基于超图和遗传算法的AI推理服务驱动的多维资源编排与调度方法

景川芳，朱晓荣
（南京邮电大学通信与信息工程学院，江苏 南京 210003）

摘　要：为解决6G网络中AI推理服务的严格性能需求与多维资源复杂耦合难题，提出了一种基于超图和遗传算法的AI推理服务驱动的多维资源编排与调度方法。首先，利用超图准确捕捉服务层AI推理任务、逻辑层算法–数据–算力–连接资源、物理层基础设施间的高阶依赖关系。其次，形成了跨层多维资源动态匹配优化问题，通过联合分配AI推理任务、涉及多维资源的微服务和智能体，实现服务需求与网络资源间的最佳匹配。同时，提出了一种改进的非支配排序遗传算法-II，利用超边预分类特性灵活缩小解维度，提高算法收敛速度。仿真结果表明，所提方法在平衡推理时延与推理精度方面优于对比算法，验证了所提算法的有效性和实用性。

关键词：6G网络；AI推理服务；多维资源；编排与调度；超图
中图分类号：TN92
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025180

Multi-dimensional resource orchestration and scheduling method driven by AI inference services via hypergraph and genetic algorithm

JING Chuanfang, ZHU Xiaorong
School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: To tackle the stringent performance requirements of AI inference services and the complex coupling of multi-dimensional (MR) resources in 6G networks, a hypergraph and genetic algorithm-based method for AI inference service-driven MR orchestration and scheduling was proposed. Firstly, hypergraph accurately captured the high-order dependency relationships among AI inference tasks in the service layer, algorithm-data-computing-connection resources in the logical layer, and infrastructure in the physical layer. Secondly, a cross-layer MR dynamic matching optimization problem was formulated. The problem aimed to achieve the best match between service demands and network resources through joint allocation of AI inference tasks, microservices involving MR, and intelligent agents. Meanwhile, an improved non-dominated sorting genetic algorithm-II was proposed. By leveraging the pre-classification characteristics of hyperedges to flexibly reduce the solution dimension, the convergence speed of the algorithm was improved. Simulation results demonstrate that the proposed method outperforms the comparative algorithms in balancing inference latency and inference accuracy, thus validating the effectiveness and practicality of the proposed algorithm.

Keywords: 6G network, AI inference service, multi-dimensional resource, orchestration and scheduling, hypergraph

收稿日期：2025–05–26；**修回日期：**2025–10–10
通信作者：朱晓荣，xrzhu@njupt.edu.cn
基金项目：国家科技重大专项基金资助项目（No.2024ZD1300400）；国家自然科学基金资助项目（No.92367102）；江苏省研究生科研与实践创新计划基金资助项目（No.KYCX22_0944）
Foundation Items: The National Science and Technology Major Project (No.2024ZD1300400), The National Natural Science Foundation of China (No.92367102), The Postgraduate Research & Practice Innovation Program of Jiangsu Province (No.KYCX22_0944)

0　引言

面向2030年及未来，人类社会将迈入以“万物智联”为核心的智能化时代[1]。作为智能化社会的信息基础设施，6G网络将不仅致力于实现人机物智慧互联与智能体高效互通，更将通过原生智能架构为用户提供无处不在的智能服务[2]。特别地，随着自动驾驶、低空网络、扩展现实等新兴垂直行业的蓬勃发展，这些行业对智能化、个性化服务的迫切需求将推动以实时视频分析和边缘智能决策为代表的人工智能（AI）推理服务逐渐成为6G网络关键应用场景[3]。与AI训练服务不同，AI推理服务核心目标是利用已有训练模型对新数据做出分析和决策[4]。此外，AI推理服务具有突发性强、单次计算量小等特点，并对6G网络通信实时性、推理精度和可靠性提出了更为严苛的要求，如1 ms时延、10 Gbit/s传输速率和99.5%的目标检测精度[5-6]。因此，复杂多样的AI推理服务性能需求将对6G网络AI服务质量（QoAIS, quality of AI service）保障构成全新挑战。

有效的多维资源编排与调度方法是解决上述挑战的一种有前景的思路。然而，如何根据AI推理用户的实际需求，灵活编排和调度通信、计算、数据和AI模型资源，确保AI推理服务稳定运行，是一个十分棘手的问题，主要体现在以下2个方面。

1) AI推理服务驱动的多维资源高阶依赖关系快速建立。AI推理服务QoAIS保障涉及算法、算力、数据和连接四大核心资源，这些资源的存在形式、实现功能和规格各不相同，并且资源之间存在复杂的非线性耦合关系。例如，高推理精度需要高复杂度算法、高质量数据和大算力，低推理时延需要低复杂度算法、大带宽、高算力等。若能在资源调度前，精准建立多维资源间的高阶依赖和交互关系，实现为面向不同AI推理服务快速提供符合其性能要求的多维资源候选集合，对于优化资源分配和提高资源利用率十分重要。

2) 多类型AI推理服务共存场景下有限多维资源高效调度。AI推理服务对时延和精度要求极高，但低时延与高精度存在天然矛盾。此外，在现实应用场景中，多种具有个性化服务需求的AI推理用户将共存于6G网络。例如，医疗影像分析与自动驾驶共存，前者对精度敏感但对时延容忍度高，后者对时延与精度需求均极为苛刻[7-8]。因此，支持多类型AI推理服务共存场景下的高效资源调度方案是十分必要的，该方案需在低时延、高精度等多目标冲突中实现动态权衡，并将任务与适当的资源进行快速匹配，以确保任务的最佳执行以及资源的灵活和有效调度。

针对多维资源高阶依赖关系建立挑战，大多数研究人员专注于在各种应用场景下直接形成优化问题。例如，文献[9]研究类人网络中包括通信、计算和内存资源在内的短期和长期联合资源分配问题。在空中边缘AI推理场景下，文献[10]联合感知功率分配、发射预编码和接收波束，提出了一种面向任务的集成感知、计算和通信方案。显然，现有研究忽略了AI推理服务与算法、数据、算力和连接资源间的高阶协同依赖关系，难以支撑6G网络为复杂多样的AI推理用户提供精细化的多维资源编排方案。近年来，基于超图的资源分配方法受到了产业界和学术界的广泛关注。超图中的超边能连接任意数量的顶点，已广泛应用于显式刻画工业互联网、智能信息物理运输系统、低空卫星等大规模复杂网络中多维顶点间的高阶关系[11-13]。此外，超边具有灵活生成特性，多类型超边包含更多顶点间的高阶交互关系，显示出其嵌套性、异质性、多层多维等多个属性特征[14]。因此，超图能克服传统图模型的局限性，支持以AI推理服务为中心的6G网络多维资源高阶复杂依赖关系建模。

针对多目标冲突下的有限资源调度挑战，研究人员也展开了多方面尝试，主要为了优化AI推理任务划分[15]、AI推理模型拆分与压缩[16-17]、AI推理模型部署[18]并联合分配网络中的通信、计算、感知和内存等资源，力求最大化所有AI推理任务的推理精度，同时最小化推理时延或能耗等性能指标。同时，为了解决这些问题，已有研究提出了李雅普诺夫优化[19]、启发式[20]、强化学习[17]等方法。相较于李雅普诺夫优化与强化学习方法，启发式方法在解决多目标优化问题时具有独特优势，其不需要构建复杂函数、训练大量样本以及频繁调整超参数等操作，因此该方法复杂度较低并且能在多项式时间内得到次优解。尽管启发式方法有上述优势，但在实际应用中仍存在不足。现有启发式方法大多仅适用于求解特定场景下单一AI推理任务的多目标冲突，未考虑端–边–云协同场景下保障多类型AI推理任务差异化QoAIS需求所带来的复杂性。

受上述挑战和研究激励，本文首先设计了统一的AI推理服务驱动的多维资源编排与调度架构，该架构由服务层、逻辑资源层和物理层组成，在每一层抽象了原子级要素，即服务层AI推理任务、逻辑资源层微服务（MS, microservice）和物理层智能体。同时，设计了跨层嵌套超图模型，以统一表达服务需求、逻辑资源与物理资源间的复杂关系。接着，考虑不同类型AI推理服务对时延和精度的性能需求，建立了跨层多维资源分配问题，其中服务层、逻辑资源层和物理层要素被整体考虑以协调资源分配。同时，本文提出了一种基于非支配排序遗传算法-II（NSGA-II, non-dominated sorting genetic algorithm II）的启发式方法，可快速提供Pareto前沿，为6G网络多维资源编排与调度提供全新的视角和解决方案。

本文主要贡献如下。

1) 设计了以AI推理服务为中心的跨层嵌套超图模型，该模型包括服务层AI推理任务–AI流超图、逻辑资源层微服务–虚拟功能（VF, virtual function）超图、物理层智能体–簇超图以及服务层–逻辑资源层–物理层跨层超图，面向不同AI推理服务的性能需求刻画了任务–算力–算法–数据–连接等多维资源间的高阶依赖关系，实现了各层元素灵活表征、动态聚合和复用。

2) 基于构建的跨层嵌套超图模型，形成了以AI推理服务为中心的层间多维资源动态调度优化问题，通过联合分配AI推理任务、微服务和智能体以最小化AI推理请求所需时延、精度与网络供给时延、精度间差值的绝对值。同时，提出了一种改进的NSGA-II来求解该问题，该算法利用预先设计的跨层超边合理划分维度并仅对跨层超边中的顶点进行种群初始化、选择、交叉和变异等操作，大大减少了算法搜索空间，加快了算法收敛速度。

3) 仿真结果表明，本文方法能为AI推理请求提供更匹配其性能需求的多维资源调度方案，从而更适用于以AI推理服务为中心的6G网络多维资源编排与调度。

1 相关工作

1.1 超图建模高阶关系

近年来，超图在表征多模态结构和数据高阶关系方面取得了很大进展，并逐渐用于复杂网络建模及网络优化。文献[11]研究了跨小区工业物联网的联合信道和功率分配问题，通过将原始问题转化为超图模型以优化信道和功率分配。在智能信息物理运输系统中，文献[12]利用超图理论对客户间的重叠干扰关系进行建模，将复杂干扰下的干扰避免资源分配推广到超图顶点强着色问题。文献[13]提出了一个多向扩展超图来准确地描述卫星网络中通信域、观测域等典型功能域资源的时空特征和功能属性。文献[21]利用超图准确捕获任务与通信资源和计算资源之间的关系，垂直统一了物联网系统的计算、通信和任务方面，并实现了资源有效匹配。文献[22]提出了一种基于超图的攻击威胁狩猎模型，该模型将输入的网络威胁情报日志图和物联网系统内核审计日志图编码为超图。文献[23]提出了面向跨域协同作战体系的复杂超图表示方法，构造了每个顶点和每条超边都伴随一个用来反映其对应要素和关系的定量、定性属性向量的赋权着色超图。文献[24]提出了基于异构图书评论信息融合模型的超图聚类方法，抽取图书和用户作为两类顶点，梳理文本评论、评分评级和标签标注3种用户行为中的异构信息，并抽取节点间的4类高阶交互关系作为超边，实现了异构数据和高阶交互关系的融合表示。针对未来6G网络场景多样化、网络要素多元化等带来的网络架构设计和部署复杂度升高的难题，文献[25]利用超图理论对6G网络业务域、逻辑域和物理域域内及域间的复杂关系进行详细建模，该研究着重关注使用超图对6G网络架构建模以及基于该模型对6G网络架构进行评估和优化，而本文着重利用超图构建差异化AI推理服务性能指标需求对多维资源的高阶依赖关系。

1.2 面向AI推理服务的资源调度算法

在即将到来的6G时代，端边云协作模式将对时延、精度和可靠性有严苛需求的AI推理服务提供有效支持。然而，如何保障AI推理服务的QoAIS并利用该模式来提高AI推理服务的性能是一个巨大挑战。针对此挑战，文献[9]考虑QoAIS要求和人工智能任务，基于类人网络服务管理和资源调度架构形成了联合人工智能代理放置的深度神经网络部署和动态带宽资源分配问题，并提出了李雅普诺夫优化求解该混合整数非线性规划问题。在边缘智能系统中，文献[15]研究了模型训练和任务推理过程，以优化终端用户的本地CPU频率、任

务拆分率和系统带宽分配，并提出了一种基于乘子交替方向法的面向分解的方法。文献[16]提出了一种用于加速边缘智能体中拆分推理的有效通信和计算资源分配算法，该算法基于梯度下降寻找推理时延和能耗之间的最优权衡。文献[17]考虑无线传感系统中的端边协同推理，制定了混合整数非线性规划来联合优化模型分裂点和系统资源分配，并分别通过深度强化学习和凸优化方法来解决组合模型分裂和资源分配问题。文献[18]研究如何分区和压缩卷积神经网络，以便在端边云协作系统中进行高精度、快速推理，实现端到端推理时延和准确性间的权衡。文献[19]首次尝试共同考虑设备推理、服务器推理或边缘设备协作推理3种模式，并通过协调不同模式下的传感、通信和计算资源，实现了在给定的推理精度和其他网络资源约束下最大限度地降低系统能源成本。文献[20]提出了一种新的服务编排多范式部署模型，并将模型中的服务部署表述为一个多目标优化问题。为了解决该问题，本文提出了一种基于加权度量的启发式方法，该方法可以有效地获得近似最优的Pareto边界。

综上所述，凸优化、深度强化学习和启发式方法已广泛用于解决端边云协作推理模式下的多维资源调度问题。然而，凸优化方法数学建模与推导过程复杂，且对动态拓扑适应性差，无法实时调整资源分配策略。深度强化学习方法虽能适应动态环境，但需要大规模训练数据和漫长的训练周期，难以满足6G网络实时推理服务需求。相比之下，启发式方法（如蚁群算法、多目标粒子群算法、NSGA-II）能够依据问题的特定结构和目标特征，通过一些启发式规则，在多项式时间内高效得到次优解。其中，NSGA-II天然具备解决多目标优化问题的能力，不需要像蚁群算法和多目标粒子群算法那样对多维目标进行加权求和或标量化处理，并且处理离散变量的能力更优，解集在Pareto前沿分布均匀。然而，在面对大规模优化问题时，随着变量数量的增加以及约束与变量的高度耦合，NSGA-II收敛性急剧下降。

为了有效解决上述问题，本文提出了结合超边预分类特性的NSGA-II改进算法，通过超边聚类将满足不同约束的变量划分为若干子集，在交叉或变异时优先保留同一超边组内的有效变量组合，约束校验时仅需校验当前超边组内的关系，避免了全局无效搜索。同时，更新后的解只需与同超边组的历史解比较支配关系，避免全局Pareto排序的重复计算，加快了算法收敛速度。此外，超边预分类对应链路带宽、算法精度等约束，生成的解自然满足“任务–算法–数据–算力–连接”多维资源联合约束。当拓扑中某类资源（如边缘节点算力）变化时，仅需更新对应超边组的变量。因此，结合超边预分类特性的NSGA-II通过局部更新即可适应新变化，并且可有效降低搜索空间维度，避免全局遍历和重复排序过程，加快了算法收敛速度。

2　系统模型

2.1　AI推理服务驱动的6G网络多维资源编排架构

面向具有复杂系统特性的智能应用场景，6G网络多维资源编排架构应屏蔽基础网络差异，为新兴AI推理服务提供通信、算力、数据和算法一体化的资源按需服务[25]。图1展示了AI推理服务驱动的6G网络多维资源编排架构，该架构利用软件定义网络与网络功能虚拟化技术抽象各层要素并封装为与场景无关的原子能力，通过开放更丰富、更细粒度的能力，实现多维资源的充分共享，快速且低成本地满足多领域新兴智能应用的差异化需求。该架构分为3层，分别为服务层、逻辑资源层和物理层，各层元素如下。

1) 服务层。将各行各业千差万别的AI推理用户性能需求转化为6G网络可理解的对AI推理服务能力的要求，并进一步将AI推理服务分解为多个AI推理任务，AI推理任务作为服务层最小粒度的原子能力。如图1所示，AI推理任务包括视频解码、文本生成、意图识别、轨迹预测等，适用于自动驾驶、工业互联网等智能应用场景[3]。

2) 逻辑资源层。将算力、连接、算法和数据资源从6G网络底层设施中抽象出来形成对应的4个子资源层。各子资源层将封装具有不同功能的微服务作为原子能力。如图1所示，连接子层提供接入、路由等微服务；算力子层提供发现、寻址等微服务；数据子层提供采集、清洗等微服务；算法子层提供推理、剪枝等微服务。

3) 物理层。如图1所示，物理层采用端–边–云协同架构，提供算法、算力、数据和连接资源的物理设备均可看作智能体。单智能体作为最小粒度的原子能力，分为：① 物理终端设备，如机器人、汽车，具备有限的多维资源供给能力；② 边缘侧设备，

如基站、交换机，多维资源供给能力中等；③ 云侧设备，如云服务器，具备较强的多维资源供给能力。

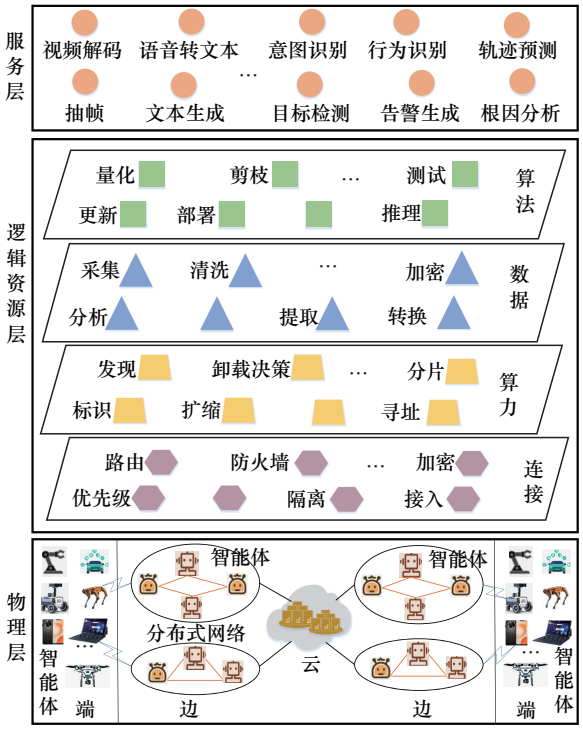


图 1 AI 推理服务驱动的 6G 网络多维资源编排架构

2.2 跨层嵌套超图

面向不同智能应用场景，6G 网络需具备个性化和定制化 AI 推理服务能力，这要求其依据 AI 推理请求性能需求，协同编排调度服务层 AI 推理任务、逻辑资源层微服务与物理层智能体。然而，服务层、逻辑资源层和物理层相互独立，层内元素的存在形式、实现功能、可提供的资源类型和规格各不相同。为了快速从细粒度和多元化的各层元素集合中挑选出满足 QoAIS 需求的元素，需提前捕捉层内不同粒度性能指标元素的按需组合关系，以及 AI 推理服务与层间不同粒度元素的适配关系。显然，传统图论中的节点间成对关系无法准确描述上述复杂高阶依赖关系。例如，若 AI 推理任务 1 满足 AI 推理服务 1 的时延和精度需求，那么这两者之间存在一条链路，表示两者适配。然而，若 AI 推理任务 1 由多个 AI 推理任务组成，则该链路并不能表示 AI 推理任务 1 与其他 AI 推理任务之间的组合关系，也不能表示该组合的性能指标粒度，更不能表示 AI 推理任务 1 与 AI 推理服务 1 的适配程度。

因此，为了捕捉上述任务–算力–算法–数据–连接等多维元素间的复杂高阶关系，本文构建了跨层嵌套超图，如图 2 所示，包括层内超图和跨层超图。层内超图用来表征不同粒度性能指标下各层元素的按需组合关系，各层顶点和超边如下。如图 2(a)所示，在服务层内，顶点为 AI 推理任务，超边为 AI 推理任务的组合方式。在逻辑资源层内，顶点为微服务，超边为微服务的组合方式。在物理层内，顶点为智能体，超边为智能体的组合方式。跨层超图用来表征 AI 推理服务与层间不同元素组合方式间的适配关系。如图 2(b)所示，跨层超图中顶点为服务层 AI 推理流、逻辑资源层 VF 和物理层簇，超边为不同组合方式下的 AI 推理流、VF 和簇与 AI 推理服务的适配程度。接下来，分别详细描述层内超图和跨层超图的构建方法。

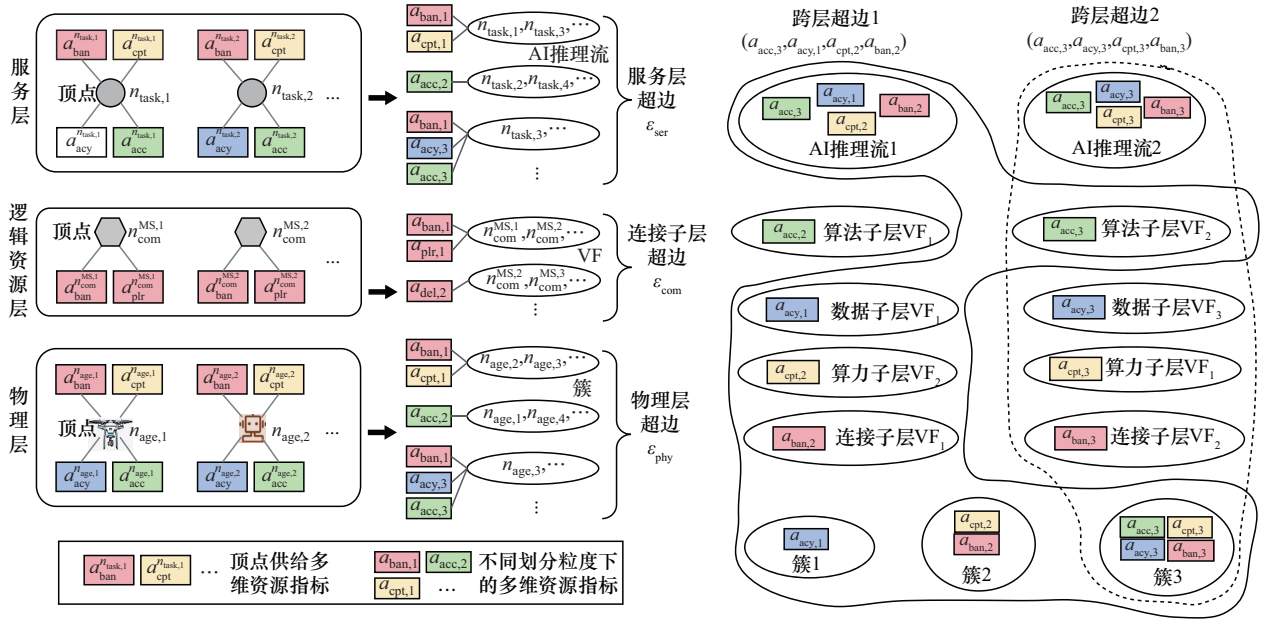
2.2.1 层内超图

1) 服务层。假设 AI 推理任务集合为 $\mathcal{N}_{\text{task}}=\{1,\cdots,N_{\text{task}}\}$，AI 推理任务 n_{task} 由二元组 $\{\text{at}^{n_{\text{task}}}_{\text{type}},\mathbf{at}^{n_{\text{task}}}_{\text{QoAIS}}\}$ 表示，其中 $\text{at}^{n_{\text{task}}}_{\text{type}}$ 表示 n_{task} 的类型，$\mathbf{at}^{n_{\text{task}}}_{\text{QoAIS}}$ 表示 n_{task} 对不同资源的性能指标需求，$\mathbf{at}^{n_{\text{task}}}_{\text{QoAIS}}$ 表示为

$$\mathbf{at}^{n_{\text{task}}}_{\text{QoAIS}}=\begin{bmatrix}\left\{a^{n_{\text{task}}}_{\text{acc}},a^{n_{\text{task}}}_{\text{F1}},a^{n_{\text{task}}}_{\text{con}},\cdots\right\}^{1}_{\text{QoAIS}}\\ \left\{a^{n_{\text{task}}}_{\text{acy}},a^{n_{\text{task}}}_{\text{rec}},a^{n_{\text{task}}}_{\text{com}},\cdots\right\}^{2}_{\text{QoAIS}}\\ \left\{a^{n_{\text{task}}}_{\text{cpt}},a^{n_{\text{task}}}_{\text{err}},a^{n_{\text{task}}}_{\text{pre}},\cdots\right\}^{3}_{\text{QoAIS}}\\ \left\{a^{n_{\text{task}}}_{\text{ban}},a^{n_{\text{task}}}_{\text{del}},a^{n_{\text{task}}}_{\text{plr}},\cdots\right\}^{4}_{\text{QoAIS}}\end{bmatrix} \tag{1}$$

其中，$\{\cdot\}^{1}_{\text{QoAIS}}$ 表示与算法资源相关的性能指标集合，包括 n_{task} 的精度 $a^{n_{\text{task}}}_{\text{acc}}$、F1 分数 $a^{n_{\text{task}}}_{\text{F1}}$ 和收敛性 $a^{n_{\text{task}}}_{\text{con}}$ 等。$\{\cdot\}^{2}_{\text{QoAIS}}$ 表示与数据资源相关的性能指标集合，包括准确性 $a^{n_{\text{task}}}_{\text{acy}}$、召回率 $a^{n_{\text{task}}}_{\text{rec}}$ 和完整性 $a^{n_{\text{task}}}_{\text{com}}$ 等。$\{\cdot\}^{3}_{\text{QoAIS}}$ 表示与算力资源相关的性能指标集合，包括计算能力 $a^{n_{\text{task}}}_{\text{cpt}}$、算力能效比 $a^{n_{\text{task}}}_{\text{err}}$ 和算力精度 $a^{n_{\text{task}}}_{\text{pre}}$ 等。$\{\cdot\}^{4}_{\text{QoAIS}}$ 表示与连接资源相关的性能指标集合，包括通信带宽 $a^{n_{\text{task}}}_{\text{ban}}$、时延 $a^{n_{\text{task}}}_{\text{del}}$ 和丢包率 $a^{n_{\text{task}}}_{\text{plr}}$ 等。

如图 2(a)所示，服务层若干 AI 推理任务按不同粒度的多维资源指标组合为 AI 推理流。例如，$n_{\text{task},1}$ 和 $n_{\text{task},3}$ 被划分进带宽为 $a_{\text{ban},1}$ 和计算能力为 $a_{\text{cpt},1}$ 的 AI 推理流中，表示其带宽与计算能力符合指标要求。因此，在服务层构建 AI 推理任务–AI 推理流超图 $\mathcal{G}_{\text{ser}}=(\mathcal{V}_{\text{ser}},\mathcal{E}_{\text{ser}})$，其中顶点集合 $\mathcal{V}_{\text{ser}}$ 包括所



(a) 层内超图　　(b) 跨层超图

图2　跨层嵌套超图

有AI推理任务，即 $\mathcal{V}_{\text{ser}}=\mathcal{N}_{\text{task}}$，超边集合 $\mathcal{E}_{\text{ser}}$ 由提供差异化QoAIS的AI推理流类型组成。定义二元变量 $\alpha_{n_{\text{task}},n_{\text{flo}}}$ 表示AI推理流 n_{flo} 是否包含 n_{task}，n_{flo} 中顶点集合定义为 $\mathcal{N}_{n_{\text{flo}}}=\{n_{\text{task}}|\alpha_{n_{\text{task}},n_{\text{flo}}}=1,n_{\text{task}}\in\mathcal{N}_{\text{task}}\}$，使用AI推理流属性的超边组 $\mathcal{E}_{\text{ser}}$ 表示为

$$\mathcal{E}_{\text{ser}}=\{\mathcal{N}_{n_{\text{flo}}}|n_{\text{flo}}\in\mathcal{N}_{\text{flo}}\} \tag{2}$$

2) 逻辑资源层。假设逻辑资源层微服务集合为 $\mathcal{N}_{\text{MS}}=\{\mathcal{N}_{\Lambda}^{\text{MS}}|\Lambda\in\{\text{alg,data,cop,com}\}\}$，其中 Λ 表示某个子层的资源属性，即算法、数据、算力和连接资源，$\mathcal{N}_{\Lambda}^{\text{MS}}$ 表示 Λ 子层的微服务集合。与服务层类似，每个资源子层的微服务均由二元组表示。例如，连接子层微服务 $n_{\text{com}}^{\text{MS}}$ 由二元组 $\{\text{at}_{\text{type}}^{n_{\text{com}}^{\text{MS}}},\mathbf{at}_{\text{QoAIS}}^{n_{\text{com}}^{\text{MS}}}\}$ 表示，其中，$\text{at}_{\text{type}}^{n_{\text{com}}^{\text{MS}}}$ 表示 $n_{\text{com}}^{\text{MS}}$ 的类型，$\mathbf{at}_{\text{QoAIS}}^{n_{\text{com}}^{\text{MS}}}$ 表示 $n_{\text{com}}^{\text{MS}}$ 提供的连接性能指标值（只与该子层资源属性相关）。那么，4个资源子层微服务共同组成的性能指标矩阵 $\mathbf{at}_{\text{QoAIS}}^{\text{MS}}$ 表示为

$$\mathbf{at}_{\text{QoAIS}}^{\text{MS}}=\begin{bmatrix}\left\{\mathbf{at}_{\text{QoAIS}}^{1},\cdots,\mathbf{at}_{\text{QoAIS}}^{n_{\text{alg}}^{\text{MS}}},\cdots\right\}_{\text{QoAIS}}^{1}\\ \left\{\mathbf{at}_{\text{QoAIS}}^{1},\cdots,\mathbf{at}_{\text{QoAIS}}^{n_{\text{data}}^{\text{MS}}},\cdots\right\}_{\text{QoAIS}}^{2}\\ \left\{\mathbf{at}_{\text{QoAIS}}^{1},\cdots,\mathbf{at}_{\text{QoAIS}}^{n_{\text{cop}}^{\text{MS}}},\cdots\right\}_{\text{QoAIS}}^{3}\\ \left\{\mathbf{at}_{\text{QoAIS}}^{1},\cdots,\mathbf{at}_{\text{QoAIS}}^{n_{\text{com}}^{\text{MS}}},\cdots\right\}_{\text{QoAIS}}^{4}\end{bmatrix} \tag{3}$$

其中，算法子层微服务 $n_{\text{alg}}^{\text{MS}}$、数据子层微服务 $n_{\text{data}}^{\text{MS}}$、算力子层微服务 $n_{\text{cop}}^{\text{MS}}$ 和连接子层微服务 $n_{\text{com}}^{\text{MS}}$ 对应的性能指标矩阵 $\mathbf{at}_{\text{QoAIS}}^{n_{\text{alg}}^{\text{MS}}}$、$\mathbf{at}_{\text{QoAIS}}^{n_{\text{data}}^{\text{MS}}}$、$\mathbf{at}_{\text{QoAIS}}^{n_{\text{cop}}^{\text{MS}}}$ 和 $\mathbf{at}_{\text{QoAIS}}^{n_{\text{com}}^{\text{MS}}}$ 分别为

$$\mathbf{at}_{\text{QoAIS}}^{n_{\text{alg}}^{\text{MS}}}=\left\{a_{\text{acc}}^{n_{\text{alg}}^{\text{MS}}},a_{\text{F1}}^{n_{\text{alg}}^{\text{MS}}},a_{\text{con}}^{n_{\text{alg}}^{\text{MS}}},\cdots\right\} \tag{4}$$

$$\mathbf{at}_{\text{QoAIS}}^{n_{\text{data}}^{\text{MS}}}=\left\{a_{\text{acy}}^{n_{\text{data}}^{\text{MS}}},a_{\text{rec}}^{n_{\text{data}}^{\text{MS}}},a_{\text{com}}^{n_{\text{data}}^{\text{MS}}},\cdots\right\} \tag{5}$$

$$\mathbf{at}_{\text{QoAIS}}^{n_{\text{cop}}^{\text{MS}}}=\left\{a_{\text{cpt}}^{n_{\text{cop}}^{\text{MS}}},a_{\text{err}}^{n_{\text{cop}}^{\text{MS}}},a_{\text{pre}}^{n_{\text{cop}}^{\text{MS}}},\cdots\right\} \tag{6}$$

$$\mathbf{at}_{\text{QoAIS}}^{n_{\text{com}}^{\text{MS}}}=\left\{a_{\text{ban}}^{n_{\text{com}}^{\text{MS}}},a_{\text{del}}^{n_{\text{com}}^{\text{MS}}},a_{\text{plr}}^{n_{\text{com}}^{\text{MS}}},\cdots\right\} \tag{7}$$

若干微服务按不同粒度的多维资源指标组合为VF，这些VF按照一定顺序（串行或并行）执行。由于逻辑资源层各子层MS组合形成VF的方式类似，图2(a)以连接子层为例，连接子层微服务 $n_{\text{com}}^{\text{MS},1}$ 和 $n_{\text{com}}^{\text{MS},2}$ 被划分进带宽为 $a_{\text{ban},1}$ 和丢包率为 $a_{\text{plr},1}$ 的VF中，表明其带宽和丢包率符合指标要求。因此，逻辑资源层构建MS-VF超图 $\mathcal{G}_{\text{loc}}=(\mathcal{V}_{\text{loc}},\mathcal{E}_{\text{loc}})$，其中顶点集合 $\mathcal{V}_{\text{loc}}$ 由MS组成，$\mathcal{V}_{\text{loc}}=\mathcal{N}_{\text{MS}}$，超边集合 $\mathcal{E}_{\text{loc}}$ 由提供差异化QoAIS的VF类型组成。假设逻辑资源层VF集合为 $\mathcal{N}_{\text{VF}}=\{\mathcal{N}_{\Lambda}^{\text{VF}}|\Lambda\in\{\text{alg,data,cop,com}\}\}$，其中 $\mathcal{N}_{\Lambda}^{\text{VF}}$ 表示 Λ 子层的VF集合。定义二元变量 $\beta_{n_{\Lambda}^{\text{MS}},n_{\Lambda}^{\text{VF}}}$ 表示 Λ 子层的虚拟功能 n_{Λ}^{VF} 是否包含微服务 n_{Λ}^{MS}，n_{Λ}^{VF} 中顶点集合定义为 $\mathcal{N}_{n_{\Lambda}^{\text{VF}}}=\{n_{\Lambda}^{\text{MS}}|\beta_{n_{\Lambda}^{\text{MS}},n_{\Lambda}^{\text{VF}}}=1,$

$n_{\mathcal{A}}^{\mathrm{MS}} \in \mathcal{N}_{\mathcal{A}}^{\mathrm{MS}}\}$，使用 VF 属性的超边组 $\mathcal{E}_{\mathrm{loc}}$ 表示为

$$\mathcal{E}_{\mathrm{loc}} = \{\, \mathcal{E}_{\mathcal{A}}, \mathcal{A} \in \{\, \mathrm{alg,data,cop,com} \,\}\} = \{\, \mathcal{N}_{n_{\mathcal{A}}^{\mathrm{VF}}} | n_{\mathcal{A}}^{\mathrm{VF}} \in \mathcal{N}_{\mathcal{A}}^{\mathrm{VF}}, \mathcal{A} \in \{\, \mathrm{alg,data,cop,com} \,\}\} \quad (8)$$

其中，$\mathcal{E}_{\mathcal{A}}$ 表示 $\mathcal{A}$ 子层使用 VF 属性的超边组。

3) 物理层。假设智能体集合为 $\mathcal{N}_{\mathrm{age}} = \{1,\cdots,N_{\mathrm{age}}\}$，智能体 n_{age} 的标识定义为二元组 $\{\mathrm{at}_{\mathrm{type}}^{n_{\mathrm{age}}}, \mathbf{at}_{\mathrm{abi}}^{n_{\mathrm{age}}}\}$，其中 $\mathrm{at}_{\mathrm{type}}^{n_{\mathrm{age}}}$ 表示 n_{age} 的类型，$\mathbf{at}_{\mathrm{abi}}^{n_{\mathrm{age}}}$ 表示 n_{age} 的拓扑结构和性能指标提供能力，$\mathbf{at}_{\mathrm{abi}}^{n_{\mathrm{age}}}$ 定义为

$$\mathbf{at}_{\mathrm{abi}}^{n_{\mathrm{age}}} = \begin{bmatrix} \left\{a_{\mathrm{loc}}^{n_{\mathrm{age}}}, a_{\mathrm{sub}}^{n_{\mathrm{age}}}, a_{\mathrm{adj}}^{n_{\mathrm{age}}}, \cdots\right\}_{\mathrm{top}} \\ \left\{a_{\mathrm{acc}}^{n_{\mathrm{age}}}, a_{\mathrm{F1}}^{n_{\mathrm{age}}}, a_{\mathrm{con}}^{n_{\mathrm{age}}}, \cdots\right\}_{\mathrm{QoAIS}}^{1} \\ \left\{a_{\mathrm{acy}}^{n_{\mathrm{age}}}, a_{\mathrm{rec}}^{n_{\mathrm{age}}}, a_{\mathrm{com}}^{n_{\mathrm{age}}}, \cdots\right\}_{\mathrm{QoAIS}}^{2} \\ \left\{a_{\mathrm{cpt}}^{n_{\mathrm{age}}}, a_{\mathrm{err}}^{n_{\mathrm{age}}}, a_{\mathrm{pre}}^{n_{\mathrm{age}}}, \cdots\right\}_{\mathrm{QoAIS}}^{3} \\ \left\{a_{\mathrm{ban}}^{n_{\mathrm{age}}}, a_{\mathrm{del}}^{n_{\mathrm{age}}}, a_{\mathrm{plr}}^{n_{\mathrm{age}}}, \cdots\right\}_{\mathrm{QoAIS}}^{4} \end{bmatrix} \quad (9)$$

其中，$\{\cdot\}_{\mathrm{top}}$ 表示拓扑结构，包括位置 $a_{\mathrm{loc}}^{n_{\mathrm{age}}}$、从属关系 $a_{\mathrm{sub}}^{n_{\mathrm{age}}}$ 和邻接关系 $a_{\mathrm{adj}}^{n_{\mathrm{age}}}$ 等。

若干智能体按不同粒度的多维资源指标组合成簇，各簇之间供给多维资源的能力不同。例如，图2(a)中物理层 $n_{\mathrm{age},2}$ 和 $n_{\mathrm{age},3}$ 被划分进带宽为 $a_{\mathrm{ban},1}$ 和计算能力为 $a_{\mathrm{cpt},1}$ 的簇中，表示其带宽和计算能力符合指标要求。因此，物理层构建智能体–簇超图 $\mathcal{G}_{\mathrm{phy}} = \left(\mathcal{V}_{\mathrm{phy}}, \mathcal{E}_{\mathrm{phy}}\right)$，其中顶点集合 $\mathcal{V}_{\mathrm{phy}}$ 由智能体组成，$\mathcal{V}_{\mathrm{phy}} = \mathcal{N}_{\mathrm{age}}$，超边集合 $\mathcal{E}_{\mathrm{phy}}$ 由簇类型组成。定义二元变量 $\gamma_{n_{\mathrm{age}},n_{\mathrm{clu}}}$ 表示簇 n_{clu} 是否包含 n_{age}，n_{clu} 中顶点集合定义为 $\mathcal{N}_{n_{\mathrm{clu}}} = \{\, n_{\mathrm{age}} | \gamma_{n_{\mathrm{age}},n_{\mathrm{clu}}} = 1, n_{\mathrm{age}} \in \mathcal{N}_{\mathrm{age}}\}$，使用簇属性的超边组 $\mathcal{E}_{\mathrm{phy}}$ 表示为

$$\mathcal{E}_{\mathrm{phy}} = \{\, \mathcal{N}_{n_{\mathrm{clu}}} | n_{\mathrm{clu}} \in \mathcal{N}_{\mathrm{clu}}\} \quad (10)$$

2.2.2 跨层超图

为了提供一次具有 QoAIS 保障的 AI 推理服务，需同时协同调配服务层 AI 推理任务和 AI 推理流、逻辑资源层各子层 MS 和 VF 以及物理层智能体和簇。虽然层内超图既包含各层元素的按需组合关系，又包含多维资源性能指标划分粒度。但是，层内超图并不涉及跨层元素的复杂组合关系以及不同组合方式下元素提供的多维资源性能指标与 AI 推理服务性能需求间的高阶适配关系。跨层要素与 AI 推理服务间是否存在高阶适配关系以及这种关系的强弱程度是由 AI 推理服务 QoAIS 需求决定的。跨层要素与 AI 推理服务的适配程度越高越能保证 6G 网络提供个性化和定制化 AI 推理服务的能力，还能提升 6G 网络多维资源的利用效率。

因此，为了表征跨层元素间的复杂组合关系以及不同组合方式提供的多维资源指标与 AI 推理服务性能需求间的高阶适配关系，本节设计了跨层超图 $\mathcal{G}_{\mathrm{int}} = \left(\mathcal{V}_{\mathrm{int}}, \mathcal{E}_{\mathrm{int}}\right)$，其中 $\mathcal{V}_{\mathrm{int}}$ 为跨层顶点集合，包括服务层 AI 推理流集合、逻辑资源层 VF 集合和物理层簇集合，$\mathcal{V}_{\mathrm{int}} = \mathcal{E}_{\mathrm{ser}} \cup \mathcal{E}_{\mathrm{loc}} \cup \mathcal{E}_{\mathrm{phy}}$，$\mathcal{E}_{\mathrm{int}}$ 为超边集合。一个跨层超边对应一个具有指定多维资源指标的 AI 推理服务。如图 2(b) 所示，2 个具有不同 QoAIS 指标的 AI 推理服务构成 2 条跨层超边，跨层超边 1 中包含提供对应 QoAIS 指标的服务层 AI 推理流 1、算法子层 VF2、数据子层 VF1、算力子层 VF1、连接子层 VF1 及物理层簇 1、簇 2 和簇 3，算法子层 VF2 同时属于跨层超边 1 和跨层超边 2。值得注意的是，每个 AI 推理服务涉及不同的虚拟功能，如数据预处理、模型部署、推理计算等。本文并不特别关注数据预处理（如清洗、归一化等）或模型部署（如分布式、集中式等）等功能采用的方式，而只关注这些功能提供的与多维资源相关的性能指标值，以便对 3 层元素进行总体安排。

3 问题形成

跨层超图能精准建立 AI 推理服务与三层多维元素间的高阶适配关系，实现为面向不同 AI 推理请求快速提供符合其性能要求的多维资源候选集合，有助于在资源受限的 6G 网络中最大限度地提高成功服务的 AI 推理请求数量。因此，本节基于构建的跨层超图，形成了以 AI 推理服务为中心的层间多维资源动态调度优化问题，通过联合分配 AI 推理任务、涉及多维资源的微服务、智能体以实现 AI 推理服务所需性能与 6G 网络供给性能间的最佳匹配。本文假设 6G 网络 AI 推理请求同时存在于端侧、边侧和云侧，端边云三侧 AI 推理请求集合为 $\mathcal{I} = \{1,\cdots,I\}$，第 i 个 AI 推理请求定义为 $U_i = \{\, D_i, B_i, T_i, \Delta t_i, \sigma_i, \Delta\sigma_i, \tau_i, \cdots\}$，其中 D_i 表示数据大小，B_i 表示带宽需求，T_i 表示时延需求，Δt_i 表示抖动需求，σ_i 表示精度需求，$\Delta\sigma_i$ 表示精度偏差需求，τ_i 表示数据质量。端边云三侧 AI 推理请求的区别在

于这些性能指标的差异。定义二进制变量$x_i^{n_\varphi}$表示元素n_φ是否分配给AI推理请求i，$\varphi \in \{\text{task},\Lambda,\text{age}\}$表示各层元素名称。当为AI推理请求$i$分配元素$n_\varphi$时，其所能提供的带宽、时延和精度必须满足用户需求，对应的带宽、时延和精度约束分别为

$$\text{C}_1: x_i^{n_\varphi} a_{\text{ban}}^{n_\varphi} \geqslant B_i, \varphi \in \{\text{task},\text{com},\text{age}\} \tag{11}$$

$$\text{C}_2: \left| \sum_{n_\varphi \in \mathcal{N}_\varphi} x_i^{n_\varphi} a_{\text{del}}^{n_\varphi} - T_i \right| \leqslant \Delta t_i, \varphi \in \{\text{task},\text{com},\text{age}\} \tag{12}$$

$$\text{C}_3: \left| \sum_{n_\varphi \in \mathcal{N}_\varphi} x_i^{n_\varphi} a_{\text{acc}}^{n_\varphi} - \sigma_i \right| \leqslant \Delta\sigma_i, \varphi \in \{\text{task},\text{alg},\text{age}\} \tag{13}$$

当用户请求映射至AI推理流时，AI推理流间无关联关系且可并行执行，但是AI推理流包含的AI推理任务有先后执行顺序。当这些AI推理任务映射到逻辑资源子层MS和物理层智能体时，智能体执行MS为AI推理任务服务的顺序也应该遵循任务执行顺序，并且应避免环路。因此，本文采用有向无环图来建模AI推理任务–MS–智能体间的先后执行顺序。如图3所示，AI推理任务1、2和3的执行顺序为1 → 2 → 3，分别映射到算法–数据–算力–连接子层MS1、MS2和MS3，这些MS再映射到智能体1、3、4和5，此时智能体执行MS为AI推理任务1、2和3服务的顺序应符合任务执行顺序，为1 → 3 → 4 → 5。假设AI推理流n_{flo}中的AI推理任务执行顺序为$n_{\text{task},1} \to n_{\text{task},2} \to \cdots$，则映射的MS集合为$\{n_\Lambda^{\text{MS},1}, n_\Lambda^{\text{MS},2}, \cdots\}$，当这些MS映射到智能体后，智能体服务AI推理任务的顺序为$n_{\text{age},1} \to n_{\text{age},2} \to \cdots$。定义二进制变量$\delta_{n_\Lambda^{\text{MS},1}}^{n_{\text{task},1}}$表示AI推理任务$n_{\text{task},1}$映射至逻辑资源层$\Lambda$子层的微服务$n_\Lambda^{\text{MS},1}$，二进制变量$\rho_{n_{\text{age},1}}^{n_\Lambda^{\text{MS},1}}$表示$n_\Lambda^{\text{MS},1}$映射至智能体$n_{\text{age},1}$。定义AI推理任务开始被智能体服务的时间点为该任务在智能体上的最早开始时间点。假设$n_{\text{task},1}$在$n_{\text{age},1}$上的最早开始时间点为$\text{ET}_{n_{\text{age},1}}^{n_{\text{task},1}}$，被$n_{\text{age},1}$服务完并传输到下一个智能体$n_{\text{age},2}$的时间点为$\text{CT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}}$，$\text{CT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}}$由$n_{\text{task},1}$在$n_{\text{age},1}$上的最早开始时间点$\text{ET}_{n_{\text{age},1}}^{n_{\text{task},1}}$、计算时间$\text{BT}_{n_{\text{age},1}}^{n_{\text{task},1}}$、排队等待传输时间$\text{WT}_{n_{\text{age},1}}^{n_{\text{task},1}}$以及$n_{\text{task},1}$到$n_{\text{age},2}$的传输时间$\text{BT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}}$组成，表示为

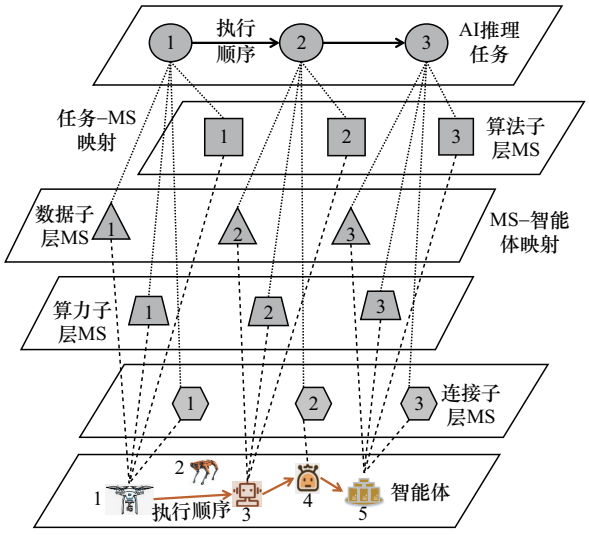


图3　AI推理任务–MS–智能体有向无环图

$$\text{CT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}} = \text{ET}_{n_{\text{age},1}}^{n_{\text{task},1}} + \text{BT}_{n_{\text{age},1}}^{n_{\text{task},1}} + \text{WT}_{n_{\text{age},1}}^{n_{\text{task},1}} + \text{BT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}} \tag{14}$$

其中，

$$\text{BT}_{n_{\text{age},1}}^{n_{\text{task},1}} = \frac{D_{n_{\text{task},1}}}{a_{\text{cpt},n_{\text{task},1}}^{n_{\text{age},1}}} \tag{15}$$

$$\text{WT}_{n_{\text{age},1}}^{n_{\text{task},1}} = \sum_{n_{\text{task}}=1}^{\mathcal{N}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}}} \text{BT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task}}} \tag{16}$$

$$\text{BT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}} = \frac{\text{DT}_{n_{\text{task},1}}}{\sum_{n_{\text{age}}=n_{\text{age},1}}^{\mathcal{N}_{n_{\text{age},1} \to n_{\text{age},2}}} a_{\text{rat},n_{\text{task},1}}^{n_{\text{age}}}} \tag{17}$$

其中，$D_{n_{\text{task},1}}$为$n_{\text{task},1}$数据大小，$a_{\text{cpt},n_{\text{task},1}}^{n_{\text{age},1}}$为$n_{\text{age},1}$分配给$n_{\text{task},1}$的计算资源，$\mathcal{N}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}}$为$n_{\text{age},1}$传输列表中比$n_{\text{task},1}$优先级高的任务集合，$\text{DT}_{n_{\text{task},1}}$为$n_{\text{task},1}$被$n_{\text{age},1}$服务完需传输的数据大小，$\mathcal{N}_{n_{\text{age},1} \to n_{\text{age},2}}$为$n_{\text{task},1}$从$n_{\text{age},1}$传输到$n_{\text{age},2}$经历的节点集合（不包括$n_{\text{age},2}$），$a_{\text{rat},n_{\text{task},1}}^{n_{\text{age}}}$为$n_{\text{age}}$分配给$n_{\text{task},1}$的传输速率。对于同一AI推理流中的先后任务$n_{\text{task},1}$和$n_{\text{task},2}$，$\text{CT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}}$应该小于或等于$n_{\text{task},2}$在$n_{\text{age},2}$上的最早开始时间点$\text{ET}_{n_{\text{age},2}}^{n_{\text{task},2}}$，表示为

$$\text{C}_4: x_i^{n_{\text{task},1}} x_i^{n_\Lambda^{\text{MS},1}} x_i^{n_{\text{age},1}} \delta_{n_\Lambda^{\text{MS},1}}^{n_{\text{task},1}} \rho_{n_{\text{age},1}}^{n_\Lambda^{\text{MS},1}} \text{CT}_{n_{\text{age},1} \to n_{\text{age},2}}^{n_{\text{task},1}} \leqslant x_i^{n_{\text{task},2}} x_i^{n_\Lambda^{\text{MS},2}} x_i^{n_{\text{age},2}} \delta_{n_\Lambda^{\text{MS},2}}^{n_{\text{task},2}} \rho_{n_{\text{age},2}}^{n_\Lambda^{\text{MS},2}} \text{ET}_{n_{\text{age},2}}^{n_{\text{task},2}} \tag{18}$$

对于映射至同一MS和智能体的任意2个任务m_{task}和n_{task}，若n_{task}优先级高于m_{task}，则m_{task}在n_{age}

上的最早开始时间点应小于或等于 n_{task} 在 n_{age} 上的计算结束时间点，表示为

$$\begin{aligned}&\mathrm{C}_5{:}\delta_{n_A^{\text{MS}}}^{m_{\text{task}}}\rho_{n_{\text{age}}}^{n_A^{\text{MS}}}\mathrm{ET}_{n_{\text{age}}}^{m_{\text{task}}}\leqslant\\&\delta_{n_A^{\text{MS}}}^{n_{\text{task}}}\rho_{n_{\text{age}}}^{n_A^{\text{MS}}}(\mathrm{ET}_{n_{\text{age}}}^{n_{\text{task}}}+\mathrm{WC}_{n_{\text{age}}}^{n_{\text{task}}}+\mathrm{BT}_{n_{\text{age}}}^{n_{\text{task}}})\end{aligned}\tag{19}$$

其中，$\mathrm{WC}_{n_{\text{age}}}^{n_{\text{task}}}$ 为 n_{task} 在 n_{age} 上排队等待计算的时间，表示为

$$\mathrm{WC}_{n_{\text{age}}}^{n_{\text{task}}}=\sum_{n'_{\text{task}}=1}^{\mathcal{N}_{n_{\text{age}}}^{n_{\text{task}}}}\frac{D_{n'_{\text{task}}}}{a_{\text{cpt},n'_{\text{task}}}^{n_{\text{age}}}}\tag{20}$$

此外，任意时刻在 n_{age} 上运行的所有任务所需计算资源不得超过其最大计算资源，即

$$\mathrm{C}_6{:}\sum_{i=1}^{\mathcal{I}}\sum_{n_{\text{task}}=1}^{\mathcal{N}_i}x_i^{n_{\text{task}}}x_i^{n_A^{\text{MS}}}x_i^{n_{\text{age}}}\delta_{n_A^{\text{MS}}}^{n_{\text{task}}}\rho_{n_{\text{age}}}^{n_A^{\text{MS}}}a_{\text{cpt},n_{\text{task}}}^{n_{\text{age}}}\leqslant a_{\text{cpt}}^{n_{\text{age}}}\tag{21}$$

同时，在 n_{age} 上传输的所有任务占用的带宽总和不能超过 n_{age} 的最大带宽，即

$$\mathrm{C}_7{:}\sum_{i=1}^{\mathcal{I}}\sum_{n_{\text{task}}=1}^{\mathcal{N}_i}x_i^{n_{\text{task}}}x_i^{n_A^{\text{MS}}}x_i^{n_{\text{age}}}\delta_{n_A^{\text{MS}}}^{n_{\text{task}}}\rho_{n_{\text{age}}}^{n_A^{\text{MS}}}a_{\text{ban},n_{\text{task}}}^{n_{\text{age}}}\leqslant a_{\text{ban}}^{n_{\text{age}}}\tag{22}$$

本文的目标是最小化 AI 推理请求所需推理时延、推理精度与 6G 网络供给推理时延、推理精度间差值的绝对值，以实现 AI 推理请求与多维资源的有效匹配。根据式(12)，AI 推理请求 i 所需推理时延与 AI 推理任务、连接子层 MS、智能体供给推理时延间差值的绝对值表示为

$$\Delta t_{i,\varphi}=\left|\sum_{n_\varphi\in\mathcal{N}_\varphi}x_i^{n_\varphi}a_{\text{del}}^{n_\varphi}-T\right|,\varphi\in\{\text{task,com,age}\}\tag{23}$$

根据式(13)，AI 推理请求 i 所需推理精度与 AI 推理任务、算法子层 MS、智能体供给推理精度间差值的绝对值表示为

$$\Delta\sigma_{i,\varphi}=\left|\sum_{n_\varphi\in\mathcal{N}_\varphi}x_i^{n_\varphi}a_{\text{acc}}^{n_\varphi}-\sigma_i\right|,\varphi\in\{\text{task,alg,age}\}\tag{24}$$

根据式(23)和式(24)，6G 网络供给推理时延差值的绝对值总和与推理精度差值的绝对值总和分别表示为

$$\Delta t=\sum_{i=1}^{\mathcal{I}}\left(\min_{\varphi\in\{\text{task,com,age}\}}\Delta t_{i,\varphi}\right)\tag{25}$$

$$\Delta\sigma=\sum_{i=1}^{\mathcal{I}}\left(\min_{\varphi\in\{\text{task,com,age}\}}\Delta\sigma_{i,\varphi}\right)\tag{26}$$

因此，服务层–逻辑资源层–物理层跨层多维资源匹配优化问题可表示为

$$\mathrm{P1}{:}\min_{\boldsymbol{x},\boldsymbol{\delta},\boldsymbol{\rho}}(\Delta t,\Delta\sigma)\tag{27}$$

$$\text{s.t.}\mathrm{C}_1\sim\mathrm{C}_7$$

$$\mathrm{C}_8{:}\sum_{n_\varphi\in\mathcal{N}_\varphi}x_i^{n_\varphi}=1,\varphi\in\{\text{task,alg,com,age}\}\tag{28}$$

$$\mathrm{C}_9{:}\sum_{n_A^{\text{MS}}\in\mathcal{N}_A^{\text{MS}}}x_i^{n_{\text{task}}}x_i^{n_A^{\text{MS}}}\delta_{n_A^{\text{MS}}}^{n_{\text{task},1}}=1\tag{29}$$

$$\mathrm{C}_{10}{:}\sum_{n_{\text{age}}\in\mathcal{N}_{\text{age}}}x_i^{n_A^{\text{MS}}}x_i^{n_{\text{age}}}\rho_{n_{\text{age}}}^{n_A^{\text{MS}}}=1\tag{30}$$

其中，约束 C_8 表示 AI 推理请求 i 提供服务时三层元素必须被协同分配。约束 C_9 表示当一个 AI 推理任务被映射到逻辑资源层某个子层时，只能映射至该层的一个微服务。约束 C_{10} 表示一个微服务只能映射到物理层的一个智能体上。优化问题 P1 符合 0~1 整数非线性规划形式，是 NP-hard 问题，难以在多项式时间内得到最优解[26]。此外，它还是个多目标优化问题，不同类型 AI 推理请求对时延和精度的需求并不一致且存在冲突，很难找到一个能够同时使所有目标达到最优的解决方案。大多数研究者根据预先给定的权衡确定线性权重，将多目标优化问题转化为单一目标优化问题，但是权重很难确定，因为精度和时延在不同应用场景下极具差异。因此，要求请求者先验地确定多个目标之间令人满意的权衡是不现实的[27]。因此，本文引入 Pareto 优势来求解问题 P1。

4 所提算法

NSGA-II 天然具备解决多目标优化问题的能力，并且处理离散变量的能力更优，其主要原因为：1) 采用非支配排序直接处理多个冲突目标间 Pareto 关系，不需要将多个目标加权求和或分解；2) 采用拥挤距离比较算子确保解在目标空间均匀分布，避免算法陷入局部最优解或产生聚集解；3) 采用精英保留策略避免优秀个体丢失。此外，Pareto 前沿的每个点都是优化问题的解决方案，决策者可根据实际需求进行选择[28]。然而，直接使用 NSGA-II 求解优化问题 P1 是不现实的。因为 NSGA-II 执行选择、交叉、变异等操作时遍历了所有种群及染色体中每个基因，导致 Pareto 前沿随机生成，而这些解大多违反约束，并且在搜索空间中分布无序。

为了解决这些问题，本文考虑结合所提跨层嵌套超图中的超边预分类特性灵活缩小解维度。超边

预分类特性是指超边能天然形成问题分解边界的能力。例如，服务层需要10 ms推理时延的AI推理任务可归类到一个超边中，物理层提供99%推理精度的物理设备可归类到另一个超边中。层内超边的合理划分可将复杂联合优化问题P1分解为若干子空间，每个子空间对应一组强相关的二进制决策向量$\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$。此外，跨层超边天然表达了层间资源的匹配规则，如GPU密集型AI推理服务必须映射到支持GPU的虚拟功能和物理设施上，这种匹配规则对应解决二进制决策向量$\boldsymbol{\delta},\boldsymbol{\rho}$。因此，在NSGA-II各个阶段利用超边预分类特性，可以为高效优化提供先验知识，加快推理速度，不仅能降低问题维度、提高算法收敛性，还能保留变量间的自然关联性。本节主要描述对NSGA-II的改进，以更好地解决优化问题P1，具体改进方面如下。

4.1　结合超边预分类的改进操作

4.1.1　解维度缩小操作

本文采用0~1编码方法，种群数量Pop即为染色体数量，被分配的三层元素的数量即为每个染色体上的基因数。针对优化问题P1，需同时为I个AI推理请求分配N_{task}个AI推理任务、N_{MS}个微服务和N_{age}个智能体，对应3个优化变量$x_i^{n_{\text{task}}}$、$x_i^{n_A^{\text{MS}}}$和$x_i^{n_{\text{age}}}$。本文利用超边预分类特性降低解的搜索空间，即确定每个跨层超边中分配的三层元素的种类和数量。假设I个AI推理服务的时延要求最大值$T_{\max}$与最小值$T_{\min}$的差为ΔT，推理精度要求最大值$\sigma_{\max}$和最小值$\sigma_{\min}$的差为$\Delta\sigma$。为了提取服务层–逻辑资源层–物理层顶点的差异化性能指标供给能力，首先将ΔT和$\Delta\sigma$分别划分为W_1和W_2个区间，其中$w_1\in\{0,1,\cdots,W_1-1\}$，$w_2\in\{0,1,\cdots,W_2-1\}$。其次，将服务层满足$|a_{\text{del}}^{n_{\text{task}}}-T_{\min}|\in\left[\frac{w_1}{W_1},\frac{w_1+1}{W_1}\right]\Delta T$且$|a_{\text{acc}}^{n_{\text{task}}}-\sigma_{\min}|\in\left[\frac{w_2}{W_2},\frac{w_2+1}{W_2}\right]\Delta\sigma$的AI推理任务集合$\mathcal{N}_{\text{task}}(w_1,w_2)$、逻辑资源层满足$|a_{\text{del}}^{n_A^{\text{MS}}}-T_{\min}|\in\left[\frac{w_1}{W_1},\frac{w_1+1}{W_1}\right]\Delta T$且$|a_{\text{acc}}^{n_A^{\text{MS}}}-\sigma_{\min}|\in\left[\frac{w_2}{W_2},\frac{w_2+1}{W_2}\right]\Delta\sigma$的微服务集合$\mathcal{N}_{\text{MS}}(w_1,w_2)$、物理层满足$|a_{\text{del}}^{n_{\text{age}}}-T_{\min}|\in\left[\frac{w_1}{W_1},\frac{(w_1+1)}{W_1}\right]\Delta T$且$|a_{\text{acc}}^{n_{\text{age}}}-\sigma_{\min}|\in\left[\frac{w_2}{W_2},\frac{(w_2+1)}{W_2}\right]\Delta\sigma$的智能体集合$\mathcal{N}_{\text{age}}(w_1,w_2)$归类到具有$(w_1,w_2)$属性的跨层超边$e(w_1,w_2)$中，$e(w_1,w_2)$表示为

$$e(w_1,w_2)=\{\mathcal{N}_{\text{task}}(w_1,w_2)\cup\mathcal{N}_{\text{MS}}(w_1,w_2)\cup\mathcal{N}_{\text{age}}(w_1,w_2)|w_1\in\{1,2,\cdots,W_1\},w_2\in\{1,2,\cdots,W_2\}\}\tag{31}$$

其中，W_1和W_2均为可调正参数。那么，面向AI推理请求的跨层超边集合$\mathcal{E}_{\text{int}}$可表示为

$$\mathcal{E}_{\text{int}}=\{e(w_1,w_2)|w_1\in\{1,2,\cdots,W_1\},w_2\in\{1,2,\cdots,W_2\}\}\tag{32}$$

跨层超边预分类伪码在算法1中给出（见附录1）。

4.1.2　其他操作细节

在划分好的跨层超边中进行选择、交叉、变异、非支配排序和拥挤距离计算操作，具体细节如下。

1) 选择。本文采用二元锦标赛选择方法，限定竞争个体来自同一超边子空间，即在同一超边覆盖的解空间内进行锦标赛。

2) 交叉。对每条超边覆盖三层变量进行两点交叉以产生新的染色体，即以交叉率ϕ在服务层交换AI推理任务对应位置值、在逻辑资源层交换微服务对应位置值、在物理层交换智能体对应位置值。同时要进行跨边保护，禁止破坏跨层超边性能指标约束$C_1\sim C_3$以及任务执行、资源竞争约束$C_4\sim C_7$的交叉。

3) 变异。对某一超边同层内变量施加关联扰动，父母的基因将根据变异率ψ随机突变以产生后代。

4) 非支配排序。在每条超边对应的目标子空间中采用标准的非支配排序方法独立计算非支配等级。

5) 拥挤距离计算。在每条超边对应的目标子空间中独立计算拥挤距离，计算每个优化目标上的归一化距离并求和，并添加边界保护，确保极端解（如最小或最大精度解）不被丢弃。

结合超边预分类特性的NSGA-II（NSGA-II-Hyper）的伪码如算法2所示（见附录2）。

4.2　算法复杂度分析

NSGA-II-Hyper的时间复杂度主要由算法1和算法2迭代过程中的目标评估以及快速非支配操作组成。算法1包括外部2个嵌套的for循环（大小为W_1、W_2）以及内部3个并列的for循环（大小为N_{task}、N_{MS}和N_{age}），时间复杂度为$O(W_1W_2(N_{\text{task}}+$

$N_{\mathrm{MS}}+N_{\mathrm{age}}$))。在算法2中，假设超边$e(w_1,w_2)$内包含$N^{\mathrm{task}}_{e(w_1,w_2)}$个AI推理任务、$N^{\mathrm{MS}}_{e(w_1,w_2)}$个微服务和$N^{\mathrm{age}}_{e(w_1,w_2)}$个智能体，即$|e(w_1,w_2)|=N^{\mathrm{task}}_{e(w_1,w_2)}+N^{\mathrm{MS}}_{e(w_1,w_2)}+N^{\mathrm{age}}_{e(w_1,w_2)}$，那么在该超边内评估$I$个AI推理请求在当前解下性能的时间复杂度为$O(N_{\mathrm{itm}}I|e(w_1,w_2)|)$，非支配排序的时间复杂度为$O(N_{\mathrm{itm}}(|e(w_1,w_2)|)^2)$。综上所述，NSGA-II-Hyper的时间复杂度为$O(N_{\mathrm{itm}}\max(I|e(w_1,w_2)|+(|e(w_1,w_2)|)^2))+O(W_1W_2(N_{\mathrm{task}}+N_{\mathrm{MS}}+N_{\mathrm{age}}))$。在理想情况下，若算法1的3个并列for循环可并行执行，同时AI推理任务、微服务和智能体被平均划分到每个跨层超边中，此时NSGA-II-Hyper并行效益最大。

此外，算法1需存储N_{task}个AI推理任务、N_{MS}个微服务、N_{age}个智能体的多维资源属性以及W_1W_2个超边分类，空间复杂度为$O(N_{\mathrm{task}}+N_{\mathrm{MS}}+N_{\mathrm{age}}+W_1W_2)$。算法2需存储$W_1W_2\mathrm{Pop}$个个体的染色体，每个个体的编码和目标值计算都依赖于I个AI推理请求，因此空间复杂度为$O(IW_1W_2\mathrm{Pop})$。综上，NSGA-II-Hyper的空间复杂度为$O(N_{\mathrm{task}}+N_{\mathrm{MS}}+N_{\mathrm{age}}+IW_1W_2\mathrm{Pop})$。

5 实验与分析

本节评估了本文所提NSGA-II-Hyper的性能。首先介绍了实验设置，包括实验环境和对比算法。然后探讨了本文算法与对比算法在收敛性、时间复杂度和推理时延、推理精度方面的性能。同时，仿真分析了超边划分区间、种群规模、交叉率、变异率参数对本文算法收敛性的影响。

5.1 实验设置

1) 实验环境。本文在Intel(R) Core(TM) i5-10210U CPU @ 1.60 GHz上使用MATLAB R2019a和Python作为评估所提算法的编程软件。服务层设置30个AI推理任务，逻辑资源层设置80个微服务，其中每个子层包含20个微服务。物理层采用包含40个智能体的端边云架构，实验场景设置在2 km×2 km的区域内，网络拓扑如图4所示，其中24个端侧智能体随机分布在300 m×2 000 m区域内，12个边侧智能体随机分布在1 200 m×2 000 m区域内，4个云侧智能体随机分布在500 m×2 000 m区域内。端侧和边侧智能体的最大通信范围分别为500 m和1 000 m。该架构可适用于智慧医疗、自动驾驶等未来典型应用场景，与6G“终端–边缘–云端”三级协同的实际部署逻辑高度契合，为不同应用场景的适配提供了扎实的拓扑基础。AI推理请求由端侧智能体随机产生，在端–边–云计算的AI推理请求的比例为5:3:2。AI推理任务、微服务和智能体供给的多维资源指标参数，包括通信带宽、计算能力、数据精度和算法精度，以及AI推理请求对多维资源的性能需求如表1所示。同时，本文算法的仿真参数也在表1中列出。

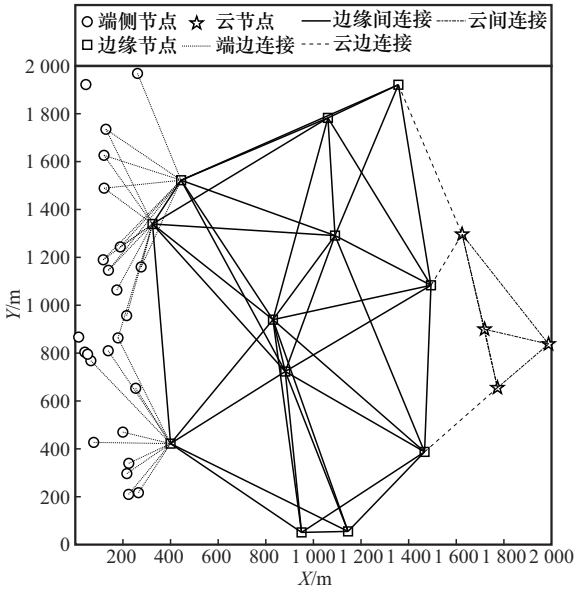


图4 物理层网络拓扑

2) 对比算法。将本文算法与以下5种具有代表性的解决多目标冲突的多维资源调度算法进行比较。每种算法运行20次。

① NSGA-II：不使用超边对解维度进行预分类，所有操作均遍历所有解。

② WCH[20]：通过加权和将多目标优化问题（即总精度差值的绝对值和总时延差值的绝对值之间的权衡）转换为一组单目标子问题，新解的产生和更新限制在邻域内。

③ MOPSO[29]：粒子的位置是优化问题P1的决策变量，飞行速度由该粒子目前搜索到的最优解和所有粒子搜索到的最优解共同决定。同时，引入变异算子对粒子进行扰动。

④ MARL[17]：该算法包含独立Q学习和策略梯度，各智能体的决策空间由优化问题P1的决策变量组成，奖励函数融合了基础奖励、Pareto前沿奖励和多样性奖励。

⑤ MAPPO：各智能体有独立的Actor网络，训练时使用集中式的Critic网络，该网络利用全局信息来评估当前联合动作的优劣，从而更有效地指导每个Actor网络的策略更新。同时，PPO的裁剪机制保证了训练过程的稳定性。

表1　　仿真参数

参数	含义
AI推理任务、微服务带宽[9]/(Gbit·s^{-1})	[0.8,10]、100
AI推理任务、微服务计算能力/TFLOPs	[0.5,15]、100
AI推理任务、微服务算力精度[9]	[0.75,1.0]
AI推理任务、微服务算法精度[9]	[0.7,1.0]
端边云智能体带宽[9]/(Gbit·s^{-1})	[0.8,2.4]、10、100
端边云智能体计算能力/TFLOPs	[0.5,2]、[5,15]、100
端边云智能体算力精度[9]	[0.75,0.85]、[0.85,0.95]、[0.95,1.0]
端边云智能体算法精度[9]	[0.7,0.8]、[0.8,0.9]、[0.9,1.0]
AI推理请求数量	60
端边云AI推理请求时延需求/ms	[10,20]、[20,30]、[30,40]
端边云AI推理请求抖动需求/ms	1、3、5
端边云AI推理请求精度需求	[0.7,0.8]、[0.8,0.9]、[0.9,1.0]
端边云AI推理请求数据大小/MB	[0.5,3]、[1,10]、[10,30]
端边云AI推理请求数据质量[9]	[0.7,0.85]、[0.8,0.95]、[0.9,1.0]
端边云AI推理请求计算密度需求/(FLOPs·B^{-1})	[100,500]、[500,1 500]、[1 000,4 000]
种群规模	50
交叉率ϕ	0.6
变异率ψ	0.05
收敛阈值ε	0.000 1
推理时延划分区间W_1	3
推理精度划分区间W_2	5

5.2　性能评估

图5和图6展示了6种算法总时延差值的绝对值和总精度差值的绝对值的收敛性。可以看出，随着迭代次数的增加，NSGA-II-Hyper、NSGA-II、MOPSO和WCH这4种启发式算法获得的总时延差值的绝对值与总精度差值的绝对值均已达到收敛状态，MARL和MAPPO算法获得的2种目标值还在动态变化。与其余3种启发式算法相比，NSGA-II-Hyper在寻找使2种目标值最小的解集时均具有较快收敛性，并且找到的解集能够使2种目标值均最小。这是因为NSGA-II-Hyper使用超边将解集进行预分类，面对不同类型的AI推理请求时，可并行计算满足性能指标的多维资源集合，降低了算法计算代价，提高了算法收敛速度。此外，虽然MOPSO和WCH算法在寻找使总精度差值的绝对值最小的解集时收敛速度更快，但这2种算法收敛时获得解集的总精度差值的绝对值大于NSGA-II-Hyper，这是因为这2种算法早熟导致目标值陷入局部最优。MAPPO相较MARL算法展现出了更稳定的收敛曲线，这种平滑性得益于PPO算法内在的裁剪机制，通过约束策略更新幅度，有效避免了训练过程中剧烈振荡，保证了学习的稳定性。

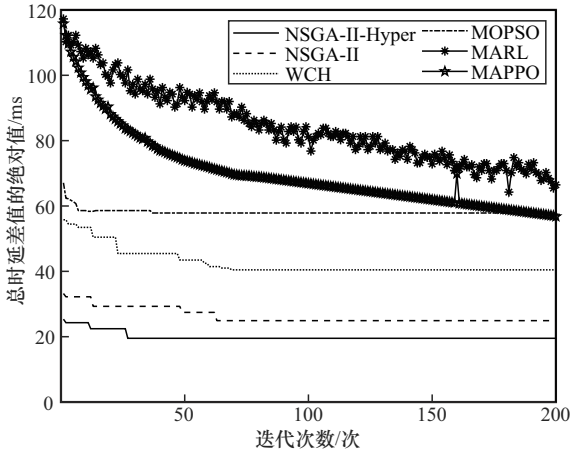


图5　6种算法总时延差值的绝对值收敛性

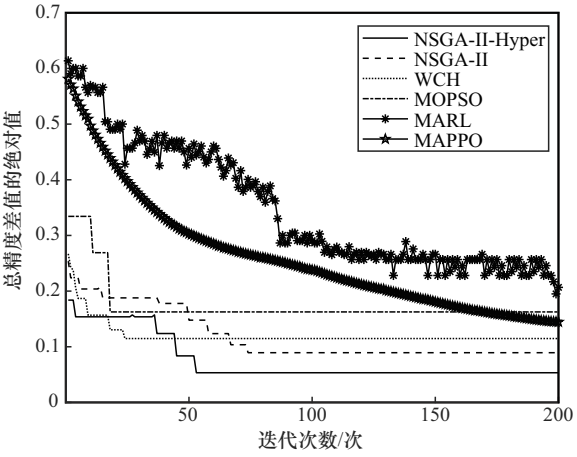


图6　6种算法总精度差值的绝对值收敛性

图7显示了6种算法获得的Pareto前沿。可以看出，本文算法获得的Pareto前沿在推理精度和推理时延方面均呈最佳结果。在相同推理时延下，NSGA-II获得的Pareto前沿的推理精度略低于本文算法获得的推理精度。MOPSO与WCH算法获得

的Pareto前沿较少。这是因为MOPSO算法对参数设置敏感导致算法过于早熟，从而获得的解集陷入了局部最优，WCH算法则依赖于权重向量并且邻域固定、搜索方向单一，从而导致获得的Pareto前沿较少。MARL和MAPPO算法获得的Pareto前沿最分散，此时2种算法还未达到收敛状态，还在探索最优Pareto前沿。图8展示了6种算法为单个AI推理请求生成多维资源调度策略的时间复杂度。可以看出，NSGA-II-Hyper较NSGA-II时间复杂度大幅度下降，WCH算法具有较低的时间复杂度，MAPPO算法时间复杂度最高。

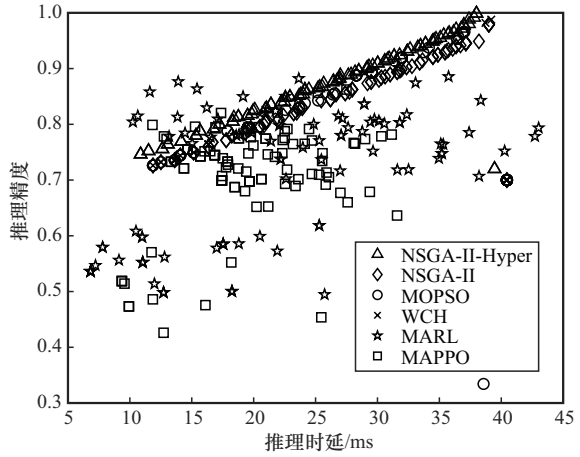


图7　6种算法Pareto前沿

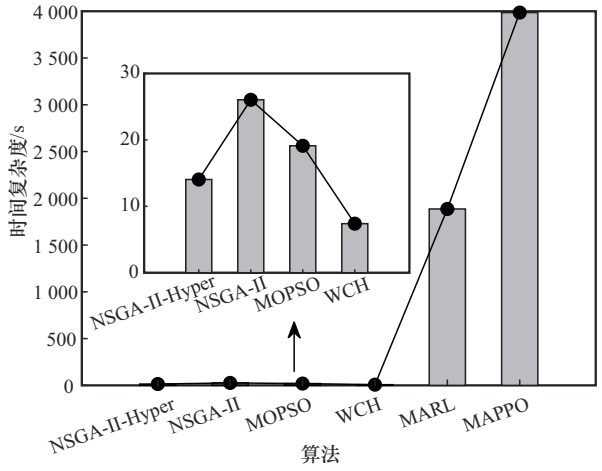


图8　6种算法时间复杂度

图9和图10分别展示了6种算法的总时延差值的绝对值和总精度差值的绝对值随AI推理请求数量变化情况。可以看出，随着AI推理请求数量的增加，6种算法的2个目标值均呈上升趋势。6种算法均在获得的Pareto前沿中为AI推理请求分配最优解，本文算法得到的2个目标值最小，NSGA-II次之。MOPSO与WCH算法获得的Pareto前沿较少，随着AI推理请求数量的增加，最优解将重复多次分配给不同的AI推理请求，因此AI推理请求等待6G网络服务层AI推理任务、逻辑资源层微服务和物理层智能体服务的时间增加，从而导致2个目标值均大幅增加。而MARL和MAPPO算法此时还未达到收敛状态，获得的2个目标值波动较大。

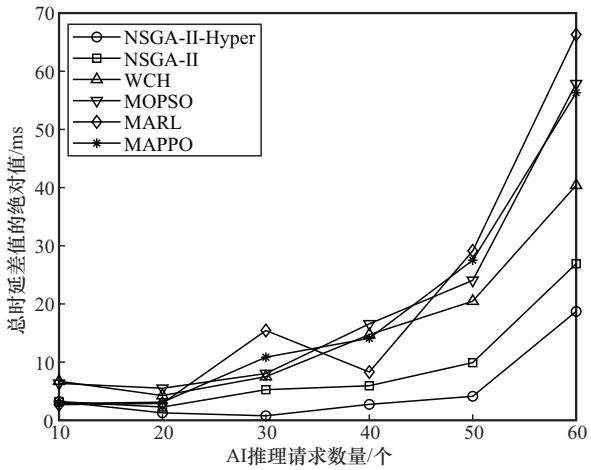


图9　6种算法总时延差值的绝对值随AI推理请求数量变化

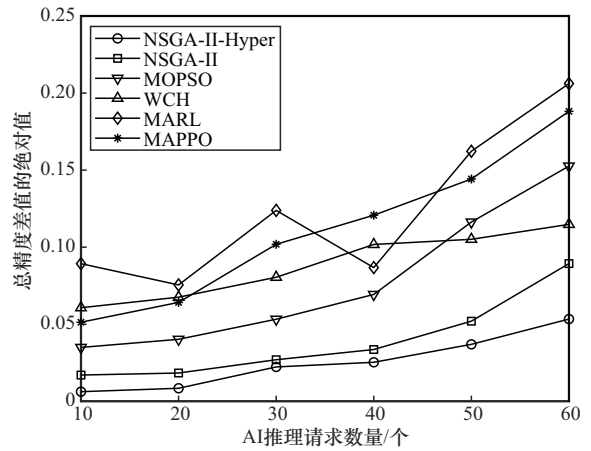


图10　6种算法总精度差值的绝对值随AI推理请求数量变化

表2展示了AI推理请求数量为60、算法运行20次时（每次运行随机生成不同资源分布/性能指标、不同AI推理请求分布/性能需求），AI推理请求成功服务比例的均值、标准差和95%置信区间。可以看出，NSGA-II-Hyper性能最优（均值79.13%最高）且最稳定（标准差3.85%最低），该算法通过并行计算Pareto最优解集，找到了更多满足AI推理请求的资源调度方案，从而在不同场景中保持

决策一致性的能力更强。同时，各算法95%置信区间进一步验证了NSGA-II-Hyper的统计稳健性。

表2　AI推理请求成功服务比例的均值、标准差及95%置信区间

算法	均值	标准差	95%置信区间
NSGA-II-Hyper	79.13%	3.85%	[77.33%,80.93%]
NSGA-II	73.92%	4.27%	[71.93%,75.91%]
WCH	68.35%	5.15%	[65.94%,70.76%]
MOPSO	61.17%	6.38%	[58.18%,64.16%]
MAPPO	57.82%	7.56%	[54.28%,61.36%]
MARL	52.44%	8.92%	[48.27%,56.61%]

5.3　参数灵敏度实验

图11和图12展示了超边划分区间W_1和W_2对2个目标值的影响。超边划分区间的精细程度会干扰NSGA-II-Hyper对服务层、逻辑资源层和物理层差异化多维资源的辨别能力，进而影响解集划分质量。例如，当$W_1 = 1$和$W_2 = 1$时，表示不对解维度进行划分，此时NSGA-II-Hyper退化为NSGA-II，并将依次对所有解集组合进行遍历，导致Pareto前沿趋于局部固定值，从而忽略了一些质量较好且空闲的解集，进而导致2个目标值较大。随着W_1和W_2的增加，2个目标值呈现先下降后上升的趋势，这是因为超边区间划分过于精细时，过多的噪声会被引入，这将误导算法带来太多的计算负担。综合考虑，本文设置$W_1 = 3$和$W_2 = 5$。

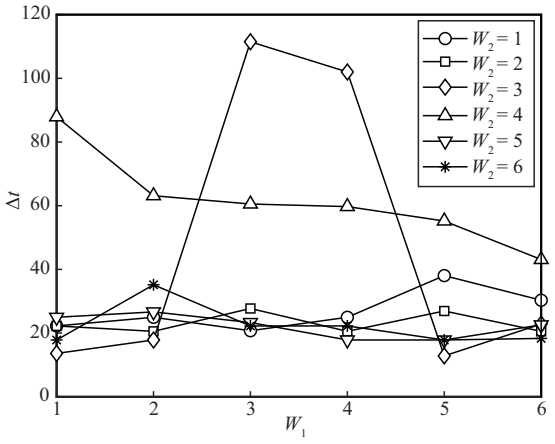


图11　时延划分区间对总时延差值的绝对值的影响

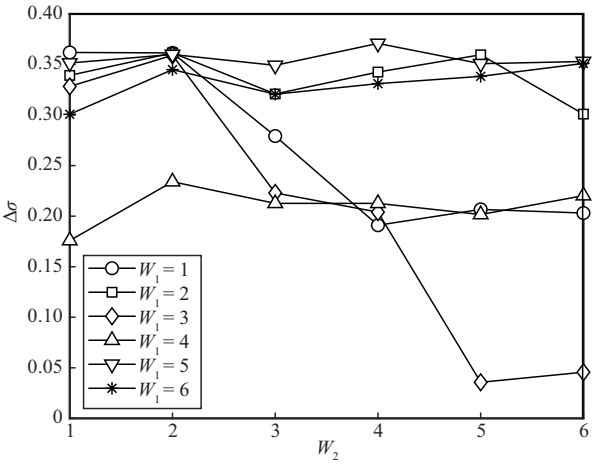
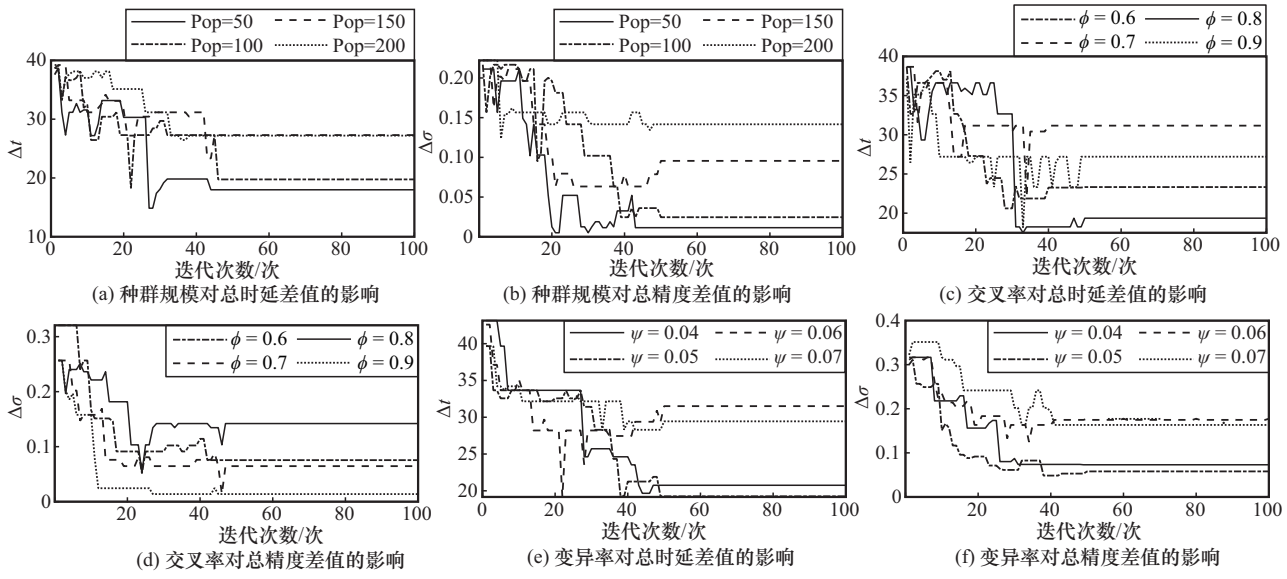


图12　精度划分区间对总精度差值的绝对值的影响

图13展示了种群规模、交叉率和变异率3种关键控制参数对总时延差值的绝对值和总精度差值的



(a) 种群规模对总时延差值的影响　(b) 种群规模对总精度差值的影响　(c) 交叉率对总时延差值的影响

(d) 交叉率对总精度差值的影响　(e) 变异率对总时延差值的影响　(f) 变异率对总精度差值的影响

图13　不同控制参数对总时延差值的绝对值和总精度差值的绝对值收敛性的影响

绝对值收敛性的影响。从图13(a)和图13(b)中可以看出，当种群规模Pop为50时，算法在迭代过程中展现出更优的收敛性，最终得到的2个目标值最小，因此Pop = 50。在图13(c)和图13(d)中，综合考量算法收敛效率与获得的2个目标值，本文折中选择0.6作为交叉率。从图13(c)和图13(d)中可以看出，当变异率为0.05时，算法在收敛过程中能有效平衡探索与开发能力，对应获得的2个目标值最小，故选择该值作为变异率。

6　结束语

本文研究了AI推理服务驱动的6G网络多维资源编排与调度方法。首先，提出了新颖的跨层嵌套超图模型以准确构建AI推理服务与6G网络服务层–逻辑资源层–物理层多维资源间的高阶协同依赖关系。接着，基于构建的跨层嵌套超图，形成了三层层间多维资源动态匹配问题，通过协同分配AI推理任务、涉及算法、数据、算力和连接资源的微服务和智能体，以最小化AI推理服务所需时延、精度与6G网络供给时延、精度间差值的绝对值。此外，提出了一种结合超边预分类特性的基于NSGA-II的启发式方法来解决该问题。仿真结果表明，本文算法优于其他基准算法，如MOPSO、WCH和MARL算法，在保障QoAIS前提下能为AI推理请求提供更匹配其性能需求的多维资源调度方案。未来工作将重点探索如何基于人工智能算法对实时业务负载特征（如时延/精度需求的分布变化）进行需求分类，以动态调整超边划分区间，构建一个自适应调度系统。

附录1　跨层超边预分类算法

算法1　跨层超边预分类

输入　$T_{\max}$，$T_{\min}$，$\sigma_{\max}$，$\sigma_{\min}$，W_1，W_2，N_{task}，N_{MS}，N_{age}

输出　$\mathcal{E}_{\text{int}}=\{e(w_1,w_2)\}$

1)计算$\Delta T=T_{\max}-T_{\min}$，$\Delta\sigma=\sigma_{\max}-\sigma_{\min}$

2)for $w_1\in\{0,1,\cdots,W_1-1\}$

3)　for $w_2\in\{0,1,\cdots,W_2-1\}$

4)　　for $n_{\text{task}}\in\{1,\cdots,N_{\text{task}}\}$

5)　　判断$|a_{\text{del}}^{n_{\text{task}}}-T_{\min}|$是否属于区间$\left[\frac{w_1}{W_1},\frac{w_1+1}{W_1}\right]\Delta T$且$|a_{\text{acc}}^{n_{\text{task}}}-\sigma_{\min}|$是否属于区间$\left[\frac{w_2}{W_2},\frac{w_2+1}{W_2}\right]\Delta\sigma$

6)　　若满足，将n_{task}归类于$e(w_1,w_2)$

7)　　若不满足，遍历下个n_{task}

8)　　end for

9)　　for $n_{\mathcal{A}}^{\text{MS}}\in\{1,\cdots,N_{\text{MS}}\}$

10)　　判断$|a_{\text{del}}^{n_{\mathcal{A}}^{\text{ms}}}-T_{\min}|$是否属于区间$\left[\frac{w_1}{W_1},\frac{w_1+1}{W_1}\right]\Delta T$且$|a_{\text{acc}}^{n_{\mathcal{A}}^{\text{MS}}}-\sigma_{\min}|$是否属于区间$\left[\frac{w_2}{W_2},\frac{w_2+1}{W_2}\right]\Delta\sigma$

11)　　若满足，将$n_{\mathcal{A}}^{\text{MS}}$归类于$e(w_1,w_2)$

12)　　若不满足，遍历下个$n_{\mathcal{A}}^{\text{MS}}$

13)　end for

14)　for $n_{\text{age}}\in\{1,\cdots,N_{\text{age}}\}$

15)　　判断$|a_{\text{del}}^{n_{\text{age}}}-T_{\min}|$是否属于区间$\left[\frac{w_1}{W_1},\frac{w_1+1}{W_1}\right]\Delta T$且$|a_{\text{acc}}^{n_{\text{age}}}-\sigma_{\min}|$是否属于区间$\left[\frac{w_2}{W_2},\frac{w_2+1}{W_2}\right]\Delta\sigma$

16)　　若满足，将n_{age}归类于$e(w_1,w_2)$

17)　　若不满足，遍历下个n_{age}

18)　end for

19)　end for

20)end for

21)返回$\mathcal{E}_{\text{int}}=\{e(w_1,w_2)\}$

附录2　结合超边预分类特性的NSGA-II（NSGA-II-Hyper）

算法2　NSGA-II-Hyper

输入　N_{task}，N_{MS}，N_{age}，W_1，W_2，最大迭代次数N_{itm}，Pop，ϕ，ψ，收敛阈值ε

输出　Pareto前沿

1)初始化网络拓扑结构、网络参数以及AI推理请求参数

2)初始化Pareto前沿为空

3)运行算法1，生成跨层超边集合

4)在各个跨层超边中均匀随机采样AI推理任务、微服务和智能体，在每个跨层超边中均随机保留Pop个组合作为初始解集

5)针对I个AI推理请求，检查每个跨层超边内初始解集是否满足约束$C_1\sim C_3$，若不满足重新执行第4) 行

6)针对I个AI推理请求，计算每个跨层超边内初始解集的总时延差值的绝对值和总精度差值的绝对值

7)非支配排序每个跨层超边内的初始解集

8)计算每个跨层超边内初始解集的拥挤距离，并按前沿等级和拥挤距离对初始解集排序

9)for $n_{\text{itm}}\in\{1,\cdots,N_{\text{itm}}\}$

10)　将当前种群Pop合并到Pareto前沿，并去除前沿中的重复解

11)　for $e(w_1,w_2)\in\mathcal{E}_{\text{int}}$

12)　　通过二元锦标赛从每个跨层超边中选择2个父代

13)　　以交叉率ϕ对父代进行两点交叉获得2个子代

14)　　以变异率ψ随机对2个子代进行单点变异

15)　　检查约束$C_1\sim C_{10}$，保留满足条件的子代

16)　end for

17)　合并父代和子代种群，计算种群中每个个体的总时延差值的绝对值和总精度差值的绝对值

18)　非支配排序合并后的种群

19)　计算拥挤距离并排序

20)　按优先级挑选前Pop个精英
21)　计算最近10代解集时延和精度变化
22)　判断时延和精度变化是否同时小于ε
23)　　若满足，执行第26) 行
24)　　若不满足，执行下次迭代
25) end for
26)返回Pareto前沿

参考文献：

[1] TERA S P, CHINTHAGINJALA R, PAU G, et al. Toward 6G: an overview of the next generation of intelligent network connectivity[J]. IEEE Access, 2025, 13: 925-961.

[2] 刘光毅, 陈天骄, 崔莹萍, 等. 面向AI服务质量保障的6G无线网络智慧内生架构[J]. 移动通信, 2025, 49(1): 2-8.
LIU G Y, CHEN T J, CUI Y P, et al. Intelligent and endogenous architecture of 6G wireless network for AI quality of service guarantee [J]. Mobile Communications, 2025, 49 (1): 2-8.

[3] PAN J P, CAI L, YAN S, et al. Network for AI and AI for network: challenges and opportunities for learning-oriented networks[J]. IEEE Network, 2021, 35(6): 270-277.

[4] ZHENG X X, ZHANG W T, HU C F, et al. Cloud-edge-end collaborative inference in mobile networks: challenges and solutions[J]. IEEE Network, 2025, 39(4): 90-96.

[5] DENG J, ZHENG Q B, LIU G Y, et al. A digital twin approach for self-optimization of mobile networks[C]//Proceedings of the 2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). Piscataway: IEEE Press, 2021: 1-6.

[6] YANG Y, MA M L, WU H Q, et al. 6G network AI architecture for everyone-centric customized services[J]. IEEE Network, 2023, 37(5): 71-80.

[7] LI X, ZHAO L, ZHANG L, et al. Artificial general intelligence for medical imaging analysis[J]. IEEE Reviews in Biomedical Engineering, 2025, 18: 113-129.

[8] YU T K, YANG H, YAO Q Y, et al. TSTNet: temporal semantic transformer-based computing power network for automatic driving in the Internet of vehicles[J]. IEEE Transactions on Artificial Intelligence, 2025, 6(10): 2669-2684.

[9] HU Y Q, LI Q, CHAI Y H, et al. AI service deployment and resource allocation optimization based on human-like networking architecture[J]. IEEE Internet of Things Journal, 2024, 11(14): 24795-24813.

[10] ZHUANG Z M, WEN D Z, SHI Y M, et al. Integrated sensing-communication-computation for over-the-air edge AI inference[J]. IEEE Transactions on Wireless Communications, 2024, 23(4): 3205-3220.

[11] ZHUANSUN C L, YAN K D, ZHANG G X, et al. Hypergraph-based joint channel and power resource allocation for cross-cell M2M communication in IIoT[J]. IEEE Internet of Things Journal, 2023, 10(17): 15350-15361.

[12] HUANG J, ZHANG S L, YANG F, et al. Hypergraph-based interference avoidance resource management in customer-centric communication for intelligent cyber-physical transportation systems[J]. IEEE Transactions on Consumer Electronics, 2024, 70(1): 1775-1786.

[13] HAO Q, SHENG M, ZHOU D, et al. A multi-aspect expanded hypergraph enabled cross-domain resource management in satellite networks[J]. IEEE Transactions on Communications, 2022, 70(7): 4687-4701.

[14] GAO Y, FENG Y F, JI S Y, et al. HGNN$^+$: general hypergraph neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3181-3199.

[15] LI X, BI S Z, WANG H. Optimizing resource allocation for joint AI model training and task inference in edge intelligence systems[J]. IEEE Wireless Communications Letters, 2021, 10(3): 532-536.

[16] YUAN X, LI N, ZHANG T, et al. High efficiency inference accelerating algorithm for NOMA-based edge intelligence[J]. IEEE Transactions on Wireless Communications, 2024, 23(11): 17539-17556.

[17] LI X, BI S Z. Optimal AI model splitting and resource allocation for device-edge co-inference in multi-user wireless sensing systems[J]. IEEE Transactions on Wireless Communications, 2024, 23(9): 11094-11108.

[18] YANG S S, ZHANG Z H, ZHAO C, et al. CNNPC: end-edge-cloud collaborative CNN inference with joint model partition and compression[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(12): 4039-4056.

[19] LIU S, WEN D Z, LI D, et al. Energy-efficient optimal mode selection for edge AI inference via integrated sensing-communication-computation[J]. IEEE Transactions on Mobile Computing, 2024, 23(12): 14248-14262.

[20] WANG L H, REN X B, ZHAO C, et al. MPDM: a multi-paradigm deployment model for large-scale edge-cloud intelligence[J]. IEEE Internet of Things Journal, 2023, 10(10): 8773-8785.

[21] ZHU B T, WANG X B. Hypergraph-aided task-resource matching for maximizing value of task completion in collaborative IoT systems[J]. IEEE Transactions on Mobile Computing, 2024, 23(12): 12247-12261.

[22] 李元诚, 林玉坤. 基于超图Transformer的APT攻击威胁狩猎网络模型[J]. 通信学报, 2024, 45(2): 106-114.
LI Y C, LIN Y K. APT attack threat hunting model based on hypergraph Transformer[J]. Journal on Communications, 2024, 45(2): 106-114.

[23] 王惠真, 韩春雷, 平超, 等. 跨域协同作战体系的复杂超图表示方法[J]. 战术导弹技术, 2024(5): 1-10.
WANG H Z, HAN C L, PING C, et al. Complex hypergraph representation method for cross-domain cooperative combat system[J]. Tactical Missile Technology, 2024(5): 1-10.

[24] 袁翠敏, 李贺, 解梦凡, 等. 基于异构图书评论信息融合模型的超图聚类研究[J]. 情报理论与实践, 2025, 48(2): 198-205, 197.
YUAN C M, LI H, JIE M F, et al. Research on hypergraph clustering based on heterogeneous book review information fusion model[J]. Information Studies: Theory & Application, 2025, 48(2): 198-205, 197.

[25] 王晓云, 陆璐, 刘超, 等. 面向6G的网络架构建模、评估及优化[J]. 通信学报, 2024, 45(7): 235-249.
WANG X Y, LU L, LIU C, et al. Network architecture modeling, evaluation and optimization for 6G[J]. Journal on Communications, 2024, 45(7): 235-249.

[26] BAHRAMI B, KHAYYAMBASHI M R, MIRJALILI S. Multiobjective placement of edge servers in MEC environment using a hybrid algorithm based on NSGA-II and MOPSO[J]. IEEE Internet of Things Journal, 2024, 11(18): 29819-29837.

[27] PAN H Y, LIU Y H, SUN G, et al. Resource scheduling for UAVs-aided D2D networks: a multi-objective optimization approach[J]. IEEE Transactions on Wireless Communications, 2024, 23(5): 4691-4708.

[28] CHU K F, CHEN T L, XIE Y, et al. Collaborative routing and charging/discharging scheduling of electric autonomous vehicles in coupled power-traffic networks: a multiobjective approach[J]. IEEE Internet of Things Journal, 2025, 12(11): 17753-17764.

[29] TRIPATHI P K, BANDYOPADHYAY S, PAL S K. Multi-objective particle swarm optimization with time variant inertia and acceleration coefficients[J]. Information Sciences, 2007, 177(22): 5033-5049.

[作者简介]

景川芳（1996– ），女，河南平顶山人，南京邮电大学博士生，主要研究方向为5G/6G通信系统、确定性网络、路由与多维资源调度等。

朱晓荣（1977– ），女，山东临沂人，南京邮电大学教授、博士生导师，主要研究方向为5G/6G通信系统、物联网、区块链等。