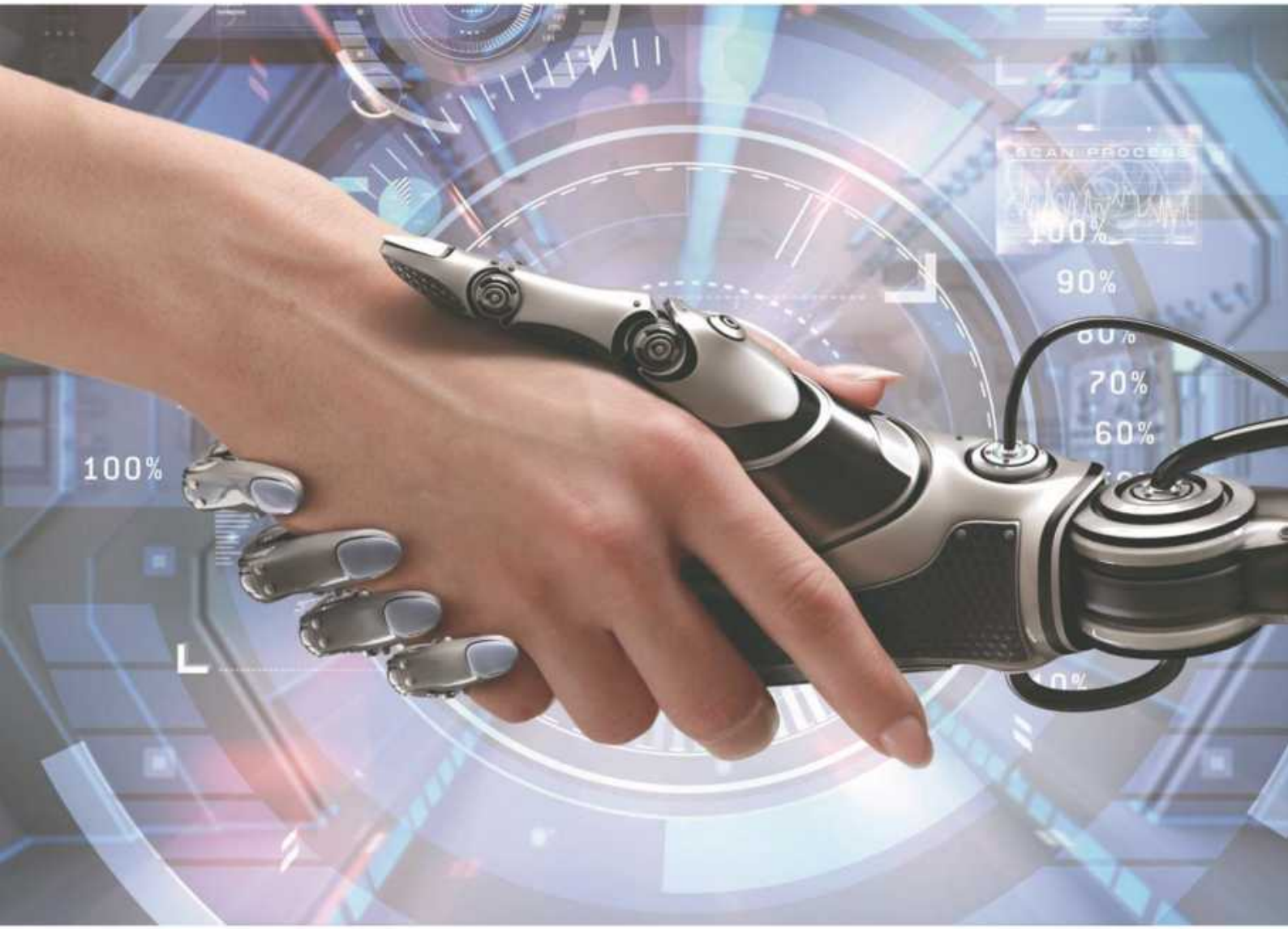
ZTE中兴



中兴通讯AIR Core技术体系白皮书

目录

图目录

表目录

1 引言

在数字化、网络化和智能化转型的今天，移动网络和人工智能(Artificial Intelligence，AI)领域都在发生重大变革。在移动网络领域，5G 已进入下半场，6G 开始启动，5G 应用发展正处于由技术驱动转向价值牵引的关键窗口期，人工智能成为业务创新和提质增效的关键驱动力；而根据 ITU 和 3GPP 的时间表，6G 预计将在 2030 年具备商用能力，通信 AI 一体化作为 IMT-2030 愿景建议书中首次提及的新场景，全球移动网络研究者都在思考未来通信范式从“万物互联”向“万物智联”演进时网络应该具有哪些新的特征。

图 1-1 ITU-R 定义的 IMT-2030 应用场景和关键能力

与此同时，人工智能领域也在发生着翻天覆地的变化。人工智能发展至今，已经经历了三次高潮，两次低谷时期。2022 年 11 月，OpenAI 公司发布的 ChatGPT 以及其采用 Transformer 算法和预训练大模型的生成式 AI 技术，使得人工智能技术发展达到了

前所未有的新高度，并由此迎来 AI 大模型技术拐点和炒作高峰。生成式 AI 技术，具有生成新内容、模仿人类创造力和创新性的能力，使其在众多领域都能发挥重要作用，从而推动了人工智能领域的繁荣和进步。规模创造奇迹，更大的模型带来更高的智能，随着 AI 技术不断发展，千行百业将可以利用 AI 更好地实现运营效率的提升和商业价值的创造，从“数字化”迈向“数智化”。

因此，移动网络和人工智能相互融合、相互促进、相互赋能，将成为今后几年业界主旋律。一方面移动网络可以利用 AI 助力网络优化、异常检测、能效改善、安全增强，多个维度推动移动网络更加智能、创新和高效；另一方面可以通过移动网络构建支持 AI 业务的端到端高效网络，实现云、边和端全场景训练和推理。

本文通过分析核心网当前发展的挑战和机遇，通过引入 AI 技术实现 AI Core 来达成业务创新和降本提效，重点思考新一代 AI 技术在重塑核心网体系架构中的作用，并对核心网未来演进方向和挑战进行了展望。

2 AI Core 发展现状

2.1 拥抱 AI+，业界积极推进核心网智能化

当前移动网络正面临着多重挑战，第一个挑战是用户流失，主要体现在 OTT 以其灵活、快速业务创新、精准个性化服务和低廉费用大幅分流传统移动网络用户；第二个挑战是移动网络业务应用单一和固化，无法满足应用特征泛化、个性化的需求；第三个挑战是移动网络流量增长放缓，基于流量的营收模式遭遇瓶颈，运营商收入增长缓慢甚

至下滑；第四个挑战是自智网络自主决策能力不足，主要体现在网络优化成效和影响、变更影响、重大逃生操作影响等评估不足。

如何通过利用人工智能技术打造 AI Core 应对当前移动网络痛点，加速网络范式转变，业界正在积极行动，持续推进核心网智能化的发展。

首先看一下宏观政策。目前各国政府都在加大对人工智能领域的投资和政策支持，例如欧盟委员会通过了“数字欧洲计划”2024 年工作计划，将为包括人工智能和网络安全在内的数字解决方案提供资金支持；中国在《政府工作报告》中提到深化大数据、人工智能等研发应用，开展 “人工智能+”行动，打造具有国际竞争力的数字产业集群。

其次再看一下标准，在 3GPP 标准持续推进下，核心网智能化正在从外挂智能向分布式智能、原生智能演进。NWDAF 作为外挂智能，从 R15 到后续版本不断发展。R15 的 NWDAF 功能单一，仅支持网络切片负载分析；在 R16 定义了集中式架构，能满足基本数据分析要求；R17 实现了训、推分离式架构，定义了分析逻辑功能（AnLF）和模型训练逻辑功能（MTLF），还构建了支持多 NWDAF 协同的分层智能架构，并引入数据管理框架提升数据采集和分析效率。未来，NWDAF 可能会在分析能力、服务范围等方面进一步拓展，以更好地适应不断增长的业务需求。未来 6G，将在设计之初将 AI 融入到架构、功能流程和协议，提供原生 AI 服务，即将其作为 6G 基本服务对内对外提供。

最后再看一下运营商，全球运营商都在积极拥抱“AI+”。国内运营商积极推进 AI+战略，以中移为例，实现全面 AI 战略，从“5G+”迈向“AI+”，已发布 23 款 AI+产品

及 20 个 AI+DICT 行业应用，5G 新通话成为中国移动 AI+战略产品；国际运营商通过多方合作探索 AI 潜力。在同行合作方面，韩国 SK 电讯、德国电信、阿联酋电信公司和新加坡电信等运营商签署人工智能合作谅解备忘录，组建“全球电信 AI 联盟”(Global Telco AI Alliance)旨在加速电信业务的 AI 转型，借助 AI 驱动模型开发新的业务增长点。在云商合作方面，美国 AT&T 协同 Meta、微软/OpenAI 共同推进基于人工智能的网络优化和故障定位等场景，并已完成商用部署；澳大利亚 Telstra 协同微软/OpenAI，开发基于 Microsoft Cloud 和 Azure OpenAI 自动生成电信网络摘要、生成式问答知识库，并已完成商用部署。Gartner 基于运营商 AI+实践，总结出如下 20 大应用场景，包括体验类如元宇宙业务分身等、效率类如故障检测/网络和资源优化等。如下图所示。

图 2-1 Gartner “AI+”电信网络应用场景

2.2 AI Core，机遇与挑战并存

通过引入 AI 技术，给核心网带来了前所未有的机遇，例如可以帮助核心网催生新的应

用场景和商业模式，提升网络价值，也可以实现网络的自感知与自优化，提升网络的可靠性和用户体验；但是也会伴随着一系列挑战，包括如何确定高价值应用场景、如何获取高质量的数据，以及数据安全和网络安全等。

- 机遇方面：
  - 业务升级：AI Core 推动了人工智能与网络技术的深度融合，如语音、消息等，这种融合可以将传统基础短信、通话业务升级到内生智能的入口式业务，例如通过引入“AI+”新通话实现视频、数据交互、AIGC、智能体 AI 入口、数智人等多种新通信能力，让传统语音业务具有了 OTT 社交软件属性，成为 APP 入口。
  - 运维变革：AI Core 可以借助 AI 大模型、数字孪生等技术，构建 Copilot/Agent 协同的运维新范式，实现对网络的智能运维和优化，提高网络资源的利用率，降低新业务创新周期，提升运维运营效率。
  - 模式转变：通过引入 AI 技术，不仅可以提升网络智能感知能力，构建 5G 差异化服务，还催生了新的商用模式，例如通过引入“AI+”连接来提升用户体验，实现流量经营到体验经营的商业模式转变。
- 挑战方面：

根据 Gartner 报告分析，在电信网络引入 AI 技术有如下挑战，主要体现在如下几个方面：

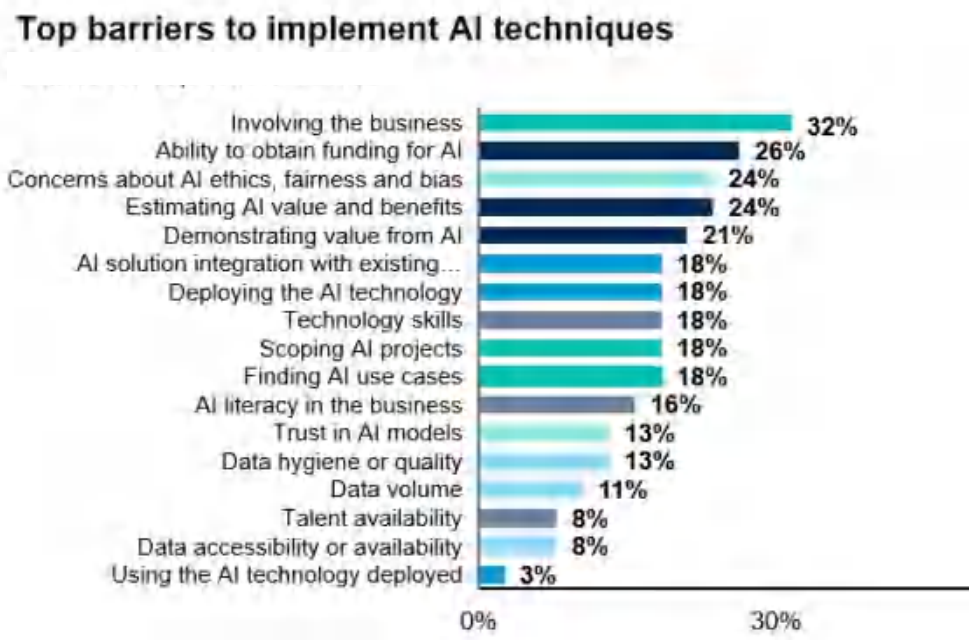


图 2-2 Gartner 实现“AI+”电信网络主要挑战

- 应用场景与商业闭环：引入 AI 技术的场景比较多，但是如何确定高价值应用场景并实现商业闭环成为比较大的挑战，例如网络优化、体验提升等差异化场景带来的价值大小不一。
- 大模型不确定性：AI 大模型缺乏可解释性、可控性，如何确保网络中引入 AI 大模型能力确定和可控。必须有相应技术手段来实现高价值场景的可落地。
- 高质量数据：大模型的能力提升依赖于高质量数据，但往往缺乏高质量数据，主要受限于数据隐私保护、数据获取成本高、数据稀缺性以及数据自身质量和数量不满足需求。
- 数据隐私与安全：AI Core 需要处理大量用户数据，如何确保这些数据的安全和隐私成为了一大挑战。必须采取有效的技术手段和管理措施，防止数据泄露和滥

用。

- 人才短缺与技能提升：AI Core 技术的快速发展对人才提出了更高的要求。然而，目前相关领域的人才储备相对不足，且技能水平参差不齐。因此，需要加强人才培养和技能提升工作，以满足技术发展的需求

综上所述，AI Core 在带来机遇的同时，也伴随着一系列挑战。为了充分发挥其潜力并应对挑战，需要各界共同努力，加强技术创新、完善法规政策、培养专业人才，并推动相关标准和规范的制定与实施。

3 AI Core 技术体系架构和发展路径

3.1 AI Core 技术体系架构

结合 AI 技术的发展趋势，以及当前核心网引入 AI 技术的机遇和挑战分析，我们知道传统的规模扩展方式逐渐放缓失效，核心网需要在业务、连接、运维、网络云基础设施四大领域进行系统化的重构，才能孵化更多业务场景，提升网络价值，发挥网络性能，降低管理运维难度和成本。为此，中兴通讯结合自身转型战略和一线客户需求，提出 AI Core 技术体系架构。

AI Core 技术体系由“三层一域”构成（如下图），分别是 AI+网络云基础设施层、“AI+”连接层、“AI+”业务应用层和“AI+”运维域。其中“AI+”网络云为核心网 AI 应用提供创新的 AI 算力底座，包括计算、存储、网络等硬件资源，以及资源管理和调度，提供裸金属、虚机和容器等多样化实例以及细粒度的资源池化能力， 在此之上搭建训

推平台提供应用 AI 使能能力；“AI+”连接提供 5GC 控制面和用户面智能化，激发流量，增强体验，挖掘数据价值；“AI+”业务提供消息和新通话的智能能力，为用户带来交互式、智能沉浸式新体验，带来无限沟通可能。“AI+”运维通过重塑运维范式，支撑自智网络向无人化演进。

图 3-1 “AI+”核心网技术体系架构

3.2 AI Core 发展路径

AI 技术还在逐步发展中，移动网络也在不断变革，网络建设也需要一个周期。为释放 AI 大模型极致潜力，同时结合网络发展的节奏，我们认为 AI Core 的发展可以分为三个阶段，如下图所示：

图 3-2 “AI+”核心网技术路径

阶段一：“启智”。这个时期最显著的特征是人工智能正从“星星之火”走向“燎原之势”。将 AI 大模型与网络深度融合，不同运营商不同业务线条都有不同的需求，而且涉及到运营商的网络规划、建设、运维、运营与 IT 等众多环节，每个环节的业务场景对模型的精度要求也不同。要找到合适的技术应用场景成为关键，需要产业中的设备提供商与愿意探索创新的先锋客户共同努力。

阶段二：“强智”。当 AI 大模型迈进超十万亿参数量规模，模型能力再次升级，AI 大模型可能真正进入到工业生产的每个环节。同时核心网经过启智阶段的积极探索，已经摸索出一整套经验和方法，可以结合 AI 大模型能力，挖掘更多的网络 AI 场景，实现 5G-A 网络的更进一步变现和 AI 应用的进一步落地，为 6G 核心网内生智能奠定坚实基础。

阶段三：“融智。这个阶段是 6G 内生智能，目标是围绕 AI 大模型和 AI agent 构建 6G 内生智能核心网。该阶段 AI Core 将内生支持通信和 AI 一体化服务，实现自主环

境感知、自主任务生成和自主执行任务的能力，赋能丰富多彩的新业务，支撑社会高效可持续发展。

我们认为，AI Core 当前已处在“启智”阶段，国内外部分运营商正在按照分层思想构建 AI Core；面向中远期，我们应重点攻关“强智和融智”的关键技术，尽快形成行业共识，加速相关核心技术和产业成熟。

4 “AI+”业务：沟通无限可能

4.1 “AI+”消息，打造可信多维智能服务

4.1.1 网络诈骗风险凸显

根据全球反诈骗联盟（GASA）2023 年的年度报告，网络诈骗给全球消费者造成了约 1.02 万亿美元的损失，相当于全球 GDP 的 1.05%。其次，根据全球经济犯罪调查（GECS），51%的受访组织在过去两年中遭遇过欺诈，这是近 20 年来的最高水平。电信网络业务，包括短信和语音/视频通话，是全球使用最广泛的通信手段，全球用户数约为 93 亿，成为网络诈骗的主要目标。

诈骗方通过频繁变换欺诈内容和各种技术手段，使得欺诈不断进化变得更加复杂，伪装形式层出不穷，电信网络诈骗进一步泛滥成灾，不但造成群众巨额经济损失，还造成恶劣的社会影响。传统治理系统主要基于静态规则，无法动态调整和识别新变化，识别能力和效率低，人工成本大幅增加，业务上线周期长，已远远无法满足通信网络的治理要求。

4.1.2 AI+消息带来技术和价值升级

对于电信网络诈骗治理现状及痛点，AI+消息多模态反诈大模型应用，开启从传统治理向 AI 治理模式的技术革新和升级，为新通信可信保驾护航，高效、准确和全面。

根据 Gartner（全球权威咨询公司）市场预测，对话式 AI 作为新兴技术正在迎来广阔前景，理想的情况下，如下图：

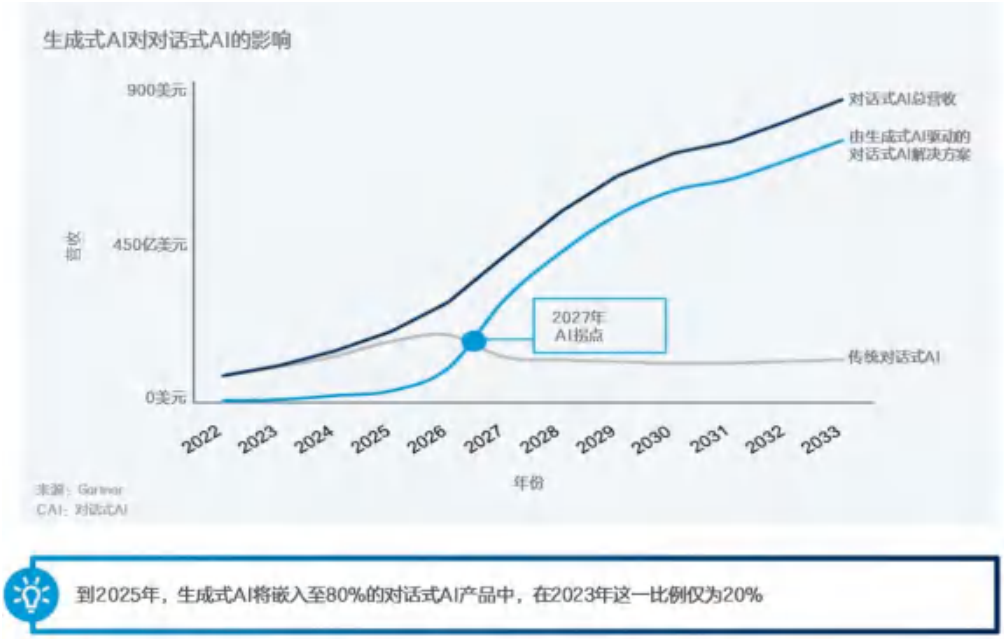


图 4-1 生成式 AI 发展趋势

- 到 2033 年，对话式 AI 平台的营收将在生成式 AI（预生成模型）的驱动下增至 848 亿美元，而 2022 年，这一数值仅为 69 亿美元，年均复合年增长率达 25.5%。
- 到 2027 年，在对话式交互解决方案中，生成式 AI 拉动的营收增长将超过传统 AI 方案创造的营收。技术供应商如果不在拐点到来之前采取行动，将面临失去市场份额的风险。

- 到 2025 年，生成式 AI 将嵌入至 80%的对话式 AI 产品中，在 2023 这一比例仅为 20%。

5G 消息是对话式 AI 的重要入口，对于 ToC，构建高频刚需入口、人人拥有一个专属智能体成为主流发展趋势，ToC 特点是以体验优先，长期发展潜力大的业务。对于 ToB，Chatbot(应用号)是 5G 消息的杀手级应用，国内总数已达 130 万+个，三年内将达数百万至千万个，千行百业 Chatbot 普遍具备强烈的 AI+能力集成和服务提供诉求，ToB 行业 Chatbot 应用普遍付费意愿高，变现快，具备很大的经济价值和发展空间；这些都给 AI+消息带来重要发展机遇。

4.1.3 三层二面，“AI+”消息打造三类服务入口

网络侧消息平台通过引入智能面与消息面无缝融合，形成三层二面的 AI+消息总体架构，如下图。

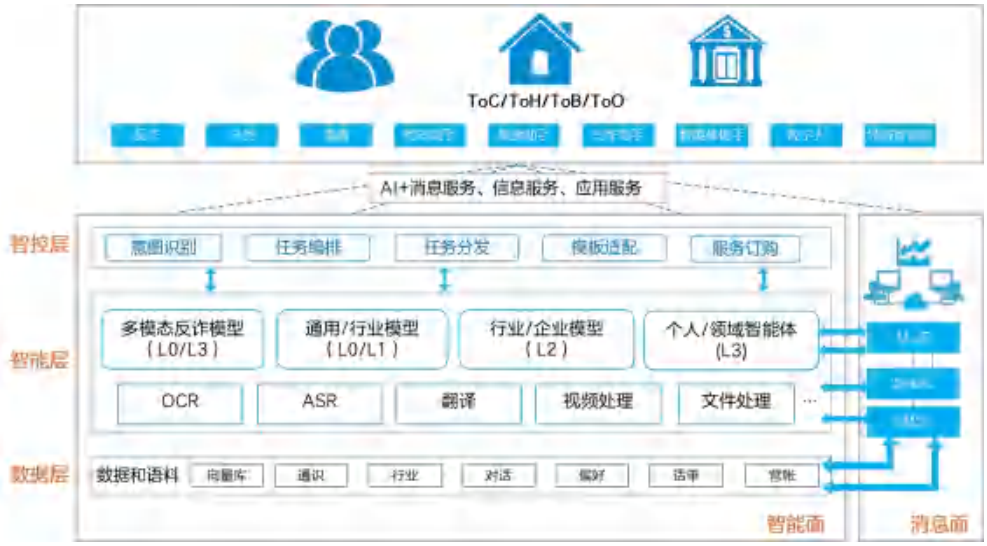


图 4-2 三层二面架构

三层二面架构以智控层意图识别、能力编排为核心大脑，聚合调度各层、各面能力，构建 AI+消息服务入口，面向个人消费者（ToC）、行业与企业（ToB）、家庭（ToH）和其他（ToO）打造“AI+消息服务、AI+信息服务、AI+应用服务”三大类综合服务。

- 三层
  - 数据层：基于网络、平台或营帐日常运营，积累提供原始话单、数据和语料，经过数据处理、清洗为智能层提供核心和长期语料。
  - 智能层：包括大小模型和算法，利用语料训练、精调或持续学习，内生各类 AI 大模型应用，与消息面紧密集成，完成业务服务流程。
  - 智控层：作为前置智慧大脑，具备目标用户意图识别、任务编排、任务分发、模板适配进和服务订购，是 AI+消息综合服务能力的核心控制和直接供给层。
- 二面
  - 消息面：移动通信消息业务和服务能力提供及通道，提供短信、彩信、5G 消息和行业消息等服务。
  - 智能面：作为新增能力面，包括智控层、智能层和数据层，与消息面集成协同完成 AI+能力升级，提供最终用户 AI+消息服务、信息服务和应用服务。

4.1.4 多模态反诈大模型，护航可信消息通信

诈骗消息作为电信诈骗最常见的途径之一，其内容、形式不断变异和升级，以穿透电

信运营商消息监控系统处理，如文本内容上通过组合变异、转义字符、谐音、形近等种种手段突破关键词规则；号码监控策略上，通过海量号码池规避流量和关键字门限；发送方式上通过拨测等方法，一点突破，海量发送；媒体格式上，越来越多的通过图片、音频、视频等方式替代文本发送。传统反诈治理方案识别准确率差、效率低，响应和处理周期长，媒体内容识别困难、语言单一等严重局限性。

AI+多模态大模型反诈治理，如下图所示，以 AI 大语言模型、CV（机器视觉）大模型和多模态大模型为基础，能够识别和处理包括文本、图像、图文、音频、视频和文件等媒体类型，并具备强大的自然语言语义分析、情感识别、逻辑推理、归纳能力大幅提升准确性，多语种能力可识别国际主流语种和地区方言，如中文、英语、法语、广东话、福建话、缅甸语等。

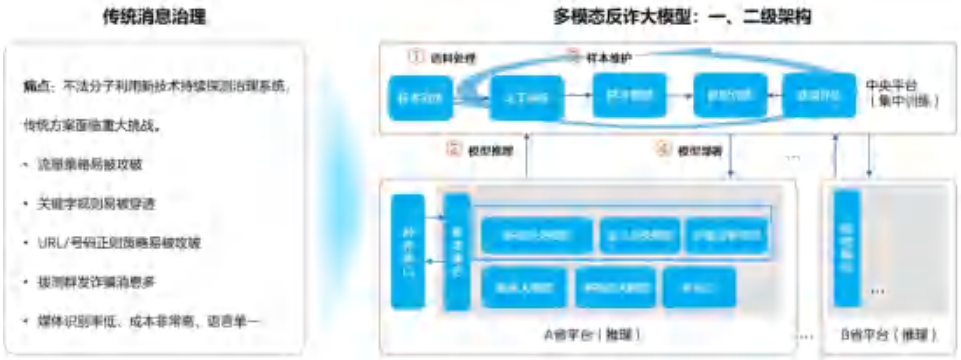


图 4-3 多模态大模型架构

多模态大模型反诈有如下关键技术和特性：

1. 模型训练：B+Z

B 表示基础预训练大模型，Z 代表反诈消息监控适配层，预训练大模型 B 通常十、百亿级参数规模，具备较强的语义分析、情感识别、逻辑推理、归纳能力。

在此基础上，叠加 Z 的“反诈专用提示词+反诈样本微调”模型训练方式，实现诈骗消息在数十万级训练样本情况下，达到高水平的诈骗识别准确率和召回率。

2. 幻觉消除

基于针对诈骗消息识别的“提示词角色引导与输出规约”技术，消除大模型常见的幻觉问题，提升大模型反诈识别的稳定性。

提示词（prompt）方法论框架为 CRISPE，CRISPE 框架由五个部分组成，分别是 Capacity and Role（角色）、Insight（洞察）、Statement（声明）、Personality（个性）和 Experiment（实验），按照 CRISPE 方法论构建反诈识别 prompt，使其能够准确、可靠地执行特定任务，消除幻觉问题。

3. MLOps

基于 MLOps（Machine Learning Operations）机器学习运营，建立工作流程和自动化流程来提高大模型的自动化和运营能力，实现全工具化和任务工作流自动化，样本数据自动导入、自动训练精调、自动部署、自动推理，实现业务流闭环迭代，无需人工干预，保持系统能力的持续更新，大幅节省维护成本和 OPEX。

4. 推理架构

部署模式上可分为训推合一和分布式推理两种架构：

(1) 训推合一架构，模型部署、训练、精调和推理都在单个局点内部署，在节省算力和服务器的情况下，训练和推理可共用算力资源；

(2) 分布式推理架构，可部署成中央训练、分布推理的二级架构，即不需要在

每个分布节点都部署训练、精调模块并进行训练、精调，只需在中央节点一个点集中完成模型训练、精调，再把完成的模型分发各分布节点直接部署和完成推理，可大大降低重复训练、精调的资源、时间和成本。

AI+多模态大模型反诈可达到高水平文本及多媒体识别准确率，远远超过传统方案，大幅降低人工审核成本、降低误拦增加收入、降低用户投诉，带来综合的经济价值。

社会价值方面，多模态大模型应用大幅提升用户的通信体验，打造可信、无忧通信服务，维护通信品牌形象和用户满意度，更多放心使用移动通信业务，在最大化减少人民群众财产损失的同时，为维护社会稳定作出科技力量带来的巨大贡献。

4.1.5 内生多维智能体，开启极智生活

作为 5G 消息杀手级应用的 Chatbot，对话是核心交互方式，目前 99%以上的 Chatbot 不支持 AI 对话能力，除预先配置的菜单或固定关键词外，无法识别用户的任意输入内容和要求，体验严重不足。

Chatbot 如通过自行升级方式支持 AI 智能对话，方式一：自行开发，门槛高，周期长，成本非常高；方式二：调用第三方大模型接口和开放的能力，需要对 API 接口定制开发、联调对接，API 收费，有一定研发成本和长期使用成本，QoS 质量依赖于公网无法保障。

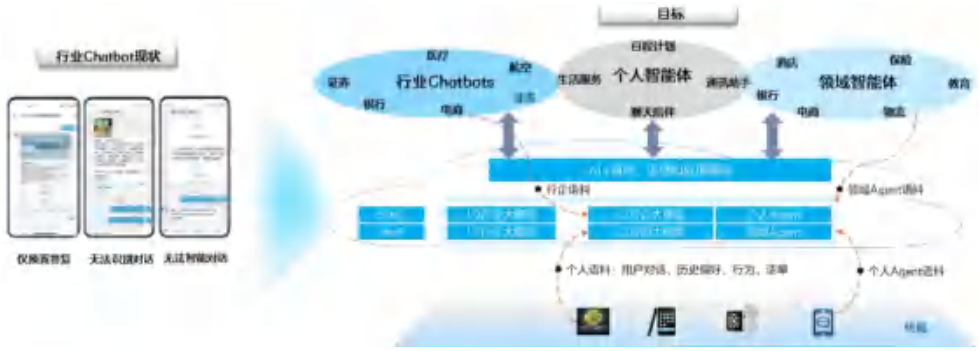


图 4-4　多维智能体目标架构

网络内生消息多维智能体，如上图，面向 ToB 的 Chatbot 和领域智能体服务具备如下特点：无感开启，无需作任何开发、改造；即时开通、可批量开通；QoS 保障，安全性高，数据不出网；成本最低。面向 ToC 方面，基于网络、平台语料和数据，长期积累和学习，提供最懂用户、千人千面、人人拥有专属的个人智能体。

基于 AI+消息网络内生多维智能体，有如下关键技术和特性：

1. 多种类型消息 AI 对话

可支持短信、彩信和 5G 消息多维智能体能力对话。

2. 双向多模态交互

多模态大型模型是结合了大语言模型与对其他模态（如图片、音、视频等）数据的理解与生成能力的模型，整合文本、图像、声音、视频等多种类型的输入和输出，提供更加丰富和自然的交互体验。

3. 互联网搜索生成

在智能体对话中，经常会问到数据有实时性特点的问题，如天气、股价等，这时候就

需要智能体首先能联网搜索获实时更新的结果或最新刷新的官方结果，再把搜索结果作为多模态大模型的输入，补充提示词后，生成得到最终结果。

另一方面，生成式 AI 与搜索的深度结合，推动搜索引擎从“检索”到“检索+生成”的升级。生成式搜索，实现信息智能整合组织、内容创作、个性化内容体验三个方面的体验提升。

4. 文档总结与归纳

利用多模态大模型技术可对文档内容进行自动分析和提炼，生成简洁明了的总结或归纳报告。这种技术可以显著提高文档处理的效率和质量，帮助用户快速把握文档的核心内容，节省时间和精力。例如包括文档关键信息提取、摘要生成、趋势分析和数据可视化等。

5. 意图识别

意图识别的目标是将用户的自然语言输入分类为具体的任务或操作，能够为系统提供准确的上下文理解，显著提高问题解决的效率，减少用户和系统之间的无效沟通，意图识别技术需要支持对用户多意图的理解以及多轮对话管理。意图识别也是智能体能力的入口和基础，通过精准的意图识别才能很好完成智能体及多智能体任务。

AI+消息的重构和升级转型，加速构建基于终端原生入口、码号直达、安全可信的多模态智能体综合入口服务，在商业模式方面，AI+消息实现商业模式升级，从传统消息以按“条”和套餐收费为主的商业模式，叠加和新增差异化 AI 服务收费模式，为下一代消息通信带来新的增长点。

4.1.6 “AI+”消息未来，智驱世界

移动通信消息业务经历了过去 2/3/4G 网络短信时代，现在 5G 网络 5G 消息时代、AI+消息时代，并往未来 6G 网络演进，AI+消息未来以消息智驱世界为目标和方向，以“消息万联、消息智联、消息智驱”为核心发展理念与思路，如下图：

图 4-5 AI+消息未来发展

意图驱动、万物智联是未来网络及 AI+发展的重要方向，而意图驱动的本质是基于消息流的智能驱动，包括意图识别、分解、执行和反馈，因此，消息智驱世界和智驱一切万物是未来的必然发展趋势。

1. 消息万联：未来消息通信将面向“人、万物、应用与行业”等一切万物消息流必达、可沟通、可理解和可执行，消息即服务，沟通人和一切万物；

2. 消息智联：消息与 AI 完全融合，AI+消息应用无处不在，消息即 AI，消息即 AI 服务；

3. 消息智驱：以高准确意图识别为基础，基于终端原生消息统一入口、多维智能体能力，完成各类消息指令、实体事务和生活工作，消息即行动。

消息智驱阶段，以消息泛在、智能泛在、智驱泛在为主要特点，构建通信入口 3.0，支

持泛在消息智能驱动一切服务和应用。

4.2 “AI+”新通话——重塑语音网络新价值

4.2.1 语音收入持续下滑，呼唤实时通信变革

随着 4G、5G 技术的普及、智能终端的广泛应用和用户通信体验需求的不断升级，以市场定位精准、创新快、内容丰富著称的 OTT 迎来了空前的发展，大幅分流着传统通信用户，已造成运营商的传统语音业务量逐年下滑，此种现象在通信发达地区的表现更为明显，见下图。

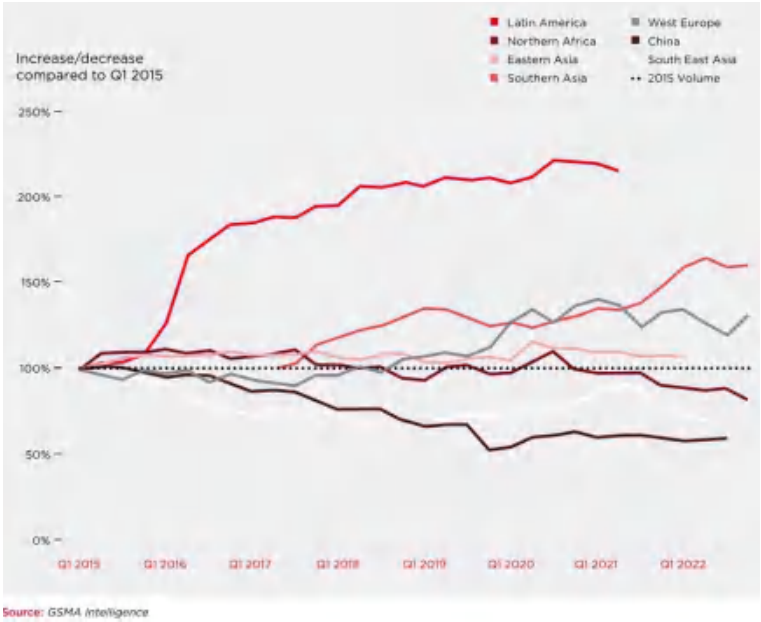


图 4-6 不同区域选定运营商的语音使用量增减百分比

随着 AI 元年到来，AI 技术呈现爆炸式发展，重塑着千行百业。AI 智能手机、OTT 凭

借着各自的优势，纷纷加入 AI 入口的争夺，经营模式快速转向智能化、个性化、虚实共生的体验经营；运营商的传统通信面临着业务重塑、经营模式转变和更大规模的用户分流等多重挑战。

新通话的出现，为全球运营商由用户经营向体验经营转型提供了重要的发展机遇。新通话融合了 AI 与实时通信，催生了多样化、智能化和个性化的应用，如点亮屏幕、明星来电、商务协同、数智人话务员、AI 代答、AI 伴聊、AI 生成数字人、智能体 AI 入口等，改变了人们对运营商的应用传统、单调的形象。

根据 GSMA《The Mobile Economy 2024》报告预测，到 2030 年，全球 5G 连接平均采纳率将达到 56%，通信发达地区此数值将超过 80%，见下图；另外，随着技术的发展和成本的降低，AI 眼镜、AR 眼镜等预计 2027 年将普适化、更适人等，为新通话的可持续发展带来更多的机遇。

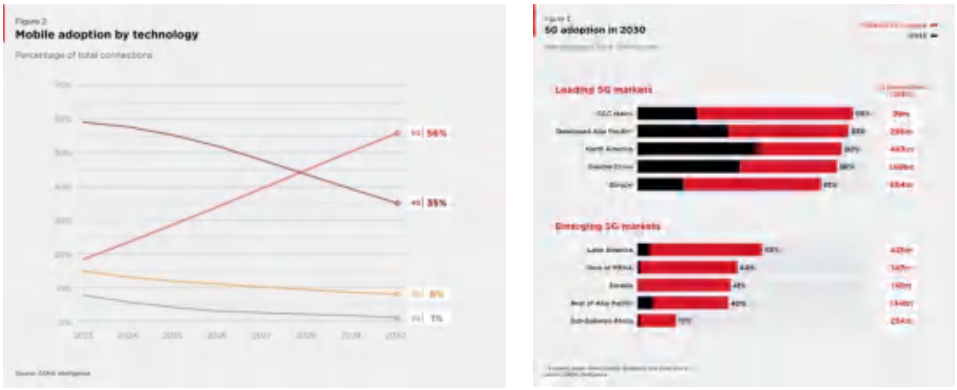


图 4-7　全球 5G 连接占比趋势 （图片来源: GSMA-The Mobile Economy 2024）

新通话具有广阔的发展前景，目前还处在市场培育阶段。移动运营商迎来重大发展机遇的同时，发展新通话也将面临着如下诸多困难和挑战：

- 运营商当前新通话的生态创新能力低，与 AI 终端、OTT 竞争乏力

新通话与 AI 终端、OTT 争夺通信 AI 入口，呈鼎立之势。新通话需要不断地推出多元的智能化、个性化的新应用来满足用户体验升级的需求，运营商的语音通信生态侧重于网络创新，应用创新能力明显不及 OTT 和 AI 终端。

- VoNR 网络封闭、架构复杂、TTM 长，无法第一时间为用户提供最新 AI 体验

VoNR 网络是一个网元多、接口多、流程复杂的封闭的网络，应用上线需要经过复杂的 FOA，无法动态加载和动态编排新能力快速生成新应用，不支持定制个性化应用和加载个人、企业的私有数字资产如私有数字人等，导致架构在 VoNR 网络之上的新通话无法第一时间上线最新的 AI 应用。

- 新通话处于发展初期，盈利模式还未形成，持续发展面临商业闭环难题

可视化、交互式、AI+、个性化的新通话的应用场景目前还在试验、用户群体正在培养，盈利模式需要继续探索。同样，AI 终端、OTT 厂商也进行着 AI 功能盈利模式的不同尝试。

- AI+、可视化、交互式、个性化的新通话引入新的安全、可信、隐私保护问题

AI 对声音、动作、形象的克隆，Data Channel 的 H5 小程序的页面共享、白板等数据交互能力和可视的新通话给用户带来通信便利的同时也引入较大的安全风险。

- 新通话终端普及和用户习惯重新培养需要一个过程，制约着新通话的推广进程

交互式新通话和多模态 AI 入口需要终端支持 Data Channel（简称：DC）通信能力，目前部分品牌的终端已经支持 DC，但是 DC 终端由上市到普及还需要一个过程。

智能、个性化、可视交互的新通话，丰富了体验的同时，也在改变用户在语音通信中

养成的习惯，需要时间去让用户适应。

4.2.2 “四新一普”方案，打造“AI+”实时通信

2023 年 OpenAI GPT 大语言模型拉开了人工智能的大幕，迎来 AI 元年；同年，中国移动正式商用全球首个新通话网络。面对人工智能的爆炸式发展和上述新通话发展实践中遇到的诸多困难与挑战，中兴通讯作为新通话解决方案的领先供应商，率先推出了基于开放生态的 AI+新通话。AI+新通话采用“四新一普”方案，打造“AI+”实时通信。新方案具有新架构、新生态、新体验、新安全和普适性“四新一普”的特点，兼顾稳定与创新，提供智能、个性化的多元体验，具有优秀的终端普适性、兼容性和网络互通性，支持网络开放、可信安全和智能内生。

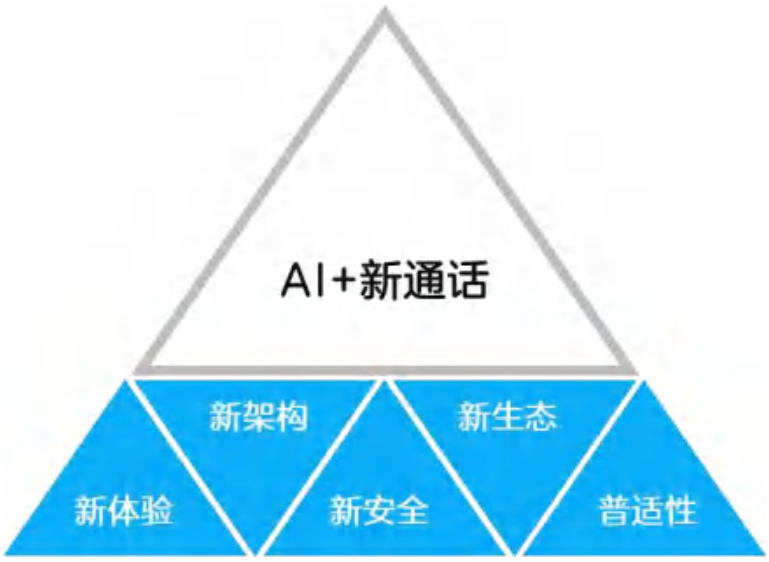


图 4-8 AI+新通话“四新一普”方案

4.2.3 双向开放新架构，敏捷上新 AI 能力

VoNR 网络复杂，导致创新慢和上新周期长。VoNR 网元多、接口多、流程复杂、业务和媒体紧耦合等给新通话应用创新造成很大困难。开发一个新通话应用除了需要设计和研发业务逻辑，还需要进行复杂的流程、接口适配和测试，导致开发周期很长；业务和媒体紧耦合的缺点是消息多、流程和业务逻辑复杂，这些缺点在业务特征单一的语音通信中表现并不明显，但是对于业务特征泛化的新通话，将是严重的；VoNR 是一个封闭的网络，不支持动态加载新能力，增加新能力需要升级系统版本，FOA 过程复杂和耗时；这些问题不仅降低了 AI+新通话的创新速度，还延长了新功能上线的周期。因此，即便移动运营商拥有了与 OTT、AI 终端厂商抗衡的创新能力，也无法第一时间为用户提供最新的 AI 业务体验。

VoNR 的多网元分设，增大了媒体时延和降低了网络可靠性。VoNR 的媒体面与新通话的媒体网关分离、新通话媒体网关与 AI 服务器分设，导致新通话的媒体流需要在多个设备之间传递、拷贝和多次编解码，既增大了时延和东西向带宽，又不利于多模态媒体多流协同；同时还增加了组网的复杂度，不利于网络稳定。

另外，VoNR 网络的封闭架构，阻碍着新通话发展。运营商需要开放新通话网络的能力助力行业开发可视、交互、智能的新通话应用，助力企业提升客户服务质量和效率。

AI+新通话采用“四层五面”架构设计，新架构至简开放、智能自治，支持个性化定制，实现“零”等待、“零”风险的快速上新，赋能千行百业。

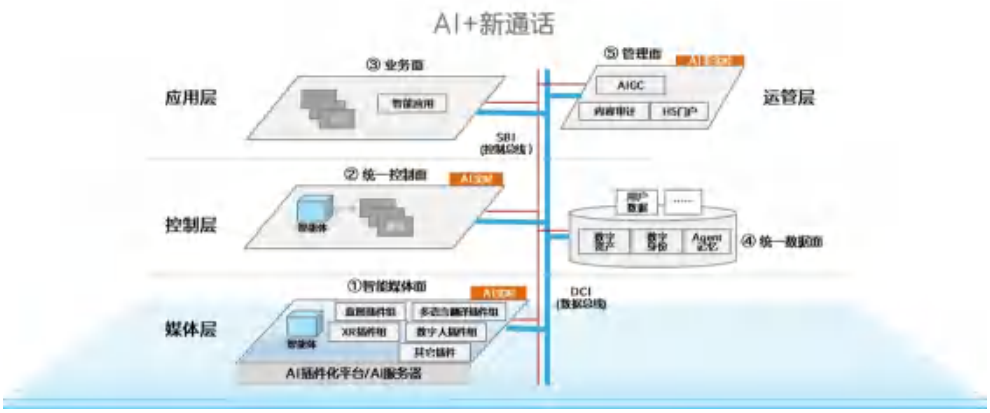


图 4-9 AI+新通话“四层五面”架构

- 智能媒体面：融合媒体和 AI，支持媒体能力、AI 能力的服务化和插件化。
- 统一控制面：融合 CSCF/SSS 和能力开放平台等，支持基本会话、路由等服务化。
- 业务面：智能实现业务逻辑。
- 统一数据面：融合用户数据、数字资产、Agent 长短期记忆、鉴权认证、数字身份管理等。
- 管理面：负责企业和个人数字资产上载、内容生成、内容审计、插件和 H5 小程序管理等。

4.2.3.1 双向开放

新架构支持 AI+新通话网络双向开放。通过插件化架构向 AI 产业开放，打开封闭的传统网络，实现第三方 AI 为新通话网络赋能；通过服务化架构将网络能力和 AI 能力向开发者和应用开放，助力 AI+新通话网络对外提供新通话业务、媒体和 AI 服务，为行业赋能。

图 4-10 双向开放架构

1. 动态可编排插件化技术赋能网络

插件化架构是一个设备和能力分离的开放框架，框架中能力封装为插件，插件与设备、插件与插件之间彼此解耦；通过插件化架构，网络设备被拆分为底座设备和能力 AI 插件两部分，能力插件运行在底座设备之中。底座设备由设备商提供，能力插件由专业的第三方供应商或设备商开发，分工协作。

图 4-11 插件化设备架构

插件化架构分为四层：服务层、框架层、能力插件层和支撑层。服务层实现媒体能力服务化，为 AS 和能力开放提供媒体服务；框架层负责插件管理、编排和执行，形成插件流水线，组合出新的媒体能力；能力插件层是框架层加载到设备中的各种能力插件的集合，包括底座设备自有插件和第三方插件；支撑层是插件的实时运行环境。

图 4-12 插件化框架

统一媒体面采用插件化框架，第三方的大模型、智能体、ASR、TTS、数字人驱动、音视频编解码、空间感知与计算、渲染、多语言翻译和手势识别等新通话能力都以插件的形式部署到媒体设备中，插件生态中的供应商强强联合，推进 AI+新通话快速创新。

插件化动态编排可以在线生成新媒体能力。插件遵循统一的接口标准，标准化的插件根据不同会话和不同应用的需要，在线编排出多样化的媒体处理流水线。流水线支持

多种执行模式，如：顺序执行、复制并行、合并执行和流切换等。基于原子插件，通过动态编排，实现“零”编码动态生成新的媒体能力。另外，插件化流水线大幅简化了媒体控制流程和 AS 业务逻辑。原先复杂的媒体调用流程，被媒体流水线替代；不同新通话应用的媒体逻辑被编排成相应的一条或多条媒体流水线，并分别服务化封装，然后为每个服务分配全局唯一的服务编号（ServiceKey）供 AS 调用；省去了复杂媒体控制逻辑，实现 AS 逻辑轻量化。流程至简和逻辑轻量化，让 AS 新应用创新更加高效。

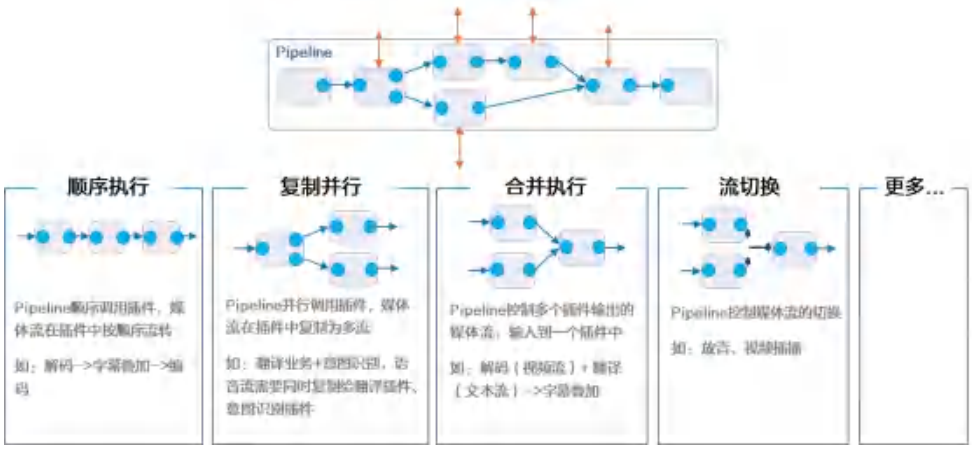


图 4-13 插件动态编排

插件化支持插件动态加载，快速上载新能力。AI+新通话商用上线后，网络增加新能力底座设备不需要升级系统版本，新能力插件通过框架层的插件管理在线动态加载、动态注册到底座设备后实时运行，实现“零”等待、“零”风险上线。插件动态加载兼顾了稳定与灵活，既保证了网络稳定运行，又能快速上线新能力，既省去了复杂的 FOA，又大幅降低了 OPEX。

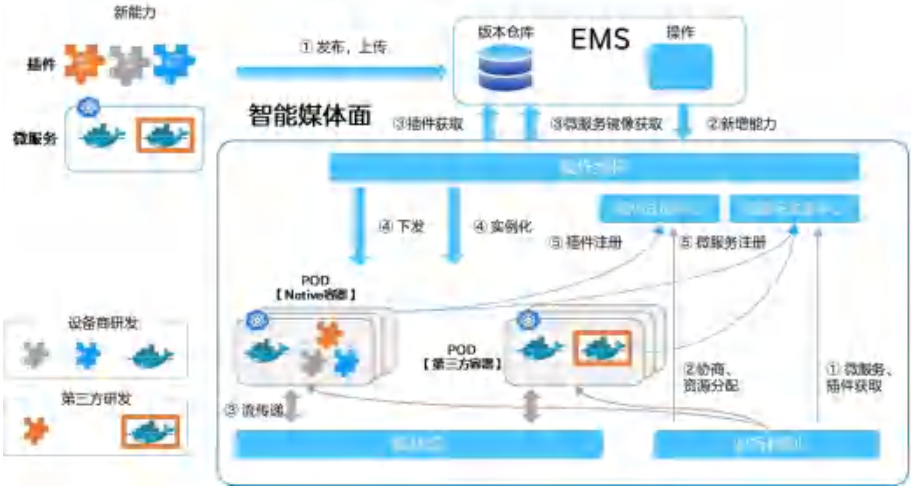


图 4-14 插件化动态加载

2. 多维度无感知能力开放技术，为行业赋能

AI+新通话多维度无感知对外开放网络能力，为行业赋能。AI+新通话通过 VoNR+能力平台和行业网关实现多维度对外开放网络能力，如业务能力和音视频、Data Channel、数字人、AI 等媒体面能力。开放的能力用于行业敏捷开发 2B2C 新通话应用，为用户提供可视、交互和智能的新通话服务，提升用户体验和服务质量，如点亮屏幕、明星来电、新通话呼叫中心、数字人坐席等。

应用与能力解耦实现接口和流程对能力变化无感。AI+新通话能力开放深度融合目标控制流程、服务化和 ServiceKey 媒体流水线技术，实现不同的 2B2C 应用的接口、调用流程的统一化和标准化；不同的 2B2C 应用采用相同的接口和流程调用业务和媒体能力。

另外，企业基于新通话 API 开发新通话呼叫中心具有诸多优势。呼叫中心使用 AI+新通话网络的媒体能力，不用在企业部署媒体设备，媒体处理在运营商网内完成，不用

迂回至企业，既降低了时延，又能帮助企业节省至少 50%以上的传输线路租赁费用和避免呼叫中心媒体设备的一次性投资。

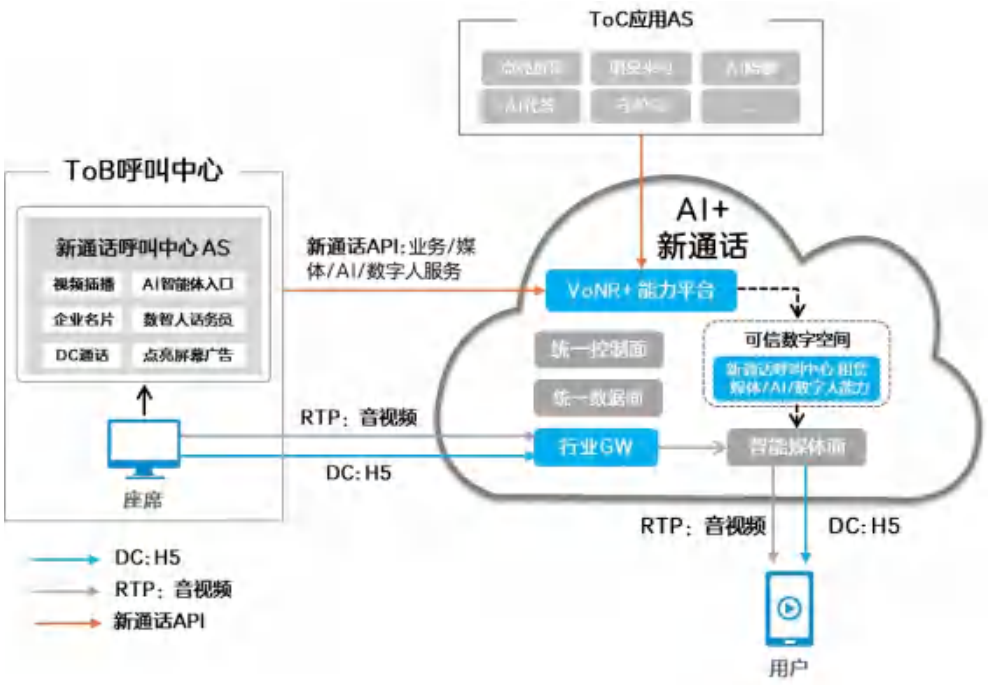


图 4-15 AI+新通话能力对外开放

4.2.3.2 智能内生

智能内生让设备具有了类人的大脑，实现设备智能自治。随着人工智能的发展，通信特征逐渐泛化，不再局限在单一的语音通信、固定的业务流程，如意图通信、XR 通信、数智人、生成式通信等。智能设备自动识别用户意图，自主完成能力插件的智能编排和调用，提供手工流水线无法预先定义的泛化服务。

零训练型意图增强技术实现设备智能自治能力和低成本持续进化。智能体、大模型、RAG 和插件化构成一个零训练型意图引擎，基于此引擎的设备不需要频繁训练，只需要动态增量式增加新的 AI 能力插件、知识库，智能体就能通过高精度检索学习快速找

到解决问题的办法，并调用新 AI 能力插件实现用户意图。另外，定期采用知识库等语料精训智能体、大模型，可以周期性提升意图识别效率和台阶。

多智能体协同与目标控制技术相结合，助力流程至简。传统通信网络中的设备没有智能体，无法自主完成决策；智能设备的智能体具有类脑能力，能够自动识别多模态媒体、信令中显性或隐性的意图，自动完成决策、任务分解，自主调用各种工具插件，与其它设备智能体之间采用简单的目标控制机制协同，不再需要传统通信设备的过程控制、主从控制的复杂流程。

AI 和媒体智能编排技术，一体化处理多模态多流媒体，提升体验。媒体加工和 AI 同一条流水线智能编排、执行，单点集中处理，媒体一次编解码、存算一体，数据无需多次复制和远距离传输，具有多流协同难度小、交互时延低、媒体无迂回、用户体验好和传输成本低的优点，大幅提升多模态通信、XR 通信的体验。

智能内生、训推一体，隐私数据网内闭环，安全可信。意图识别、语义提取等 AI 处理的私有数据资产不出网，在统一控制面、统一数据面、智能媒体面内完成推理，训练和内容生成在管理面实现，可以有效防止用户数据泄露和保护用户的隐私。

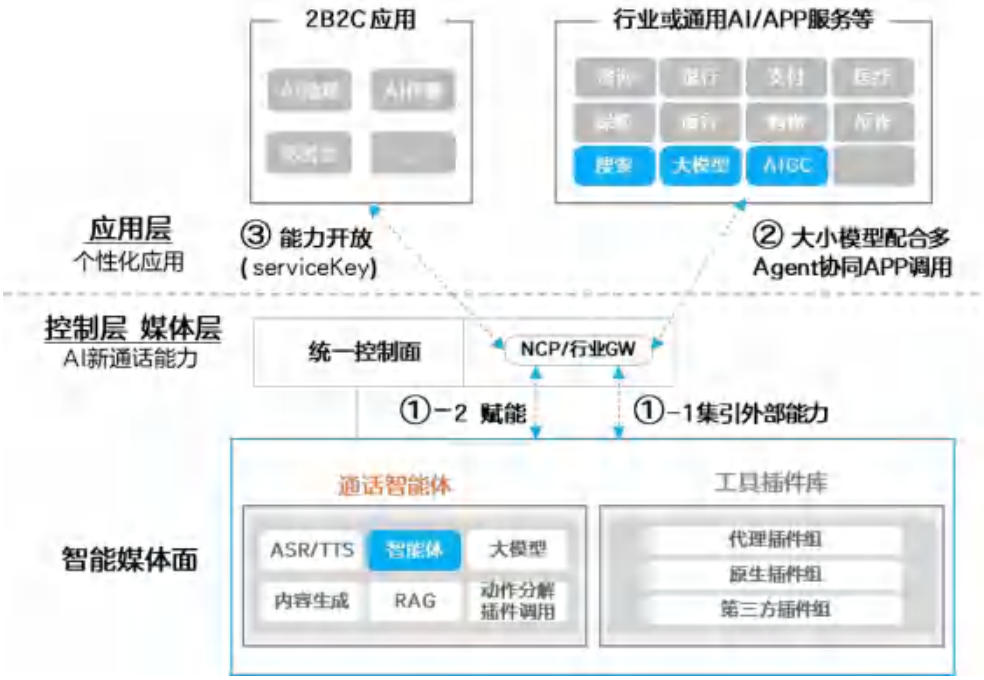


图 4-16 AI+新通话的控制层和媒体层智能内生

4.2.3.3 网络至简

AI+新通话采用多制式深度融合技术，实现网络架构简化、流程至简。至简新架构使得应用创新更加高效、网络更加稳定。

无状态并发业务处理和 NFs 融合技术,实现控制面网元归一和流程至简。高频业务 AS、控制面网元、能力开放网关等多网元通过 NFs 技术融合，简化组网；网元归一屏蔽了复杂的接口和调用关系，对外提供简洁高效的路由、业务等能力服务。无状态并发业务触发、编排、冲突管理和调用等新技术，降低了多个业务之间的耦合度、简化了流程。控制面统一，应用与网络解耦，让应用创新更加高效。

智能一体机技术，融合了智能与媒体，实现媒体面集约化部署。高通量、强计算、低

时延、边网协同的 XR、AI 等 AI+新通话应用需要 AI 和媒体设备边缘部署，与终端之间的传输距离越短越好。智能媒体面采用一体机技术，融合裸金属容器、插件化、服务化技术，实现 AI 网元、媒体网元集约化部署为一个网元，具有物理网元组网简单和资源池扩展灵活的双重优点。另外，智能媒体面设备内置 GPU，实现 AI 和媒体单点处理和深度融合。

泛在数据管理技术，标准化和融合用户数据、AI 数据和数字资产的管理与存储。AI 数据和数字资产的管理、存储的标准化，是 AI+新通话普适性、个性化的数据基础，为数据和新通话用户在不同设备、不同地区、不同运营商网络之间共享和迁移提供了保障。数据融合提供一站式数据服务。统一数据面采用分布式计算和存储技术、向量数据库等多模数据存储技术和 RDMA、文件传输等多种传输技术，通过 SBI 控制总线、DCI 高速数据总线为网络提供全网数据的生产、消费等全生命周期的管理、存储和数据增值服务。数据管理提供用户鉴权、数字身份认证和业务签约等管理服务，数据存储提供 AI 数据、数字资产、用户信息等数据的采集、处理和存储等服务。

至简新架构支持与现网兼容和互通。新架构引入融合、服务化的思想，来简化网络架构、流程、业务逻辑和减少接口数量。新架构实现网络至简的同时，支持与现有的传统网络的兼容和互通。

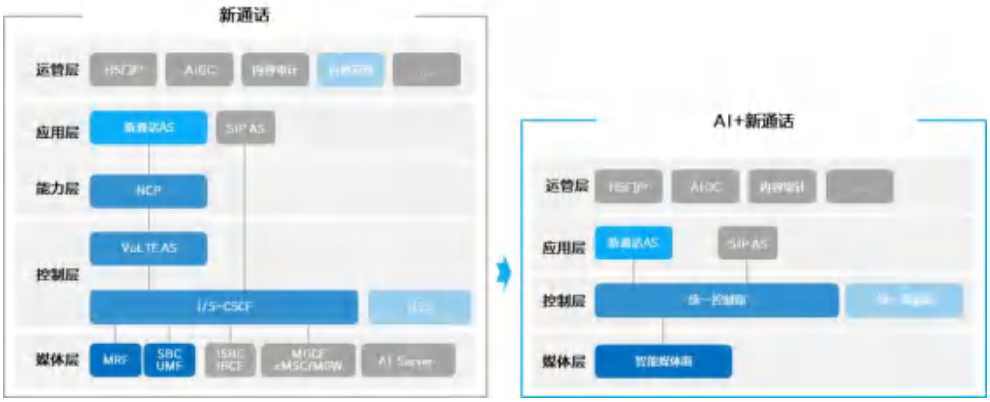


图 4-17 AI+新通话的至简网络

4.2.4 两大产业链，构建快速创新生态

AI+新通话生态设计了两条开放产业链，提升创新能力和效率。AI+新通话融合通信和智能，涉及媒体加工、ASR、TTS、多语言翻译、智能体、NLP/CV/多模态大模型推理和训练、联邦学习、图片/视频/文本 AIGC、数字人驱动和生成、化身、3D 数字人、XR、裸眼 3D、渲染、AI 反诈、内容审计、分布式计算、H5 小程序等众多高新技术领域；同时，AI+新通话打破了 CT 和 IT 的界限，除了提供可视、交互式的通话功能外，还支持通过 AI 入口调用 IT 应用，提供导航、订票、天气、购物、检索、咨询等服务。目前的语音生态只是 AI+新通话生态的子集，庞大的新生态中体验创新需要多个行业分工协作，高效的创新离不开科学机制的保障，为此 AI+新通话设计了两条端到端开放产业链：插件化产业链和 H5 小程序产业链。另外，AI+新通话支持 CICD、灰度升级和提供数字孪生平台，快速孵化新应用。

1. 插件化产业链

生态圈中的不同厂家的 AI、数字人、XR、多语言翻译、音视频等能力以插件或容器的

形式发布，通过插件工厂组装测试后发布到运营商的运管中心，按需部署到 AI+新通话的网络，部署过程无需网络设备版本升级。

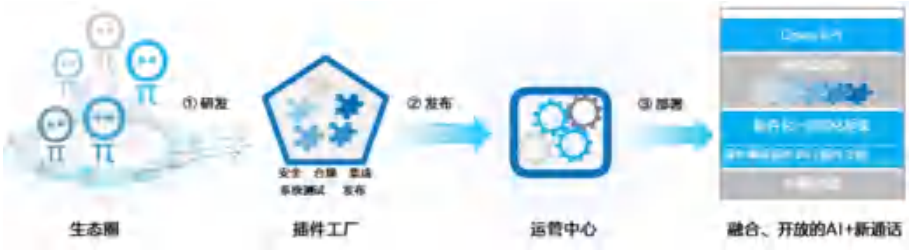


图 4-18 插件化产业链

2. H5 小程序产业链

新通话无需终端安装 APP，部分新应用需要终端和 AS 协同完成，终端侧的新增功能以 H5 小程序的形式通过 Data Channel 自动下载到终端运行。H5 小程序、H5 新应用 AS 开发商入驻 H5 生态，新产品通过 H5 工厂测试后发布上线。

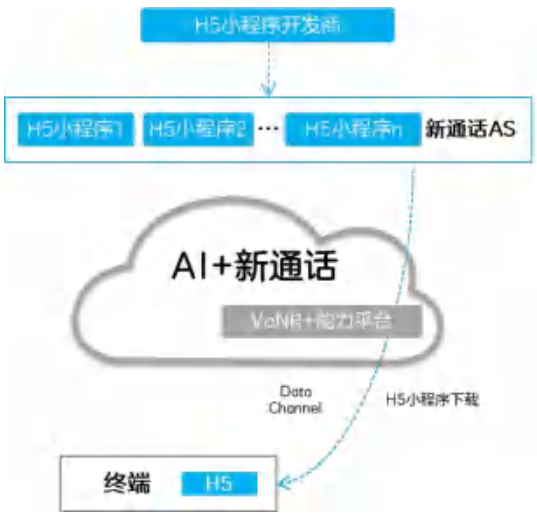


图 4-19 H5 产业链

4.2.5 三大智能新体验，助力新通话商业闭环

AI+新通话基于插件化生态、H5 生态和开放智能的新架构打造出“可视、交互、AI+、个性化”的 2B2C 新体验。做实 VC/DC 通信，与数字人驱动、数字人生成、智能体、多语言翻译等 AI 能力深度融合，推出“单向视频+”、“AI 入口”、”AI+ToB 新通话”等端到端方案，构建内容变现、ToC 应用智能、ToB 智能体交互式商业通信闭环服务的盈利模式。另外，AI 翻译、AI 商务速记、AI 手势识别、趣味通话等 AI 应用，也为用户提供了有趣、实用的体验。

4.2.5.1 多要素融合单向视频增强技术，智能变现视频内容

单向视频+私有数字人、单向视频+广告，不改变用户通信习惯的同时，提升用户可视体验，并且保护用户隐私。个人或企业通过 H5 portal 上传的图片等素材，由运营平台的 AIGC 生成私有数字人、广告素材，与单向视频+流水线配合，推出通话前、通话中的个性化可视服务，实现视频内容变现，如：企业名片、城市名片、企业数字人、个人语音驱动数字人等。

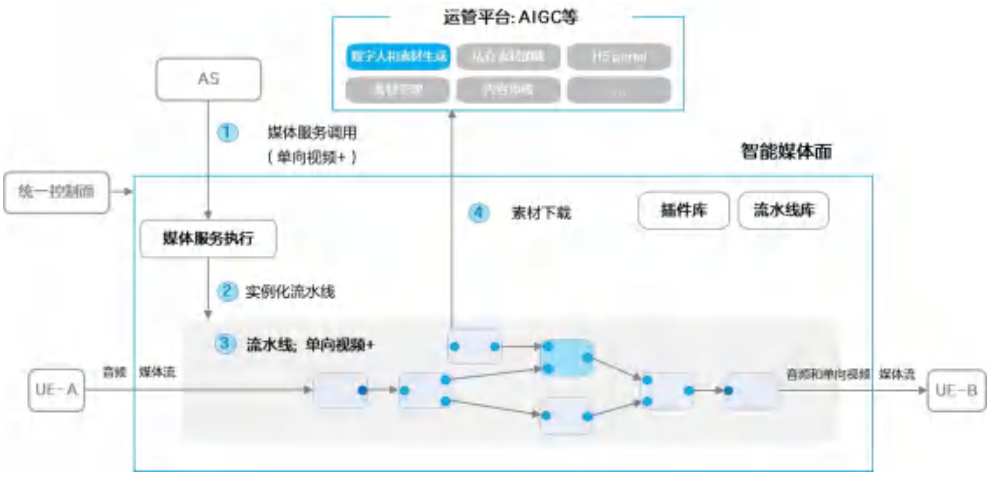


图 4-20 AIGC+个人语音驱动数字人应用

4.2.5.2 AI 入口多模态多 Agent 协同技术，业务智能重塑

AI 入口智能重塑业务，拓展通信边界，提供智能化、人性化通信应用，如：AI 助理、AI 伴聊、语音智能体入口、APP 服务入口、企业呼叫中心入口等服务。

AI 助理支持在用户忙、不在线、免打扰时，代接电话和处理广告、骚扰、快递、诈骗等电话，提升 ARUP 值；AI 伴聊支持人与 AI 之间的通信，如明星来电，提供情绪价值，服务粉丝经济、银发经济。

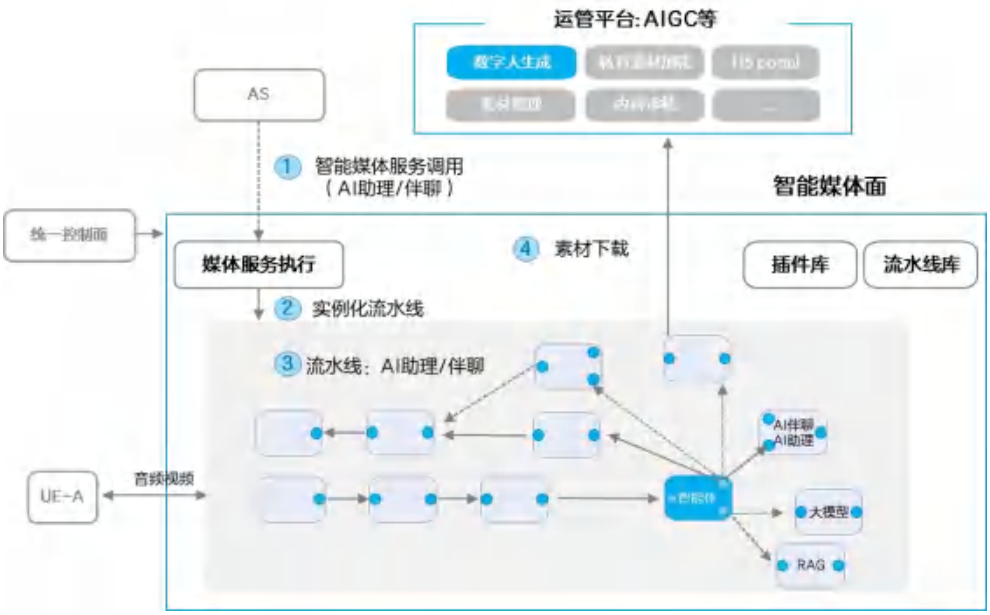


图 4-21 AI 助理、AI 伴聊

APP 服务入口提供一种 APP 使用的简单模式。用户输入意图，入口自动输出结果，屏蔽 APP 复杂的操作，提供管家服务、高效服务，服务不愿意使用 APP 的人群，如银发族。

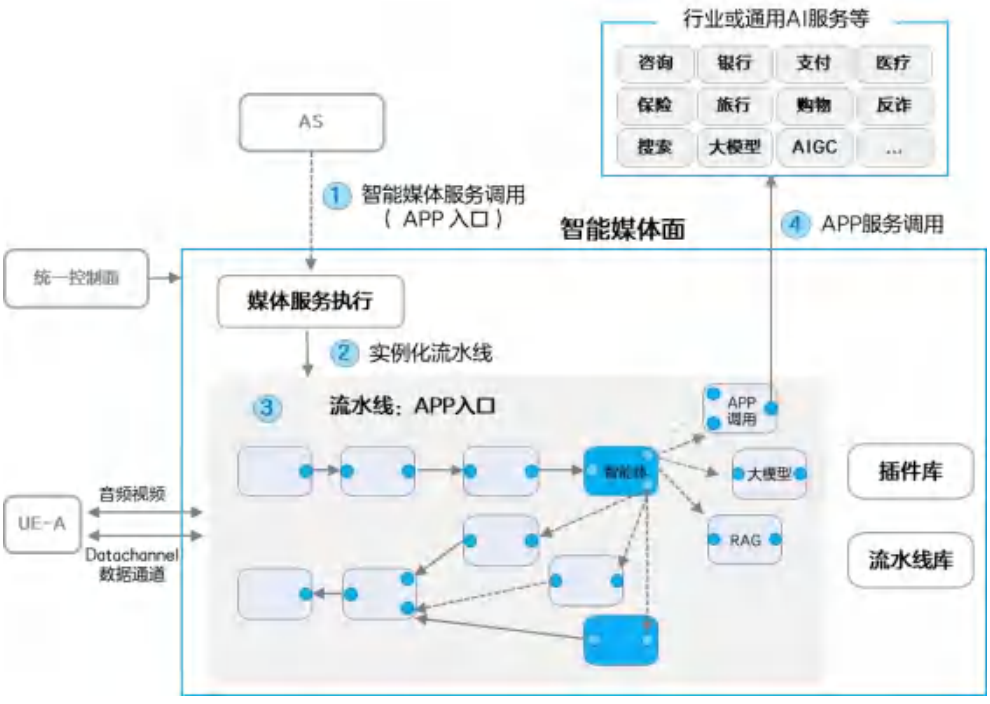


图 4-22 语音智能体入口和 APP 服务入口

4.2.5.3 AI+ToB 新通话数智化呼叫中心技术，提供智能体交互式商业服务

运营商在 B2C、工作、陌生人通信中占据主导地位，并且新通话 AI 能力全、普适性和互通性好，与 AI 终端、OTT 相比，新通话在 ToB 市场中占据绝对优势。ToB 新通话具有“听、视、触、意图、数字人”全方位能力，DC+H5 小程序通信触摸式闭环，助力销售和服务；数智人坐席 24 小时在线，降低服务成本；单向视频，提升沟通效率、产品宣传和保护用户隐私等。

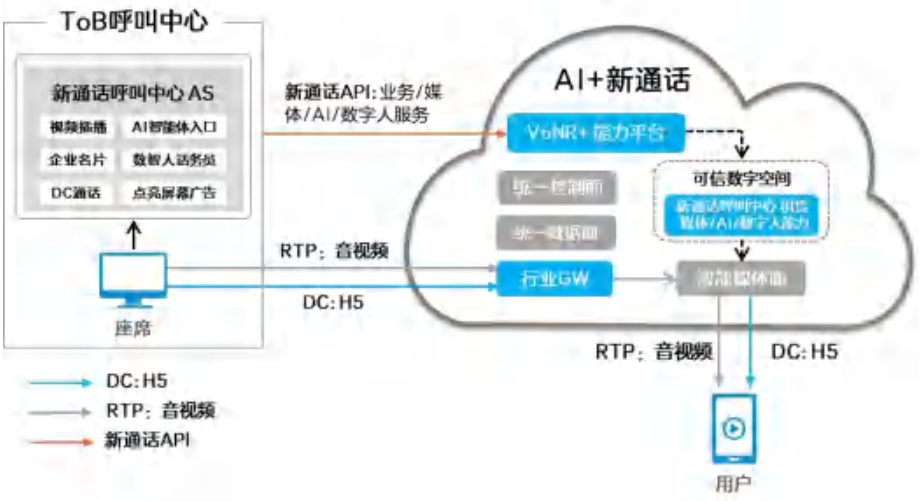


图 4-23 AI+新通话呼叫中心

4.2.6 六边形新安全机制，打造盈利可信锚点

可信是新通话的锚点。新安全从可信空间、内容合规、通话安全、隐私保护、传输安全和运行安全六个方面打造端到端六边形安全机制，守护 AI+新通话安全。AI+新通话的可视、交互、AI+和个性化的应用，相比语音通信，具有海量的私有数字资产需要保

护，AI 生成和 AI 克隆让用户难辨真假，可能导致电诈数量上升，交互式 DC 新通话方便了通话中交易操作也让电诈有了更多可乘之机，可视通话也可能会不经意泄露用户隐私。ToB 应用是新通话的关键盈利模式，可信是交易的前提，合同签订、开户、签名、支付等，都需要新通话提供端到端可信安全机制。

图 4-24 AI+新通话新安全六边形

AI+新通话提供端到端安全保障。新安全中，私有智能体、大模型、数字资产运行在专用的可信数字空间中，内容不出网，有效防止个人隐私泄露；数字身份认证、水印等为通信双方提供可见的可信服务；单向视频，保护用户隐私；私有资产、会话信令、多模态媒体加密传输；内容审核平台，负责对生成的内容、通过 H5 portal 上传的个人或企业素材的内容进行合规审计；部署 AI 反诈大模型插件，提供守护宝应用，守护新通话通信安全。语音通信是运营商的基础通信业务，AI+新通话提供 Bypass 方案、智能流控、应用防火墙等容灾安全机制，保障新通话安全运行的同时，确保不影响语音通信的稳定。

图 4-25 AI+新通话端到端安全架构

4.2.7 普适性新通话，加速实时通信价值重塑

AI+新通话的不同应用对终端有不同的能力要求。视频新通话、交互式新通话，分别需要终端支持单向或双向视频能力、Data Channel 能力。为了不改变“电信运营商应用不需要安装 APP”的用户习惯，交互式新通话普及需要等待支持 DC 能力的手机大量上市。全球通信市场发展不平衡，针对不同发展阶段的通信市场、用户习惯、两种能力终端的普及程度，我们设计了不同的发展策略和配套的普适性应用，加快新通话的发展，重塑实时通信价值。

1. 先视频后交互、先单向后双向

为了不改变用户的语音通信的习惯，可以通过“单向视频+”多种应用，来培养用户视频通话的习惯，如点亮屏幕、主叫名片、视频插播、趣味通话、ToB 视频菜单式新通话呼叫中心等服务。

为了提升用户使用视频通话的粘性，单向视频与 AI、内容生成融合，推出个性化数字人生成、语音驱动数字人、数字人坐席、明星来电（陪聊）、AI 代答等服务。

2. 先意图交互，后 DC 交互

DC 手机普及需要时间，DC 交互新通话应用短时无法大范围推广。交互式新通话是 ToB 通信的重要需求，意图新通话中通过语音 AI 入口的智能体来代替 DC，智能体根据用户语音输入的意图自动完成各种 APP 调用，并把网页等 APP 返回的结果转换为单向视频发给用户，交互简单高效，同时免去复杂的 APP 操作。

DC 手机普及后，通过 DC 交互，可以提供拖拽、触摸式应用，ToB 新通话商务操作将更加便利。另外，意图和 DC 配合可以提供多模态超级智能体 AI 入口服务，多模态交互让通信更加高效和随心。

4.2.8 下一代实时通信，挑战及发展建议

AI 和通信融合、沉浸多感、泛在连接、虚实共生已经成为 6G 发展的共识。面向 6G 的下一代实时通信是一个由数智人、化身、XR、意图通信、AIGC 等智能应用和 AI 眼镜、AI 手机、XR 眼镜、具身智能等智能设备组成的类人智能、沉浸、虚实共生的通信。下一代实时通信正处于市场培育期，面临诸多挑战；AI+新通话只是语音通信向下一代实时通信演进的一个中间阶段。目前，数字人、意图通信、AIGC、XR、具身智能的算力成本太高、产业链不成熟、技术分工界面不明确，DC 终端还未普及，XR 终端的关键技术和人性化需要继续攻关，XR、化身、AI+网络、DC 终端 SDK 等标准正在制定，下一代实时通信需要定义新的架构和标准，智能应用盈利模式尚需探索，

用户新习惯培养需要一个过程等，制约着下一代实时通信的发展进程。

根据当前标准现状、产业链和技术成熟度、终端普适性和用户习惯培养过程，下一代实时通信的演进可以大致划分为三个阶段：可视阶段、意图阶段和泛在阶段。~2025 年为可视新通话阶段，重点发展单向视频通信，不改变用户语音通信操作模式，逐渐培养用户视频通信习惯；2026-2027 为意图（AI+）新通话阶段，打造智能体验，重点发展 AI 入口、数智人等 AI 应用，重构语音价值；2028~泛在实时通信阶段，聚焦沉浸通信、泛在实时通信（URCN：Ubiquitous Real-time Communication）。

图 4-26 下一代实时通信演进

下一代实时通信的可持续发展，离不开引领性的标准、端到端完整的产业链、开放的生态、明确的分工。通信的升级换代不可能一蹴而就，需要分阶段进行；富集大量 CT/IT 高新科技、产业链更长、生态圈更大，转向体验经营的下一代实时通信更需要标准先行，来协同生态圈中千行百业有序发展，打破目前 AI+高科技技术壁垒，打造普适性新通话；随着 AI/XR、VC/DC 终端的普及，下一代实时通信将代替语音通信成为日常的基本通信。

5 “AI+”连接：挖掘数据价值

5.1 5G 进入平稳期，后继发展乏力

自 2018 年 5G 建设启动伊始，5G 建设已从高速发展期步入高质量稳定发展期。2024 年二季度，全球 5G 用户总数达到 18.7 亿，东亚地区（中日韩）5G 用户规模最大，达到 10.57 亿，其中，中国 5G 用户数达到 9.27 亿。北美地区 5G 用户数约 3.17 亿，欧洲地区 5G 用户约 2.23 亿，其他国家地区 5G 用 户数约 2.81 亿。

然而，5G 带来的纯流量红利消耗殆尽，并且随着视频技术的发展，单用户 DOU 在未来不增反降，如视频编解码 H266 相对于 H265，码率下降 49%，带宽需求降低；1080P 高清业务受投资回报比影响，普及率被压制。同时，5G 正迈入 5G-Advanced 阶段，新技术的引入，将赋能低空经济、车路协同等新兴领域，拓展更广阔的市场空间，但在敏捷性、可拓展性等方面也对网络提出了更高的需求。此外，以 OpenAI 发布的 ChatGPT 为标志，人工智能迈入新的阶段，大模型的应用日益广泛，深刻影响着包括电信行业在内的各行各业，既带来了新的发展机遇，也提出了前所未有的挑战。

图 5-1 2018－2023 年移动数据流量与收入发展情况（来源：工业和信息化部网站）

- 收入增长停滞，需挖掘新的增长点
  - 流量增长放缓，基于流量的营收模式遭遇瓶颈，运营商收入增长缓慢甚至下滑。
  - 新兴业务未能为运营商带来应有的商业价值。
  - 终端用户消费模式固化，不利于新业务的推广。
- 管道差异化能力欠缺，精细化经营困难
  - 高价值用户体验缺乏针对性保障，业务差异化服务未能体现。
  - 运营商不了解用户业务喜好，难以精准营销，网络价值降低。
  - 运营商缺乏用户体验数据，无法主动优化网络。
- 通用设备面向公网，无法契合行业多样化需求

- 工业领域需要专用工业 UPF，满足确定性要求，并实现智能化编排。
- 低空产业需要内置 AI 算法和算力的 UPF，实现通信和感知的融合。
- 智慧交通领域需要独立的车联网 UPF，支持上行超大带宽和下行超低时延，并实现 AI 动态编排。
- 低时延、丢包敏感类业务在移动网络中体验差，阻碍 5G 向行业拓展。

- ICT 技术融合不足，亟需提效降耗
  - 资源利用率低，能耗高，不符合绿色发展路线，需要引入新技术节能降耗。
  - 云化技术和 AI 技术在移动网络中的融合度不够，资源编排和分配主要依赖静态规划。

- 信令风暴频发，构成巨大风险
  - 5G 网络分层解耦，架构相对复杂；终端种类多、物联网终端数量庞大；新业务、新应用层出不穷，信令消息激增。
  - 运营商虽已部署预防方案，但效果评估不足，应对效果不佳。

5.2 三层架构四种能力，赋能智能连接

在 5G 核心网中引入 AI，能够有效应对当前的诸多挑战，给连接注入新活力。 通过引入 NWDAF 为大脑，PCF（AM、SM）为策略支撑点，各个网元内置的 AI Engine 为执行体，协同无线和终端，来构建端到端的 AI+连接。

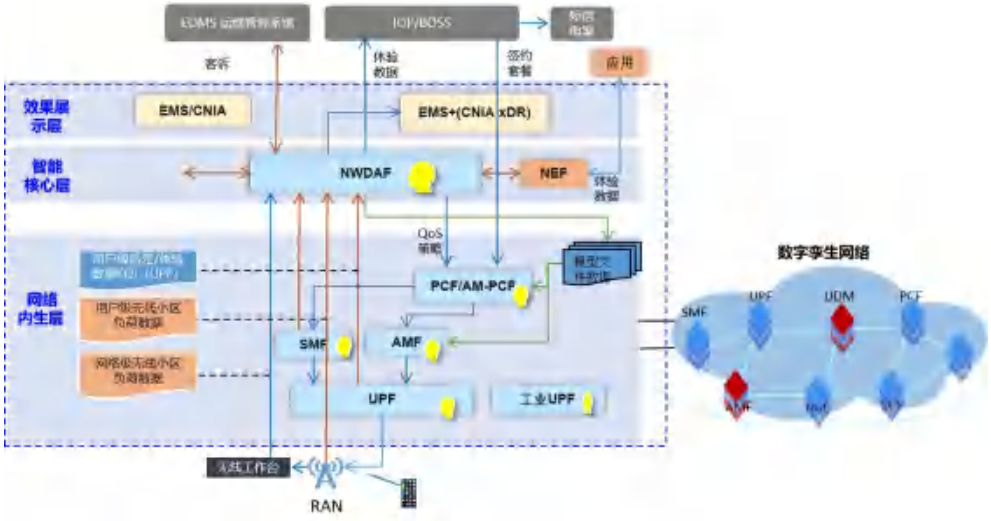


图 5-2 AI+连接智能“3+4”架构

AI+连接智能网络采用 3 层架构：

- 网络内生层：对应核心网网元，在当前的核心网网元功能基础上集成智能化处理单元/功能，并为功能业务应用场景部署独立的工业 UPF；同时，采用数字孪生技术，孪生一个虚拟的数字网络层。
- 智能核心层：包含 NWDAF 和 NEF， NWDAF 作为智能面大脑收集网络内生层网元上报的数据，并基于 AI 算法执行预测以及策略决策，决策后的策略下发到网络内生层执行，形成闭环；
- 效果展示层：包含网络运维展示界面，展示智能化执行效果。

增强的 4 种能力包含：

- 商业运营能力：从传统的围绕流量的经营模式，拓展到体验经营模式；
- 多行业支撑能力：网络能力增强，可以提供更多的能力（比如确定性能力）和资

源（比如算力资源），可以更好的为行业服务。

- 多技术融合能力：各种技术结合 AI 后，更好的发挥各种技术的优势；
- 网络安全能力：安全和 AI 结合，让网络变得更有韧性。

5.3 流量经营到体验经营，创新连接商业模式

一直以来移动网络通过静态签约来为用户提供不同等级的服务质量保障，然而，随着技术的进步和各类应用的激增，一方面用户需要更具针对性的业务体验保障，并愿意为此付费，例如，直播用户需要在热门地点进行无卡顿的高清直播；电竞用户，需要降低时延，在竞技时快人一步；商旅用户为了高效的沟通，需要更好的通话效果和数据传输速度；另一方面运营商需要准确地感知不同场景的用户需求和网络状况，动态调配网络资源，实现价值变现。因此，针对网络中不同等级的用户，不同业务应用，不同的网络场所提供差异化的体验保障，包括：

- 对于网络中高价值用户提供高等级的体验；
- 对于网络中 Top N 的体验敏感应用（比如游戏，直播等）作为开展体验保障的重点业务，推出各种保障套餐，供用户选择购买；
- 对于用户集中的特定场所（比如演唱会，体育馆，高铁等），提供针对特定用户（比如安保，VIP 等）的保障；

图 5-3 体验经营示意图

1. UPF 基于内置 AI 提供两大关键能力：智能软硬协同业务识别技术 和 全息 KQI 体验度量技术；

2. NWDAF 基于收集的数据（比如来自智能 UPF 的用户行为、体验和质差数据，来自 RAN OAM 和网元的负荷），通过多维度跨层跨域画像技术，实现如下能力：

   (1) 业务画像：帮助运营商发现网络中热门应用以及各种应用的体验数据，辅助运营进行业务套餐设计

   (2) 用户画像：帮助运营商寻找目标客户，精准向用户推荐其喜好应用的体验提升套餐

   (3) 网络画像：帮助运营商分析网络资源使用情况，尤其是重点场所（人员集中区域）的网络拥塞情况，帮助运营商更好的运营体验保障套餐；比如在重度拥塞区域部署多频基站，通过 RFSP 将保障用户分流到特定频段，避免建立 GBR 专载进一步挤占普通用户的资源。

5.3.1 多维度跨层跨域画像技术

多维度画像可以辅助运营商更好的开展体验经营，比如运营商期望针对商旅人士开展一个包含某种 Top 视频业务的体验提升套餐；则可以针对网络中的用户，基于移动性信息画像发现经常大范围移动的商旅人士，进一步分析用户常使用的 Top 视频业务，以及视频业务使用体验信息；精准发现某些商旅人士喜欢使用某一款视频，且由于其经常出入人口集中的 CBD 区域导致出现了体验不好；针对画像出来的这部分用户群，可以针对性的触发邀约短信，邀请其签约体验提升套餐，提升其观看视频的体验。

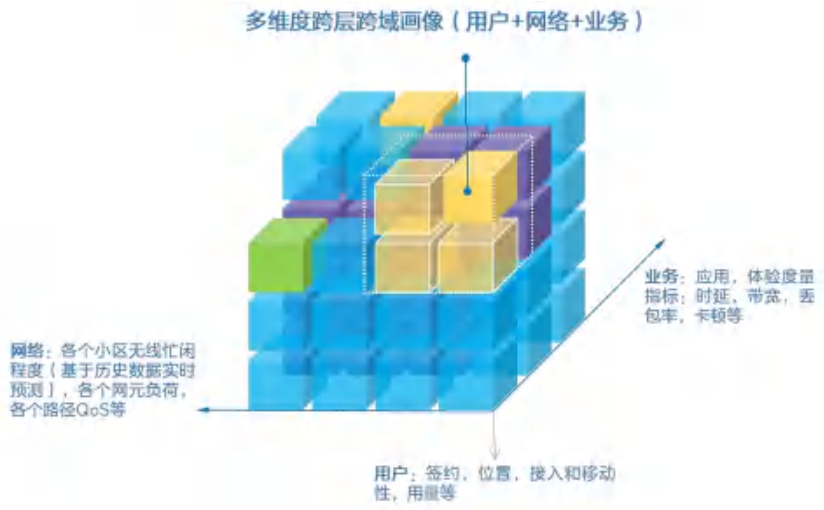


图 5-4 多维度画像

多维度跨层跨域画像，包含了用户维度+网络维度+业务维度，画像后的信息体现了特定用户在特定场所下使用特定业务的体验。为了完成精准的画像以及基于画像的体验套餐运营，需要跨越跨层协同：

- 无线和 NWDAF 之间的数据采集，NWDAF 从无线侧获取到无线拥塞信息以辅助完成无线网忙闲维度的画像；

- NF 和 NWDAF 之间的数据收集，NWDAF 从网络层收集用户使用业务的体验数据以及用户位置，移动性等数据，辅助完成用户和业务维度的画像
- NWDAF 和运营域的接口，NWDAF 基于画像信息，向运营域推送信息，辅助体验套餐运营，比如目标客户，套餐体验结果等等。

5.3.2 智能软硬协同业务识别技术

移动网络中加密数据流占比越来越高，对于加密数据流需要基于报文的特征来识别，传统的方法是从网络中获取数据流，离线分析后生成特征库，然后将生成的特征库加载到网络中完成对于加密数据流的识别；采用这种方式，从新应用出现到完成特征库生成和加载需要较长的周期，对于爆款应用，可能错过针对该应用开展体验经营业务的最佳时间。

互联网各种应用的版本更新频繁，采用特征库识别方式，需要持续更新特征库以匹配新版本应用，受限于特征库从分析识别到完成加载的时间周期长，导致针对开展体验经营应用的识别率下降，从而影响体验经营的效果。

面对移动网络数据加密、应用版本易变等挑战，传统 SA 特征库识别的方式在准确率和及时性有较大挑战，无法满足体验经营的要求；需要引入基于 AI 的应用识别技术来匹配这种新需求。

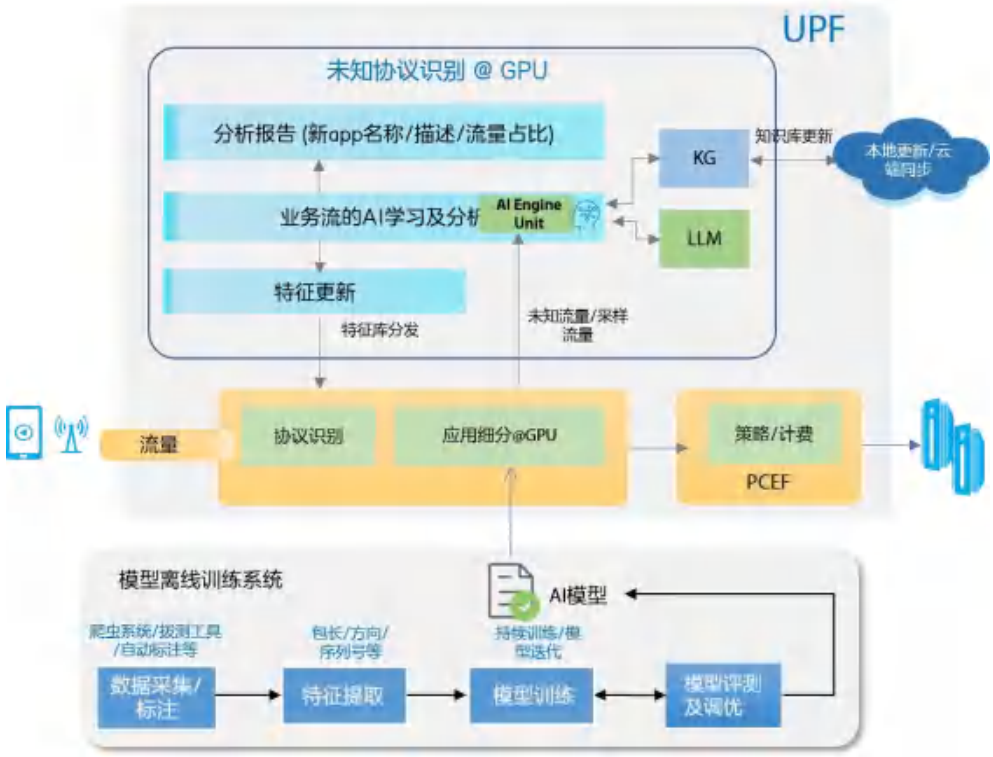


图 5-5 智能软硬协同业务识别架构

上述架构，借助 GPU 提供的超强算力，运行两个关键智能组件：

1. 未知协议识别，具体包括：

   – 智能分析：业务流量自动分析，识别网络新应用及已有应用更新，分析报文特征，进行流特征聚类

   – 智能标注：借助大语言模型，分析同类应用的不同业务流关联性，实现业务的聚合及标注；

   – 特征生成：生成协议特征，并进行有效性验证；生成的新特征库在线加载到 UPF，实现特征库在线更新，特征库更新周期缩短到天。

2. 未应用细分，具体包括：

   - 离线训练细分业务识别 AI 模型
   - 在线完成数据流推理，识别业务流归属的细分应用，比如社交软件中的语音通话，视频通话，文本传递， 文件传递等。

开展体验经营业务时，可以针对细分的应用，分配对应的 GBR 保障带宽，实现精准的资源配置，最大化使用无线资源。

5.3.3 全息 KQI 体验度量技术

在推进体验经营套餐的过程中，精准衡量用户业务体验成为了关键所在，这要求我们采用一种能立体、真实地反映用户体验的度量技术——全息 KQI 度量技术。该技术通过多维度的 KQI 指标，实现了对用户体验全方位、深层次的度量。

以往，体验度量主要依赖于传输数据、播放器消息数据和操作系统信息数据的综合收集与分析。然而，在运营商开展的体验经营场景中，由于数据流加密等复杂因素，直接获取播放器和操作系统等详细数据变得极为困难。这一局限性在传统的度量方式下构成了巨大的挑战。

全息 KQI 度量技术则打破了这一瓶颈。它采用 AI 算法，通过分析数据流，提取数据流的各种特征（比如包长分布，时延，流量等），通过建模的方式，生成各种不同用户体验的应用数据流对应的数据模型；后续网络中的真实数据流，同样提取数据流的特征，通过模型推理，获取到数据流对应的各种真实体验指标（KQI）。这相当于全方位提取数据流的特征，将这些特征投影到模型中，模型通过推理运算，还原用户的真实

体验指标。

借助该技术，运营商可深入解析视频、游戏等数据流，即便在数据加密的情况下，也能准确捕捉并提炼出对应的 KQI 体验指标。这种技术不仅具备高度的精准性，而且将用户体验的多个细节都立体、真实地呈现出来，为运营商准确衡量并优化用户体验提供了有力的技术支持。

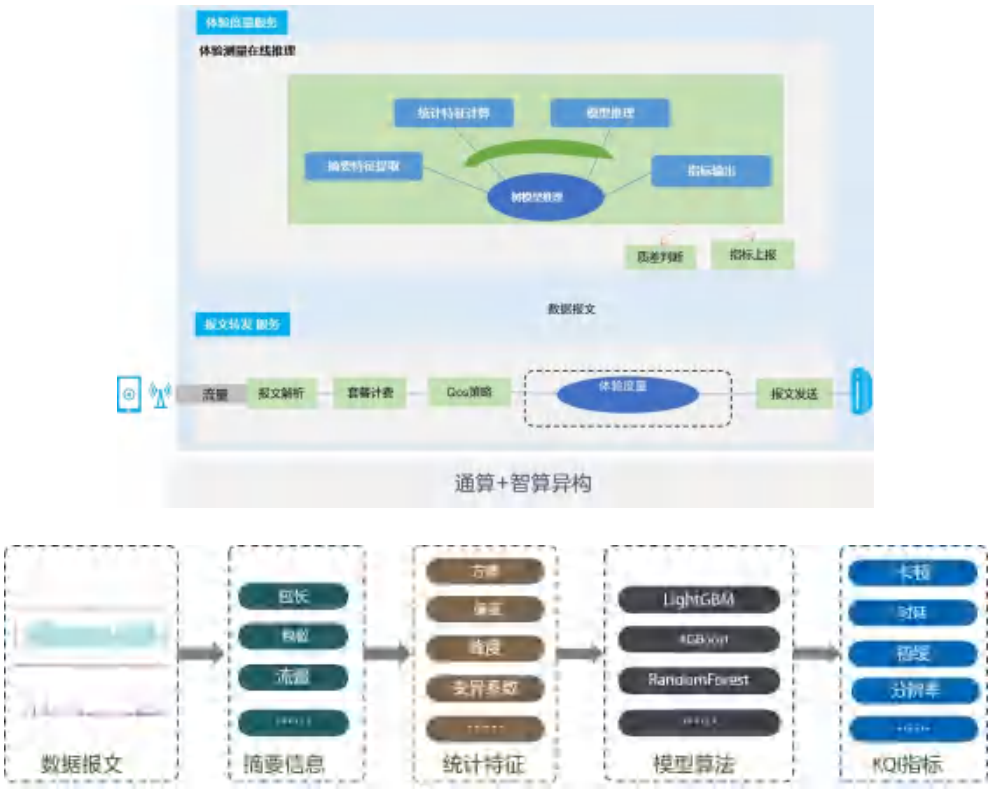


图 5-6 全息 KQI 体验度量技术

5.4 赋能行业，拓展连接空间

无线连接取代有线连接，需要满足千行百业的需求，总结起来，行业应用对于连接的主要诉求如下：

1. 行业应用需要连接提供确定性，比如保证工业生产过程中的指令按照预期有序执行；无线网络提供确定性连接有两种方式，一是外置 TSN 网关保证数据传输的确定性，这会带来更好的成本，二是内置 TSN 网关，提供内生确定性，这对于网络的端到端编排提出了较高的要求，如果依赖于手工配置，会带来大量的网络配置和维护工作失去了引入无线连接提升网络灵活性的价值，有必要引入 AI 实现智能编排。

2. 行业应用有低时延需求，比如工业生产，这要求 UPF 能够下沉到工业园区，同时由于工业应用机房环境有限因此对于 UPF 设备提出了更多的环境等需求，需要部署适用于工业环境的工业 UPF，同时工业检测等领域需要做图像检测等处理，这需要 AI 算力，因此在有限的工业机房中需要提供集成 AI 算力的工业 UPF。

5.4.1 AI for 确定性，拓宽 OT 域服务广度

面对工业 4.0 时代对生产效率与精度的极致追求，传统网络架构已难以满足工业生产中对于低时延、高可靠性和确定性传输的迫切需求。为解决这一难题，我们创新性地在工业生产现场部署了 OT-UPF，旨在为企业提供便捷的就近接入服务。通过内置 AI 能力，OT-UPF 打破了传统 5G 网络“尽力而为”的传输局限，为工业生产量身定制了确定性传输通道。这一变革性举措不仅避免了传统 TSN 网络配置复杂、运维繁琐的困境，更以生产任务为导向，智能优化端到端转发调度，确保整个传输路径上的确定性，为工业生产的数字化转型与效率提升奠定了坚实基础。

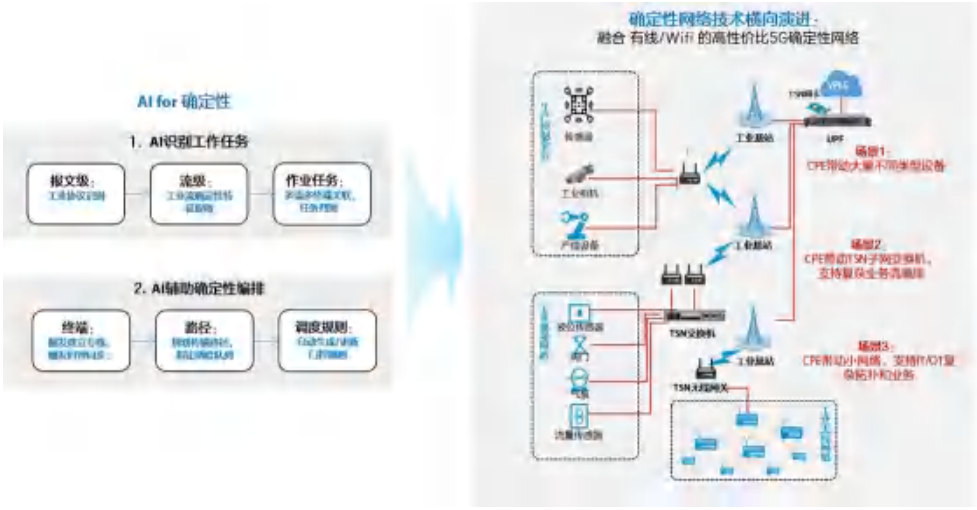


图 5-7 AI for 确定性网络

1. 在工业生产现场部署 OT-UPF，为企业提供就近接入；

2. OT-UPF 内置 AI 能力为企业提供确定性传输通道；传统的 5G 网络是一种尽力而为的传输通道，满足不了工业确定性需求，为了满足工业确定性指标要求，需要引入 TSN 并进行端到端的配置规划，这给网络运维带来了巨大的复杂性；引入 AI 以后，以工业生产任务为主线，在每个主线条上自动进行转发调度（设置门控调度规则），保障端到端的协同从而达到整个传输路径上的确定性。

5.4.2 网智融合，全面支撑边缘应用

面对边缘应用中的算力分散、模型泛化不足及数据安全隐患，边缘智算 UPF 应运而生。作为新一代网络基础设施，它集成 AI 算力，为智能业务和装备提供一站式算网底座，吸引并壮大智能工业/园区等 5G 专网生态圈。从专用 AI 模型到通用大模型，边缘智算 UPF 的 AI 能力持续增强，且随着 AI 板卡能力的提升，实现从离线到在线训推一体，

确保实时动态调整并避免数据外流，为园区智能化转型提供坚实保障。

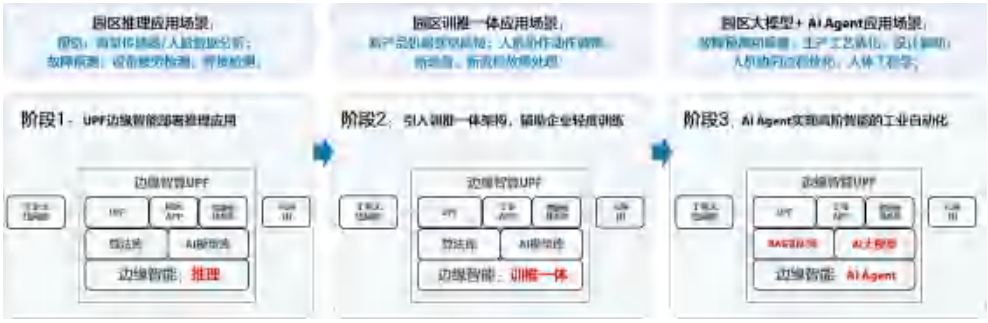


图 5-8 边缘智算 UPF

1. 边缘智算 UPF 集成 AI 算力，为智能业务和装备 提供一站式算网底座，吸引并壮大智能工业 5G 专网生态圈

2. 边缘智算 UPF 从提供针对特定场景（比如机器视觉模型）的专用，逐步泛化到支持通用场景的 AI 大模型；AI 模型的泛化能力逐步增强。

3. 随着嵌入的 AI 板卡能力的提升，边缘智算 UPF 从离线训练+在线推理，逐步增强到在线训练+在线推理（训推一体）实现实时动态调整，并有效避免数据离开园区。

5.5 多技术融合，提升连接效率

5.5.1 AI+云化技术，高效节能

核心网引入虚拟化技术后，可以通过降频/休眠方式实现空闲时节能；进一步引入 AI 后可以提升对于网络空闲的预测准确度，实现动态的精准节能，提升节能效果。

场景：NF 的负荷在不同时间段存在波动，在低负荷区间可以通过降频/休眠实现节能，在高负荷区间需要取消降频/休眠：

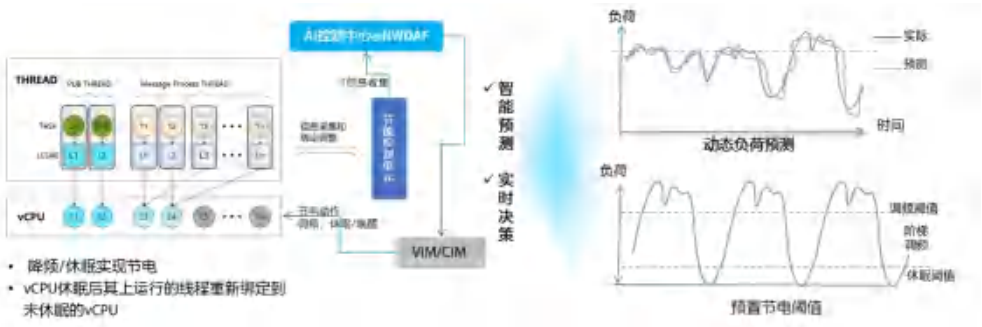


图 5-9 智能节电

AI 应用：AI 控制中心基于从 NF 收集的历史负荷数据，使用 AI 算法推算在后续时间段内 NF 的负荷数据；

当采用休眠方式时，将原来运行在多个 vCPU 上的线程，迁移到集中的少量线程中，由于不减少运行的线程数量，因此不改变原有的负荷分担，业务基本不感知，用户体验不受影响。

5.5.2 AI+网络技术，优化网络信令负荷

寻呼信令在网络信令中占比很大，因此网络发展过程中持续在做寻呼信令的优化，过程中始终需要考虑减少寻呼信令和减少寻呼成功时间的平衡，通常情况下，大范围寻呼可以减少寻呼成功时间，先小范围寻呼再扩大范围逐步寻呼可以减少寻呼的信令，因此减少寻呼信令和减少寻呼成功时间是一对矛盾，传统采用固定策略的方式很难达到最优的平衡；引入 AI 技术，基于用户轨迹画像，可以实现最优平衡。

智能寻呼方案，AMF 内置 AI，基于用户的历史活动轨迹推算出当前寻呼时间用户最可

能停留的基站，然后向几率最高的一个或者几个基站发起寻呼，达到寻呼效率和寻呼时间的最优平衡。

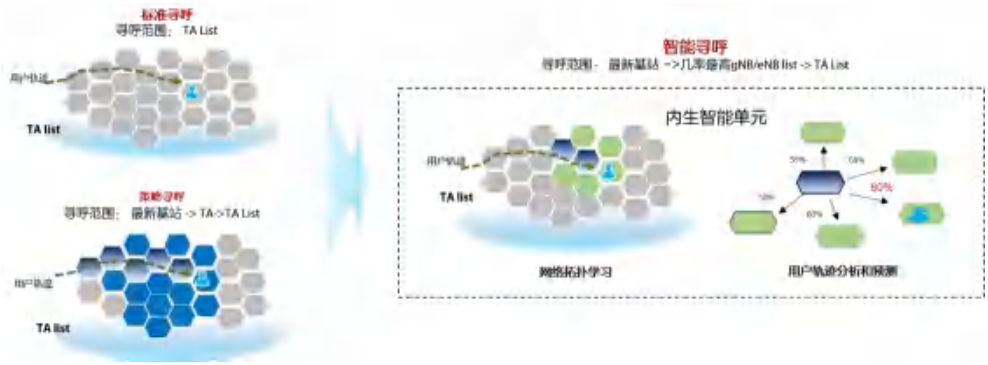


图 5-10 智能寻呼

5.6 安全内生，增加连接韧性

随着移动网络发展，连接规模越来越庞大，海量连接容易触发信令风暴，给网络安全带来巨大的风险，例如，加拿大 R 运营商发生大规模网络中断，时长约 19 个小时；日本 K 运营商发生全网业务瘫痪，中断时间长达 62 个小时。业务稳定是通信产业的根基，网络瘫痪不仅会造成运营商直接收入损失和巨额赔偿，还会影响到运营商的品牌声誉乃至生存发展。因此预防并应对信令风暴成为运营商的共识，而借助 AI 技术能够有效预防和高效应对信令风暴。

5.6.1 智能识别异常终端，预防信令风暴

随着 5G 网络的快速发展，网络攻击手段日益复杂多变，异常终端的频繁出现成为网络安全的重大隐患。因此，通过智能识别技术及时发现并隔离异常终端，有效预防网

络被攻击，是确保网络稳定运行、避免信令风暴发生的关键举措。

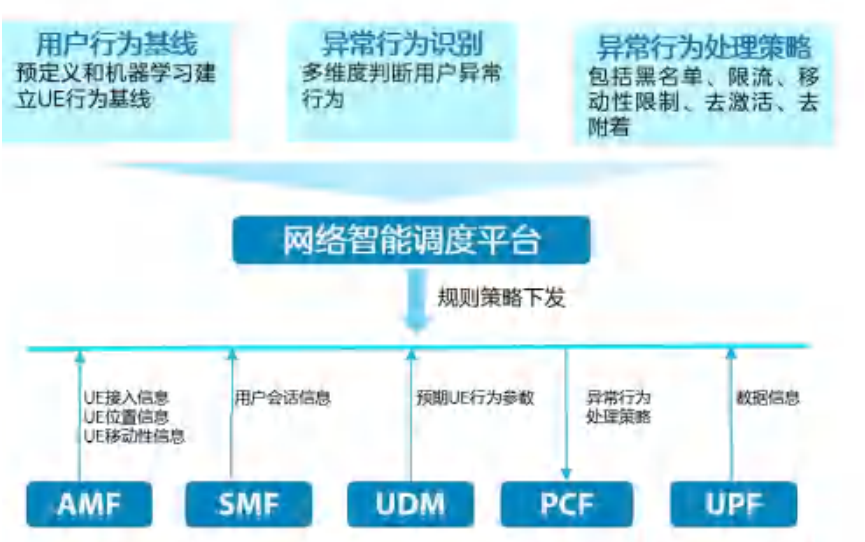


图 5-11 异动终端识别和防护

1. NWDAF 内置智能调度平台，通过收集各个网元上报的控制面和用户面数据，通过 AI 算法训练并生成正常用户的行为基线；

2. NWDAF 基于行为基线进一步推理判断当前活跃的用户是否超越了基线，从而判定用户为异动终端，对于判定的异动终端，基于预定义的处理策略对于用户执行禁止接入一段时间（Parking 用户），禁止用户使用数据业务（服务禁止），分离用户等。

5.6.2 基于数字孪生，生成信令风暴应对预案

在人工智能技术的推动下，数字孪生网络为我们提供了前所未有的网络问题洞察与预防能力。数字孪生网络通过精准复制生产网络的信令处理能力，结合 AI 算法预测话务

增长与信令变化，提前识别网络瓶颈，模拟异常场景下的信令冲击，为优化网络性能、预防信令风暴提供科学依据，确保在面临网络压力时，能够迅速响应并给出针对性的优化建议。

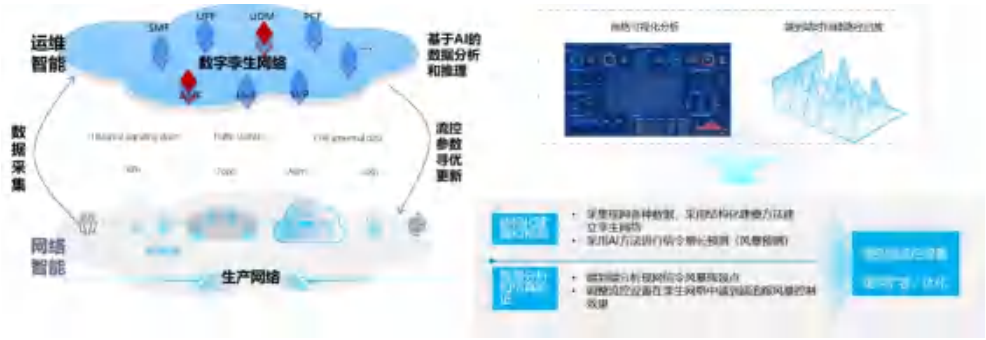


图 5-12 基于数字孪生的信令风暴模拟和防护

1. 部署数字孪生网络，网络中各个孪生节点的信令处理能力遵从生产网络中的各个节点的能力

2. 从生产网络中采集话务数据，并采用 AI 方法进行端到端信令增长预测，比如当 5G 注册增加时，相应的 UDM 侧的交互信令也会同步增加；由此分析，随着用户数和话务模型的增长，生产网络中率先出现瓶颈的节点和接口；针对性的提前部署应对措施；

3. 模拟网络异常（比如某个节点发生故障）情况下的突发信令冲击，模拟该信令冲击下网络的恢复时间和各个节点的恢复时间，对于收到冲击最大的节点，给出优化建议；比如网络出现故障时，由于大量用户的重新注册，导致 UDM 受到巨大冲击，为了减缓冲击，建议限制 AMF 向 UDM 发起的消息数量，并结合 UDM 的

能力给出具体的建议值。

5.7 场景化到全面 AI，面向 6G 持续演进

5G 在现有的网络连接基础上引入 AI，以应对一些当前网络面临的挑战；采用这种叠加方式引入 AI 好处是：（1）针对具体的场景应用 AI，模型更有针对性，更高效；（2）对于现有架构和流程冲击小，更容易部署； 但是，采用叠加方式缺点也是很明显的：（1）模型通用性/泛化能力差，每次出现新需求场景都需要重新开发模型，开通周期长；（2）模型数量多，模型管理复杂。 因此，网络演进到 6G 后，需要网络内生 AI，从 6G 系统的各个单元收集数据并完成系统级的大模型训练，将训练出的通信大模型应用到 6G 系统的各个决策点，从而使得网络决策变得更智能、更高效，比如资源分配，比如负荷均衡，比如路径选择等等。

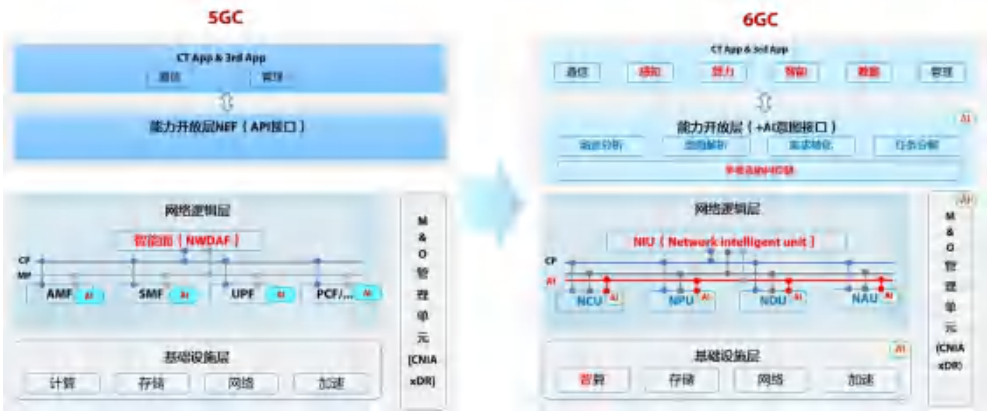


图 5-13 从 5G 智能架构演进到 6G 智能架构

从 5G AI+连接演进到 6G 连接内生智能是一个持续演进的过程，演进过程中，AI 应用场景不断扩大，比如从人网扩展到行业网；AI 技术的不断增强，比如从简单的机器学

习到水平和垂直领域的联邦学习，从离线训练+在线推理到在线训练+推理。

图 5-14 网络智能化演进

6 “AI+”运维：重塑运维范式

通信领域与千行百业的数字化转型不断加速，新网络、新业务和新技术持续涌现，网络结构趋于复杂，业务应用多样化，运维管理更为复杂，网络安全与隐私保护的重要性也越发突显。传统人工运维方式受效率和能力的限制，已无法适应当前的运维需求。在此背景下，运维自动化和智能化成为业界的共识，智能运维成为保持竞争力的关键所在。

意图网络、大模型、数字孪生等新兴技术的问世为智能运维带来新的契机。这些技术具有诸多优势，如人性化的人机交互模式，可处理海量格式化数据，能进行高精度的分析与预测，还可实施高逼真度的仿真验证等。引入 AI 大模型、意图网络和数字孪生等技术，构建“AI+”运维，可重塑核心网运维范式，推进运营商从 TMF 自智网络的 L3 级运维朝着 L4 +高阶自智运维发展。

6.1 网络运维挑战：三多、三新、三跨

随着 5G 网络的发展和虚拟化云化技术的引入，通信网络愈发复杂，通信业务日益多样，基础设施和业务系统面临多种复杂状况，主要表现如下：

- 三多：多接入、多类型、多网元，维护成本高。运营商大多存在 2G/3G/4G/5G 多接入网络、数据和语音网络、运营商公网和专网并存，使得网元数量和接口数量成倍增加，导致运维难度增加，OPEX 成本上升。

- 三跨：跨层、跨域、跨厂家，运维复杂，人工运维难以应对。虚拟化架构包含硬件层、虚拟层和网络功能层，每层的产品可以由多厂家提供；端到端网络也存在无线域、承载域、传输域、核心网域等多域并存。如要实现端到端的运维和故障定界定位，需跨层、跨域、跨厂家相互协作，导致运维管理复杂度和难度指数级上升。人工操作周期长、效率低、出错率高，难以应对复杂场景的运维。

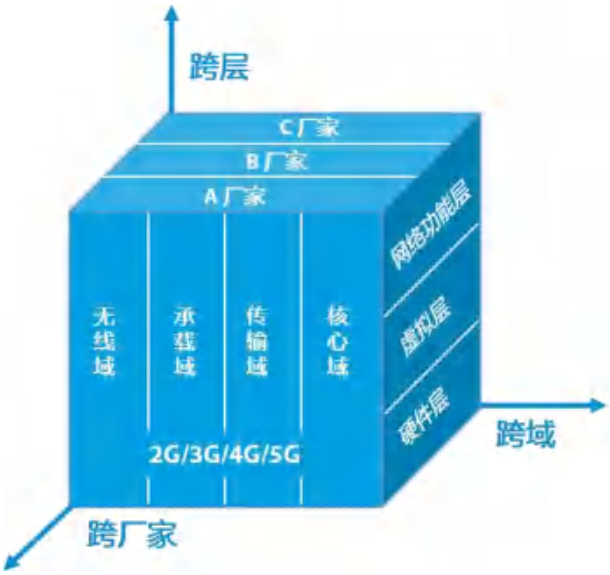


图 6-1 网络运维面临的三跨挑战

- 三新：新业务、新架构、新技术，需重塑运维范式。
  - 新业务：5G 网络不断涌现新业务，如新通话、XR、元宇宙、工业互联、无人机等。这些业务上市时间紧，需求更新频繁，还往往伴随着海量数据，并且对实时性和可用性要求极高。此外，不同行业客户的业务需求与服务水平协议（SLA）要求也存在差异。因此，迫切需要智能运维对各个网络进行精准管理，按照不同网络的需求开展智能运维工作，保障高度稳定、安全，确保业务持续运行，提升用户满意度。
  - 新架构：SBA（基于服务的架构）的服务化将网络功能拆分成多个独立的服务，需要对每个服务单独进行监控、管理和故障排查。当业务流程改变，例如新增一种业务场景需要不同服务组合时，能够及时调整服务编排策略。传统的手动监控方式无法满足这些需求，需要自动化的监控手段实时获取各个网元的状态信息，如性能指标、故障告警等，同时还要借助智能化手段分析网元之间的关联，便于出现问题时迅速确定故障根源。
  - 新技术：AI 大模型、意图网络、数字孪生等新兴技术出现后，行业标准组织和运营商都在积极探索如何将这些新技术应用到网络运维领域，从而实现核心网的智能化运维。例如，可以利用 AI 大模型的强大分析能力预测网络故障，也可运用数字孪生技术对网络进行模拟和优化等。

6.2 “四层一体”AI+运维新架构，迈入高阶自智

为积极应对核心网运维的诸多挑战，同时兼顾业界对自动化和智能化的需求，需构建

基于意图网络、大小模型和数字孪生体的“四层一体”核心网智能运维新架构，实现网络、数据、模型和应用的全开放解耦，推进网络迈入高阶自智阶段。如下图所示，该架构通过大小模型、异构模型的解耦协同与平滑演进，打造灵活编排的数智技术底座，赋能高效运维；通过数据、模型、应用一体化的数字孪生体为规划、建设、运维和优化赋能，精准构建高稳网络；通过涵盖监控中心、排障中心、应急中心和变更中心的统一运维门户，赋能全闭环网络变更、全闭环监控排障、全闭环投诉处理等高价值运维场景，推动网络运维从被动应急模式转变为主动高效模式，持续削减 Opex 运维成本。

图 6-2 “四层一体”核心网智能运维架构

此系统架构包括四层一体：网络层、数据层、模型层、应用层和数字孪生体。

1. 网络层：涵盖 2G/3G/4G/5G/IMS 的核心网现有原子网元，这些原子网元是构建整个核心网的基础要素，为网络的运行提供了最基本的物理或逻辑实体支持。

2. 数据层：在整个架构中扮演着提供高质量语料数据的关键角色，其中包含通用知

识库、大数据、OMC、MANO 和 CHR/UDR 等网元。这些网元通过收集、整理和存储各类数据，为上层的智能分析和运维决策提供了丰富的数据资源。

3. 模型层：是核心网智能运维架构的核心部分之一。它以大/小模型为强大智能引擎，致力于打造一个可组装、可编排且能够自主迭代的数智技术底座。通过引入大小模型构建起以大模型为基础核心、小模型 Agent 为智能应用的新型使能运维系统。通过叠加各产品专家多年经验沉淀精调语料集，能够提供面向交互、分析、生成等多方面的大模型服务。

4. 应用层：基于智能层构建的数智技术底座，能够智能编排生成自智应用、AI Agents 专家集群、Copilot 助手团队等多应用能力。通过意图交互、内容生成、多任务串接等功能相结合，能够轻松应对核心网在规划、建设、维护、优化、运营等多复杂运维场景下的能力要求。例如，在网络规划阶段，可以根据业务需求和资源状况智能编排网络建设方案；在运维阶段，可以快速定位故障并提供解决方案。

5. 数字孪生体：借助数字孪生仿真平台，构建“业务模型 + 孪生应用”相结合的架构。这种架构有助于实现孪生应用的快速部署、业务的敏捷创新和高稳网络的精准构建。例如可以在数字孪生环境中仿真网元本体，构建网元孪生体，同时仿真业务流程，提前发现潜在问题并优化解决方案。

依托“四层一体”架构，核心网智能运维系统支撑内生智能，具备基于 AI 大模型智能体的智能网络感知与监控、智能网络分析与诊断等能力；具备基于数字孪生体的智能网络评估与决策、智能网络变更与执行等能力，并形成全闭环控制机制，助力网络实现自配置、自优化和自修复。

6.3 大小模型协同，赋能高效运维

AI 大模型，可基于海量数据进行高精度的数据分析与预测，以自然语言快速回答用户的各种问题，提升工作学习效率，改善人机交互体验。还能高效处理文本，启发创新方案，如文档的生成与分析，在多方面为用户创造便利和价值。

随着通讯网络和业务变得越发复杂，核心网运维人员需要掌握多种产品的专业知识，并具备很强的故障分析与解决能力。但面对海量数据和众多故障点时，人工运维就难以快速定位和排查故障。

为降低运维人员的学习和运维成本，提升系统智能运维分析和方案生成能力，“四层一体”核心网智能运维系统依托通信领域大语言模型能力与 AI Agent 智能体服务，结合通讯通用知识和各产品专家多年经验沉淀的精调语料集，构建智能交互、智能分析、智能生成等大模型服务，以满足规、建、维、优、营等复杂运维场景的要求，推动网络向 L4+高阶自智演进。

此系统通过智能体协同的检索增强生成技术（Agentic RAG）构建统一智能交互门户，实现通讯知识问答和网络运行状况检查等功能；通过泛化意图理解与精准运维相结合的大小模型协同技术，实现快速、高效的智能分析，提供故障诊断、网络优化和例行检查等功能；通过基于强化学习的智能生成技术，提供高效的文本处理能力，能够快速生成重大操作方案和观察报告等文档。

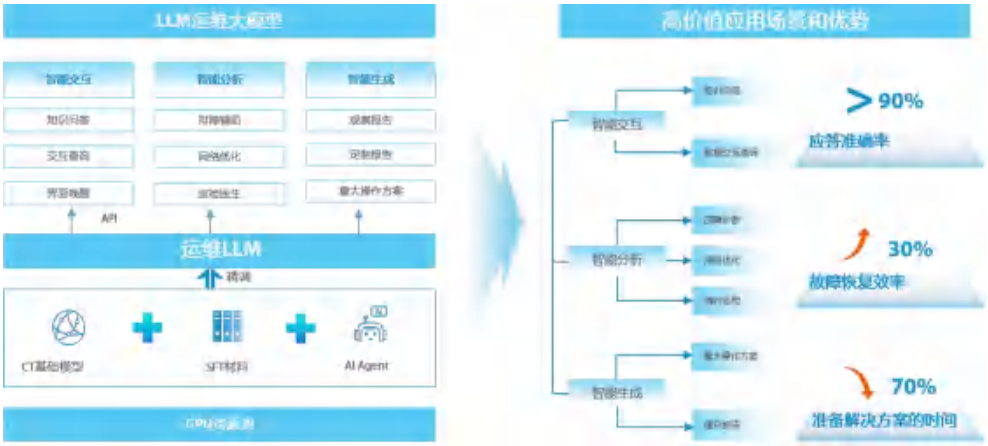


图 6-3　基于 LLM 大模型的高价值应用场景

- 智能体协同的检索增强生成技术（Agentic RAG）

首先，该技术通过通用 RAG 能力构建起统一的知识获取门户，凭借基于 LLM 大模型的知识问答与解答，使用户快速掌握 2G、3G、4G、5G 等各类产品的专业知识。

其次，增强型 Agentic RAG 把 AI 智能体融入 RAG 流程，协调各组件以执行超出简单信息检索和生成的更多操作。这使得回答准确率可达 95%以上，并且答案内容在元素、图片、文字等形式上更加丰富，既方便检索又提高检索效率，还降低了运维人员的学习成本。

同时，采用 Agentic RAG 技术可自动识别运维人员对 KPI（关键绩效指标）的诊断和分析意图，自适应调用不同运维系统的应用程序接口（API），借助 NL2SQL（自然语言到结构化查询语言）快速获取数据。这种高效灵活且场景个性化的方式有助于打造智能交互门户，让人人都能瞬间成为“通信专家”。

- 泛化意图理解与精准运维结合的大小模型协同技术

通过提供轻量级算法引擎、迭代数据学习和业务体验注入，构建可视知识地图，提高运维事件关联规则的建立效率。经实践验证，基于 AI 的迭代学习和专家经验，可在 1 小时内完成 100 万告警或事件规则的挖掘；基于关联告警事件聚合，可大大减少告警数量，告警聚合压缩率达到 80%以上。

大模型通常解决泛化的意图理解，而小模型用于精准的解决实际运维断点。通过大小模型协同，可实现故障检测、故障定位、故障逃生与恢复的全闭环，并协助完成网络优化和例行检查等操作。实践证明，故障检测时间可缩短为 1 分钟，故障定位时间可减小 50%，故障恢复时间可减少 50%人力。

- 基于强化学习的智能生成技术

基于强化学习的智能生成技术为新手解决了文档编辑的难题。对于新手来说，编辑诸如重大解决方案、智能报告等文档是极为繁琐且极具挑战性的工作。

该技术借助基于 LLM（大型语言模型）的生成能力，能够把全球网络规划和建设方案归档到数据湖中。然后运用大模型学习历史规划数据，进而生成电信领域的规划模型参数和解决方案等，可用于客户现场交流，提升用户体验。

经实践验证，利用该技术准确生成解决方案的时间能够降低 70%，其生成的方案具有较高的准确性和可靠性。

6.4 一体化数字孪生，构建高稳网络

“四层一体”核心网智能运维系统，凭借数字孪生技术，融合大语言模型 LLM、智能

体服务 AI Agent 与多模态大模型（MMLM），构建数字孪生体，提供核心网系统、设备和组件的数字孪生模型，具备跨厂家模型的管理能力。它可通过编排或低代码开发构建孪生模型，依据厂家模型信息调用能力获取计算结果，还具备呈现能力（以可视化方式在用户门户呈现孪生体计算结果）和北向能力（供运营支撑系统调用）；还可为运维提供可视、仿真、预测、策略反馈等能力，以低成本提供定性与定量分析能力，支撑从人工决策向机器决策转变，赋能高阶自智演进，进而实现运维自动化闭环。

同时，为构建高稳网络，数字孪生体搭建了数据、模型、应用一体化的系统架构，为规划、建设、运维、优化等应用场景赋能。在移动通信网络中，核心网处于中枢位置，是终端与业务服务器连接的必经通道，其稳定性和可靠性直接影响网络业务的可用性。若核心网出现信令风暴，必然致使网络长时间无法使用，业务也会长期中断，影响范围广泛。为增强核心网抵御信令风暴的能力，可运用数字孪生技术对核心网的状态和网元行为进行孪生，开展网络故障、事件模拟以及信令风暴过程的演练，进而发现并优化网络薄弱之处，提升网络的抗风险能力。

- 多业务仿真验证与优化

数字孪生体利用构建的孪生体仿真网络，开展多业务场景与解决方案的仿真验证工作。比如，模拟高流量业务场景下网络的性能情况，或者验证新的网络优化方案是否可行。

运用 AI 算法和统计方法对现网数据进行分析处理，获取网元机理、网络信令、网络拓扑、路由权重、网络话务、终端恢复行为、网元恢复行为等关键网络信息模型。然后结合孪生技术，完成核心网组网状态和配置的数字化仿真建模。

通过灵活编排孪生核心网的网元运行状态、网络信令流量等故障事件，模拟多种异常

场景引发的网络信令风暴冲击，分析并找出信令风暴和业务恢复过程中的网络瓶颈点及其影响程度，从而实现核心网的孪生。

鉴于现网中不同厂商的网元、终端在运行机理和恢复行为方面存在差异，所以有必要针对现网各厂商的网络信息进行个性化学习与建模。

- 层次化模型构建与映射

数字孪生模型是数字孪生网络的重要组成部分。它以感知到的网络数据为基础，将物理实体的网络映射到虚拟空间，构建出与物理实体相同的孪生数字网络。

数字孪生模型的构建，重点在于对多源异构数据进行分类与归并。针对不同的网络域构建相应网络层级的数字孪生模型，这些模型包含基础属性信息和场景功能信息，并且具备接受指令与反馈事件的能力。

为了适应不同场景下模型的共享性和差异性，采用了层次化模型构建与映射的方法。这种方法将数字孪生模型划分为不同层次，每个层次有各自的功能和特点，它们共同实现对物理世界全面的映射和模拟，有助于清晰地理解和构建数字孪生的各个部分。

数字孪生模型构建包含基础模型构建和业务模型构建两部分，如下图所示。基础模型层是针对单个物理网络实体的模型定义，从多个维度构建数字孪生模型，如运行规则模型、属性定义模型、数据模型等，这与物理网络实体的类型密切相关；业务模型则是针对特定的孪生应用场景，对基础模型进行组装融合，利用采集到的网络数据，从不同维度构建或扩展数字孪生模型。按照实现功能来划分，可分为拓扑规则模型、流量拟真模型、网络路径模型、事件检测模型、网络质量模型、孪生体管控模型等。

图 6-4　数字孪生模型目标方案

- 孪生体自主评估与决策机制

数字孪生网络目标的实现与落地并非一蹴而就，而是需要持续的目标牵引并不断迭代推进。

在数字孪生技术中，孪生体的自主决策与协同机制是数字孪生体系达到高级阶段的重要标志，这体现了数字孪生在模拟、监控、分析的基础上朝着更智能化的方向发展。

对数字孪生网络能力进行闭环评估时，构建量化评估指标体系并建立“评估 - 分析 - 提升 - 再评估”的闭环机制，是检验自身能力建设成效、推动数字孪生网络能力提升以及确保系统规划顶层设计得以执行落地的有效手段。

在智能容灾倒换、信令风暴仿真评估场景中，采用定性和定量相结合的方法，首先对数字孪生网络信令冲击仿真分级标准各级别特征和能力、场景需求进行分解，梳理出客观、可量化的关键成效关键指标，制定评价方法，然后分别对各场景的网络冲击孪

生仿真能力的建设成效评价，再结合网络智能运营场景覆盖率，最终对运营商数字孪生网络能力建设成效综合评估。通过“评估-分析-提升-再评估”的闭环机制识别短板差距，建立问题清单并跟进问题解决，持续推进数字孪生网络能力提升和全场景智慧化运营水平。

6.5 全闭环运维管理，降低运维成本

全闭环运维管理，是一种从问题发现、分析、解决到反馈优化的全方位、无死角管理模式。它不仅仅关注于故障的即时处理，更注重通过数据分析与预测，提前识别潜在风险，实现运维工作的前置化、智能化。这一模式的核心价值是在于，通过集成监控、自动化工具、AI 算法等多维度技术手段，构建了一个自我学习、自我优化的运维生态系统。

针对核心网网络面临的挑战，综合 TMF AN L4 推荐的关键价值场景与运营商实际生产需求，当前的全闭环运维管理主要包括全闭环网络变更、全闭环监控排障、全闭环投诉处理三大高价值场景，其流程和成效目标如下图所示：

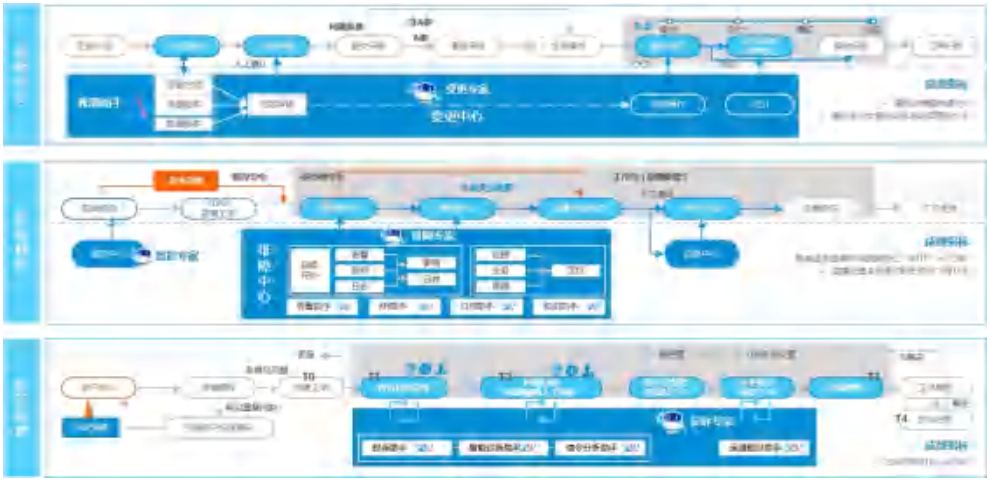


图 6-5　全闭环运维管理的高价值场景流程和成效目标

- 全闭环网络变更

在通信设备的运维流程中，全闭环网络变更是极为关键的端到端场景，涉及网络架构、设备、配置、安全等多方面的调整与优化。

网络变更从制定计划开始，要历经变更前、变更中、变更后的诸多步骤，包括方案制作、提交审核、发布操作、操作执行、业务验证、值守分析，直至最后的变更工单归档等环节。在这个端到端的流程中，存在人工扭转、审核等步骤，这会影响变更操作的时效性。

而智能网络变更与执行能够实现流程可编排、操作自动化。例如：升级、割接、扩容等重大操作可实现全流程自动化，能进行分钟级的模板设计，任务可定时、批量执行；各网元操作有监控大屏，流程可定义，过程可监控，业务指标能分钟级展示，还能快速定制巡检任务，实现一键式巡检。借助智能网络变更流程，整体操作的故障隐患可降为 0，操作步骤的自动化率能够逐年递增 15%以上。

- 全闭环监控排障

全闭环监控排障涵盖从源头到终端的整个数据传输过程。在此过程中，通过监控系统实时追踪通信设备运行状态，收集与分析运行数据，及时发现潜在问题或故障并采取相应解决措施，确保通信设备稳定运行。网络监控排障一般以监控流程开始，开展多维度、多层次健康监控，进行智能指标异常检测与故障/隐患智能识别工作。并借助对指标波动性、周期性、趋势性等数据的分析，自动构建指标异常识别模型，通过智能故障分析，提供告警聚合、KPI 异常分析等能力，实现网元多场景多维度智能分析定位故障、网络跨层故障诊断，并自动输出诊断图和根因。

- 基于告警大模型的告警聚合方案，提供轻量化算法引擎，数据 AI 迭代学习，业务经验注入，建立关联告警汇聚规则。通过聚合规则、空间关系聚合关联告警，提升告警规则挖掘效率，大幅减少告警数量，提升故障定位效率。
- 基于 KPI 大模型的异常分析，自动理解运维人员 KPI 分析意图，自动适配和调用不同运维系统 API，在数分钟内辅助运维人员完成异常 KPI 异常分析，实现自助异常分析高效运维，效率可提升 50%。在个性化的异常分析场景可以任意调度任意串接，诉求覆盖率提升 1 倍以上
- AI 安全生产方案，提供网元级、网络级安全保护方案，建立内生安全机制，通过安全策略中心实现资源智能微隔离、入侵主动检测、病毒防护以及资产安全管理，实现智能化微隔离，对异常流量主机快速的安全隔离；内生安全对 5G 网络入网、运行进行主动防御、主动检测，增强安全策略自动部署，加强 VNF 内生安全能力。

- 应急中心，一键故障逃生机制，可预置多种故障应急预案，统一管理、集中控制，自动化、可视化执行应急操作，缩短故障恢复时长。

- 全闭环投诉处理

在通信设备的运维流程中，投诉处理属于重要的端到端场景。其通常包括投诉接收与分析、投诉问题分拣与派单、投诉内容分析、投诉结果回单等环节。在通讯领域，要对用户投诉涉及的 XDR 话单和相关信令进行提取分析，从而精准定位用户投诉的问题。

然而，由于这一过程步骤繁多、调用系统复杂、问题定位难度大，往往需要较长时间来完成问题分析，用户体验不佳。为改善这种情况，提高投诉处理效率，有必要在关键步骤引入更优的处理机制。

- 以意图驱动投诉处理工作，借助相关静态经验知识、投诉案例检索、投诉关联数据查询、信令智能分析、单域原子能力调用以及大模型综合推理，将专业室能力前移到监控室，突破专业室技能的限制，从而缩短投诉分析与处理的时长。

- 以时间范围、业务类型、具体接口、用户号码等作为查询条件，检查现网数据采集是否完整、时钟是否同步等情况，确保数据质量达到可进行下一步分析的要求。

- 依据 XDR 话单，根据失败的 XDR 关联出失败信令，解析出失败原因码，将失败原因翻译出来，进而给出解决方案。

在上述典型的全闭环运维场景中，“四层一体”架构下的监控中心、排障中心、应急

中心和变更中心这四个中心门户系统被充分利用。这四个系统强有力地支撑核心网的 AI +智能运维能力，实现智能网络感知与监控、智能网络分析与诊断、智能网络评估与决策、智能网络变更与执行这四项核心能力，推动网络实现自愈、自优化和作业自动化，从而降低 Opex 运维成本。

6.6 持续演进，目标“无人化”AI+运维

TM Forum 定义了自智网络等级，用于指导网络和业务实现自动化与智能化，评估自智网络业务的价值和增益，指导运营商和厂商的智能化升级，并详细介绍了自智网络分级评估方法和流程、任务评估标准、评分方法等，各运营商都在积极推进自智网络的实践。目标实现 L5 级全流程系统自动化，所有场景均将交由系统自动完成，实现理想的“无人化”智能运维新范式。

然而，在实际应用中，“无人化”运维会遭遇技术复杂性、安全与隐私、人为干预的必要性以及监管与合规性等挑战。因此，未来运维的演进可在多维智能体协作可编排、基于多模态大模型的自适应智能体协同、基于 AI 和数字孪生双轮驱动的融合底座等关键技术方面持续增强，并使其有机融合，实现更加高效的运维，打造更加高稳的网络，不断降低运维成本。

自智网络等级	L0： 人工运维	L1： 辅助运维	L2： 部分自智网络	L3： 条件自智网络	L4： 高度自智网络	L5： 完全自智网络
执行	P	P/S	S	S	S	S
感知	P	P/S	P/S	S	S	S
分析	P	P	P/S	P/S	S	S
决策	P	P	P	P/S	S	S
意图/体验	P	P	P	P	P/S	S
适用性	N/A	选择场景				所有场景

P：人（手工）　　S：系统（自动）

图 6-6　TM Forum 定义的自智网络等级

- **多维智能体协作可编排技术，**是“无人化”运维应用的创新保障。

智能普惠与连接智能是未来网络的重要愿景。未来网络除了作为连接基础设施之外，在架构层还应基于原生设计支持 AI，为用户提供 AIaaS 服务，其服务范围不仅包含连接服务，还涉及内生的计算、数据、AI 等服务。

在运维过程中，必然要将 AI、数字孪生、大模型等关键技术进行有机组合与编排。当多个智能体向网络发出服务请求后，按照需求编排服务流程，将其部署到满足能力要求的网络节点运行，最终输出全局决策来指示智能体。这些技术能够为智能运维应用场景中的大量多智体协同作业需求（例如故障自我发现、自愈修复、自动报告等）提供支持。

在多维资源联合寻优方面，引入强化学习技术感知生产网络和资源的动态变化，实现与用户需求的最优匹配，这也是实现“无人化”运维应用的创新保障。

- **基于多模态大模型的自适应智能体协同技术，**是“无人化”运维模型能力提升关键因素。

此技术融合了多模态大模型与自适应智能体协同控制。该技术借助多模态大模型的能力，让智能体能够理解并处理多种模态的信息，依据这些信息以及环境的改变自动调整策略与行为，进而达成多个智能体间的协同工作。

多模态大模型可处理和理解多种类型的信息，如文本、图片、音频、视频等，在通信运维领域还涵盖表格、日志、图形码流等模态数据。这种模型可以执行更为复杂和智能的任务，如视觉问答、图文生成、语音识别与合成、视频理解与生成等。与自适应智能体相结合后，智能体能够根据环境变化自动调整策略和行为，实现更好的协同。在自适应协同控制中，每个智能体都有自己的目标和行为策略，当受到其他智能体影响时，可调整自身策略以适应新情况，这使得多智能体系统更具灵活性和适应性。

该技术可应用于多个领域，如机器人控制、智能家居、智慧城市等。在这些场景中，智能体需要理解和处理不同模态的信息，并根据信息和环境变化进行自适应调整与协同工作。在运维场景下，无论是网络变更、故障处理还是投诉处理，多个智能体面临的输入输出要求日益增多。采用多模态大模型与自适应智能体协作的技术，可以让智能体更全面地理解运维场景中的不同环境和任务要求，并根据环境变化自动调整策略和行为，以实现更好的协同效果。

然而，该技术也面临诸多挑战，如数据对齐和融合、模型选择和训练以及计算资源需求等，需要在系统底层技术不断演进过程中逐步优化增强。

- **基于 AI 和数字孪生双轮驱动的融合底座**，是“无人化”运维的智能引擎。

首先，单独的 AI（如神经网络）有一定的局限性。神经网络在使用时，要求训练数据和测试数据遵循相同的分布，并且数据集合要全面、平衡，这种情况下它通常适用于

单一指标预测、网络异常检测等场景。然而，一旦数据分布发生变化或者故障数据不充足时，神经网络的有效性就会大幅降低。而数字孪生则可解决这一问题，通过在虚拟孪生空间创建高保真动态孪生模型，其高保真环境能够产生仿真故障数据，且不会对实际网络造成损坏。在网络资源管理、容量优化等场景中，AI 与数字孪生有机结合的机制能够让仿真、验证、预测等能力相互协作、彼此支撑。

其次，在构建数字孪生的数字化表达时会涉及特征选取。此时，可以运用无监督/自监督的深度学习方法，这种方法能够在没有先验知识的情况下，从大量未标记数据中提取具有代表性的特征，而 AI 算法将对这一整体过程的运行实施起到辅助作用。

未来网络中，“无人化”运维演进的目标方向主要聚焦于智能化、自动化以及高效的数据处理与分析能力，以实现运维工作的全面优化和升级。“无人化”运维新范式旨在运用智能化与自动化手段，最大程度减少人工干预，以此提高运维效率、降低成本，提高系统的可靠性与稳定性，进而充分发挥其在提高效率、降低成本以及增强系统稳定性方面的巨大潜力。

7 “AI+”网络云：重塑算力底座

随着 ChatGPT 横空出世，人工智能（AI）技术在短时间内呈现涌现态势，核心网智能化转型成为必然趋势。作为核心网的算力基础设施平台，网络云的智能化转型是其中的关键环节。

由于 AI 训练任务以及推理应用对算力有着高性能、大规模并行、低时延互联的要求，

导致网络云从传统的 CPU 为中心的通用计算演进到 DPU/GPU/NPU 为中心的异构计算，支持算力池化编排调度、高性能并行存储访问及高通道无损网络等技术，保障资源供应的高效和稳定成为关键。同时，屏蔽底层 GPU 异构资源细节，解耦上层 AI 框架应用和底层 GPU 类型的算力原生技术也是未来演进的方向。

在部署形态方面，由于核心网网元应用同时需要通算和智算资源，因此 AI+网络云的智算和通算资源混池部署是一个重要的特性。另外，网络云智算资源的中心预训练、区域精调以及边缘推理的分布式部署及协同模式，和传统通算网络云的中心+区域+边缘分布式部署架构完全一致。因此，网络云分布式架构智能化平滑升级，可完全满足核心网智能化的需求。

7.1 资源池化技术，提升基础设施资源利用率

智算资源池化，主要是指算力、内存资源池化以及这些池化资源的连接访问技术。智算资源池化是构建高效、灵活、可扩展的智算中心的关键。以下是对这些池化技术及其连接访问技术的详细解析：

1. 算力池化

随着 AI、大数据等技术的快速发展，GPU 作为一种重要的计算资源，在数据中心中的应用越来越广泛。然而，传统的 GPU 使用方式存在资源利用率低、弹性扩展性差等问题。据公开数据统计，传统模式下的智算中心 GPU 利用率平均数值低于 30%。除此之外，不同厂家 GPU 之间存在的软硬件绑定竖井屏障，进一步加剧了 GPU 利用率低下的问题。

因此 GPU 资源池池化技术应运而生，它本质是通过软件定义硬件加速的方式，将多厂家物理 GPU 资源通过软件抽象成一个统一的虚拟 GPU 资源池，通过 GPU 虚拟化、多卡聚合、远程调用、动态释放等多种能力，实现更加高效灵活的聚合、调度以及释放海量 AI 加速算力，精准保障 AI 模型开发、训练、部署、测试、发布端到端算力配给，使能资源可被充分利用，降低碎片概率，提高总体有效算力，降低智算中心算力服务提供成本，提升智算中心整体效能。

图 7-1　算力池化能力层级

如上图所示，从对异构算力使用的成熟度及灵活性角度出发，当前算力池化技术可划分为以下能力层级：

- 静态管理：将单物理 GPU 按固定比例切分成多个虚拟 GPU，如 1/2 或 1/4，每个虚拟 GPU 的显存相等，算力轮询。该技术可以解决虚拟机共享和使用 GPU 资源的问题，典型的包括 2021 年英伟达在部分 Ampere 系列 GPU 上提供了 MIG 技术，可以将 A100 切分成最多 7 份；
- 动态管理：以单物理 GPU 为目标，支持物理 GPU 从算力和显存两个维度灵活切分，实现自定义大小（通常算力最小颗粒度 1%，显存最小颗粒度 1MB），满足 AI 应用差异化需求。同时，该技术可充分适应当前应用云原生化趋势，实时响应上层应用对资源需求的变化，实现 vGPU 资源基于 Scale-Up/Scale-Down 的动态

伸缩，并通过资源动态挂载动态释放实现 GPU 资源超分；

- 远程调用：AI 应用与 GPU 服务器分离部署，支持通过高性能网络远程调用 GPU 资源。AI 应用可以部署到数据中心的任意位置，无论应用部署节点上是否有 GPU，只要网络可达，都可以调用 GPU 资源。此时资源纳管范围就从单个节点扩展到了整个数据中心；
- 资源池化：支持 CPU 通用算力及 GPU 智能算力独立成池，两种资源池内汇聚的资源独立按需调用、动态伸缩、用完释放。借助池化能力，AI 应用可以根据负载需求调用任意大小的 GPU，甚至可以聚合多个物理节点的 GPU 资源。在容器或虚拟机创建后，仍然可以调整虚拟 GPU 的数量和大小。除此之外，池化管理技术可引入服务质量管理技术，按任务优先级，优先分配本地资源，次选远程调用，任务资源不足时将 AI 任务进行队列化管理，等待释放出充足资源时再运行；

1. 内存池化

大模型训练任务对内存和显存带来较大挑战，数据需要在 Cache、显存、内存设备之间频繁移动，缺乏统一内存空间的寻址会导致编程模型变得复杂，也会限制设备之间的协作，必须通过手动管理数据传输和复制，因此增加了开发难度和错误率。同时，在内存和显存之间数据需要多次转换，不同的 CPU/GPU 异构设备既无法直接共享数据，也无法充分发挥各自的优势，这些因素都限制了系统整体性能的提升。

为了降低以上问题对 AI+网络云整体运行效率的影响，需要引入基于计算总线协议的统一内存池化技术。通过构建统一内存池技术，实现一致性的内存语义和空间寻址能力，将多个物理显存、内存设备及资源整合到一个逻辑内存池中，可以实现对内存资

源的统一调度、监控和管理。这种技术能够动态地分配和释放内存资源，根据应用需求进行灵活的调整，从而优化系统的响应速度和数据处理能力。

CXL（Compute Express Link）是一种开放标准的高速互联协议，专为高性能计算、数据中心和存储应用而设计。CXL 技术通过高速通道连接多个处理器、加速器、存储和其他设备，提供更高的带宽和更低的延迟，以消除计算密集型工作负载的传输瓶颈。在智算中心中，CXL 技术可以用于构建内存池，实现 CPU 与加速器（如 GPU、FPGA）之间共享和一致性地访问内存，并保持内存的一致性，这意味着数据在不同设备之间传输时不需要频繁复制或同步，从而提高了性能。但是目前看来，CXL 还需要重点在以下几个方面进行增强及优化才能实现产业落地：

- 完善满足内存池化技术的计算总线协议及子协议实现。完整、高效地实现 CXL.io 和 CXL.mem 协议，为设备之间的 I/O 通信和内存访问提供通道，优化数据传输和复制机制，降低内存池化引入的额外性能损失，确保系统高效运行；
- 加快 GPU 支持基于 CXL 实现内存一致性机制。引入内存池技术将减少数据在计算和存储设备之间协议转换频度，通过实现内存一致性机制，优化内存、显存、缓存之间的一致性算法，确保共享内存中的数据同步更新，使得设备之间数据具有一致性和可用性。同时，实现健壮的纠错机制，确保内存池系统稳定可靠运行；
- 加快制定多异构设备与内存池之间的统一接口，并具备隔离保护能力。提供多异构设备之间的协同工作接口，聚焦设备间高效协作和共享计算能力，减少数据传输和复制所带来的延迟和能耗。同时，强化安全措施，确保只有授权的处理器能访问内存池，防止访问冲突。

算力、内存资源池化技术，按照需求动态分配及回收资源，快速适应变化的工作负载，共享和复用资源，可以提升资源利用率、提升系统性能、降低硬件投资和维护成本。除此之外，还可以实现异构算力（如不同品牌、型号的 GPU）的灵活、动态、高效调用和分配，这有助于打造多厂家算力的合作开放生态，构建更加灵活、可扩展的 AI 加速算力资源池，以适应不断变化的技术演进、工作负载和应用需求。这些优势使得资源池化技术成为人工智能领域中的重要技术之一。

7.2 智算存储，满足训推任务高性能、高并发核心挑战

在大模型开发端到端的多个环节中，都对存储提出了创新需求。具体包括：

- 多元存储：视频、图像、语音等多模态数据集带来块、文件、对象以及大数据等多元存储以及协议互通的要求；
- 海量存储：为保证大模型训练的精准性，数据集通常为参数量的 2-3 倍，在当前大模型从千亿到万亿飞速发展的时代，存储的规模是一个重要的指标；
- 并发高性能：大模型并行训练场景下，多个训练节点需要同时读取数据集。在训练过程中，训练节点需要定时保存检查点（CheckPoint）以保障系统的断点续训能力。这些读写操作的高性能能够大大提升大模型训练的效率。

图 7-2 智算存储需求及架构

因此，如上图所示，作为智算存储需要具备以下能力：

- 统一存储：构建统一的存储，满足 AI 流水线不同阶段的需求，提供多元数据存储能力以及块（iSCSI）/文件（NAS）/对象（S3）/大数据（HDFS）多协议互通能力；
- 硬件加速：具体包括 DPU 卸载存储接口协议以及去重/压缩/安全等操作，以及数据按热度自动分级及分区存储；
- 软件加速：具体包括分布式缓存、并行文件访问系统/私有客户端等技术。同时，采用 NFS over RDMA 以及 GPU 直接存储（GDS）技术也能够大大降低数据访问的时延；
- 降低数据熵：减少不必要的数据移动和复制，优化存储和访问策略，降低“数据熵税”。通过去重、压缩等技术，减少数据传输和存储开销。

智算存储通过提供高性能、高扩展性、多元统一的存储解决方案，除了能够轻松应对大规模数据处理的需求、提升 AI 模型训练和推理效率、优化 AI 系统成本和功耗以及加速 AI 创新和应用落地之外，分布式智算存储系统能够完善地支持分布式 AI 架构的部署和运行。例如，通过使用分布式存储系统，可以实现数据的分布式存储和访问，提高跨域训练任务数据处理的并行性和可扩展性。同时，智算存储还可以提供跨节点的数据复制和备份功能，确保数据的安全性和可靠性。

7.3 开放高通道无损网络，降低并行计算通信开销

随着人工智能技术的飞速发展，AI 大模型的参数规模正以超越摩尔定律的速度急剧扩张，AI 大模型训练对计算能力提出了前所未有的挑战。为应对这一需求，企业纷纷构建智算集群，并引入并行计算技术，以加速模型训练。然而，尽管并行计算提升了整体计算效率，它也带来了同步开销和通信延迟的问题。在此背景下，探索如何在超大规模智算集群中实现服务器内 Scale Up 以及服务器间 Scale Out 的高速互联，从而显著提高 GPU 的利用率，已成为行业面临的重要挑战。

图 7-3 高通道无损网络架构

2. Scale-Up 网络互联趋势

随着大模型训练对算力需求的不断提升，传统的机内 Scale-Up 网络点对点 Full Mesh 互联架构逐渐暴露出其扩展性不足的弊端。Full Mesh 架构虽然能够提供高带宽和低延迟的通信能力，但其扩展能力有限，尤其是在 GPU 数量增加时，点对点的通信方式难以实现线性扩展。通常，Full Mesh 架构最多只能支持单机 8 张 GPU 卡，这大大限制了大模型的训练效率。

为了突破硬件系统的扩展性，同时解决现有 GPU 互联私有总线协议的封闭性，实现多厂家芯片互联兼容性，我们创新性地提出了基于交换拓扑的 GPU 高速开放互联 OLink

技术。通过这种技术，GPU 之间的通信从传统的点对点互联模式转向交换互联模式，显著提升了单机的扩展性和通信带宽，突破单机 8 卡的限制。通过 OLink 技术，可以打造更大规模的高带宽域(HBD)，从而大幅提升集群算力。同时，促进了多厂家生态的繁荣，还为企业提供了更加灵活的选择。这一技术开放性为行业带来了更大的灵活性和可持续性，有助于推动智算技术的多元化发展。

3. Scale-Out 网络互联趋势

超节点服务器之间的 Scale-Out 互联网络，对解决模型训练中的通信带宽和时延等技术瓶颈，提升模型训练的整体效率同样非常重要。当前业界主要有 IB 和 RoCE 两大主流技术路线。其中 IB 网络，即 Infiniband 网络，是英伟达独家提供的封闭网络解决方案，性能优异但价格高昂，而 RoCE 是基于标准以太协议的开放解决方案，但是各厂家有自己的增强方案，不同厂家都锚定自身的交换设备做了拥塞控制、端网协同等优化，难以与网络设备解耦。

智算资源管理平台和 RoCE 网络管控系统间协同，完成参数面网络的自动化部署，以及基于开放的 RoCE 协议进行增强，提供通用、开放、高性价比的高性能无损方案，是解决上述困难的有效解决思路，但生态构建面临极大的困难和挑战，需要产业界共同长期努力推动。

基于 RoCE 提供一套开放、完善的 RoCE 解决方案是业界的目标。目前业界对于 RoCE 的拥塞控制制定基础协议外，解耦主要考虑的是网卡侧和 RDMA 网络侧进行解耦，即服务器与网络设备的解耦。目前 RoCE 组网解耦存在一定的困难，对于较大规模的组网，通常需要更复杂的拥塞控制算法或者流量调度策略，目前业界主要有两类解决思

路：一类是通过增强端网协同来实现更精细化的拥塞控制，比如 HPCC 等拥塞控制算法，依赖于网卡侧和交换机侧协同才能完成；第二类是交换机网络侧提供更好的流量调度能力，从而尽量避免流量在交换网络上发生拥塞。通常需要结合以上两类解决方案才能更好的解决大规模无损网络下的拥塞控制难题，目前 OLink 正在基于以上思路，推动业界共同努力实现标准化。

7.4 算力原生，打造异构算力解耦生态

随着智算技术的发展以及新兴应用的涌现，在 intel、NVIDIA、AMD 等传统行业巨头推出 AI 芯片的同时，一些创新芯片厂商也纷纷推出 AI 芯片解决方案。不同厂家围绕自身芯片架构构筑各自的软件生态，导致了各厂家软件生态的碎片化和竖井化。多厂家竖井化的封闭生态带来了跨架构的应用优化部署开发成本高，异构算力的合理规划和应用的动态迁移难度大等问题，从而导致资源利用率低以及难以构建良性发展的生态等问题。

因此，基于多种基础架构环境、多种 GPU 卡类型的异构开放环境是未来演进的方向。在初期，可以通过异构混池技术，提升资源利用率。同一资源池支持不同厂家的 GPU 资源管理及编排，资源池将不同厂家的卡进行分类管理及编排。应用在申请 GPU 资源时，不同框架的应用按需调度到兼容的厂家 GPU 资源上。

在下一个阶段，可以通过构建标准统一的算力抽象模型及编程范式接口，打造开放灵活的开发及适配平台，实现各类异构硬件资源与计算任务有效对接、异构算力与业务应用按需适配、灵活迁移，充分释放各类异构算力协同处理效力、加速智算应用业务

创新，实现异构算力资源一体池化、应用跨架构无感迁移、产业生态融通发展的算力原生架构。此时，真正实现了底层 GPU 异构资源细节的屏蔽，上层 AI 框架应用和底层 GPU 类型完全的解耦分离。

具体来说，算力原生包括了算力池化层和算力抽象层两部分。

算力池化层将各类硬件资源一体池化，并且通过构建底层异构硬件的统一抽象模型，重定向应用调用底层算力资源的请求，实现了通过统一定义的抽象智能算力度量值申请算力，从而屏蔽了异构硬件的差异。同时为应对智算业务的潮汐效应，算力池化层可根据业务需求及算力负载情况提供算力资源弹性扩缩容的能力。

算力抽象层由原生堆栈和原生接口组成，其中原生堆栈主要包括统一编程模型、跨架构编译和原生运行时，统一编程模型、跨架构编译可将基于特定芯片编程的应用程序转译为与底层硬件架构无关的算力原生中间表示（Intermediate Representation）；原生运行时可实现对底层算力资源的感知和控制，完成原生程序的加载、解析，保障计算任务与本地计算资源的即时互映射，按需执行。原生接口基于原生算力抽象接口及多模混合并行编程模型，构建原生算力统一 API、原生编程模型范式及原生编译优化部署工具，形成可嵌入式融入用户业务的开发环境，辅助用户生成可跨架构流转、无感迁移与任务式映射执行的算力原生程序。

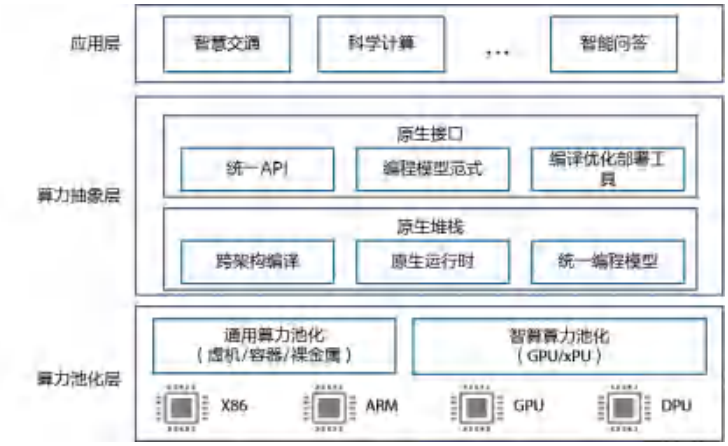


图 7-4 算力原生架构

7.5 分布式混池部署，满足核心网应用综合资源需求

由于核心网网元对通算及智算基础设施资源都有需求，同时训推应用也存在分布式部署的要求。因此通算和智算的混池部署以及分布式部署，成为 AI+网络云部署的特点。

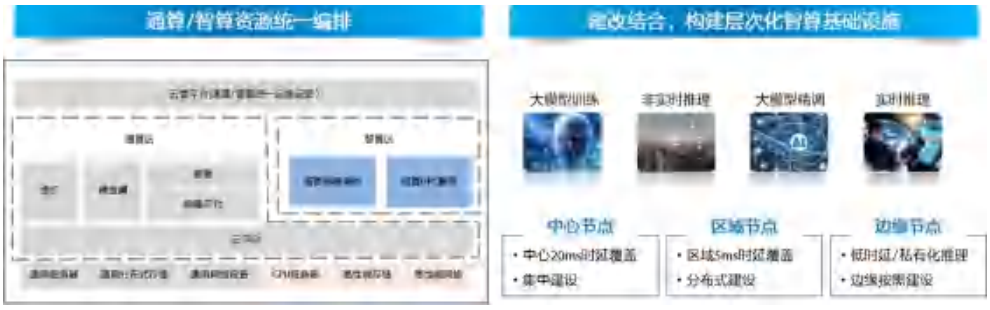


图 7-5 AI+网络云部署模式

网络云由通算资源池平滑升级到智算资源池，同时通算智算混池编排管理是网络云的一个重点特征。通常采用集中的云平台统一管理通/智算的算力、存储以及网络等基础设施资源。在初期，通算资源由传统通算网络云管理平台编排，智算资源由智算资源

运维运营平台编排。在系统成熟后，通算和智算资源由升级后的网络云管理平台统一编排。

由于基础大模型预训练、行业大模型精调以及客户场景大模型微调对算力特征及部署位置的要求均不同，结合运营商网络云层次化分布的架构，AI+网络云部署也呈现枢纽大模型训练中心、区域训推融合资源池、边缘训推一体机三级部署模式。其中枢纽大模型训练中心为集中新建的超大规模 GPU 集群，满足基础大模型预训练要求，提升大规模集群算效及能效、提升训练可靠性是其关键挑战；区域训推融合资源池通常以现网通算资源池按需扩展升级落地，满足大模型精调、推理以及应用部署为主，需要重点关注提升多样化资源管理效率和资源利用率、开放解耦能力和应用生态的挑战；边缘训推一体机主要满足企业私域数据的精调、以及推理及应用的部署，满足政企客户入驻场景快速部署，提供智算一体化服务是其主要特征。

8 AI Core 部署关键要素

当前业界核心网主要采用云化建设方案，以中心、区域、边缘三级数据中心为基础的物理网络架构，云化的核心网网元功能可按照场景部署在网络相应的位置上。当核心网引入“AI+”之后，需要考虑大模型的选择、场景化部署和端到端合规安全等几个关键要素。

8.1 360°评估体系，选择最优 AI 模型

AI 大模型在在自然语言处理、图像识别、语音识别等领域展现出了强大的能力，然而，对于运营商来说，大模型层出不穷，面对如此多的选择，如何根据核心网需求选择合

适的大模型？建议建立 360 度大模型评估体系，分别从模型评价指标、大模型框架及平台层面，以及大模型赋能应用场景等方面综合评估，帮助运营商选择合适的大模型进行部署。

在模型评价指标方面，模型的复杂度、泛化能力、稳定性、可控性等要素是评价模型并选择的主要指标。模型复杂度是模型前向的计算量和参数总量，复杂度越高模型的表达能力及拟合能力越强。泛化能力是模型在训练数据外的新数据上的性能表现，通常用泛化误差来衡量；稳定性是指模型在数据异常数据下的表现能力，稳定的模型不会因输入数据的波动而大幅波动；可控性是指模型的可解释能力、人类意志对齐能力以及法律道德伦对齐能力。具体选择时应基于任务的复杂度和实际需求来综合考虑，此外，模型是否开源、是否可商用、支持的语言种类、公开的测评结果等也可作为评估指标。

在大模型框架及平台层面，需要引入高质量大模型预训练及精调端到端流程，包括数据处理、预训练、精调、评估及推理部署全流程。数据管理及标注平台提供大模型预训练数据高效清洗，精调数据标注、评估功能；预训练平台提供一键创建多机多卡预训练作业、3D 并行方案自动规划能力；模型精调平台提供数据处理、预置主流大模型、支持多种精调手段，一键启动全流程自动化， 大幅降低精调复杂度并缩短周期；模型评估平台提供评测流水线平台，内置多个基准测试数据集，建立大模型自动化评测体系，一键输出评测报告；模型推理部署平台，提供模型量化、模型稀疏手段降低推理成本提升推理速度，并兼容多种后端平台，并支持大模型分布式推理部署。

同时，通信领域核心网大模型应用层面，应赋能业务智能化、网络智能化、运维智能

化、云基础设施智能化等各应用层次。在业务智能化方面，消息业务反欺诈、消息业务多模态行业大模型，新通话趣味交互、商务助理等均可带来创新的业务价值；在网络智能化领域，大模型可应用在精准的业务识别、重点业务保障、业务运营支撑分析等方面；在核心网及网络云资源基础设施层面，大模型可重点赋能复杂运维场景，需要各个智能体 Agent 灵活编排、通过 LUI 的方式进行人机交互、利用规划和工具调用能力，充分融合现网 AI 小模型能力及网络原子能力，实现运维体系的平滑演进。

8.2 建改结合，构建层次化智算基础设施

从人工智能应用场景来看，主要有训练和推理场景。训练和推理需要智算数据中心，智算数据中心与通算数据中心存在较大区别。相对于通算数据中心来说，智算数据中心主要由基于 GPU、NPU、ASIC 等 AI 芯片的加速算力组成，需要配套大功率机架、风冷或液冷散热，对网络时延和丢包要求比较高，一般都要求无收敛组网，资源管理需要支持任务和集群等新型管理机制。因此核心网引入“AI+”后，需要考虑如何在现有中心、区域、边缘三级通用算力数据中心布局架构下，如何将智算资源引入进来，支持核心网全场景的 AI 应用。

为了满足核心网 AI 引入需求，建议在核心网现有数据中心布局基础上，建改结合，将智算中心融入到通算数据中心，构建层次化智算基础设施。

中心节点：该类型节点主要用于核心网大模型预训练和推理，可以考虑新建和改造结合来支撑智算的升级。针对大模型预训练场景，采用新建智算数据中心方案，物理位置尽可能与通算数据中心共站址，即与核心网控制面部署位置相同；由于大模型预训

练对智算中心要求比较高，尤其是散热模式、组网方案与通算数据中心不同，因此通过集中新建大规模智算集群，采用液冷散热模式和 RoCE 组网，可快速低成本满足核心网大模型预训练使用。针对中心节点的推理场景，考虑到推理池和通算池在组网、存储、管理等方案商差别不大，尤其是推理智能算力可以在现有通算服务器上增加 PCIE 类型的 GPU 即可支持，因此可以通过改造现有通用算力数据中心，包括通算池扩容、升级、新建并举等方式，来支撑核心网控制面智能推理能力。

区域节点：该类型节点主要用于核心网控制面的大模型精调和推理，其特点是分布式按需扩展，多样化算力应用。因此可以通过改造现有通用算力数据中心，包括通算池扩容、升级、新建并举等方式，来支撑核心网控制面的精调和推理能力，同时通过编排器实现灵活的网络服务部署和动态的网络资源协同，从容应对各式各样的业务快速发展。

边缘节点：该类型节点主要为核心网用户面/媒体面或者垂直应用提供网络边缘计算和推理能力。针对纯核心网用户面场景，当前通用算力提供采用专用硬件方式提供，因此在引入推理算力后，可以采用通算智算一体机，为核心网用户面提供数据、模型、应用的一体化服务；针对 MEC 场景，主要为行业客户提供私域数据精调、推理和应用，因此可以为用户提供训推一体机，软硬一体规格化配置，快速入驻政企客户机房，端到端一体化服务，实现数据及行业大模型不出园区，提供差异化的体验，提高运营商增值收入。

8.3 层次化纵深防御安全体系，打造安全合规 AI

AI 技术与核心网的深度融合和风险叠加，对 AI 安全性提出更高要求。随着 AI 技术应用范围不断拓展， AI 本身的伦理道德风险、内容安全风险、隐私数据泄漏等威胁也日渐凸显，安全性是 AI 技术部署时的关注焦点；核心网作为关键基础设施，承载重要公共通信和提供信息服务，一旦遭到破坏、丧失功能或者数据泄露，会对国家安全、国计民生、公共利益产生重大影响。

通过构建以下层次化的纵深防御安全体系，可为 AI Core 体系提供防护能力：

- 基础设施层打造可信环境。针对多租户共用智算资源的安全风险，可部署多个边界安全设施，防止通用资源池中常见的安全攻击，满足法律法规中的安全防护能力要求；利用基于机密计算的隔离能力，构建智算能力所需的安全运行环境，确保智算关键信息可用不可见。
- 数据层面确保全程合法合规。为防止出现知识产权或版权等纠纷、个人隐私或商业数据泄露、数据污染投毒危害训练过程等安全风险，需构建完整的数据安全检查手段确保数据流转全程合规：数据来源方面，进行合法性检查，确保阻拦高风险数据，防止违规数据搜集；数据内容方面，通过专项内容核查，剔除违规、隐私和存在版权风险的数据；数据审计方面，对数据建立跟踪溯源、完整性检查机制，确保数据风险可审计可检查，及时处置问题数据源。
- AI 模型本身提供安全保障。针对模型和算法本身存在的可解释性差、窃取篡改和对抗样本攻击等风险，通过贯穿人工智能全过程的安全治理规范和风险控制措施进行安全防范：安全治理流程上，在设计、研发、部署、维护过程中建立并实施

安全开发规范，尽可能消除模型算法存在的安全缺陷、歧视性倾向；供应链安全管理上，跟踪人工智能涉及的软硬件产品漏洞、缺陷信息并及时采取修补加固措施，保证供应链安全性；模型风险管控上，为系统内部构造、推理逻辑、技术接口、输出结果提供明确说明，正确反映系统产生结果的过程，不断提高人工智能可解释性、可预测性。

- 内容安全保障输入输出合法。针对人工智能存在的非法查询与生成、违规内容输出等风险，利用多个安全机制确保内容合规合法：内容管控方面，建立围栏机制，重点输入和输出两个关键点，建立可定制化和持续更新的非法问法和非法内容集，确保用户无法进行非法内容询问和获取危害性答复；内容检测方面，具备文本、图片、视频等多种类型内容的检测能力，并可检测用户多个会话和同一会话上下文之间的内容安全，动态评估用户安全信誉度防止持续危害发生。

- 持续测评审计实战增强安全力：为了对人工智能产品进行实际安全表现检验与度量，基于已有的中兴通讯安全风险治理体系框架，利用 CEval、HumanEval 等数据集进行准确性和可靠性评估；基于 MITRE ALTA 等多种攻防模型和渗透性测试，对数据安全把控能力、算法模型攻击对抗能力、内容输入输出检测能力等多个方面进行安全性测试和度量，通过快速迭代改进实现默认安全能力的不断增强。

9 AI Core 实践

9.1 全球首个组装式”AI+”5G 新通话网络

随着 AI 技术的迅猛发展，终端厂商、OTT（Over-The-Top）服务商以及运营商纷纷加码 AI 应用，争夺智能化入口，推动产业价值重塑。运营商在这场竞争中，凭借庞大的用户规模和海量通话量，具备了成为 AI 入口的独特基础。与 OTT 应用不同，通话业务作为一种传统且普遍的通信方式，天然具备了不依赖 APP 安装、实时互动、低延迟等显著优势，这使得通话业务在 AI 时代拥有巨大的潜力和创新空间。

中兴通讯通过与运营商的紧密合作，将 AI 技术与通话业务深度融合，成功部署了全球首个组装式“AI+”新通话网络。目前已经开通了翻译、趣味通话等 6 种 AI+应用，旨在为运营商和用户提供更为智能、便捷的通信体验，极大地丰富了通话业务的功能和场景。这些应用不仅提升了用户的通话体验，也为企业提供了更高效的智能化客服解决方案。项目已开通实时翻译、字幕翻译、AI 速记、手势动效、表情语、背景替换、虚拟头像等 AI 应用；语音驱动数字人、点亮屏幕、人像风格、新通话座席、AR 标记等将陆续开通。

通过与运营商的深度合作，中兴通讯在全球首个组装式“AI+”5G 新通话网络中实现了 AI 与通信业务的深度融合。凭借开放的生态架构、智能原生能力、智能编排与动态加载的创新机制，以及丰富的 AI 应用，运营商能够在增强用户体验的同时，为企业和行业赋能，推动整个通信行业的智能化发展。

9.2 业界首个分层分级 VIP 用户保障商用

随着用户需求的日益个性化和多样化，传统的流量经营模式已无法满足现代运营商在激烈竞争中脱颖而出的要求。运营商不仅要关注流量的增长，还必须注重用户体验的保障，特别是在关键业务场景下（如直播、游戏、高铁专网等），如何提升网络服务的差异化和品质，成为决定运营商品牌价值和用户满意度的关键挑战。中兴通讯联合运营商成功验证了“AI+”连接的业务体验保障方案，实现了直播业务保障试点、高铁专网驻留保障试点，解决了现网中体验保障不足、业务优先级调度不精细等问题。

高优先级用户保障：对于高价值、VIP 用户，系统会自动提升其网络保障等级，确保他们在任何时刻都能享受到最佳的网络服务体验。针对重点业务（如直播、游戏、视频会议等），方案为每类业务设定了不同的 5QI 等级，以此实现速率保障、时延优化等针对性网络调度。例如，速率要求高的直播类业务将被优先保障带宽，而时延敏感的游戏业务则重点保障低时延。

高铁专网驻留保障：针对高铁场景，方案实现了低速区用户画像，识别高铁 VIP 用户，针对高铁用户下发无线保障策略，确保 VIP 用户在高速移动中的专网驻留，保证其数据服务的稳定性。此外，在经停站台场景、高速并线场景也进行了 VIP 用户的保障体验。经停站台场景：方案能够自动识别 VIP 用户在站台的驻留，优先保障其数据和语音通信的高优先级调度，避免用户体验受到干扰；高速并线场景：确保高铁专网内的 VIP 用户稳定驻留，快速识别并清除不必要的入侵用户，保障高铁专网的网络资源不被浪费。

9.3 网络云自智网络 L4 故障处理场景落地实践

- 实践 1：基于数字孪生的资源池倒换量化评估方案

资源池或数据中心（DC）级别的容灾备份机制在保障用户业务连续性和网络运行安全方面起着至关重要的作用。然而，当进行资源池或 DC 级别的容灾操作时，原资源池上的网元所服务的用户必须迅速迁移到其他可用的资源池或 DC。在这一迁移过程中，大量用户短时间内重新连接网络，不可避免地会引发信令冲击，从而导致 CPU 和网络资源出现急剧的高负载。

因此，中兴通讯与运营商合作首次提出了基于数字孪生的资源池倒换量化评估体系，结合网元和资源池之间的关系，对资源池倒换过程进行仿真和量化分析。

- 自动化数据采集与处理，评估速度快

通过全面自动化的数据采集与分析处理流程，我们将容灾评估时间压缩至 10 分钟以内。这一技术创新不仅大幅提高了决策响应速度，还有效减少了人工干预的复杂性，显著提升了数据处理效率。特别是在紧急情况下，运维人员能够迅速获取评估结果并做出决策，确保网络能够在最短时间内恢复正常，提升了网络的稳定性和应急响应能力。

- 秒级仿真粒度，容灾精准决策

引入秒级粒度的仿真计算，精准模拟网元与资源池的负荷冲击。与传统方法相比，这种仿真方法能够更真实地还原网络中实际冲击波形，从而大幅提高了评估准确性，准确度达到 95%以上。通过这种高精度的仿真技术，我们提升了倒换过程的可靠性，为紧急故障恢复提供了更加精确的决策依据，增强了系统的容灾能力。

– 高精度自适应模型，全程透明可解释

核心网和资源池冲击联动评估方案通过白盒仿真方式对网元层进行透明、可解释的建模。该模型具备强大的扩展性，可以根据不同倒换场景调整和优化评估方法，确保模型能够适应多种网络拓扑和流量变化。通过基于信令数量和拓扑映射的转换资源池流量仿真算法，提升了故障评估的准确度和自适应能力，进一步提升了容灾系统的智能化水平和应用范围。借助理论模型与 AI 学习结合的方法，我们解决了核心网信令冲击评估中的终端行为和省份业务差异问题，通过模型自适应调整，显著提高了评估准确性，确保了网络评估的精度高于 95%。这一方法不仅修正了传统计算偏差，还提高了网络在多变环境下的适应能力，为决策提供了可靠的数据支持。

通过能力开放嵌入到生产流程中，构建“预防+感知+应急+溯源”的立体防御体系，具备了事前预防能力，事中问题快速洞察恢复，事后精准溯源能力。

本案例在运营商现网进行了部署验证，通过 DC 容灾倒换验证，孪生系统从仿真计算的准确性，仿真评估时长均达到设计目标。孪生系统通过能力开放和上级网管对接，将能力嵌入生产流程。商用系统基于核心网标准信令交互冲击流程以及云化资源池北向数据上报规范进行倒换模型以及框架开发，系统可推广应用至各云化核心网容灾倒换场景，具有良好的复制转化能力。

- 实践 2：基于大模型与多智能体的网络故障诊断

现代电信网络日益复杂，随着 5G、AI 等技术的不断发展，网络云故障的预警、定位和处理也面临了前所未有的挑战。现有的故障管理系统难以应对大规模、动态变化的网络环境，导致了以下几个主要运维痛点：

在电信网络运维过程中，故障管理的效率和响应速度是关键因素。然而，目前的运维体系面临着多方面的挑战。首先，故障无法提前预警，大多数问题依赖客户投诉或上报告警才能被发现，导致决策延迟，从而错失了处理故障的最佳时机。其次，故障传递链路长，传递效率低，在复杂的网络环境中，故障的定位需要经过多个环节，通常需要较长时间来分析和定位问题，这不仅增加了故障定位的耗时，也加大了重大业务故障的风险。

即使是已知问题，现有系统依然需要依赖大量专家人力进行处理，造成 80%问题的重复劳动。由于缺乏智能化的人机交互式原因诊断功能，这些问题无法快速形成闭环，导致大量的时间和精力浪费。与此同时，现网日常运维报告生成往往需要每个环境 2-3 小时的人工操作，进一步加剧了人力资源的压力，并影响了运维效率。

在此基础上，重大操作方案的编写通常具有较强的专业性，涉及细节较多，且编写和修改周期长。一旦在细节上有所疏漏，可能导致操作失败，给网络运营带来严重影响。因此，整个故障管理流程面临着预警不及时、定位效率低下、专家资源浪费和报告生成效率低等多重问题，亟需通过智能化的手段来提升整体的运维效率和准确性。

为此我们提出了基于基于大模型与多智能体的网络故障诊断解决方案，通过智能化手段提升整体的运维效率和准确性，确保能够实时响应故障并进行快速修复。

- 多层多维数据感知，可视全息网络状态

网络云的运行状态涉及到多个维度的数据，包括物理层、虚拟层。通过多层次、多维度的数据融合技术，大模型可以整合来自不同层次的数据源，形成一个更加精确和全面的网络状态视图。

– 故障根因快速诊断，精准生成修复建议

在通用语言模型上通过混合生成领域语料与收集的推理语料进行 FT+SFT，训练成领域大模型，对于领域下游任务准确率提升 6%。基于大模型的思维链智能分析能力，系统可以快速诊断出故障的根本原因，大幅减少了误报和告警的数量，确保运维人员专注于真正的故障问题。通过中兴星云大模型、多智能体技术及 RAG 增强，系统提供精准的故障信息和针对性修复建议，帮助站点工程师快速定位和解决问题，大大缩短了工单处理时长，提升了故障响应效率。

– 故障自修复闭环，减少人工操作失误

基于中兴运维智能体，通过自我学习和自适应调整，可以自主调用相关工具或接口，自动完成修复闭环。这一自修复机制不仅能够在无人干预的情况下迅速恢复系统功能，还能确保修复过程的完整性和准确性，提升了网络的自动化运维水平，减少了人工操作错误。业界首创多智能体能够完成超过 70 次以上的任务动态分解，准确率 90%以上，解决了领域复杂任务智能，准确完成及落地。

9.4 消息反诈大模型助力涉诈短信案件量降低 64%

随着电信网络诈骗形式的不断变化，传统的反诈治理方案已无法灵活应对新型诈骗手段，且在诈骗短信的拦截和识别效率方面存在较大瓶颈。诈骗短信的内容和形式日益复杂，使得传统的基于关键词匹配的技术难以满足实时性和精准性要求。此外，许多反诈系统由于没有足够的智能化与自适应能力，导致对新型诈骗手段的识别较为滞后，拦截效果不尽人意。为了应对这一挑战，亟需引入先进的人工智能技术，特别是大模

型技术，以提高诈骗短信的识别准确率和拦截效率，从根源上减少诈骗信息对社会的危害。中兴通讯与上海移动合作，提出并实施了以“反诈大模型”为核心的创新性短信反诈治理方案。该方案通过引入泛化特征神经网络、SNS 社交特征分析等先进技术，从诈骗短信的意图、语义等深层次维度进行分析，为电信运营商提供了一种智能化、全方位的反诈治理解决方案。

1. 真实数据构建反诈大模型，模型可信

反诈大模型的可信性首先源自于其训练数据的质量和广度。为了确保模型在实际应用中的高效性和准确性，中兴通讯通过大量真实数据的采集和标注，构建了反诈大模型的核心数据集。这些数据包括了来自各个渠道的百万级诈骗短信样本，涵盖了各种诈骗类型、话术及形式。通过对这些诈骗短信内容的深度分析，模型能够学习并捕捉到诈骗短信的细微特征，并在此基础上建立起精准的分类和识别能力。

多样性多渠道数据收集：为了确保数据的代表性，所使用的诈骗短信样本来源于多个渠道，包括真实用户反馈、运营商反诈骗数据库、公共举报平台等。这样可以确保模型覆盖到尽可能多的诈骗场景和手法。

反诈大模型在开发过程中，通过交叉验证和真实场景的测试来验证其准确性。特别是对模型在不同运营商网络、不同地区和文化背景下的表现进行了测试，确保模型能够适应各种变异和地域性差异。

实时反馈与优化：在实际应用中，反诈系统会持续从用户和运营商的反馈中收集数据，动态优化模型，进一步提升其可信性。

2. 深度微调，减少大模型幻觉

大规模预训练的反诈大模型通常具备强大的语义理解和预测能力，但在应用过程中，可能会出现一些“幻觉”现象，即模型根据其学习的模式错误地生成不准确或无关的判断。为了减少这些幻觉现象，反诈大模型采用了深度微调技术，以提高其在特定反诈任务中的准确度。优化推理准确率与输出格式，幻觉发生概率低于百万分之一。

3. 语义挖掘，提升识别能力

诈骗短信的文本内容和表达方式具有很强的变异性，这意味着传统的基于关键词匹配的反诈技术很容易被欺骗。因此，为了提高反诈系统的识别能力，反诈大模型利用语义挖掘技术，深入分析短信的深层语义，从而更准确地识别潜在的诈骗信息。

自该系统上线后，境外涉诈案件数量明显降低，为减少人们财产损失、维护社会和谐做出贡献。未来，中兴通讯将持续加强新技术研究，深化合作和应用实践，进一步增强反诈大模型能力，助力运营商构建智能化、高安全的通信网络。

10 未来核心网智能化演进展望

面向 6G，中兴通讯坚持 AI+发展理念，以创新引领和降本增效为目标，聚焦业务、连接、运维和网络云基础设施等方面，持续不断研究业务创新技术。一方面实现 AI Core 助力运营商业务创新和降本增效；另一方面充当 AI 赋能者，满足行业用户智能化转型需求，助力国家实现新质生产力目标。当前 AI 创新技术层出不穷，在 5G 网络与 AI 技术深度融合实现智能化的转型过程中，需要产业界合作伙伴通力协作，探索智能化的平滑演进路径，携手促进智能化难题的解决：

1. 持续推进 NWDAF 功能的演进

在 3GPP 标准下，NWDAF 从 R15 到后续版本不断发展。R15 的 NWDAF 功能单一，仅支持网络切片负载分析。而 R16 及之后版本（R17、R18、R19）在不同层面进行了增强优化。例如，R16 定义了集中式架构，能满足基本数据分析要求；R17 实现了训 - 推分离式架构，定义了分析逻辑功能（AnLF）和模型训练逻辑功能（MTLF），还构建了支持多 NWDAF 协同的分层智能架构，并引入数据管理框架提升数据采集和分析效率。未来，NWDAF 可能会在分析能力、服务范围等方面进一步拓展，以更好地适应不断增长的业务需求。

2. 推进自智网络进一步发展

自智化运维通过运用大数据、人工智能等技术，实现网络运维的智能化、自动化和高效化。目前已经能够对网络数据实时收集分析，利用机器学习算法进行模式识别和预测分析，未来可能会在提升运维效率、网络稳定性等方面进一步优化。例如，在提升运维效率方面，可能会进一步减少人工操作依赖，更精准地自动执行运维任务；在提升网络稳定性上，可能会采用更先进的实时监控和智能分析技术，更迅速地处理网络故障。

3. 适应多样化业务的智能化发展

随着业务场景不断丰富，如不同的移动互联网数据业务（视频直播、视频会议、游戏等）对网络资源占用和体验保障需求不同，核心网智能化需要能够精准识别并保障不同业务的需求。例如，在无线网络资源负载较重时，核心网能够实时精准了解保障业务的客户体验 QoE。未来可能会发展出更智能的资源分配和保障机制，以适应更多样化的业务场景。核心网将加强与垂直行业的合作，满足不同行业的需求，推动数字化

转型和社会进步。智能化的核心网能够利用人工智能和大数据技术对海量数据进行处理和分析，挖掘潜在价值，优化资源配置和决策支持，未来可能会针对不同垂直行业的特殊需求，定制更智能化的网络服务。

4. 6G 核心网内生智能演进

6G 核心网连接平面将在 5G 核心网已有控制平面和用户平面基础上进一步增强，实现可编程能力和 6G 原生的 AI 能力。其演进的控制平面将沿用服务化架构并进一步解耦网络功能、实现网络服务能力的灵活按需调用。同时，6G 网络将支持分布式组网架构，网络功能发现和选择机制将进一步扩展，除网元级别外还将支持网络间的发现和选择能力。6G 核心网将在传统连接平面基础上引入新的数据平面和计算平面，以支撑 6G 核心网能力内生的演进需求，实现面向数据、计算、智能等服务资源的多要素协同服务能力，从而满足千行百业差异化需求和智算融合，促进云网边端业协同和产业生态繁荣发展。

总结，未来 AI Core 将在 5G-A 核心网智能化发展成果的基础上，朝着 6G 核心网内生智能演进，在架构、功能、运维等多方面不断发展，以满足日益多样化的业务需求和垂直行业需求，实现更高效、智能、灵活的网络服务，为运营商带来更多的价值空间。

11 缩略语

表 11-1 缩略语

缩略语	全称
2B	To Business 面向行业
2C	To Customer 面向消费者
3GPP	3rd Generation Partnership Project 第三代合作伙伴计划
5G	5th Generation Mobile Communication Technology 第五代移动通信技术
5GC	5G Core 5G 核心网
6G	6th Generation Mobile Communication Technology 第六代移动通信技术
AI	Artificial Intelligence 人工智能
AIGC	Artificial Intelligence Generated Content 人工智能生成内容
AGV	Automated Guided Vehicle 自动导引车辆
AMF	Access and Mobility Management Function
API	Application Programming Interface 应用编程接口
AR	Augmented Reality 增强现实
B2B	Business-to-Business 企业对企业
B2C	Business-to-Customer 企业对消费者
CIFS	Common Internet File System 公共互联网文件系统
CPU	Central Processing Unit 中央处理器
CV	Computer Vision 计算机视觉
CXL	Compute Express Link 计算快速链路
DPU	Data Processing Unit 数据处理单元
GPU	Graphics Processing Unit 图形处理器
GSMA	Global System for Mobile Communications Association 全球移动通信系统协会
FOA	First Office Application 首次商用测试
FPGA	Field - Programmable Gate Array 现场可编程门阵列
FTP	File Transfer Protocol 文件传输协议
GDS	GPU Direct Storage GPU 直接存储

缩略语	全称
H5	HTML5，HTML 第五代标准
HBD	High Band Domain 高带宽域
HDFS	Hadoop Distributed File System Hadoop 分布式文件系统
IB	InfiniBand
IMT-2030	International Mobile Telecommunications for 2030 面向 2030 的国际移动通信
I/O	Input/Output 输入/输出
iSCSI	Internet Small Computer System Interface 互联网小型计算机系统接口
OTT	Over The Top 视频及数据服务业务
POSIX	Portable Operating System Interface for UNIX 可移植操作系统接口（UNIX）
HPCC	High Precision Congestion Control 高精度拥塞控制
NFS	Network File System 网络文件系统
NAS	Network - Attached Storage 网络附属存储
NPU	Neural - Processing Unit 神经处理单元
NWDAF	NetWork Data Analytics Function 网络数据分析功能
MIG	Multi-Instance GPU 多实例 GPU
RDMA	Remote Direct Memory Access 远程直接内存访问
RoCE	RDMA over Converged Ethernet
ToC	To Consumer 个人消费者
ToB	To Business 行业与企业
ToH	To Home 家庭
ToO	To Other 其他
vGPU	Virtual GPU 虚拟 GPU
VoNR	Voice over New Radio 新空口语音通话
VR	Virtual Reality 虚拟现实