**文章编号** 1672-6634(2020)06-0040-07 **DOI** 10.19728/j.issn1672-6634.2020.06.005

# 基于深度强化学习的OFDMA-PON三维资源分配研究与性能分析

陈 斌[1] 顾家骅[2] 朱 敏[2] 晏春平[3] 周怡君[4] 顾萍萍[3]

(1.东南大学 电子科学与工程学院,江苏 南京 210096;2.东南大学 移动通信国家重点实验室,江苏 南京 210096;
3.太仓市同维电子有限公司 江苏 太仓 215400,4.东南大学 机械工程学院,江苏 南京 210096)

**摘 要** 由于具有大容量、高效灵活的多地址访问、高频谱效率、动态带宽分配等优点,正交频分复用接入无源光网络(OFDMA-PON)成为了下一代光接入网络的最有潜力的选择之一.在OFDMA-PON中,不同的光网络单元(ONU)可以共享子载波资源来支持网络资源管理和有效带宽分配.在上行传输中,多个ONU可以在整个传输周期内的不同时隙(TS)内共享正交低比特率子载波(SC)来传输上行数据.本文提出了一种基于深度强化学习(DRL)的动态子载波分配(DSA)策略.该策略以动态灵活的方式联合分配OFDMA-PON中时隙、子载波和调制格式等三维资源,通过采用合适的调制格式,同时优化业务延迟和ONU发射功率.将本文提出的基于DRL的DSA算法与传统的二维DSA算法进行仿真比较,结果表明,本文提出的DSA算法不仅大大降低了业务延迟,还可以节省ONU发射功率.

**关键词** OFDMA-PON;DRL;动态子载波分配;低延迟;节能

**中图分类号** TN915.6 **文献标识码** A

## 0 引言

随着各种新兴多媒体业务对网络带宽需求的日益增长,高效经济的无源光网络(PON)已经成为了"最后一公里"宽带接入的一种成熟技术,在全球范围内得到了广泛铺设[1].近年来,正交频分复用(OFDM)技术在光网络研究领域获得了令人瞩目的发展势头[2].OFDM技术利用正交性,将发送的信号划分为几十个,乃至数百个低速率、部分重叠但互不干扰的子载波信号[3].由于具有大容量、高效灵活的多地址访问、高频谱效率等优点,基于OFDM技术的正交频分多址无源光网络(OFDMA-PON)已经成为下一代光接入网络有前途的解决方案之一[2,3].

通常,OFDMA-PON中每个子载波所承载的比特率远低于单个波长的比特率,也远低于一个光网络单元(ONU)的平均速率[2].这就意味着,要为每个ONU提供所需的带宽,这就需要将多个子载波组合在一起,共同为该ONU提供载波服务.在可用频谱资源有限的情况下,服务于每一个ONU的子载波数量就要加以控制,否则会影响OFDMA-PON所提供的服务质量(QoS),如延迟性能.我们知道,采用高阶调制格式可以提高频谱利用率,有助于减少所需的子载波的数量,进而改善延迟性能[5].然而,为了保证一定的传输质量,高阶调制格式往往需要更多发射能量.据研究报道,ONU的功耗占OFDMA-PON能耗的60%至70%[6].在实际应用中,尽可能降低网络能耗,也是人们不断追求的目标之一.因此,在OFDMA-PON中,考虑到调制格式配置的动态子载波分配(DSA)算法会极大地影响网络性能,如信道利用率、业务延迟和网络能耗等[4].

**收稿日期**:2020-05-05

**基金项目**:国家自然科学基金项目(61771134)资助

**通讯作者**:朱敏,男,汉族,博士,副教授,研究方向:光网络与光通信,E-mail: minzhu@seu.edu.cn.

为了提高 OFDMA-PON 资源分配的效率，早期工作[3,4,7-9]主要基于二维资源的联合 DSA 算法被提出，即时隙(TS)和子载波(SC)的分配. 文献[3]在 OFDMA-PON 提出了一种加权 DSA 调度算法来减少终端无线数据包延迟. 文献[4]提出的算法在动态带宽分配上结合了流量预测技术来降低延迟. 文献[7]针对 OFDMA-PON 的上行资源分配问题，利用离线调度框架来分析子载波信道利用率和总授权时间. 文献[8]提出了一种在距离自适应 OFDMA-PON 中的公平感知 DSA 算法. 文献[9]提出了一种异构 OFDMA-PON 中的动态带宽分配框架，并开发了基于权重分布的 ONU 调度新算法. 但是，以上这些 DSA 算法都没有考虑子载波调制格式的灵活分配，也没有考虑 ONU 发射功率的优化配置.

文献[5]考虑了 OFDMA-PON 中时隙 TS，子载波 SC 和调制格式这三维资源，通过在每个时隙中实现子载波和调制格式的最佳分配，来最小化 ONU 的发射功率. 文献[6]同样研究了虚拟子载波(VS)，TS 和调制格式的联合分配，通过多维资源的灵活重配置来最大程度地节省能耗. 文献[10]通过共享 OFDM 调制模块来提高波分复用正交频分复用-无源光网络(WDM-OFDM-PON)的能量效率. 文献[11]提出一种距离自适应带宽分配方案，实现低成本大容量长距离 OFDMA-PON. 但是上述这些方案并未考虑 OFDMA-PON 所需满足的服务质量，如 ONU 请求业务延迟.

最近，深度强化学习(DRL)已成功地应用于资源管理的一些复杂决策问题，在提高通信网络性能方面引起了学术界和工业界的广泛关注. 文献[12,13]中研究了 5G 网络中基于深度强化学习的切片准入策略，以最大程度地提高基础架构提供商的利润. 文献[14]从广义的角度，针对网络多种资源优化配置问题，演变成“装箱”问题，并通过 DRL 工具来解决，以最大程度地减少工作延迟. 文献[15]提出了一种基于 DRL 的 C-RAN 中的联合 BBU 布局和路由策略，以最大程度地利用资源. 文献[16]提出了一种基于 DRL 的策略来提高弹性光网络环境下的网络整体性能.

在本文中，据我们所知，我们首次应用 DRL 技术来解决 OFDMA-PON 中的动态子载波分配 DSA 问题. 提出的基于 DRL 的 DSA 算法可以根据不同 ONU 请求的带宽需求，联合动态分配可用的子载波数量、时隙和调制格式，以最大程度地降低 ONU 功耗和 ONU 请求的延迟.

## 1　系统模型与问题建立

图 1 显示 OFDMA-PON 系统的物理架构. OFDMA-PON 系统具有 3 个组成部分：位于中心局(CO)的光线路终端(OLT)，基于光分离器(Splitter)的无源光分配网络(ODN)和位于用户端的多个 ONU. OLT 通过 ODN 将来自中心局的下行业务数据流广播给每个 ONU. ONU 有选择地接收由 OLT 广播的下行数据，并将其发送给用户. OLT 通过 ODN 从每个 ONU 收集上行数据. 上/下行数据业务通过多个 OFDM 子载波进行传输. 在本文中，我们专注于 OFDMA-PON 系统的上行链路传输，其中 OFDM 符号由正交子载波承载，为不同的 ONU 选择不同的调制格式，分配不同的时隙和子载波信道，每个子载波信道包括一个或多个子载波 SC.

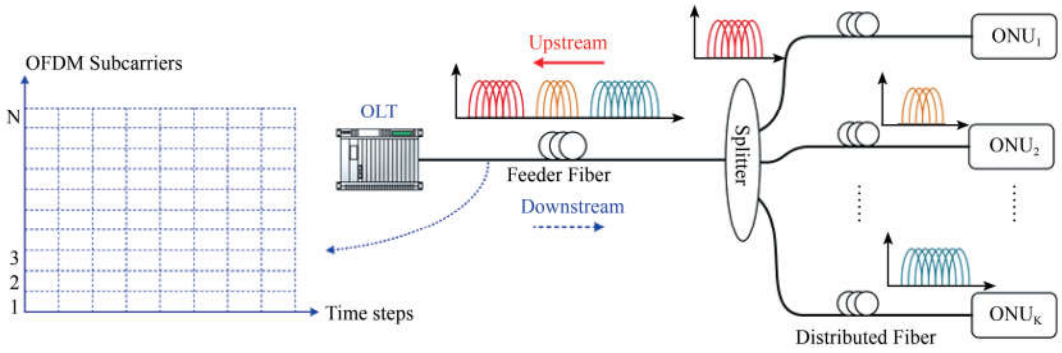



**图 1　OFDMA-PON 系统结构**

首先本文对 OFDMA-PON 系统进行建模，一共有 $N$ 个子载波 SC 和 $K$ 个 ONU，并且每个 SC 在每个时隙内只能被一个 ONU 占用. 一个 ONU 所使用的多个 SC 必须是相邻的，并且这些 SC 的开始/结束时间均相等. 因此，如图 2 所示，可以将分配每个 ONU 的 SC 和 TS 资源，表示为一个矩形. 而且，对于某一个 ONU 业务请求来说，采用不同的调制格式，所需的子载波的数量会发生变化，表示分配资源的矩形也会发

生变化.因此,ONU 的发射功率和整体平均延迟也将有所不同.例如,如果 ONU 选择一个低阶的调制格式,虽然 ONU 的发射功率会较低,但分配给该 ONU 的子载波数量就会增加,从而产生较高的业务延迟.反之亦然.由此可见,调制格式的不同选择,使得 ONU 的发射功率和 ONU 的平均延迟是相互联系的.要在 OFDMA-PON 中实现高效的 DSA 算法,需要谨慎的时隙 TS,子载波 SC 和调制格式的联合分配.

$b_k$表示为分配给第$k$ 个 ONU 的某种调制格式,同时也表示为在这种调制格式下,每一个 OFDM 调制符号所代表的比特数.$b_k$取值为 1,2,…,$M$,其中 $M$ 是每个 OFDM 调制符号代表的最大比特数.这表明对应的调制格式是从 BPSK 到$2^M$-QAM.本文假设分配给一个 ONU 的所有子载波的调制格式都相同.

如文献[6]中所述,电功率占 ONU 的总发射功率的很大一部分,因此本文也同样忽略了 ONU 的光功率.在一个时隙 TS 内,$P_k$表示第 $k$ 个 ONU 支持给定误码率(BER)$P_e$下的$b_k$比特/符号,单个子载波 SC 所需的发射功率[6]

$$P_k = \frac{N_0}{3} \cdot \left[Q^{-1}\left(\frac{P_e}{4}\right)\right]^2 \cdot (2^{b_k} - 1), \tag{1}$$

其中 $Q(x) = (1/\sqrt{2\pi}) \int_x^{\infty} e^{-t^2/2} dt$,$N_0$为噪声功率谱密度,根据文献[5],$\frac{N_0}{3} \cdot \left[Q^{-1}\left(\frac{P_e}{4}\right)\right]^2$ 的值等于 0.4039.因此,第 $k$ 个 ONU 的发射能耗可以简化表示为

$$E_k = 0.4039 \cdot (2^{b_k} - 1) \cdot T_k \cdot ceil\left(\frac{R_k}{b_k \cdot f_{SC}}\right), \tag{2}$$

其中 $T_k$是第 $k$ 个 ONU 请求的持续时间,ceil 上取整函数表示为第 $k$ 个 ONU 所需的子载波数,$R_k$是第 $k$ 个 ONU 的数据速率请求(单位为比特),$f_{SC}$是每一个子载波 SC 所占据的频谱带宽(单位为 Hz).式(2)可见 $b_k$的值越大,说明采用越高阶的调制格式,频谱利用率越大,所需的子载波频谱资源 SC 就越少,可以让更多的 ONU 业务请求得到 SC 资源分配,从而降低业务的平均延迟时间;但要满足一定的 BER 要求,所需要 ONU 信号发射功率也会增加(可由式(1)所示),反之亦然.

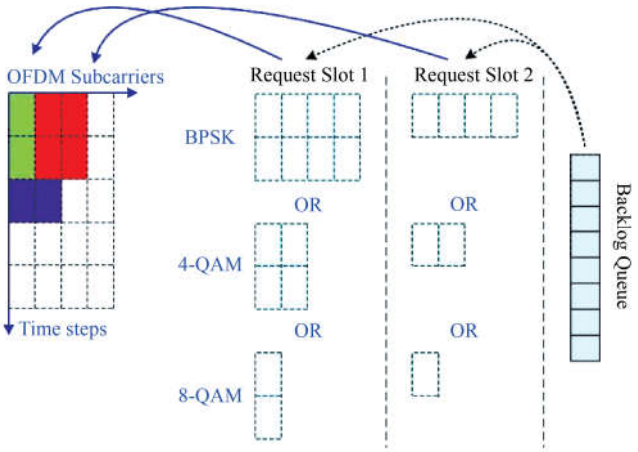



**图 2 三个可选调制格式下的两个待处理的 ONU 请求的状态表示示例**

我们的优化目标是,为每一个 ONU $k \in \{1,2,\cdots,K\}$分配最优的 $b_k$,从而联合最小化 ONU 请求的平均等待时间和平均发射功率.

$$Minimize\left(\alpha \cdot \sum_{k \in K} \frac{c_k - T_k}{T_k} + \beta \cdot \sum_{k \in K} \frac{P_k - P_k^{fix}}{P_k}\right), \tag{3}$$

其中$\alpha$ 和$\beta$是为调整两项的重要性而引入的因素,应根据实际情况进行设置.第一项反映归一化的总业务延迟;第二项表示归一化的总 ONU 发射功率;$c_k$是第 $k$ 个 ONU 请求的完成时间,包括请求的等待时间和请求实际持续时间 $T_k$,$P_k^{fix}$是第$k$ 个 ONU 采用指定调制格式所需的发射功率,而该指定调制格式可以满足所有 ONU 业务需求.

## 2 深度强化学习模型

如图 3 所示,深度强化学习方法是一个典型的马尔可夫决策过程[13].强化学习的目标是:给定一个马尔可夫决策过程,寻找最优策略.DRL 中的学习者或决策者被称为代理,与代理交互的代理外的所有部分,被称为环境.代理选择某些动作,然后环境响应这些动作并向代理反映新的环境.代理和环境在一系列离散时间步骤中相互作用相互影响.具体地说,在每一个时间步 $t$,代理会观察一些状态 $S_t$,并在当前状态的基础上选择一个动作 $A_t$.在一个时间步以后,作为该动作的结果,代理接收到新的奖励 $R_{t+1}$,并且环境的状态转换为 $S_{t+1}$.在马尔可夫决策过程中,$S_{t+1}$和 $R_{t+1}$的每个可能值的概率仅取决于紧接在前的状态 $S_t$ 和动作 $A_t$.状态必须包括有关过去的代理和环境互动所有方面的信息.

本节利用 DRL 策略网络，经过不断地迭代训练，为每一个 $ONU_k$ 分配最优调制格式 $b_k$，以最大程度地减少 ONU 请求的平均业务延迟和平均功耗. DRL 算法可以生成大量训练数据，同时将复杂的系统和决策建模为深度神经网络. 下面将定义 DRL 策略网络的三要素：状态，动作和奖励.

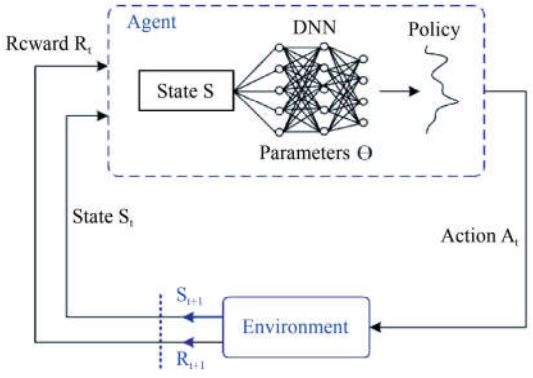



**图 3　强化学习中基于深度神经网络(DNN)策略的代理-环境交互**

状态：不同的图像表示系统的状态，包括当前已分配的系统资源情况，等待配置的 ONU 请求信息，以及在待办事项队列(Backlog Queue)中的候选配置的 ONU 请求信息. 图 2 中最左边的图像表示已分配的 ONU 请求，并从当前时间步开始，持续 T 时间步，直到所有 ONU 请求配置完成. 这些图像中的不同颜色表示不同的 ONU 请求. 例如，图 2 的已分配方案中，红色图块表示已成功配置的 2 个 TS、2 个 SC 的 ONU 请求. 等待配置的 ONU 请求图像表示采用不同的调制方式所需分配的 SC 和 TS 资源. 例如，图 2 右侧为等待配置中的 ONU 请求图像，当采用不同调制格式时，所需的 SC 和 TS 资源要求也不同. ONU 请求 1(Request Slot 1)要求两个 TS 的持续时间，当采用 BPSK 调制格式时，需要 4 个 SC 资源，当采用 4-QAM 时，则需要 2 个 SC 资源，或采用 8-QAM 时，则仅仅需要 1 个 SC 资源. 按照先来先处理的原则，按序处理 Backlog Queue 中最先到达的 $d$ 个 ONU 请求，以使代理中的神经网络输入可以表示为有限且固定的状态(图 2 中 $d=2$)[14]. 这样，不仅可以减少延迟，还可以限制动作空间，从而使强化学习更加有效.

动作：在每个时间步，代理中的调度程序可以调度 $d$ 个 ONU 请求的任何子集，并有 $M$ 种可选的调制格式(一个 ONU 请求仅选择一种调制格式). 这就需要 $2^{(d\times M)}$ 的动作空间，这个动作空间非常大，可能会使强化学习非常具有挑战性. 在图 2 中，为了大大降低动作空间的规模，可以允许调度程序在每个时间步执行多个动作. 给定动作空间由 $\{\varnothing,1\times1,1\times2,\cdots,i\times j,\cdots,d\times M\}$ 表示，其中元素 $a=i\times j$ 表示调度程序选择第 i 个请求槽中 ONU 请求，并采用第 $j$ 种调制格式，并试图把该请求的资源块放置在 SC 和 TS 资源图像中的适当位置；$a=\phi$ 表示在当前时间步中调度程序选择到无效动作，即不选择任何 ONU 请求进行资源配置. 当调度程序选择到无效动作或当前可用资源不能满足 ONU 请求时，时间步长向前移动一步，可用资源图像也向上移动一步. 新到达的 ONU 请求将通知调度程序并同时对请求槽状态进行更新. 这样，调度程序可以在同一时间步执行多个动作，完成多个 ONU 请求的配置，使得动作空间保持线性($d\times M$)[14].

**表 1　训练算法伪代码**

| 算法　深度强化学习中深度神经网络训练代码 |
|---|
| 输入：可区分的策略参数化 $\pi(A_t\mid S_t,\theta)$ |
| 结果：子载波资源分配策略 |
| 1　策略参数的初始化：$\theta\leftarrow 0$ |
| 2　For 每个 ONU 请求集： |
| 3　　For 每一轮 $i=1,\cdots,N$： |
| 4　　产生 $S_0^i,A_0^i,R_1^i,\cdots,S_{T-1}^i,A_{T-1}^i,R_T^i$，遵循策略 $\pi(A_t\mid S_t,\theta)$. |
| 5　　For 每一轮的每一步 $t=0,1,\cdots,T-1$： |
| 6　　$G^i\leftarrow\sum_{k=t+1}^{T}R_k^i$， |
| 7　　$b^i\leftarrow\frac{1}{N}\sum_{i=1}^{N}G^i$， |
| 8　　$\theta\leftarrow\theta+\alpha^{\theta}\gamma^t(G^i-b^i)\nabla ln\pi(A_t\mid S_t,\theta)$， |
| 9　　End |
| 10　End |
| 11　End |

奖励：通过奖励来给代理提供反馈，以寻求实现所需目标的最佳策略. 优化目标是通过为所有 ONU 请求联合分配时隙 TS、子载波 SC 和调制格式，尽可能减少 ONU 请求的平均业务延迟和平均功耗. 在单个时间步 $t$ 中，强化学习奖励设置为

$$R_t=-\alpha\cdot\left(\sum\nolimits_{j\in J}\frac{1}{T_j}-\sum\nolimits_{k\in K'}1\right)-\beta\cdot\sum\nolimits_{k\in K'}\frac{P_k-P_k^{fix}}{P_k},\tag{4}$$

其中 $J$ 是当前时间步下的系统中的所有 ONU 请求集，包括已分配的 ONU 请求集，等待配置的 ONU 请求

集,以及候选配置的 ONU 请求集,$K'$ 是在当前时间步下已分配的 ONU 请求集. 在时间步 $t$,代理观察到一些状态 $S_t$,并根据该奖励 $R_t$ 选择一个动作 $A_t$. 再经过一个时间步后,代理会收到一个新的奖励 $R_{t+1}$,并且环境状态会转换为 $S_{t+1}$. 请注意,在每个时间步长中,代理都不会获得任何中间动作的奖励. 通过与环境交互,代理尝试选择一种动作,以最大化将来收到的折扣奖励的总和. 也就说,强化学习的目的是最大化预期的累积折扣奖励:$E[\sum_{t=0}^{\infty}\gamma^t R_t]$,$\gamma\in(0,1]$. 在本文中,将折扣率$\gamma$设置为 1. 累积奖励最大化就是使平均 ONU 请求延迟和平均发射功率最小化.

本文通过很多轮迭代来训练代理中的策略网络. 在每一轮迭代中,固定数量的 ONU 请求到达并根据策略进行资源配置. 当所有 ONU 请求都执行完成时,本轮训练终止. 表 1 显示了深度神经网络训练算法的伪代码. 为了训练出通用的策略,训练过程中随机生成多个 ONU 请求集(第 2 行),对每个 ONU 请求集进行多轮探索(第 3 行),使用当前探索策略,以得到可能的动作概率空间,选择某一种动作,并使用产生的奖励值来进一步改进探索策略. 具体地说,我们记录每轮探索所有时间步的状态,动作和奖励信息,并使用这些值来计算每一轮探索每个时间步 $t$ 的累积折扣奖励.

## 3 实验结果

### 3.1 仿真参数

仿真参数设置:ONU 请求根据伯努利过程到达,到达率 $\lambda$(即每个时间步到达一个新的 ONU 请求的概率)从 0 到 1 变化,步长为 0.1. 本文考虑 32 个 SC 通道,总带宽为 1.28 GHz,则每一个子载波 SC 通道带宽为 0.04 GHz. 每一个 ONU 请求可选择调制格式为 4 种:BPSK,4-QAM,8-QAM 和 16-QAM. 优化目标中两个指标的权重设置为相同,$\alpha=\beta=0.5$. ONU 请求的持续时间设置为:80%ONU 请求的持续时间在 $1t$ 和 $3t$ 之间均匀选择;而其余的 20%ONU 请求从 $10t$ 到 $15t$ 之间均匀选择. ONU 的带宽需求 $R_k$ 设置在(0.32,4.48)Gb/s 范围内均匀分布.

在该算法中,本文使用具有 33 个神经元的完全连接的隐藏层和总共 532323 个参数的神经网络. DRL 代理使用的“图像”长 20 个时间步,每次仿真持续 50 个时间步. 在当前时间步,代理只调度最先达到的 $d$ 个 ONU 请求($d=8$),采用不同调制格式,同时也不断更新在待办事项队列中的 ONU 请求. 待办事项队列的长度设置为 64 个 ONU 请求. 在每次训练迭代中,本文使用 50 个不同的请求集,并对每个请求集并行运行 10 个蒙特卡洛模拟进行探索. 更新策略网络参数的学习率被设置为 0.001.

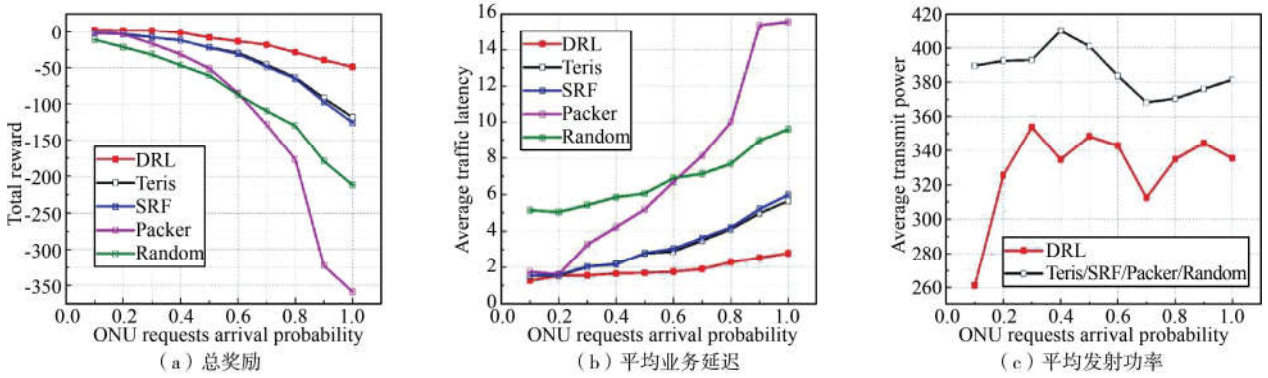
提出的灵活选择调制格式的 DRL 方案与四种固定调制格式的基准方案进行比较:(1) 随机 Random 算法,它随机选择请求;(2) 最短请求优先算法(SRF)[6],它按 ONU 请求的持续时间升序排列,(3) Packer 算法[17],它根据工作需求和资源可用性之间的排列顺序分配资源;(4) Teris 算法[17],综合了 SRF 算法和 Packer 算法的优势. 这四个基准启发性算法采用固定的调制格式,该调制格式是满足 ONU 最大带宽需求的最小阶调制格式,以此来尽可能减少发射功率.

### 3.2 仿真结果分析

由于 ONU 的数据速率请求 $R_k\in[0.32,4.48]$Gb/s,4 种基准算法的固定调制格式被设置为 16-QAM 才能满足所有 ONU 的数据速率需求. 图 4 比较了 ONU 请求到达率变化时的总奖励、业务延迟和发射功率. 与 Packer 和 Random 算法相比,SRF 算法在奖励和业务延迟方面均具有更好的性能. 在低负载下,SRF 性能类似于 Packer 算法. 随着负载的增加,SRF 和 Packer 算法之间的差异不断增加,SRF 接近 Teris. 因为尽管 Packer 为大带宽需求的 ONU 保留的资源比 SRF 多,但大带宽需求的 ONU 却更多,这直接导致 Packer 的延迟性能最差. Teris 结合了它们的优势,胜过 SRF 和 Packer 算法. 如图 4 所示,在高负载条件下,DRL 在这三个指标方面的表现要优于上述 4 种启发式算法. 这是因为 DRL 学会了为不同带宽需求的 ONU 请求灵活分配调制格式的能力,以节省功率;并为将来的 ONU 请求保留一些资源,以降低 ONU 请求平均等待时间,因此总奖励也是最高的.

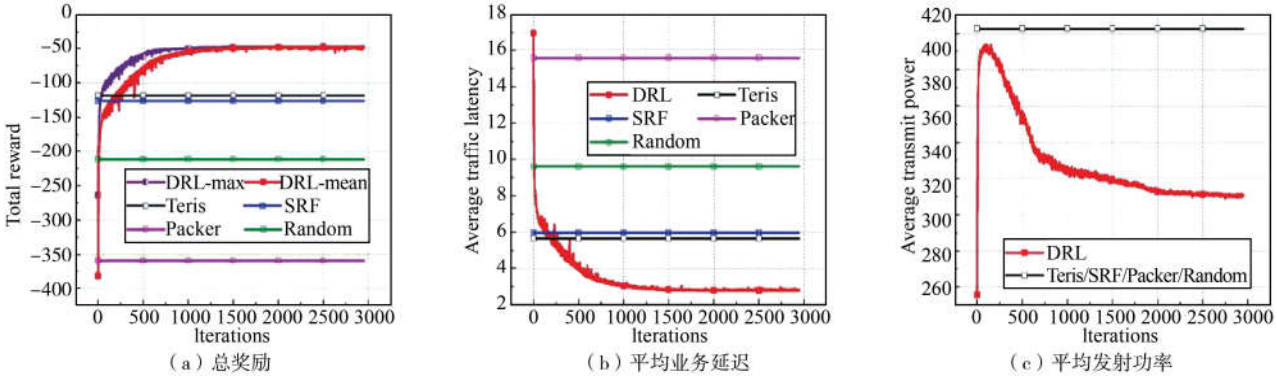
图 5 描述了当 ONU 请求到达率为 1 时,DRL 代理如何学习训练迭代. 在迭代开始时,DRL 代理没有任何先验知识. DRL 代理的行为类似于随机策略,并且行为比基准算法差. 随着迭代的进行,$DRL_{max}$ 和 $DRL_{mean}$ 的值都随着 DNN 的连续训练而增加. 经过约 100 次训练迭代后,DRL 得知可以通过为一些小请求保留一些

资源并使用更低阶的调制格式来增加总奖励，然后 DRL 继续尝试增加总奖励，直到经过 1500 次迭代后，$DRL_{max}$ 和 $DRL_{mean}$ 之间的差距越来越小并逐渐收敛到稳定值，这表明此时系统已达到最佳状态. 图 5(b)和(c)中的仿真结果表明，DRL 方案实现了更好的 ONU 业务延迟，又尽可能地降低发射功率. 值得一提的是，这四种基准启发性算法，不需要上述的迭代学习过程，但性能明显低于我们提出的基于 DRL 的 DSA 算法；同时，我们也看到，提出的 DSA 算法在迭代 1500 次之后就很快达到最佳状态，算法复杂度较低，收敛速度较快，所付出的时间成本在 OFDMA-PON 动态资源调度过程中可忽略不计.




(a) 总奖励　(b) 平均业务延迟　(c) 平均发射功率

**图 4　ONU 的数据速率请求 $Rk\in[0.32,4.48]$ Gb / s 和 $\alpha=\beta=0.5$ 的测试结果**




(a) 总奖励　(b) 平均业务延迟　(c) 平均发射功率

**图 5　ONU 的数据速率请求 $R_k\in[0.32,4.48]$ Gb/s 和 $\alpha=\beta=0.5$ 的训练结果**

## 4　结论

本文在 OFDMA-PON 中提出了一种基于 DRL 的新颖的三维 DSA 算法，联合配置了 ONU 请求的时隙 TS，子载波 SC 和调制格式，该算法同时优化了 ONU 请求的平均延迟和平均功耗. 从仿真结果可以看出，与 SRF 等四种基准调度算法相比，本文提出的基于 DRL 的 DSA 算法可以显着减少平均延迟和平均功耗，并且可以通过直接从经验自学策略中提高自身配置性能，是一种非常灵活的资源优化配置工具.

## 参　考　文　献

[1] 李贵鑫，孙卿，张圣羽，等. 光接入网中低延迟高能效动态带宽分配算法研究[J]. 聊城大学学报(自然科学版)，2019，32(5)：1-6.

[2] Kanonakis K，E Giacoumidis，I Tomkos. Physical-layer-aware MAC schemes for dynamic subcarrier assignment in OFDMA-PON networks [J]. Journal of Lightwave Technology，2012，30(12)：1915-1923.

[3] Wansu Lim，Konstantinos Kanonakis，Pandelis Kourtessis，et al. Flexible QoS differentiation in converged OFDMA-PON and LTE networks[C]. //2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)，2013.

[4] Wansu Lim，Pandelis Kourtessis，John M Senior，et al. Dynamic bandwidth allocation for OFDMA-PONs using hidden markov model [J]. IEEE Access，2017，5：21016-21019.

[5] Weizhi You，Lilin Yi，Silu Huang，et al. Power efficient dynamic bandwidth allocation algorithm in OFDMA-PONs [J]. IEEE/OSA Journal of Optical Communications and Networking，2013，5(12)：1353-1360.

[6] Bi M H，S L Xiao，L Wang. Joint subcarrier channel and time slots allocation algorithm in OFDMA passive optical networks [J]. Optics Communications，2013，287：90-95.

[7] Yumiko Senoo, Kota Asaka, Takuya Kanai, et al. Fairness-aware dynamic sub-carrier allocation in distance-adaptive modulation OFDMA-PON for elastic lambda aggregation networks [J]. IEEE/OSA Journal of Optical Communications and Networking, 2017, 9(7): 616-624.

[8] Wansu Lim, Pandelis Kourtessis, Konstantinos Kanonakis, et al. Dynamic bandwidth allocation in heterogeneous OFDMA-PONs featuring intelligent LTE-A traffic queuing [J]. Journal of Lightwave Technology, 2014, 32(10): 1877-1885.

[9] Gong X, L Guo, Q Zhang. Joint resource allocation and software- based reconfiguration for energy- efficient OFDMA-PONs [J]. IEEE/OSA Journal of Optical Communications and Networking, 2018, 10(8): 75-85.

[10] Xiaofeng Hu, Liang Zhang, Pan Cao, et al. Energy-efficient WDM-OFDM-PON employing shared OFDM modulation modules in optical line terminal [J]. Optics Express, 2012, 20(7), 8071.

[11] Xiaofeng Hu, Pan Cao, Jiayang Wu, et al. High-capacity and low-cost long-reach OFDMA PON based on distance-adaptive bandwidth allocation [J]. Optics Express, 2015, 23(2): 1249-1257.

[12] Muhammad Rehan Raza, Carlos Natalino, Peter Öhlen, et al. Reinforcement learning for slicing in a 5G flexible RAN[J]. Journal of Lightwave Technology, 2019, 37(20): 5161-5169.

[13] R S Sutton, A G Barto. Reinforcement Learning: An Introduction [M]. Second Edition England: The MIT Press, 2018.

[14] Hongzi Mao, Mohammad Alizadeh, Ishai Menache, et al. Resource management with deep reinforcement learning [J]. Proceedings of the 15th Acm Workshop on Hot Topics in Networks, 2016, 45(1): 50-56.

[15] Zhengguang Gao, Jiawei Zhang, Shuangyi Yan, et al. Deep reinforcement learning for BBU placement and routing in C-RAN[C]. // 2019 Optical Fiber Communications Conference and Exhibition (OFC), 2019.

[16] Xiao Luo, Chen Shi, Liqian Wang, et al. Leveraging double-agent-based deep reinforcement learning to global optimization of elastic optical networks with enhanced survivability[J]. Optics Express, 2019, 27(6): 7896-7911.

[17] Robert Grandl, Ganesh Ananthanarayanan, Srikanth Kandula, et al. Multi-resource packing for cluster schedulers [J]. Acm Sigcomm Computer Communication Review, 2014, 44(4): 455-466.

# Deep Reinforcement Learning for 3D Resource Allocation in OFDMA-PON and Performance Analysis

CHEN Bin[1] GU Jia-hua[2] ZHU Min[2] YAN Chun-ping[3]
ZHOU Yi-jun[4] GU Ping-ping[3]

(1. School of Electronic Science and Engineering, Southeast University, Nanjing 210096, China; 2. State Key Laboratory of Mobile Communications, Southeast University, Nanjing 210096, China; 3. TAICANG T&W Electronics Co. Ltd. Taicang 215400, China; 4. School of Mechanical Engineering, Southeast University, Nanjing 210096, China)

**Abstract** Due to the advantages of large capacity, flexible multiple address access, high spectrum efficiency and dynamic bandwidth allocation, orthogonal frequency division multiplexing access passive optical network (OFDMA-PON) has become one of the most potential choices for the next generation optical access network. In OFDMA-PON, it allows different optical network units (ONUs) to share subcarriers (SCs) to support network resource management and effective bandwidth allocation. In uplink transmission, multiple ONUs can share orthogonal low bit rate SCs to transmit data in different time slots (TSs) during the entire transmission cycle. In this paper, a dynamic subcarrier allocation (DSA) strategy based on deep reinforcement learning (DRL) is proposed. The strategy jointly allocates time slots, subcarriers and modulation formats in a dynamic and flexible manner. By using the optimal modulation format, the delay service quality is ensured and the ONU transmit power can be reduced. The DSA algorithm using DRL is compared with the traditional two-dimensional DSA algorithm. The simulation results show that the proposed DSA algorithm using DRL reaches lower traffic latency with energy saving.

**Key words** OFDMA-PON; DRL; dynamic subcarrier allocation; low latency; energy saving