基于深度强化学习的Wi-Fi多接入点协同调度

戴维·努涅斯*、弗朗西斯科·威廉米*、马克西米利安·沃伊纳尔♯、卡塔兹娜·科塞克-绍特♯、希蒙·绍特♯和鲍里斯·贝拉尔塔*，

*Wireless Networking, Universitat Pompeu Fabra, Barcelona, Spain
♯AGH University of Krakow, Poland

Abstract多接入点协同（MAPC）是IEEE 802.11bn标准的核心特性，对未来Wi-Fi网络发展具有重要影响。该技术通过跨多个接入点（AP）联合调度决策，可显著提升密集Wi-Fi部署场景中的吞吐量、时延和可靠性。然而在重叠基本服务集（OBSS）环境下，面对多样化业务流与干扰条件，实现高效调度策略仍具挑战性。本文提出一种网络级最坏时延最小化方法：将MAPC调度建模为序列决策问题，并采用深度强化学习（DRL）机制来优化OBSS部署中的最坏时延表现。具体而言，我们在兼容802.11bn标准的Gymnasium环境中，使用近端策略优化（PPO）算法训练DRL智能体。该环境提供队列状态、时延指标及信道条件等观测数据，使智能体能通过空间复用（SR）组调度多组AP-站点对同步传输。仿真结果表明，所提方案在各类网络负载和业务模式下均优于现有启发式策略，训练所得的机器学习（ML）模型始终保持着更优的99%分位时延表现，较最佳基线方案提升幅度高达30%。

Index Terms—协调空间复用，IEEE 802.11bn，机器学习，多接入点协调，强化学习，调度，Wi-Fi 8。

一、引言

无线流量的快速增长、对延迟敏感应用的兴起以及用户设备密度的不断增加，正对Wi-Fi网络提出前所未有的要求。传统的基于竞争的频道接入机制在流量模式多样且同频干扰严重的密集部署场景中无法有效扩展[1]。这些局限性导致了服务质量(QoS)低下、延迟大幅增加以及频谱利用效率不高。未来的Wi-Fi技术必须采用更具确定性的频道接入方式和延迟敏感调度机制，才能满足增强现实、工业自动化和高密度公共场所等新兴应用场景的需求[2]。

在此背景下，IEEE 802.11bn提出的多接入点协调（MAPC）标志着频谱共享从无协调到协调的根本性架构转变。通过实现跨多个接入点（AP）的集中式调度决策，MAPC促进了时间、频率和空间资源的联合优化。协调式空间复用协作（Co-SR）作为一种特定的多接入点协作（MAPC）技术应运而生，它通过协调网络中的空间复用（SR）机会，允许多个接入点-站点（STA）[1]对实现并发传输[3]。与时分或频分协调技术[4]不同，Co-SR旨在通过允许满足干扰约束的重叠传输来最大化频谱效率。在各类MAPC策略中，Co-SR在复杂性与性能之间实现了卓越的平衡，使其特别适合高密度场景下对延迟敏感的应用，同时保持较低的实施成本。因此，我们的分析聚焦于采用Co-SR作为底层协调机制的MAPC网络。然而，基于Co-SR的MAPC网络效能不仅取决于并行传输的可行性与质量，更依赖于在不同流量和信道条件下做出及时、明智的调度决策的能力。设计高效的MAPC调度器仍是一项具有挑战性的任务。调度器需综合流量、信号质量和干扰等因素为各独立基本服务集（BSS）做出决策，由此产生的复杂关系难以通过启发式算法（如Co-SR中采用的最大数据包数（MNP）、最旧数据包（OP）和流量对齐追踪器（TAT）等调度策略[5]）来准确捕捉。

在本工作中，我们通过将MAPC调度问题建模为序列决策任务，并提出一种能够从实时网络观测中学习调度策略的模型来解决该问题。所提出的解决方案包含一个深度强化学习（DRL）智能体，该智能体通过标准化的Gymnasium接口与环境交互。[2] 我们的主要目标在于证明，通过学习到的策略能够有效适应多样化的拓扑结构、负载及流量模式，从而持续超越基于启发式的调度器（如文献[5]先前提出的方案）。实验结果表明，与性能最佳的非机器学习基线相比，所提模型在最坏情况下可降低高达30%的延迟。本文的主要贡献总结如下：

[1]In the IEEE 802.11 standard, client devices are referred to as non-AP STAs, which we refer to as STAs in the remainder of this paper.
[2]https://gymnasium.farama.org/

- 我们将802.11bn MAPC调度问题表述为一个序列决策优化问题。
- 我们提出了一种深度强化学习（DRL）解决方案，旨在基于实时指标（如整个重叠基本服务集（OBSS）的排队延迟）来学习动态MAPC调度策略。
- 我们开发了一种新颖的Gymnasium兼容仿真环境，该环境将AI驱动的决策与符合802.11bn标准的详细模拟器相结合，能够在多样化的流量和信道条件下训练和评估基于深度强化学习(DRL)的MAPC调度器。
- 我们展示了所提出的基于DRL的调度解决方案的性能，并在全面的模拟场景中将其与MNP、OP和TAT等启发式基线进行比较。

本文的其余部分结构如下。第二节深入探讨了MAPC和Co-SR领域的最新解决方案。第三节讨论了IEEE 802.11bn MAPC的配置，以及我们提出的从多个BSS中创建SR兼容设备组的方案。随后，第四节详细介绍了所提出的机器学习（ML）框架和基于深度强化学习（DRL）的调度方案。第五节提供了评估设置和仿真细节，而第六节则展示了仿真结果。最后，第七节对全文进行了总结。

二、相关工作

自IEEE 802.11ax标准引入以来，空间复用(SR)技术获得了广泛关注，许多研究尝试通过采用机器学习(ML)技术来提升其性能。例如文献[6]–[9]中采用多臂老虎机(MAB)方案驱动SR参数选择。由于简单高效的特性，MAB天然适合在线决策场景，因此与Wi-Fi的MAC层优化需求高度契合。

基于学习的方法也被探索用于应对MAPC网络固有的复杂性和组合动作空间。文献[10]中，作者展示了基于强化学习（RL）解决方案的潜力，包括分散式和协调式方法，用于驱动密集Wi-Fi部署中SR的优化。随后，作者提出了一种多智能体多臂老虎机（MA-MAB）方法，基于协调（共享）奖励策略，联合配置多个共存AP的SR参数——特别是分组检测阈值和发射功率[11]。他们的结果显示，与无协调基线相比，在吞吐量和公平性方面取得了显著提升。类似地，文献[12]引入了一种分层MAB框架，用于IEEE 802.11bn中的Co-SR组选择。该研究利用RL选择兼容的AP子集，发现基于上置信界（UCB）的老虎机算法能在异构拓扑中实现快速收敛和持续性能。最后，文献[13]的工作通过提出基于MAB的实用解决方案（包括平面和分层变体），解决了IEEE 802.11bn中的Co-SR调度问题。仿真和测试平台实验表明，分层MAB能将聚合吞吐量提升高达80%。在传统IEEE 802.11基础上，不减少每个STA分配的传输机会数量。

关于Wi-Fi调度这一本文的核心议题，近期研究也聚焦于机器学习方向。文献[14]提出并实现了一种基于强化学习的调度框架，通过商用设备优化Wi-Fi网络的应用层服务质量(QoS)。该方法利用历史QoS反馈训练的Q网络动态调整竞争窗口大小和吞吐量限制，在真实测试环境中较增强型分布式信道接入(EDCA)取得显著提升。文献[15]则采用深度学习技术，仅依据节点位置进行传输调度，绕过了信道估计环节并有效应对密集干扰场景。与本文紧密相关的文献[16]提出基于深度强化学习的解决方案，用于优化802.11bn网络中的空间复用(SR)机制。但[16]的研究思路与本文存在本质差异——其通过深度强化学习决策执行信道接入和链路适配，而本文则聚焦于作为深度强化学习多接入点协调(MAPC)调度器的一部分，选择最优SR组进行传输。该问题在文献[5]中已有探讨，但仅评估了非人工智能(AI)调度器。本文通过论证机器学习解决方案的潜力推动领域发展，这类方案有望有效应对多接入点场景下日益增长的复杂度挑战。

三、Wi-Fi 8中的多AP协作：概述与系统模型

A. Main Operation

基于对Wi-Fi 8统一MAPC框架的持续讨论[17]，图1展示了一个在四个AP组中实现协同短距离传输(Co-SR)的可行机制，这些AP可能向各自对应的STA进行传输。根据IEEE 802.11bn中的MAPC规范，我们仅考虑下行链路传输。该图说明了协调AP在两个传输机会(TXOP)期间的交互过程。所有AP均处于相互通信范围内，并共享同一无线信道，通过分布式协调功能(DCF)竞争信道接入——该机制包括选择随机退避值，在感知信道空闲时递减该值，直至获得信道传输权限。

为了促进接入点（AP）间的协调，一旦退避计数器归零，赢得竞争的AP将承担共享AP的角色——在TXOP #1中为AP_1，在TXOP #2中为AP_4（图1）。共享AP负责通过发送MAPC初始控制帧（MAPC-ICF）来发起信道预留。该控制帧除其他功能外（如指示要使用的MAPC特性），此处还用于向MAPC协调范围内的邻近AP（称为共享AP）征询缓冲区状态信息。作为响应，这些共享AP会发送MAPC初始控制响应（MAPC-ICR）消息，其中包含其参与意向，并且为实现本文提出的机制，还包括缓冲数据包数量ϱ_i以及队首数据包的时间戳。

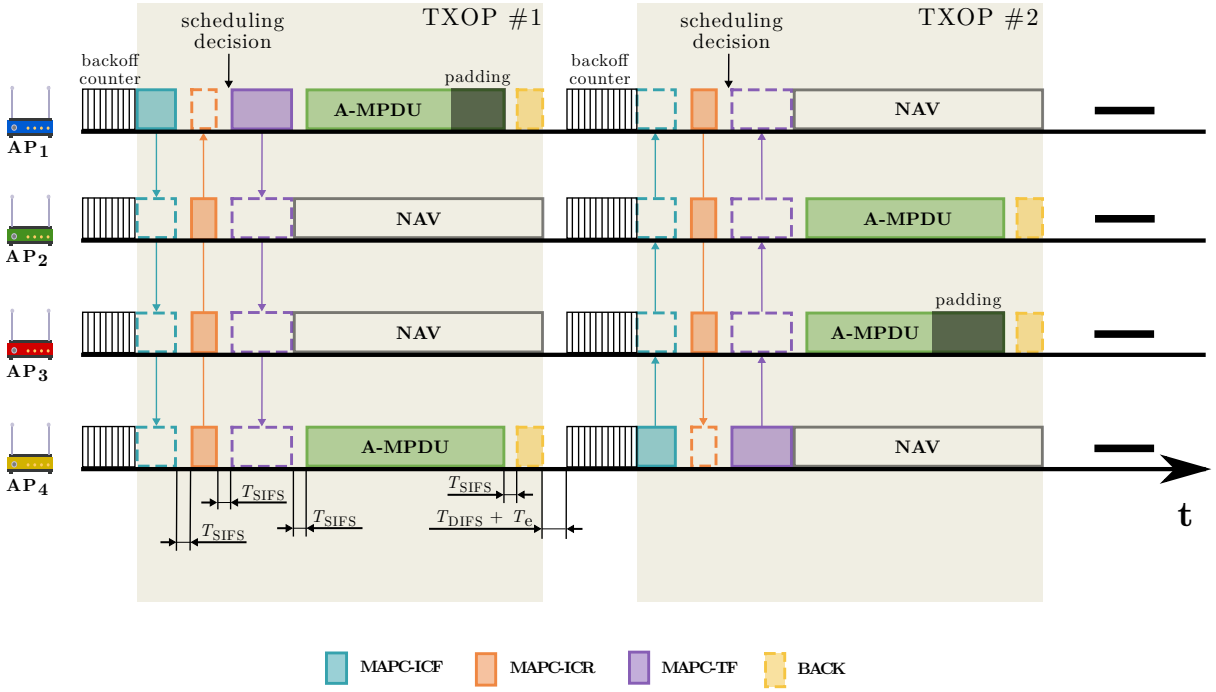


图1：提出的MAPC网络。

(HoL) 为每个关联的STA分配一个ε_i数据包。采用正交频分多址接入(OFDMA)技术，以实现MAPC-ICR帧的同步传输。

基于竞争的接入方式虽然实现了灵活的介质共享，但也可能引发冲突和延迟波动。当在等于$T_{\text{MAPC-ICF}} + T_{\text{SIFS}} + T_{\text{MAPC-ICR}} + T_{\text{DIFS}} + T_{\text{e}}$的超时时间内未收到任何MAPC-ICR响应时，即判定发生冲突——其中T_{SIFS}和T_{DIFS}分别表示短帧间间隔(SIFS)和DCF帧间间隔(DIFS)，T_{e}代表退避时隙时间。这种情况表明OBSS内有两台或更多设备选择了相同的退避值，此时每台受影响的设备会增大其竞争窗口(CW)，并在更新后的CW范围内重新选取随机退避值。随后，系统将使用新的退避值重新启动竞争流程。

另一方面，若信道接入尝试成功，共享接入点将解析接收到的MAPC-ICR消息，并根据既定调度策略选择在当前传输机会（TXOP）期间下行链路中服务的STA群组。该决策在每个TXOP中独立作出。

一旦选定了一组潜在的SR传输（包括多个AP及其对应的STA），共享AP会广播一个MAPC触发帧（MAPC-TF），该帧标识出选定的STA（根据调度器判断，若单个STA能带来最佳效果，则可能仅包含一个STA）并指定其传输参数，包括带宽、调制与编码方案（MCS）以及持续时间。

在TXOP之后，被选中的设备使用聚合MAC协议数据单元（A-MPDU）发送其缓冲的数据，而其余设备则设置其网络分配向量（NAVs）以推迟接入，直到信道预期再次空闲。需要注意的是，在所提出的框架中，信道接入遵循传统的竞争机制，即DCF。然而，被选中进行传输的设备集合不一定包括共享AP，如图1中的TXOP #2所示。最后，每个STA在成功接收后返回其对应的块确认（BACK）。任何未成功接收的帧将保留在缓冲区中，并在STA的下一个预定TXOP中重新尝试发送。

B. SR Group Creation

为了实现多AP网络中的协作空间复用(Co-SR)，我们预先计算了一组可行的SR分组，方法如文献[5]所示。[3]每个分组包含可被调度进行并发传输的AP-STA对子集，同时确保传输质量。通过估算组内n每个STAi的传输速率来评估分组的可行性，由此得出

$$R_{i,n}^{\text{CoSR}} = \frac{N_{\text{bps}} R_{\text{c}} N_{\text{sc}} N_{\text{ss}}}{T_{\text{OFDM}} + T_{\text{GI}}}, \tag{1}$$

3其他空间复用组创建方法，如[12]、[13]中提出的方案，同样可行。但我们采用[5]提出的方法，以确保基于机器学习的模型与启发式算法采用完全相同的组选择流程。

其中N_{sc}、N_{ss}、T_{OFDM}和T_{GI}分别表示数据子载波的数量、使用的空间流数量、正交频分复用(OFDM)符号的持续时间以及保护间隔的持续时间。数值N_{bps}和R_{c}对应于为站点i所选调制编码策略(MCS)定义的每符号比特数和编码率，该策略根据预期的信号与干扰加噪声比(SINR)在每次传输时动态选择。

设$\mathcal{M}_n$为包含组n中STA的子集。为了评估并发传输的可行性，将$\mathcal{M}_n$中每个STAi的性能与其单传输情况（记为R_i^{ST}）进行比较，该情况对应于STA单独服务（不受BSS间干扰影响）的场景。若满足以下条件，则接纳该组：

$$|\mathcal{M}_n|\frac{R_{i,n}^{\mathrm{CoSR}}}{R_i^{\mathrm{ST}}} \geq 1, \quad \forall i \in \mathcal{M}_n. \tag{2}$$

请注意，(2)中的等式定义了STA被纳入SR组n的最低可接受传输速率。这一条件确保在长期持续高流量负载下，每个STA的平均吞吐量至少与单次传输所达到的水平相当。该表达式捕捉了因组内并发传输导致的干扰增加与同一TXOP内服务多个STA所获效率之间的权衡。然而，仅满足吞吐量要求并不能保证延迟需求得到满足。因此，需要智能调度机制来有效管理缓冲流量并最小化延迟。

四、基于DRL的MAPC调度方案

本节概述了将深度强化学习（DRL）智能体集成至Wi-Fi MAPC调度操作的整体解决方案。如图2所示，我们定义了一个包含三个核心模块的系统：*i)*网络环境、*ii)*符合Gymnasium标准的接口，以及*iii)*学习智能体。接下来我们将逐一详述各模块定义。

A. Environment

该环境集成了一个用Python实现的802.11模拟器[4]，用于表征特定的802.11部署场景，并管理诸如流量到达、信道条件和传输事件等网络操作。在集成802.11模拟器与深度强化学习（DRL）智能体时，我们设定每个训练周期模拟一个包含多个接入点（AP）及其关联站点（STA）的特定配置部署（包括随机流量生成实例）。模拟过程按步骤推进，每一步对应一个传输机会（TXOP），如图1所示，其中可调度来自不同AP的多路传输。关于场景配置、信道与流量模型以及调度策略的更多细节将在第五节详述。

[4]该模拟器可在https://github.com/dncuadrado/11bn-py-Simulator获取。

B. Gymnasium

为了实现基于学习的调度器的训练与评估，我们将第四-A节中的802.11模拟器集成至Gymnasium应用编程接口（API）——一个广泛应用于强化学习环境的标准化接口[18]。这一抽象层使得智能体与网络环境之间的交互遵循观察、行动与奖励的固定循环，同时将学习逻辑与模拟器内部实现解耦。

每一轮次开始时，`reset()`函数会以全新的部署配置初始化环境。返回给智能体的初始观测反映了当前网络状态，包括各STA（站点）的特征，如队列大小、延迟指标以及信道质量指示器{*}。

作为体育馆与环境交互的一部分，`step()`函数被体育馆用于发送代理选定的动作（见第IV-C节），该动作对应一个有效SR组的索引，并被模拟器用于接下来的TXOP中。环境在执行所提议的动作后，会更新队列状态和延迟指标，并返回下一个观察结果、奖励信号以及情节终止标志（terminated和truncated）。这一反馈机制使得代理能够逐步学习哪些调度决策是有效的。

C. DRL-based MAPC scheduling agent

在MAPC网络中高效调度传输的问题可建模为一个马尔可夫决策过程(MDP)[19]，其中智能体通过与环境的交互学习最优策略。作为解决方案的一部分，我们采用近端策略优化(PPO)算法对智能体进行训练，该算法基于演员-评论家框架实现[20]。该方法维护两个独立的神经网络：演员网络将当前观测s映射为动作a的概率分布，评论家网络则评估从观测s出发的预期回报。两个网络同步优化，在确保学习动态稳定的同时提升策略质量。该深度强化学习框架的核心组件包括观测空间、动作空间和奖励函数，具体阐述如下。

1) Observation Space: 在每个时间戳为t的决策步骤中，智能体会接收到一个观测向量s，该向量包含了网络中所有N个STA的归一化特征。最终的观测向量$s = [\bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\varrho}}, \bar{\boldsymbol{h}}] \in [0,1]^{3N}$由以下信息构成：

- 延迟向量$\bar{\boldsymbol{\delta}} = [\delta_1, \ldots, \delta_N]/T_{\mathrm{sim}}$，其中$\delta_i = (t - \varepsilon_i)$、$i = 1, ..., N$表示自HoL数据包到达其对应AP队列以来，各STA所经历的时间，该时间已通过事件持续时间T_{sim}进行了归一化处理。
- 队列大小向量$\bar{\boldsymbol{\varrho}} = [\varrho_1, \ldots, \varrho_N]/\varrho_{\max}$，表示每个STA传输队列的当前占用情况，$\varrho_i$ (以帧数)为单位，并经过最大缓冲区长度归一化处理，$\varrho_{\max}$。

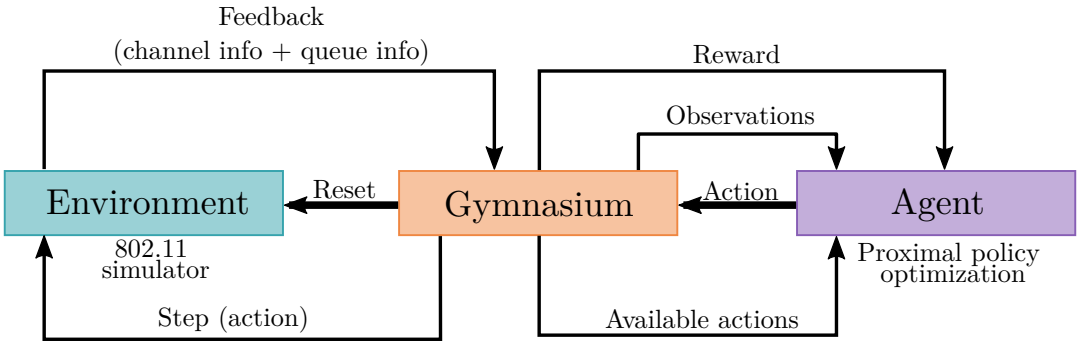


图2：将机器学习集成到802.11网络中的提议框架。

- 信道系数向量$\bar{\boldsymbol{h}} = [h_1, \ldots, h_N]/h_{\max}$，其中包含每STA链路质量，即在视距(LoS)条件下归一化至1米距离路径损耗的信道系数$h_{\max}$。

所考虑的观测数据通过一个共享的特征提取器处理，该提取器由两层多层感知机（MLP）构成，其中每个隐藏层包含64个单元，并采用双曲正切（Tanh）激活函数。生成的特征向量随后作为输入传递给两个输出头[21]：

- 演员计算动作逻辑值，这些逻辑值通过分类分布转化为离散动作空间上的概率分布。
- 评论家输出一个标量值，用于估计从观测{v*}出发的预期累积奖励，通过减少梯度估计的方差来指导策略更新。

2) Action Space and Masking: 离散动作空间$\mathcal{A}$由$Z >$个有效动作构成，即有效的SR组。为了提高效率，每一步都会通过一个动态二进制掩码$m \in \{0,1\}^Z$来过滤掉不可行或无益的动作。[5]若满足以下任一条件，则动作a_z会被掩码（即$m_z = 0$）：

1) *Offline filtering:* 如果该组中至少有一个接收方不满足条件(2)，则在整个部署实现过程中该组将被预先丢弃。
2) *Online masking:* 在每次调度决策时，如果目标接收方各自的发射队列中没有待处理的数据包，则该组将被屏蔽。

掩码机制确保智能体在每一步仅从有效且高效的调度动作子集中进行选择，这通过引导智能体避开无效或低效动作，显著减少了探索空间并加速了收敛。

3) Reward Design: 为了引导智能体减少延迟并加速学习，我们构建了一个奖励函数，通过如下奖励塑形方法平衡长期反馈与即时信号：

$$r = r_{\text{sh}} + r_{\text{lg}},$$

5通过sb3_contrib库中的MaskablePPO实现：https://github.com/Stable-Baselines-Team/stable-baselines3-contrib。

其中r_{sh}是一个奖励塑造组件，它为选择与当前网络状态一致的动作提供即时反馈。相比之下，r_{lg}反映了先前决策对系统整体延迟的累积影响。尽管r_{lg}在每一步都会提供（即它不是稀疏的），但当最坏情况下的排队延迟增加时，其幅度会急剧减小。下文将对这两项进行讨论。

奖励塑形：该组件使得智能体即使在整体延迟仍然较高的情况下，也能识别出有效动作。当所有STA在前一决策点观察到的队首数据包在当前TXOP期间成功传输时，它会贡献一个正向奖励，从而产生：

$$r_{\text{sh}} = \min\{\boldsymbol{\varepsilon}\} - \min\{\boldsymbol{\varepsilon}'\},$$

其中$\boldsymbol{\varepsilon}' = [\varepsilon'_1, \ldots, \varepsilon'_N]$和$\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_N]$分别表示当前TXOP前后的队首延迟。若同一数据包仍停留在队列头部，表明其未被调度，则$r_{\text{sh}} = 0$。

长期奖励：这一项鼓励智能体在长期运行中保持较低的排队延迟：

$$r_{\text{lg}} = \min\left(\frac{\beta}{t - \min\{\boldsymbol{\varepsilon}\} + \nu}, 1\right),$$

其中$\beta > 0$为缩放参数，$\nu > 0$用于防止除零导致的不确定性。分母部分反映了系统范围内的最大排队延迟，该项会随延迟增加而递减。

在训练的早期阶段，智能体收集的知识可能仍不足以确定最佳行动，从而导致延迟逐渐累积。术语r_{sh}使智能体能够识别出带来显著改进的特定行动（如清除最早缓冲的数据包）。随着策略的成熟，r_{lg}变得更具信息量，鼓励可持续的低延迟操作。因此，综合奖励引导学习过程从初始探索转向有效的长期调度策略。

五、评估设置

本节详细阐述了我们针对IEEE 802.11bn协议的仿真设计及所采用的调度技术，相关仿真参数汇总于表I。

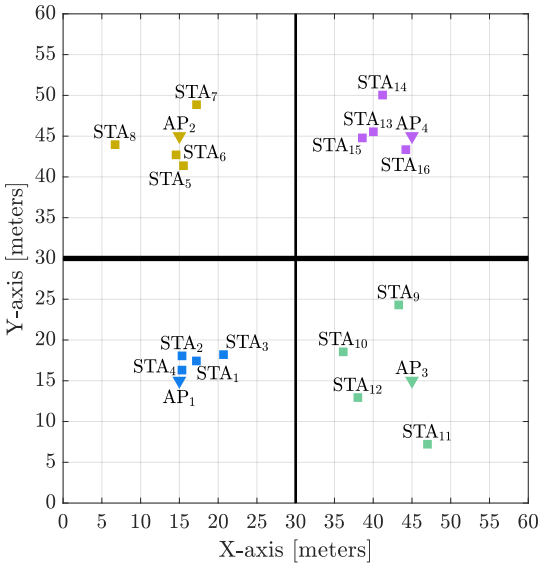


图3：包含4个AP和16个STA的示例部署。

表 I：仿真参数。

Parameter	Description	Value
d_{STA}	Distance between AP and STAs [meters]	[1, 10]
B_{p}	Break-point distance [meters]	10
B	Bandwidth [MHz]	80
f_{c}	Carrier frequency [GHz]	6
σ	Standard deviation of shadowing [dB]	5
N_{SC}	Number of data subcarriers	980
N_{SS}	Number of spatial streams	2
T_{OFDM}	OFDM symbol duration [μs]	12.8
T_{GI}	Guard interval duration [μs]	0.8
T_{max}	Max TXOP duration [ms]	5
$T_{\text{MAPC-ICF}}$	MAPC Initial Control Frame duration [μs]	74.4
$T_{\text{MAPC-ICR}}$	MAPC Initial Control Response duration [μs]	88
$T_{\text{MAPC-TF}}$	MAPC Trigger Frame duration [μs]	74.4
T_{BACK}	Block ACK [μs]	100
T_{SIFS}	Duration of a SIFS slot [μs]	16
T_{DIFS}	Duration of an DIFS slot [μs]:	34
T_{e}	Duration of an empty slot [μs]	9
CW_{min}	Min contention window	15
CW_{max}	Max contention window	1023
P_{max}	Max transmission power [mW]	200
ϱ_{max}	Max buffer size	10^4
h_{max}	Max channel gain	10^{-3}
W	Noise power [Watts]	3.2×10^{-13}
CCA	Clear channel assessment threshold [dBm]	-82
T_{ON}	Bursty traffic ON period mean duration [ms]	1
T_{OFF}	Bursty traffic OFF period mean duration [ms]	10
T_{sim}	Simulation duration [s]	5
L	Length of single data frame [bits]	12×10^3

A. Scenario

我们考虑一个802.11企业场景[22]，其中多个接入点（AP）共存于同一公共区域。如图3所示，为评估Co-SR性能，我们仅关注同一频段上J =4个AP组成的子集。这些AP部署在由墙壁分隔的不同办公室内，AP间距为30米。每个APj关联了N_j个站点（STA），这些STA随机分布在距其服务APd_{STA} ∈[1,10]米的范围内。为简化模型，我们假设所有AP关联的STA数量相同。

假设所有传输均采用固定的发射功率P_{max}，我们认为所有接入点（AP）处于同一通信范围内，因此它们依据分布式协调功能（DCF）竞争信道。一旦某个AP赢得信道访问权，便会选择一个空间复用（SR）组进行传输。每种部署场景中可用的空间兼容组集合均通过第III-B节所述策略预先计算得出。

我们仅考虑下行链路流量，且各站点（STA）的流量模式具有异质性。具体而言，每个STA的流量特征独立地从泊松模型或突发流量模型（第V-C节）中等概率抽取。此外，站点i、ω_i的负载量独立且均匀地从区间[$\omega_{\text{min}}, \omega_{\text{max}}$]中采样。结果中仅包含至少有一种调度机制能使99%分位延迟低于100毫秒的部署场景，其余情况被视为过载并排除在分析之外。这类部署通常由不利的站点布局或流量模式导致，不适合进行有效的延迟评估。每种评估案例中会报告被剔除部署场景的比例。

B. Channel Model

STA i感知到的SINR，记作ξ_i，取决于同时活跃的干扰发射机子集$\mathcal{K}$（如有），其计算公式为：

$$\xi_i = \frac{P_j h_{i,j}}{W + \sum_{k \in \mathcal{K}} P_k h_{i,k}}, \tag{3}$$

其中P_j表示接入点(AP)j为向站点(STA)i进行目标传输所消耗的功率，W代表噪声功率。项P_k表示干扰APk发射的功率，该功率构成了STAi所承受的干扰。STAi与APj之间的信道增益$h_{i,j}$定义为$h_{i,j} = 10^{-\frac{P_{\text{L}}(d_{i,j})}{10}}$，该函数与收发器间距$d_{i,j}$相关。同理，$h_{i,k}$表示基于距离$d_{i,k}$的AP$k$至STA$i$的信道增益。项$P_{\text{L}}(d_{i,j})$代表以分贝为单位的路径损耗，其建模采用TGax规范针对企业环境的设定[22]：

$$\begin{aligned} P_{\text{L}}(d_{i,j}) = & 40.05 + 20 \log_{10}\left(\frac{\min(d_{i,j}, B_{\text{p}}) f_{\text{c}}}{2.4}\right) \\ & + \mathbb{1}_{d_{i,j} > B_{\text{p}}} P' + 7 W_{\text{n}} + \mathcal{X}, \end{aligned} \tag{4}$$

其中$d_{i,j} \geq$1表示以米为单位的距离，f_{c}为载波频率（GHz），W_{n}为穿透的墙体数量，而P' =35 $\log_{10}(d_{i,j}/B_{\text{p}})$用于计算超出断点距离$B_{\text{p}}$的额外损耗。项$\mathcal{X}$模拟了对数正态阴影效应，其中$\ln(\mathcal{X}) \sim \mathcal{N}(0, \sigma)$。

我们采用IEEE 802.11be的MCS集（预计802.11bn不会有重大变化），并确定合适的MCS

为每次传输建立索引，利用预生成的包错误率（PER）与信噪比（SNR）曲线关系，这些曲线数据源自MATLAB仿真实验[23]。该映射关系反映了我们后续研究中的仿真条件，包括固定包大小L、带宽B、信道模型及天线配置。选择标准确保仅采用满足PER$<\leqslant 10^{-2}$的MCS值（基于估计值ξ_i）。因此，靠近接入点（AP）的站点（STA）会被分配更高的MCS值，而距离较远的站点则使用更稳健（更低）的MCS索引。

每次TXOP发送给站点i的聚合帧数量（A-MPDU大小）由U_i表示，其取决于所选MCS和TXOP持续时间。并发传输产生的干扰——即公式(3)中的$\sum_{k\in\mathcal{K}} P_k h_{i,k}$——会直接影响MCS选择，在将站点分组以实现SR兼容性时必须予以精确考量。

STA i成功解码的MAC协议数据单元(MPDU)数量被建模为一个二项随机变量，即$\mu_i \sim B(U_i, q)$，其中q =1−PER表示成功概率，U_i为传输帧数。STA i未能解码的帧将保留在其对应AP的缓冲区中，并在为STA i安排的下一个TXOP期间重新尝试[6]传输。

C. Traffic Model

考虑下行链路流量，每个接入点$_j$为其关联的N_j站点维护独立的逻辑传输队列，接收到的数据包将被缓冲。我们假设每个站点支持单一数据流，$\varrho_i(t)$表示在时间t站点i的队列深度。对于每个传输机会（TXOP），队列动态变化遵循以下规律

$$\varrho_i(t+T_\mathrm{s}) = \max\{\varrho_i(t) - \mu_i(T_\mathrm{s}), 0\} + \lambda_i(T_\mathrm{s}), \qquad (5)$$

其中$\mu_i(T_\mathrm{s})$表示在持续时间为T_s (的TXOP期间成功传输的帧数，该时间从期望值为$\bar{\mu}$)的随机变量中抽取；而$\lambda_i(T_\mathrm{s})$代表同一时间间隔内的到达帧数（从期望值为$\bar{\lambda}$的随机变量中抽取）。

以下两种流量生成模型用于评估系统性能，在所有情况下，STA i 的负载定义为 ω_i [Mb/s]。

1) *Poisson*: 数据包到达遵循泊松过程，导致每个STA的到达间隔时间呈指数分布。 2) *Bursty*: 该模型描述了间歇性流量，其中数据包突发发生在与OFF周期分离的ON周期内。突发到达被建模为一个两态（ON/OFF）马尔可夫过程，ON和OFF持续时间分别服从均值为T_ON和T_OFF的指数分布。在ON周期内，流量负载聚合了来自等效泊松过程的所有到达。

[6]We assume that retransmissions are not subject to any maximum attempt limit.

我们假设每个STA队列内的所有数据包具有相同的优先级，并采用先进先出（FIFO）调度策略来管理每个队列。

D. MAPC Scheduling Mechanisms

我们评估了五种不同的MAPC调度机制的性能，包括三种启发式方法和两种基于学习的智能体:

- *Maximum Number of Packets (MNP):* 该调度器选择当前TXOP中数据包数量最多的SR组。对于每个可行组，可调度数据包的数量是基于当前队列长度估算的，不考虑数据包的年龄。
- *Oldest Packet (OP):* OP策略优先考虑包含最高HoL数据包延迟的STA的SR组，其目标是通过在每个决策步骤中针对延迟最大的流来减少最坏情况下的延迟。这一机制隐含地促进了公平性，但当延迟最大的流无法并发调度时，可能导致TXOP容量的利用不足。
- *Traffic-Alignment Tracker (TAT):* 最初在[5]中提出的TAT算法，旨在通过选择能最小化延迟对齐成本（即组内所有参与者的队首包到达时间差{v*}）的SR组，来平衡最坏情况下的延迟降低与效率。因此，TAT倾向于在异构部署环境中保持稳定的延迟分布。
- *Machine Learning-General Agent (ML-G):* 这种基于DRL的调度器经过训练，能够泛化适用于广泛的流量和部署场景。在每一个决策步骤中，智能体观察所有STA的归一化队列长度、队头延迟以及信道系数，并通过PPO算法[20]训练的策略选择一个有效的SR组。智能体的目标是最小化所有数据包和STA间系统范围内的最坏情况延迟。通过动作掩码技术动态过滤掉不可行或非有益的SR组。
- *Machine Learning-Expert Agent (ML-E):* ML-E与ML-G共享相同的架构、观测空间及掩蔽流程，但其训练环境特殊：采用固定部署拓扑（对应图3）并涵盖多样化的流量场景。这种针对性训练使智能体能够针对单一环境的空间结构与干扰模式微调度策略，同时保持对不同流量条件的泛化能力。

六、性能评估

本节深入探讨了基于DRL的调度机制的性能评估。我们首先展示了ML调度器的训练过程，随后检验了它们在新（未见过的）部署环境中的表现。

A. Training

DRL智能体的训练采用PPO算法进行，通过来自{v*}的MaskablePPO实现。

sb3_contrib库[24]。该智能体在之前描述的自定义Gymnasium环境中进行训练，每个训练周期都包括一次Wi-Fi部署的模拟，其中涉及STA（站点）的布局及其对应的信道条件，以及生成新的流量模式，每个STA的负载为ω_i [Mb/s]，这些负载在整个周期内保持不变。鉴于信道模型仅考虑路径损耗（作为发射器与接收器距离的函数）和阴影效应，我们假设信道条件在每个周期内保持静态。此外，每个周期的持续时间与模拟的持续时间相同，即T_{sim}。

在模型训练过程中，我们采用余弦退火策略逐步衰减学习率，这有助于提升策略的泛化能力和性能表现。模型会定期进行评估，当智能体收敛至稳定策略（即连续20次评估未观察到改进）或达到预定义的环境步数n_{total}时，训练即告终止。模型检查点与性能指标通过Weights & Biases平台进行记录[7]。为提升计算效率并加速收敛，我们使用SubprocVecEnv封装器实现并行化训练，以实例化多个独立环境[24]。据此，智能体在每次 rollout 阶段会与总计n_{env}个并行环境进行交互。经调参后，训练使用的主要超参数值详见表II。

接下来，为评估奖励函数的重要性，图4展示了两个ML-G智能体的长期奖励演变情况——一个采用奖励塑形进行训练，另一个则不使用——每个智能体在10次独立训练运行中评估，每次运行10^7步。如图所示，采用奖励塑形的智能体在初始学习阶段表现出显著更快的速度，在前400万步内即获得更高奖励。这表明额外反馈引导早期探索朝向更具潜力的调度动作。虽然两个智能体最终在600万至1000万步左右收敛到相近的性能水平，但塑形奖励在最终训练性能稳定性和峰值奖励方面仍保持轻微优势。这些结果表明，奖励塑形不仅能加速收敛，还能在训练后期带来更优的策略精炼效果。

类似地，图5展示了两种智能体在有无奖励塑形训练时的实际最差延迟性能（以各周期所有STA中最差99$^{\text{th}}$百分位延迟计）。阴影区域表示先前分析的相同20次训练（10次有奖励塑形，10次无奖励塑形）的最小-最大边界，而实线则代表平均值。随着训练推进，两个智能体均逐步降低延迟；但采用奖励塑形的模型收敛速度显著更快，且在整个训练过程中达到的最差延迟值更低。未使用塑形的智能体表现出更高的波动性和较慢的收敛速度，尤其在训练初期阶段。

[7]https://wandb.ai/site

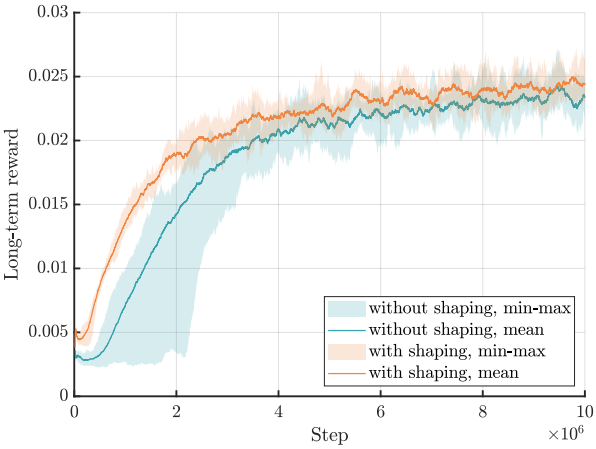


图4：ML-G智能体训练过程中，有/无奖励塑形时的长期奖励演变。

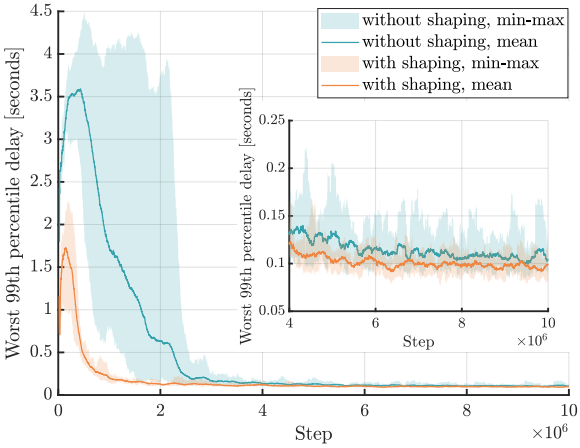


图5：在使用/不使用奖励塑形的情况下，ML-G代理训练过程中最差99$^{\text{th}}$百分位延迟的演变。

右下角的插图为稳定区域的特写（从4× 10^6到10^7步），其中奖励塑形的优势保持了一致性。

B. Inference Results

为验证我们的方案，首先评估了训练好的DRL智能体及启发式算法在图3所示部署场景中的性能表现。随后，我们将实验扩展到其他随机生成的部署场景，以验证方案的适应性和泛化能力。最后，我们分析了基于机器学习的智能体在用户数量从N =8增至N =20时的扩展性表现。

1) Sample Deployment: 图6展示了在如图3所示的部署场景中，所有评估调度器在100种不同流量实现下所达到的99$^{\text{th}}$百分位延迟与平均延迟，该场景包含4个接入点和16个站点。在每次流量实现中，每个站点被独立分配突发流量或泊松流量，其提供负载ω_i从

表II：训练参数。

Parameter	Description	Value
`gamma`	Discount factor	0.99
`gae_lambda`	GAE (Generalized Advantage Estimation)	0.92
`n_steps`	Steps per environment per update	128
`batch_size`	Minibatch size	256
`clip_range`	Clipping parameter	0.2
α_{init}	Initial learning rate	6.5×10^{-4}
α_{schedule}	Learning rate schedule	Cosine decay
n_{env}	Number of parallel environments	10
n_{total}	Total number of steps	10^7
β	Scale factor for long-term reward	10^{-3}
ν	Factor to avoid indetermination	10^{-6}

范围[10, 90]——对应平均负载为50 Mb/s，与本节稍后讨论的中高流量状态相当。在启发式基线中，MNP和OP表现出最高的最坏情况延迟，超过240毫秒。这证实了它们在高度负载或突发场景下的适应性较差，其中优先考虑总队列长度（MNP）或队首数据包（OP）可能导致饥饿和低效的组选择。相比之下，TAT明确平衡了流量对齐和高效传输，将99$^{\text{th}}$百分位延迟显著降低至103.85毫秒。

基于学习的调度器ML-G和ML-E均以显著优势超越所有启发式方法。值得注意的是，ML-E实现了最低的最坏情况延迟72.98毫秒，相比ML-G降低了8%，较TAT提升了30%，相对MNP和OP更是减少了70%以上。这些结果表明，深度强化学习智能体能在多样化流量变化下有效调度群组，同时平衡最坏情况延迟与传输效率。在两种场景中，基于机器学习的调度器始终比所有启发式机制获得更低的平均延迟。

图7展示了在如图3所示的部署场景下，针对单次流量实现所测量的各调度策略按优先级指数的归一化选择频率。优先级指数从低(L)到高(H)分布，其中L对应所选动作至少包含具有最低队首时延的STA，H则对应具有最高时延的STA。需注意由于缓冲区状态会随着传输持续演变，同一动作在连续TXOP中可能对应不同优先级水平。纵轴表示各优先级水平被选中的TXOP占比。可以看出，MNP的选择分布相对均匀，体现了其面向吞吐量的设计；而OP始终集中于最高优先级，持续服务最旧数据包，但代价是同时也会服务于所选SR组内低优先级的STA。TAT同样偏向高优先级，但其平滑的分布曲线反映了时延平衡机制。相比之下，ML-G在优先级谱系上呈现渐进式增长，表明其学习到的策略能在时延降低与调度多样性之间取得平衡。类似地，ML-E展现出高度多样性，但对高优先级水平存在略为明显的偏向。

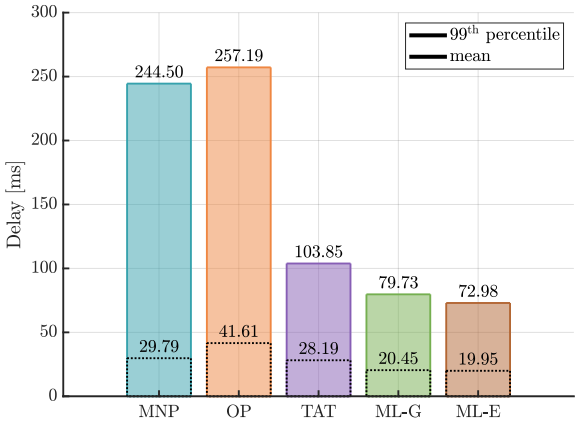


图6：样本部署中所有调度策略的第99$^{\text{th}}$百分位数和平均延迟，经过100次流量实现后，$\omega_i \in [10, 90]$且丢弃了10%的流量实现。

由于其在此部署中的专业性。

2) Random Deployments: 图8通过展示100次随机部署中最坏情况延迟的分布，提供了详细的性能分析。每次部署中，4个AP的位置如图3所示，而每个AP周围随机放置N_j =4个STA（总计N =16个站点）。我们在三种每STA流量负载模式下评估所有调度策略：低负载ω_i ∈[10,30]、中负载ω_i ∈[30,50]和高负载ω_i ∈[50,70]。对于低和中流量负载，所有部署均纳入分析。然而在高负载条件下，41%的部署因网络过载被排除。此场景中省略了ML-E的结果，因为其在训练所用部署之外的性能表现显著下降。

在低负载情况下，如图8a所示，所有调度器均能保持延迟有界，但MNP表现出显著更高的波动性，偶尔会产生超过30毫秒的最坏情况延迟。基于机器学习的方法总体上展现出更严格的延迟上限，但在低负载条件下并未超越OP或TAT的表现，其中OP成为该场景下最高效的调度器——这是因为网络利用率不足，此时最优解即为服务队首（HoL）数据包。

随着负载增加，图8b显示非机器学习调度器开始出现性能分化。MNP和OP方案呈现出更高的中位数延迟及更广泛的异常值分布。TAT保持相对稳定，但ML-G始终优于所有基线方案——其四分位距更窄、异常值更少，这归因于它以不同方式处理最坏情况：如图7所示，该算法并非总是调度最早到达的数据包。

在高负载情况下，如图8c所示，ML-G与基线方法之间的性能差距显著扩大。MNP和OP频繁遭遇超过100毫秒的最坏情况延迟。TAT虽表现出更平缓的性能下降，但仍落后于ML智能体。ML-G通过采用{v*}策略，既保持了较低的中位延迟，又实现了更紧凑的延迟分布。

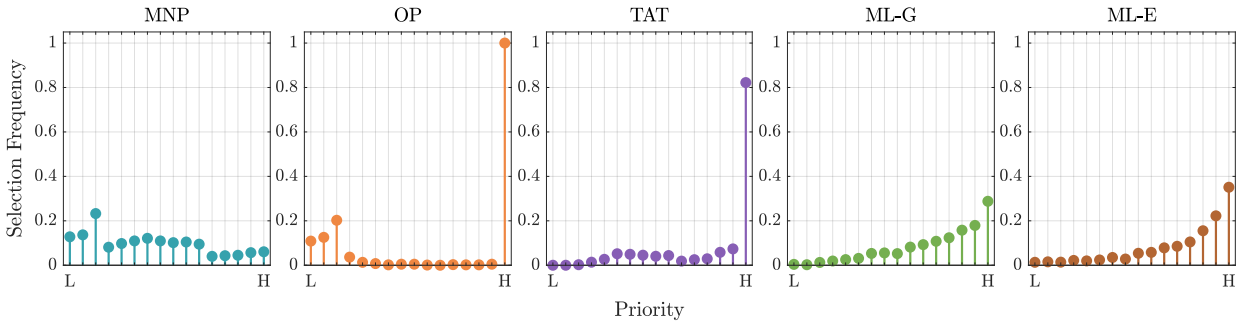


图　图3部署中每种评估调度策略按优先级指数的归一化选择频率和 o 交通流量实现。

一种更灵活的处理延迟策略，而非在每个决策步骤严格服务最早的数据包，这证实了学习策略在网络压力下的韧性。

最后，图9展示了四种调度策略——MNP、OP、TAT和ML-G——在100次部署实现中达到的第99百分位延迟和平均延迟，其中$\omega_i \in [10, 90]$（与ML-G代理训练时使用的范围相同），且有15%的部署因过载被排除在分析之外。结果表明，所提出的基于机器学习的调度器（ML-G）在平均延迟和最坏情况延迟方面均优于所有启发式基线。具体而言，ML-G实现了最低的第99百分位延迟64.54毫秒，比表现最佳的启发式策略TAT（75.92毫秒）提升了15%。在平均延迟方面，ML-G同样表现优异，达到16.87毫秒——略高于MNP的15.55毫秒，但显著低于OP（23.20毫秒）和TAT（23.00毫秒）。

这些结果共同表明，基于学习的调度器不仅在平均延迟和尾部延迟指标上优于启发式基线，还在不同负载条件下展现出更优的稳定性。其捕捉跨场景调度模式的能力使其相较于固定优先级策略具有显著优势。

3) Scalability: 图10展示了四种调度策略——MNP、OP、TAT和ML-G——在最坏情况下的延迟分布，随着用户数量从J =4个AP和$N \in \{8,12,16,20\}$个用户逐渐增加。每个箱线图汇总了100次随机部署实现的结果，其中部分部署因过载被舍弃（比例从8个STA时的4%到20个STA时的24%不等）。所有场景和部署实现中，网络总负载保持恒定（平均值），平均网络负载为ω_{net} =800 Mb/s，标准差为σ_{net} =92.4 Mb/s。需注意的是，此设置下每用户流量分布与先前分析的16个STA案例相当，即$\omega_i \in$[10,90] Mb/s。为匹配每种配置，使用相应的网络负载分布训练了四个不同的ML-G代理。

在不同数量的STA中，ML-G智能体相比启发式基线方法，始终能实现更低的最坏情况延迟。随着STA数量的增加，这一性能差距变得更为显{v*}

随着STA数量的增加，这一优势更为显著，反映出ML-G在高密度场景下更优的适应性。虽然TAT在$N \in \{8, 12, 16\}$用户数时表现出竞争力，但在用户数更多（N = 20）的场景下性能下降。相比之下，ML-G在所有评估场景中均能将延迟分布（包括异常值）严格控制在100毫秒以内，展现了不同网络密度下稳健的队列管理与调度决策能力，证实了基于机器学习的方法在延迟性能方面能随用户密度提升而良好扩展。

七、结论

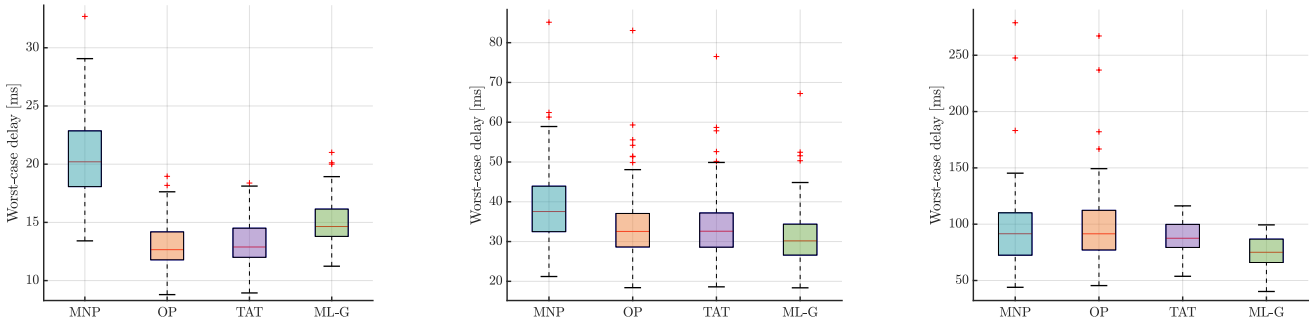
本研究探讨了利用深度强化学习（DRL）解决IEEE 802.11bn Wi-Fi网络中多接入点协作（MAPC）调度问题的方法。我们设计并实现了一个基于DRL的智能体，该智能体能够观测各站点最差延迟、队列大小及信道条件，并做出旨在最小化最坏情况延迟的调度决策。该智能体通过近端策略优化（PPO）算法，在一个兼容Gymnasium的Wi-Fi仿真环境中进行训练，该环境模拟了协作式空间复用（Co-SR）及异构流量场景。

我们的结果表明，所提出的模型不仅能够泛化至多样化的部署场景与流量模式，还持续优于三种成熟的MAPC调度启发式方法（MNP、OP和TAT）。这些发现凸显了基于学习的方法在提升协调Wi-Fi网络时延敏感性能方面的潜力，为下一代无线标准中更智能的调度铺平了道路。

先进的调度策略——如涉及群组管理和传输功率控制的策略——以及多链路操作（MLO）等新兴功能的集成，将留待未来工作。此外，开发一个能够扩展至任意网络规模、无论接入点（AP）和站点（STA）数量多少的通用模型，仍然是一个开放的研究方向。

八、致谢

本文由CHIST-ERA Wireless AI 2022年度MLDR项目（编号ANR-23-CHR4-0005）资助，并部分获得AEI和NCN的经费支持，项目编号分别为PCI2023-145958-2和2023/05/Y/ST7/00004。D.的工作



(a) Low load (b) Medium load (c) High load

图8：100次随机部署中最坏情况下的延迟分布。

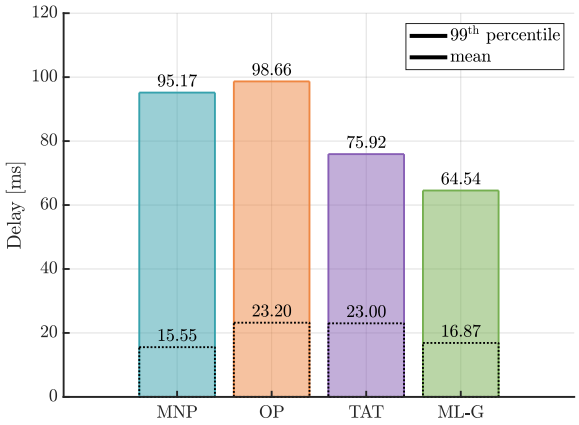
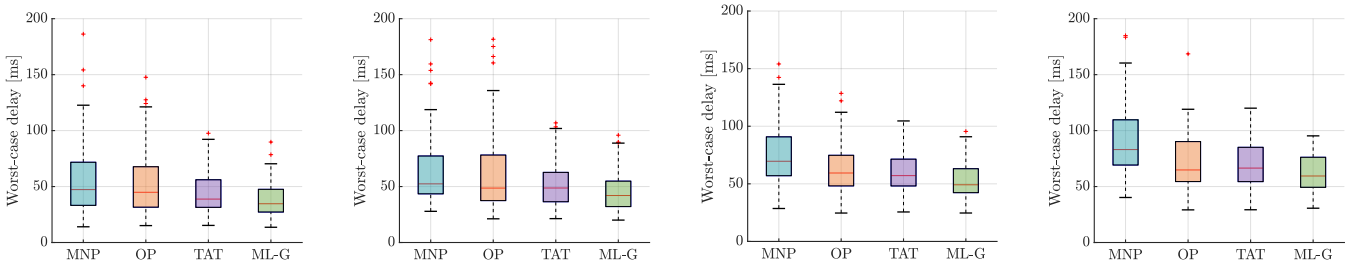


图9：100次部署实现后所有调度策略的99$^{\text{th}}$百分位数和平均延迟，其中$\omega_i \in [10, 90]$且15%的部署被舍弃。

努涅斯、F. 威廉米和B. 贝利亚尔塔还部分获得了Wi-XR PID2021-123995NB-I00（MCIU/AEI/FEDER,UE）、MCIN/AEI在玛丽亚·德·梅兹图卓越单位计划（CEX2021-001195-M）下的资助，以及ICREA Academia 00077的支持。

参考文献

[1] R. Costa, P. Portugal, F. Vasques, C. Montez, 与 R. Moraes, “IEEE 802.11 DCF、PCF、EDCA及HCCA处理实时流量的局限性”, 收录于 *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*。IEEE, 2015年, 第931–936页。 [2] L. Galati-Giordano, G. Geraci, M. Carrascosa, 与 B. Bellalta, “Wi-Fi 8将是什么? IEEE 802.11bn超高可靠性初探”, *IEEE Communications Magazine*, 第62卷第8期, 第126–132页, 2024年。 [3] D. Nunez, F. Wilhelmi, L. Galati-Giordano, G. Geraci, 与 B. Bellalta, “IEEE 802.11bn协调多AP无线局域网中的空间复用: 吞吐量分析”, 2024年。[在线]。可访问: https://arxiv.org/abs/2407.16390 [4] I. Val, D. López-Pérez, A. Kijanka, S. Schelstraete, L. Muñoz, D. Arlandis, 与 M. Martínez, “Wi-Fi 8揭秘: 关键特性、多AP协调及C-TDMA的作用”, 2025年。 [5] D. Nunez, P. Imputato, S. Avallone, M. Smith, 与 B. Bellalta, “通过多AP联合调度实现Wi-Fi 8的低延迟可靠性”, *IEEE Open Journal of the Communications Society*第6卷, 第2090–2101页, 2025年。[6] F. Wilhelmi, C. Cano, G. Neu, B. Bellalta, A. Jonsson, 和 S. Barrachina-Muñoz, “无线网络中基于自私多臂老虎机的协作式空间复用”, *Ad Hoc Networks*, 第88卷, 第129–141页, 2019年。[7] F. Wilhelmi, S. Barrachina-Munoz, B. Bellalta, C. Cano, A. Jonsson, 和 G. Neu, “多臂老虎机在WLAN分散式空间复用中的潜力与陷阱”, *Journal of Network and Computer Applications*, 第127卷, 第26–42页, 2019年。[8] A. Bardou, T. Begin, 和 A. Busson, “改进IEEE 802.11ax WLAN的空间复用: 一种多臂老虎机方法”, 收录于 *Proceedings of the 24th International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2021年, 第135–144页。[9] Y. Huang, “提升无线局域网空间复用的强化学习方法”, 博士论文, 伍伦贡大学, 2022年。[10] F. Wilhelmi, S. Szott, K. Kosek-Szott, 和 B. Bellalta, “机器学习与Wi-Fi: 迈向AI/ML原生IEEE 802.11网络的路径”, *IEEE Communications Magazine*, 2024年。[11] F. Wilhelmi, B. Bellalta, S. Szott, K. Kosek-Szott, 和 S. Barrachina-Muñoz, “协调多臂老虎机提升Wi-Fi空间复用”, 2025年。[在线]。可用: https://arxiv.org/abs/2412.03076 [12] M. Wojnar, W. Ciezobka, K. Kosek-Szott, K. Rusek, S. Szott, D. Nunez, 和 B. Bellalta, “基于层次多臂老虎机的IEEE 802.11bn多AP协调空间复用”, *IEEE Communications Let- ters*, 第29卷第3期, 第428–432页, 2025年。[13] M. Wojnar, W. Ciężobka, A. Tomaszewski, P. Choda, K. Rusek, K. Kosek-Szott, J. Haxhibeqiri, J. Hoebeke, B. Bellalta, A. Zubow, F. Dressler, 和 S. Szott, “IEEE 802.11 MAPC网络中基于机器学习的协调空间复用调度”, *IEEE Journal on Selected Areas in Communications*, 第1–1页, 2025年。[14] Q. Li, B. Lv, Y. Hong, 和 R. Wang, “ReinWiFi: 基于强化学习的WiFi网络应用层QoS优化”, 2025年。[在线]。可用: https://arxiv.org/abs/2405.03526 [15] Cui, Wei 和 Shen, Kaiming 和 Yu, Wei, “空间深度学习在无线调度中的应用”, *IEEE Journal on Selected Areas in Communications*, 第37卷第6期, 第1248–1261页, 2019年6月。[在线]。可用: http://dx.doi.org/10.1109/JSAC.2019.2904352 [16] M. Du, R. Yan, P. Liu, Z. Guo, 和 X. Sun, “基于深度强化学习的IEEE 802.11bn空间复用”, 收录于*2025 IEEE Wireless Communications and Networking Conference (WCNC)*。IEEE, 2025年, 第1–6页。[17] “IEEE 802.11-25/0502r0: 统一MAPC框架详情”, 2025年。[18] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG *et al.*, “Gymnasium: 强化学习环境的标准接口”, *arXiv preprint arXiv:2407.17032*, 2024年。[19] R. BELLMAN, “马尔可夫决策过程”, *Journal of*



(a) $N = 8$ STAs (b) $N = 12$ STAs (c) $N = 16$ STAs (d) $N = 20$ STAs

图 10: 不同用户数量 $N \in \{8, 12, 16, 20\}$ 在100次随机部署下的最坏情况延迟分布 t
实现 每种情景下的应用。

*Mathematics and Mechanics*第6卷第5期，第679–684页，1957年。[在线]。可访问：http://www.jstor.org/stable/24900506 [20] J. Schulman、F. Wolski、P. Dhariwal、A. Radford与O. Klimov合著，“近端策略优化算法”，2017年。[在线]。可访问：https://arxiv.org/abs/1707.06347 [21] V. Konda与J. Tsitsiklis合著，“演员-评论家算法”，收录于*Advances in Neural Information Processing Systems*，S. Solla、T. Leen与K. Müller编，第12卷。麻省理工学院出版社，1999年。[在线]。可访问：https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf [22] S. Merlin*et al.*，“TGax仿真场景”，2015年11月，文件编号：IEEE 802.11-14/0980r16。[23] The MathWorks公司，“802.11be EHT MU单用户分组格式的包错误率仿真”，美国马萨诸塞州纳蒂克，2024年。[在线]。可访问：https://www.mathworks.com/help/wlan/ug/802-11be-packet-error-rate-simulation-for-eht-mu-single-user-packet-format.html [24] A. Raffin、A. Hill、A. Gleave、A. Kanervisto、M. Ernestus与N. Dormann合著，“Stable-Baselines3：可靠的强化学习实现”，*Journal of Machine Learning Research*，第22卷第268期，第1–8页，2021年。[在线]。可访问：http://jmlr.org/papers/v22/20-1364.html