QoAIS 指标体系研究报告

6GANA

2023-09-09

前言

面向 6G 的智能普惠愿景，6G 网络将支持原生智能，将通信、信息和数据技术以及工业智能深度集成到无线网络，提供适应不同应用场景的智能能力。不同于传统的移动网络，6G 网络将基于原生智能架构，为用户提供泛在的智能服务，实现 6G AI 即服务（AIaaS）。相应地，需要从多个维度来构建 6G AIaaS 的性能指标和服务质量保障体系，即智能服务质量 QoAIS（Quality of AI Service）。

提供有服务质量保证的 AI 服务对于网络实现泛在普惠智能具有重要意义，通过建立合适的 QoAIS 指标体系及评估和保障机制，可以量化和评估 AI 服务的关键指标，如性能、可靠性、安全性和用户体验等，从而按需提供有质量保障的AI服务，提升网络对全社会、全行业、全生态业务的适应能力，真正赋能千行百业。可以确保 6G 网络能够按照不同智能应用场景的需求，提供高质量的 AI 服务，在实现智能随取随用的同时提高用户体验，并推动智慧工厂、智慧交通等可靠性要求高的关键行业快速发展，带动起社会效益和经济效益。

随着通信网络架构的演进，现有的各类 QoS 机制无法应用于 6G 网络，因此需要研究 QoAIS 指标体系，并进行相应的调整和优化，确保在内生 AI 网络架构中，AI 服务的质量能够得到准确的评估和有效的优化。随着通信技术的不断创新，QoAIS 指标体系也需要持续更新以适应新技术的发展，并确保充分体现服务质量。

当前，业界已广泛地开启攻关无线网络 QoAIS 相关技术，积极开展原创性、先导性的关键技术研究，努力形成具有我国自主知识产权及产业把控力的 QoAIS 技术体系，巩固和提升我国在大规模无线网络智能普惠方面的国际领先水平，促进我国 ODICT 产业可持续、安全的发展。

本报告对网络 AI 服务质量指标体系，即 QoAIS 指标体系进行研究。本报告由中国移动联合华为、亚信、大唐、vivo、上海诺基亚贝尔等企业以及北京邮电大学、重庆邮电大学等高校共同撰写。

文档作者列表

贡献者	单位
邓娟、岳烈骧、华美慧	中国移动
王君、张宽	华为技术有限公司
戴翠琴、李职杜	重庆邮电大学
杨忱逊、胡焕然	北京邮电大学
王首峰	亚信科技（中国）有限公司
孙万飞	大唐移动通信设备有限公司
周通、袁雁南	维沃移动通信有限公司
沈钢	上海诺基亚贝尔股份有限公司

1. 背景介绍

面向“智慧泛在”的未来社会发展愿景，6G 网络需助力千行百业的数智化转型，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[1]。未来，AI 将与 6G 进行更为全面的原生的结合。通过在 6G 网络架构的设计中充分考虑 AI 的算法，算力和数据以及网络连接等诸多要素，6G 将成为融合连接和算力的新型基础设施，从而极大提高 AI 资源的使用效率并使 AIaaS（AI as a Service）成为可能。

考虑到不同的智能应用场景（如网络高水平自治，用户智能普惠、用户极致业务体验、网络内生安全等）对 AI 服务质量有着不同的需求。需要一套指标体系通过量化或分级的方式表达差异化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。对此，提出 AI 服务质量，即 QoAIS（Quality of AI Service）的概念，QoAIS 反映了网络对 AI 服务质量的保障能力，即网络能够按照用户提出的指标提供相应的 AIaaS 并保障用户的体验。

6GANA 在前期发布的《6G 网络内生 AI 网络架构十问》白皮书中，从技术特征内涵、必要性分析、可行性分析、对网络架构的影响方面解释了为什么需要 QoAIS[2]。QoAIS 包括指标体系、评估体系和保障体系，本文将在《6G 网络内生 AI 网络架构十问》白皮书的基础上主要研究 QoAIS 指标体系。在指标体系设计上，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 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”、“碳达峰”政策的实施、全球智能应用行业对计算能效、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显化，成为评估服务质量的新维度，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。

2.驱动力

相比云 AI，6G 网络 AI 具有泛在普惠、移动性支持、带有用户属性、极致性能、内生安全隐私、可信等技术优势。QoAIS 指标体系是体现 6G 网络 AI 优势，解决用户智能需求、公平性与有限资源矛盾的重要技术组件。

首先，运营商需要一套指标体系帮助运营商更好地评估和优化 AI 服务质量，从而在竞争激烈的市场中脱颖而出，并促使运营商持续关注用户需求，为用户提供更优质、更便捷的 AI 服务。

(1) 广谱 AI 服务质量：基于网络 AI 和云 AI 的联合，运营商具有构建广谱 AI 服务质量的基础设施和技术优势[3]。需要一套指标体系量化谱系中差异化服务质量。

(2) 极致性能的保障：与云 AI 供应商相比，移动通信网络运营商可以提供性能更好的 AI 服务，尤其是在用户移动和空口信道质量不稳定的场景下。运营商需要一套指标体系监测和优化 AI 服务质量，确保用户在各种场景下的 AI 服务体验。

(3) 泛在普惠智能：大规模泛在网络基础设施支持智能服务的普及。运营商需要一套指标体系评估网络智能服务的泛在普惠程度，了解不同地区、时段和网络条件下智能服务质量的达标情况，从而针对性地进行网络智能服务覆盖率的优化和扩展[4]。

（4）新网络生态满意度：6G 网络原生支持各类 AI 应用，能够构建新的网络生态，并实现以新型网络使用者为中心的业务体验，运营商需要一套指标体系评估新网络生态中各类用户的业务体验满意度[5]。

其次，网络需要基于一套服务质量指标体系，解决用户需求、公平性与有限资源之间的矛盾。基于该指标体系对 AI 服务的资源分配策略进行优化，确保各类用户在有限资源下获得公平的服务。

然而，现有服务质量指标设计方案均无法直接使用。首先，通信 QoS 体系无法直接使用，不同于传统通信业务，智能服务用户具有隐私、模型性能等新的需求维度，同时引入了数据、计算和模型资源维度，需要引入新的 QoS 指标。其次，云 AI 服务指标方案无法直接用，因为云资源一般不受限，以尽力而为的模式提供服务，其性能指标一般体现为AI业务的性能，不体现网络保障能力，不进行从用户体验指标到资源QoS指标的映射[6]。第三，面向泛在普惠的服务供应，现有方案的 QoS 没有考虑网络级服务的可获得性和性价比（如能效等）度量。

随着通信技术创新，QoAIS 指标体系需要不断更新以适应新技术的发展，并确保服务质量得到充分体现。随着通信网络架构的演进，需要 QoAIS 指标体系进行相应的调整和优化。随着人工智能技术进步，人工智能服务的性能和能力不断提高。QoAIS 指标体系需要与人工智能技术的进步保持同步，以评估和优化服务质量。

3. QoAIS 定义与内涵

3.1 QoAIS 指标体系

QoAIS 指标体系是网络对 AI 服务的质量进行评估和保障所使用的一套指标体系。是“面向单次服务 QoS 保障的指标”，是将用户 SLA 转译之后的网络可理解和执行的一套指标体系。QoAIS 指标体系包含 AI 服务的 QoS、AI 任务的 QoS、AI 资源的 QoS，三层指标间具有映射关系。

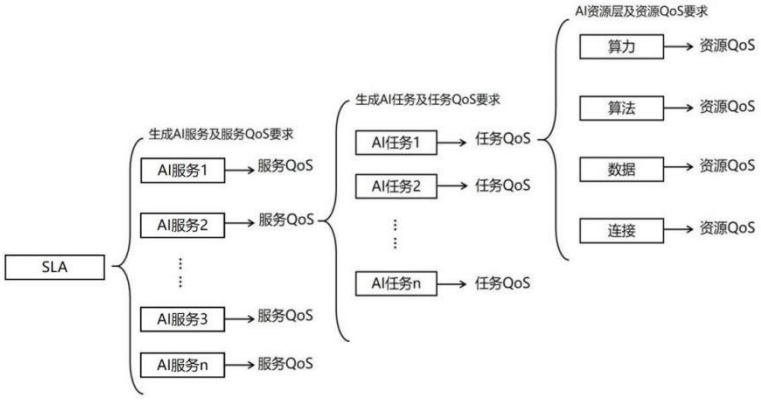


图 1 AI 服务 QoS、任务 QoS 和资源 QoS 间的逻辑关系

当用户向网络提出一次服务请求，其中可能涉及到一种或多种 AI 服务（如 AI 训练、推理、数据和验证服务）；而 AI 任务由 AI 服务分解而来，是指网络新能力涉及到多节点场景下连接、计算、数据和算法资源的协同和调配，以共同完成某个特定的智能服务目标[7]。

AIaaS 既为 To B 和 To C 用户提供所需智能，也为网络自身提供所需智能，因此 QoAIS 指标体系需体现以上场景中对网络 AI 的质量需求。由于不同层次的 QoS 指标可能

由不同的子系统实现和维护，比如服务 QoS 位于管理编排体，任务 QoS 和资源 QoS 位于控制面，不同场景中的需求由哪个层次上的 QoS 指标体现有待进一步研究。

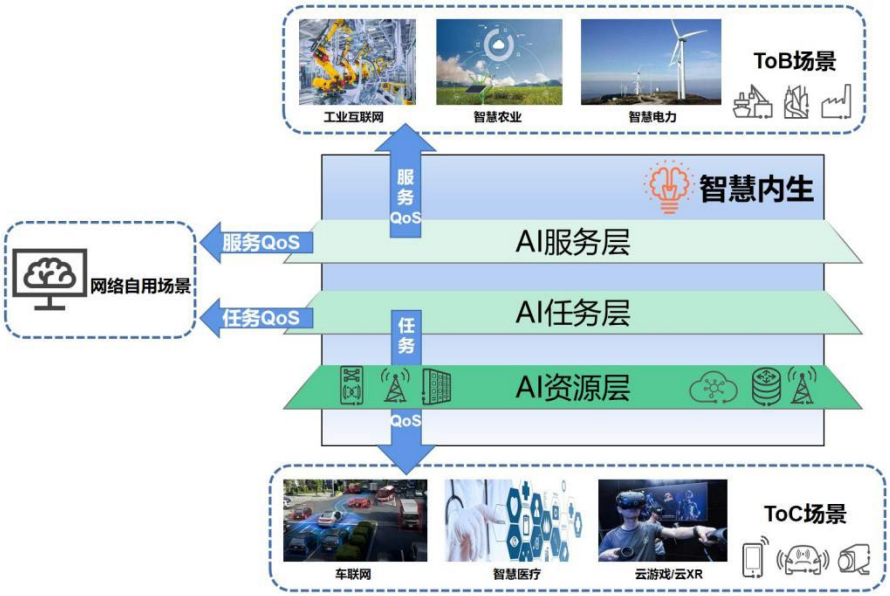


图 2 QoAIS 和 AIaaS 关系图示例

3.2 三层 QoS 定义

3.2.1 AI 服务 QoS 的定义

根据 6GANA TG1 给出的定义，“AI 服务（AI Service）”是指按需向被服务方提供 AI 技术、业务或 AI 三要素等。“AI 服务”范畴通常要大于“AI 业务”，AI 服务可涉及到 AI 三要素资源能力，AI 相关功能技术和 AI 业务等不同层面的内容。

在 IT 领域中，AI 服务没有统一定义，一般是指通过预先经过训练的 ，为用户的应用程序和工作流程提供现成的智能功能，为用户提供的各种服务。服务的形式分为平台层提供 AI 算力服务，服务层的训练，推理服务，应用层远程或者云端提供的自然语言理解 (NLU)、自动语音识别 (ASR)、视觉搜索和图像识别、文本转语音 (TTS) 等。

AI 服务 QoS 是指 AI 服务提供的性能、可用性、安全性等方面的质量指标。

3.2.2 AI 任务 QoS 的定义

根据 6GANA TG1 给出的定义，“AI 任务”是指协同计算、算法、连接和数据完成某个特定的目标，该目标来源于 AI 服务。AI 服务到任务的映射过程可以是灵活的，AI 服务

可以分解为一个或多个 AI 工作流，而 AI 工作流可进一步分解为一个或多个任务。因此 AI 任务在 AI 服务/任务/资源的三层结构中承担着承前启后的作用：首先，一个 AI 任务是上一层 AI 服务轻量化的子集；其次，AI 任务对于下一层的 AI 资源四要素的分配和执行限定了范围。

AI 任务的 QoS 是指衡量网络对一项 AI 任务编排质量和完成质量的指标集。AI 任务 QoS 重点关注两个方面：任务编排的优先级（部署阶段）和四要素保障优先级（执行阶段）。

3.2.3 AI 资源 QoS 的定义

AI 资源的定义是：保障任务实现涉及的四要素资源（计算/算法/数据/连接），其中不仅包含诸如 CPU 算力或空口时频域资源等物理资源，也包括实现的方法。

AI 资源的 QoS 是指网络对于四要素资源质量的一系列指标，其中包括独立资源内的优先级，物理资源分配和实现方法的应用情况。计算 QoS 可包括计算时延，计算可靠性，计算并行度等；算法 QoS 可包括训练收敛速度，泛化性，可解释性等；数据 QoS 可包括共线性度，F1 值，安全隐私等级等；连接 QoS 可包括优先级、时延、丢包率等。

3.3 网络 AI 服务类型

当前 IT 领域的 AI 服务包括自然语言理解功能服务、语音识别、计算机视觉、文本转换为语音，语音合成、机器学习等，多采用云资源服务供应模式。6G 网络 AI 具有泛在普惠、移动性支持、带有用户属性、极致性能、内生安全隐私、可信等优势，价值场景与云 AI 各有侧重，可从网络自治、To B 和 To C 三类场景分析所需 AI 服务。

1) 网络自治

网络智能化实践对 AI 的需求场景可分为运维智能、网络智能和网元智能三类。运维智能对 AI 能力的要求为利用机器学习的方法，对运维数据进行计算和分析，通过 AI 与运维流程相结合，实现运维效率的提升。网络网元智能对 AI 能力的要求为基于 AI 能力持续迭代优化，提升系统性能和用户体验，最终实现自动化闭环流程系统。

对 AI 服务的需求可包含业务场景识别，数据管理、模型分析、模型训练，决策运行自动化闭环。业务场景识别包含基于业务目标构建机器学习问题，根据自身数据分析或者外部导入的方法进行 AI 用例生成，网络能根据 AI 用例描述调配网络元素；数据管理包含数据采集、处理等；模型分析包含对数据的智能化分析以及对业务问题的建模；模型训练

指针对问题建模使用数据训练合适的 AI 模型；决策运行表示模型具备可信评估以及自助决策能力，模型能够自主评估优劣并且自动部署和运行。

2) To B 场景

6G 网络 AI 的典型 To B 场景包括时延保障类的智慧城市、智慧工厂、智慧医疗、网络金融，大带宽类的普智教育等。这些场景由于对服务的时延保障要求较高、或需要大带宽传输，云 AI 服务无法较好满足 6G 网络 To B 业务需求，需要 6G 网络 AI 为其提供包括模型生成服务、模型推理服务、模型共享服务、模型验证服务、模型部署服务、数据服务和计算服务等服务

模型生成服务包含了模型训练、模型优化和模型组合。模型训练是在神经网络模型初始化和训练数据收集、预处理之后，以数据驱动的方式进行神经网络参数配置。模型优化是根据模型需求，有针对性地对模型精度、泛化性、模型规模等指标进行优化。模型组合利用 6G 网络内现有的模型，自动进行 AI 算法模块组合。模型推理服务是在已有的模型基础上，通过输入相关数据，从中提取特征，并获得推理结果。模型共享服务是指在相同的场景、需求或任务下，复用已有 AI 模型的服务。模型验证服务是评估 AI 模型在新数据或网络环境上的表现。模型部署服务是将训练好的 AI 模型在云端、边缘、终端位置合理部署。数据服务包括数据获取、数据预处理、数据脱敏、数据存储、数据开放等。计算服务是指为不同算力需求的 AI 应用提供相应的算力支撑，例如分布式算力集群的算力分配与调度。

3) To C 场景

6G 网络 AI 的典型 To C 场景包括位置移动类的车联网、室外机器人协作、空中高速上网，时延保障类的无人机自治系统，大带宽类的基于全息通信的 XR 互动游戏、数字孪生人。这些场景由于具有较强移动性、对服务的时延保障要求较高、或需要大带宽传输，单纯的云 AI 限制了用户对云服务设备、AI 应用程序和 AI 数据的访问权限，降低了服务灵活度。此外，所有数据集中在云端将增加通信传输成本，还会导致较高的个人数据安全和隐私风险，因此需要 6G 网络 AI 为其提供服务。

To C 场景的应用由于面向个人消费者，智能终端的计算、存储和电量资源有限，因此 To C 场景所需 AI 服务还应考虑模型压缩服务。模型压缩就是利用网络结构的冗余特性对 AI 模型进行重构并简化，在不影响原任务完成度的情况下，得到参数更少、结构更轻量级的模型。模型压缩服务可以归为上述模型生产服务中。

综上，6G 网络 AI 提供的服务可分为模型服务、数据服务和计算服务三种类型。

4. 设计原则

为在满足多样化服务质量需求的同时，体现网络 AI 优势、提高网络执行效率、保证兼容性，降低复杂度，QoAIS 指标体系的设计需要遵从如下原则：

1) 特殊性原则：QoAIS 指标体系需体现 6G 网络 AI 的技术优势，包括泛在普惠、移动性支持、带有用户属性、极致性能、内生安全隐私、可信等。

2) 统一框架原则：所有 AI 服务和任务共同一套指标体系，不因为具体服务和任务的内容不同单独成立多套框架；资源维度指标则根据四要素本身的类型进行区分。

3) 重用性原则：为保证兼容性和合理性，6G 网络的 QoAIS 设计应尽可能保留原有评价指标。与通信领域已有指标定义接近的，参考通信领域的指标定义方式；与 IT 领域已有指标定义接近的，参考 IT 领域的指标定义方式。新定义的指标和量化方案，尽量可拆解为已有指标，从而复用已有指标定义和量化方案；

4) 非必要不增加原则：在新增指标和参数时慎重考虑其必要性，非必要不增加。

5) 前向兼容原则：6G 网络 AI 保障算法、算力、数据、连接等资源 QoS。其中连接 QoS 应尽量与原有通信网络连接 QoS 指标有机统一，降低网络复杂度。

6) MECE 原则：不同指标之间完全正交，不重不漏。

7) 安全性原则：5G QoS 机制中并不涉及安全，传输安全通过其他机制保障。6G 引入 AI 服务后，网络不再仅仅是一个不触碰内容的管道，因此安全是必须要考虑的因素。而在不同用户在不同场景下，AI 服务的安全隐私的要求是差异化、动态化的，所以把安全性放在 QoS 机制中进行保障较为合理。

8) 三层 QoS 指标间可拆分可映射：QoAIS 能够在不同资源维度上分拆并映射，以便控制、管理并协同不同资源，从而在机制上保障内生 AI 服务质量

9) QoAIS 指标与 6G QoS 指标有机统一：在 AI 服务提供和保障过程中，QoAIS 指标设置及流程机制设计，应该尽量与 6G 网络 QoS 指标和流程机制有机统一，降低网络复杂度。在 6G QoS 指标中可共用的部分，各类服务 QoS 指标项应保持一致。

5.现有方案

5.1 通信 QoS 和 SLA

通信 QoS 是指通信服务质量，是网络通信业务能力的综合体现，其设计目的主要有两方面：一、匹配业务和业务所需的能力；二、保障用户间的差异性和公平性。QoS 管理是端到端的过程，需要所有网络节点（UE<—>基站<—>核心网）共同协作，来保障服务质量。

随着移动通信网络技术的发展和网络架构的更迭，自 3G 网络提供基于金银铜分级粗粒度的 QoS 差异化保障之后，4G、5G 网络也纷纷基于自身的网络架构提出了不同的 Qos 机制。4G 网络中的 QoS 架构基于扁平化的网络结构设计提供了端到端的 QoS 保障。5G 网络的 QoS 模型是以 QoS 流为基础，支持 GBR、non- GBR 以及 Delay Critical GBR。5G 网络中使用 5QI 区分不同的 QoS 等级，其值代表标准的或者预设的 5G QoS 特性。总体而言，5G QoS 是基于会话（PDU session）管理，最小管理粒度为 QoS flow，限于篇幅，不详细展开。

本小节主要介绍 QoS 的指标设计。以 5G 为例，QoS Profile 分为两个层级——QoS parameters 和 QoS characteristics。如下图所示：

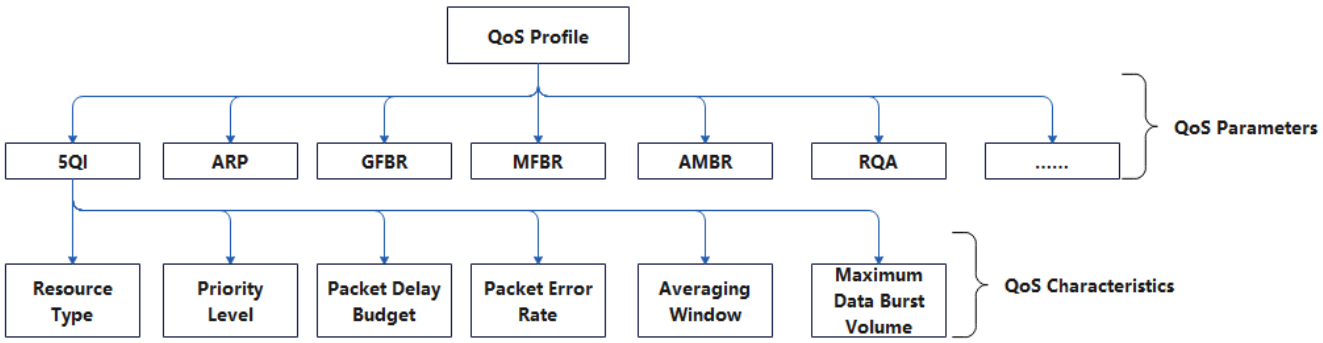


图 3 5G QoS Profile 分级图

5QI——以资源类型、优先级、时延、丢包率、最大数据突发量等指标维度，对通信业务进行了切分，不同的业务对应不同的 5QI 值，每个 5QI 值对应上述指标维度的组合；

ARP——表示分配保持优先级，指示一个 DRB 相对另一个 DRB 的资源分配和保持的优先级，ARP 取值越大，则优先级越低，在网络拥塞时容易被准入拒绝；

GFBR/MFBR——表示 QoS Flow 能够提供的保证流比特速率和最大流比特速率，类似于 LTE 的 GBR 和 MBR；

AMBR——表示聚合最大比特速率，是 Non-GBR QoS Flow 特有的 QoS 参数，该 QoS 参数限制了共享这一 AMBR 的所有 QoS Flow 所能提供的总速率；NR 的 AMBR 分为 Session-AMBR 和 UE-AMBR；

RQA——RQA 是反射 QoS 属性，用于 UE 发送上行用户面数据时，在核心网没有提供映射规则的情况下，将上行用户面数据映射到 DRB，节省了映射的信令开销[8]；

对于具有超高速率、超低延迟和大规模连接特征的 5G 网络，将其运用在不同的场景中 QoS 需求不尽相同。比如高清视频和流媒体[9]应用场景一般选择 5QI 为 6-8 的 QoS Flow，无人驾驶和工业自动化这类应用选择 5QI 为 85 的 QoS Flow，以满足其严格的服务质量需求，mMTC 关注的 5G QoS 参数主要要求为：连接密度在 10^6 设备/km2；误帧率（FER）1~10^{-5}，可靠性必须做到接近 100%；空口时延 1 毫秒(ms)，端到端时延必须达到毫秒(ms)量级。

SLA 是服务商与用户之间约定的一种双方认可的协定，该协议定义了服务商为用户提供的服务类型、服务质量以及对用户保障服务的性能和可靠性的承诺等内容。5G 重点关注 To B 业务的 SLA，其中主要包括三个方面：影响业务中断，高可靠和确定性。指标如下表所示：

表 1 5G 业务 SLA

业务 SLA	
类型	指标
影响业务中断	网络可用度（根据 TL9000 系统统计中断时长计算）
可靠性	RTT 时延可靠性、接收包间隔可靠性、OTT 时延可靠性、RTT 平均时延、满足时延要求的连接数
确定性	单终端速率、单连接速率满足度、满足速率要求的连接数、单/多终端在线率、多工业终端在线率满足度

无线接入网、传输网和核心网通过端到端协同实现网络切片来保证 SLA。RAN 侧主要的实现方式是对时频域资源进行切分，具体方法有两类，一种是通过载波隔离的方式硬切，

另一种是基于 QoS 调度（通过 SLA 指标中对于优先级、时延、丢包率等要求找到对应的 QoS Profile）和 RB 资源预留方式实现软切。传输网主要通过 FlexE 切分传输带宽，实现不同程度的硬管道。核心网侧主要通过 VNF 的切分保障 SLA。

具体 SLA 是如何和 QoS 关联的，以上述 To B 业务中 RTT 平均时延为例，SLA 中该值某一确定要求可以通过 UDM 或者 AF/NEF 传递给 PCF，PCF 生成 PCC Rule 传递给 SMF，SMF 完成 QoS Flow binding 后给 RAN 传递 QoS profile，其中该 To B 业务涉及 QoS Flow 的 5QI 值为 86，对应 package delay budget 为 5ms[10,11]。

综上可知，当前通信的 QoS 设计无法满足 AI 服务的需求，原因主要有：

- 资源的多维性： 6G 新增 AI 服务和原有的通信服务相比引入了新的资源维度，包括算力资源和存储资源；同时 AI 服务涉及到四要素（连接，计算，数据，算法）协同无法在原有的 QoS 指标体系中得到合理的评估；
- 机制的分歧性：对于 6G 的指标体系而言，已不再适合沿用 5G 基于会话的生命周期管理机制，需要基于完整的任务生命周期的管理机制。因此，对应的 QoS 机制和 QoS 指标设计已经无法沿用 5G 的 QoS 架构。
- 安全性需求增长：随着“碳中和”“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显化，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。

5.2 云 AI 服务指标体系

当前云 AI 提供不同的服务类型（IaaS，PaaS，SaaS 等），对应不同的服务指标体系，以及不同的计费模式。

IaaS（基础设施即服务）：用户使用云服务提供商的基础设施部署运行各种软件，但无权访问和管理底层基础设施。在这种服务模式下，服务的计费方式主要考虑的是物理资源的获取，比如 CPU/GPU 的类型，带宽，使用个数，使用时长，资源线性加速比=多卡全局吞吐/（单卡吞吐*卡数），计算资源（GPU）利用率=计算时间/（计算时间+等待时间）等。云服务提供商的计费量纲通常为$/CPU、$/带宽等。

PaaS（平台即服务）：用户使用云服务提供商支持的编程语言，库以及开发工具等来开发应用程序并部署在相关的基础设施上。在云 AI 服务中，PaaS 逐渐被 FaaS（功能即服务）取代，和 PaaS 不同的是，FaaS 下应用程序是由大量的 functions 组装而成，因此具有无限的自动拓展能力以及比 service 更细的粒度。相关性能指标包括：

平台数据存储读性能=读取随机打散后 batch 数据时间总和的高低。（为了算法达到比较好的鲁棒性、加快收敛速度，训练数据集需要随机打散）

平台写数据性能=保存训练结果时间的高低。（在模型训练的过程中，即使模型训练中断，也可以基于 checkpoint 接续训练和故障恢复，因此需要不断地保存训练结果（包括 epoch、模型权重、优化器状态、调度器状态））

平台训练效率=训练时间/(训练时间+数据集随机打散时间+读取时间+保存检查点时间)。

平台调度性能=训练串行等待时间高低（在一个完整的训练流程里，除了计算的部分，还有很多其它的操作和步骤，训练之间也可能存在串行、并行等关系）

平台 GPU 卡之间通信性能等。

FaaS 下的计费是以每个函数实例运行的时间为基准的，常用的计费量纲是（$/s）。

SaaS（软件即服务）：用户直接使用云服务提供商提供的应用程序，可以通过客户端接口轻量定制化，整体来说服务具有很强的限定性。相关评估指标包括：模型，环境部署过程难易程度（是否托管的 Jupyter 环境内置主流算法框架和软件库，可以无需配置环境，点开即用，训练模型一键发布，进行推理部署等），支持的 AI 计算框架（TensorFlow、PyTorch、Kaldi、MXNet、飞桨等），提供的训练数据集丰富程度（MNIST，ImageNet，OpenImage，LFW，MegaFace，Free Spoken Digit 等）。容灾自愈、负载均衡，7x24 小时运营运维指标。服务推理请求每分钟请求数，最高并发数，请求成功率，请求返回时间延迟等。这种服务模式下，通常的计费模式是按照服务的调用次数进行计费，常用的计费量纲为图像分析（$/次）、时频分析（$/次）、chatGPT（$/token）等。

通过对上述服务类型的计费方式的分析，可知云 AI 服务指标设计的三个层次，分别是：资源层，直接的物理资源或逻辑资源，包括算力，存储和连接；服务层，是直接物理

资源的二次抽象，将物理资源能力以服务的方式暴露（如 FaaS 中即为函数的实例运行时间）； 应用层，基于单个服务或多种服务的组合，提供客户最终的应用或者产品包。

云 AI 重点关注 To C 业务的 SLA，其中主要包括两个方面：可用性，服务响应成功率。指标如下表所示：

表 2 云 AI 业务 SLA

业务 SLA	
类型	指标
可用性	服务可用率 = （服务周期总时间 - 服务周期未达标时间）/ 该服务周期总时间 * 100% 其中，AI 开发平台服务：需要选择资源套餐运行的服务，如单个 Notebook 实例、单个训练任务、单个在线服务等。 AI 开发平台服务未达标：设备故障、设计缺陷或操作不当，发生的单个 AI 开发平台服务连续超过五分钟无法访问和使用。 服务未达标时间：指一个服务周期内单个服务未达标的时间的总和，按分钟计算。
服务响应成功率	响应成功率=1 -（服务月度内的服务不可用次数/服务月度内的发送请求数）x 100%。 其中，服务不可用的定义为每一次调用发起 API 接口调用且尝试失败，则视为该次调用该服务不可用。

可用性保证代表云 AI 服务提供商保证其服务的可用性百分比。例如，一个 SLA 可能保证 99.9%的可用性，这意味着服务每个月不应超过 43 分钟的停机时间。服务响应成功率是云 AI 服务提供商保证服务请求在客户端发送后，服务端能够成功响应的比率，能够反映出服务的可靠性和稳定性。总的来说，SLA 是云服务提供商和客户之间的协议，它规定了服务提供商向客户提供的服务承诺，帮助客户了解服务的质量和预期结果。

云 AI 的实现本身是 Best Effort 模式的，并没有闭环的 QoS 保障机制，依赖于用户本身的感知和参与。云 AI 的分层设计和 SLA 机制是可以为 QoAIS 借鉴的。

5.3 网络智能化测评指标

5.3.1 网络网元智能化测评指标

网元智能化是在 5G 系统/5G 演进系统的网元（如基站、核心网网元）功能基础上，引入数据、算法、算力等 AI 能力，提高网络效率及性能。

无线网元智能化的评估对象为用于模型训练和模型推理的基站设备、无线 OMC，也可能包括其它通用设备及组件。核心网网元智能化的评估对象包括独立的智能网元（如 NWDAF）、内置 AI 模块的智能网元，以及这些网元设备的相关组件。

评测指标上，可从 AI 系统和通信系统两个维度拟定测评指标。从 AI 系统角度，从 AI 架构的可扩展性、服务化能力、算法库能力、数据安全和自动化采集能力、模型动态更新及分发能力等方面评估。从通信系统角度，评估该通信系统对智能化应用的支持数量、性能表现、算力等资源消耗，以及人工对该智能化应用的干预程度等。

鉴于以上评估指标针对不同的智能化应用，其表征方法均不相同，因而与智能化应用用例强相关，称为场景化智能化评估指标。如，无线及核心网场景化智能化评估指标主要包括以下方面：

- “量”：网元支持的智能化应用数量；
- “性”：网元智能化应用的性能表现；
- “算”：网元智能化应用的资源及算力消耗情况；
- “分级/自治度”：人工对该网元智能化应用的干预程度

下面以无线基站负荷预测为例，对相关智能化指标进行评估。负荷预测任务为典型的回归任务，因此从以下几个角度评价负荷预测智能化水平：

1. 数量

智能化应用支持数量评分值$\mathrm{F}_{\mathrm{num}}$表示如下，其中$\mathrm{Num}_{\mathrm{ref}}$是测试方已定义的智能化应用全集的数量，$\mathrm{Num}$ 是被测方实际被验证的支持的智能化应用的数量。

$$\mathrm{F}_{\mathrm{num}} = 5 \cdot (\mathrm{Num}/\mathrm{Num}_{\mathrm{ref}})$$

2. 性能

性能指标中，可以直接测量平均绝对百分比误差大小，泛化性和鲁棒性需要新的外部数据来获得。

平均绝对百分比误差：设平均绝对百分比误差的敏感系数为$\alpha_{\mathrm{MAPE}}=0.5$，平均绝对百分比误差的目标值为$\mathrm{MAPE}_{\mathrm{ref}}=0.05$，实测平均绝对百分比误差$\mathrm{MAPE}=0.14$， 则平均绝对百分比误差根据以下公式计算：

$$\begin{aligned}\mathrm{F}_{\mathrm{MAPE}}&=\min\{5\cdot\exp(\alpha_{\mathrm{MAPE}}(1-\mathrm{MAPE}/\mathrm{MAPE}_{\mathrm{ref}}),5)\}\\&=\min\{5\cdot\exp(0.5\cdot(1-0.14/0.05),5)\}=2.03\end{aligned}$$

即$\mathrm{F}_{\mathrm{Performance}}=2.03$。

3. 开销

在模型训练及推理过程中，对被测基站的小区数、小区峰值、RRC 连接用户数等规格指标进行同步测试，根据指标规格与传统设备规格的差距评分。

4. 分级

分级得分复用自智网络分级标准的得分，根据分级标准评定后，形成[0,5]的分级分值。根据分级评估方法得到，$\mathrm{F}_{\mathrm{Auto}}=1.84$。

5.3.2 运维智能测评指标

由于自智网络分级标准无法完全契合对基站、核心网等网元智能化能力的评估和评测，2023 年 3 月 ITU-T SG13 正式立项关于自智网络成效评估指标的标准项目 ITU-T Y.IMT-2020-MEVE-req-frame：Future networks including IMT-2020: requirements and framework for measurement of effectiveness and value evaluation of autonomous networks。该标准项目旨在规范定义自智网络的价值度量指标体系和计算方法。

自智网络的价值度量指标体系主要为了带动各网络、服务领域能力提升，根据各领域的应用特点，制定应用效能指标，体现自动化、智能化的应用效能。ITU-T M.3385 标准给出了自治网络信任（TiAN）的测评方法，表 3 为标准中的测评指标，其中委托方指有权授权网络和相关实体进行自主管理的一方，受托方指具有自主能力的网络或网络相关实体，其可以被授权在极少甚至没有人为干预的情况下管理自身。

表 3 自治网络信任的测评指标

评估指标	子指标	描述
准确性	复现性	执行后与上次相同的复现结果占所有复现结果之比
	精确性	受托方在执行过程/步骤期间产生精确结果的交互与所有交互之比
	时效性	受托方在TiAN评估的特定期限内产生的动作数与总动作数之比
	有效性	受托方的有效输出与所有输出之比
	资源	数据、知识或相关输入的合规资源与所有资源之比
稳定性	中断	在整个TiAN评估过程中中断交互的时间长度与总时间之比
	事故	整个TiAN评估过程中的事故数与所有动作数之比
	成熟度	交互中的实际成熟度水平与TiAN评估中的最高成熟度水平之比
	可变性	TiAN评估中受托方的自我变化交互与总交互数之比
可控性	可预测性	委托方可以预测或在委托方预期范围内的受托方的决定/行动/反应/反馈的百分比
	监督	委托方可以受托方进行监督的时间与总时间之比
	接管	委托方可以接管受托方的处理步骤数与所有步骤数之比
恢复性	备份	流程中备份点的加权分数
	回退	受托方可以在必要时成功回退到备份点的次数与总回退次数之比
	重置	受托方可以在必要时成功重置为原始状态
可解释性	透明度	受托方在数据处理、算法等方面的透明步骤占所有步骤之比
	可翻译性	受托方在数据处理、算法等方面可以翻译成某种语言（包括机器语言或人类语言）的步骤占所有步骤之比
	可理解性	受托方在数据处理、算法等方面的可理解步骤占所有步骤之比
	解释准确性	受托方准确的解释占所有解释之比
	解释完整性	解释的完整性，建议解释涵盖任务，包括但不限于执行、意识、分析、决策和意图处理。
	解释再现性	受托方的可复现解释占所有解释之比
适应性	灵活性	受托方可以在不影响受托方满足相关和特定要求的能力的情况下更改的处理
	调整	更改后决策/行动/反应/反馈仍能满足相关场景或用例要求的处理

5.4 算力网络指标体系

6G 通信系统呈现出通信、感知、计算融合的趋势，算力则是通感算融合的基石[12]。算力网络是一种以算为中心、网为基础的新型计算信息基础设施，利用网络感知业务需求与算力资源状态，进行算网的统一编排调度，以实现业务需求与算网资源的匹配。

应构建统一的算力网络评估指标体系，从多维度全面选取评估因素，用以衡量算力网络的资源、需求、性能。依据参考文献[13-15]进行总结和拓展，表 4 给出了算力网络在不同维度下的相关指标及定义。

表 4 算力网络相关指标及定义

维度		指标	定义	单位
资源 QoS	算力	算力	双精度浮点数操作数（FLOPS）	
		算力节点服务能力	采用算力分级 + 能力隶属度的方式评估，由｛算力分级、算力类型｝表示	
	连接	调度控制能力	算力和网络的衔接能力	
		带宽	单位时间内能够发送/接收的最大数据量	MB/s
		丢包率	传输时丢失数据包所占比例	
	数据	存储容量	可存储的二进制数据量	TB/GB
		存储带宽	单位时间可存取的最大数据量	MB/s
		IOPS	单位时间读写次数	
服务 QoS		时延	用户从请求服务到获得服务的总时延	ms
		调度有效期	业务所能承受的最大调度时长	ms/us
用户 SLA		业务算力需求	由｛算力大小、算力类型｝表示	
		安全性	业务的安全性要求（划分不同等级）	
		可用性	业务的可用性要求（划分不同等级）	

算力节点服务能力：算力可分为通用算力（CPU）、并行算力（GPU）、智能算力（TPU/NPU）和定制化算力（FPGA），不同算力擅长不同的业务，为全面评估节点服务能力，采用算力分级＋能力隶属度的评估方式。首先将算力节点分级，然后采用多属性群决策算法或者人工智能算法获得处理能力隶属度。最后，算力节点服务能力用｛算力分级、算力类型｝表示。

调度控制能力：算网编排及算力调度的能力，可从算网能够编排、调度算力节点的数量、地理范围大小衡量，如跨区域调度、多个云之间调度等。从效率的角度来看，也可从编排、调度消耗的时间考虑。

调度有效期：进行业务需求与算网资源的匹配以及调度资源过程的最大时间，如果超出调度有效期则认为当前算网状态无法满足业务需求。

安全性：安全性指标主要从算力隔离、数据加密、操作审计三方面考虑。算力隔离是指业务需要专用的算力资源，与其他业务进行隔离，包括逻辑隔离和物理隔离。

可用性：指业务的连续性要求，可分为三类：故障时仍可用、故障后可恢复、数据备份。更具体地可考虑恢复时间目标（RTO）和恢复数据目标（RPO）两个指标。RTO 指业务因故障中断到业务恢复运营的时间要求；RPO 指发生故障后数据必须恢复的时间点要求，相当于能够承受的最大数据丢失量。

QoAIS 包含 AI 服务的 QoS、AI 任务的 QoS、AI 资源的 QoS 三个层次，算力网络指标体系在用户 SLA、算力、连接等维度上能够与 QoAIS 指标体系相对应，对 QoAIS 指标体系的建立具备重要参考和指导意义。

6. QoAIS 指标设计方案

6.1 映射模型

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服务 QoS 进行分解，映射到任务 QoS，再映射到数据、算法、算力、连接的资源 QoS 要求上。

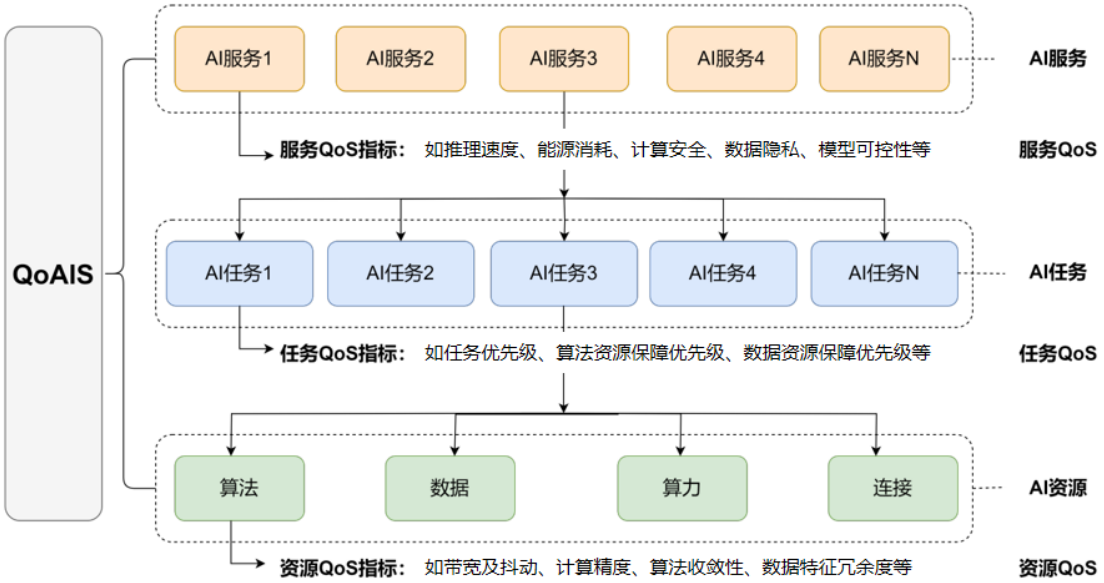


图 4 QoAIS 指标分解到各资源维度上的 QoS 指标

6.2 指标设计

当前，业界仍处于对 QoAIS 指标体系研究的发散期，映射模型和具体指标设计均尚未达成共识。因此，本报告初步收集了来自参与单位的五种典型的指标设计方案。可以看出不同方案从不同的设计角度出发，有共性的部分，也有反映对“AI 服务质量”概念理解的独特之处，对于后续深入研究讨论并形成共识具有较高的技术参考价值。

6.2.1 区分 AI 服务 QoS 的多层 QoS 设计方案

6.2.1.1 指标描述

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 5 提供了一种 QoAIS 指标体系的设计方案，按 AI 训练服务、AI 推理服务、AI 数据服务、AI 验证服务四种 AI 服务类型将 QoAIS 指标分别分类。

表 5 QoAIS 指标体系设计方案

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
AI 推理	性能	推理速度、并发度、推理误差、推理精度等
	自治	模型可控性，如推理任务自动调度、推理资源自动分配等
AI 数据	性能	数据质量、数据时效性、数据可用性、数据准确性等
	自治	数据归属权、数据可控性，如数据采集任务自动调度、数据存储自动分配、数据清洗和标注自动化
AI 验证	性能	验证速度、验证精度、可靠性、鲁棒性等
	自治	验证可控性、验证自监测能力、自纠错能力、可复现性等

注*：不同类型 AI 服务间的共同评估指标

其中，以 AI 训练为例，“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）。

根据上一节图 4，QoAIS 映射到各资源维度上的 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 6 所示。

表 6 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	

QoAIS 在性能、开销、安全、隐私、自治五个评价维度都有相应的指标，特别是在性能层面，数据、算法、算力、连接四大资源维度下均有适合量化评估的服务质量指标映射。因此，下面将在多小区天线波束联合赋形的场景用例下，对内生 AI 网络架构中基于 QoAIS 的 AI 工作流编排方案（包括集中式、分布式和协同式）展开分析，选取多个典型的可量化 QoAIS 指标进行介绍。

6.2.1.2 指标计算示例

多小区天线波束联合赋形旨在解决在多个基站覆盖的区域中出现人群聚集场景时，基站天线的波束权值动态调整方案。目前常用的解决方案是通过使用历史数据，学习人群运动的规律，指导基站天线进行决策。但是由于人群聚集属于突发事件，不出现在历史的轨迹数据中，机器学习较难准确地预知人群的分布位置，因而做出的基站天线决策有偏差。利用 6G 智慧内生 AI 架构可以解决这一问题。内生 AI 架构能够综合利用本地边缘节点和全局中心节点的优势，可以在遇到人群热点时，通过本地的模型预测和全局的推理计算，对人群分布作出准确的预测，及时指导基站天线作出决策。

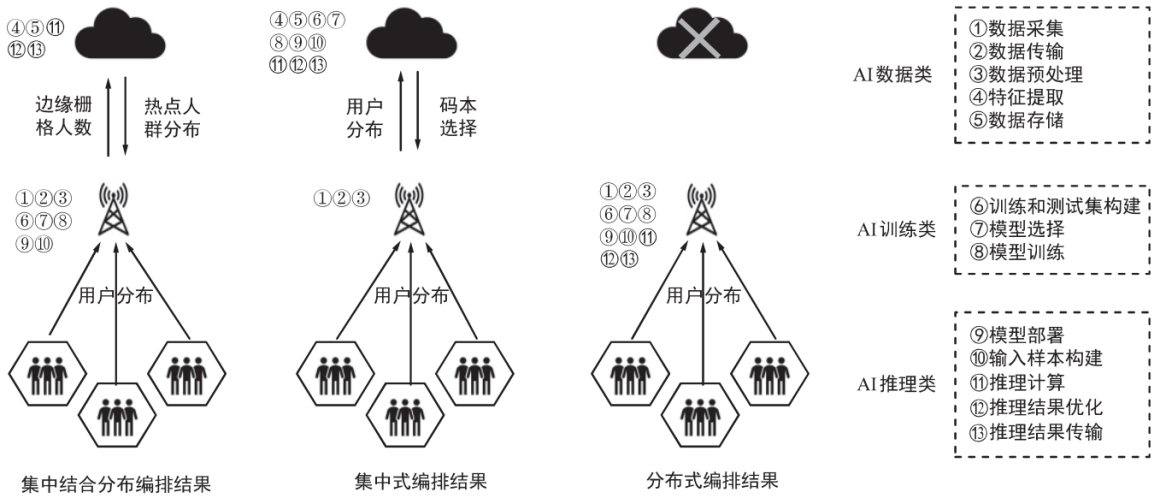


图 5 三种不同编排方案的 AI 任务分布图

图 5 展示了集中式、分布式、协同式三种 AI 工作流编排方案下，上述 AI 任务在集中式云脑和多个分布式边脑中的分布情况。该用例利用 6G 智慧内生 AI 架构解决在多个基站覆盖的区域中出现人群聚集场景时，基站天线的波束权值动态调整方案。因此，设置三种不同的场景，场景一、场景二和场景三分别含有 10%、50%和 100%的动态用户。对于每一种场景，分别使用集中式、分布式和协同式三种方案进行波束选择，并分析三种方案对 QoAIS 相关指标的满足程度。

本场景用例所需调用的 AI 服务包括数据类、训练类和推理类。网络资源部署方式是集中式云脑和多个分布式边脑(基站)。其中 AI 数据类服务的工作流包括数据采集、数据传输、数据预处理、特征提取、数据存储；AI 训练类服务的工作流包括训练和测试数据集构建、模型选择、模型训练；AI 推理的工作流包括模型部署(含模型优化)，输入样本构建，推理计算、推理结果优化、推理结果传输等 AI 任务。

指标计算示例：

性能指标界：在该场景用例下，以用户分布预测准确度作为衡量性能指标界的关键因素。将区域栅格化，用户分布预测准确度相当于利用三种编排方案预测结果与真实分布结果数值相同的栅格数与真实分布中的栅格数之比。

三种编排结果的性能指标界如表 7 所示，协同式方案的预测准确度上界最高，分布式方案的下界最低。

表 7 性能指标界对比

性能指标界	用户预测分布准确度/%		
	集中式方案	分布式方案	协同式方案
上界	99.75	97.50	99.77
下界	98.47	96.40	97.59

优化目标匹配度：一种量化优化目标匹配度的方式是计算模型训练过程中损失函数的参数变量对 AI 用例优化目标指标的覆盖程度。比如，在本用例场景下，若优化目标指标包括 RSRP 覆盖性能和 SINR 覆盖性能，而损失函数的设计仅包含 RSRP，则并非完全匹配。具体的计算公式可设计如下：

$$\theta = \alpha \times f(\mathrm{RSRP}) + \beta \times f(\mathrm{SINR})$$

其中，α代表 RSRP 的权值，β代表 SINR 的权值，（α+β=1），θ代表优化目标匹配度。f（RSRP）与 f（SINR）作为损失函数是否包含相应优化指标的 0 - 1 函数（包含则为 1，否则为 0），在本用例中，三种编排结果的优化目标都是相同的，故而三种编排结果的匹配度相等。

鲁棒性：鲁棒性用来衡量方案结果的抗干扰性，针对该场景用例，以用户分布预测模型准确度的方差表示方案的鲁棒性程度，预测结果的方差越小，鲁棒性越高。三种编排结果的鲁棒性分别为：集中式方案 31.58×10 - 7，分布式方案 25.14×10 - 7，协同式方案 1.082×10 - 7。集中式方案的鲁棒性差于分布式方案，协同式方案的鲁棒性最好。

传输、存储、算力开销：对于该场景用例中的 QoAIS 开销类指标，从数据传输、存储和算力开销三方面对比。

从表 8 可以看出，集中式方案在传输数据上传量和训练数据的存储量上需要较大的数据开销，同时需要较多的算力资源。分布式方案因为数据决策都在本地执行，所以没有传输数据量，数据的存储和算力消耗也比较少。协同式方案由于云边之间的反馈，所以传输数据下发量需要的开销较大，其余指标与分布式方案相同。

表 8 开销对比

指标列项	开销

		集中式方案	分布式方案	协同式方案
传输数据量	上传数据量	3330kB	0kB	463kB
	下发数据量	1.2kB	0kB	1924kB
存储数据量	训练数据	12.3GB	1.99GB	1.99kB
	推理数据	9990kB	9990kB	9990kB
	模型大小	1kB	17kB	17kB
算力消耗 (单位：MFLOPS)		1150	268.6	268.6

6.2.2 统一 AI 服务 QoS 的层次化 QoS 设计方案

6.2.2.1 指标描述

QoAIS 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，反应了用户层面对于 AI 服务质量的需求，并将需求量化，导入 6G 内生智能网络，使网络能够基于用户提出的指标提供相应的服务质量保障。AI 服务的性能同时取决于系统底层所提供的 AI 模型相关能力、用户和网络侧的计算能力、用户与网络连接的通信能力、以及数据的质量等因素。基于此，可构建以 AI 服务为中心的层次化性能指标体系架构，如下图所示。

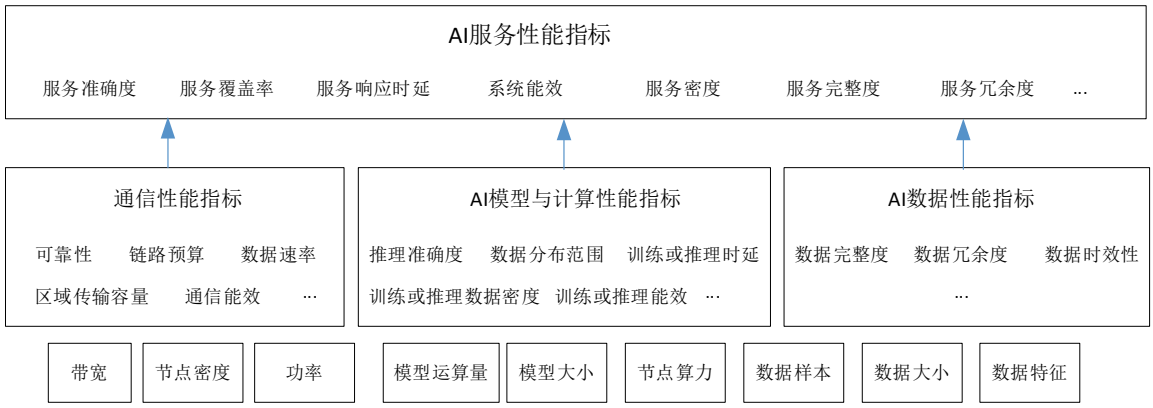


图 6 以 AI 服务为中心的性能指标体系架构

作为上层指标，AI 服务性能指标可根据通信性能指标、AI 模型与计算性能指标以及 AI 数据性能指标等下层指标联合确定，而下层指标则受带宽、节点密度、功率等实际系统参数影响。下表列出了部分 AI 服务性能指标的定义及计算参考公式。

表 9 AI 服务性能指标

性能指标名称	定义	计算参考公式
服务准确度	用户体验的 AI 服务准确度（%）	推理准确度*链路可靠度
服务覆盖率	给定链路预算下，满足服务准确度的区域面积占比（%）	满足服务准确度的区域面积/总服务面积
服务响应时延	端到端 AI 服务响应时延（ms）	终端计算时延+端边链路时延+边缘计算时延+云边链路时延+云计算时延
系统能效	单位能量消耗所能处理的数据量（bit/J）	任务大小/（终端计算能耗+端边链路能耗+边缘计算能耗+云边链路能耗+云计算能耗）
服务密度	单位区域内的服务速率（bps/平方公里）	min{传输容量，训练或推理数据密度}/区域面积
服务完整度	数据服务的完整度（%）	数据的完整度*子任务完成比例
服务冗余度	数据服务的冗余度（%）	数据的冗余度*通信冗余度*计算冗余度

在 6G QoAIS 指标体系中，通信性能指标与 5G 类似，其定义和参考计算公式如下表所示：

表 10 AI 服务性能指标

通信性能指标名称	定义	计算参考公式
可靠性	传输链路的可靠度（%）	该链路传输成功次数/总的传输次数

链路预算	达到特定数据速率时基站与终端之间的最大耦合损耗（dB）	信道衰落+噪声+干扰
数据速率	用户体验的数据速率(特别是上行)（bps）	成功传输的数据量/传输时间
区域传输容量	单位区域内的数据吞吐量（bps/平方公里）	区域内成功传输的数据量/传输时间/区域面积
通信能效	单位能量消耗所能传输的数据量（bit/J）	成功传输的数据量/通信能耗

AI 模型/计算性能指标包含 AI 模型性能指标及算力指标，其定义和参考计算公式如下表所示：

表 11 AI 模型/计算性能指标

性能指标名称	定义	计算参考公式
推理准确度	模型推理结果的准确度（%）	推理数据集正确样本数量/推理数据集总样本数量
数据分布覆盖范围	达到模型推理准确度的数据分布所能覆盖的范围（主体数或节点数）	min{节点数：推理准确度(节点数)>准确度阈值}
训练或推理时延	模型训练达到目标准确度的时间（ms）	计算公式：min{时间：训练准确度(节点数)>准确度阈值}
训练或推理数据密度	单位区域单位时间内节点能处理的比特数（bps/平方公里）	节点计算量/计算时间/区域面积
训练或推理能效	单位能量消耗所能支持的训练或推理的浮点运算次数（FLOPs/J）	任务训练和推理所需浮点运算数/计算能耗

AI 数据性能指标刻画 AI 服务的数据质量，其定义和参考计算公式如下表所示：

表 12 AI 数据性能指标

AI 数据性能指标	定义	计算参考公式
数据完整度	数据集的无损数据占比（%）	数据集可用样本量/数据集样本总量
数据冗余度	数据集的冗余数据占比（%）	数据集冗余样本量/数据集样本总量
数据时效性	信息年龄(s 或 ms)	当前时刻-模型最新接收数据的产生时刻

6.2.2.2 指标计算示例

6G 网络可基于云边智能协同的方式为用户提供知识和内容服务，以 AR/VR 视频业务为例，云平台基于全网用户数据训练获得兴趣偏好预测模型、轨迹预测模型以及网络状态模型，并通过云边协同的方式在边缘网络侧部署轻量化预测模型，赋能边缘网络预测短期 AR/VR 视频业务请求，从而提前完成热点视频内容和渲染服务的缓存，同时通过边缘协同方式提前规划边缘网络资源，应对业务请求突发情况。与此同时，边缘侧模型可通过联邦学习等方法实现数据不出本地情况下的模型微调优化，不断提升用户服务的准确性和用户体验。

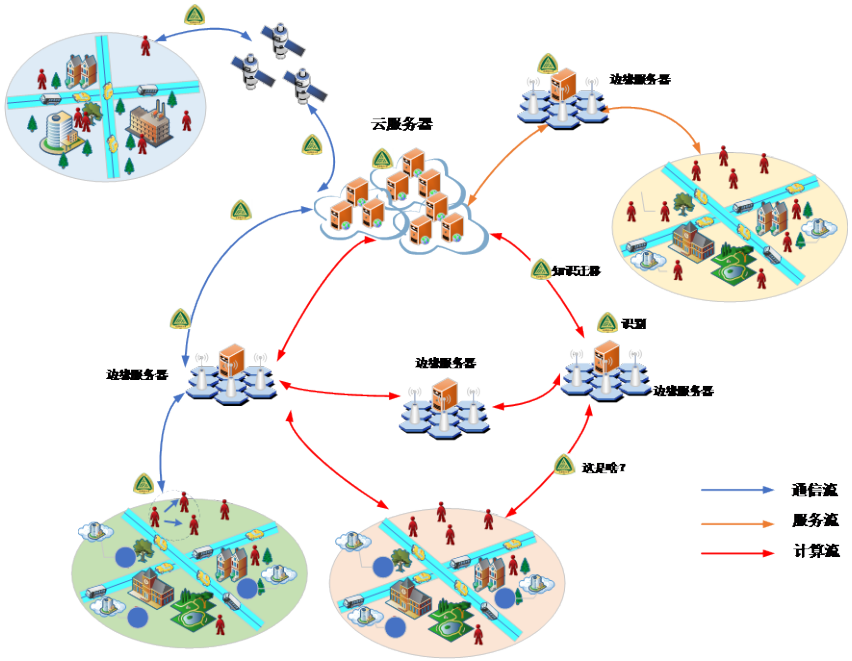


图 7 云边智能协同知识和内容服务场景

本场景用例所需调用的 AI 服务包括数据类、训练类和推理类。其中 AI 数据类服务的工作流包括用户位置、请求、观看、评价 AR/VR 视频内容的数据采集、传输、预处理、提取、存储等；AI 训练类服务的工作流包括云平台对兴趣偏好预测模型、轨迹预测模型以及网络状态模型等预测模型的训练，边缘网络轻量化模型的联邦学习训练过程等；AI 推理类服务的工作流包括边缘网络侧内容缓存和网络资源优化等。

指标计算示例：

(1) 推理时延

以 AR/VR 视频渲染为例，其渲染过程为 AI 推理类服务。假设边缘网络侧基于用户请求预测模型获得的缓存推理决策准确率为 p^I，原始视频任务的大小为 l^{2D} 比特，通过 AI 模型渲染 1 比特内容需要 α 的模型计算量，渲染后的 AR/VR 视频大小为 l^{3D} 比特，边缘服务器计算能力为 f^e，本地设备计算能力为 f^l，下行链路·速率为 r^l，假设原始视频任务可通过分割由边缘侧和终端协同渲染，其在边缘服务器侧处理的任务比例为 β。若边缘缓存没有命中用户请求，则从云平台侧调取请求内容，云边链路时延假设固定为 t^c。

基于上述场景，可计算 AR/VR 服务平均响应时延为：

$$t = \min\{p^I(\frac{(1-\beta)l^{2D}}{r^l}+\frac{(1-\beta)l^{2D}\alpha}{f^l})+(1-p^I)(\frac{(1-\beta)l^{2D}}{r^l}+\frac{(1-\beta)l^{2D}\alpha}{f^l}+t^c),$$
$$p^I(\frac{\beta l^{3D}}{r^l}+\frac{\beta l^{2D}\alpha}{f^e})+(1-p^I)(\frac{\beta l^{3D}}{r^l}+\frac{\beta l^{2D}\alpha}{f^e}+t^c)\}$$
$$= \min\{(1-\beta)l^{2D}(\frac{1}{r^l}+\frac{\alpha}{f^l}),\beta(\frac{l^{3D}}{r^l}+\frac{l^{2D}\alpha}{f^e})\}+(1-p^I)t^c$$

(2) 推理准确度：

以视频安全监控为例，假设 N 个智能摄像头均能对同一场景进行监控，并通过设备部署的 AI 模型进行事件推理，对于摄像头 i，其模型推理准确度为 p_i，链路可靠性为 l_i，因此该设备推理服务准确度为 $p_i l_i$，则该推理类服务总的准确度为：

$$Acc = 1-\prod_{i=1}^{N}(1-p_i l_i)$$

6.2.3 区分 AI 服务的资源 QoS 设计方案

6.2.3.1 指标描述

前文尽可能全面的深入介绍了 QoAIS 所涉及的指标，本节将从另一个最低限度的角度分析，6G 提供 AIaaS 服务时，网络在连接维度、资源维度及模型维度，需要最低限度提供的 QoS 保障指标，该指标体系概述如表 13 所示。这个角度将以最低限度定义 QoS 指标数量，在十分必须加入另外的指标时，再逐个加入扩充指标。这部分工作与前文的工作，将共同推动 QoAIS 进一步的研究。

网络提供特定 AI 服务过程中，该指标体系的模型维度将表征模型的性能保障需求；资源维度将表征网络资源保障需求；连接维度将表征服务使用过程中相关数据的传输保障需求。其中资源维度的资源类型，将包含数据、算法、计算、存储等资源，可以进一步分类型设计表征方式。

表 13 精简式 6G QoAIS 指标

维度	指标
模型维度	模型准确率、模型复杂度、模型推理开销、模型并发性能、模型泛化性能
资源维度	资源数量（注:包括资源类型）、资源时长
传输维度	传输带宽、传输时延、传输抖动、传输可靠性、传输优先级

该指标体系希望统一的表征 AIaaS 所有服务的 QoS 指标需求，并尽可能与 6G 网络其它服务，如算力服务、数据服务等，做到指标维度和参数的有机融合，将 6G 网络 QoS 保障指标体系和保障机制尽可能统一，提高 6G 网络服务效率，降低 6G 网络管理复杂度。

在目前的 AIaaS 的研究中，较为认可的服务包括模型训练服务、模型推理服务、模型部署服务等，对于特定服务的 QoS 指标参数设计如下。

- 模型训练服务：

  模型维度：模型准确率、泛化性、模型复杂度[可选]、模型推理开销[可选]

  资源维度：资源数量、资源时长

[可选]传输保障维度：传输可靠性、传输时延、传输带宽

- 模型推理服务：

  模型维度：并发性能（数量、密度、可靠性）

  资源维度：资源数量，资源时长

  传输维度：传输时延、传输可靠性、传输抖动[可选]

- 模型部署/开放服务：

  模型维度：并发性能、开放范围

  资源维度：资源数量

  [可选]传输保障维度：传输可靠性、传输时延、传输带宽

- 数据管理服务

  模型维度：开放范围，可靠性

  资源维度：资源数量，并发性能

6.2.3.2 指标计算示例

6G 星地融合网络具备全域覆盖的能力，网络的无线资源包含了地面的基站无线资源和天上的卫星无线资源。当区域内同时存在上述两种资源时，面对区域内用户的网络服务需求，6G 星地融合网络需要合理选择和分配无线资源来服务用户，实现最大化满足服务需求和无线资源利用。但是，卫星的高动态性，基站资源的不均衡性，以及用户需求的未知变化性，对 6G 星地融合系统的无线资源统一调度是一项巨大的挑战。而基于强化学习进行的卫星和基站资源联合分配，是解决该难题的一种方法。

6G 星地融合系统无线资源示例如图 8 所示。

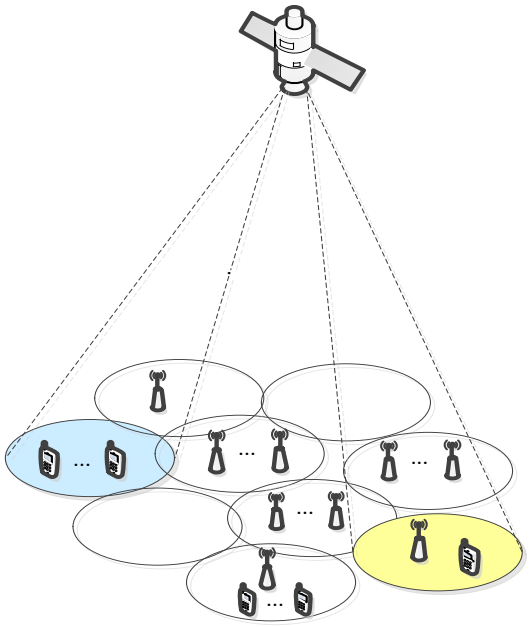


图 8 星地融合系统无线资源分布示例

在上述资源分配场景基于强化学习的方案中，首先需要统计波位范围用户需求资源表征列表、波位内基站无线资源利用情况、卫星无线资源利用情况作为 Space 空间；而后，智能体产生周期内资源分配策略，其中，包括卫星跳波束策略和系统的无线资源分配策略的 Action 动作为；之后，根据系统吞吐量和用户需求满足函数的 Reward 奖励，不断迭代优化。最终实现卫星覆盖区域内，不同用户服务需求数量和基站/卫星负载变化下，系统资源合理分配，达到系统资源高效使用和用户需求最大化满足的目标。

本场景下，用到的 AIaaS 服务包括模型训练服务和模型推理服务。在星地融合资源联合分配场景下，模型训练过程在本地训练，因此指标只包含模型维度和资源维度，模型推理服务包含指标体系中的模型、资源和传输三个维度。具体指标定义如下。

- 训练服务
  - ◆ 模型维度：
    - ✧ 模型准确率：本场景下没有可准确计算的模型准确率，以用户需求满足率和系统资源利用率联合表示。

$$\theta = f(requirement, resource)$$

- ✧ 模型泛化性：模型泛化性以卫星服务的多个区域（如 M 个）的模型准确率均方差值表示。

$$P = \sqrt{\frac{\sum_{i=1}^{M}(\theta_i - \theta_{ave})^2}{M}}$$

其中，θ_i 为模型在区域 i 的准确率，θ_{ave} 为 M 个区域的模型平均准确率，表示如下。

$$\theta_{ave} = \frac{\sum_{i=1}^{M}\theta_i}{M}$$

- ◆ 资源维度：
  - ✧ 算力资源数量：分配算力大小，以 FLOPS 表示。
  - ✧ 存储资源数量：分配存储空间，以 MB 表示。

■ 推理服务

- ◆ 模型维度：
  - ✧ 并发性能，以并发推理数和推理响应时间表示。
- ◆ 资源维度：
  - ✧ 算力资源数量：分配算力大小，以 FLOPS 表示。
  - ✧ 存储资源数量：分配存储空间，以 MB 表示。
- ◆ 传输维度：
  - ✧ 传输时延：执行体至推理模型的数据传输时延，以 ms 为单位。
  - ✧ 传输可靠性：执行体至推理模型的数据传输可靠性。

6.2.4 AI 多维资源 QoS 设计方案

6.2.4.1 指标描述

6G 引入 AIaaS 后，业务保障的要素从纯连接演变为连接，计算，数据和模型四要素，且原先 UE-RAN-CN 端到端结构演变为多网络节点的复杂拓扑，考虑到 QoS 保障的维度和方式相对 5G 通信 QoS 有了很大的变化，6G QoAIS 指标体系纵向上分为服务 QoS、任务 QoS 和资源 QoS 三个层次，横向上分为四个维度，分别对应连接，数据，计算，和算法四要素。

指标的具体设计上，参考 5G QoS，分为 QoS parameters 和 QoS characteristics 两级。QoS Parameters 的指标则定义了 QoS 管理对象和管理方式，如连接的核心是对于速率进行管理（GFBR/MFBR/AMBR），计算的核心是对于计算资源进行管理（比如 FLOPS）等。不同节点根据具体的 parameters 来进行 QoS 保障，如根据 6QI 和 ARP 管理时频域资源分配，根据 notification control 管理是否在 QoS 达不到要求时进行上报等。QoS Characteristics 对应 6QI，根据不同要素（连接/数据/计算/算法）的不同业务属性（如数据，按照去除唯一性/降噪/特征提取/缺失值处理业务的不同）进行分类，划分后由唯一的 6QI 值进行对应。

6.2.4.2 指标计算示例

该场景为网络基于实时的路况提供通行时长预估。业务流程如下：1）车辆向 6G 网络运营商订购了 AIaaS 服务（签约时已经确定 SLA 要求，比如轻算法，重数据）；2）车辆（终端）发起 AI 业务请求；3）NAMO 根据该车辆（终端）的签约信息中的 SLA 转化后形成的服务 QoS 作为输入完成该 use case 的服务编排，生成 workflow 和 task，选择合适的任务锚点（RAN/UE），下发任务内容。编排过程同时完成服务 QoS 到任务 QoS 的分解（从指标上看，non-separable 指标不变，separable 指标分段）。4）各任务执行节点根据任务 QoS，按照四要素资源维度进行划分，按照资源 QoS 要求执行数据采集处理，模型训练，推理，验证（在计算执行体上按照计算资源 QoS 执行计算）等；5）各任务锚点完成任务后，得出最终的推理结果并返回给该车辆。6）该车辆完成行程后返回实际通行时间给网络，网络完成对预测模型的验证，并释放任务上下文。

相关 AI 服务包括 AI 数据服务，AI 训练服务，AI 推理服务，AI 验证服务。

QoS characteristics 对应 6QI，和 5QI 从连接角度对于不同业务进行划分不同，6QI 是从连接/计算/算法/数据四个维度进行 AI 业务的分类的，具体如下所示：

连接维度：

- Resource type：连接拓扑（区别于5G UE-RAN-CN唯一形式）；连接分类（纯连接/计算连接/算法连接/数据连接）；GBR/non-GBR
- Default priority level
- Packet delay budget
- Packet error rate
- Default maximum data burst volume
- Default averaging window

计算维度：

- Resource type：计算拓扑；计算分类（纯计算/算法计算/数据计算）；GFLOPS/non-GFLOPS；
- Default priority level
- Computing latency
- Computing accuracy
- Degree of parallelism
- Default averaging window

数据维度：

- Resource type：数据拓扑；数据源类型（sensing data/AIML data/IoT data/network operation data）；guaranteed I/O or non-guaranteed I/O; guaranteed volume/non-guaranteed volume；
- Default priority level
- Data processing latency
- Degree of parallelism

算法维度：

- Resource type：算法拓扑；模型类型；guaranteed training precision/non-guaranteed training precision；guaranteed inference accuracy/non-guaranteed inference accuracy；
- Default priority level
- Model inference latency
- Model training latency
- Degree of parallelism

6.2.5 AI 模型 QoS 设计方案

6.2.5.1 指标描述

针对 AI 模型，6G 新系统要求具备模型训练，模型评估，模型认证，模型推理，模型存储等功能和服务。相应的 QoAIS 指标可以包括 AI 模型性能、AI 模型时延、AI 模型存储等方面。

表 14 AI 模型相关 QoAIS 指标

QoAIS 指标	描述
AI 模型性能	模型的性能主要考虑 AI 模型的固有属性，如准确率，召回率，泛化性，鲁棒性，复用性和可解释性等。涉及模型训练、推理和迁移等。根据模型应用场景，用例门类和模型类型的不同，模型性能的具体属性可能不同。
AI 模型时延	产生模型时延的因素有模型的传输，模型再训练和模型推理等方面，涉及模型参数量，存储位置，再训练模式，再训练数据，推理数据，连接性能和网络算力配置等，体现 AI 服务总指标和连接，算力，数据和算法等分指标之间统分关系。
AI 模型存储	6G 新系统的架构充分考虑了云、网、边、端 等异构资源的发现和编排。须考虑如何在异构的存储资源中最优地存储 AI 模型，以保障时延和用户体验等。

以 AI 模型性能为例，根据 AI 模型的智能类型，AI 模型的性能评估指标存在着不同的共性或特性，如图像处理中的目标检测，可采用物体实际区域与推测区域的交并比（IoU=物体实际区域与推测区域的重合面积/两个区域整体所占的面积）来评价；异常检测等分类问题的性能可采用常用的准确率、召回率或 F1-score 来衡量；多小区流程预测这类指标预测问题，可采用图神经网络作为基础的 AI 算法，模型性能可通过预测值和真实值之间的平均绝对误差（MAE）或均方根误差（RMSE）等指标体现。

6.2.5.2 指标计算示例

场景描述：考虑通过强化学习的方式进行多小区联合参数优化，以提升网络质量，例如基于多智能体强化学习的多小区 MIMO 天线权值优化。具体地，该方法包括智能体环境和优化算法两个主要模块。智能体环境首先在计算设备上构建的一个虚拟通信环境（可由数字孪生实现）。该虚拟环境基于现场环境信息，可快速输出不同基站位置、不同天线参数等状态下，UE 的信号强度和信噪比。智能体环境基于基站位置和天线参数等状态，生成一个动作发送给虚拟通信环境，虚拟通信环境通过计算，返回信号覆盖值等仿真结果作为模型的奖励。算法根据当前的奖励选择下一个动作。 经过多次循环，优化算法从智能体环境中获取动作和奖励反复学习，最终学习到一个最优的结果。

相关 AI 服务：AI 数据服务，AI 训练服务，AI 推理服务。

指标计算

性能指标界：在实际应用中，该方法基于数字孪生环境等虚拟通信环境生成配置组合，并部署到现场执行，我们通过计算 RSRP 覆盖性能和 SINR 覆盖性能的提高，评价配置组合（如相对人工经验方法优化效果的提升比率），可计算配置组合评价的期望作为 AI 模型的性能指标参考标准。例如，RSRP 覆盖性能提升可通过以下公式计算，

$$\Delta RSRP = \frac{RSRP_{RL} - RSRP_{ori}}{RSRP_{exp} - RSRP_{ori}} \times 100\%,$$

其中$RSRP_{RL}$表示通过强化学习优化后的 RSRP，$RSRP_{exp}$表示基于专家经验优化后的 RSRP，$RSRP_{ori}$是优化前的 RSRP。同时强化学习策略模型的学习训练需要较大的开销进行多轮训练，其训练耗时和传输也是评价模型性能的关键指标。将优化后的策略部署在近端计算环境（如基站等）中，可进一步降低推理时延。

开销：模型的训练结合实际环境参数和虚拟通信环境，在数据传输上将产生较大的开销。在建设好虚拟通信环境的基础上，基于强化学习的模型训练仍产生较大的算力开销。因此，模型的训练适宜在算力等资源充足的环境，如云端进行。模型在部署时，产生将策略模型从云端传输到近端计算环境的开销，及相应的存储开销。

安全隐私：模型需收集地形，建筑物，用户分布等信息并传递给虚拟通信环境，涉及到用户信息，需进行脱敏处理。

自治：结合数字孪生技术，massive MIMO 天线权值可逐步实现自动化配置。具体地，流程初始阶段，需要结合人工经验和配置的性能指标，逐步确定部署策略回传虚拟通信环境更新的时间，以保证服务效果并积累更新相关信息数据，用于训练策略更新模型，将服务从半自动化升级为自动化。

综上所述，上述场景相关 AI 服务的 QoAIS 指标通过以下公式抽象表示。

$$\text{AI 数据 QoS} = f_{data}(\text{算力资源，传输资源，数据描述，缓存机制，安全级别}),$$

$$\text{AI 训练 QoS} = f_{train}(\text{算力资源，传输资源，模型大小，模型超参，缓存机制，安全级别}),$$

$$\text{AI 推理 QoS} = f_{inf}(\text{算力资源，传输资源，模型大小，缓存机制})。$$

其中算力资源包括计算和存储的硬件设施，如 GPU 类型和显存大小等；传输资源包括传输方式，速率，吞吐量等；数据描述包括特征类型，存储方式，数值精度等；模型超参包括智能体数量，训练轮数等。

7.待研讨问题

（1）指标需求

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？

网络在引入 AI 服务后，用户对 AI 服务安全性和隐私性上会存在哪些不同的需求选项？如何对其进行分级和量化？

（2）设计原则

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），如何分阶段进行指标设计和引入？

在 AI 服务层，由于每一类服务存在共性指标，也有一些跟服务特性相关的指标，服务 QoS 是否需要统一？针对不同 AI 服务类型是否设计特异性服务 QoS？

（2）指标设计

QoAIS 体系包括 AI 服务、任务和资源三个层次，每层 QoS 指标参数是否会有重叠？任务 QoS 与服务 QoS、资源 QoS 的映射关系如何？

不同场景的需求是否导入到不同层面的 QoS 上？从逻辑角度，服务到任务的映射、任务到资源的映射均可以是一对一的，因此，不同场景的需求可以统一到服务 QoS 上；从功能部署角度，由于服务 QoS 可能位于管理系统中，任务 QoS 和资源 QoS 可能位于控制面，对于有不同时延需求的场景，导入到不同层面的 QoS 效率更高。

8. 总结与展望

6G 网络将构建内生 AI，建立可应用于各种智能场景的能力体系。不同的应用场景对 AI 服务的质量有着不同的要求，因此需要制定一套全面完整的指标体系，通过量化或分级的方式体现用户业务的需求，并对 AI 服务质量进行评估和保障。

本白皮书旨在为 6G 网络服务质量保障体系设计提供科学依据，首先介绍了 QoAIS 指标体系的研究背景，并从运营商、网络、方案多方面分析了 QoAIS 的驱动力。然后阐述了 QoAIS 的定义与内涵，将其分为 AI 服务的 QoS、AI 任务的 QoS、AI 资源的 QoS 三层。接下来研究了 QoAIS 指标体系的设计原则以及已有的相关方案，在此基础上，打破了传统仅考虑通信资源指标的 QoS 体系，从算力、连接、算法、数据等不同维度设计 QoAIS 模型和指标，给出了多种可能的方案，并针对具体场景展开分析。最后对未来可能的研讨问题进行了总结。

展望未来，6GANA 将联合合作企业及高校持续对 QoAIS 指标体系进行深入研究，旨在为 6G 网络内生 AI 服务质量的保障和评估提供参考，助力千行百业的数智化转型，推动构建智慧内生的网络体系。

参考文档

[1] 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.

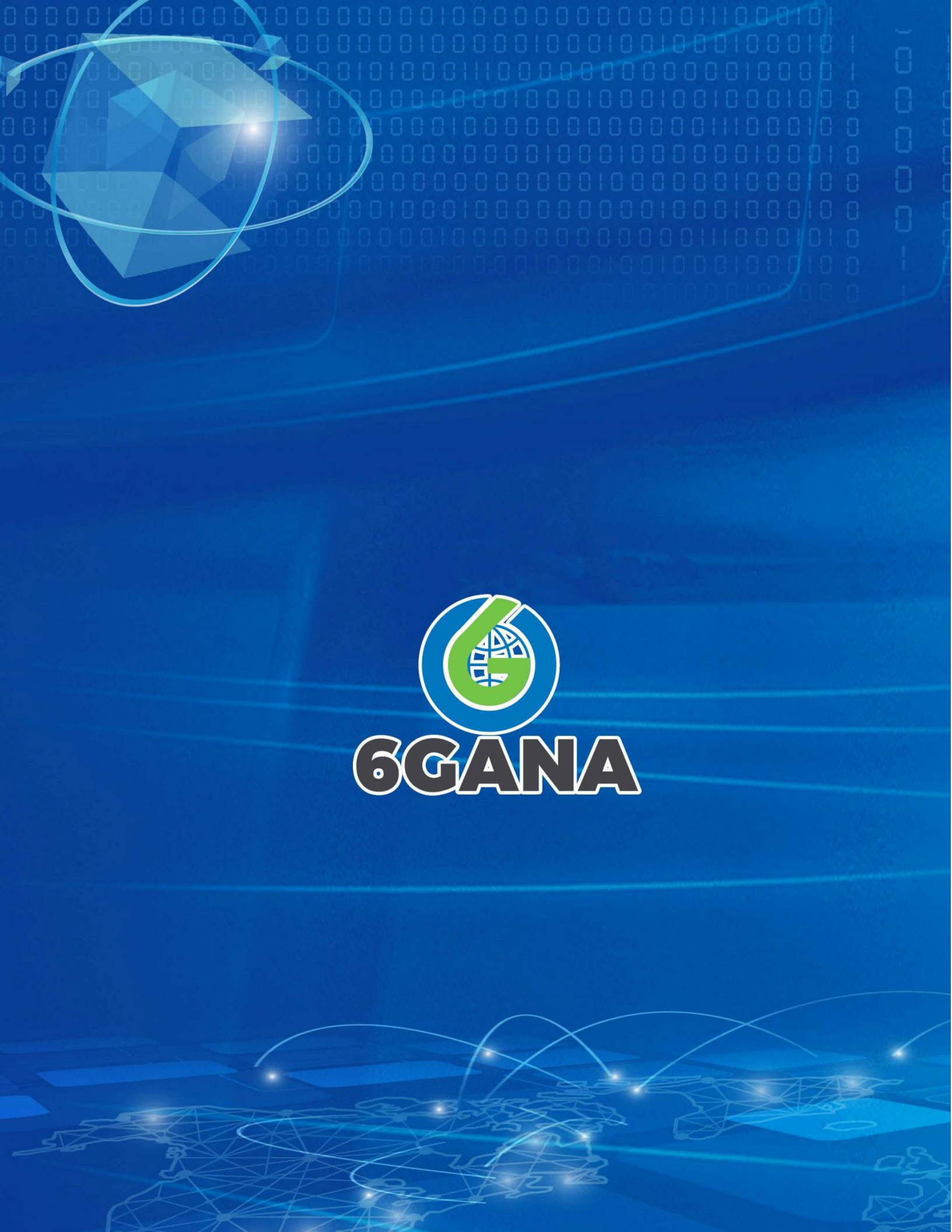
[2] 6GANA. 6G 内生 AI 网络架构十问[R], 2022.
[3] 孙子剑,廖逸玮,鲁智敏. 面向 6G 智能内生的隐性语义认知通信[J]. 移动通信, 2023, 47(04): 7-13.
[4] 王晴天, 刘洋, 刘海涛. 面向 6G 的网络智能化研究[J]. 移动通信, 2022, 38(09): 151-160.
[5] 王瑜新, 章秀银, 徐汗青. 6G 需求、愿景与应用场景探讨[J]. 电子技术应用, 2021, 47(03): 14-17.
[6] 赵军辉, 李一博, 王海明. 6G 定位的潜力与挑战[J]. 移动通信, 2020, 44(06): 75-81.
[7] 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(4): [LI Tingting,XIN Yutong,RAN Peng,et al. Channel State Information Feedback for Massive MIMO System Based on Asymmetric Convolution [J].Radio Communications Technology, 2022,48(4):]
[8] 3GPP TS 23.501: "System Architecture for the 5G System; Stage 2".
[9] 朱婷婷,张勇. 5G 网络实现 4K/8K/VR 高清视频业务的标准化模型和应用场景研究[C]// 中国标准化协会. 第十七届中国标准化论坛论文集. 2020:6.
[10] 3GPP TS 23.502: "Procedures for the 5G System (5GS) Stage 2".
[11] 3GPP TS 23.503: "Policy and charging control framework for the 5G System (5GS) Stage 2".
[12] 郭凤仙,孙耀华,彭木根.6G 算力网络:体系架构与关键技术[J]无线电通信技术,2023,49(1) :21-30.
[13] 武振宇,肖子玉,刘鹏等.算力网络业务分级模型与指标体系研究[J].电信工程技术与标准化,2023,36(03):8-13.
[14] 李一男,唐琴琴,彭开来,等.以服务为中心的算力网络度量与建模研究[J].信息通信技术与政策，2023，49(5):21-29.
[15] 李宁东,邢玉萍,马新翔.我国算力网络发展评估体系研究[J].信息通信技术与政策,2023,49(05):15-20.

缩略语

缩略语	全称	中文释义
5QI	5G QoS Identifier	5G 服务质量指示符
AF	Application Function	应用功能
AIaaS	AI as a Service	AI 即服务
AMBR	Aggregate Maximum Bit Rate	聚合最大比特率
ARP	Allocation and Retention Priority	分配和保留优先级
FaaS	Function as a Service	功能即服务

FER	Frame Error Rate	误帧率
FLOPS	Floating Point Operations Per Second	每秒浮点运算次数
GBR	Guaranteed Bit Rate	保证比特速率
GFBR	Guaranteed Flow Bit Rate	保证流比特率
IaaS	Infrastructure as a service	基础设施即服务
IOPS	Input/Output Operations Per Second	单位时间读写次数
LTE	Long Term Evolution	长期演进
MAE	Mean Absolute Error	平均绝对误差
MBR	Maximum Bit Rate	最大比特速率
MFBR	Maximum Flow Bit Rate	最大流比特率
MIMO	Multiple-Input Multiple-Output	多输入多输出
mMTC	massive Machine Type Communication	大规模机器类通信
NAMO	Network AI Management & Orchestration	网络 ai 管理与编排
NEF	Network Exposure Function	网络开放功能
PaaS	Platform as a service	平台即服务
PCF	Policy Control Function	策略控制功能
QoAIS	Quality of AI Service	AI 服务质量
QoS	Quality of Service	服务质量
RAN	Radio Access Network	无线接入网
RMSE	Root Mean Squared Error	均方根误差
RPO	Recovery Point Object	恢复数据目标
RQA	Reflective QoS Attribute	反射式 QoS 属性

RSRP	Reference Signal Receiving Power	参考信号接收功率
RTO	Recovery Time Object	恢复时间目标
SaaS	Software as a service	软件即服务
SINR	Signal to Interference plus Noise Ratio	信干噪比
SLA	Service-Level Agreement	服务等级协议
SMF	Session Management Function	会话管理功能
UE	User Equipment	用户设备
UDM	Unified Data Management	统一数据管理


6GANA