

基于强化学习的天地融合网络算力路由发展与挑战

王明乾¹, 吴晓梅², 杨 桃², 雷广旺¹, 张婷婷¹, 刘子深¹

(1.北京理工大学, 北京 100081;

2.中国铁塔股份有限公司, 北京 100195)

摘 要: 天地融合网络通过协同卫星、空中平台和地面设施构建立体化通信架构, 其动态拓扑和异构算力资源对路由机制提出了全新挑战。面对这一复杂场景, 强化学习因其强大的环境适应能力和智能决策优势成为解决算力路由问题的研究热点。首先, 系统梳理并深入探究强化学习在算力路由中的应用实践; 然后, 从算法范式与优化目标两个维度解析其核心设计原理; 最后, 深入剖析并前瞻性地指出其在天地融合网络环境下所面临的关键挑战。

关键词: 天地融合; 算力路由; 强化学习

中图分类号: TP393

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2025034

Development and Challenges of Computing Power Routing in Space-Ground Integrated Networks Based on Reinforcement Learning

WANG Mingqian¹, WU Xiaomei², YANG Guang², LEI Guangwang¹, ZHANG Tingting¹, LIU Zishen¹

1. Beijing Institute of Technology, Beijing 100081, China

2. China Tower Corporation Limited, Beijing 100195, China

Abstract: Space-air-ground integrated networks (SAGIN) establish a multi-dimensional communication architecture by synergistically coordinating satellites, aerial platforms, and terrestrial facilities. The inherent characteristics of dynamic topologies and heterogeneous computing resources within this architecture present novel challenges for routing mechanisms. In response to this complex scenario, reinforcement learning (RL) has emerged as a prominent research focus for addressing computing-aware routing problems, owing to its robust environmental adaptability and intelligent decision-making capabilities. This paper systematically reviews and investigates the application of RL in computing-aware routing. It analyzes the core design principles from the dual perspectives of algorithmic paradigms and optimization objectives. Finally, this paper deeply analyzes and prospectively identifies the key challenges that RL faces in the context of SAGIN environments.

Keywords: space-integrated-ground, computing power routing, reinforcement learning

0 引言

随着低轨卫星技术的迅速发展, 天地融合网络逐步形成并完善, 未来将呈现卫星、地面云计算中心及边缘节点协同的多层次网络架构。同时, 人工智能技术的快速发展使智慧城市、无人驾驶、灾难救援等复杂应用场

景对网络算力的实时响应能力提出了更高的要求。这些新兴应用不仅需要高速的数据传输, 更要求网络能够实时、高效地分配和调度分布式计算资源。然而, 传统网络路由技术主要关注数据传输路径的优化, 难以应对天地融合环境中异构资源与动态负载带来的挑战。同时, 卫星网络拓扑结构的动态性与网络资源的持续变化, 对

收稿日期: 2025-06-19; 修回日期: 2025-08-14

基金项目: 国家自然科学基金 (No.62401047)

Foundation Item: The National Natural Science Foundation of China (No.62401047)

网络管理和资源分配提出了更高要求^[1]。

基于强化学习 (reinforcement learning, RL) 的方法通过将每个具备计算能力的资源节点视为智能体, 实现分布式自主学习和协作决策, 实时感知网络状态和资源负载, 动态调度计算任务位置。此外, 基于软件定义网络 (software defined network, SDN) 技术的集中控制方式, 也可利用单一智能体对全网资源进行集中管理与优化分配^[2]。无论是分布式还是集中式方法, 基于强化学习的自主决策和持续学习能力都能有效地适应网络环境的动态变化, 降低网络时延, 提高带宽利用率和能源效率。本文将综述基于强化学习的天地融合网络算力路由技术, 分析其发展现状、关键技术, 并探讨其未来发展机遇与挑战。

1 天地融合网络架构与算力系统模型

1.1 天地一体化算力网络架构

随着人工智能、大数据和物联网的发展, 网络正由“以数据为中心”向“以计算为中心”演进。为突破地面基础设施在覆盖与算力部署上的限制, 实现全球无缝连接与泛在智能计算, 天地一体化算力网络应运而生。该体系融合天基算力的卫星网络、空基算力的高空平台/无人机网络与陆基算力的地面通信网络, 构建覆盖陆、海、空、天, 集通信与计算于一体的多维立体化基础设施^[3], 如图1所示。

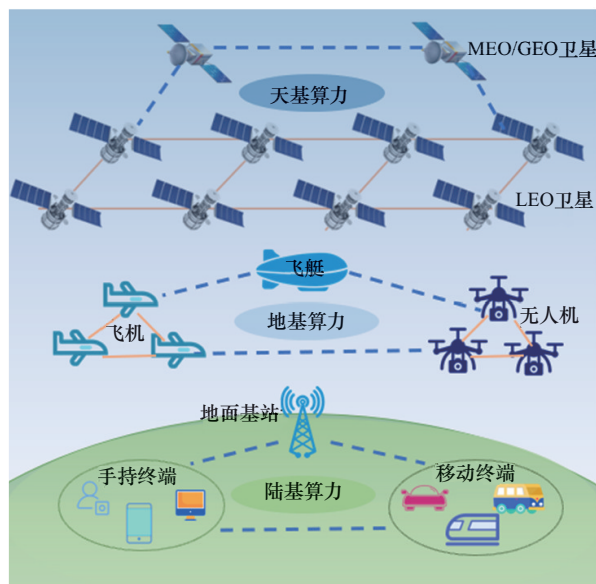


图 1 天地融合算力网络架构

1.1.1 地基算力网络

地基算力网络由分布在全球的光纤骨干网、核心路

由交换设施、移动通信基站、边缘数据中心及大型云数据中心构成, 是算力供给的主体。它具备计算能力强、存储容量大、能源供应稳定等特点, 可为大规模数据处理、模型训练及后台分析提供支撑。同时, 地基算力网络通过高速回传链路与空基、天基网络互联, 能够实现全球范围的算力资源调度与协同。

1.1.2 空基算力网络

空基算力网络部署在平流层高度的高空平台、飞艇及长航时无人机等载体上^[4-5], 具有机动灵活、快速部署和广域覆盖的优势。其算力节点可根据任务需求临时投放或调整位置, 为地面网络无法覆盖或受灾受阻地区提供临近的计算与通信能力。空基算力网络不仅可在任务现场直接完成数据预处理与智能推理, 还可作为中继节点, 连接天基算力网络与地基算力网络, 实现跨层级的低时延算力协同。

1.1.3 天基算力网络

天基算力网络是实现全球无缝覆盖的核心, 由部署在不同轨道的多颗卫星组成星座系统, 可将网络服务延伸至海洋、沙漠、极地、深山等传统网络难以触达的区域。根据轨道高度不同, 天基算力网络通常分为3类。

地球静止轨道 (geostationary orbit, GEO) 卫星位于约 36 000 千米高度, 与地球自转同步, 3 颗 GEO 卫星即可覆盖全球绝大部分区域, 适合广播业务及作为稳定的管理与控制节点。然而, 其往返时延超过 500 ms, 不适用于低时延计算与实时交互。

中地球轨道 (medium earth orbit, MEO) 卫星位于 2 000~20 000 千米高度, 在覆盖范围与传输时延之间实现平衡, 相较 GEO 卫星时延显著降低, 且单星覆盖范围大于低地球轨道 (low earth orbit, LEO) 卫星。

低地球轨道卫星位于 500~2 000 千米高度, 单跳星地时延可低至数十毫秒, 链路损耗小, 能提供接近地面光纤的体验。然而, 其约 7.8 km/s 的高速运动使网络拓扑频繁变化, 对路由协议、移动性管理和波束切换提出极高要求。目前, LEO 星座已成为巨型星座研究与部署的核心方向, 典型代表包括 Starlink、OneWeb 和国内正在规划的“国网”等项目。LEO 卫星凭借低时延和广域覆盖的特性, 具备在全球范围内直接为终端用户提供宽带接入与计算服务的潜力。为了从传统通信网络无缝演进至算力网络, 现代 LEO 卫星正逐步引入星上处理单元, 搭载高性能处理器、存储模块以及可编程硬件 (如 FPGA、GPU), 实现数据的本地计算、缓存与智能路由。这一转变使其从单纯的“弯管”中继器演化为具备一定自治能

力的“空中数据中心”，能够在轨执行数据预处理、边缘推理、分布式计算分片等任务。更重要的是，星间链路（inter-satellite link, ISL）的引入使LEO星座内部能够构建高速、低时延的空间骨干网，实现跨区域算力的直接调度与任务迁移。

这种多层级轨道体系不仅实现了覆盖互补，还天然构成一个分层、异构的算力资源池：地基算力网络算力充足且成本低，但对偏远用户时延高；空基算力网络提供低时延的瞬时算力，但覆盖范围与持续性有限；LEO卫星在广域覆盖下可提供低时延算力，但资源有限且宝贵；GEO卫星虽时延高，不适合实时计算，却可胜任网络管理与非实时任务。

1.2 网络化算力系统模型

1.2.1 算力路由的概念与内涵

为了精确把握算力路由的内涵，有必要将其与传统IP路由进行对比。传统IP路由的核心目标是在网络层实现数据包的高效转发。路由器通过运行路由协议（如OSPF、BGP）获取网络拓扑信息，并生成一张路由表，记录到达各目的网络地址（IP前缀）的最优下一跳路径，其决策依据是纯粹的网络度量，如带宽、时延、跳数、链路负载与成本等。整个过程对上层应用透明，其终极目标是将数据包从源IP地址高效、可靠地送达目的IP地址。相比之下，算力路由（computing power routing, CPR）的目标不仅是找到一条网络路径，更是实现业务请求与算力资源的最优匹配与高效协同，是“算网一体”理念在路由层面的具体体现。它是一种更高维度的决策过程，将网络资源信息与算力资源信息有机地整合在一起进行通告、分发和决策。算力路由决策不仅要考虑网络的传输性能（时延、抖动、带宽），更要感知和评估计算节点的实时状态，如CPU/GPU的负载、内存/存储的可用性、是否部署了特定服务（如AI推理模型、数据转码服务）等，其最终目标是为用户的业务请求选择一个最优的执行节点，并建立一条满足服务质量（QoS）要求的网络连接，实现端到端的服务交付。

1.2.2 网络化算力系统模型设计的相关问题

为了将强化学习等前沿数学工具应用于天地一体化算力网络这一高度动态的复杂系统，首要任务是构建一个能够精确刻画其多维度资源与约束的形式化模型。在通信层面，整个网络可以抽象为一个时变的图 $G_t = (V, E_t)^{[6]}$ ，其中 V 是所有节点（包括卫星、地面站、边缘服务器等）的集合， E_t 是时刻 t 存在的通信链路集

合。进一步地，模型需要量化每条链路 $e \in E_t$ 的动态特性，包括可用带宽 $E_e(t)$ 、传播时延 $D_{\text{prop}}(t)$ 、排队时延 $D_{\text{queue}}(t)$ 、多普勒效应（尤其对于星地链路）以及链路中断概率 $P_{\text{fail}}(t)$ ，这些参数共同决定了数据传输质量。在计算层面，模型需要对每个计算节点 $v \in V$ 的计算资源进行抽象化建模^[7]，包括CPU处理频率 F_v 、内存大小 M_v 、存储空间 S_v 等资源容量，以及时刻 t 的实时状态，如CPU利用率 $\mu_v(t)$ 、内存占用率 $\eta_v(t)$ 和待处理任务队列长度 $Q_v(t)$ 。对于能源受限的设备，能耗模型需考虑处理能耗（与CPU频率的平方或立方呈正比）和通信能耗（与数据传输量和发射功率相关）。最后，网络服务的对象-计算任务 J_k ，则可用元组 $\langle I_k, W_k, D_{k\text{max}} \rangle$ 描述，其中 I_k 是输入数据大小， W_k 是总计算量， $D_{k\text{max}}$ 是最大可容忍时延。此外，任务的可分割性（是否可并行处理）也会影响路由和调度策略。

基于前述通信、计算、能耗与任务的耦合模型，原先概念性的“天地一体化算力路由”便可被精确地形式化为一个动态资源分配与调度的随机优化问题。其核心是在瞬息万变的天地一体化网络中，为每个新到达的计算任务，实时地决定其最佳计算卸载位置（在本地、边缘、云端还是星上处理）以及为其分配的最优通信与计算资源。其最终目标是学习一个能够最大化系统整体效能的智能决策策略，该策略必须在多个相互冲突的子目标间做出精妙权衡：在追求最低任务时延和最高完成率的同时，兼顾全网的负载均衡以及卫星等能源受限节点的功耗效率，从而在性能、成本和效率之间找到最佳平衡点。

2 强化学习理论概述

强化学习为不确定环境下的序贯决策问题提供了强大的理论框架，与动态算力路由问题天然契合。在RL范式中，智能体在每个决策时刻 t 观测环境状态 S_t （如网络负载与链路质量），执行一个动作 A_t （如路由决策），随后环境转移至新状态 S_{t+1} 并反馈一个奖励信号 R_{t+1} 。智能体的目标是学习一个最优策略 π^* ，以最大化长期累积奖励。近年来，RL在网络路由领域的应用逐渐增多，主要是因为网络环境具有动态性、不确定性和高复杂性等特点，难以通过传统静态或人为定义的规则获得最优解。RL无须事先准确建模，通过智能体的自主试错和持续优化，能够有效适应网络状态变化，实时调整路由决策，从而实现更高效、更灵活、更鲁棒的网络性能。根据智能体数量的不同，RL可分为单智能体强化学习与多智能体强化学习（multi-agent reinforcement learning, MARL）

两大类^[8]。两者在建模方法、学习目标、策略协同以及环境稳定性等方面存在显著差异，主流的强化学习算法分类如图2所示。

单智能体强化学习是最基本的RL形式，其假设环境为一个马尔可夫决策过程（Markov decision process, MDP），即智能体的当前状态与下一个状态之间的转移概率仅依赖于当前状态和动作。在该框架下，智能体通过不断探索环境并获得奖励信号，从而优化策略以最大化长期累计奖励。代表性算法包括Q学习、SARSA、深度Q网络（deep Q-network, DQN）、PPO、深度确定性策略梯度（deep deterministic policy gradient, DDPG）等，已被广泛应用于路径规划、自动控制、游戏对抗等领域。由于不存在其他智能体干扰，单智能体RL的环境具备稳定性，训练过程相对收敛性强，易于理论分析与性能验证。

MARL旨在研究多个智能体在共享环境中并行决策与交互的复杂问题，其理论框架通常被形式化为马尔可夫博弈；在部分可观测设定下，该框架可推广为部分可观察马尔可夫博弈（partially observable Markov game, POMG），而在应用广泛的完全协作场景中，则通常采用其重要特例，即分布式部分可观察马尔可夫决策过程（Dec-POMDP）进行建模。MARL的核心挑战在于非平稳性：由于所有智能体的策略在训练中同步演化，从单一智能体视角看，环境的转移概率与奖励函数会因他人策略的改变而持续漂移，这严重破坏了传统单智能体依赖平稳环境的收敛前提。为应对此挑战，集中式训练、分布式执行（centralized training, decentralized execution, CTDE）已成为当前主流范式，它在训练阶段利用全局信息（如联合状态、动作）构建一个集中的评估模块以稳定学习过程，而在执行阶段，智能体仅依赖本地观测进行独立决策。在此框架下，主流技术路径各辟蹊径：一

种是构建集中式评论家（例如MADDPG）将DDPG范式扩展至多智能体领域，其评论家能获取所有智能体的动作以进行精确的联合动作价值评估；COMA则通过计算一个反事实基线来回答“若只有该智能体采取不同动作，结果会如何？”，从而实现了更精细的个体信用分配；而作为当前许多协作任务性能标杆的MAPPO，则将近端策略优化的稳定策略步长更新机制与集中式价值函数相结合，在保证稳定性的同时提升了样本效率。与此并行，值函数分解方法致力于构建全局团队价值与个体价值之间的明确关联，其中VDN假定全局 Q 值是局部 Q 值之和，结构简单但表达能力有限；QMIX通过一个非负权重的混合网络施加了更具表达力的单调性约束，确保了个体-全局-最大化原则，即个体价值的提升必然带来全局价值的提升；QTRAN则通过一致性约束进一步放宽了单调性假设，提供了更大的灵活性。当智能体间需要显式协调时，通信与图结构方法被广泛引入，早期方法（如DIAL和CommNet）引入可学习的通信信道让智能体自主共享信息，而更先进的基于图神经网络（graph neural network, GNN）的方法则将智能体交互拓扑显式建模为图结构，通过图卷积或图注意力机制高效聚合邻域信息，实现了可扩展的结构化协调。此外，为处理大规模智能体系统的可扩展性挑战，源于物理学的平均场方法（如MF-Q/PG）将复杂的多体交互简化为个体对群体平均的交互，将复杂度从指数级降低为线性级，为大规模MARL应用提供了可行方案。综上所述，这些技术路径分别通过“全局价值评估”“结构化信用分解”“显式信息共享”与“群体动态近似”等核心机制，从不同维度强有力地缓解了由策略共演化引发的非平稳性问题，显著提升了MARL算法在复杂协作与竞争博弈中的训练稳定性、信用分配质量与最终性能。

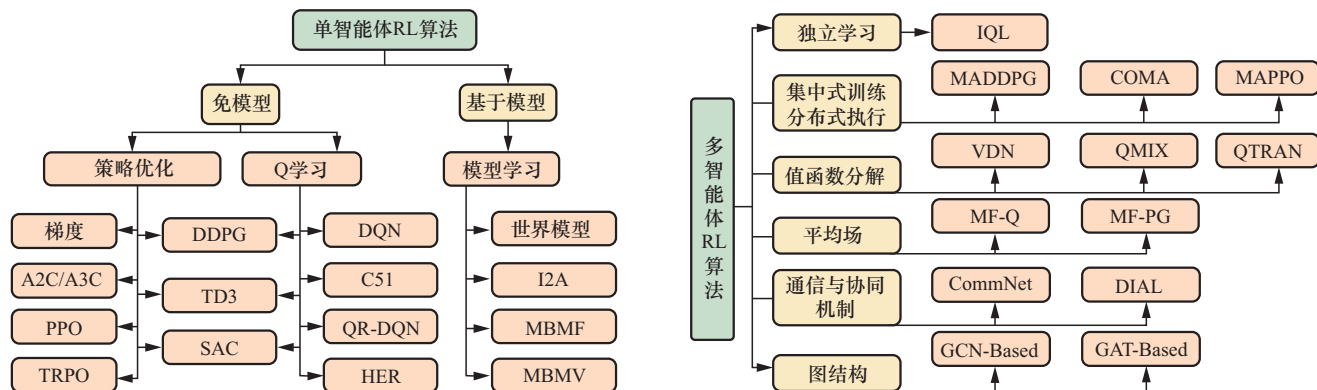


图2 强化学习算法分类

3 基于强化学习的天地融合网络算力路由技术

在LEO卫星网络中,计算资源的空间异构性与拓扑的高动态性共同构成了算力路由的核心技术瓶颈,其关键在于实现传输与计算的跨域联合优化。为攻克此难题,早期的探索性工作相继展开,以Cao等^[9]的研究为例,他们提出了一种面向计算感知的路由算法:首先构建了一个无执照的动态网络模型,通过动态映射卫星和地面虚拟节点的关联关系,将时变的资源与拓扑特性刻画为加权图结构;随后,将算力路由建模为带节点权和边权的动态单源最短路径问题,在路径规划过程中同时考虑传输时延与计算开销。在此框架下,为化解实时搜索空间爆炸带来的挑战,算法进一步引入基于遗传算法的启发式搜索机制,从而在复杂星座中实现了较高的任务调度效率与自适应能力。然而,随着卫星规模持续扩大、网络状态愈发频繁变化,此类基于规则或启发式的传统方法在全局最优性与泛化能力方面逐渐显露局限。为克服此局限,研究范式转向了RL,其核心框架如图3所示。该框架将算力路由这一通信与计算的联合优化问题,重塑为一个由多智能体协作的闭环决策过程。在此过程中,智能体基于对网络拓扑、链路状态、节点算力及任务属性的实时观测,从候选节点集中选择最优动作(即卸载目标),并通过一个目标为最小化时延、保障QoS并最大化资源利用率的组合式奖励函数进行学习。

3.1 基于算法范式的算力路由技术

在天地融合网络的算力路由中,深度强化学习

(deep reinforcement learning, DRL)已成为解决该问题的关键技术,其算法范式主要沿着单智能体和多智能体两个路径发展。

单智能体强化学习将整个路由决策过程从一个中心化或节点化的视角建模为MDP,其算法主要可分为值函数型与策略梯度型^[10]。值函数型方法以DQN为代表,通过神经网络逼近状态-动作价值函数进行离散决策;而策略梯度型方法以深度确定性策略梯度为代表,更适用于需要进行卸载比例、功率分配等决策的连续控制场景^[11]。在具体应用中,文献[7]便以DQN为核心,其状态向量融合了卫星剩余算力、链路负载矩阵、节点间距离和业务时延约束等高维信息,动作为“选择下一跳邻居”以同时决策传输路径与计算落点,奖励函数则由时延、时延违约惩罚及拓扑引导3项加权构成;算法主体采用一个两层全连接深度Q网络,并结合目标网络、经验回放和 ϵ -贪婪衰减等技术实现稳定收敛。为进一步提升智能体对网络拓扑结构的感知能力,谢明熹^[12]提出了一种基于GCN-DQN架构的自适应路由算法。具体而言,该方法通过链路带宽利用率和节点算力负载构建图结构状态矩阵,结合业务优先级构造动态奖励函数,并利用图卷积神经网络(graph convolution neural network, GCN)实现网络-算力多维特征的深度融合,在预筛选的 k 条最短路径中自适应优化路由决策。此外,该工作还设计了基于服务类型的权重调节机制,显著提升了智能体在不同优先级任务中的泛化性能。

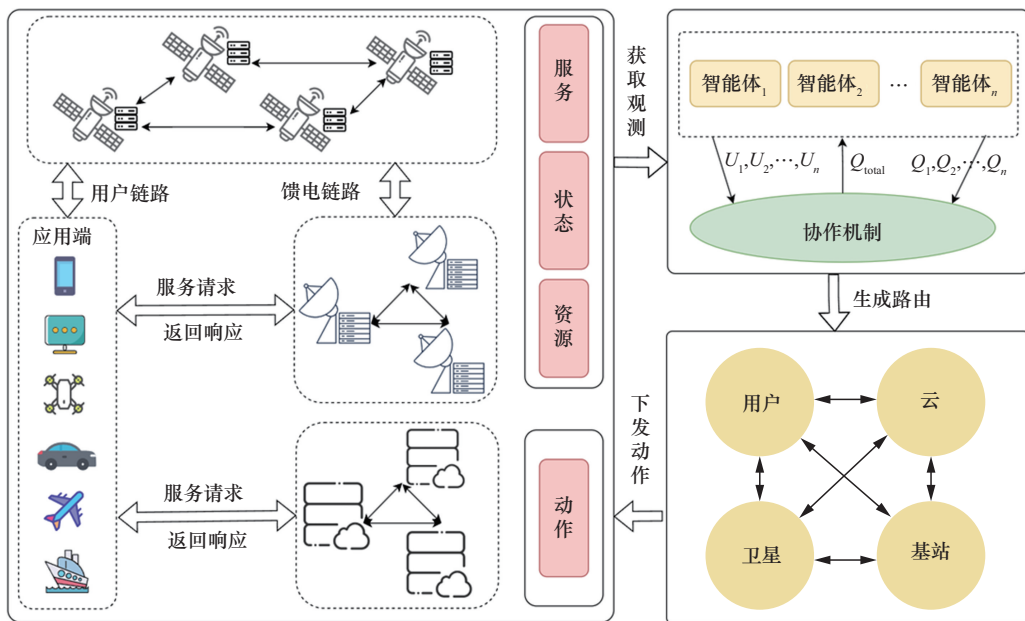


图3 基于强化学习的算力路由算法流程

然而,当面临多跳卫星网络与边缘协同计算等复杂性更高的场景时,单智能体范式在可扩展性和协同性上逐渐受限,因此研究重心自然地 toward MARL 演进。MARL 将网络中的关键节点(如卫星、边缘服务器)视为独立的智能体,它们各自感知本地链路和算力状态,通过集中式训练、CTDE 的策略协调优化资源分配与任务转发路径,能够有效应对频繁的链路切换和节点高移动性等挑战。Chen 等^[13]提出的 ComputingRoute 系统便是一个代表性应用,它在 SDN 架构下,基于多智能体软行为者-评论家(multi-agent soft actor-critic, MASAC)算法进行算力路由优化。在该系统中,每个节点作为一个智能体,利用共享经验池进行策略学习,并通过引入策略熵来增强探索性,同时采用软更新机制提升训练的稳定性,最终实现了跨节点的自适应协同路由。

在信息不完备或反馈时延场景下,常采用 POMDP 模型,通过历史轨迹或循环神经网络(recurrent neural network, RNN)推断隐含状态以增强决策鲁棒性。例如在多无人机协作中,可将问题抽象为去中心化 POMDP,并采用 CTDE 架构:训练阶段跨智能体共享全局信息进行联合优化,执行阶段则基于本地观测独立决策^[14]。

3.2 基于优化目标的算力路由技术

不同算力路由算法侧重的优化目标各异,可分为时延敏感型、能耗优化型和可靠性/负载均衡型等类别,部分工作则追求多目标折中。时延敏感型方法以最小化任务完成时延为主要目标,将每轮决策的总时延作为奖励的惩罚项^[15]。例如,有研究面向物联网任务调度,通过强化学习策略选择卫星、无人机或地面边缘进行卸载,显著降低了端到端时延^[16]。能耗优化型则关注节约设备和网络能量开销,例如以总能耗或“能耗+时延”加权和作为优化指标^[13]。这类方法通过奖励函数权衡任务完成速度与能耗,在保证时延需求的同时最大限度降低能量消耗。可靠性和负载均衡型方法则注重网络稳健性和资源利用率,如通过多路径路由提高容错能力^[17],或利用 Q 网络动态调整链路权重避免拥塞。例如, Dong 等^[18]采用 DQN 进行卫星网络负载均衡路由,在降低拥塞和丢包的同时提高了吞吐量。针对多目标优化场景,有研究通过设定加权奖励或引入层次化决策来兼顾多种性能,例如 Qiu 等^[19]构造综合代价函数同时考虑带宽分配、传输功率和计算卸载决策,提出类型匹配的 DRL 算法框架,实现了调度开销总体最小化且性能优于其他算法;又如在无人机移动边缘计算(mobile edge computing, MEC)中引入多目标 DRL 权衡时延与能耗,仿真显示该算法在

能耗和时延间取得了良好折中^[20]。

可以看出,算力路由技术的发展呈现出明显的阶段性特征。从技术演进路径来看,早期研究主要采用静态配置和启发式算法,随着网络动态性加剧,逐步转向基于强化学习的动态感知方法;优化目标从单一时延指标扩展到多维度 QoS 协同优化;系统架构从集中式控制向分布式自治转变^[12];方法论层面则实现了从人工规则驱动到数据自主学习的范式升级。这一过程反映了天地融合网络从基础通信能力保障向智能化服务效能跃升的核心转型。

4 挑战与技术展望

4.1 非平稳环境下的训练收敛挑战

天地融合网络的拓扑和业务负载高度动态化,导致强化学习环境非平稳,智能体策略在训练中面临“不断移动的靶标”。特别是在多智能体场景下,每个智能体策略的更新都会影响他人,传统单智能体容易发散。尽管 CTDE 等方法缓解了非平稳性,仍无法完全消除训练的不稳定因素。此外,卫星/UAV 高速运动造成链路时延和频繁拓扑变化,使得经验样本的时效性差、方差大。解决该问题需要发展更稳定的算法和训练范式,例如,基于元学习的策略快速适应、多智能体博弈论框架来保证策略收敛。在样本效率和训练开销方面:DRL 往往需要大量交互数据,然而在 SAGIN 中真实采集数据的成本高、风险大(不当动作可能危及网络服务)。仿真环境虽然可以生成数据,但高保真模拟 SAGIN 非常复杂,训练一个 DRL 模型也需巨大计算开销。Huang 等^[15]指出在训练初期智能体能获取的样本有限,易出现训练困难和结果不准确的问题,而且直接在真实网络上探索会有性能下降的风险。为提升样本效率,一些研究引入先验知识和模型辅助,例如在 A3C 训练中融入网络领域知识来引导策略搜索,或利用专家策略和模仿学习热启动智能体。另一个趋势是采用数字孪生技术构建逼真的虚拟环境供智能体离线训练和测试,再将策略部署到真实网络,从而避免直接试错的代价。此外,训练过程中需要解决稀疏奖励问题,可考虑 shaping 奖励或分层分解任务降低学习难度。

4.2 信息不完备条件下的决策挑战

在天地融合网络中,单个智能体往往无法获取全局网络状态,观测具有局部性和时延性。这种部分可观测性要求算法具备记忆和推断能力,否则难以及时正确决策。此外,卫星和远距离 UAV 通信的长传播时延会导致

智能体基于陈旧的信息更新策略，影响决策有效性。为应对这些问题，未来可能需要将时延敏感机制纳入强化学习，例如，利用循环神经网络或Transformer结构让智能体记忆历史信息、推断当前隐含状态，以及设计能够容忍一定信息时延的鲁棒决策策略。

4.3 网络规模引发的维度与扩展性挑战

天地一体化网络节点数众多、状态维度极高（包括拓扑连接、信道质量、计算资源余量等各类信息）且状态变化迅速。随着网络规模扩大，传统DRL的神经网络规模和训练难度呈指数级增长，出现扩展性瓶颈。为提高模型可扩展性，新兴的方法是将GNN与DRL结合。GNN能提取拓扑结构的特征并进行归纳推理，使训练得到的决策策略可泛化到不同规模的网络而无须重新训练。另外，层次化决策也是应对高维状态的一种思路，即将任务分解为分层子任务，各层次分别训练智能体决策，从而降低单个智能体面临的状态-动作空间复杂度。

5 结束语

展望未来，随着6G及天地一体化网络的演进，智能算力路由将呈现几个趋势：其一，强化学习将与其他AI技术深度融合，如结合联邦学习在分布式节点上协同训练模型，结合元学习实现策略的快速迁移适应新环境，结合大模型/生成式AI利用预训练知识辅助决策等^[12]，这些融合有望缓解纯RL方法训练开销大、冷启动慢的问题；其二，天地融合的通信-计算感知一体化趋势下，RL策略可能同时考虑通信调度、计算分配和环境感知，形成联合资源优化框架，例如将通信网络切片、计算资源编排和任务调度统一在一个多目标强化学习框架中处理；其三，在工程应用上，开发开放的仿真平台和测试床来训练和验证RL算法将十分关键；其四，随着理论和算力发展，实时在线学习有望部分实现，让部署在卫星或无人机上的智能体持续学习环境变化，自主演进策略，从而赋能天地一体化网络具备自适应、自主优化能力。在可以预见的将来，深度强化学习驱动的智能算力路由将成为6G天地融合网络的重要基石，支撑起真正普适、高效、可靠的空天地一体化信息服务。

参考文献

- [1] 吴巍. 天地一体化信息网络发展综述[J]. 天地一体化信息网络, 2020, 1(1): 1-16.
WU W. Survey on the development of space-integrated-ground in-

- formation network[J]. Space-Integrated-Ground Information Networks, 2020, 1(1): 1-16.
- [2] 曹素芝, 孙雪, 王厚鹏, 等. 星地融合网络智能路由技术综述[J]. 天地一体化信息网络, 2021, 2(2): 11-19.
CAO S Z, SUN X, WANG H P, et al. Review of intelligent routing technology in integrated satellite-ground network[J]. Space-Integrated-Ground Information Networks, 2021, 2(2): 11-19.
- [3] 田厉锋, 张佳宇, 李昕, 等. 空天地海一体化算力资源管理技术研究及展望[J]. 天地一体化信息网络, 2025, 6(2): 1-10.
TIAN L F, ZHANG J N, LI X, et al. Research and prospect of integrated computing resource management technologies for space-air-ground-sea networks[J]. Space-Integrated-Ground Information Networks, 2025, 6(2): 1-10.
- [4] SERRANO P, GRAMAGLIA M, MANCINI F, et al. Balloons in the sky: unveiling the characteristics and trade-offs of the google loon service[J]. IEEE Transactions on Mobile Computing, 2023, 22 (6): 3165-3178.
- [5] LU D, XIONG C K, LYU B Z, et al. Multi-mode hybrid aerial underwater vehicle with extended endurance[C]//Proceedings of the 2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO). Piscataway: IEEE Press, 2018: 1-7.
- [6] 许柳飞, 罗志勇. 面向低轨卫星网络的算力路由策略[J]. 移动通信, 2025, 49(6): 35-42.
XU L F, LUO Z Y. Computing power routing strategy for low earth orbit satellite networks[J]. Mobile Communications, 2025, 49(6): 35-42.
- [7] 孔梦燕, 张亚生, 董飞虎. 基于深度强化学习的低轨卫星网络算力路由研究[J]. 计算机测量与控制, 2025, 33(2): 286-292.
KONG M Y, ZHANG Y S, DONG F H. Research on computing power routing for LEO satellite networks based on deep reinforcement learning[J]. Computer Measurement & Control, 2025, 33(2): 286-292.
- [8] ALBRECHT S V, CHRISTIANOS F, SCHAFER L. Title of the monograph[M]. Cambridge: MIT Press, 2024.
- [9] CAO J, ZHANG S, CHEN Q, et al. Computing-aware routing for LEO satellite networks: a transmission and computation integration approach[J]. IEEE Transactions on Vehicular Technology, 2023, 72(12): 16607-16623.
- [10] NGUYEN T H, PARK L. A survey on deep reinforcement learning-driven task offloading in aerial access networks[C]// Proceedings of the 2022 13th International Conference on Information and Communication Technology Convergence (ICTC). Piscataway: IEEE Press, 2022: 822-827.

- [11] FU J, LIANG L, LI Y, et al. Deep deterministic policy gradient algorithm for space/aerial-assisted computation offloading[C]// Proceedings of the International Conference on Communications and Networking in China. Cham: Springer International Publishing, 2021: 523-537.
- [12] 谢明熹. 算力网络中基于深度强化学习的路由算法研究[D]. 北京: 北京邮电大学, 2024.
XIE M X. Research on routing algorithm based on deep reinforcement learning in computing power network[D]. Beijing: Beijing University of Posts and Telecommunications, 2024.
- [13] CHEN Z Y, ZHANG S C, TANG Y, et al. Optimization of computing power network routing strategies based on multi-agent soft actor-critic[C]//Proceedings of the 2024 9th International Conference on Intelligent Computing and Signal Processing (ICSP). Piscataway: IEEE Press, 2024: 426-430.
- [14] LI Y, LIANG L, FU J, et al. Multiagent reinforcement learning for task offloading of space/aerial-assisted edge computing[J]. Security and Communication Networks, 2022(1): 4193365.
- [15] HUANG L, SUN R, CHENG N, et al. Delay-oriented knowledge-driven resource allocation in SAGIN-based vehicular networks[C]// Proceedings of the 2023 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway: IEEE Press, 2023: 1-6.
- [16] 杨桂松, 黄业超, 何杏宇. 低轨卫星网络中面向时延优化的跨域边缘计算卸载策略[J]. 天地一体化信息网络, 2025, 6(2): 65-75.
YANG G S, HUANG Y C, HE X Y. Cross-domain edge computing offloading strategy for delay-optimized in low earth orbit satellite network[J]. Space-Integrated-Ground Information Networks, 2025, 6(2): 65-75.
- [17] HAN C, XIONG W, YU R H. Deep reinforcement learning-based multipath routing for LEO Mega constellation networks[J]. Electronics (Switzerland), 2024, 13(15): 3054.
- [18] DONG F, SONG J, ZHANG Y, et al. DRL-based load-balancing routing scheme for 6G space-air-ground integrated networks[J]. Remote Sensing, 2023, 15(11): 2801.
- [19] QIU Y, LUO X, NIU J, et al. Category attribute-oriented heterogeneous resource allocation and task offloading for SAGIN edge computing[J]. Journal of Sensor and Actuator Networks, 2025, 14(4): 81.
- [20] LIU X, CHAI Z Y, LI Y L, et al. Multi-objective deep reinforcement learning for computation offloading in UAV-assisted multi-access edge computing[J]. Information Sciences, 2023, 64(2): 119154.

[作者简介]



王明乾 (1999-), 男, 北京理工大学博士生, 主要研究方向为强化学习、卫星通信、路由优化。



吴晓梅 (1968-), 女, 中国铁塔股份有限公司通信技术研究院副总经理、教授级高级工程师, 主要研究方向为移动通信、卫星通信、新能源、绿色低碳、算力网络等。



杨桄 (1989-), 男, 中国铁塔股份有限公司通信技术研究院工程师, 主要研究方向为卫星通信、物联网。



雷广旺 (2002-), 男, 北京理工大学硕士生, 主要研究方向为卫星通信、路由优化。



张婷婷 (1989-), 女, 北京理工大学预聘助理研究员, 主要研究方向为天地融合网络、软件定义网络、网络功能虚拟化、智能组网。



刘子深 (2000-), 男, 北京理工大学硕士生。主要研究方向为AI赋能无线通信、卫星通信、贝叶斯推断。