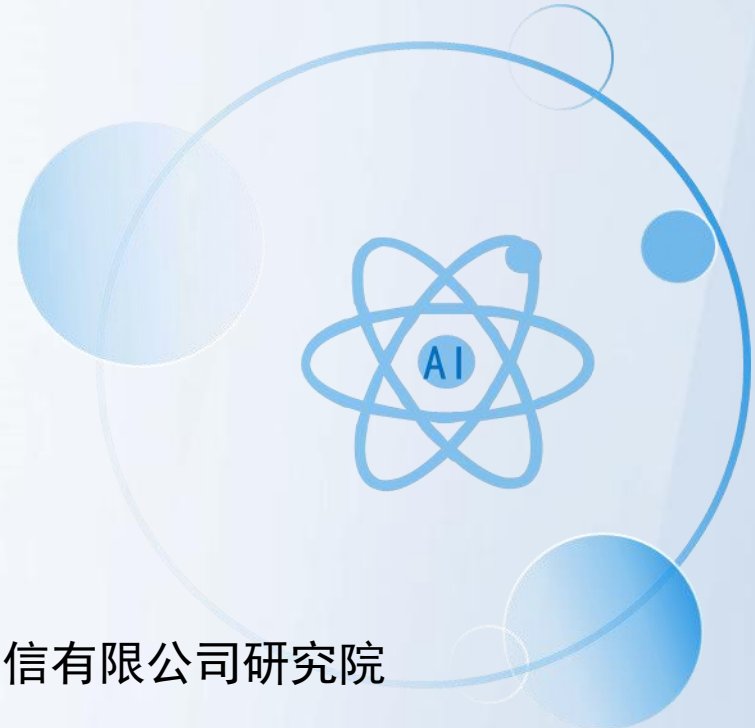
中国移动 China Mobile | 研究院 CMRI

6G 无线内生 AI 架构与技术白皮书

（2022）



中国移动通信有限公司研究院

目录

1 前言

人工智能（Artificial Intelligence）在最近十年发展迅猛，在挖掘大数据样本的非线性规律、与环境交互的在线精准决策等方面快速超越了以人工为主的专家经验（Human Intelligence）模式，在计算机视觉、自然语言处理、机器人控制等领域取得了巨大的成功。究其缘由，一方面得益于以深度学习、强化学习等为代表的人工智能算法能力的突破；另一方面，以 GPU 为代表的人工智能算力成本的快速下降和普及，也加速了这一趋势。

从 5G 开始，人工智能在移动通信网络中逐渐得到了广泛的应用，比如从网管级别的网络配置优化到网元级别的资源调度优化，甚至空口物理层的智能化，此外终端侧的智能化应用也越来越多。面向未来，6G 网络需要助力千行百业的数智化转型，需要满足和提供相比云端智能实时性更高、性能更优的智能化服务。对于运营商而言，需要大幅降低网络运营成本，网络运营维护需要从局部的智能化运维向高水平的网络自治演进。

目前的人工智能主要是以云端智能为主，在云端汇聚大量的数据，利用集中的算力对数据进行预处理，AI 模型训练和验证等。但是在网络中传输大量的原始数据，一方面会对网络的传输带宽带、性能指标（比如时延）来巨大压力，另一方面对数据隐私保护也会带来很大的挑战。此外终端侧的智能化应用由于算力，算法模型，数据等不足，目前还有较大的提升空间。

面对以上挑战，在网络中引入内生 AI 的能力，摒弃外挂 AI 打补丁的方式，在架构层面实现通信连接、计算、数据和 AI 算法模型的深度融合，充分利用网络中分布式的算力和数据，引入多节点间以及终端与网络间协同机制，实现分布与集中处理的融合。这种方式一方面保护了数据隐私，另一方面也提升了数据处理效率、决策推理的实时性和网络节点的利用效率。

本白皮书首先介绍了内生智慧的驱动力和需求场景，从现有网络智能化现状，到 6G 时代对网络高水平自治、智能普惠、高价值的新型业务和极致业务体验、网络安全可信等的需求出发引出内生 AI。然后阐述了内生 AI 的定义和内涵，提出了 AI 算力、数据、算法与网络连接功能的深度融合。接下来从 AI 服务质量、全生命周期编排、计算与通信融合、与数字孪生的融合几个方面介绍了 6G 内生 AI 的新理念;随后详细介绍了内生 AI 驱动的新架构，包括数据面、智能面和扩展的控制面和用户面，和新技术，包括模型编排、分布式模型训练、分布式模型推理、数字孪生的预验证和优化，最后对后续研究方向进行了展望。

2 驱动力

人工智能技术在 5G 网络中的应用促进了移动通信网络和垂直行业的智能化发展，但以“打补丁”和“外挂”的应用模式阻碍了 AI 应用效果的发挥。同时，人工智能在各行各业的应用探索，对未来网络新的基础能力提出了需求，面向智慧泛在的未来愿景，6G 网络需要具有内生 AI 能力。

2.1 5G 网络智能化面临的挑战

5G 时代，网络智能化需要将 AI 等智能化技术与 5G 通信网络的硬件、软件、系统、流程等融合，利用 AI 等技术助力通信网络实现规划、建设、维护、优化、运营流程智能化，达到提质、增效、降本的效果，促进网络自身的技术和体系变革，使能业务敏捷创新，推动构建智慧网络，包括云网设备智能、网络运营智能、网络服务智能。5G 网络智能化主要面向通信连接及服务过程中进行优化，虽然引入了服务云，但由于 5G 架构、协议功能和流程已经定型，只能在现有架构方案上做增量迭代，网络和云的融合偏松耦合。

5G 网络智能化大多使用外挂 AI 的模式，基于外挂设计的 AI 应用，一般是采用打补丁等方式进行，面临如下的挑战：

- 缺乏统一的标准框架，导致 AI 应用缺乏有效的验证和保障手段，AI 应用效果的验证是在事后进行，这样端到端的整体流程长并且很复杂，中间过程一般需要大量的人力介入，对现网的影响及改动也比较大，这导致了目前 AI 在应用到现网的过程中难以迅速推广。
- 外挂模式难以实现预验证、在线评估和优化的全自动闭环。AI 模型训练通常需要准备大量的训练数据，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，训练开销较大、收敛慢、模型泛化性差等问题。
- 外挂模式下，算力、数据、模型和通信连接属于不同技术体系，体系间并未定义规范的接口和交互规则，对于跨技术域的协同，只能通过管理面拉通进行，通常导致秒级甚至分钟级的时延，服务质量也难以得到有效保障。

2.2 6G 智慧泛在新场景

内生 AI 是指在架构层面通过内生设计模式来支持 AI，而不是叠加或外挂的设计模式。对于内生设计模式的驱动力，主要包括如下几个方面：

- **网络提供泛在 AI 服务：**面向智慧泛在的未来愿景，6G 网络需助力千行百业的数智化转型，实现“随时随地”智能化能力的按需供应。相比云服务供应商，6G 网络需提供实时性更高、性能更优的智能化能力服务，同时提供行业间的联邦智能，实现跨域的智慧融合和共享。另一方面，由于终端存在大量数据，终端的计算能力也越来越强，考虑到数据隐私需求，需要内生智能协同网络和终端的算力、通信连接和算法模型等资源，比如算力卸载、模型编排等，为 2C 客户提供极致业务体验和高价值新型业务。
- **AI 为网络优化提供服务：**6G 网络需实现高水平自治和安全可信。目前网络自治水平不高（自动驾驶网络等级约为 2.2 级），需要引入网络内生 AI 能力支持实现对运营商和用户意图的感知和实现，实现网络的自我设计、自我实施、自我优化、自我演进，最终实现网络的高水平自治。此外未来网络将承载更多样化的业务，服务更多的应用场景，承载更多类型的数据，因此网络将面临大量新的、复杂的攻击方式。基于内生 AI 的安全能力在 6G 网络的各环节嵌入，实现自主检测威胁、自主防御或协助防御等。

从以上驱动力分析可以看出，6G 网络除了满足基本的通信需求之外，还需要考虑计算、数据、模型/算法等多方面的融合，即 6G 需要通过架构层面的内生 AI 设计，来满足网络

AI 多样化的新业务场景和网络自治优化等需求，包括应用于网络自身优化和用户体验的 AI（如用 AI 重写的空口），也包括第三方所需的各类 AI 服务。

3 定义和内涵

6G 在设计阶段考虑和 AI 的深度融合，不同于 5G 通过 AI 功能叠加、外挂等方式，6G AI 内生将算力、数据和模型进行端到端编排和控制，在架构层面支持连接、计算、数据和 AI 算法/模型等元素的深度融合，支持将 AI 能力按需编排到无线、传输、承载、核心等，为高水平网络自治和多样化业务需求提供智能化所需的基础能力。即 6G 的内生 AI 能力，将可以使得网络智能化更高效、性能更优，同时，网络智能化的内涵也随之扩展，不仅能助力网络性能持续优化，还能提供智能化的服务能力，助力千行百业的数智化转型。网络智能化将在 6G 时代持续演进，推动构建真正的智慧内生的网络。

3.1 6G 内生 AI 的定义

6G 网络内生 AI 是在 6G 网络架构内部提供数据采集、数据预处理、模型训练、模型推理、模型评估等 AI 工作流全生命周期的完整运行环境，将 AI 服务所需的算力、数据、算法、连接与网络功能、协议和流程进行深度融合设计。6G 网络内生 AI 为网络高水平自治、行业用户智能普惠、用户极致业务体验、网络内生安全等提供所需的实时、高效的智能化服务和能力。

3.2 6G 内生 AI 的内涵

现有的移动通信网络主要是面向连接的数据传输，需要实现以 QoS（Quality of Service）（比如速率，时延等）为基础的传输链路保障，而内生智慧需要实现对算力，模型和数据端到端的控制和编排，这两者对网络设计和实现的需求差异巨大，需要在 6G 网络设计之初就考虑如何融合设计，一方面考虑引入基于 AI 的服务质量保障体系，AI 工作流的端到端编排，计算与通信连接的融合设计等新理念，另一方面考虑引入数据面，智能面，扩展的控制面和用户面等新架构设计。

内生 AI 需要构建 AI 的服务质量评估和保障体系，在此基础上实现基于服务质量的 AI 全生命周期编排，包括算力、模型、数据和连接。

内生 AI 需要实现计算与通信的深度融合。考虑到内生 AI 的能力需要分布到网络节点中，分布式的网络节点通常数据、算力、带宽和时延受限，需要考虑计算和通信资源的深度融合设计。此外，需要重构网络架构、协议和功能，适应空口传输，优化内生 AI 的性能。

4 新理念

如何在 6G 网络设计之初将 AI 与网络进行融合，从而构建全新的内生 AI 系统是一个多层面的复杂问题，是两个技术领域的碰撞和渗透。需要打破传统通信网络的设计思路，融入 AI 元素和理念。我们认为，AI 服务质量的评估与保障、AI 全生命周期编排管理以及 AI 计算与通信的深度融合，将成为构成内生 AI 系统的基本理念。

面向不同行业和场景对 6G 网络内生 AI 千差万别的需求，我们需回答的第一个问题即是：如何将用户的需求转化为网络可以理解的对网络 AI 服务能力的要求？对此，我们提出 AI 服务质量，即 QoAIS（Quality of AI Service）的概念，并认为网络应提供对 QoAIS 的评估和保障体系。紧接着，网络作为 AI 服务的提供者，如何评估和持续地满足 QoAIS，实施 QoAIS 保障，则需要从内生 AI 的管理、控制、业务流等多个层面展开研究。从管理角度，我们提出可以通过对 AI 全生命周期工作流的编排管理，半静态地使相关资源要素（算力、数据、算法、连接）的配给满足 QoAIS 要求；从控制和业务角度，则需要多维资源的融合，协同控制前述资源要素的调配，以实时、持续的满足 QoAIS，其中 AI 计算和通信的深度融合是主要理念。

4.1 AI 服务质量（QoAIS）

QoAIS（Quality of AI Service）是对 AI 服务质量进行评估和保障的一套指标体系和流程机制[1]。6G 网络将构建内生于网络的 AI 能力，形成一套可服务于多种智能应用场景的能力体系，即 AIaaS。考虑到不同的智能应用场景（如网络高水平自治、行业用户智能普惠、用户极致业务体验、网络内生安全等）对 AI 服务的质量将有着不同的需求，因此需要一套指标体系通过量化或分级的方式表达用户层面的需求以及网络编排控制 AI 各要素(包括算法、算力、数据、连接等）的综合效果。

6G 网络内生的 AI 服务可以分为 AI 数据类、AI 训练类、AI 推理类和 AI 验证类，每一类 AI 服务均需要一套 QoAIS。传统通信网络的 QoS 主要考虑通信业务的时延和吞吐率（MBR、GBR 等）等与连接相关的性能指标。6G 网络除了传统通信资源外，还将引入分布式异构算力资源、存储资源、数据资源、AI 算法等 AI 服务编排的多种资源元素，因而需要从连接、算力、算法、数据等多个维度来综合评估网络内生 AI 的服务质量。同时，随着“碳中和”和“碳达峰”政策的实施、全球智能应用行业对数据安全性和隐私性关注程度的普遍加强，以及用户对网络自治能力需求的提升，未来性能相关指标将不再是用户关注的唯一指标，安全、隐私、自治和资源开销方面的需求将逐渐深化，成为评估服务质量的新维度，而不同行业和场景在这些新维度上的具体需求也将千差万别，需要进行量化或分级评估。因此，QoAIS 指标体系从初始设计时，即需要考虑涵盖性能、开销、安全、隐私和自治等多个方面，需从内容上进行扩展。

表 4.1-1 提供了一种针对 AI 训练服务的设计方式。

表 4.1-1： AI 训练服务的 QoAIS 指标体系

AI 服务类型	评估维度	QoAIS 指标
AI 训练	性能	性能指标界、训练耗时、泛化性、可重用性、鲁棒性、可解释性、损失函数与优化目标的一致性、公平性
	开销*	存储开销、计算开销、传输开销、能耗
	安全*	存储安全、计算安全、传输安全
	隐私*	数据隐私等级、算法隐私等级
	自治	完全自治、部分人工可控、全部人工可控

注*：不同类型 AI 服务间的共同评估指标

其中，“性能指标界”是评估模型性能好坏指标的上界和下界，如模型错误率、查准率、召回率等性能指标的范围。“泛化性”指模型经过训练后，应用到新数据并做出准确预测的能力。“可重用性”是模型在应用场景变化时能够继续使用的能力。“鲁棒性”指在输入数据受到扰动、攻击或者不确定的情况下，模型仍然可以维持某些性能的特性。“可解释性”是指模型能支持对模型内部机制的理解以及对模型结果的理解的程度。“损失函数与优化目标的一致性”是指模型训练过程中，对损失函数的设计与 AI 用例的优化目标的一致程度，比如函数中考虑的变量个数是否完全覆盖智能优化场景的优化目标指标。“自治”指对 AI 数据/训练/验证/推理服务的工作流中自主运行部分和人工干预部分的要求，反映了用户对 AI 服务自动化程度的要求。自治分为三个等级：完全自治（全流程自动化的 AI 服务，全程无需人工干预）、部分人工可控（AI 服务的工作流在部分环节自动化，部分环节要求人工辅助）、全部人工可控（AI 服务工作流的各环节均要求人工参与）。

除了上表所示的评估维度,QoAIS 也可以包括智能应用的性能指标。以信道压缩为例[2]，可以选择归一化均方误差（Normalized mean square error, NMSE）或是余弦相似度作为信道恢复精度的 KPI，也可以选择链路级/系统级指标（如误比特率或吞吐量等）作为反映信道反馈精度对系统性能影响的 KPI。此外，QoAIS 还可以包括 AI 服务的可获得性、AI 服务的响应时间（从用户发起请求到 AI 服务的首条响应消息）等与 AI 服务类型无关的通用性评价指标。

QoAIS 是网络内生 AI 编排管理系统和控制功能的重要输入，网络内生 AI 管理编排系统需要对顶层的 QoAIS 进行分解，再映射到对数据、算法、算力、连接等各方面的 QoS 要求上。

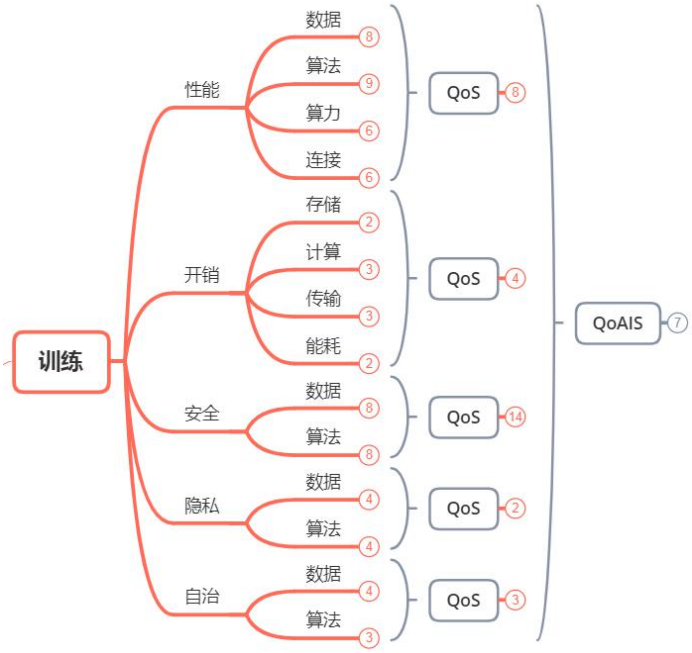


图 4.1-1：QoAIS 指标分解到各资源维度上的 QoS 指标

上图展示了 QoAIS 各指标维度和各资源维度上的 QoS 之间的映射关系。AI 服务的 QoAIS 整体指标拆解到各指标维度上的 QoAIS 指标，再进一步映射到各资源维度上的 QoS 指标，由管理面、各资源维度的控制面和用户面机制进行保障。图中各资源维度上 QoS 指标可分为适合量化评估的指标（如各类资源开销）和适合分级评估的指标（如安全等级、隐私等级和自治等级）。在前一类指标中，有部分指标的量化方案已成熟或较容易制定（如训练耗时、算法性能界、计算精度、各类资源开销等），部分指标目前尚无定量评估方法（如模型的鲁棒性、可重用性、泛化性和可解释性等），如表 4.1-2 所示。因此，如何在起始阶段设计出足够开放包容的网络架构以便后续逐步引入上述指标的成熟量化技术是需要思考和研究的问题。

表 4.1-2：AI 训练服务性能 QoAIS 到各资源维度的映射

指标维度	QoAIS 指标	资源维度	可量化指标	尚无量化方案指标
性能	性能指标界、训练耗时、泛化性、可重用性、鲁棒性、可解释性、优化目标匹配度、公平性	数据	特征冗余度、完整度、数据准确度、数据准备耗时	样本空间平衡性、完整性、样本分布动态性
		算法	性能指标界、训练耗时、是否收敛、优化目标匹配度	鲁棒性、可重用性、泛化性、可解释性、公平性
		算力	计算精度、时长、效率	
		连接	带宽及抖动、时延及抖动、误码率及抖动、可靠性等	

在质量评估和保障机制上，5G 网络的 QoS 机制仍存在一些问题，如业务区分颗粒度较粗，优化调整的周期较长，空口资源配置无法灵活适配网络与业务的实时动态变化等。因此在 6G 网络中提出评估 AI 服务的 QoAIS 指标的同时，也需要考虑如何设计端到端 QoAIS 机制和流程以更加高效准确。

延伸问题：

1. 网络在引入 AI 服务后，用户对 AI 服务安全性和隐私性上存在不同的需求选项，如何打破传统通信服务中 QoS 体系和安全体系分开独立设计的模式，使这种需求的差异性得到更好的满足?
2. 当前，部分 QoAIS 指标尚无成熟的量化评估方式（如模型的泛化性、可解释性、可重用性[3]），如何在起始阶段设计出足够开放包容的网络架构以便后续逐步引入上述指标的成熟量化技术?

4.2 AI 全生命周期编排管理

AI 生命周期是指网络中 AI 工作流的生命周期，即一条 AI 工作流的产生、执行、监测、评估、优化、完成及删除。网络内生 AI 工作流（Network Native AI Workflow）是指网络为完成一项 AI 服务需要分步骤完成的一项或多项工作任务。当前，AI 在各行业应用中具有类似的端到端工作流程[4]，可分为数据管理、模型学习、模型验证和模型部署四个环节，图 4.2-1 展示了一种网络环境中通用的 AI 端到端工作流程设计模式。

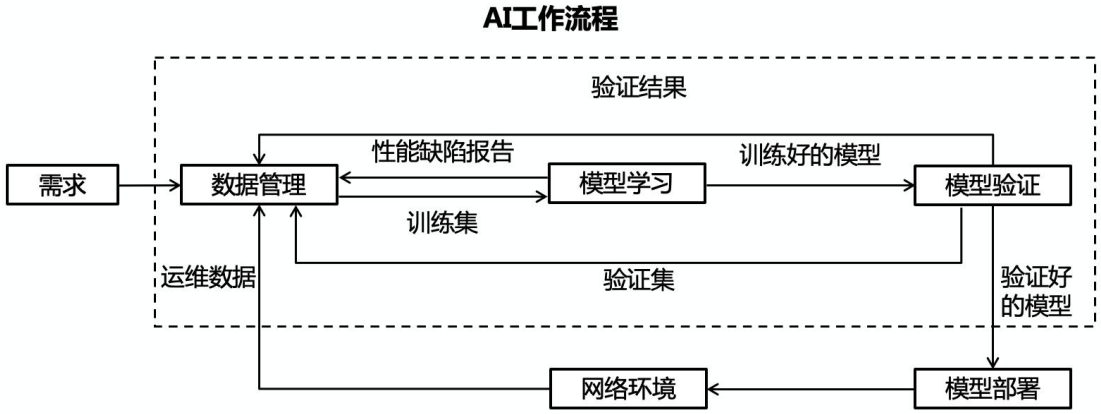


图 4.2-1：网络环境中通用的 AI 端到端工作流程

当前，在 5G 网络智能化实践中，AI 工作流程的大部分环节位于线下，与网络运行环境割裂，不同智能应用场景间采用“烟囱式”研发模式（即针对每个智能应用场景均独立进行全部环节的研发，不同场景间缺乏资源协同和共享），效率低下，成本高昂。6G 网络将为 AI 工作流的端到端各环节提供完整的运行环境，以一套全新的架构和技术体系，满足其全生命周期的运行需求，以一套统一的服务需求导入、分解、评估和保障系统，为网络自身和行业各种智能应用场景提供不同质量的 AI 服务。

6G 网络内生的 AI 工作流依据 AI 服务类型的不同，包含的工作任务也有所不同，任务数量有多有少，“流”有长有短，并非均为端到端工作流。例如，AI 数据服务的工作流仅包含数据管理环节相关任务；AI 验证类工作流则可以既包含数据管理环节，也包含模型验证环节的相关任务，或者仅包含模型验证环节的任务；AI 训练类工作流则可以仅包含模型学习环节，也可以同时包含数据管理和模型学习环节，取决于用户提供的数据是否已满足质

量要求。AI 推理类工作流可以仅包含模型部署相关任务，也可以同时包含数据管理和模型部署环节的相关任务。对于一个同时请求了多项 AI 服务的智能应用场景（例如，同时请求了 AI 训练、验证和推理服务），其对应的工作流可能是端到端的。图 4.2-2 展示了 6G 网络内生 AI 工作流与 AI 服务的关系。

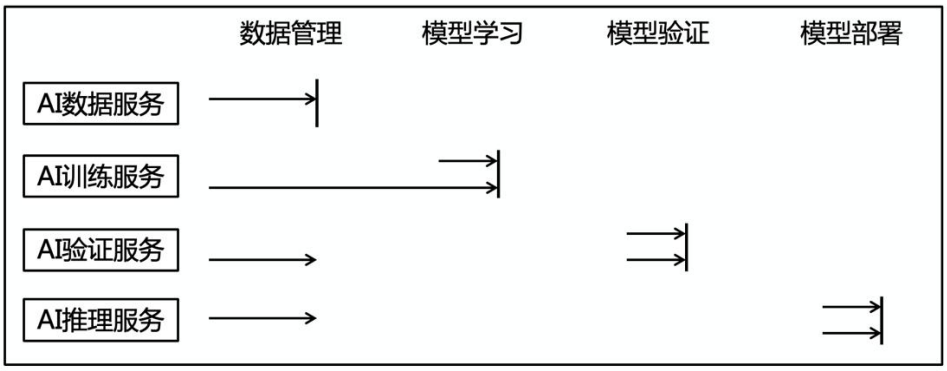


图 4.2-2：6G 网络内生 AI 服务及其工作流示意图

6G 网络为每一项 AI 服务生成其所需的工作流，对该工作流中每项任务所需的资源（算力、算法、数据、连接等）进行编排，对工作流的全生命周期进行管理，以保障该 AI 服务的 QoAIS 持续达成。在这一过程中，管理面采集工作流各环节的性能监测数据，评估 QoAIS 的达成情况，学习出工作流的任务设计和资源编排方案对 QoAIS 的影响，从而不断优化方案和策略，实现智能化的编排管理。

延伸问题：

1. 为保障 AI 服务的 QoAIS 持续达成，仅依靠管理面对工作流所需资源进行编排是否足够？是否还需要控制面的参与？管理与控制如何协同？界面如何区分？

4.3 AI 计算与通信深度融合

为保障 AI 服务的 QoAIS 持续达成，除了在管理面上实现 AI 工作流全生命周期的智能化编排管理，也需要在控制面和用户面上实现 AI 计算与通信的深度融合。

传统通信网络中的算力资源主要服务于通信业务，算力资源集成在设备处理板卡内，按照通信业务的处理流程进行算力资源的部署和分配。与通信业务不同，AI 业务是高算力需求业务，近年来各种处理器架构（GPU、NPU、DPU、TPU 等）不断涌现以提高计算效率，降低能耗。6G 网络内生 AI 服务对算力的核心需求为高计算效率、低能耗和低时延。虽然云端集中式算力资源堆放的计算效率较高,但往往无法满足边缘 AI 应用场景对实时性的需求。端和边侧虽然单节点算力资源有限，但规模庞大，实时性较好，在与云端算力进行协同编排调度后，有望满足各类 AI 服务对计算性能的需求。

5G MEC 方案中引入了边缘计算能力，用于提供低时延的计算服务，但其网络和计算部分是松耦合设计，在效率、部署成本、安全和隐私保护等方面存在进一步提升的空间。例如，在 5G MEC 方案中[5]，核心网用户面网元 UPF 可以与 MEC 合设，但在逻辑架构层面，及控制管理机制上，都还是两套相对独立的系统，当需要同时调整连接和算力时，是通过管理面进行协同，调整时延较大。另一方面，云、边和端侧部署的算力资源是分布式异构的，其协同调度需要实时适配网络动态复杂的通信环境，需要深入到控制面和用户面进行实时支持，这与云端单纯的计算环境完全不同。

以无线网络为例，AI 计算和通信的深度融合在控制面上存在三种可能的模式，图 4.2-3 是示意图。

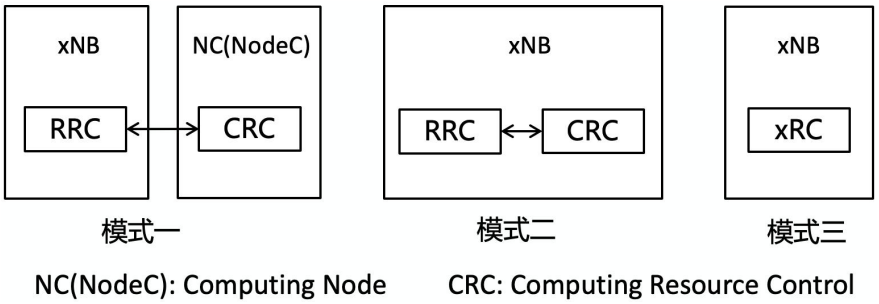


图 4.2-3：AI 计算和通信在控制面上融合的三种模式

模式一：无线网架构中引入新的逻辑计算单元，与基站独立，RRC（Radio Resource Control）和 CRC（Computing Resource Control）通过标准接口进行控制面交互。这种模式的好处是允许基站设备和计算单元设备间异厂商连接，部署方式更灵活，运营商可选择范围更大，缺点是外部接口时延较长，较难满足控制面实时性要求。

模式二：逻辑计算单元内置于基站内，属于基站功能范畴，RRC 和 CRC 通过内部接口进行控制面交互。这种模式的好处是基站内部接口实现性能较好，且无线通信资源和计算资源独立控制、按需调用，便于依据资源特性设计专用控制流程，也便于统计资源状态。

模式三：逻辑计算单元内置于基站内，属于基站功能范畴，RRC 和 CRC 融合成统一的资源控制实体（xRC），同时对连接和计算资源进行控制。这种模式的好处是同时决定连接和计算资源的控制决策，资源控制的协同和实时性最佳，但联合控制机制的设计较复杂，也不便于分别统计资源状态。

管理面基于 QoAIS 需求对算力和连接资源的编排，其优势在于对网元连接关系、各类资源状态具有宏观视角，可保证资源利用率或其他网络级性能指标较优。

计算和通信在控制面上的深度融合为 QoAIS 目标的持续达成提供了较高实时性的保障手段，其优势在于当发现 QoAIS 指标发生恶化时，可快速调整，例如，当连接带宽受限，但本地算力充足时，增加本地计算量，对所需传输的 AI 数据进行高保真度的压缩；当连接带宽充足、质量稳定，但本地算力受限时，减少本地计算量，通过增加周边节点的协作，共同完成该任务。在用户面上，AI 计算和通信的深度融合主要体现为对 AI 计算协议和通信协议的联合设计和优化，以同时满足性能和开销上的需求。在计算协议方面，对于同一项 AI 计算任务，异构的算力资源在计算精度、架构和流程上均可能存在不同的协议和配置参数，影响计算结果的准确度和计算耗时。在通信协议方面，考虑带宽及信道状态的不稳定性而对 AI 任务数据（如模型参数、模型计算中间结果、模型梯度等）进行的各种处理，如信源和信道的编解码，也存在多种协议参数的配置选项，影响传输时延和质量。由于 AI 任务的计算和通信常在时间上串行，共同影响 AI 任务的质量，这就为其联合设计和优化提供了可能，值得进一步思考和研究。

延伸问题：

1. 如何将管理面计算与通信的融合机制与控制面计算与通信的融合机制有效的结合起来，以便在满足 AI 服务 QoAIS 需求的同时，达到网络资源分配的均衡，资源和能耗效率较优？
2. 如何在用户面上对 AI 任务的计算协议和通信协议进行联合设计和优化，以同时满足性能和开销上的需求？

5 新架构

6G 网络从架构设计上融入 AI 要素是 6G 内生 AI 的最基本特点。随着 AI 三要素（数据、算法和算力）与网络连接一样成为网络内部的基本资源，其在网络架构的设计中就不能仅体现为局部的机制和流程创新，而是贯穿于 AI 全生命周期中的完备的功能、交互机制和信令流程。每种资源要素在具有自身内部的管理、控制、处理和传输机制之外，还会与其他资源之间协同，共同完成 AI 任务，满足 QoAIS 需求。因此，不同于 5G 网络，6G 网络将新增数据面、智能面、计算面，并产生维度大幅扩展的控制面和用户面。下图所示为 6G 多维融合网络的逻辑架构。

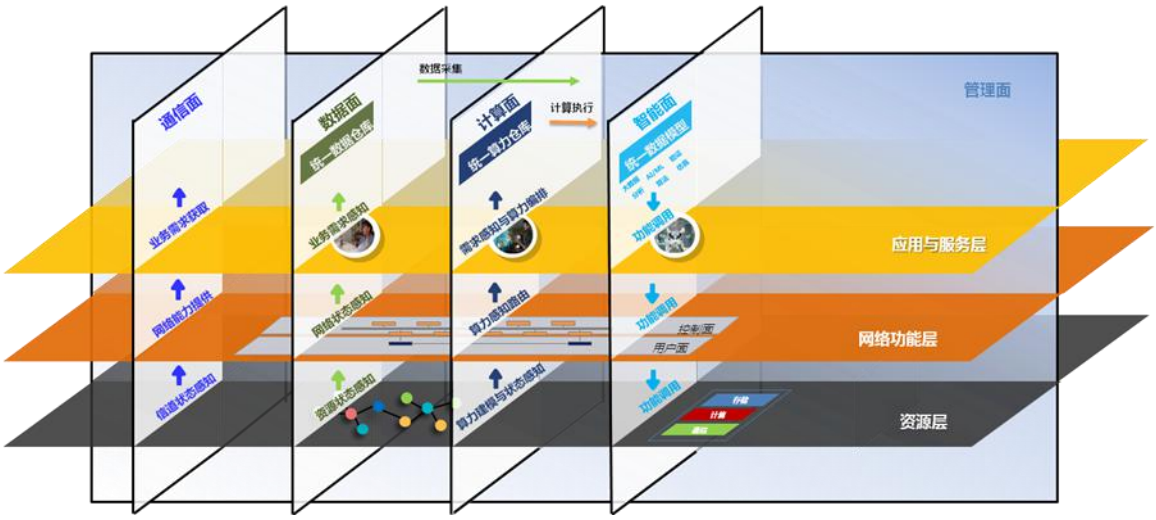


图 5-1：维融合网络逻辑架构

在横向逻辑上，6G 网络可划分为资源层、网络功能层和应用与服务层（“三层”）。资源层提供无线、计算、存储等底层资源，并为网络功能层的功能生成提供相应的支持和服务。网络功能层形成特定的网络功能，或将一种或多种网络功能组合在一起，提供最基本的网络服务能力，以满足应用与服务层的需求。应用与服务层为客户的业务和应用提供相应支持，实现服务定制化。在纵向逻辑上，除了承载传统通信业务的“通信面”，6G 网络将新增数据面、计算面和智能面。数据面负责端到端网络中数据的采集、清理、处理和存储，并向其他层和面提供数据服务。计算面提供统一的算力仓库，感知算力需求，编排计算任务，提供算力路由、算力建模与状态的感知，为其他层和面提供计算服务。智能面提供内生 AI 全生命周期所需的完整运行环境，调用数据面、计算面提供的服务，为其他层和面提供智能服务。管理面则是对其他所有层和面进行管理。

对于 6G 内生 AI 系统而言，前述新理念的实现主要体现在“三层”以及智能面、数据面、计算面和管理面上。值得一提的是，控制面和用户面隶属于网络功能层，传统意义上的控制面和用户面是网络功能层为支持传统通信业务提供的控制机制和业务数据的传输机制，在 6G 新增数据面、计算面和智能面后，这些面将产生新的业务数据（如数据面上采集和传输的各类数据、计算面上计算任务的输入输出和中间数据、智能面上 AI 模型的参数等），成为网络需要支持的新“业务”，因此网络功能层需要扩展控制面和用户面以提供相应的支持。

本章节重点介绍其中的数据面、智能面和扩展的控制面和用户面。

5.1 数据面

5G 网络智能化实践经验[8]表明，数据的获取非常困难，数据质量难以保证。因为在先前的网络架构和协议设计中没有预定义数据收集的接口，而当前基于实现的数据收集服务器/设备，例如深度包检测或数据探测无法及时提供足够的数据。基于网管的数据收集也存在数据种类较少，采集周期较长（15min）、异厂商数据格式、命名、计算方式不统一，南向网管数据难以开放的问题。同时，由于数据在设备内部采集的不稳定性、传输链路有损，网管设备存储空间有限，标签难获得，获取的数据常存在缺失、串行、无标签或标签错误等质量问题，在 AI 模型训练之前，需要花费大量的时间和人力成本对数据进行预处理。

针对上述挑战，6G 将通过在网络架构中新增“数据面”[7]来提供解决方案。数据面中的数据元素将涵盖网络内部和外部数据，具体包括业务数据、用户数据、网络数据、感知数据、外部数据、资源层数据等。基础数据服务包括数据采集、数据预处理、数据存储、数据访问、数据共享与协同等，基础数据服务具有如下技术特征：支持可信的认证、授权、访问，高效的数据存储和管理，按需动态的数据采集、数据预处理和聚合，对外能力开放和注入等。图 5.1-1 展示了 6G 网络数据面逻辑功能架构。

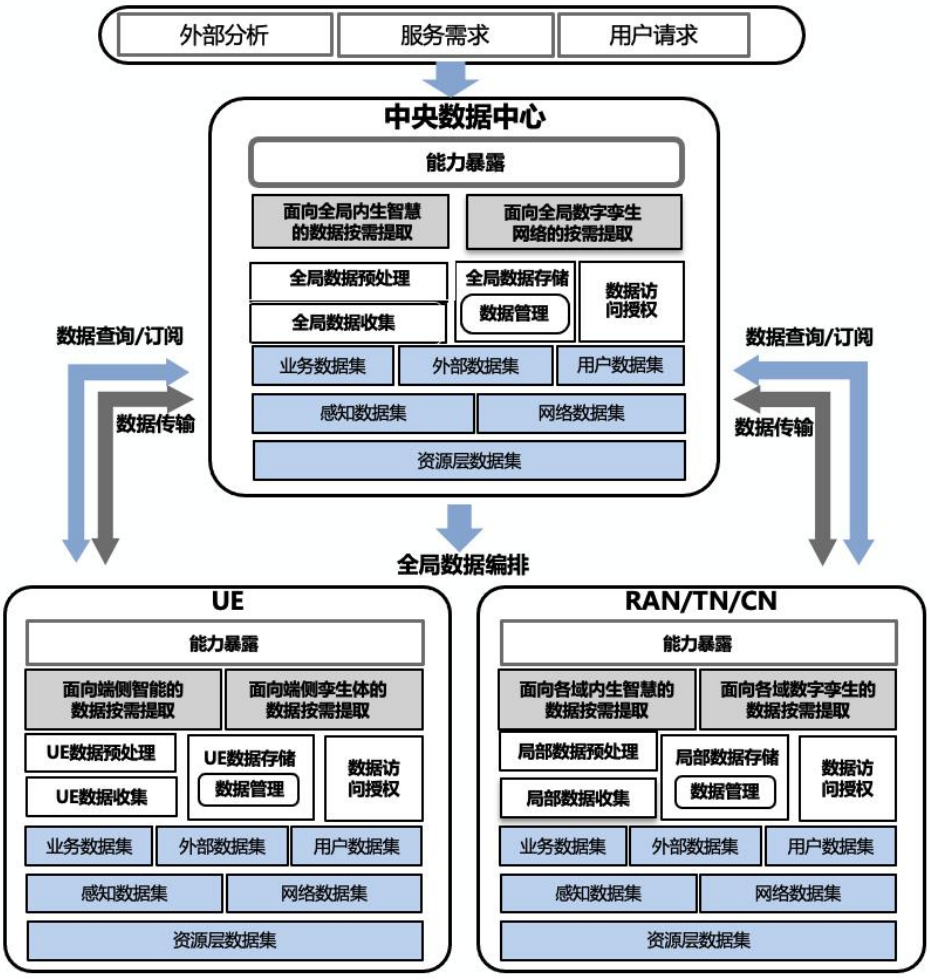


图 5.1-1：6G 网络数据面功能架构

6G 网络的数据面架构由中央数据中心和各域本地数据中心组成，是集中式与分布式相结合的架构。中央数据中心存储网络端到端全局数据，按需对数据进行全局编排；各域内部的本地数据中心存储和管理从本地网络采集到的数据，为域内各类上层应用提供数据服务；

网元内部存储和管理网元产生的实时数据，为网元内部和周边网元的上层应用提供数据服务。

由于 AI 服务所需要和所生成的数据也属于数据面数据元素，比如训练样本、AI 模型参数、模型中间计算结果、模型梯度、推理样本和推理结果等，因此 6G 网络的数据面为内生 AI 提供基础数据服务，可为各类内生 AI 服务所调用，贯穿于 AI 工作流全生命周期。比如 AI 服务通过调用数据面可信服务来保障 AI 服务 QoAIS 中的可信要求[9]；通过调用按需动态的数据采集和预处理服务，减少计算和传输开销，满足 QoAIS 中的开销要求；通过调用对外数据能力开放和注入服务，与行业用户进行数据交互，导入 AI 服务所需数据，提交服务生成的结果。

6G 网络中，“可信”将成为用户对数据服务的重要需求[10]。数据服务的可信主要体现在数据采集、数据存储、数据访问、数据共享与协同等阶段。数据采集阶段需要考虑数据的隐私性、公平性、数据采集的再现性、鲁棒性等。数据隐私性主要通过一些数据流程或技术来保障，如 debias 采样和注释，追朔数据源（包括数据来源、数据依赖关系等）等定性方法，以及数据匿名化、差分隐私等定量的方法。数据公平性主要通过定量的指标来评估，比如变量的相关系数、损失函数、完整笛卡尔积等。数据采集的再现性和鲁棒性可以通过数据朔源来保障。

延伸问题：

1. 如何从网络架构层面支持网络和网元深度数据的开放和使用？
2. 如何从网络架构层面支持内生 AI 对数据的按需动态提取？包括采集数据类型、采集数据量、采集方式、数据预处理方式等。

5.2 智能面

我们在上一章中提出的新理念涉及到管理面、控制面和用户面上新机制的设计，这些新机制为各类 AI 工作流的全生命周期提供了完整的运行环境，满足各类 AI 服务的 QoAIS 要求，我们将这一完整的运行环境称为 6G 网络的“智能面”。图 5.2-1 为 6G 网络智能面功能架构设计。

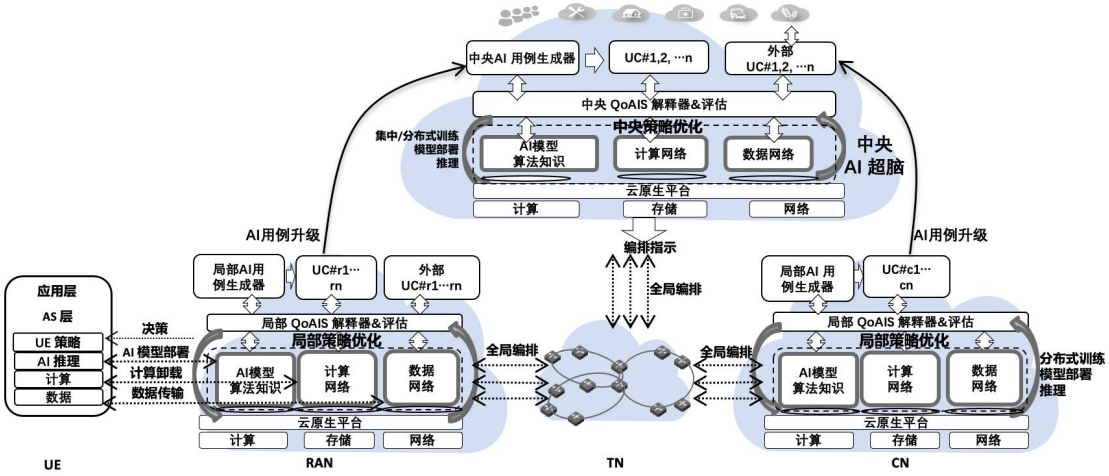


图 5.2-1：6G 网络智能面功能架构

6G 网络内生 AI 的智能面架构具有如下技术特征：

第一项技术特征是 AI 用例的自生成和导入。AI 用例是用户向网络一次性提出的 AI 服务请求，一个 AI 用例可能涉及到一类或多类网络内生 AI 服务（如 AI 训练、验证和推理服务）的调用。AI 用例描述是对用户所需 AI 服务在网络实操层面的框架性或辅助性信息描述。从该描述中，网络可获知在智能应用场景、输入输出数据、模型选择、模型训练、模型验证优化、以及实施模型输出的结论/决策等方面的信息。网络可通过基于自身数据分析或外部导入的方式，生成 AI 用例描述。管理面负责管理所有 AI 用例，调度实施 AI 用例，生成该用例所需的 AI 服务、AI 工作流和 QoAIS 要求，按需调配网络元素（包括数据、算法、算力、连接等）。图 5.2-2 展示了 AI 用例、AI 服务、AI 工作流和 AI 任务之间的逻辑关系。

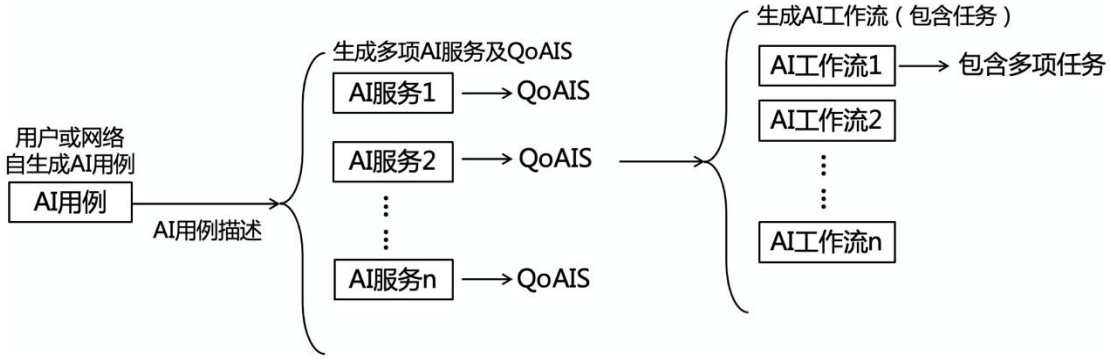


图 5.2-2：AI 用例、AI 服务、AI 工作流和 AI 任务之间的逻辑关系

其次，QoAIS 的生成。QoAIS 是网络内生 AI 服务的质量评估指标体系。一个 AI 服务对应一套 QoAIS，一个 AI 用例对应的 QoAIS 由其包含的所有 AI 服务对应的 QoAIS 组合构成。当网络收到一个 AI 用例后，需要获知该用例对应的 QoAIS 要求，以便分解到对各类资源的编排、调度和控制的具体要求上。获知的方式有两种：一种是外部导入，比如在外部导入 AI 用例描述的同时，即包含 QoAIS 要求；一种是内部生成，比如对于网络根据上层意图信息生成 AI 用例的场景，网络也可以根据意图信息同时生成 QoAIS 的指标要求。

第三，AI 工作流全生命周期承载于网络内部。网络管理面为 AI 服务生成其所需的各类 AI 工作流（包括数据采集、预处理、数据扩展、数据分析；模型选择、训练、调参；模型验证、集成、监测和更新等）、编排所需资源、监测工作流状态、优化工作流以满足 QoAIS 要求、直至所有任务完成。AI 工作流的全生命周期过程均在网络内部完成。在要求可信 AI 的场景下，需少量的人工干预。

第四，管理面、控制面和用户面协作保障 QoAIS 的持续达成，主要通过对 AI 的三要素（算法、算力、数据）以及网络要素（连接）的编排和控制达成。管理面负责起始阶段的资源编排和过程中较慢速的资源分配调整，控制面和用户面进行实时的 QoAIS 保障，根据网络环境的动态变化，进行过程中连接、算法、算力和数据的快速调整。

第五，AI 集中式与分布式架构相结合。中央 AI 超脑算力充足，存储量大，数据抓取范围大，适用于模型规模大（如大规模通用 AI 模型）、性能要求高、实时性要求较低的智能应用场景，所需数据跨域的场景，包含用户相关数据的场景。无线、传输和核心网各域内 AI 小脑作为域内集中式 AI 引擎节点，负责本地域内可完成的 AI 用例。各域内分布式部署的网元节点算力和存储有限，将通过网元间协作，支持本地实时性要求较高的智能应用场景。当本地域内 AI 用例的 QoAIS 无法在域内达成时（比如缺少其他域的特征数据、缺少算力资源），则该用例上升到中央 AI 超脑，通过全局资源编排来达成。这种分级分域的部署架构可减轻单一集中的超脑面临性能压力，并兼顾到各种智能应用场景的性能需求。

5.3 扩展的控制面和用户面

既有移动通信网络的控制面和用户面是面向传统通信业务（包括语音、数据包传输、流媒体等）的质量需求设计的，其主要目的是为数据传输提供连接、支持用户移动性、保证其业务体验。在资源类型上，采用专用算力资源，对计算和存储资源的需求量均不高。与传统通信业务不同，AI 服务属于数据和计算密集型业务，AI 服务的内生将为 6G 网络引入新的资源维度（包括异构的算力资源和存储资源、新型的计算任务（AI 算法）以及 AI 所需的和生成的数据（后称“AI 数据”），因此，需要设计新维度资源的管理和控制机制，同时需要面向 AI 服务的输入、输出和过程中数据设计高效的用户面机制，即 AI 服务将成为 6G 网络的一种特殊“用户”。这些将大幅扩展传统移动通信网络中的控制面和用户面。

我们将 6G 网络为了支持 AI 服务质量的达成而设计的新的控制面和用户面机制、协议和流程分别叫做“AI 的控制面（AI CP）”和“AI 的用户面（AI UP）”。表 5.3-1 展示了 AI CP 和 AI UP 与传统移动通信网络中 CP 和 UP 的对比。

表 5.3-1：AI CP 和 AI UP 与传统移动通信网络中 CP 和 UP 的对比

传统通信业务	**内生 AI 服务**				
连接	连接	算力	算法	数据	多维度资源
NF CP	AI 连接的控制机制	AI 算力的控制机制	AI 算法自优化的控制机制	AI 按需动态的数据采集和处理控制机制	AI CP
NF UP	AI 数据的传输机制	AI 计算任务的执行机制	AI 算法自优化的处理流程	AI 所需数据的处理机制	AI UP

传统的“控制面”和“用户面”之所以是“面”是因为传统通信业务的端到端连接特性使得相关控制和传输机制渗透到终端、无线、传输与核心网各个域，不同域的控制和传输机制组合成控制面和用户面。然而，AI 服务的数据流不再具有“端到端”的传输特性，取而代之的是根据管理面的编排，在一定范围的网元节点内，采用不同维度资源间的组合来完成 AI 服务中的一项具体的工作任务。因此，AI 的控制面是由多种维度资源上的控制机制组成的，AI 的用户面是由多种维度资源上的数据处理机制组成的。“面”的内涵与传统通信业务不同。随着，AI 计算与通信在控制面和用户面上越来越深度的融合，AI“面”的有机性将越加明显。

虽然都是对连接资源的控制和使用，由于 AI 数据（比如训练样本、推理结果、模型参数、训练/推理的中间计算结果、模型梯度等）的传输模式、数据类型、数据量大小、数据元素间的结构和关系、对信道变化的鲁棒性、对用户移动性的支持方式，用户的参与模式，及其对 AI 服务质量的影响与传统通信业务有所不同，现有网络各域通信连接的控制和传输机制是否依然适用尚未知，可能需要针对 AI 数据设计专门的连接控制机制和数据传输协议，也可能可以用同一个功能模块同时服务于传统通信业务和 AI 服务。图 5.3-1 展示了 AI 连接与传统连接在控制机制和数据传输协议上的两种关系模式。

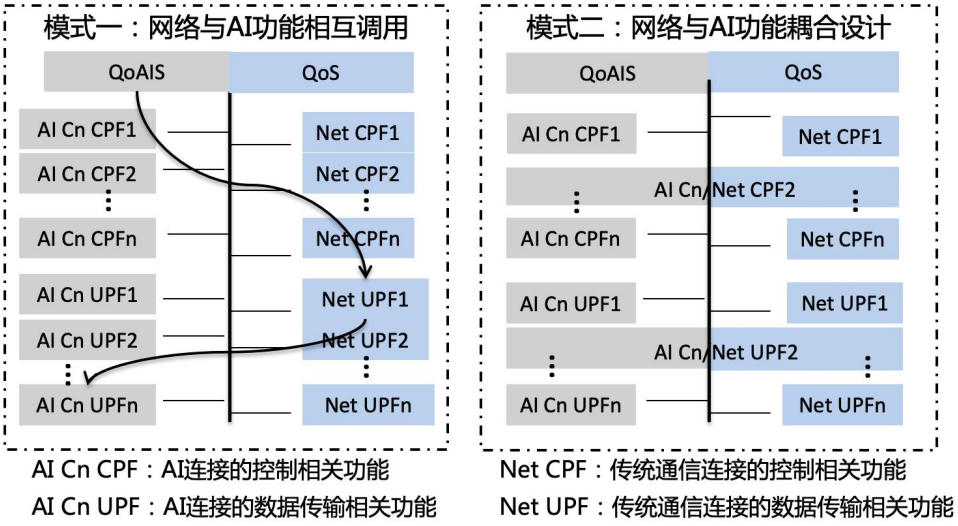


图 5.3-1：AI 连接与传统连接在控制机制和数据传输协议上的关系模式

延伸问题：

1. 通过应用层方式提供 AI 服务存在哪些弊端？采用当前通信连接的控制和用户面协议提供 AI 连接是否可满足 AI 服务的质量需求？需在哪些方面进行改进或创新？
2. 连接、算力、算法和数据资源在数据处理流程和控制机制上，有何关联关系？如何协同优化？

6 新技术

当 AI 全生命周期内生于网络后，AI 各阶段，包括数据采集、模型选择、模型训练、推理、性能评估和闭环优化等各环节均需要相应的技术来支撑。区别于当前在资源集中环境中的技术实现，在大规模节点分散部署模式的无线网络中则需要多资源节点间的分布式协作完成 AI 任务，需要设计计算与通信融合的具体机制。根据 AI 任务的不同，协作方式与融合机制也有所不同。另一方面，为了保证 AI 训练过程和推理结果的实施对网络性能不产生负面影响，为内生 AI 建立一套真实物理网络的数字化镜像网络环境显得尤为重要。基于网络的数字孪生构建的数字化网络环境或可满足此需求。本章前三节介绍了模型编排、训练和推理任务相关的新技术，最后一节介绍了内生 AI 与孪生的数字化网络之间的交互技术。

6.1 AI 模型的选择与再训练

AI 模型的编排技术是 AI 全生命周期编排管理的关键技术之一。在模型训练、模型验证、模型部署运行和模型迁移的 AI 模型全生命周期管理环节中，考虑到模型的复杂程度差异巨大，导致训练开销，推理开销，迭代优化？开销和推理时延等的不同，需要结合不同用例的需求，算力和通信资源的动态变化，进行模型的选择和再训练。

通过理论分析和实验发现，AI 模型的深度和复杂度对模型性能及效果影响明显，AI 模型越复杂，找到更优解的概率越大[11]。但是模型越复杂，收敛训练所需的开销就越大。为了降低对训练开销（如训练数据大小，算力）的要求，现有通常的做法是引入模型的再训练，

即选取源域大数据集上训练得到的性能良好的人工智能基础模型结构和权重,然后在基础模型上利用目标域数据进行再训练，学习和优化源域和目标域的分布偏差。

模型的再训练需要解决的很重要的问题就是选择一个合适的基础训练模型结构和权重。将 AI 模型的选择和再训练应用到无线通信系统中，需要考虑以下因素：

1. 由于实际系统中收发信机器件的非线性功放差异，复杂的空口物理环境，终端的分布和移动等因素导致无线信道数据多变使得需要引入再训练提升模型的适应性更加迫切，但是无线数据的采集困难，需要考虑选择在目标域小样本数据集上迁移性能较好的基础训练模型。
2. 需要考虑在分布式节点上进行基础模型的再训练，甚至在终端上进行。在分布式节点/终端上进行再训练，由于算力、功耗等受限，不可能遍历所有的基础 AI 模型，如何在保证较低的再训练开销下，选择合适的基础模型成为一个很重要的问题。

表 6.1-1 列出了三种基础训练模型下学习的性能和开销比较。其中基础学习基于基础模型结构和随机的权重值；迁移学习利用源域上的大量数据训练基础模型的权值；元学习[12]利用源域的数据进行“学会学习”的元知识能力学习，学习源域数据的迁移特征和能力。

表 6.1-1: 三种基础训练模型下学习的性能和开销比较

	基础学习	**迁移学习**	**元学习**
源域模型	基础模型结构和随机权重	基于大量数据学习模型结构和权重	基于大量数据学习元知识
源域模型训练开销	无	开销较大(1 倍)	最大 (10 倍)
目标域模型性能	小样本性能低 大样本性能高	分布差异大：小样本性能中 分布差异小：小样本性能高	小样本性能高
从源域到目标域的再训练开销	开销大	开销很小	开销很小

在系统设计中引入 AI 再训练的基础模型编排功能，来实现 AI 基础模型的选择。引入源域和目标域数据分布差异评估功能来选择 AI 基础模型的训练方法，比如分布差异大且充足的训练资源,考虑在源域数据集上使用元学习来训练基础模型,在使用元学习训练学习“元知识”过程中，为了更好的学习数据样本间的迁移特征，对源域数据集进行数据分布的分析，动态编排元学习的迁移特征数据集合，同时采用分布式来加速模型的训练。如果分布差异较小且训练资源相对受限，则考虑使用迁移学习来训练基础模型。此外，在源域性能最优的基础模型在目标域的性能不一定最优，我们可以获取目标域的数据特征和分布，来优选基础训练模型结构和初始权重。

6.2 终端与网络协作的 AI 模型训练

分布式 AI 模型训练是指利用云、边和端侧部署的分布式算力资源协同进行 AI 模型的训练，在不同应用场景下，可达到提高计算资源利用率、提升模型性能、保护数据隐私的目的。如章节 4.3 所述，随着 AI 计算与通信的深度融合，分布式异构算力资源的协同调度需

要实时适配网络动态复杂的通信环境，需要控制面和用户面进行实时支持，进而产生章节 5.3 所述“扩展的控制面和用户面”。

对于数据驱动的 AI 模型训练，传统方式包括基于集中式算力和数据的模型训练和基于分布式并行计算模式的模型训练。后者通常在计算机集群中，将数据或模型分割到不同的计算节点上并行计算，并通过集中计算节点进行聚合处理产生最终结果[13]。由于计算机集群网络环境比较稳定可靠，训练数据集的分布已知且可操作，模型训练过程的理论建模比较容易，性能有所保障。然而，在移动通信网络中，面临着网络环境动态变化，信道质量不稳定，用户移动性、数据源无法由网络单方决定，训练数据非独立同分布等等现实复杂的问题，因此 6G 网络内生 AI 的分布式模型训练技术必然会比计算机集群网络环境下更加复杂。

当前，业界已提出较多分布式 AI 模型训练的技术框架，比如（分层）联邦学习[14]、群学习[15]、多智能体学习[16]、基于模型分割的学习[17][18]等。但大多是在较理想的环境下基于一定的理论假设对算法或模型的性能进行研究，缺乏复杂网络环境下系统性能的理论建模。在此情况下，对于 AI 而言，模型学习的性能是否能得到保证？对于通信网络而言，通信资源开销和效率是否可以接受？都是有待研究和确认的问题。

我们基于对无线通信网络的理解和 AI 模型训练过程的研究，初步认为在无线网络中进行终端与基站协作的分布式 AI 模型训练时，训练过程会产生大量中间计算结果，需频繁占用空口无线资源进行传输，传输的时延和误码率情况会对训练结果产生影响。为了在保证模型收敛的同时，提高空口无线资源的利用率，引入效率更高的高阶模型学习算法是一种值得考虑的思路[19][20][21]。由于不同阶数（零阶、一阶随机梯度下降、二阶牛顿方法等）的模型学习算法在训练速度和资源开销上各有优劣势，可以考虑根据无线信道状态动态调整学习算法，即多种学习算法间的动态转换机制。图 6.2-1 是原理示意图，图 6.2-2 展示了为引入这种动态转换机制设计的功能交互。

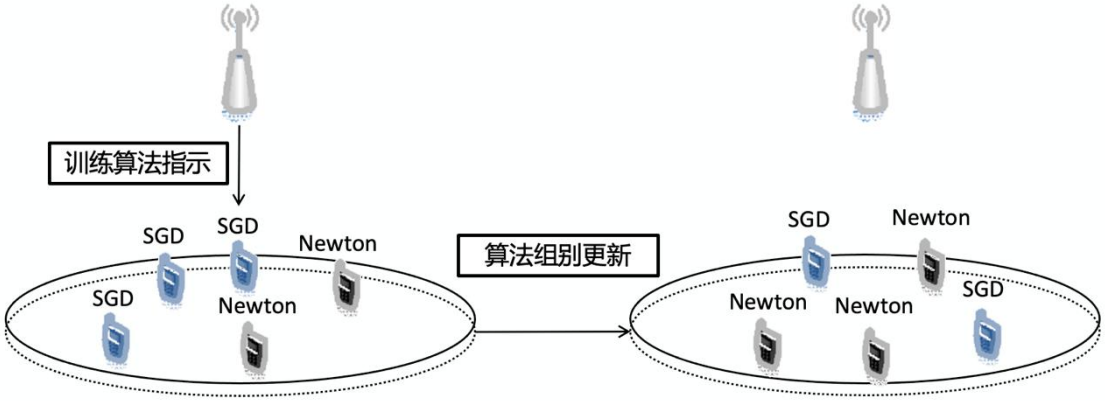


图 6.2-1：多种学习算法动态转换原理示意图

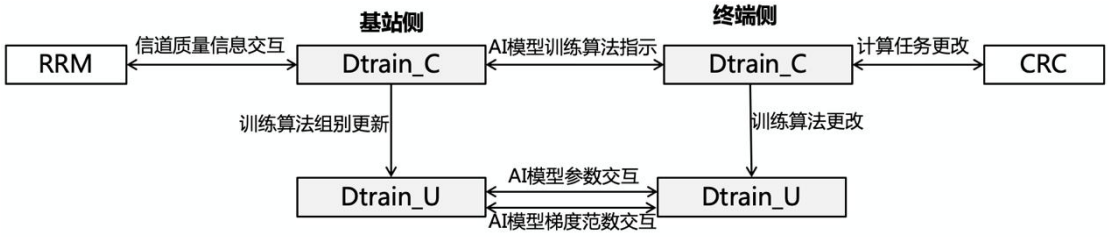


图 6.2-2：多种学习算法动态转换机制的功能交互

如 5.3 中所述，上述新技术需要在空口引入针对 AI 连接的控制机制和数据传输协议，在上图中分别用 Dtrain_C 和 Dtrain_U 来表示。其中，Dtrain_C 是负责控制终端与基站协作

进行 AI 模型训练的控制功能实体，在前述场景中该实体根据信道质量的变化，动态调整参与分布式训练的终端采用的模型学习算法。Dtrain_U 是负责终端与基站协作进行 AI 模型训练的业务面功能实体其包含有在基站与终端间传输模型参数、梯度或梯度范数等信息所需的专用协议栈。由于上述信息数据的传输模式与传统通信业务不同，对空口信道可靠性的要求也不同，因此需要针对性设计其空口传输协议。

延伸问题：

1. 将网络中的连接特性与 AI 模型训练特性相结合而做出的技术改进和优化，理论上可提升 6G 网络内生 AI 分布式训练模式的可行性和性能，但本质上未能改变基于数据驱动的 AI 模型训练范式，是否有可能探索基于模型驱动的新训练范式，从而实现算法/模型的自生长？

6.3 终端与网络协作的 AI 模型推理

分布式 AI 模型协作推理是指利用云、边和端侧部署的分布式算力资源协同进行 AI 模型的推理，可提高计算资源利用率、应对端侧算力不足、保护数据隐私。如章节 4.3 所述，随着 AI 计算与通信的深度融合，分布式异构算力资源的协同调度需要实时适配网络动态复杂的通信环境，需要控制面和用户面进行实时支持，进而产生章节 5.3 所述“扩展的控制面和用户面”。

基于模型分割的端边协作推理是近年来业界提出的一种无线网络中的分布式协作推理框架。当终端需完成一项模型推理任务，而自身算力又不够时，可通过该框架获得网络侧算力资源的协助，共同完成推理任务，满足推理的性能需求。需要做出的决策是如何切分模型，即对于如下有向无环深度神经网络而言，决定从哪两层之间进行切分，将切分点左部分放在终端侧计算，右部分放在网络侧计算。

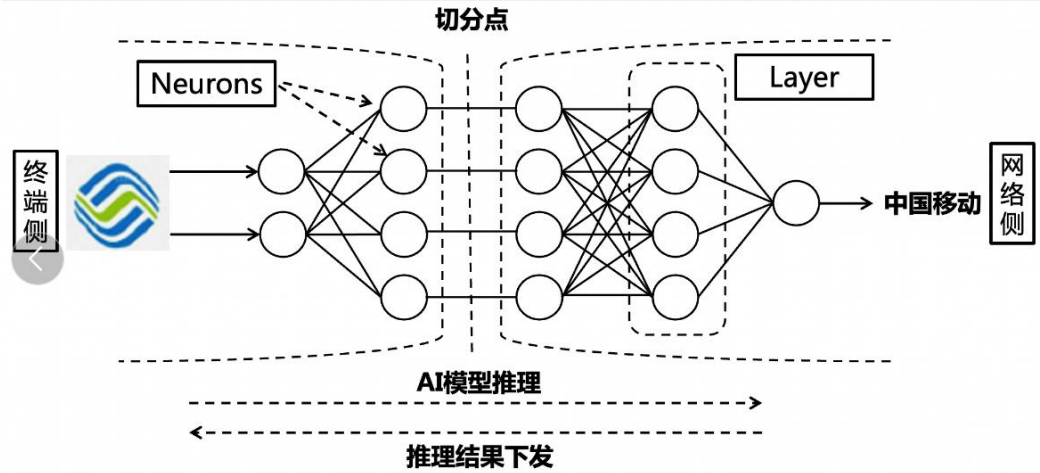


图 6.3-1：端边协作推理示意图

时延和准确度是 AI 推理服务的两项重要性能指标。模型分割点的决策将影响终端侧和基站侧分别负责的计算任务量，以及空口需要传输的数据量。因此，影响这两个指标的包括模型分割点、终端侧计算资源分配、空口无线资源分配和基站侧计算资源分配等环节，需要进行通信与计算资源的协同调度。尤其是当终端和基站侧均存在多种异构计算资源时，计算单元类型和数量的分配会对推理时延产生直接影响。因此，在 AI 连接的控制机制设计上需要考虑如何同时决定模型分割点、终端侧和基站侧的计算资源分配方案，并随着网络环境的

变换，进行动态调整，从而保障推理性能的持续达成。可考虑通过强化学习给出较优决策方案，相关变量设计如下表所示。

表 6.3-1：基于强化学习的端边协作推理方案相关变量设计

变量	基站侧	终端侧
状态（State）	基站内各类型计算资源的剩余可分配算力、计算单元间传输带宽（可选）、基站侧剩余可分配上下行空口信道传输资源、终端上行信道质量	终端内各类型计算资源的剩余可分配算力、计算单元间传输带宽（可选）、终端下行信道质量
动作（Action）	基站侧负责计算的模型部分参数、上下行带宽分配、基站侧计算资源分配	终端侧负责计算的模型部分参数、终端侧计算资源分配
奖励（Reward）	基站侧推理能耗等	推理性能指标、终端推理能耗等

图 6.3-2 展示了为引入这种机制需设计的功能交互。如 5.3 中所述，上述新技术需要引入针对 AI 连接的控制机制和数据传输协议，在下图中分别用 Dinfer_C 和 Dinfer_U 来表示。控制实体 Dinfer_C 可以根据无线通信资源和计算资源的状态变化，动态配置模型分割点、无线资源和计算资源的联合分配，数据传输实体 Dinfer_U 包含有在基站与终端间传输模型推理计算的中间结果和推理最终输出结果所需的专用协议栈。

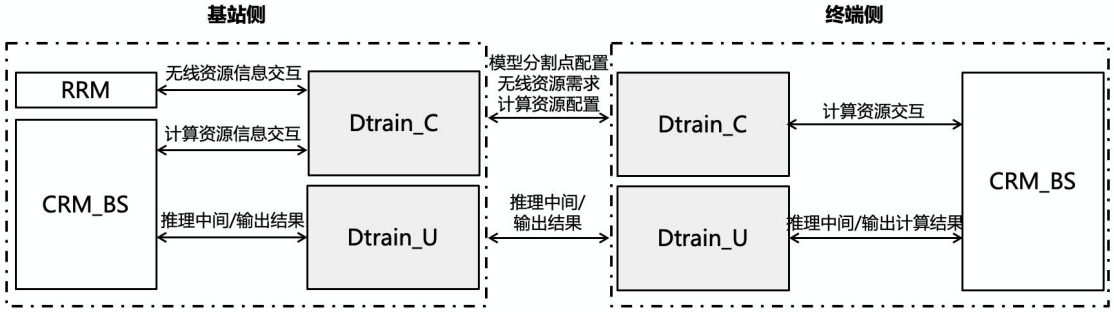


图 6.3-2：端边协作推理方案的功能及交互

延伸问题：

1. 对于对实时性要求或时延稳定性要求较高的 AI 服务（如推理），如何对计算资源进行精准分配？是否可以对异构计算资源的类型、算力容量、架构、连接带宽等特征与计算性能（时延、能耗、精度等）之间的关系进行抽象建模，以依据模型实现对算力资源的精准分配？

6.4 基于网络数字孪生的 AI 性能预验证技术

数字孪生网络是一个由物理网络实体及其孪生的数字化网络构成的，且物理网络与孪生的数字化网络间能进行实时交互映射的网络系统。物理网元对应的孪生的数字化网元可以通过各种数据采集和仿真模拟手段来构建，进而在数字域形成网元的数字孪生体和网络的数字孪生体。在该系统中，各种网络管理与应用可以利用网络的数字孪生体，基于数据和模型对物理网络进行高效的分析、诊断、仿真和控制[6][7]。

AI 模型生命周期管理中的模型验证技术使用验证数据集对训练好的模型集合进行选型，但是验证数据集和训练数据集通常同分布，缺乏分布多样化的验证数据集用于模型验证和选型。如何在孪生的数字化网络环境中提升模型泛化性能是需要解决的关键技术问题之一。在该环境中产生比物理环境更多场景下的样本数据，可以降低对物理网络数据采集开销和性能影响，并通过性能预验证迭代训练鲁棒性更强，性能更优的 AI 模型。一方面，在物理环境中利用内生 AI 对数据进行蒸馏，将原始数据分布抽象和压缩成 AI 模型，降低数据传输的需求，同时实现数据增广，满足孪生网络层数据样本多样性扩展的需求；另一方面，在孪生网络层中引入样本空间的探索来进一步增加样本的多样性，达到性能预验证的目的。

条件对抗生成网络 CGAN [22]可以通过动态改变环境条件，来动态生成符合特定分布的环境模型，环境模型可以包括用户分布模型，用户信道模型，用户业务模型，网络状态模型，网络资源分配模型等。如图 6.4-1 所示，在物理网络中引入条件对抗生成网络，将随机序列和一定语义的环境条件作为输入，通过生成模型和辨识模型的博弈训练，达到纳什均衡来生成符合特定数据分布的模型。物理网络将对抗训练生成的环境模型发给孪生网络层，后者可以有选择性的改变环境条件来生成符合特定分布的环境数据，从而实现对环境数据的蒸馏和增广。

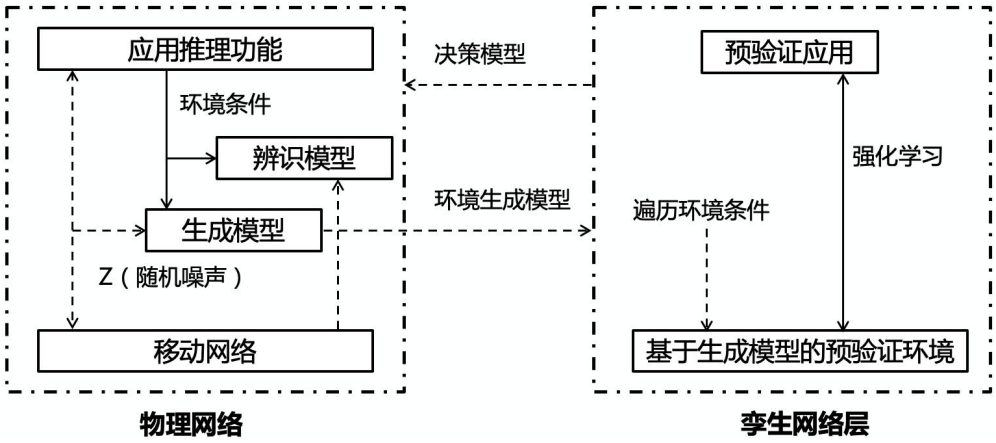


图 6.4-1：基于网络数字孪生的模型性能预验证流程

除了引入条件对抗生成网络来产生多样化的数据样本之外，为了实现更多场景的预验证，孪生网络层引入强化学习来进行数据样本空间的搜索和遍历。考虑到状态和动作空间可能非常巨大，探索样本空间的遍历成本高，一方面可以引入分布式的搜索代理加快遍历速度，另一方面也需要引入高效的样本空间搜索算法，其中结合物理网络的反馈来确定搜索的维度和方向，另外也可以在部分维度上引入静态或者半静态的环境模型，减少样本空间的探索开销。

此外，在物理网络和孪生网络层之间引入双向闭环优化机制，孪生网络层的预验证应用通过与预验证环境的交互采集了完善的样本数据后，对 AI 模型进行训练，生成物理网络需要的决策模型。另外决策模型在物理网络的推理过程中生成与决策模型预期产生的偏差，并传给孪生网络层用于预验证环境的纠偏。

需要进一步考虑的是，对抗生成网络目前主要应用在图像领域，模型收敛和评估的指标目前主要考虑了图像的多样性和还原度（比如 IS, FID），需要针对移动网络数据的分布特点设计合适的指标。

7 总结与展望

从 5G 网络智能化实践中，我们发现外挂式 AI 存在诸多短板和性能不足，但由于 5G 架构、协议功能和流程已经定型，只能在现有架构方案上做增量迭代，效果有限。同时，面向 6G，智能的需求场景将更加广泛，除服务于网络自身的高水平自治、性能优化、为 to C 客户提供极致业务体验外，还将助力千行百业的数智化转型，这将极大丰富智能应用场景，对智能服务种类的需求将更加多样化，对智能服务性能的需求也将更加明确和多维。这些都要求改变当前外挂式 AI 的应用模式，实现内生式 AI。

6G 网络实现内生 AI 需要新理念的提出、新架构的设计以及新技术的支撑。在新理念方面，本白皮书提出 6G 网络将具有 AI 服务质量（QoAIS）的评估指标体系和闭环保障机制，将提供包括数据采集，数据预处理，模型训练，模型推理，模型评估等 AI 工作流全生命周期的完整运行环境，不仅对 AI 全生命周期中各工作流所需资源进行编排管理，还将 AI 服务所需的算力、数据、算法、连接与网络功能、协议和流程进行深度融合设计，以满足实时高效的性能需求。同时，本白皮书还提出将内生 AI 与数字孪生网络融合，后者对 AI 模型和工作流的效果进行预验证，前者对数字孪生网络的性能进行优化。在新架构方面，本白皮书提出 6G 网络将新增数据面和智能面，并大幅扩展传统的控制面和用户面。其中，数据面将为内生 AI 和数字孪生网络提供基础数据服务，智能面将提供 AI 工作流全生命周期的完整运行环境，扩展的控制面和用户面则是从与传统通信业务对比的角度，提出了由于网络中引入 AI 服务所需的多维度资源而产生的新的控制机制和数据处理机制。在新技术方面，本白皮书列举了在模型编排、训练、推理、内生 AI 与数字孪生网络的融合等方面的关键技术，以支撑架构中的创新性设计。

当前，业界对于 6G 网络内生 AI 的需求、概念和内涵已逐渐达成共识，内生 AI 网络架构和关键技术体系仍在积极研究和讨论中，尚无统一的设计方案。面向 6G 时代丰富的智能应用场景设计智慧内生的网络架构不仅需要具备对传统移动通信网络的深入理解，还需要准确把握未来各类潜在用户对智能服务的质量需求，同时具备对 AI 全生命周期工作流编排管理、AI 服务所需数据、算力、算法和连接等各类资源的融合控制等方面的深入理解，是一项巨大的挑战。

为此，我们联合来自运营商、设备商、互联网服务商、高校等 18 家单位于 2020 年 12 月共同发起成立了 6GANA（6G Alliance of Network AI））论坛。6GANA 定位为全球性论坛，专注于 6G 网络 AI 相关技术、标准化、监管和产业的持续探索和推广。它旨在通过整个生态系统的联合研究，包括 ICT（如芯片制造商、网络基础设施供应商、移动网络运营商）、垂直行业、人工智能服务提供商、人工智能解决方案提供商、人工智能学术界和其他利益相关者，形成业界共识，推动 AI 能够成为 6G 网络全新的能力与服务[23]。6GANA TG2 是 6GANA 下负责研究基础网络架构的工作组，其将识别 6G 网络内生 AI 的基本技术特征，研究其对 6G 网络架构的影响，对标准化的影响，构建 6G 网络内生 AI 整体框架，定义基础架构，并对涉及的关键使能技术进行探讨。面向该目标，TG2 成员单位经过全面收集和充分探讨，与 2021 年 12 月凝练出现阶段业界广泛关注的、对 6G 网络架构存在潜在影响的十大核心技术问题，形成《6G 内生 AI 网络架构十问》白皮书。本白皮书内容为《6G 内生 AI 网络架构十问》中部分核心技术问题提供了重要的参考答案。

最后，我们倡议产业链各方合作伙伴携手创新，聚焦如下关键技术问题开展更深入的研究和广泛的探讨：

- 未来多样化智能应用场景对 AI 服务的质量需求（QoAIS）体现在哪些方面？相比传统 QoS 会出现哪些新的评估维度？如何从网络架构上支持上述需求指标的获得和评估？
- 为保证 QoAIS 的持续达成，如何分别在管理面、控制面和用户面上设计不同资源维度（数据、算力、算法、连接）的处理机制，并进行协同和融合？
- 如何从网络架构层面支持网络和网元深度数据的开放和使用？如何从网络架构层面支持内生 AI 对数据的按需动态提取和处理？
- 采用传统通信连接的控制和用户面协议提供 AI 连接是否可满足 AI 服务的质量需求？需在哪些方面进行改进？
- 内生 AI 与数字孪生网络之间是什么关系？如何从网络架构上支持两者的深度融合？
- 若网络为保障 QoAIS 的达成使用了多种 AI 技术，导致出现网络问题，如何进行问题的溯源和恢复处理？
- 模型的可信如何保障？若模型评估阶段效果较好，但遇到数据问题或突发情况，造成模型效果变差，该如何及时发现，如何处理异常？

缩略语

缩略语	英文全称	中文全称
AI	Artificial Intelligence	人工智能
ML	Machine Learning	机器学习
QoS	Quality of Service	服务质量
QoAIS	Quality of AI Service	AI 服务质量
AIaaS	AI as a Service	AI 即服务
NMSE	Normalized mean square error	归一化均方误差
KPI	Key Performance Indicator	关键绩效指标
GPU	Graphics Processing Unit	图形处理器
NPU	Neural-network Processing Unit	嵌入式神经网络处理器
DPU	Data Processing Unit	数据处理器
TPU	Tensor Processing Unit	神经网络加速器
MEC	Mobile Edge Computing	边缘计算技术
UPF	User Plane Function	用户面功能
CPF	Control Plane Function	控制面功能
RRC	Radio Resource Control	无线资源控制
CRC	Computing Resource Control	计算资源控制
xRC	x Resource Control	资源控制实体
CP	Control Plane	控制面
UP	User Plane	用户面
Dtrain_C	Distributed Training ControlPlane Unit	分布式训练控制面单元
Dtrain_U	Distributed Training UserPlane Unit	分布式训练用户面单元
CGAN	Conditional Generative Adversarial Networks	条件生成式对抗网络

编写人员

本白皮书由中国移动通信有限公司研究院如下人员共同编写：

未来研究院：邓娟、李刚、郑青碧、温子睿、潘成康、王启星、刘光毅

人工智能与智慧运营中心：李光宇、蔡海涛、梁燕萍、赵鹏、余立

参考文献

[1] 刘光毅,邓娟,郑青碧,李刚,孙欣,黄宇红.6G 智慧内生：技术挑战、架构和关键特征[J].移动通信,2021,45(04):68-78.

[2] Wen C K, Shih W T, Jin S. Deep learning for massive MIMO CSI feedback[J]. IEEE Wireless Communications Letters, 2018, 7(5): 748-751.

[3] Liu H, Wang Y, Fan W, et al. Trustworthy ai: A computational perspective[J]. arXiv preprint arXiv:2107.06641, 2021.3GPP TS 38.323, “NR; Packet Data Convergence Protocol (PDCP) specification.”

[4] Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges[J]. ACM Computing Surveys (CSUR), 2021, 54(5): 1-39.

[5] 马洪源,肖子玉,卜忠贵,赵远.5G 边缘计算技术及应用展望[J].电信科学,2019,35(06):114-123.

[6] Deng J, Zheng Q, Liu G, et al. A Digital Twin Approach for Self-optimization of Mobile Networks[C]//2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, 2021: 1-6.

[7] Liu G, Li N, Deng J, et al. 6G Mobile Network Architecture-SOLIDS: Driving Forces, Features, and Functional Topology[J]. Engineering, 2021.

[8] 中国移动：中国移动自动驾驶网络白皮书[R/OL].(2021)[2021-11-28]. https://www.sohu.com/a/492610271_121015326

[9] Li B, Qi P, Liu B, et al. Trustworthy AI: From Principles to Practices[J]. arXiv preprint arXiv:2110.01167, 2021.

[10] Toreini E, Aitken M, Coopamootoo K, et al. The relationship between trust in AI and trustworthy machine learning technologies[C]//Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020: 272-283

[11] Choromanska A, Henaff M, Mathieu M, et al. The loss surfaces of multilayer networks[C]//Artificial intelligence and statistics. PMLR, 2015: 192-204.

[12] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning. PMLR, 2017: 1126-1135.

[13] Shukur H, Zeebaree S R M, Ahmed A J, et al. A state of art survey for concurrent computation and clustering of parallel computing for distributed systems[J]. Journal of Applied Science and Technology Trends, 2020, 1(4): 148-154.

[14] Liu L, Zhang J, Song S H, et al. Client-edge-cloud hierarchical federated learning[C]//ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020: 1-6.

[15] Warnat-Herresthal S, Schultze H, Shastry K L, et al. Swarm Learning for decentralized and confidential clinical machine learning[J]. Nature, 2021, 594(7862): 265-270.

[16] Xu X, Li R, Zhao Z, et al. Stigmergic Independent Reinforcement Learning for Multiagent Collaboration[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[17] Shao J, Zhang J. Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems[C]//2020 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2020: 1-6.

[18] Teerapittayanon S, McDanel B, Kung H T. Distributed deep neural networks over the cloud, the edge and end devices[C]//2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017: 328-339.

[19] Li T, Sahu A K, Zaheer M, et al. Feddane: A federated newton-type method[C]//2019 53rd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2019: 1227-1231.

[20] Hua S, Yang K, Shi Y. On-device federated learning via second-order optimization with over-the-air computation[C]//2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall). IEEE, 2019: 1-5.

[21] Ghosh A, Maity R K, Mazumdar A, et al. Communication efficient distributed approximate Newton method[C]//2020 IEEE International Symposium on Information Theory (ISIT). IEEE, 2020: 2539-2544.

[22] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[23] 6GANA Whitepaper

数字孪生 智慧泛在