# 6G内生AI架构及AI大模型

刘光毅

2023年7月

# 目录

# 泛在智能的需求

中国移动 China Mobile

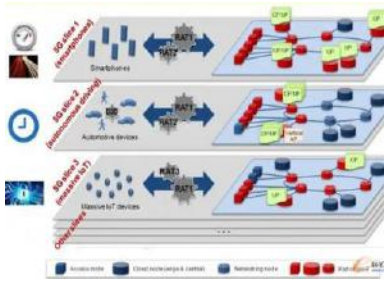
## 人工智能已成为新一轮产业升级的核心驱动力，产业的自动化、数字化、智能化需要泛在智能

### 网络自治需要AI

运行与维护

应急通信

智能覆盖



定制网络

### 用户需要AI

声纹识别


机器翻译

智能导航

个性化推荐

### 企业需要AI

医疗识别

安全监控

机器人救援

智能制造

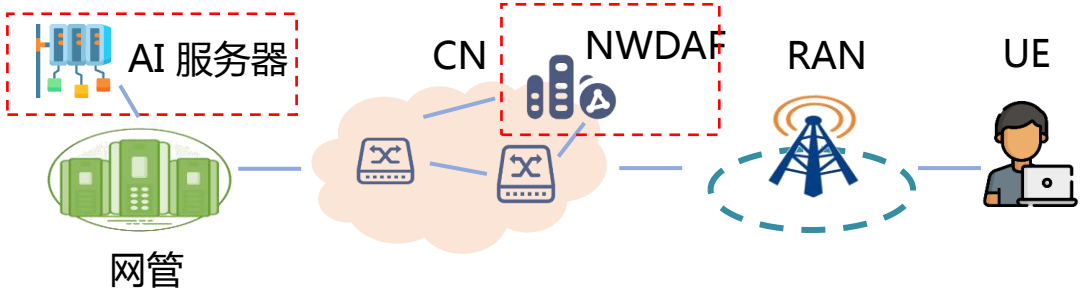
**6G网络需要高效地为网络自治、ToC和ToB提供AI服务!**

# 5G时代网络和AI融合的启示

## 网络智能化的启示

### 外部和叠加AI

- **模式一：将具备AI资源的服务器接入网管设备，为网络提供AI模型。**
- **模式二：在核心网络中增加AI作为新的网络功能，如NWDAF。**

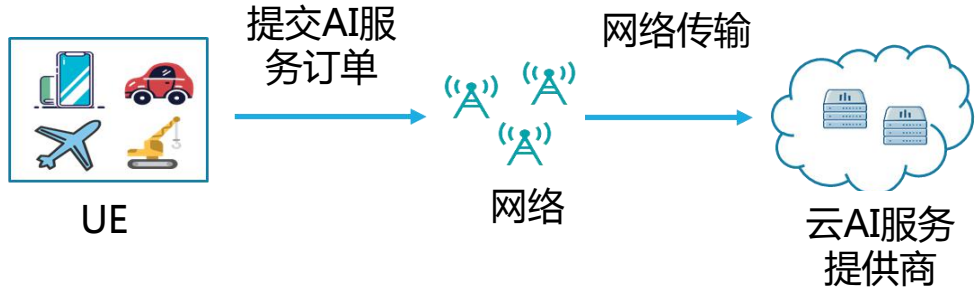



- **6G需要一个统一的框架：**以支持有效的AI性能验证和保障方法。
- **6G需要高效的数据采集和传输：**以实现AI预验证、在线评估和优化的全自动闭环。
- **6G需要计算、数据、模型和连接的协同：**以提供更低的延迟和有保障的QoS。

## 云AI服务提供商的启示

### 云AI服务提供商

- **云AI服务提供商在用户提交订单后提供尽力而为的AI服务。**




- **6G需要充分利用网络的人工智能相关资源：**以感知网络状态，利用网络广泛分布的计算、数据、算法资源，提供高效的人工智能服务。
- **6G需要为AI服务提供QoS保障：**以提供满足用户特定需求的AI服务。
- **6G需要保护数据隐私和安全数据：** 在提供可靠AI服务的同时，防止数据泄露。

# 5G智能到6G的转变——内生AI

## 面向6G泛在智能的愿景，网络与AI的融合需要有三大转变，6G网络将是内生AI

### 1. 从 烟囱式开发 到 泛在智能的统一网络AI框架


工业互联网


智慧能源


智慧农业


车联网


智慧医疗


云游戏/云XR

网络自治
网络极简

AIaaS

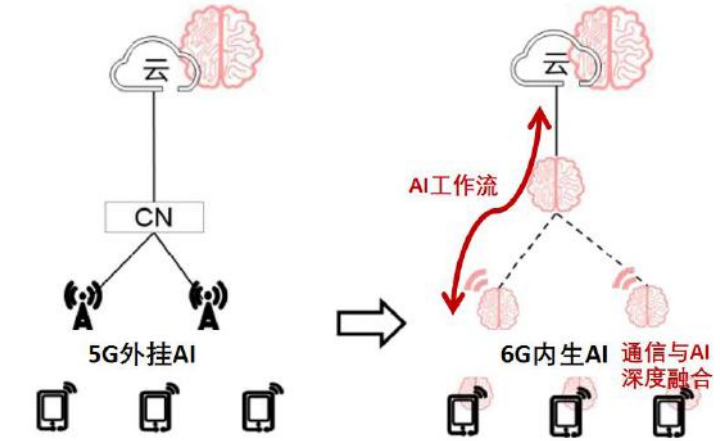
QoAIS

QoAIS

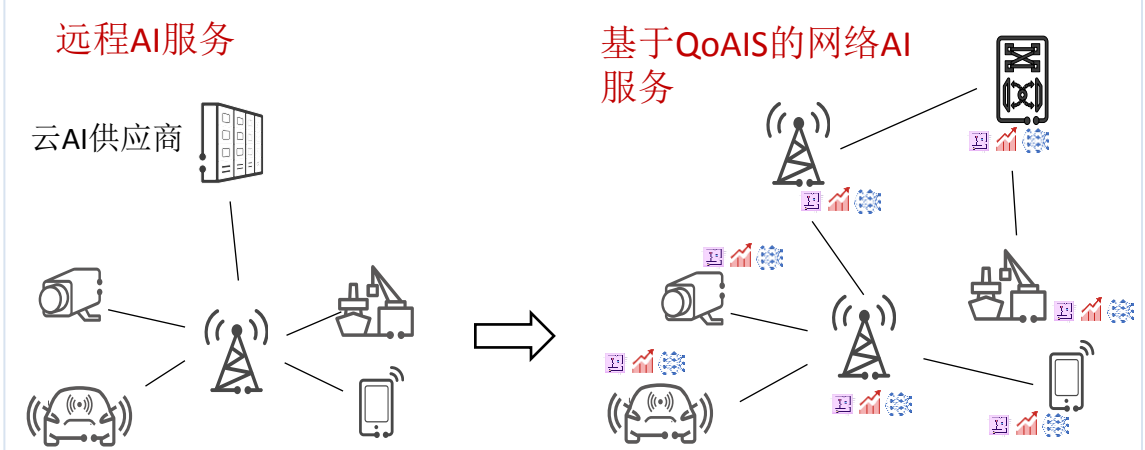
AIaaS

内生AI

### 2. 从 外挂叠加 到 内生一体




### 3. 从 尽力而为 到 按需可得



# 目录

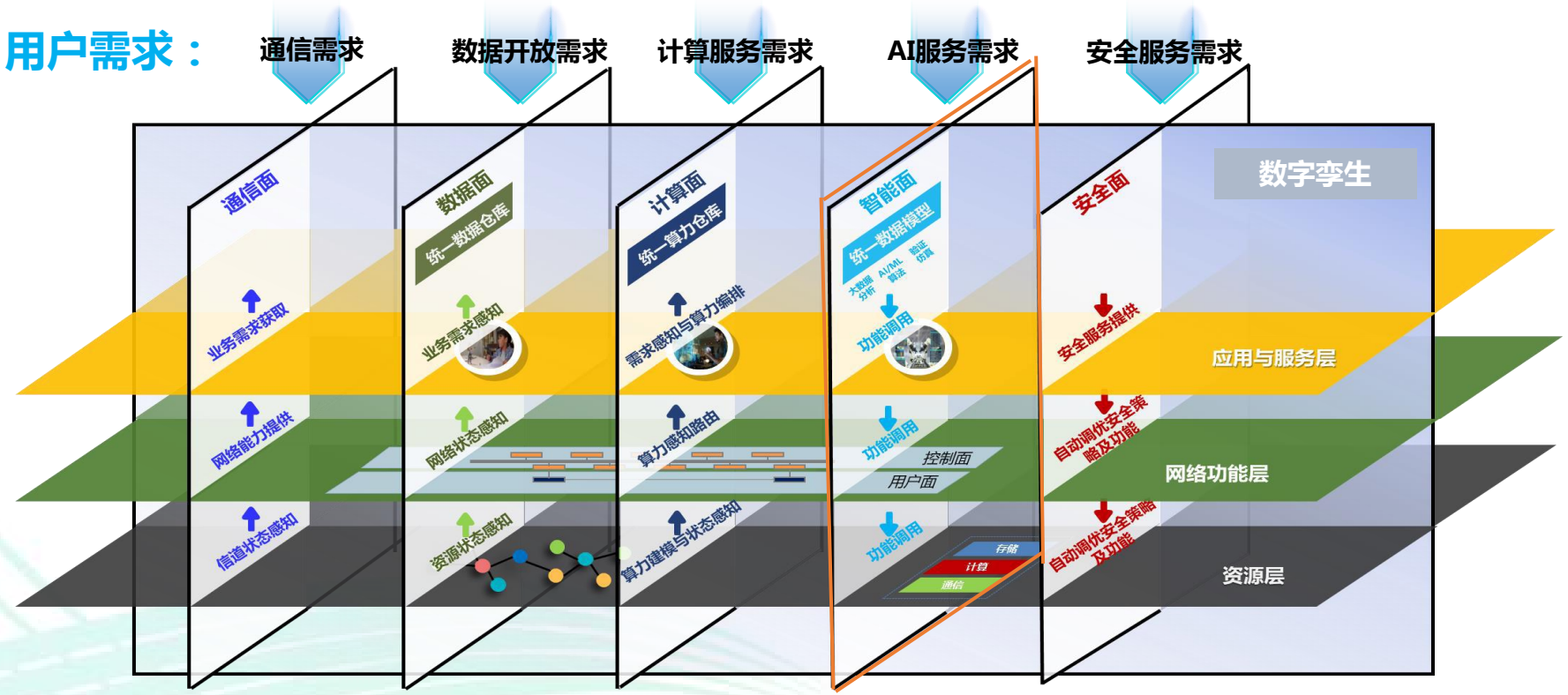
# 6G网络逻辑架构

中国移动 China Mobile

**面向泛在智能等多种服务需求，6G将新增多个逻辑面，提供通信、感知、计算、AI、大数据、安全等一体融合的多维网络能力，以及平台化、一体化的服务体系**

**与5G网络不同，6G网络将定义新的数据面、智能面、计算面等，**

**并有望扩展传统的控制面和用户面。**




**数据面：**
管理网络数据，提供数据服务

**计算面：**
管理计算并提供计算服务

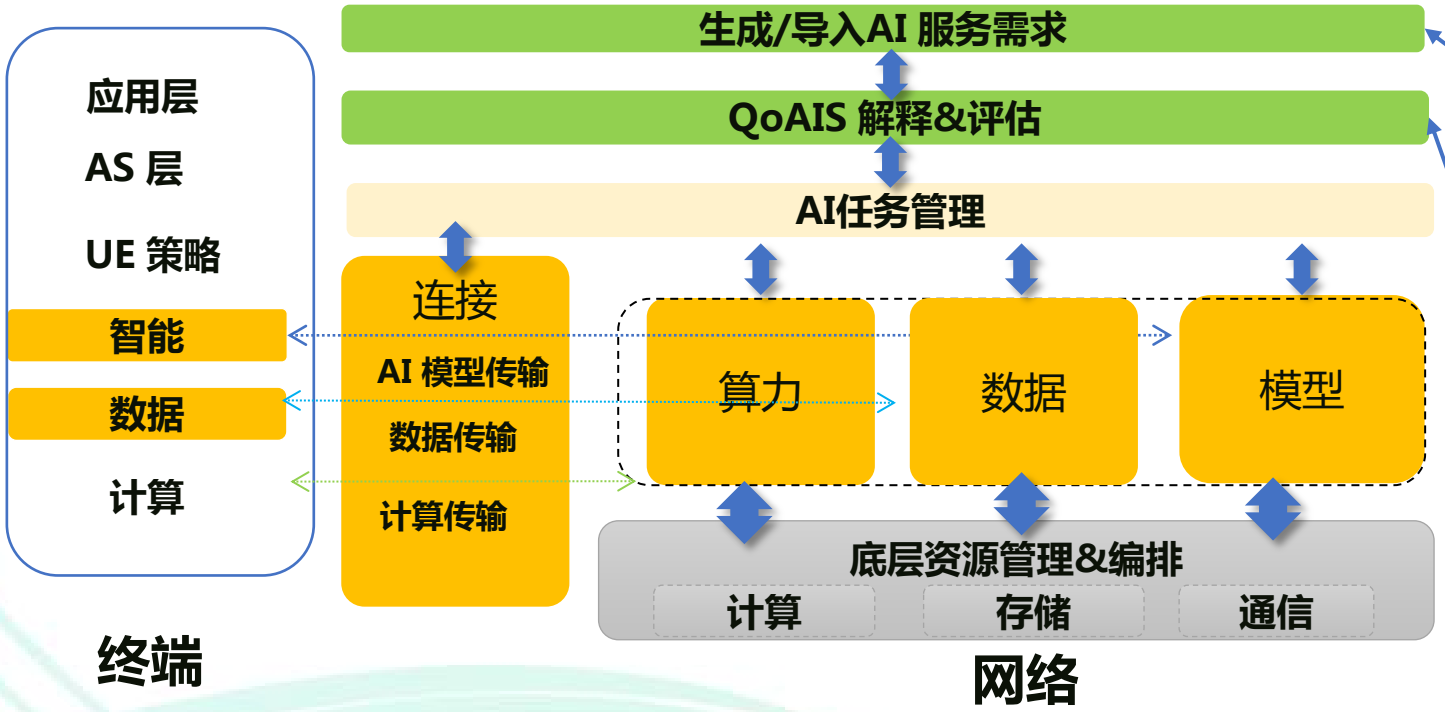
**智能面：**
为原生AI提供全生命周期的运行环境

# 6G网络逻辑架构——内生AI的具体设计

中国移动 China Mobile

**AI业务的实现需要通信、计算、数据和模型服务的支撑，需要不同逻辑面之间复杂的协同机制。**

内生AI将AI 三要素（数据、算法和算力）与网络连接一样下沉为网络内部的基本资源，使网络通过多维资源的协同，直接、便捷地为用户提供高质量的AI服务。




1. 泛在智能的统一网络AI框架

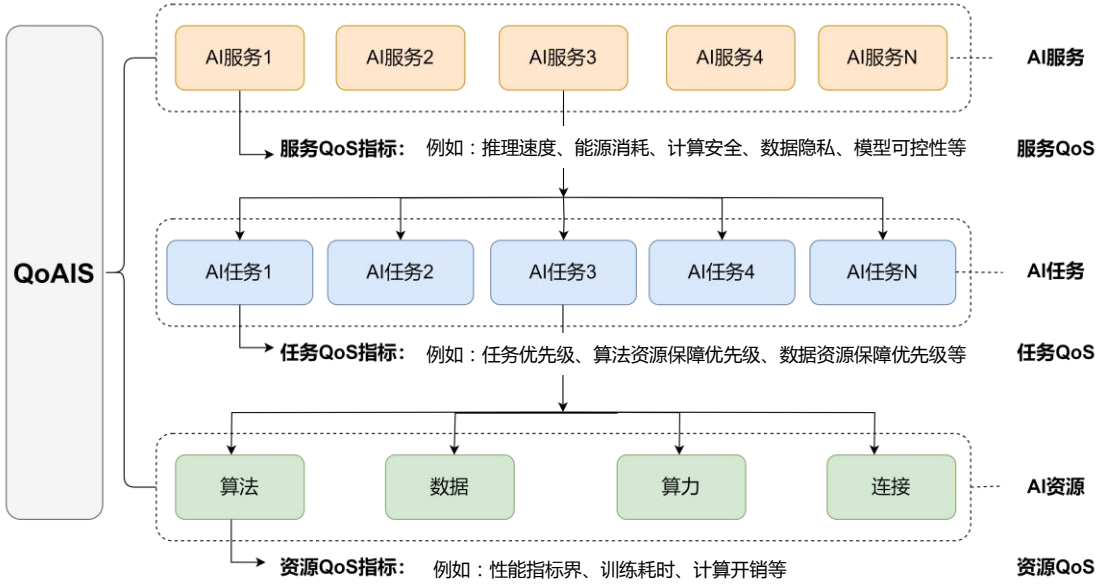
2. 从外部叠加向内部集成转换

3. 从尽力而为转向按需应变转换

# 关键技术1：AI服务质量（QoAIS）

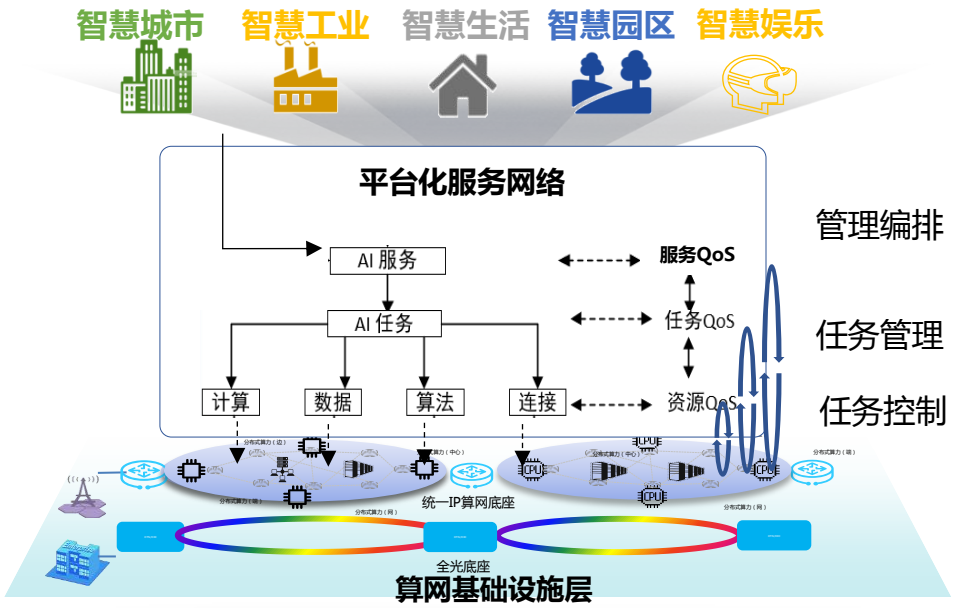
**AI服务质量（QoAIS）指标体系，突破传统QoS体系中以会话和连接性能为主要关注指标，将安全、隐私、自治和资源开销作为新的评估维度纳入，形成一套标准化的AI服务质量评价体系，为用户AI服务质量的衡量和保障机制的设计提供了统一的依据**

## AI服务质量（QoAIS）指标体系




- QoAIS指标体系是网络对AI服务的质量和效果进行保障所使用的一套指标体系
- QoAIS包含AI服务的QoS、AI任务的QoS、AI资源的QoS三个层次上的指标，三层指标间具有映射关系

## QoAIS保障机制




- QoAIS是管理编排和任务管理/控制的重要输入，管理编排需要将服务QoS分解为任务QoS，再映射到对连接、计算、数据和算法等各方面的资源QoS要求上
- 为保障QoAIS的达成，需要“三层闭环”的保障机制

# 关键技术2：AI计算与通信深度融合

**传统网络中提供AI服务需要通信和计算协议之间频繁的交互与协调，需要设计一套通算融合的内生AI协议，实现对计算和通信的协同管控与承载，满足AI所需的连接和分布式计算服务、以及基于AI的连接和计算融合控制需求。**

**管理面、控制面、用户面三个维度实现计算和通信深度融合**

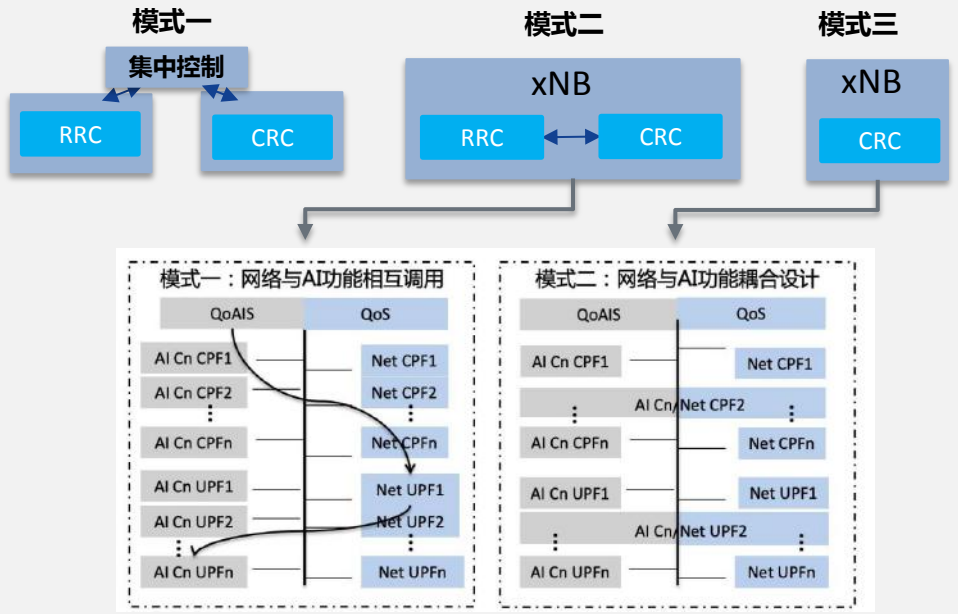
**5G MEC边缘计算缺点**

管理面融合的松耦合设计
效率低成本高
安全性隐私性不够

**6G内生AI的算力需求**

高计算效率
低能耗、低时延
满足各类AI场景差异化QoAIS需求

**控制面：计算和通信深度融合的三种模式**




**优势**：当QoAIS指标发生恶化时，可快速调整，保障QoAIS目标的持续达成。

**管理面**：计算和通信资源的协同编排管理。
**优势**：宏观网元连接关系、各类资源状态，保证网络级性能指标较优。

**用户面**：
联合设计“计算协议+通信协议”
满足QoAIS+均衡分配网络资源，
满足“性能 + 开销”上的需求。

CRC：Computing Resource Control
NC：Node Compute

# 目录

# 6G与AI的融合迎来新机遇：通用大模型时代

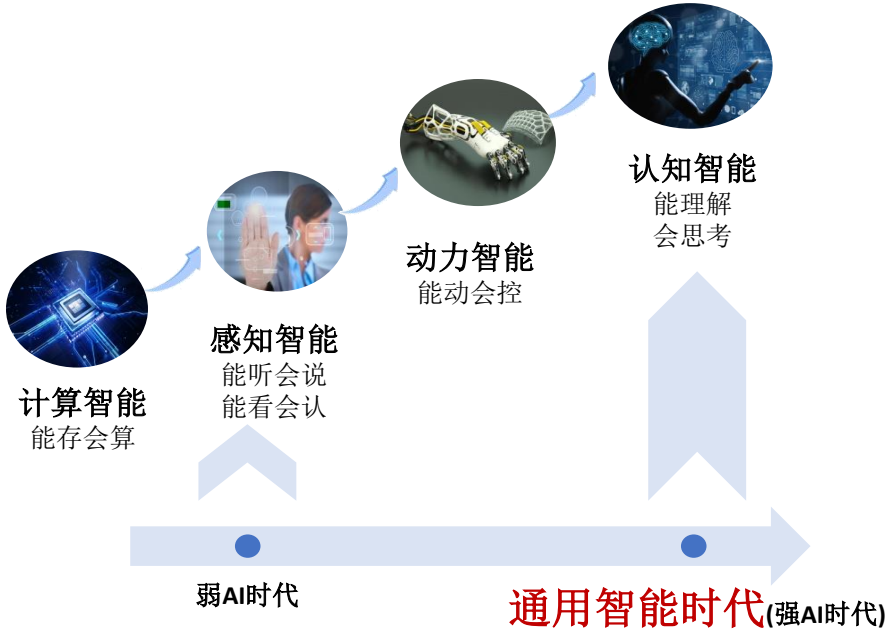
中国移动 China Mobile

## AI迈入通用智能时代，大模型的出现将为6G与AI的融合带来巨大变革

### 发展模式新跃迁

**ChatGPT现象级事件，标志着人工智能进入通用智能时代**

从“能听、会说、能看、会控”，走向“能理解、会思考、会创作”，甚至能“自主决策、自主处理问题”

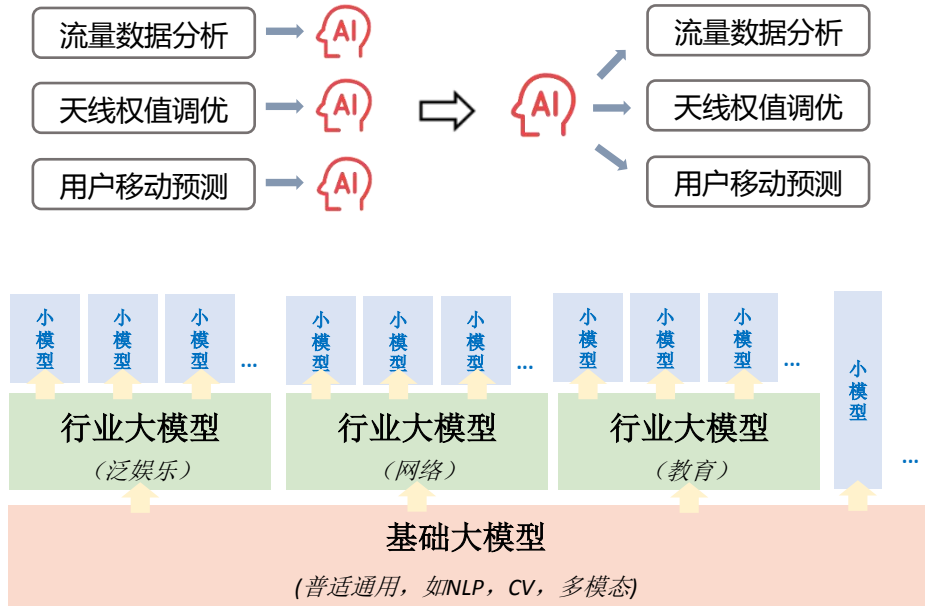



**6G网络内生AI如何使能AI大模型？**

### 当前网络AI泛化性有限

**从小模型到大模型，生产效率跨越式提升**

基础通用大模型具有泛化性，网络智能化将从用例驱动转变为能力驱动，迅速降低应用开发门槛，加速AI工程化、规模化落地




**如何设计赋能网络的AI大模型？**

# 网络使能AI大模型：成本与资源的挑战

## AI大模型在训练、推理、储存方面开销极大，网络难以支撑

### 训练成本

■ **GPT-3训练一次的成本约为140万美元**

Azure 构建  训练 ChatGPT

微软云服务　　超过1万枚A100

### 推理成本

■ **假设访量2500万/日，10个问题/用户，30字/问题**

ChatGPT 换算  消耗 

每日需要729,167个A100运行小时　　每日需要30,382个A100　　3,798个服务器 电费：30万/日

### 储存成本

GPT-3 需要  另加 

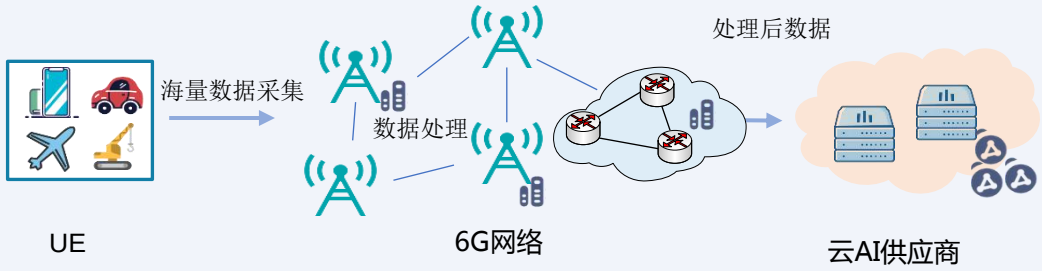
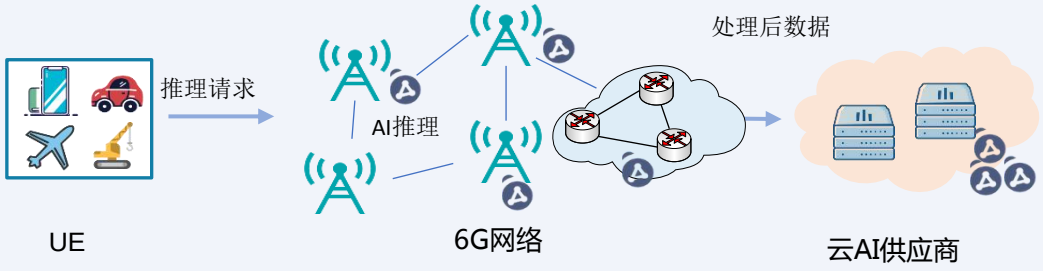
1750亿个参数　　700GB的参数储存　　45TB的数据集

数据来源：OneFlow，国盛证券研究所

| | 模型名称 | 参数 | 领域 | 功能 |
|---|---|---|---|---|
| | BERT | | NLP | 语言理解与生成 |
| | LaMDA | | NLP | 对话系统 |
| 谷歌 | PaLM | 5400亿 | NLP | 语言理解与生成、推理、代码生成 |
| | 1maeen | 110亿 | 多模态 | 语言理解与图像生成 |
| | Parti | 200亿 | 多模态 | 语言理解与图像生成 |
| 撇软 | Florence | 6. 4亿 | CV | 视觉识别 |
| | | 170亿 | NLP | 语言理解、生成 |
| Faoebook | 0PT-175B | 1750亿 | NLP | 语言模型 |
| | M2M-100 | 150亿 | NLP | 100种语言互译 |
| | Gato | 12亿 | 多模态 | 多面手的智能体 |
| DeepMind | Gooher | 2800亿 | NLP | 语言理解与生成 |
| | AIohaCode | 414亿 | NLP | 代码生成 |
| | CLIP&DALL-E | 120亿 | NLP | 图像生成、跨模态检索 |
| OpenA1 | Codex | 120亿 | 多模态 | 代码生成 |
| | ChatGPT | 175B | NLP | 语言理解与生成、推理等 |
| | NLP大模型 | | NLP | 语言理解、生成 |
| | CV大棋型 | | CV | 图像试别 |
| 百度 | 跨棲态计算大棋型 | 千亿级别 | 多模态 | 语言理解与困像生成 |
| | 生物计算大模型 | | CV | 化合物表征学习、分子结构预测 |
| 阿里巴巴 | M6 | 万亿级别 | 多模态 | 语言理解与图像生成 |
| 腾讯 | 混元大馍型 | - | NLP | 语言理解与生成 |
| 京东 | K-PLUG | - | NLP | 语言理解与生成、推理、代码生成 |
| 三六零 | | - | NLP | 智能搜索 |
| 字节跳动 | DA | - | NLP | 语言理解 |
| 科大讯飞 | 中文預训练棋型 | - | NLP | 语言理解与生成、语言互译 |
| 百度 | 文心一言 | 千亿级 | NLP | 对话互动，回答问题，协助创作，获取信息 |

# 网络使能AI大模型：可能模式

**6G内生AI为AI大模型的训练过程提供链接、数据服务，为推理过程提供链接、计算、模型拆解/分发服务。**

| | AI训练服务 | AI推理服务 |
|---|---|---|
| |  |  |
| **特点** | AI大模型训练通常在数据中心的多台服务器中，服务期间需要用高速率光纤连接，难以将AI大模型的训练过程部署到无线网络中。 | AI大模型需要较大的储存空间和较强的AI推理芯片，单基站无法满足该需求。 |
| **服务** | 从用户、网络中采集大量数据，数据的预处理和流量的调度，支撑AI大模型的训练。 | 在实现合理模型分割的前提下，可将大模型分布式的部署到无线网络中，提供AI推理服务。 |
| **潜在增益** | 6G网络承担数据采集、预处理等数据服务，节省网络中的数据传输，为云AI训练AI大模型提供更好的数据服务。 | 在6G网络中分布式部署AI大模型，更靠近用户侧，可能在时延方面具有优势。 |
| **未来问题** | AI大模型需要哪些特殊的数据分析技术？数据处理功能分布式部署后网络中数据该如何合理调度？ | 分布式部署导致推理时延增加和靠近用户侧传输时延减少该如何权衡？大模型分割、压缩、加速等技术是否可行；大模型分布式部署后节点之间数据如何合理调度？ |

# AI大模型赋能网络：可行性

中国移动 China Mobile

## AI赋能网络的场景主要分为网络运维和网络运行两大类，能否用一个AI大模型解决多种场景的问题？

| | 场景实例 | 所需数据（特征） | 数据来源 |
|---|---|---|---|
| 网络运行 | 编码调制、语义+编码调制、波形、多址、MIMO、干扰消除 | 非标准化数据：上下行信道 、语义信息、语义特征、小区内/间干扰 | 网络设备内部数据、语义信源数据 |
| | 基于无线栅格的切换、智能AMC、网络流量检测和拥塞控制、网络流量预测和调度优化 | 标准化数据：MR测量数据、MDT数据；<br>非标准化数据：链路级BLER、端口数据流信息数据（端口流量、时延等） | 终端测量上报；<br>网络设备内部数据 |
| | 业务识别和感知、异常行为监测 | 标准化数据：业务数据流信息数据（IP五元组、URL、PFD等） | 网络设备内部数据 |
| | 无线组网动态负载均衡、无线组网动态干扰规避、网络节能、智能寻呼、IP网络智能路由 | 标准化数据：MR测量数据、MDT数据、KPI监控数据（PRB利用率、小区吞吐量等）、控制面信令数据、业务数据流信息数据 | 网络设备内部数据<br>网管系统数据 |
| | 东数西算类、算网融合类、超算智算类等算网服务场景；智能需求分析、智能策略匹配、智能服务优化等 | 标准化数据：基础资源状态、拓扑、性能、成本、能耗、告警等数据；业务数据流信息数据、KPI监控数据、XDR、运行日志，告警等数据 | 网管/云管系统数据 |
| 网络运维 | 感知类：智能业务识别<br>诊断类：智能故障处理<br>预测类：智能扩容规划 | 标准化数据：KPI等监控数据、XDR数据、告警数据、MR数据、拓扑等资源数据<br>非标准化数据：日志数据、图片数据、文档/案例数据等 | 网管系统数据 |

**AI大模型赋能网络场景十分多样，需分析：数据是否可用？如何构建大模型？**

# AI大模型赋能网络：数据获取和处理的挑战

**网络运维的数据是以分钟/小时粒度数据为主，来源较为统一；网络运行的数据时间粒度、标准化程度、数据来源更为多样和复杂，获取较为困难**

## 数据是AI大模型的基础，如何获取适合AI大模型训练的数据面临极大挑战

**数据获取难**

- 物理层等数据源缺失，应用难开展
- 采集数据粒度不统一，数据难应用

**数据质量差**

- 数据记录不完整，应用难优化
- 数据记录不准确，应用难商用

### 标准化

联合业界共同制定新增数据采集规范，
制定按需动态数据采集粒度方案

### 实时校验

研发数据实时校验能力，
推动质量及时改进

### 数据开放

持续梳理和积累网络智能化数据集，对外开放，构建智慧网络创新系列生态，助力研究

### 闭环稽核

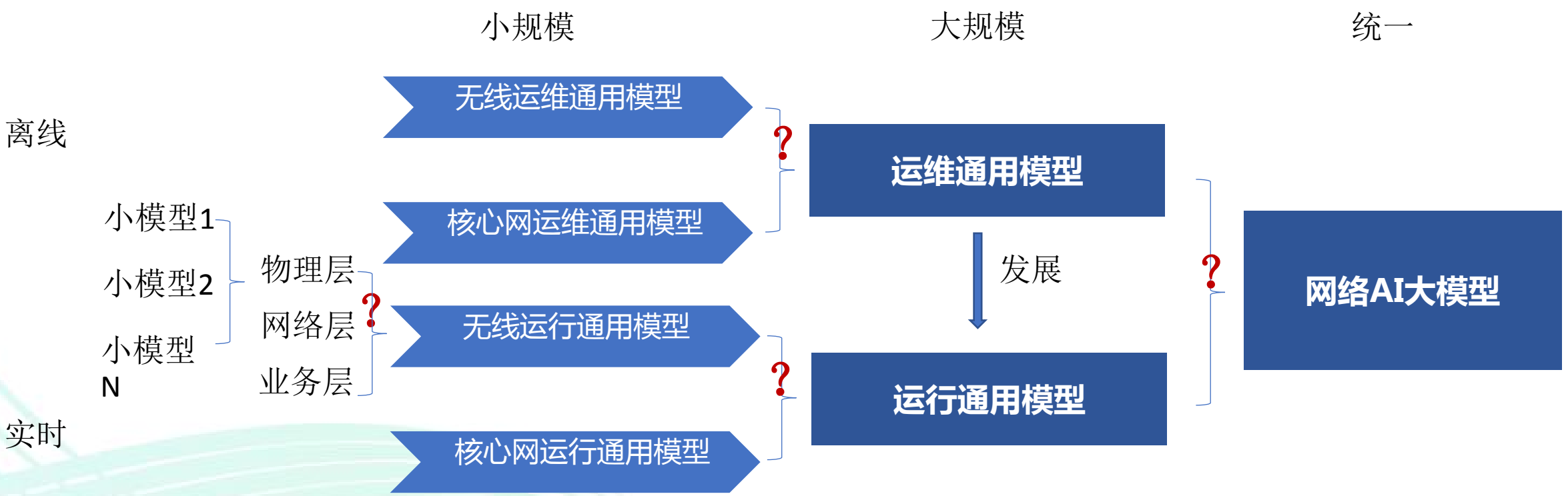
研发数据闭环稽核能力，
保证数据准确可靠

# AI大模型赋能网络：大模型的构建路径

**与ChatGPT不同，网络中存在大量结构化数据，且网络不同问题间的共性不清晰，网络AI大模型面临较大挑战**

**面向上述场景，可考虑分阶段探索，首先探索网络运维人工智能大模型**

**从小规模、离线入手，向大规模、实时发展，最终探索是否可以实现统一**



# 问题思考

中国移动 China Mobile

**应用场景**

- 如何探索网络使能AI大模型的潜在应用场景，挖掘场景价值？
- 如何评价AI大模型赋能网络的价值和增益？

**数据/模型**

- 如何构建统一的网络大模型？
- 对于网络而言，AI领域如何建立可解释性理论模型，保障网络中AI大模型决策的有效性和可靠性?
- 如何解决数据离散、设备数据获取难等问题？
- 如何利用数字孪生网络生成高质量数据，并对AI大模型进行验证？

**算力**

- 如何利用算力的泛在和流动性，使能大模型，如chatGPT、语义大模型？

**架构**

- 如何细化网络使能AI和AI赋能网络的统一架构，实现智能面/计算面功能、接口及流程高效设计？
- 架构如何支持AI大模型的分布式训练？

# 共同推进全球6G AI合作研究

中国移动 China Mobile

**6GANA——全球 6G Network AI 交流平台**

- 海量网络容量，零时延体验
- 全连接世界的神经网络
- 从原生虚拟化到原生智能化

**5G**
**产业转型**
**改变社会**

- Vertical Transformation
- Change the Society

**6G**
**智能普惠**
**人类挑战**

- Pervasive Artificial Intelligence
- Human Challenges

**6GANA**

**欢迎参加7.16-17的**
**6GANA TG 联合研讨会！**

**营造国际高水平学术交流平台**

IEEE WCNC — 2021

IEEE/CIC ICCC™ — 2022

IEEE ICC™ — 2023

IEEE GLOBECOM® — 2023(征稿中！)

**各位专家和学者，欢迎投稿 IEEE GLOBECOM 2023 Workshop 11 on Intelligent 6G Architecture: Towards Network Simplicity and Autonomy，探讨6G架构创新的新进展！**

# 谢 谢 ！