources, computing resources, AI models, and data resources belong to different domains. In addition, the nodes of edge computing lack a coordination mechanism in the application of the 5G field. To address the above issues, it is proposed to introduce an AI service management network element in the 6G core network to be responsible for the unified scheduling of AI tasks, computing power, bandwidth, and AI models, forming a new architecture for external AI services. With the support of the computing power network, it provides more stable and efficient services for AI algorithms. First, the AI service models and architectures in 5G networks are summarized and analyzed, and the characteristics and defects of current edge computing are also concluded. Subsequently, in response to the requirements of intelligence, low latency, and real-time performance of 6G's external AI services, a new network architecture is proposed. Finally, the development prospects of the integration of AI inherent in 6G and computing power networks, as well as the potential difficulties and challenges, are summarized.

Keywords: Native intelligence, 6G, AI, Edge computing