高速铁路5G-R网络切片自适应资源调度策略

高云波[1]，张琦芸[1]，谢健骊[2]
（1.兰州交通大学自动化与电气工程学院，甘肃 兰州 730070；2.兰州交通大学电子与信息工程学院，甘肃 兰州 730070）

摘　要：针对5G-R/FRMCS网络切片在业务异构、节点高速移动及激励缺失的资源调度挑战，本文提出一种融合分层联邦学习（HFL）、Stackelberg博弈与多任务近端策略优化（MT-PPO）的协同智能调度框架。首先构建三层HFL架构实现非独立同分布（Non-IID）数据高效处理与资源状态预测，其次通过双层动态博弈提升节点参与度与协同效率，最后MT-PPO算法融合二者输出完成差异化调度。仿真结果表明，所提框架在业务QoS、资源利用率及吞吐量等方面较现有基线方法具有较好的性能。

关键词：5G-R；FRMCS；网络切片；资源分配；多任务近端策略优化
中图分类号：TN929.5
文献标志码：A

Adaptive Resource Scheduling for 5G-R Network Slicing in High-Speed Railways

Gao Yunbo[1], Zhang Qiyun[1], Xie Jianli[2]
1. School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
2. School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

Abstract: Aiming at the resource scheduling challenges of 5G-R/FRMCS network slicing under service heterogeneity, high-speed node mobility and insufficient incentives, a collaborative framework integrating HFL, Stackelberg game and MT-PPO was proposed. A three-layer HFL was designed to handle Non-IID data and predict resource states. A two-layer Stackelberg game was adopted to improve node participation and coordination efficiency. MT-PPO was employed to fuse both outputs for differentiated scheduling. Simulation results show that the proposed framework outperforms baseline methods in QoS, resource utilization and throughput.

Keywords: 5G-R, FRMCS, Network Slicing, Resource Allocation, Multi-Task Proximal Policy Optimization

0　引言

高速铁路的蓬勃发展对铁路专用移动通信系统的带宽、时延与可靠性提出了更高要求。为替代传统的GSM-R系统，国际铁路联盟（UIC）提出了面向未来的铁路移动通信系统（the future railway mobile communication system, FRMCS）愿景，我国也同步积极推进基于5G的铁路移动通信系统（5G-Railway, 5G-R）的技术标准体系建设。5G网络中，网络切片作为面向多业务需求的资源虚拟化技术，通过逻辑隔离方式在同一物理基础设施上构建多个独立端到端网络实例，以满足差异化业务需求。5G-R场景中，网络切片根据其支持的业务类型和服务质量（quality of service, QoS）需求[1]，被标准化地分为以下三类：增强移动宽带（enhanced mobile broadband, eMBB）切片、超可靠低时延通信（ultra-reliable and low-latency communications,

收稿日期：2026–03–06；修回日期：2026–03–23
通信作者：张琦芸，12241623@stu.lzjtu.edu.cn
基金项目：国家自然科学基金资助项目（No.62161016）；甘肃省自然科学基金（No.4JRRA227）
Foundation Items：The National Natural Science Foundation of China (No.62161016), Natural Science Foundation of Gansu Province (No.4JRRA227)

URLLC）切片以及海量机器类通信（massive machine-type communications, mMTC）切片。然而，铁路场景的资源调度在隧道、弯道等复杂线路环境中，空间传播条件与覆盖边界效应加剧覆盖不均与切换困难，工程上需通过专用天线布设或中继形态增强覆盖并降低切换终端风险[2]；同时，多切片动态资源调度还需在吞吐、时延与可靠性之间进行实时权衡。在上述工程约束叠加下，面向网络切片的资源调度还面临高动态性与强异构性的双重挑战。首先，高速移动导致网络拓扑和信道状态频繁且剧烈变化，终端设备在高速移动过程中采集的业务数据差异显著，数据呈现非独立同分布（non-independent and identically distributed, Non-IID）特性[3]，从而影响模型收敛；其次，业务负载即各切片业务的数据量与任务请求数呈现非均匀性与突发性；再者，边缘节点在算力、能耗与存储资源上存在差异，缺乏有效激励机制来保障其持续参与分布式协同优化；最后，多业务的QoS需求存在本质冲突，eMBB切片追求高带宽与高吞吐量，URLLC切片要求低时延与高可靠性，而mMTC切片则侧重海量连接与低功耗。

当前研究多在动态资源调度研究中引入深度强化学习算法（deep reinforcement learning, DRL），文献[4]将资源调度问题建模为马尔可夫决策过程（Markov decision process, MDP），推动了静态配置到动态优化的演进。文献[5]将近端策略优化（proximal policy optimization algorithm, PPO）算法应用于车载边缘计算网络，实现多智能体频谱资源动态分配，但其奖励函数面向单一优化目标。此类单智能体DRL方法难以实现异构业务的差异化QoS需求。为此，研究者引入了多智能体学习。文献[6]采用基于多智能体DRL的双时间尺度框架分别优化带宽分配与功率控制，但其模型收敛性能无法适应高速移动场景。文献[7]在高速铁路多用户移动边缘计算场景中，提出多智能体深度确定性策略梯度（multi-agent deep deterministic policy gradient, MADDPG）算法，联合优化计算卸载与频谱资源分配，有效降低了系统总成本。文献[8]构建多任务深度强化学习（multi-task deep reinforcement learning, MDRL）算法，通过共享网络结构，使单一模型能够适应多种调度场景，以提升决策效率。文献[9]提出双策略网络的协作双行动者PPO（cooperative dual-actor proximal policy optimization algorithm, CDA-PPO）算法，但其奖励函数面向固定任务，缺乏对多业务权重动态调整的机制。现有MDRL方法通过固定权重平衡多目标，但5G-R/FRMCS场景中难以适应业务优先级的动态变化。

为实现高效的分布式协同学习与环境感知，联邦学习（federated learning, FL）及其分层架构被引入。文献[10]提出移动感知的联邦DRL辅助无线电接入网（radio access network, RAN）切片方案，将FL与DRL结合，以降低通信开销，但其边缘聚合延迟影响实时调度。文献[11]提出了分层联邦学习（hierarchical federated learning, HFL）客户端选择与资源联合优化方法，进一步优化了多层网络拓扑，但在高速铁路场景业务数据量较大的情况下存在显著的传输延迟。文献[12]提出资源感知HFL方法，通过修剪技术降低带宽需求，但静态聚合周期难以适应业务负载快速变化。面向Non-IID与通信开销瓶颈，文献[13]在HFL分层结构下对客户端到边缘服务器的协同训练给出理论支撑。针对业务动态变化下的资源需求预测与实时调度的协同挑战，文献[14]引入长短期记忆（long short-term memory, LSTM）网络预测切片资源需求，并结合基于多智能体协作的近端策略优化（multi-agent proximal policy optimization, MAPPO）算法实现小时间尺度分配，但对分布式学习过程中边缘节点贡献差异与参与意愿考虑不足，文献[15]以“HFL+强化学习”的融合视角讨论分布式环境下的策略学习与开销控制，并指出HFL可在不共享原始数据的情况下协同学习策略。

此外，联邦学习中边缘节点参与意愿直接影响模型性能，因此激励机制的引入成为关键。文献[16]提出基于Stackelberg博弈的多因素激励机制，但未根据节点实时状态动态调整价格。文献[17]提出了基于Stackelberg博弈的质量感知激励机制，通过设计综合评估指标，动态激励车辆客户端参与，有效缓解Non-IID数据带来的挑战，但未考虑5G-R/FRMCS中基础设施管理者与铁路运营商之间的双层博弈关系。文献[18]构建两阶段Stackelberg博弈激励机制并评估任务难度，但未支持差异化业务定价。文献[19]为解决FL中计算资源的不足以及移动设备的异构性，提出一种基于DRL的异构Stackelberg方法，实现动态环境中对异构参数的控

制。现有激励机制未涉及多业务QoS冲突与切片级资源调度，难以实现全局最优的激励效果。

综上所述，现有研究已在HFL、Stackelberg博弈及深度强化学习算法方面分别取得了一定进展，并已围绕5G-R网络切片的部署与运维管理开展了相关探索[20]，但在5G-R/FRMCS网络切片资源调度场景下仍存在一定局限性。一方面，已有方法多聚焦于单一环节或其中两类机制组合优化，或仅对其中两类机制进行组合设计，难以协同应对高动态环境下的精准预测、异构节点的高效激励与多冲突目标的合理决策这一交叉性挑战。另一方面，若对业务负载预测、节点激励和资源调度等环节分别进行优化而缺乏协同，易导致各环节之间信息缺乏交互，从而使系统整体性能受限[21]；对于网络切片资源调度场景而言，分离式设计还会进一步削弱资源调度的协同效率[22]。具体而言，当前仍缺乏一个能够同步感知环境趋势、动态调整节点参与和自适应权衡多业务QoS的协同智能调度框架。

为此，本文提出一个融合HFL、Stackelberg博弈与多任务近端策略优化（multi-task proximal policy optimization, MT-PPO）的协同智能调度框架，以实现感知、激励与决策的闭环协同优化。本文的主要贡献包括：首先，设计面向高速移动与Non-IID数据的三层HFL架构，通过集成业务负载预测的自适应聚合机制，提升模型环境预测能力与收敛稳定性；其次，构建了一个以网络切片提供商为领导者、应用服务提供商为跟随者的双层动态Stackelberg博弈模型，通过差异化效用函数实时调整激励策略，以提升边缘节点参与度；最后，将HFL的预测结果与博弈的激励系数共同融入MT-PPO智能体的状态空间，并结合广义优势估计（generalized advantage estimation, GAE）与并行任务训练，使资源调度策略同时具备环境前瞻性、激励响应性与业务自适应性，最终实现eMBB、URLLC与mMTC三类异构切片资源的差异化高效调度。

1　系统模型及问题表述

1.1　系统模型

本文提出面向5G-R/FRMCS异构业务场景，构建“终端训练-边缘协调-中心策略”三层架构，融合HFL、Stackelberg博弈激励机制与MT-PPO强化学习算法，实现资源的协同优化与智能调度。系统模型如图1所示。

第一层为终端训练层，该层由海量异构业务终端构成。令$\mathcal{N}=\{n_1,n_2...,n_N\}$为终端业务集合，每个终端隶属于相关网络切片$k\in\{eMBB,URLLC,mMTC\}$，各终端在列车及沿线场景中分布，通过基站接入网络，并基于本地数据集在每轮接收来自上层的全局模型参数$w_j^{(i)}$，执行本地模型更新以得到本地参数$w_n^{(i+1)}$，并上传结果以参与全局模型训练。在终端训练层与边缘聚合层之间，引入基站作为无线接入节点。设基站集合为$\mathcal{B}$，其沿铁路线路呈连续分布，负责终端与MEC服务器之间的模型参数及控制信息传输。

第二层为边缘聚合层，由多个边缘计算（multi-access edge computing， MEC）服务器组成，记为集合$\mathcal{J}$，每个MEC服务器与覆盖区域内的基站形成映射关系，且集成三项核心功能：HFL边缘聚合功能对终端本地模型进行聚合并生成边缘模型$w_j^{(i+1)}$；激励执行代理响应上层激励策略并反馈本地资源状态与贡献度；资源策略执行功能负责将上层下发的全局资源调度策略在本地执行，控制终端训练层的资源使用。

第三层为中心决策层，负责全局协同。其中，HFL全局聚合与预测模块负责聚合所有边缘模型生成新的全局模型，并据此预测未来业务负载与QoS趋势，为决策提供前瞻信息；Stackelberg博弈模块动态生成用于激励边缘节点积极参与的激励系数以调节节点行为；MT-PPO智能决策模块则综合分析实时网络状态、预测信息及激励反馈，输出全局资源分配策略。

整体系统模型可表示为：

$$\mathcal{M}=\{\mathcal{N},\mathcal{K},\mathcal{R},\mathcal{A},\mathcal{U}\} \tag{1}$$

其中，$\mathcal{N}$为节点集合，$\mathcal{K}=\{k_1,k_2,k_3\}$表示三类切片（eMBB、URLLC、mMTC），$\mathcal{R}$为可分配资源集合，将资源抽象为计算、存储与网络三类[23]，$\mathcal{A}$为可执行的调度与接入策略，$\mathcal{U}$为实际消耗资源集合。

1.2　问题表述

1.2.1　分层联邦学习建模

为适应5G-R/FRMCS网络切片中异构节点的动态特性，本文采用HFL方法。HFL构建了包含

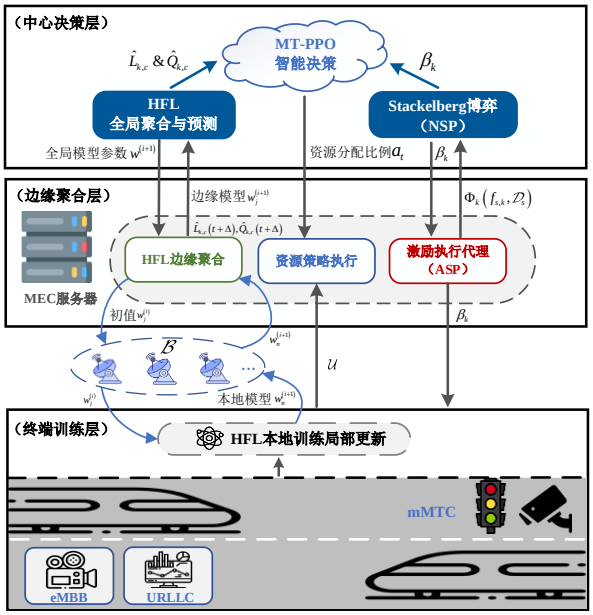


图1　“终端训练-边缘协调-中心策略”三层架构模型图

中心决策层、边缘聚合层和终端训练层的多级架构，该方法包含三类操作。

（1）局部更新

各终端设备 n 以 MEC $j \in \mathcal{J}$ 在第 i 轮时 MEC j 的边缘模型 $w_j^{(i)}$ 为初值，基于本地数据 D_n 计算损失函数 $L_n(\cdot)$，并执行随机梯度下降，得到本地模型参数 $w_n^{(i+1)}$，其计算式为：

$$w_n^{(i+1)} = w_j^{(i)} - \eta \nabla L_n\left(w_j^{(i)}\right) \tag{2}$$

其中，η 为本地学习率，$L_n(\cdot)$ 为基于 D_n 的损失函数。

（2）边缘聚合

MEC j 对终端 n 上传的本地模型参数进行加权平均，聚合得到边缘模型 $w_j^{(i+1)}$，然而，由于 Non-IID 数据与节点参与率不稳定且数据质量具有显著差异，本文在 HFL 聚合阶段引入综合质量感知权重 ε_n，将数据质量评分 q_n（由本地损失收敛度量化）以及节点参与稳定性 ρ_n 纳入联合计算，从而使贡献更高、数据质量更优、参与更稳定的节点在聚合中获得更合理的权重占比。

设终端设备 x 在时隙 t 的上行传输延迟为 d_x^t，由当前信道条件与列车速度共同决定；链路丢包率为 p_{loss}，受多普勒效应与阴影衰落影响，在高速场景下呈动态变化特征。此外，MEC j 的算力上界记为 C_m，在三类切片业务高负载时段，聚合计算时延随有效参与节点数增大而增加，当 C_m 不足时将导致部分节点更新超出时延上界而被丢弃。时隙 t 有效参与的终端集合为 $S_t^j = \left\{x \middle| d_x^t \leqslant T_{end} \wedge \xi_x^t = 1 \wedge \varphi_x^t\left(C_m\right) = 1\right\}$，其中 T_{end} 为截止时延，$\xi_x^t \in \{0,1\}$ 为丢包指示变量，其中 1 表示上传成功，$\varphi_x^t\left(C_m\right) \in \{0,1\}$ 为 MEC 算力约束指示变量，取 0 表示该节点更新被丢弃。

因此，边缘聚合模型表示为：

$$\begin{aligned} w_j^{(i+1)} &= \sum_{n \in S_t} \frac{\varepsilon_n}{\sum_{k \in S_t} \varepsilon_k} w_n^{(i+1)}, \varepsilon_n \\ &= \left|D_n\right| \cdot q_n \cdot \rho_n \cdot e^{-\frac{d_x^t}{T_{end}}} \cdot \xi_x^t \cdot \varphi_x^t\left(C_m\right) \end{aligned} \tag{3}$$

其中，$\left|D_n\right|$ 为终端 n 的本地数据集大小。

（3）全局聚合

中心决策层接收所有边缘服务器上传的边缘模型参数，再次根据 ε_n 进行加权平均，得到全局模型参数 $w^{(i+1)}$，计算式为：

$$w^{(i+1)} = \sum_j \frac{E_j}{\sum_j E_j} w_j^{(i+1)}, E_j = \sum_{n \in S_t^j} \varepsilon_n \tag{4}$$

除了完成参数聚合外，HFL 还对未来时隙的业务负载与 QoS 需求进行预测。

设 $\hat{L}_{k,c}(t+\Delta)$ 与 $\hat{Q}_{k,c}(t+\Delta)$ 分别表示在小区 c、切片 k 的业务负载与 QoS 的预测值，预测窗口为 Δ。该预测由聚合后的模型计算得到：

$$\begin{aligned} \hat{L}_{k,c}(t+\Delta) &= f_{HFL}\left(L_{k,c}(t), w^{(i+1)}\right), \hat{Q}_{k,c}(t+\Delta) \\ &= g_{HFL}\left(Q_{k,c}(t), w^{(i+1)}\right) \end{aligned} \tag{5}$$

其中，$w^{(i+1)}$ 为全局聚合后的模型参数，$\hat{L}_{k,c}(t+\Delta)$ 表示小区 c、切片 k 的未来业务负载预测；$\hat{Q}_{k,c}(t+\Delta)$ 表示小区 c、切片 k 的未来 QoS 需求预测。预测结果不仅能辅助资源供需匹配，还将在后续的 MT-PPO 状态输入中提供前瞻信息。

1.2.2　Stackelberg 博弈激励机制建模及均衡分析

为解决在 5G-R/FRMCS 网络切片场景中由于边缘节点资源有限且异构分布导致的 HFL 节点参与度不均问题，本文构建一个以网络切片提供商（network slice provider, NSP）为领导者、应用服务提供商（application service provider, ASP）为跟随者的 Stackelberg 博弈模型[18]，如图 2 所示。该模型

旨在通过NSP设计的差异化资源定价与动态激励，精准引导ASP优化其资源分配，并提升边缘节点参与HFL训练的积极性。

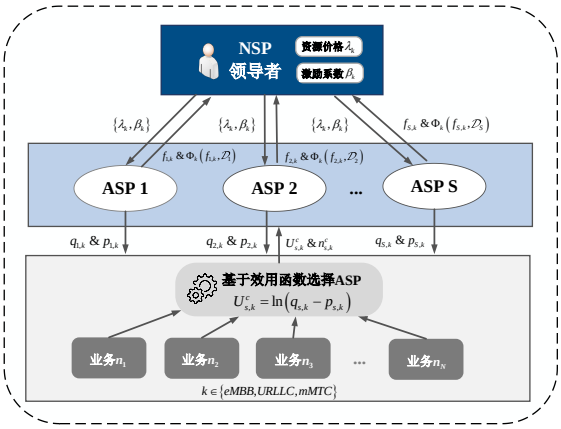


图2　基于Stackelberg博弈的异构业务激励

将NSP设定为“领导者”，作为物理及虚拟网络基础设施的管理者，制定面向不同业务类型$k\in\{eMBB,URLLC,mMTC\}$的单位资源价格λ_k与激励系数β_k；ASP设定为“跟随者”，记为$S=\{1,2,...,S\}$，每个ASP$s(s\in S)$作为租用切片以承载各类业务的边缘节点实体，在观测到NSP的策略后，确定向业务k投入的资源量$f_{s,k}$以及向用户收取的服务订阅价格$p_{s,k}$，以最大化自身效用。因此，NSP将先行公布全局策略$\Lambda=\{\lambda_k,\beta_k\}$，各ASP据此决策最优资源分配方案$F=\{f_k\}$，以最大化自身效用。

将小区c中，一个用户选择由ASPs提供的业务k所获得的效用定义为：

$$U^c_{s,k}=\ln\left(q_{s,k}-p_{s,k}\right) \tag{10}$$

其中，$q_{s,k}=Q_k\left(f_{s,k}\right)$是业务提供的服务质量，其中$Q_k(\cdot)$是单调递增函数，$p_{s,k}$是ASP$s$对业务$k$的订阅定价，对数函数反映了用户满意度的边际效用递减效应。

ASP的效用函数如下：

$$U^{(s)}_{ASP}=\underbrace{\sum_k p_{s,k}\cdot n^c_{s,k}}_{\text{业务收入}}-\underbrace{\sum_k \lambda_k\cdot f_{s,k}}_{\text{资源成本}}-\underbrace{\mathcal{P}\left(f_{s,k}\right)}_{\text{QoS违约惩罚项}}+\underbrace{\sum_k \beta_k\cdot\Phi_k\left(f_{s,k},\mathcal{D}_s\right)}_{\text{核心激励项}} \tag{11}$$

其中，$n^c_{s,k}$是连接到ASPs的服务类型用户k的数量，λ_k为资源单价，$P\left(f_{s,k}\right)$是QoS违约惩罚项，$\mathcal{D}_s$是ASP的本地数据集，$\Phi_k\left(f_{s,k},\mathcal{D}_s\right)$为贡献度函数，其量化了ASP在资源投入以及HFL训练的贡献，并提升联邦学习模型的训练预测能力。

对于任意给定的NSP策略$\Lambda=\{\lambda_k,\beta_k\}$，由于效用函数$U^{(s)}_{ASP}$关于其决策变量$\left(f_{s,k},p_{s,k}\right)$为严格凹，ASP最大化效用函数存在唯一最优解。该解定义了ASP对NSP策略的最优响应函数R，表示如下：

$$f^*_{s,k}=\mathrm{R}^f_{s,k}\left(\lambda_k\cdot\beta_k\right),p^*_{s,k}=\mathrm{R}^p_{s,k}\left(\lambda_k\cdot\beta_k\right) \tag{12}$$

NSP作为领导者，在制定策略时能够预见跟随者ASP将依据其响应函数采取行动。因此，NSP的效用函数将基于此预见进行构建。其效用函数如下：

$$U_{NSP}=\underbrace{\sum_{s,k}\lambda_k\cdot\mathrm{R}^f_{s,k}\left(\lambda_k,\beta_k\right)}_{\text{资源租赁收益}}-\underbrace{\mathcal{L}\left(\left\{\mathrm{R}^f_{s,k}\left(\lambda_k,\beta_k\right)\right\}\right)}_{\text{系统性能损失}}-\underbrace{\frac{\theta}{2}\sum_k{\beta_k}^2}_{\text{激励成本}} \tag{13}$$

资源租赁收益为NSP通过向ASP出租切片资源所获得的总收益，$\mathcal{L}(\cdot)$是保障网络整体服务质量的惩罚项，促使NSP在定价时兼顾网络性能，以防止过度激励。

其次，激励因子$\eta(t)$为所有ASP在激励系数β_k作用下的整体贡献响应，定义如下：

$$\eta(t)=\frac{1}{|S|}\sum_{s\in S}\sum_k\beta_k\cdot\Phi_k\left(f_{s,k},\mathcal{D}_s\right) \tag{14}$$

该因子结合激励强度与节点实际贡献，直接反映激励机制的整体有效性。

定义1（Stackelberg均衡）：策略$\left(\Lambda^*,\left\{f^{**}_{s,k}\right\},\left\{p^{**}_{s,k}\right\}\right)$称为该博弈的一个Stackelberg均衡，当且仅当满足：

（1）对于所有s、k，有$f^{**}_{s,k}=\mathrm{R}^f_{s,k}\left(\lambda^*_k,\beta^*_k\right)$且$p^{**}_{s,k}=\mathrm{R}^p_{s,k}\left(\lambda^*_k,\beta^*_k\right)$。

（2）给定跟随者的响应函数R，策略$\Lambda^*=\left(\lambda^*_k,\beta^*_k\right)$是NSP效用函数最大化的全局最优解。

2　基于MT-PPO的自适应调度

在HFL提供预测、Stackelberg博弈下发激励的基础上，为应对5G-R/FRMCS网络切片资源调度的高动态性和异构性所带来的挑战，本研究将设计一种基于MT-PPO的自适应调度框架。该框架将调

度问题建模为马尔科夫决策过程（markov decision process，MDP）[24]，并引入HFL预测结果与Stackelberg博弈下发的激励系数作为状态输入，同时将Stackelberg博弈的激励因子融入奖励函数，使得调度策略具备前瞻性与激励响应能力。

2.1 马尔可夫决策过程

在动态场景下，如何根据实时状态选择资源调度策略，是本研究的核心优化目标。本研究将该过程建模为一个马尔可夫决策过程[25]：

$$M = \left(S,A,P,R,\gamma\right) \tag{15}$$

其中，S表示状态；A表示为动作，对应资源调度决策；P为状态转移概率；R表示奖励函数，用于评估状态价值；$\gamma\epsilon(0,1)$为折扣因子，衡量未来QoS对当前决策的重要性。MT-PPO智能体根据每一时刻当前状态$s_t \in S$选择动作$a_t \in A$，系统根据转移概率转移到下一个状态s_{t+1}，并获得奖励r_t。通过不断交互，目标是学习最优策略以最大化累积折扣奖励。

2.1.1 状态空间设计

系统状态反映当前实时网络与业务状态，本文创新性地融入了HFL预测结果与Stackelberg博弈激励系数，使MT-PPO具备了环境前瞻感知与节点激励状态感知的双重能力，从而为其决策提供了兼具实时性、前瞻性与激励感知的综合决策依据，具体如下：

$$s_t = \left[L_{k,c}(t), Q_{k,c}(t), \underbrace{\hat{L}_{k,c}(t), \hat{Q}_{k,c}(t)}_{HFL预测}, \underbrace{\beta_k}_{Stackelberg激励} \right] \tag{16}$$

其中，$\hat{L}_{k,c}(t)$、$\hat{Q}_{k,c}(t)$分别表示通过HFL预测得到的业务负载与QoS趋势，β_k为Stackelberg博弈中NSP下发的激励系数，使智能体感知当前激励策略。

2.1.2 动作空间设计

动作空间为连续向量，表示对不同切片类型的资源分配比例，令第k类切片在时隙t的资源分配比例为a_t^k，其可以表示为：

$$a_t^k = \left\{\alpha_1,\alpha_2...\alpha_N\right\}, \sum_{i=1}^{N}\alpha_i = 1, a_i \in [0,1] \tag{17}$$

该比例将进一步转化为系统可执行的具体资源分配方法，并通过MT-PPO框架的训练，算法学习得到最优策略参数θ^*，从而确定最优策略π_{θ^*}。

2.1.3 奖励函数设计

针对不同切片QoS指标不一致的问题，首先将QoS进行归一化处理。设第k类切片在时隙t的QoS为QoS_t^k，对应的需求阈值为QoS_k^{req}，则定义其归一化QoS为$QoS_t^{k\prime} = \min\left\{\frac{QoS_t^k}{QoS_k^{req}}, 1\right\}$，可将各类业务的$QoS_t^k$映射到[0,1]区间，以实现统一度量。最终奖励函数表示为：

$$r_t^k = \sum_{k\in\mathcal{S}}\left[\omega_1 \cdot QoS_t^{k\prime}\right] + \omega_2 \cdot \eta(t) - \omega_3 \cdot Penalty_t^k \tag{18}$$

其中，$\eta(t)$为激励因子，将Stackelberg博弈的激励效果转化为可量化的奖励信号，使MT-PPO能够学习到提升节点参与度的调度策略，从而实现激励与调度的协同优化，$Penalty_t^k$为QoS违约惩罚，随违约程度增加而增大，权重系数ω_1、ω_2、ω_3用于平衡各项优化目标。

2.2 MT-PPO算法优化设计

MT-PPO作为改进型的策略梯度方法，通过限制新旧策略的变化幅度，在有效避免策略更新过快的同时，通过多次迭代更新以提高样本利用率。本文所提的5G-R/FRMCS网络切片资源调度问题，本质目标是在满足资源分配可行性和QoS约束的前提下，联合优化$QoS_t^{k\prime}$、资源违约惩罚和激励机制。资源调度的长期优化目标可表示为：

$$\begin{aligned}
&\mathrm{P}: \max_{\pi_\theta} U_{total} = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T}\gamma \cdot r_t^k\right] \\
&\text{s.t. C1}: \sum_{k\in\mathcal{K}} a_t^k = 1, \forall t \\
&\text{C2}: a_t^k \geqslant 0, \forall k,t \\
&\text{C3}: QoS_t^{k\prime} \geqslant QoS_k^{\min}, \forall k,t \\
&\text{C4}: Penalty_t^k \leqslant P^{\max}, \forall k,t
\end{aligned} \tag{19}$$

约束条件C1表示各时隙资源分配和为1，约束条件C2表示各类业务切片资源比例非负，约束条件C3表示各类切片的QoS必须不低于最小QoS门限$Q_k^{\min}$，约束条件C4表示每时隙的服务违约惩罚不超过容忍上限$P^{\max}$。

由于上述优化问题具有非凸性，本文基于MT-PPO构造多任务损失函数，并通过对该损失函数的迭代优化，实现对原始资源调度问题的近似求解：

$$L^{MT-PPO}(\theta) = \sum_{k\in\{eMBB,URLLC,mMTC\}} w_k \cdot L_k^{PPO}\left(\theta_k\right) \tag{20}$$

其中

$$L_k^{PPO}(\theta_k)=\hat{\mathbb{E}}_t\left[\min\left(r_t^k(\theta_k)\hat{A}_t^k,clip\left(r_t^k(\theta_k),1-\varepsilon,1+\varepsilon\right)\hat{A}_t^k\right)\right] \tag{21}$$

其中，θ_k 为 k 类业务的专属策略头参数，$\theta=\{\theta_k\}_{k\in\{eMBB,URLLC,mMTC\}}$ 为完整参数集，w_k 为业务权重系数，反映了业务优先级，ε 为裁剪参数，$clip(\cdot)$ 函数用于限制新旧策略概率比 $r_t^k(\theta_k)$，从而约束策略更新幅度，$\hat{A}_t$ 为优势函数。具体地，$r_t^k(\theta_k)=\dfrac{\pi_{\theta_k}(a_t^k|s_t)}{\pi_{\theta_k^{old}}(a_t^k|s_t)}$，其中将调度策略记为 π_{θ_k}，整体调度策略记为 $\pi_\theta=\{\pi_{\theta_k}\}_{k\in\{eMBB,URLLC,mMTC\}}$。

为了进一步提升优势函数估计的精确性与稳定性，本研究在 MT-PPO 中引入广义优势估计（generalized advantage estimation，GAE）[26]，计算公式如下：

$$\hat{A}_t^{GAE(\gamma,\lambda)}=\sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V \tag{22}$$

其中，$\delta_t^V=r_t^k+\gamma V(s_{t+1})-V(s_t)$ 为时序差分（Temporal-Difference，TD）误差，用于衡量当前值函数估计的偏差，$V(\cdot)$ 为状态值函数，用于估计当前状态下预期能获得的累计奖励，$\lambda\in[0,1]$ 为 GAE 超参数，用于在估计的偏差与方差之间进行权衡。

2.3　算法流程

本研究所提出的基于 HFL 预测与 Stackelberg 激励的 MT-PPO 资源调度算法是一个周期性的迭代过程。

如图 3 所示，首先通过状态观测模块收集环境反馈的异构切片负载、HFL 预测信息及 Stackelberg 激励系数，构成 MT-PPO 智能体的状态输入。MT-PPO 智能体基于该架构并利用上述状态生成面向三类异构业务的资源分配动作，并通过价值网络提供状态价值估计。每一轮决策执行后，通过奖励函数综合 QoS、违约惩罚与 Stackelberg 激励因子计算奖励值。MT-PPO 策略优化模块通过 GAE 实现方差-偏差权衡，并借助含裁剪机制的优化目标函数更新网络参数，最终形成“观测-决策-优化”的训练闭环，直至收敛至最优调度策略 π_{θ^*}。本文提出的 MT-PPO 资源调度算法的伪代码如算法 1 所示。

算法 1　基于 HFL 预测与 Stackelberg 激励的 MT-PPO 资源调度算法

输入 折扣因子 γ、GAE 参数 λ、裁剪系数 ε、策略网络参数 θ、价值网络参数 v

输出 最优调度策略参数 θ^*

1) 初始化 θ，v

2) for 每一轮训练回合 $n=1,2,\ldots$ do

3) HFL 获得业务负载与 QoS 预测结果：$\hat{L}_{k,c}(t+\Delta),\hat{Q}_{k,c}(t+\Delta)\leftarrow f_{HFL}\left(w_j^{(i)}\right)$

4) 激励机制下发激励系数 β_k

5) for 决策时隙 t do

6) 生成资源分配动作 a_t^k

7) 执行动作并分配切片资源

8) 根据奖励函数计算即时奖励 r_t^k

9) 状态转移至 s_{t+1}

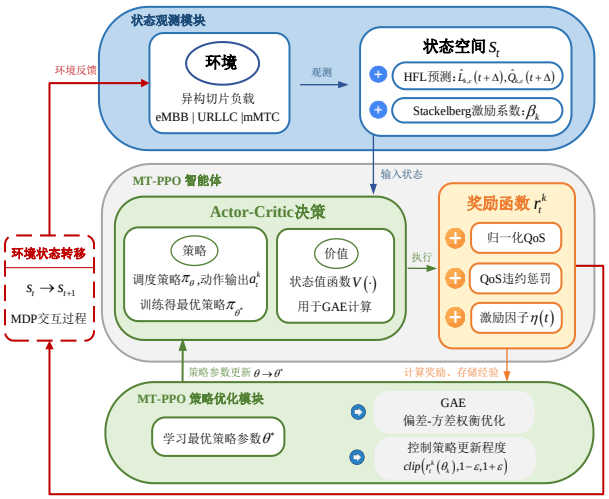


图 3　基于 MT-PPO 优化算法示意图

10) end for

11) 计算误差：$\delta_t^V = r_t + \gamma V\left(s_{t+1}\right) - V\left(s_t\right)$

12) 计算优势函数 $\hat{A}_t^{GAE(\gamma,\lambda)}$

13) 构建 MT-PPO 裁剪函数 $L_k^{PPO}\left(\theta_k\right)$

14) 更新多任务策略网络参数 θ

15) 更新价值网络参数 v

16) end for

17) 输出最优策略参数 θ^*

算法 1 的时间复杂度分析如下。第 1)行到第 2)行完成模型初始化，仅执行一次，时间复杂度为 $O(1)$。第 3)行通过 HFL 模块得到业务负载与 QoS 预测结果，设 HFL 模型参数规模为 N_p^{HFL}，其复杂度为 $O\left(N_p^{HFL}\right)$，并可在 N_{MEC} 个边缘节点并行执行。第 4)行 Stackelberg 博弈根据最优响应函数计算激励系数，参与 Stackelberg 博弈的 ASP 数为 $|S|$，设均衡求解迭代次数为 I，复杂度同为 $O\left(I\cdot|S|\right)$。第 5)到第 10)行为在线决策过程，设输入层维度为 n_1，共享层维度为 n_2，三类切片专属策略头维度分别为 $n_3^{(1)}$、$n_3^{(2)}$、$n_3^{(3)}$，T 个时隙总复杂度为 $O\left(T\left(n_1n_2 + n_2\sum_{i=1}^{3} n_3^{(i)}\right)\right)$。第 11)行到 15)为策略更新阶段，策略与价值网络参数规模分别为 N_p^{π} 和 N_p^{V}，更新轮次为 K，复杂度为 $O\left(KT\left(N_p^{\pi}+N_p^{V}\right)\right)$。综上单轮训练复杂度为 $O\left(N_p^{HFL} + I\cdot|S| + T\left(n_1n_2 + n_2\sum_{i=1}^{3} n_3^{(i)}\right) + KT\left(N_p^{\pi}+N_p^{V}\right)\right)$。此外，从工程实现角度来看，实际部署中策略与价值网络更新的复杂度近似为 $O\left(N_p^{\pi}+N_p^{V}\right)$，即与网络参数规模线性相关，在 MEC 计算能力范围内可高效完成，避免系统开销的显著增加，在保证收敛性能的同时减少训练轮次，因此整体具备实际部署可行性。

3　仿真分析

3.1　仿真场景与仿真参数设置

为验证所提“HFL+Stackelberg+MT-PPO”方法框架在 5G-R/FRMCS 场景下的性能，本节构建一段长度为 L=50km 的高速铁路线路，沿线均匀部署 N_{bs}=25 个 5G-R 基站（gNB），基站间距 d_{bs}=1.5km，以确保高速列车在最高运行速度 $v_{\max}$=360km/h 下的连续覆盖与无缝切换。系统架构包含 1 个中心云服务器与 N_{mec}=6 个沿线路旁部署的边缘 MEC 服务器，以构成云边协同的计算网络。铁路沿线按密度 ρ_{sensor}=200 个/km 部署轨旁物联网终端，构成大规模机器类通信（mMTC）业务源。为保证算法收敛性和训练效率，强化学习超参数在所有实验中保持统一：学习率设置为 η=0.03 避免优化震荡；折扣因子 γ=0.99，以使智能体注重长期优化。探索率 ε 从 0.90 线性衰减至 0.05，实现从广泛探索到精准利用的平滑过渡。仿真实验中将系统总带宽设置为 100MHz，该参数基于 UIC 对未来铁路业务更宽频谱资源的考虑[27]。仿真基于 Python 3.12 实现，具体仿真实验参数设置如表 1 所示。

在业务模型方面，eMBB 切片承载每辆列车并发数量不等的高清视频监控与实时信息发布业务，单流带宽需求为 8Mbps 且业务到达过程服从泊松分布，URLLC 切片保障紧急告警等安全关键业务，数据包大小在 100-500Bytes 中均匀分布，要求端到端时延严格低于 10ms 且可靠性不低于 99.999%，mMTC 切片模拟大规模轨旁基础设施监测，总计 10000 个轨旁物联网终端，平均每 30s 上报 100Bytes 的监测数据，要求大规模连接下成功率高于 95%。

3.2　实验结果分析

为验证本文所提“HFL+Stackelberg+MT-PPO”方法框架在 5G-R/FRMCS 动态多业务场景下的整体有效性及其核心组件的贡献，本节首先设计了消融实验以验证 PPO 相关改进算法、HFL 与 Stackelberg 博弈激励机制的独立作用与协同增益，其次通过对比实验将所提方法与多种基线算法从 QoS 满意度、系统吞吐量、算法收敛性等多个维度进行比较，综合验证所提方法的优越性与鲁棒性。

3.2.1　消融实验

图 4 通过 PPO 消融实验验证了所提方法中各模块的贡献及其协同作用。实验结果表明，相比传统 PPO 方法[5]，引入 GAE 机制[26]后性能提升约 10.1%，主要归因于改进的价值估计降低了策略梯度方差，进一步采用多任务 PPO[28]结构后，QoS 满足率与资源利用率分别提升至 89.6% 与 86.8%，证明其能够有效缓解业务间的资源竞争，但仍存在训练波动较大的问题，在此基础上整合 MT-PPO+GAE 框架，QoS 满足率和资源利用率进一步提升

表1　　仿真参数

符号	说明
铁路线路长度L/km	50
基站数量N_{bs}/个	25
基站间距d_{bs}/km	1.5
系统总带宽（每基站）B_{total}/MHz	100
资源块总数（每基站）N_{rb}	273
边缘服务器数量N_{mec}	6
列车速度v/(km/h)	[200,360]
载波频率f_c/GHz	2.1
系统总带宽B_{total}/MHz	100
子载波间隔Δf/kHz	30
切片初始带宽比β_{eMBB}:β_{URLLC}:β_{mMTC}	0.5:0.3:0.2
业务负载监测周期T_m/ms	100
学习率η	0.03
折扣因子γ	0.99
ε - *greedy* 探索率起止值ε_{start}，ε_{end}	0.90,0.05
PPO 裁剪参数ε_{clip}	0.2
PPO 算法更新参数轮次N_{epoch}	10
一次经验学习抽取的数据量B_{batch}	64
总训练时间步T_{total}/ms	1×10^6
截止时延T_{end}/ms	50
链路丢包率p_{loss}	0.2
MEC 服务器算力上限C_m	0.25

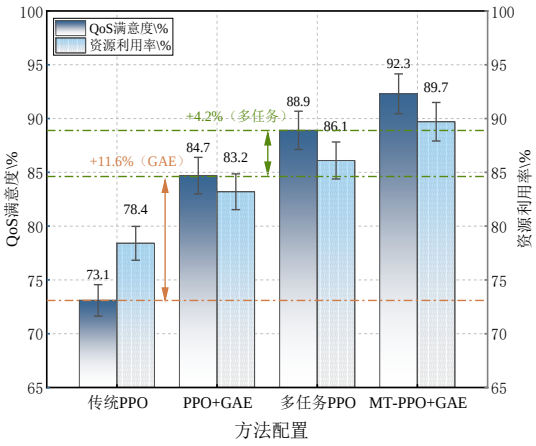


图4　PPO 改进方法性能对比

至 93.5% 与 90.4%。该结果表明，多任务结构与 GAE 机制协同作用，使系统在保障各类切片 QoS 的同时，有效提升训练收敛性与策略鲁棒性。

图 5 展示了 PPO 改进方法在训练过程中的归一化累积奖励收敛曲线。实验结果表明，所提出的 MT-PPO+GAE 方法更快地收敛至 0.95 左右，这一效率优势得益于 GAE 机制的低方差优势估计与多任务学习的梯度共享的协同效应。从收敛曲线可见，传统 PPO 在训练初期震荡剧烈，反映其价值估计方差较高；GAE 的引入显著改善了训练稳定性，但单目标优化限制了收敛速度；多任务 PPO 虽加快了学习进程，但训练后期仍存在性能震荡；而所提方法不仅实现了最快收敛，更在达到收敛后保持了最小的性能波动，放大的子图中呈现的关键收敛阶段进一步证实了这一优势。快速而稳定的收敛特性使所提方法适配于 5G-R/FRMCS 中列车高速移动、频繁切换等动态特点，显著降低了实际部署中的训练成本和时间开销。

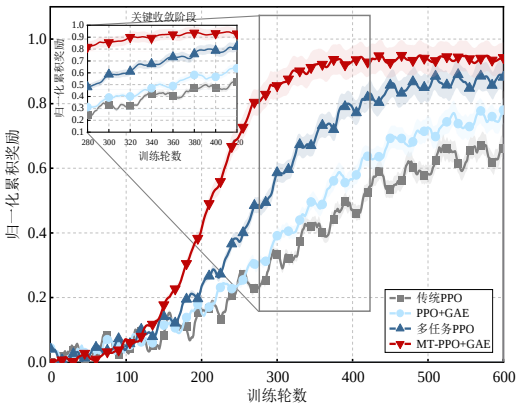


图5　PPO 训练收敛曲线对比

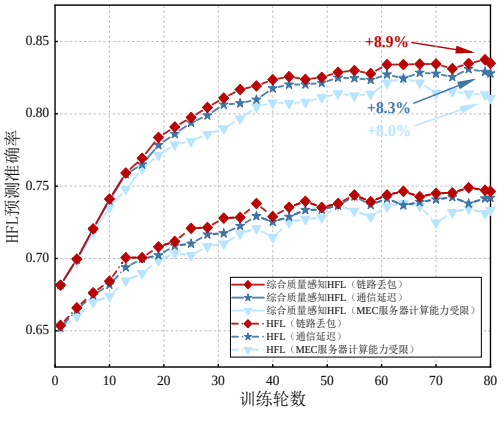


图6　HFL 聚合引入综合质量感知权重效果对比

图 6 展示了三种通信受限场景下 HFL 聚合阶段引入ε_n后 HFL 预测准确率的效果对比。在链路丢包场景中，引入ε_n后 HFL 预测准确率达到 0.83，而 HFL 仅约为 0.74，主要得益于ε_n中丢包指示变量ξ_x^t的过滤作用，降低低质量、高延迟节点的权重因

子。在通信延迟场景中，引入ε_nHFL 预测准确率总体提升约 8.1%，收敛更为平稳。在 MEC 服务器计算能力受限场景中，两者差距约 7.7%，验证了算力不足后ε_n仍能通过q_n和ρ_n维持聚合准确率。上述结果表明，引入ε_n能够在高动态铁路通信环境下有效提升 HFL 聚合的预测准确率和鲁棒性。

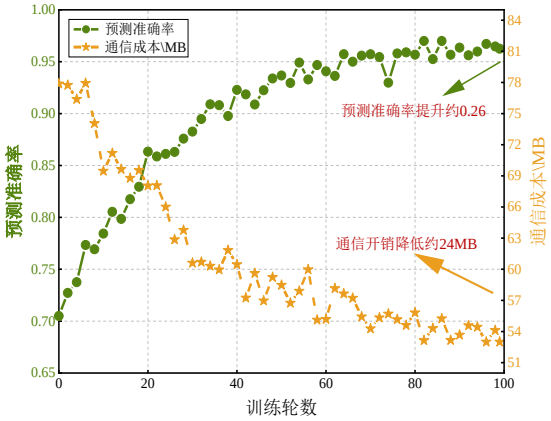


图 7　HFL 分层聚合效果

图 7 体现了 HFL 层级聚合机制的性能表现。随着训练轮次的增加，HFL 预测准确率稳步提升，从初值 0.70 提升至 0.96，增长约 0.26，前 37 轮快速收敛并持续稳步增长，同时通信成本由 78MB 降至 54MB，表明了 HFL 三层架构通过局部模型聚合，显著减少了原始数据向中心决策层的直接传输，ε_n的引入进一步降低了低贡献节点的无效传输占比，在保证模型性能的前提下有效降低通信开销，证明其在模型性能与通信效率之间能够取得良好平衡。

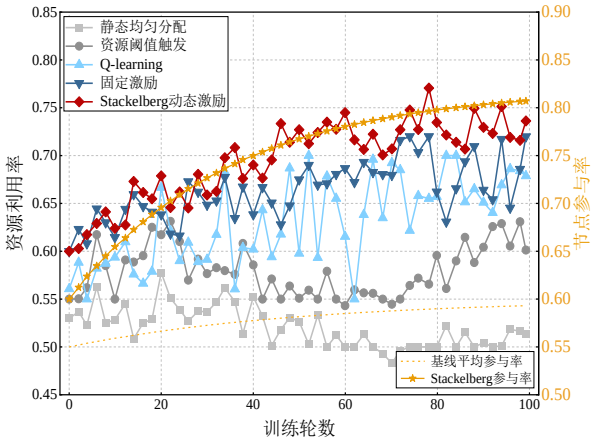


图 8　Stackelberg 激励效果对比

图 8 验证了 Stackelberg 激励机制的优势。在资源利用率方面，Stackelberg 动态激励方法从初始的 0.60 快速上升，最终稳定在 0.77 左右，固定激励次之，最终收敛至 0.72 左右，但波动较为明显，Q-learning 收敛至 0.70 左右，波动较大且不够稳定，资源阈值触发方法性能中等，而静态均匀分配从初始 0.53 逐渐下降，说明缺乏动态调整机制导致资源浪费严重。节点参与率方面，Stackelberg 动态激励机制通过 NSP 与 ASP 之间的博弈，实现资源利用率和节点参与率的双重优化，显著优于静态分配、阈值触发和固定激励等传统方法，验证了激励机制在资源分配中的有效性。

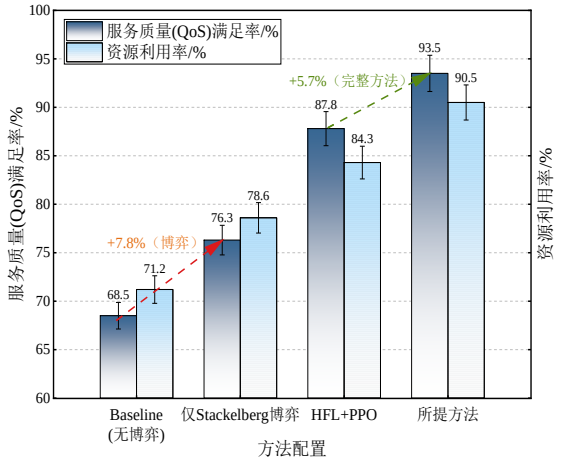


图 9　所提方法各组件性能

图 9 衡量了所提方法各组件对系统性能的贡献。由图可知，各组件都使得 QoS 满足率有所提升，而完整方法使 QoS 满足率提升约 5.7%，效用提升约 15%，说明 HFL 的预测能力、博弈激励带来的参与稳定性以及 MT-PPO 对三类业务需求的差异化调度共同增强了系统性能。综上，本实验验证了本文提出的方法能够有效解决 5G-R/FRMCS 场景中高动态性、强异构性、资源受限的核心挑战，使得系统保持稳定、高效且具备强鲁棒性的资源调度能力。

3.2.2　综合性能对比实验

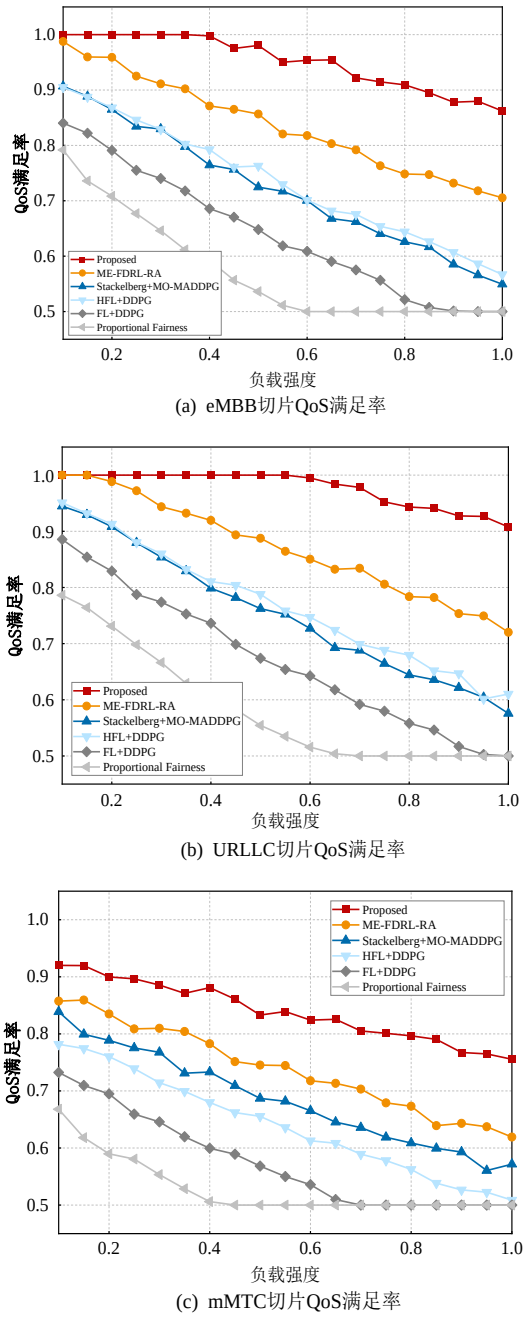
为了进一步验证本文所提方法的优势，本研究将选取以下四种具有代表性的算法进行对比。其中，移动感知与能效优化的联邦强化学习辅助资源分配（mobility-aware and energy-efficient federated deep reinforcement learning-assisted resource allocation , ME-FDRL-RA）算法[10]采用了多智能体强化学习进行资源分配，但未引入 Stackelberg 博弈激励机制，HFL+深度确定性策略梯度（deep determin-

istic policy gradient，DDPG）算法[15]具备 HFL 架构，且其决策核心为 DDPG，在应对 5G-R/FRMCS 高速移动以及多异构业务共存的场景下存在局限性，Stackelberg+多目标多智能体深度确定性策略梯度（multi-objective multi-agent deep deterministic policy gradient，MO-MADDPG）算法[29]虽然集成了博弈激励机制与多智能体架构，但缺乏 HFL 层级聚合机制以应对 Non-IID 数据，联邦学习（federated learning，FL）+DDPG[30]算法采用传统单层架构的联邦学习与 DDPG 相结合的方式。作为进一步对照，比例分配（proportional fairness）算法作为一种经典的静态启发式策略，不具备任何在线学习与动态调整能力。该综合性能对比实验将全面评估各算法在 QoS 满足率、系统吞吐量、收敛速度及 URLLC 实验保障等关键指标上的性能。

图 10（a）-（c）分别展示了在不同业务负载强度下，三类网络切片 QoS 满足率的变化趋势，其中“负载强度”量化为 5G-R/FRMCS 专网场景下三类切片业务的总业务请求强度。整体来看，所提方法在三类业务场景中均保持更优且更稳定的 QoS 保障能力，其优势在高负载条件下尤为显著。

对于 eMBB 业务（图 10（a）），随着负载强度增加，所提方法的 QoS 满足率始终保持在较高水平，平均约 94%，主要得益于 Stackelberg 博弈激励机制对边缘节点参与行为的有效引导。而 Proportional Fairness 方法由于采用静态资源分配策略，难以动态协调异构网络切片资源调度，性能下降明显。相比未引入激励机制的 ME-FDRL-RA 算法，其在负载强度为 0.5 时 QoS 满足率下降至 83% 左右，而 HFL+DDPG 与 Stackelberg+MO-MADDPG 方法均在高负载下性能退化，表明对业务异构型的适应能力不足。

对于 URLLC 业务（图 10（b）），所提方法在低负载条件下即可实现 98% 以上的 QoS 满足率，通过 HFL 聚合及 MT-PPO 差异化优化，使其在负载升高时下降幅度较小，在满负载下仍保持约 0.9。而 Proportional Fairness 算法与 ME-FDRL-RA 算法的 QoS 满足率在高负载下快速下降，HFL+DDPG 算法虽具备一定稳定性，但在高负载时 QoS 满足率仍低于所提方法约 23%。该结果表明，在 URLLC 对时延与可靠性高度敏感的场景下，静态资源分配策略难以在高负载条件下维持稳定 QoS。



(a) eMBB切片QoS满足率

(b) URLLC切片QoS满足率

(c) mMTC切片QoS满足率

图 10　不同负载强度下的 QoS 满足率对比分析

对于 mMTC 业务（图 10（c））。所提方法在高负载区间仍表现出较强的稳定性，当负载接近 1.0 时，仍可达到 0.75 左右，而 ME-FDRL-RA 算法的 QoS 满足率约为 0.61，Stackelberg+MO-MADDPG 算法与 HFL+DDPG 算法接近于 0.58 与 0.50，相比之下，FL+DDPG 与 Proportional Fairness 算法在中高负载强度下均退化至 0.50 左右。进一步对比可得，ME-FDRL-RA 算法在负载升高过程中由于缺乏 HFL 机制，QoS 满足率持续下降，表明其对大

规模终端并发接入的适应能力有限，Stackelberg+MO-MADDPG 算法由于其仅依赖博弈激励在高负载下性能退化更为明显，HFL+DDPG 算法在负载增大后下降幅度较大，反映出单任务强化学习难以有效应对 mMTC 业务的高连接密度特性。

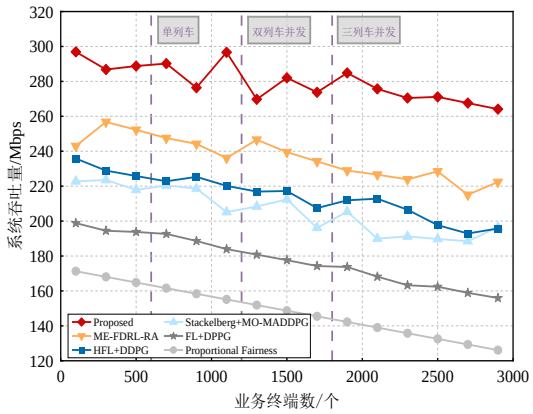


图 11　系统吞吐量对比实验

图 11 展示了不同方法随着业务终端数的增加系统吞吐量的变化趋势。本文所提方法平均吞吐量达 275.8Mbps，在单列车场景下保持 290Mbps，双列车并发时达 296Mbps 峰值，三列车并发时虽有所下降但仍稳定在 265-275Mbps 区间。ME-FDRL-RA 算法平均吞吐量为 238.6Mbps，相比本文所提方法下降 13.5%，并且随着终端数增加至 1800 以上时吞吐量衰减至 215Mbps，主要因缺少 Stackelberg 博弈激励导致边缘节点在高负载下资源协同效率不足；HFL+DDPG 与 Stackelberg+MO-MADDPG 算法性能相近，而前者受益于 HFL 分层聚合，在 600-1500 业务终端数下系统吞吐量变化更为稳定；FL+DDPG 算法平均吞吐量仅为 178.4Mbps 左右，随着业务终端数的增长，吞吐量快速衰减，验证了 HFL 的优势；Proportional Fairness 算法采用静态资源分配，平均吞吐量仅为约 147.2Mbps，充分说明动态智能调度的优势。综上，仿真结果表明所提方法具有更强的资源调度稳定性和负载均衡能力。

图 12 展示了六种算法的平均奖励值随通信轮次的变化曲线，所提方法在收敛速度与最终性能方面表现最优，在第 50 轮左右完成收敛，最终稳定在 91.35 左右，超过目标奖励值约 6.35。ME-FDRL-RA 方法表现出较好的学习稳定性，然而最终性能较低且波动较为明显，Stackelberg+MO-MADDPG 方法虽然具备博弈激励机制，但缺乏

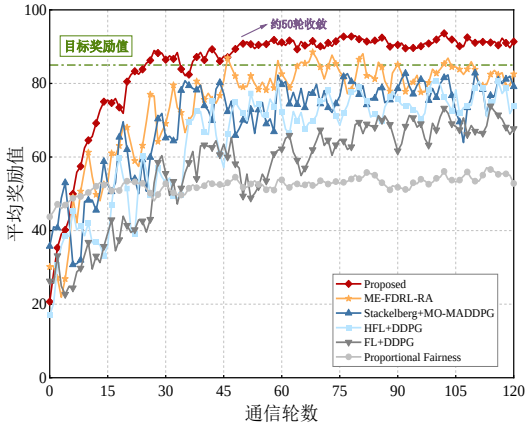


图 12　平均奖励收敛对比实验

HFL 分层聚合导致收敛波动较大，FL+DDPG 方法策略更新相比所提方法曲线收敛较慢且无法达到目标奖励值，Proportional Fair 算法不具备在线学习能力，其奖励值收敛至 55 左右的较低水平，验证了引入智能决策机制的必要性。仿真结果表明，所提方法通过 HFL、博弈机制以及 MT-PPO 能更高效地实现资源调度算法性能的显著提升。

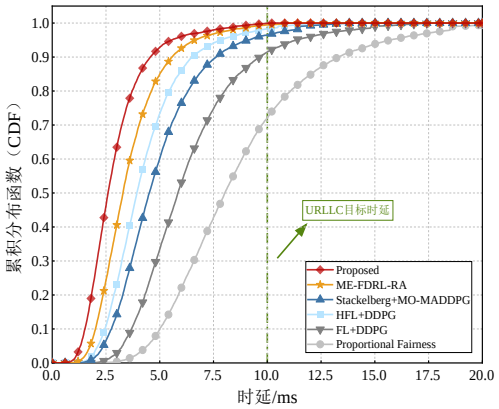


图 13　URLLC 时延 CDF 曲线

为验证所提资源分配算法对于极端服务质量要求的保障能力，图 13 展示了对超高可靠低时延通信（URLLC）切片采用累积分布函数（CDF）进行评估，CDF 曲线能够直观呈现算法在不同时延下的可靠性，即对于给定的时延阈值，URLLC 业务能够被成功满足的比例。

由图 13 可得，本文所提方法平均时延 2.91ms，99.91% 的数据满足 10ms 的时延要求，ME-FDRL-RA 方法平均时延 3.61ms，URLLC 达标率达 99.45%，相比本文所提方法相差 0.46%，Stackelberg+MO-MADDPG 算法因多智能体通信开销导致

平均时延 4.93ms，约 3.5% 的数据超出 URLLC 要求，FL+DDPG 性能显著低于 HFL+DDPG，证明了 HFL 架构对降低时延波动的作用，Proportional Fairness 算法达标率仅 72.02%。综合平均时延和达标率，仿真结果充分验证了所提方法于 URLLC 切片业务的优越性。

4　结束语

本文面向 5G-R/FRMCS 高动态场景中业务异构、节点高速移动与激励不足导致的网络切片资源调度问题，构建了一种融合 HFL、Stackelberg 博弈与 MT-PPO 的协同智能调度框架。研究结果表明，联合建模可有效提升资源调度在动态环境中的前瞻性与自适应能力，面向 Non-IID 数据的分层协同学习与负载预测有助于提高状态感知精度与训练稳定性，差异化激励机制能够改善分布式协同中节点参与不均的问题，而将预测信息与激励因子融入策略学习过程，使得调度策略更适配异构业务 QoS 需求。仿真表明该协同机制在系统吞吐量、资源利用率及异构切片 QoS 满足率上均具优势，并收敛迅速，验证了“预测-激励-决策”协同优化思路在 5G-R/FRMCS 网络切片资源调度中的适用性。

参考文献：

[1] 孙君,雳振宇.基于差异性隔离和复用的网络切片无线资源分配方案[J].通信学报,2025,46(03):109-121.
Sun J, Ai Z Y. Wireless resource allocation scheme for network slicing basedon differentiated isolation and multiplexing[J]. Journal on Communications, 2025,46(03):109-121.

[2] Le H N, Duong D T, Phuong N T, et al. Dynamic resource allocation for RIS-assisted network slicing in millimeter-wave high-speed rail communications[J]. Physical Communication, 2026, 74: 102955.

[3] 刘淼,林婉茹,王琴,等.车联网联邦学习的数据异质性问题及基于个性化的解决方法综述[J].通信学报,2024,45(10):207-224.
Liu M, Lin W R, Wang Q, et al. Survey on data heterogeneity problems and personalization based solutions of federated learning in Internet of vehicles[J]. Journal on Communications, 2024,45(10):207-224.

[4] Cai Y, Cheng P, Chen Z, et al. Deep reinforcement learning for online resource allocation in network slicing[J]. IEEE Transactions on Mobile Computing, 2023, 23(6): 7099-7116.

[5] 赵佳楠,胡晓辉,杜欣欣.基于近端策略优化算法的车载边缘计算网络频谱资源分配[J].数据与计算发展前沿,2022,4(04):142-155.
Zhao J N, Hu X H, Du X X. Spectrum Resource Allocation of Vehicle Edge Computing Network Based on Proximal Policy Optimization Algorithm[J]. Frontiers of Data & Computing, 2022,4(04):142-155.

[6] Lu B, Wu Y, Qian L, et al. Multi-agent DRL-based two-timescale resource allocation for network slicing in V2X communications[J]. IEEE Transactions on Network and Service Management, 2024, 21(6): 6744-6758.

[7] Gao X, Zhao J, Zhang Q, et al. Optimization of resource allocation strategy for high-speed railway based on deep reinforcement learning[J]. Physical Communication, 2024, 66: 102455.

[8] Tang T, Chai S, Wu W, et al. A multi-task deep reinforcement learning approach to real-time railway train rescheduling[J]. Transportation Research Part E: Logistics and Transportation Review, 2025, 194: 103900.

[9] Baltes J, Akbar I, Saeedvand S. Cooperative dual-actor proximal policy optimization algorithm for multi-robot complex control task[J]. Advanced Engineering Informatics, 2025, 63: 102960.

[10] Azimi Y, Yousefi S, Kalbkhani H, et al. Mobility aware and energy-efficient federated deep reinforcement learning assisted resource allocation for 5G-RAN slicing[J]. Computer Communications, 2024, 217: 166-182.

[11] Wu B, Fang F, Wang X, et al. Client selection and cost-efficient joint optimization for NOMA-enabled hierarchical federated learning[J]. IEEE Transactions on Wireless Communications, 2024, 23(10): 14289-14303.

[12] Pervej M F, Jin R, Dai H. Hierarchical federated learning in wireless networks: Pruning tackles bandwidth scarcity and system heterogeneity [J]. IEEE Transactions on Wireless Communications, 2024, 23(9): 11417-11432.

[13] Yan X, Zuo S, Fan R, et al. Sequential federated learning in hierarchical architecture on non-IID datasets[J]. IEEE Transactions on Mobile Computing, 2025, 24(10): 11110-11124.

[14] 夏玮玮,王博业,夏雅星,等.基于多时间尺度协同的无蜂窝 RAN 切片资源分配算法[J].通信学报,2025,46(07):60-77.
Xia W W, Wang B Y, Xia Y X, et al. Cellular-free RAN slicing resource allocation algorithm based on multi-timescale collaboration[J]. Journal on Communications, 2025,46(07):60-77.

[15] Sagar A. S. M. S.; Haider A.; Kim H. S. A hierarchical adaptive federated reinforcement learning for efficient resource allocation and task scheduling in hierarchical IoT network. Computer Communications, 2025, 229: 107969.

[16] Chen Y, Zhou H, Li T, et al. Multifactor incentive mechanism for federated learning in IoT: A Stackelberg game approach[J]. IEEE Internet of Things Journal, 2023, 10(24): 21595-21606.

[17] Wang S, Xia W, Zhao H, et al. Stackelberg Game-Based Hierarchical Incentive Mechanism for Clustered Vehicular Federated Learning[J]. IEEE Transactions on Communications, 2025, 73(10): 9071-9086.

[18] Datar M, Altman E, Cadre H L. Strategic resource pricing and allocation in a 5G network slicing Stackelberg game[J]. IEEE Transactions on Network and Service Management, 2023, 20(1): 502-520.

[19] Ren B, Yang P, Du M, et al. Energy efficient heterogeneous federated learning over mobile devices: a deep reinforcement learning based stackelberg game approach[J]. IEEE Transactions on Network Science and Engineering, 2026, 13: 3534-3550.

[20] 董从堂，丁建文，孙斌，等. 5G-R 网络切片部署及运维管理技术方案研究 [J]. 铁道通信信号，2026, 62 (3): 1-7.
Dong C T, Ding J W, Sun B, et al. Research on Technical Solutions for Deployment and Operation Management of 5G-R Network Slicing[J]. Railway Signalling & Communication, 2026, 62 (3): 1-7.

[21] Ebrahimi S, Bouali F, Haas O C L. Resource management from single-domain 5g to end-to-end 6g network slicing: A survey[J]. IEEE Com-

munications Surveys & Tutorials, 2024, 26(4): 2836-2866.

[22] Ruihan Wen, Gang Feng, Chengjie Li, et al. AI-RAN resource configuration for non-collaborative cross-domain slicing[J]. Journal of Information and Intelligence, 2026, 4(1): 38-53.

[23] Vidhya P, Subashini K, Sathishkannan R, et al. Dynamic network slicing based resource management and service aware Virtual Network Function (VNF) migration in 5G networks[J]. Computer Networks, 2025, 259: 111064.

[24] 景小荣,彭喆,陈前斌.基于深度强化学习的分层协同干扰资源分配方案[J].中国科学:信息科学,2025,55(09):2371-2396.
Jing X R, Peng Z, Chen Q B. Hierarchical cooperative jamming resource allocation scheme based on deep reinforcement learning[J]. Sci Sin Inform, 2025,55(09):2371-2396.

[25] Zhang N, Liu B, Zhang J. Dual Resource Scheduling Method of Production Equipment and Rail-Guided Vehicles Based on Proximal Policy Optimization Algorithm[J]. Technologies, 2025, 13(12): 573.

[26] Jin Y, Song X, Slabaugh G, et al. Partial advantage estimator for proximal policy optimization[J]. IEEE Transactions on Games, 2024, 17(1): 158-166.

[27] OFCOM. Statement and further Consultation: Future authorisation of the 1900 – 1920 MHz band [R/OL]. (2025 – 10 – 24) [2026 – 01 – 16].

[28] Zhang C, Wu C, Lin M, et al. Proximal policy optimization for efficient D2D-assisted computation offloading and resource allocation in multi-access edge computing[J]. Future Internet, 2024, 16(1): 19.

[29] Abishu H N, Seid A M, Erbad A, et al. A multi-agent DRL-based framework for optimal resource allocation and twin migration in the multi-tier vehicular metaverse[J]. IEEE Transactions on Vehicular Technology, 2025: 1-16.

[30] 陈晓,仇洪冰,李燕龙.边缘辅助的自适应稀疏联邦学习优化算法[J].电子与信息学报,2025,47(03):645-656.
Chen X, Qiu H B, Li Y L. Edge-assisted Adaptive Sparse Federated Learning Optimization Algorithm[J]. Journal of Electronics & Information Technology, 2025, 47(3): 645-656.

高云波（1980— ），男，陕西眉县人，兰州交通大学教授、硕士生导师，主要研究方向为轨道交通无线通信、可重构智能表面传感和通信、无线网络资源管理等。

张琦芸（2002— ），女，甘肃兰州人，兰州交通大学硕士，主要研究方向为铁路通信、5G-R 网络切片资源调度及其智能优化算法等。