引用本文：王亚男，杨雪，庄浩涛，等. 基于强化学习的资源最优化逻辑拓扑映射算法[J]. 光通信技术，2020，44(6)：46-50.

基于强化学习的资源最优化逻辑拓扑映射算法

王亚男[1]，杨 雪[2]，庄浩涛[3]，朱 敏[2]，康 乐[2]，赵永利[3]

（1.中国电力科学研究院有限公司，北京 100192；2.国网四川省电力公司，成都 610041；
3.北京邮电大学 信息光子学与光通信国家重点实验室，北京 100110）

摘要：光传送网（OTN）中光节点波长复用/解复用器以及光开关矩阵可实现任意结构的逻辑拓扑在物理拓扑上的映射，不合理的映射方案将消耗额外端口资源。提出一种基于强化学习（RL）的逻辑拓扑最优化映射算法，将预处理后的拓扑状态和逻辑通道数据用于训练 RL 模型，以对逻辑通道进行全局波长资源分配，最终达到资源最优化目的。仿真结果表明：所提算法有效减小逻辑拓扑映射过程中的资源消耗，从而最小化网络部署成本。

关键词：光传送网；逻辑拓扑映射；强化学习；网络资源分配

中图分类号：TN915　**文献标识码**：A　**文章编号**：1002-5561(2020)06-0046-05

DOI：10.13921/j.cnki.issn1002-5561.2020.06.011　开放科学（资源服务）标识码（OSID）：

Algorithm of logical topology mapping for resource optimization based on reinforcement learning

WANG Ya'nan[1], YANG Xue[2], ZHUANG Haotao[3], ZHU Min[2], KANG Le[2], ZHAO Yongli[3]

(1. China Electric Power Research Institute, Beijing 100192, China; 2. State Grid Sichuan Electric Power Company, Chengdu 610041, China; 3. State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100110, China)

Abstract: The optical node wavelength multiplexer / demultiplexer and optical switch matrix in optical transport network (OTN) can map the logical topology of any structure on the physical topology, the unreasonable mapping scheme will consume additional port resources. A logic topology optimization mapping algorithm based on reinforcement learning (RL) is proposed. The preprocessed topological state and logical channel data are used to train the RL model, so as to allocate the global wavelength resources to the logical channel, and finally achieve the purpose of resource optimization. Simulation results show that the proposed algorithm can effectively reduce the resource consumption in the process of logical topology mapping, thus minimizing the cost of network deployment.

Key words: optical transport network; logical topology mapping; reinforcement learning; network resource assignment

0 引言

光传送网（OTN）具有带宽容量大、交换能力强等优点，是下一代骨干网络的主要基础设施。从信号承载的角度，OTN 体系架构可分为光层和电层。其中，光层是基于波长选择开关（WSS）的业务疏导体系，支持多方向、多速率的波长级颗粒的疏导；电层是基于光通路数据单元交换机制（ODU*k* Switch）的多向性电层疏导体系，支持 ODU0/1/2/3 多颗粒疏导，通过电层交叉实现子波长调度，解决波长冲突、传输限制和跨域互联等问题[1]。

OTN 可利用波长复用/解复用器以及波长路由技术（OMS）实现任意结构的逻辑拓扑在物理拓扑上的映射[2]。其电交换结构支持更细粒度的电层逻辑通道，即光层的一个波长可复用多条电层通道，从而提高频谱利用效率；当需要电处理时，同一波长上的所有逻辑通道也可以被解复用。与同步数字体系（SDH）分层网相比，OTN 能显著提高网络功能虚拟化（NFV）、物联网等高带宽应用的适应性和传输效率。其中，NFV 最

收稿日期：2019-12-10。

基金项目：国家电网科技项目“电力通信光传输网络分布式仿真及优化技术研究”（5442XX180003）资助。

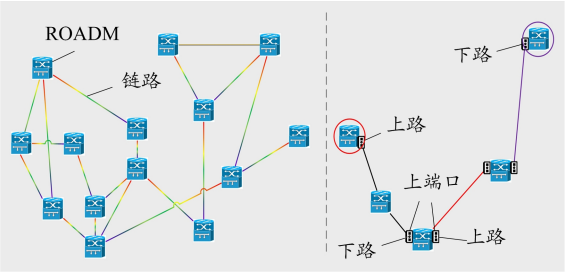
作者简介：王亚男（1992—），女，2014 年获得河北大学通信工程专业学士学位，2017 年获得北京交通大学通信与信息系统专业硕士学位。2017 年 8 月就职于中国电力科学研究院，当前研究方向包括电力通信网络仿真技术和光传输技术。

大的挑战之一是如何以最小代价且最高效率将虚拟网络拓扑(逻辑拓扑)映射到物理网络抽象出的拓扑上[3]。所谓物理拓扑是 OTN 光节点的物理连接关系,即光节点与光缆链路的集合,因而物理拓扑一般不随业务改变而改变。逻辑拓扑指的是节点之间业务的分布情况,由一系列端到端的光通道构成的拓扑结构,即在物理拓扑基础上负责分层路由的光层逻辑结构[2]。

在逻辑拓扑映射过程中,为了实现光/电转换,必须在交换结构上分配一个或多个光/电(O/E)端口(下文简称为端口),以便在光域和电域间转换信号。不合理的逻辑通道映射将消耗额外的端口,这将大大增加运营商的成本。优化的逻辑拓扑映射方案是解决 OTN 网络资源消耗最小化问题的关键。资源消耗主要表现在波长资源分配时新端口的引入,从而导致部署成本的增加。目前,虚拟网络映射问题已得到了广泛的研究[5-8],但针对 OTN 逻辑拓扑映射过程资源优化问题的研究还很少。文献[9]提出了基于拓扑排序的修正算法,但未将所涉及的网络拓扑状态作为网络资源最优化的影响因素;文献[10]利用最小 k-cut 问题形式设计了 2 种启发式算法,用于智能平衡电层资源和频谱资源使用率以达到资源消耗最小化目标,但未从网络状态出发考虑全局最优化。因此,为了综合考虑有限波长资源、端口数等条件对逻辑拓扑资源分配限制的问题,本文提出一种基于强化学习(RL)的逻辑拓扑最优化映射算法,以降低 OTN 中拓扑映射过程中额外端口引起的成本。

1 网络模型与评价指标

在进行映射算法的研究之前,先设定算法网络模型以及评价标准。在拓扑映射问题中,给定的 OTN 网络由 1 组 OTN 节点和 1 组 OTN 链接组成。本文可用无向图 $G=(V,E,W,C)$来表示该物理拓扑。其中,V 是物理节点的集合,E 是光纤链路的集合,W 是每条链路上可用波长资源的集合,C 是链路的资源容量权重值。共享 1 条光纤链路的光通道数不得超过 W。P_p 为所有物理路径的集合,$p_p(src,dst)\in P_p$ 表示从源节点 src 到目的节点 dst 的物理路径集合。P_l 为逻辑通道集合,$p_l(src,dst)\in P_l$ 表示从源节点 src 到目的节点 dst 的逻辑路径集合。本文假设初始状态下每条物理链路具有相同的波长资源,每条逻辑通道映射带宽相同且为每波长带宽的四分之一。物理拓扑态和逻辑通道态图如图 1 所示。其中,蓝色图标代表拓扑中的普通节点,红圈中的节点是源节点,紫圈中的节点是宿节点。链路中节点旁的长方框代表节点的端口,即代表此节点可以进行逻辑通道汇聚或作为中继节点等。



(a) 图像化的 OTN 拓扑态　　(b) 逻辑通道态

图 1 物理拓扑态和逻辑通道态

逻辑拓扑由一组全光的逻辑通道组成。一般地,业务流需“尽可能”地直接在光域上传送(“单跳”技术),若从一条光通道接入到另一条光通道时,需要引入端口(“多跳”技术),即从 src 到 dst 的逻辑通道可直接映射到单一全光通道。而当某段物理链路出现波长信道资源不足时,不可能在所有的源宿节点对之间都建立起光通路,则其逻辑通道需在不同光通道间经多跳转接来实现。结合“单跳”和“多跳”方式的网络结构可充分利用各种物理资源,尽可能避免端口的引入,在高效提供所需的逻辑通道连接的前提下达到虚拟网络部署成本最优化的目标。

2 基于 RL 的逻辑拓扑智能化映射算法

2.1 RL 基本理论

基于马尔科夫决策过程(MDP)[11]的 RL 是一种重要的机器学习算法[12],在智能控制及评估推荐等领域有许多应用。RL 算法的基本结构如图 2 所示。RL 过程包括 4 个要素:状态空间(State Space)、动作空间(Action Space)、状态转移概率(Transition Probability)分布和奖赏函数(Reward Function),记为$\langle S,A,P,R\rangle$。其中,$S_s=\{s_1,s_2,s_3,\cdots\}$,$A=\{a_1,a_2,a_3,\cdots\}$,$P^a_{ss_1}=P[S_{t+1}=s_1|S_t=s,A_t=a]$为 s 状态下执行 a 转换到 s_1 状态的概率,$R=r(s,a)$。RL 模型中学习体(Agent)以“试错”

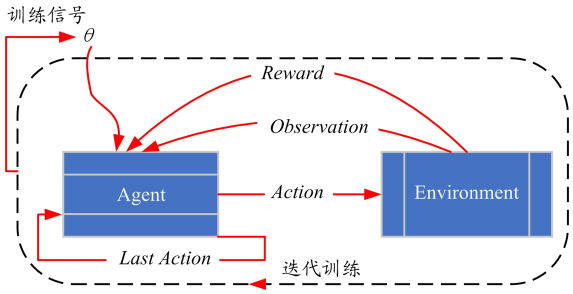


图 2 RL 算法基本结构

形式学习,在此过程中不断与环境(Environment)进行交互并获得 Reward 值。这个过程也促使 Agent 做出目标动作的趋势加强,从而获取到最优策略。在交互过程中,状态 s_T 下累积的 Reward 值设为从时刻 T 开始执行动作后所获得的 Reward 值,如式(1)所示。

$$R_T=\sum_{i=0}^{\infty}\gamma^i R_{T+i+1} \tag{1}$$

其中,$\gamma\in(0,1)$为折扣因子,用于调节当前 Reward 和长期 Reward 平衡性,取值大小与对长期 Reward 的重视程度成正相关。RL 中由 Environment 提供的强化信号会对 Agent 所发出动作的好坏做出评价(通常为标量信号),而非直接指导强化学习系统(RLS)如何做出正确的动作。由于每一次迭代训练过程中外部环境反馈的信息较少,RLS 须靠自身经历进行学习,在行动-评价的环境中获得知识,进而改进行动方案以适应环境。

RL 模型最终目标为学习获取最优策略 π,从而最大化累积 Reward 期望,即:

$$\pi(a|s)=\text{argmax}E[R] \tag{2}$$

RL 模型中引入 State Value 函数与 Action Value 函数以便对 State 和 Action 的优劣程度做出评估,表达式分别如下:

$$V_\pi(s)=E_\pi[R_{t+1}+\gamma V(S_{t+1})|S_t=s] \tag{3}$$

$$Q_\pi(s,a)=E_\pi\left[\sum_{i=0}^{\infty}\gamma^i R_{t+i+1}|S_t=s,A_t=a\right] \tag{4}$$

式(3)和式(4)的求解过程可理解为创建一张 State-Action 表[13],通过不断地循环迭代更新表中的值。该方法属于有模型法,空间复杂度为 $O(S_sA)$,在状态空间 S_s 比较大时其求解难度将急速上升。为了解决以上算法在离散行为空间上的局限,Deep Mind 提出了深度确定性策略梯度(DDPG)算法[14],借鉴了深度 Q 网络算法的深度神经网络、经验回放[15],并设置 target 网络使确定性策略梯度中的 Actor-Critic 算法更容易收敛。前人的研究表明,DDPG 算法在多种连续行为决策问题上表现良好。

2.2 端口最少化的逻辑拓扑映射算法

传统的启发式算法是通过选择一条逻辑通道并遍历其中每段链路来判断是否增加端口,因此只能保证当前通道所用端口数最少,即仅考虑局部最优,全局仍有很大优化空间。为了让算法达到全局最优效果,本文将全局物理拓扑状态(网络状态)作为 RL 模型训练的 Environment。通过不断与全局网络环境互动学习,RL 模型最终学到最佳策略,并做出全局最优决策。

2.2.1 拓扑数据预处理

本文先对逻辑通道状态和物理拓扑状态进行预处理,并将预处理后链路态和拓扑态作为 RL 模型的输入。拓扑态预处理过程主要包括逻辑通道属性的设定以及逻辑通道路由的计算。涉及到的属性包括链路重要度、源宿点一致性和链路带宽占用情况等。逻辑通道路由采用最短路径算法。根据通道属性的设定对链路进行分类和排序,并按序对逻辑通道进行映射。通过对逻辑通道映射请求次序进行排序,RL 模型的学习效率将一定程度上得到提升[16]。本文采用的排序规则是将链路重要度相同、同源同宿,且带宽占用情况相同的逻辑通道分为一类,重要度最高的逻辑通道优先进行映射;在链路重要度相同情况下,对相同源宿节点对的逻辑通道按照宽资源占用率由大到小的次序进行映射;链路重要度、占用带宽资源相同而源宿对不同的逻辑通道,按照数量由多到少的次序进行映射。

2.2.2 逻辑拓扑智能映射流程

逻辑拓扑映射问题主要包括:①将逻辑拓扑中的每一条 $p_l(src,dst)\in P_l$ 映射到一些现有的某条 p_p $(src,dst)\in P_p$,为 p_l 通道中每段链路分配波长资源;②为新请求的逻辑通道响应光路映射;③自动分配端口以支持某些逻辑通道所需的光/电转换。对于③,在以下状况下需在节点上加端口进行光/电转换:◆在源节点处;◆到宿节点处;◆在同一波长上带宽资源不足,需进行波长转换。另外,当物理链路通信环境不佳需对通道中信号进行补偿时,也需要添加端口。

本文算法中引入的 RL 方法为 DDPG,其由 Actor 网络和 Critic 网络组成[14],如图 3 所示。它们之间的协同流程为:在从 Environment 中获取 State 的观测量 S,Agent 根据 Actor 网络做出决策 Action,Environment对此 Action 做出反应,也即给出 Reward 值及新状态值 S_,然后根据 Reward 值更新 Critic 网络,最后DDPG模型依据 Critic 所建议的方向进行 Actor 网络的更新。其

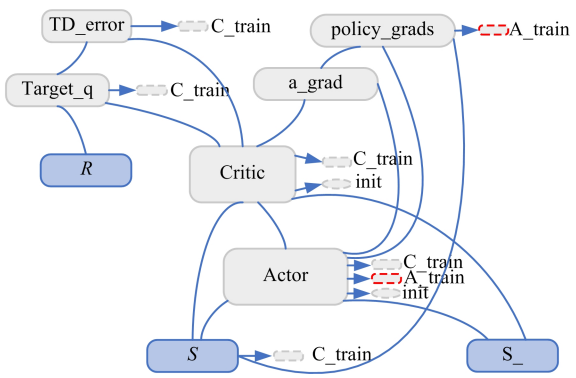


图 3 DDPG 主结构

中,a_grad 表示动作梯度,来自 Critic 的 a_grad 表示此次 Actor 最优 Action 移动方向,来自 Actor 的 a_grad表示 Actor 自身参数修改方式,使得 Actor 更有可能采取这个 Action。两者贡献的 a_grad 则表示 Actor 要朝着更有可能获取大 Reward 的方向修改动作参数了。Target_q 表示目标 Q 网络,其依据下一状态,用 Actor 来选择动作。TD_error 表示误差,在计算出误差后进行反向误差传递。C _train 表示 Critic 网络的训练过程;A _train 表示 Critic 网络的训练过程;init 表示整个算法的初始化过程;蓝色箭头表示左边实线框的数据输入到右边虚线框中。通过循环迭代,直至最终训练出最优的 Actor 网络。

本文算法首先初始化网络状态并将预处理好的物理拓扑态和逻辑通道态作为 DDPG 模型的输入,物理拓扑的状态 State 设定为 RL 中的 Environment。RL 模型的输出为一个 Action 和执行此 Action 所返回的一个价值,分别对应 Actor 网络的输出以及 Critic 网络的输出。本算法中 Action 的设置为:

①添加端口,分配资源;②不加端口,分配资源;③不采取任何动作;④分配资源或添加端口。

Actor 负责选择可用的波长或不执行任何请求;而 Critic 则负责评估光网络的状态,它可以帮助 Actor 提取物理拓扑的基本特征。

由于 RL 算法是一种非监督式学习,即 Agent 在学习训练过程中输入的训练数据并没有监督学习算法中标有正确的标签,因而 RL 的训练方向具有随机性。为了评估 Agent 的动作选择的随机程度,本文引入了动作熵 $H(X)$ 作为评估指标。动作熵直接与一个 Agent 在给定策略中所采取的不可预测性有关,其方程为:

$$H(X)=E_X[I(x)]=-\sum_{x\in X}p(x)\log p(x) \tag{5}$$

动作熵值越低,Agent 的动作选择随机程度越低,即 Agent 所作决策的方向性越强。在 RL 中,Agent 的目标一般被体现为长期总 Reward 值的最大化。这意味着学习过程中的 Agent 将逐渐采取特定的 Actions 序列并排除其它可能的 Actions 序列以实现 Reward 值的最大化。这样将会导致 Actions 选择政策熵逐渐减小。

算法目标是在确保所请求的每条逻辑通道映射成功的同时,尽可能减少 O/E 端口的数量。对应地,本文所提算法是在保障所有逻辑通道成功映射的情况下进行端口数的优化。为此,算法在执行的过程中需为 RL 模型设置适当的 Reward 机制,以指导算法在逻辑拓扑映射期间减少端口数量。例如,在为通道分配波长资源时不增加新端口的 Reward 值大于增加新端口的 Reward 值。通过训练,RL 模型将趋向于执行正确行动以获得更高的回报。对于不同的目标,Reward 机制的设计也应该有所不同。在不同条件下对不同行为的 Reward 值如表 1 所示。

表 1 RL 模型 Reward 值设置情况

物理拓扑状态	采取的动作	Reward 值
所达节点有可用波长资源	无动作	−1
所达节点有可用波长资源	添加端口,分配资源	1
所达节点有可用波长资源	不加端口,分配资源	3
所达节点无可用波长资源	无动作	0
所达节点无可用波长资源	分配资源或添加端口	−3

3 仿真设置及结果分析

3.1 仿真设置

本文设置 9 节点 16 链路的矩形拓扑作为仿真拓扑,基于 DDPG 的逻辑拓扑映射智能算法流程如图 4 所示。为降低仿真复杂度,我们假设 3×3 网络拓扑中每段链路均有 5 个可用波长。图 4 中紫色部分是 RL 算法中的 Environment,具体设定为一个 OTN 网络环境,实现了图 2 中的 Environment,其中左边是一个包含 9 节点 16 链路的矩形拓扑,右边是 ROADM 节点上、下路的简单说明;下半部分是 DDPG 算法的 2 个主要模块:Actor 网络和 Critic 网络,这是图 2 中 Agent 的具体实现。RL 模型的 Environment 仿真采用基于

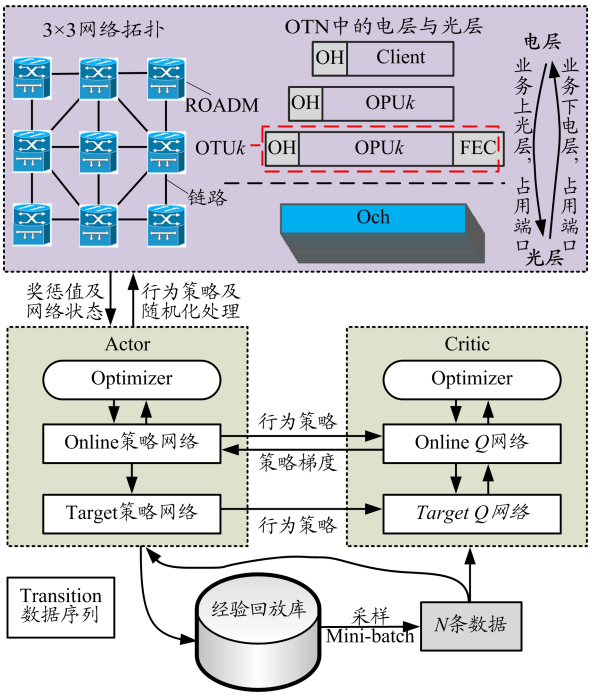


图 4 基于 DDPG 的逻辑拓扑映射智能算法流程

Networkx 的 python 库。Networkx 是一个机器学习中常用的复杂网络建模工具,其中涵盖了网络仿真中常用的图算法以及复杂网络分析方法,例如路由最短路径算法,便于对复杂网络进行仿真建模等工作。RL 模型中使用 Root Mean Square Prop optimizer 作为梯度下降过程的优化器,设置其基本学习率为 8×10^{-6}。请求映射的逻辑拓扑随机生成,其通道总数设定为 50。逻辑拓扑中每一条通道的路由使用 K 路最短路径算法(KSP)算法来计算。本文结合 KSP 算法与 First-Fit(FF)算法(KSP+FF)[17],作为本文算法对比的基准。

3.2 结果验证与分析

训练期间端口数量的变化情况如图 5 所示。蓝色实线为本文算法的仿真结果,记录了 RL 模型训练每隔 1000 步所添加的端口数量;红色虚线表示 KSP+FF 算法。KSP+FF 算法算出的端口添加数在一定范围内均衡波动,因而直接采用其平均值进行比较。由于带宽资源充足,所以不考虑业务阻塞问题。

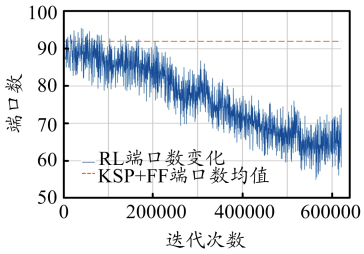


图 5 不同算法下端口数变化曲线

从图 5 中可见大约迭代 210000 次后,蓝色实线稳定低于红色虚线,也即 RL 模型在经过一定训练后其性能超过了 KSP+FF 算法。RL 模型中 Agent 的动作熵变化图如图 6 所示。可以看出,大约经过 500000 次迭代过后,Agent 的动作选择熵值出现了收敛趋势,这表明 RL 模型已经学习到了一些决策规则,也即模型具备做出更好决策的能力。

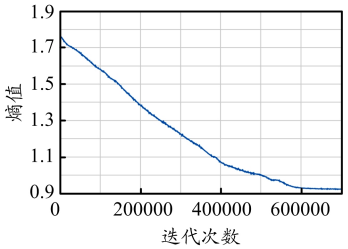


图 6 本文算法中 RL 模型的动作熵值变化图

4 结束语

为了减少 OTN 中逻辑拓扑映射过程额外端口的数量,本文提出了一种经济高效的逻辑拓扑智能化映射算法。仿真结果显示:与常规启发式算法(KSP+FF)相比,本文算法能够根据全局网络环境特性做出更好的映射决策,以避免额外端口的引入。本文拓扑映射算法中通道资源分配采用 RL 算法,而逻辑通道算路过程采用 KSP 算法,并未完全实现逻辑拓扑映射过程的智能化。后续工作将着重研究基于 RL 的算路与资源分配算法,实现逻辑拓扑映射过程的完全智能化。

参考文献:

[1] 刘玉洁,肖峻,丁炽武,等. OTN 最新研究进展及关键技术[J]. 光通信技术,2009,33(6):5-7.

[2] 张杰,顾畹仪,李国瑞,等. 光传送网的最优虚拓扑设计原理[J]. 现代有线传输,1999(4):30-33.

[3] YU M, YI Y, REXFORD J, et al. Rethinking virtual network embedding: substrate support for path splitting and migration [J]. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 17–29.

[4] FISCHER A, BOTERO J F, BECK M T, et al. Virtual Network Embedding: A Survey [J]. IEEE Communications Surveys & Tutorials, 2013, 15(4): 1888–1906.

[5] FENGQING L, QINGJI Z, XU Z, et al. Virtual Topology Reconfiguration in WDM Optical Networks[J]. ACTA PHOTONICA SINICA, 2003, 32(10): 1175–1180.

[6] CHOWDHURY M, RAHMAN M R, BOUTABA R. Virtual network embedding algorithms with coordinated node and link mapping [J]. IEEE/ACM Transactions on networking, 2011, 20(1): 206–219.

[7] CHENG X, SU S, ZHANG Z, et al. Virtual network embedding through topology-aware node ranking [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(2): 38–47.

[8] CHOWDHURY N M, RAHMAN M R, BOUTABA R. Virtual network embedding with coordinated node and link mapping [C]// IEEE INFOCOM 2009, Apr 19–25, 2009, Riode Janeiro, Brazil. Piscataway: IEEE, 2009: 783–791.

[9] YOU X, WANG X, CHEN A, et al. A coordinated algorithm with resource evaluation for service function chain allocation [C]// 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Oct. 8–10, 2016, Atlanta, USA. Piscataway: IEEE, 2016: 45–49.

[10] WANG Y, LU P, LU W, et al. Cost-efficient virtual network function graph (vNFG) provisioning in multidomain elastic optical networks [J]. Journal of Lightwave Technology, 2017, 35(13): 2712–2723.

[11] WHITE, DOUGLAS J. Dynamic programming, Markov chains, and the method of successive approximations [J]. Journal of Mathematical Analysis and Applications, 1963, 6(3): 373–376.

[12] SUTTON R S, BARTO A G. Introduction to reinforcement learning [M]. Cambridge: MIT press, 1998.

[13] WATKINS C J C H, DAYAN P. Technical Note: Q-Learning [J]. Machine Learning, 1992, 8(3-4): 279–292.

[14] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint: 1509.02971, 2015.

[15] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518 (7540): 529–533.

[16] 何源,侯韶华. WDM 光网络虚拟映射协同 RWA 算法研究[J]. 计算机技术与发展,2017,27(3):122-125.

[17] MOKHTAR A, AZIZOGLU M. Adaptive wavelength routing in all-optical networks [J]. IEEE/ACM Transactions on Networking, 1998, 6(2):197–206.