引用本文：季晨阳，毕美华，周钊，等. 基于深度强化学习的多租户 PON 在线带宽资源分配算法[J]. 光通信技术，2021，45(9)：36-39.

基于深度强化学习的多租户 PON 在线带宽资源分配算法

季晨阳[1]，毕美华[1]，周 钊[2]，陈天宁[1]，林嘉芊[1]，徐志威[1]

（1.杭州电子科技大学 通信工程学院，杭州 310018；2.国家电网湖南省电力有限公司 信息通信分公司，长沙 410007）

摘要：为了解决运营商共享网络资源带来的资源分配问题，提出了一种基于深度强化学习（DRL）的在线带宽资源分配算法。该算法将多租户无源光网络（PON）系统映射到 DRL 模型中，DRL 代理通过与环境交互，为各个待处理的带宽请求和当前剩余带宽做决策，并不断更新策略参数直至模型收敛，从而完成算法优化。搭建了仿真系统，对该算法进行了可行性验证，仿真结果表明所提的算法可以有效提高带宽资源利用率。

关键词：多租户；网络共享；带宽资源分配；深度强化学习

中图分类号：TN915.6　**文献标志码**：A　**文章编号**：1002-5561(2021)09-0036-04

DOI：10.13921/j.cnki.issn1002-5561.2021.09.009　开放科学（资源服务）标识码（OSID）：

Online bandwidth resources allocation algorithm for multi-tenancy PON based on deep reinforcement learning

JI Chenyang[1], BI Meihua[1*], ZHOU Zhao[2], CHEN Tianning[1], LIN Jiaqian[1], XU Zhiwei[1]

(1. School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China;
2. Information and Communication Branch of State Grid Hunan Electric Power Co., Ltd., Changsha 410007, China)

Abstract: In order to solve the problem of resources allocation caused by operators sharing network resources, this paper proposes an online bandwidth resources allocation algorithm based on deep reinforcement learning (DRL). The algorithm maps multi-tenancy passive optical network (PON) system to DRL model. The DRL agent interacts with the environment to make decisions for each pending bandwidth request and the current remaining bandwidth, and constantly updates the policy parameters until the model converges, so as to complete the algorithm optimization. A simulation system is built to verify the feasibility of the algorithm. The simulation results show that the proposed algorithm can effectively improve the utilization of bandwidth resources.

Key words: multi-tenancy; network sharing; bandwidth resources allocation; deep reinforcement learning

0 引言

近年来，随着网络服务迅速发展[1-2]，网络运营商面临着网络需求（特别是接入网带宽需求）快速增加的挑战。无源光网络（PON）是一种提供宽带服务高效的接入网解决方案，已成为现在接入网部署的主流技术[3-4]。然而，传统接入网的不同运营商需要在同一地区部署各自的网络设施来服务其用户，独立部署 PON 系统将会造成大量的带宽浪费。为此，在多个网络运营商之间共享网络资源即多租户共享的解决方案应运而生[5]，其中，基础设施供应商负责物理网络的部署和维护，网络运营商向基础设施供应商租借现有的资源，并为自己的用户服务[6]。于是，网络共享带来的资源分配问题成为亟待解决的难题。文献[7]提出了一个多运营商间的带宽资源共享框架，并在该框架中引入了一个片调度器，充当电路开关，在一帧的持续时间内将整个 PON 容量分配给某一个虚拟网络运营商（VNO）。然而，在该框架中，动态切换方法导致带宽短缺并增加负载，使得 VNO 之间的隔离性变差。与此同时，深度强化学习（DRL）已成功地应用于资源管理的决策问题，特别是在提高通信网络性能方面的应用引起了学术界和工业界的关注。文献[8]提出利用 DRL 进行资源

收稿日期：2020-11-25。

基金项目：国家自然科学基金（No.61501157）资助；浙江省自然科学基金（No. LY20F050004）资助；浙江省教育厅一般科研项目（No. Y201942093、No.Y201942104、No.Y202044258）资助。

作者简介：季晨阳（1995—），男，硕士研究生，现就读于杭州电子科技大学通信工程学院电子与通信工程专业，本科毕业于武汉科技大学通信工程专业，主要从事无线接入网中带宽分配方法的研究及其仿真验证工作。

*** 通信作者**：毕美华（1981—），女，博士，副教授，主要从事高速光通信与光网络的安全性研究工作。

管理，在解决多资源集群调度问题时优于启发式算法，能够有效减少作业的平均等待时间。多租户 PON 系统对请求的等待时间要求并不严格，而是更加关注 PON 系统的带宽利用率。因此，本文以提升共享网络的带宽利用率为目标，提出一种基于 DRL 的在线带宽资源分配算法。

1 基于 DRL 的多租户调度模型

本文针对 PON 结构中多租户资源调度问题进行研究，其调度模型如图 1 所示。在该调度模型中，租户们共用 PON 系统的分布式网络结构。与普通的 PON 系统相同，该 PON 系统由光线路终端(OLT)、光分配网络(ODN)和光网络单元(ONU)组成。在带宽分配过程中，PON 系统的带宽请求由 ONU(即租户)发出，模型中用“请求”表示；OLT 负责为 ONU 分配带宽，模型中用“集群”表示。

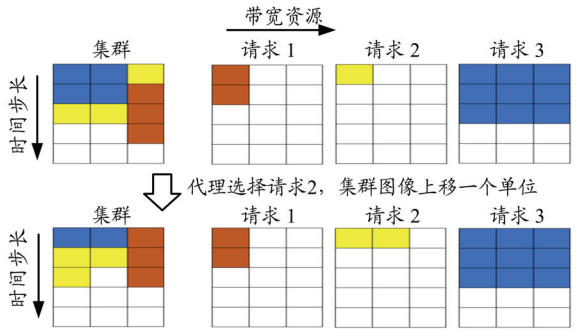


图 1 基于 DRL 的调度模型

对于一个具有带宽资源的集群(即 PON 系统中的 OLT)，传入的租户带宽请求以离散时间步长和在线方式到达。在每个时间步长的开始时刻，集群即 OLT 将带宽时频资源图的第一行清空并将全部已分配资源上移一个单位，代表上一个时间步长的结束，然后算法选择一个或多个等待的请求进行调度。为简单起见，本文假设每个带宽请求到达时就知道资源需求，一旦分配了某个请求，就必须从请求开始执行一直到完成该请求为止，而不会发生资源被抢占或中断。在此，集群图像代表基础设施供应商当前时刻的带宽分配结果，不同颜色的色块代表来自不同租户的资源请求大小。每个请求的行和列分别代表请求所需的时间步长和带宽大小。例如，图 1 中请求 1 代表该请求的执行需要 2 个时间步长和 1 个单位的带宽资源，请求 3 代表该请求的执行需要 3 个时间步长和 3 个单位的带宽资源。由于带宽具有时效性，当前未分配的带宽将无法再利用，不同的带宽分配选择将会导致不同的带宽利用率。为此，本文提出基于 DRL 的调度算法，用以提高系统的带宽利用率。

2 基于 DRL 的调度算法

为了有效处理针对多租户网络中的带宽分配问题，本文提出了一种基于 DRL 的在线带宽资源分配算法，为多租户网络设计的基于 DRL 的调度算法框架如图 2 所示，其中 DRL 代理与环境交互。在每个时间步长，DRL 代理都会从多租户网络(即环境)观察当前状态，并被要求选择一个操作，然后环境状态会转换为下一个状态。作为反馈，DRL 代理从环境中获得奖励，该奖励的值指示操作的良好程度。状态转换和奖励是随机的，并假定具有马尔可夫性质[9]，即该过程的未来状态的条件概率分布仅取决于当前状态，而不取决于它之前的事件顺序。该框架的目标是使预期的累积折扣奖励最大化。本文将介绍针对带宽分配问题提出的基于 DRL 的解决方案。

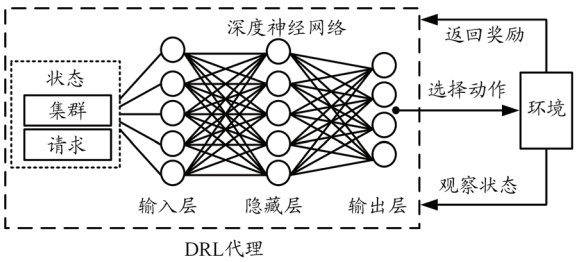


图 2 基于 DRL 的调度算法框架

多租户网络的状态包括集群中租户的带宽资源的当前状态以及请求集合中租户请求的带宽和时间资源。假设有 M 个待处理的 ONU 请求，对于每个 ONU 请求有选择或不选择 2 种方式，需要 2^M 的动作空间，这使得学习的过程非常具有挑战性。为了使动作空间保持线性，本算法允许同时处理多个 ONU 请求，并将动作空间表示为{Φ,1,2,…, M}，当选择 Φ 时即表示当前时间步长不再处理更多的请求；当选择 1~M 时则代表当前时间步长所要处理请求的个数，从而保持较小的动作空间。策略可以表示为一个以状态为输入的神经网络输出所有可能动作的概率分布。这样就可以通过神经网络获得{状态，动作}作为策略对，在每个有效操作且是第一个可能的时间步长中，为接受的租户请求分配其所需的带宽资源。在选择无效操作后，集群图像上移一个时间步长，新的租户请求到达。

学习的目标是使预期的累积折扣奖励最大化。为了实现最大化带宽利用率这一目标，本算法在每个时

间步长上都设计了一个奖励 $\sum_{j=1}^{J}(\frac{1}{T_j}+B_j)$，其中，$J$ 是当前系统中的请求集合，T_j 和 B_j 分别代表当前请求的时间和带宽资源需求量，该奖励使得算法更偏向于选择带宽资源需求更大的请求，从而提高带宽利用率。当提出的算法偏向选择时间较短和带宽较大的请求时，随着时间的推移，最大化累积奖励可以获得最大的带宽利用率。

本文提出的算法使用小批量梯度下降方法将神经网络训练为策略[10]，伪代码如图 3 所示。在每次训练迭代中，模拟 N 个回合探索使用当前策略可能动作的概率空间，并利用返回的奖励值改进策略。策略 $\pi_\theta(s,a)$ 定义为在状态 s 时选择动作 a 的概率，即 DRL 代理首先观察系统中所有 ONU（即租户）的带宽请求以及带宽资源使用情况，从动作空间中选择一定数量 ONU 的带宽请求进行处理。该算法记录每个回合中所有时间步长的状态、动作和奖励信息，即 $\{s,a,r\}$，并使用这些值计算每个回合的每个时间步长的折扣累积奖励 v，衡量策略的优劣并更新策略参数 θ，以便在未来得到更好的决策以提高带宽利用率。在每个回合的初始状态 s_1，算法根据策略选择动作 a_1，得到奖励 r_1，并更新策略参数 θ。然后，状态转移到 s_1，再次使用更新后的策略选择动作，迭代直至完成本回合训练。马尔可夫决策过程可以表示为 $\{s_1^i,a_1^i,r_1^i,\cdots,s_{L_i}^i,a_{L_i}^i,r_{L_i}^i\}\sim\pi_\theta$。

对于每次迭代：
  $\Delta\theta\leftarrow 0$
  对于每个请求集合：
    对于训练回合 $i\in[1,N]$：
      $\{s_1^i,a_1^i,r_1^i,\cdots,s_{L_i}^i,a_{L_i}^i,r_{L_i}^i\}\sim\pi_\theta$
    计算回报：$v_t^i=\sum_{s=t}^{L_i}\gamma^{s-t}r_s^i$
    对于每个时间步长 $t\in[1,L]$：
      计算基准值：$b_t=\frac{1}{N}\sum_{i=1}^{N}v_t^i$
      对于训练回合 $i\in[1,N]$：
        $\Delta\theta\leftarrow\Delta\theta+\alpha\nabla_\theta\log\pi_\theta(s_t^i,a_t^i)(v_t^i-b_t^i)$
  参数更新：$\theta\leftarrow\theta+\Delta\theta$

图 3 训练算法伪代码

在以上步骤中，策略参数 θ 通过梯度下降法更新：

$$\Delta\theta\leftarrow-\Delta\theta+\alpha\nabla_\theta\log\pi_\theta(s_t^i,a_t^i)(v_t^i-b_t^i) \quad (1)$$

其中，α 代表学习率，即每次移动的距离；$\nabla_\theta\log\pi_\theta(s_t^i,a_t^i)$ 指示了改变参数的方向，其中 $i\in[1,N]$ 代表当前训练的回合数，$t\in[1,L]$ 代表当前训练回合的时间步；$(v_t^i-b_t^i)$ 决定了当前修改步长的大小，v_t^i 为每个时间步长返回的奖励值（即最大化预期累积回报），根据 $\sum_{s=t}^{L_i}\gamma^{s-t}r_1^i$ 计算得到，$\gamma\in(0,1]$ 是未来回报的折扣因子，L_i 为迭代一个回合的长度；基准值 b_t^i 根据当前迭代所有的奖励值计算平均值得到，即 $b_t=\frac{1}{N}\sum_{i=1}^{N}v_t^i$。

3 算法仿真结果分析

3.1 基于 DRL 的调度算法仿真参数

为了验证提出算法的可行性，本文进行了仿真验证。仿真参数的配置具体如下：为了更好地模拟实际情况，假设各个租户的请求按照伯努利过程到达，并将请求的业务负载设置为 0.7。80%请求的时间需求在 1~3 个时间步长内均匀分布；剩余请求的时间需求在 10~15 个时间步长内均匀分布。另外，在总带宽为 10 MHz 的条件下，每个请求的带宽需求在 2.5~5 MHz 之间均匀分布。DRL 代理的集群在时间上持续 20 个时间步长，每次实验的时间为 50 个时间步长。代理从 M 个请求的集合中选择合适的请求进行分配。在每次训练迭代中，使用 50 个不同的请求集，动作空间设置为 M=8，对每个请求集并行运行 20 个蒙特卡洛模拟。本文使用 rmsprop 算法[11]更新策略网络参数，设置学习率为 0.001。

为了进一步验证算法特性，本文将基于 DRL 的在线带宽资源分配算法与 3 种基准启发式算法进行了比较：①随机选择（Random）算法，该算法随机选择等待的租户请求，每个租户请求的机会是相等的；②最短请求优先（SJF）算法，该算法在多个租户请求处于等待状态时，优先选择时间最短的请求；③资源打包（Packer）算法[12]，该算法根据请求需求和可用资源之间的顺序选择请求。

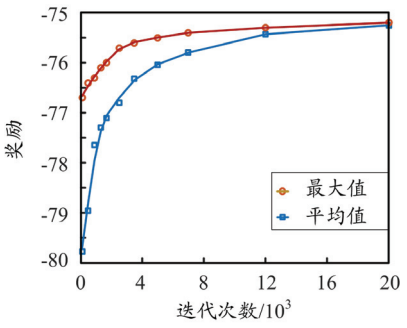
3.2 基于 DRL 的调度算法结果分析

基于 DRL 的算法需要通过训练不断优化策略网络参数。在迭代开始时，DRL 代理没有当前系统的任何先验知识，此时的策略与随机选择算法类似，因此该策略在训练迭代次数较少时并不会达到较优的效果。策略的输出是所有可能动作的概率分布，随着每一次迭代更新策略参数，向着奖励变大的方向前进。每次迭代的所有蒙特卡洛运行中奖励最大值和平均值如图 4(a)所示。由于在迭代开始时策略的奖励与随机选择算法类似，因此每次迭代的奖励平均值远低于当前迭代的最大值。如预期的一样，迭代中所有蒙特卡洛运行的奖励最大值和平均值都会随着迭代次数的增加而增加。当奖励平均值远低于最大值时，就意味着该策略需要进一步学习。如果某个操作路径的奖励比平均操作路径的奖励大得多，则可以通过调整策略参数来增加奖励。相反，当二者的差距缩小直至接近相

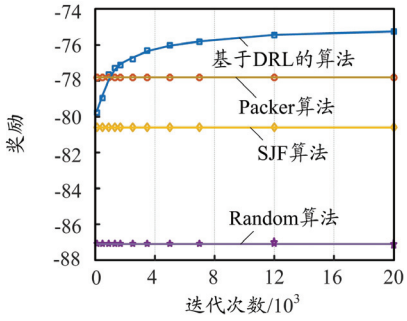
等时，如图中迭代次数达到 20000 次时，认为此模型收敛，训练完成。

各算法每次迭代奖励与训练迭代次数的关系如图 4(b)所示。由于基准启发式算法无法训练，因此结果随着迭代次数的增加奖励保持不变。而基于 DRL 的在线带宽资源分配算法的结果随着迭代次数的增加而增加，直至收敛趋于稳定。在迭代次数较低时奖励较低，如迭代 0 次的启动策略并不比 Packer 算法好，但是经过 4000 次迭代后，基于 DRL 的算法奖励超过 3 种对比算法。当迭代次数到达 20000 次时，基于 DRL 算法的奖励明显优于 3 种对比算法，且能保持稳定。

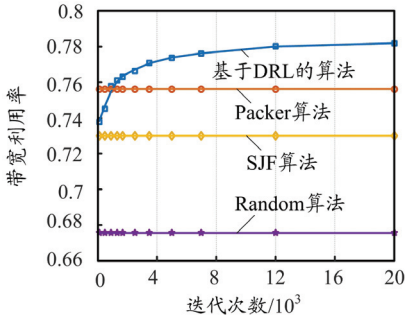
各算法每次迭代带宽利用率与训练迭代次数的关系如图 4(c)所示。可以看出，当训练迭代次数较少时，基于 DRL 的算法得到的带宽利用率低于 Packer 算法，略优于 SJF 算法。随着迭代次数的增加，基于 DRL 的算法得到的带宽利用率快速提高，而后趋于稳定。当带宽利用率快速提高时，表明当前模型仍有较大改进空间，迭代次数增加后，模型趋于收敛，模型参数趋于稳定，此时基于 DRL 算法的带宽利用率高于 3 种对比算法。这表明基于 DRL 的算法经过训练，能够有效提高系统的带宽利用率。在模型收敛和算法稳定后，相对于 Random、SJF 和 Packer 算法，使用 DRL 的算法的奖励和带宽利用率均提高分别达 11.85%、5.35%和 2.55%，表明该模型能够有效改善系统性能。



(a) 每次迭代的所有蒙特卡洛运行中奖励的最大值和平均值



(b) 各算法每次迭代奖励与训练迭代次数的关系



(c) 各算法每次迭代带宽利用率与训练迭代次数的关系

图 4 基于 DRL 的在线带宽分配算法与基准启发式算法性能比较

4 结束语

多租户共享网络资源的方案能有效缓解网络运营商的成本压力和投资风险。面对网络共享带来的资源分配问题，本文提出了针对多租户网络的基于 DRL 的在线带宽资源分配算法。仿真结果表明：随着训练迭代次数的增加，基于 DRL 的在线带宽资源分配算法带宽利用率不断提高直至模型收敛，比基准启发式算法的带宽资源利用率高。

参考文献：

[1] 成刚. 25G/50G PON 技术发展趋势[J]. 光通信技术，2020，44(1)：35–38.
[2] ANAND A, VECIANA G D, SHAKKOTTAI S. Joint scheduling of URLLC and eMBB traffic in 5G wireless networks [C]//IEEE INFOCOM 2018-IEEE Conference on Computer Communications, Honolulu, USA. New York: IEEE, 2018: 1970–1978.
[3] 蚁泽纯，何培森. 面向 5G 的 PON 承载技术与应用 [J]. 信息通信，2020(5)：248–249.
[4] 李承俊. 软件定义光接入网带宽资源管理策略研究[D]. 上海：上海交通大学，2017.
[5] SCHNEIR J R, XIONG Y. Cost analysis of network sharing in FTTH/PONs[J]. IEEE Communications Magazine, 2014, 52(8): 126–134.
[6] CICERI O J, ASTUDILLO C A, FONSECA N L S D. Dynamic bandwidth allocation with multi-ONU customer support for ethernet passive optical networks[C]//2018 IEEE Symposium on Computers and Communications (ISCC), June 25–28, 2018, Natal, Brazil. New York: IEEE, 2018: 230–235.
[7] LI C, GUO W, WANG W, et al. Bandwidth resource sharing on the XG-PON transmission convergence layer in a multi-operator scenario [J]. Journal of Optical Communications and Networking, 2016, 8(11): 835–843.
[8] MAO H, ALIZADEH M, MENACHE I, et al. Resource management with deep reinforcement learning [C]//15th ACM Workshop on Hot Topics in Networks, November 9–10, 2016, Atlanta, USA. New York: ACM, 2016: 50–56.
[9] HASTINGS W K. Monte Carlo sampling methods using markov chains and their applications[J]. Biometrika, 1970, 57(1): 97–109.
[10] RAZA M R, NATALINO C, HLEN P, et al. A slice admission policy based on reinforcement learning for a 5G flexible RAN[C]//2018 European Conference on Optical Communication (ECOC), September 23–27, 2018, Roma, Italy. Roma: ECOC, 2018: 1–3.
[11] HINTON G. Overview of mini-batch gradient descent[EB/OL]. [2020–10–30]. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
[12] ROBERT G, GANESH A, SRIKANTH K, et al. Multi-resource packing for cluster schedulers [C]//2014 ACM conference on SIGCOMM, August 17–22, 2014, Chicago, USA. New York: ACM, 2014: 455–466.