Multi-Agent Deep Reinforcement Learning in Cognitive Inter-Domain Networking with Multi-Broker Orchestration

Xiaoliang Chen[1], Baojia Li[2], Roberto Proietti[1], Zuqing Zhu[2], S. J. Ben Yoo[1]

1. University of California, Davis, Davis, CA 95616, USA, Email: sbyoo@ucdavis.edu, xlichen@ucdavis.edu
2. University of Science and Technology of China, Hefei, Anhui 230027, China, Email: zqzhu@ieee.org

Abstract: This paper proposes, for the first time, a cognitive inter-domain networking framework with multi-broker orchestration and multi-agent deep reinforcement learning for multi-domain optical networks. Simulation results show $> 17\%$ blocking reduction compared to the baselines.

OCIS codes: (060.1155) All-optical networks; (060.4251) Networks, assignment and routing algorithms.

1. Introduction

The rapidly expanding cloud applications demand for today's Internet backbone of multi-domain optical networks (MD-ONs) to support dynamic and high-capacity end-to-end services. Due to the autonomy constraints and the delicate resource allocation mechanisms in MD-ONs, realizing efficient inter-domain networking and coordinating operations of multiple domains are becoming very challenging. Previous works have proposed several flat or hierarchical architectures for the network control and management (NC&M) of MD-ONs [1, 2]. While flat architectures may lead to poor resource efficiency due to lack of coordination, hierarchical architectures suffer scalability and autonomy issues. Therefore, we envision more scalable and efficient inter-domain networking based on a multi-broker framework where broker agents participate in multi-domain service provisioning due to market-driven incentives [3].

Service provisioning in multi-broker orchestrated MD-ONs essentially can be modeled as incomplete information games (either cooperative or noncooperative) with multiple players, dynamic environments and huge yet complex strategy spaces [3, 4]. Optimizing the service provisioning strategies with traditional game-theoretic approaches, e.g., analyzing the Nash equilibrium, thus becomes intractable for such scenarios. On the other hand, recent breakthroughs in deep reinforcement learning (DRL) have demonstrated self-learning capabilities and exceptional performance of DRL for online control problems [5, 6]. Lately, DRL has been extended to the multi-agent learning domain (leading to multi-agent DRL, MADRL) to tackle multi-agent collaboration or competition tasks with complex system states [7].

In this paper, we leverage MADRL and propose a cognitive inter-domain networking framework for multi-broker orchestrated MD-ONs. The proposed framework allows broker agents to parameterize service provisioning policies from high-dimensional network states with deep neural networks (DNNs) and learn the optimal policies asynchronously through dynamic network operations. An advantage actor-critic (A2C) based approach is designed for the training of the agents. Evaluation results show that with MADRL, brokers can learn successful policies and reach a point where the interests of brokers (i.e., reward) and domain managers (i.e., inter-domain throughput) are jointly optimized.

2. Cognitive Inter-Domain Networking Framework

Fig. 1(a) depicts the architecture of multi-broker orchestrated multi-domain optical networks. A broker plane consisting of multiple incentive-driven brokers is introduced as a new NC&M hierarchy for coordinating the operations of domain managers (DMs) and assisting inter-domain networking. While DMs operate their domains autonomously, they can subscribe to multiple brokers for diversified and better services. Thus, brokers and DMs work together to constitute an inter-domain service provisioning market, where brokers may cooperate or compete freely for inter-domain service requests from DMs motivated by incentives such as revenue, reputation, and etc. Quality of service and domain autonomy are achieved through service level agreements (SLAs) between brokers and DMs. By signing different SLAs, DMs may advertise different degrees of intra-domain information (i.e., domain abstractions [3]), making a tradeoff between domain privacy and attainable performance.

We design the service provisioning operations of the broker plane with MADRL to realize cognitive inter-domain networking, and Fig. 1(b) shows the schematic of the proposed framework. Basically, each broker serves as an autonomous learning agent and attempts to learn the optimal service provisioning policies from its experiences of dynamic multi-domain operations, i.e., interactions with DMs and peer brokers. The DRL agent is the brain of each broker. Upon servicing each inter-domain request, it observes and analyzes the current network state and generates a provisioning policy with deep learning models to assist the service provisioning manager taking an action (i.e., determining a service scheme). Here, the network state includes *(i)* the pending service request, *(ii)* the information of in-service requests from the traffic engineering database, *(iii)* the multi-domain abstractions reported by DMs, as well as *(iv)* the observations and information exchanges from peer brokers. DMs then try to set up the inter-domain service

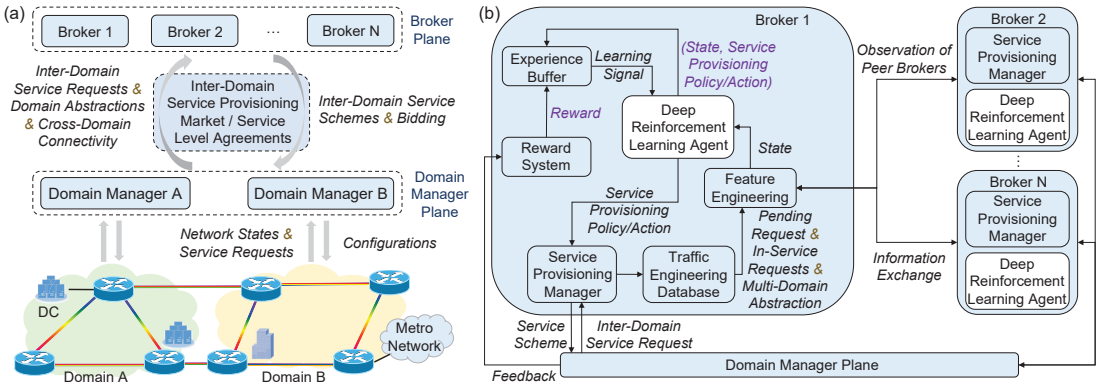

Fig. 1. (a) Multi-broker orchestration architecture and (b) principle of inter-domain networking with MADRL.

with the recommended service scheme and return feedback, indicating the performance of the service scheme (for instance, whether the request has been successfully provisioned, the resource costs and so forth). The reward system translates the feedback into an immediate reward according to the objective of the broker. The reward, together with the observed state and the action being taken are pushed in the experience buffer, which in turn generates and passes a learning signal to the DRL agent. The DRL agent is continuously trained with the learning signals to reinforce advantageous actions and approach the optimal provisioning policies. Note that, since the multiple agents share the same multi-domain infrastructure, the policy of each of them is also part of the environment of the others. Therefore, each agent has to learn not only the rule of the substrate network but also how to adapt to the evolutionary policies of its opponents, making MADRL much more challenging than single-agent learning scenarios.

3. MADRL Algorithm Design

We design an MADRL instance for the problem of inter-domain routing and spectrum assignment (RSA) [1]. Let $G = \{G_n, 1 \le n \le N\}$ represent a multi-domain optical network with N domains. Each of the brokers is subscribed by a set of clients spreading throughout G. For a broker i, an inter-domain lightpath request arriving at step t is denoted as $\chi_i^t(o,d,b,T)$, where o and d are the origin and destination nodes, b is the demanded number of frequency slots (FS's) and T is the service duration. To service χ_i^t, broker i constructs a multi-domain virtual topology (VT) G_i' by abstracting each G_n as consisting of the domain edge nodes inter-connected through a virtual node. We assume that brokers can access the detailed states of inter-domain links, e.g., spectrum utilization, length and etc. Figs. 2(a) and (b) show an example of domain abstraction. With the above preliminaries, the MADRL design is detailed as follows.

Objective & Reward: The objective of each broker is to maximize the long-term throughput for its clients' services. Therefore, we set the immediate reward r_i^t as 1 if χ_i^t is successfully serviced, otherwise, $r_i^t = -1$, and set the learning target of each DRL agent as maximizing the total discounted reward R_i^t collected within an episode, i.e., $R_i^t = r_i^t + \gamma r_i^{t+1} + \gamma^2 r_i^{t+2} + \cdots, 0 < \gamma < 1$. Here, to make R_i^t be finite and thus a valid learning target, we define an episode as the servicing of M inter-domain requests.

State: The network state s_i^t is defined as a $1 \times (2|G'| + 2 + 3k)$ array, where $|G_i'|$ represents the number of nodes in G_i' and k is the number of candidate routing paths broker i calculates (with G_i') for χ_i^t. Specifically, s_i^t contains *(i)* o and d in the one-hot form ($1 \times |G'|$ each), *(ii)* b, *(iii)* a counter c indicating the number of requests already being serviced in the current episode, and *(iv)* the number, the average and maximum sizes of available FS blocks on each of the candidate paths ($1 \times 3k$). We normalize all the fields in s_i^t before inputting it to the DRL agent. Meanwhile, we do not consider the communications among brokers.

Action: The set of actions A_i that broker i can take is selecting from one of the k candidate paths. Then, the related DMs perform spectrum allocation on the selected path domain-by-domain according to the procedures discussed in [1].

DRL agent: We employ two DNNs for each DRL agent to parameterize the provisioning policy $\pi_i^t(s_i^t, A_i)$ (i.e., the probability distributions for action selection) and the value (i.e., the estimation of R_i^t, denoted as $\hat{R}_i^t$), respectively. Let $f_{\theta_i^p}(\cdot)$ and $f_{\theta_i^v}(\cdot)$ denote the policy and value DNNs of broker i, where θ_i^p and θ_i^v are the sets of parameters, we have $\pi_i^t(s_i^t, A_i) = f_{\theta_i^p}(s_i^t, A_i)$ and $\hat{R}_i^t = f_{\theta_i^v}(s_i^t)$.

Training: We design the training scheme for MADRL based on the A2C algorithm derived from [8]. Recall that every service provisioning instance (s_i^t, a_i^t, r_i^t) of broker i is stored in its experience buffer. The DRL agent triggers a training process each time an episode terminates. In particular, the agent first calculates R_i^t for each instance and obtains the advantage of a_i^t as $\delta_i^t = R_i^t - f_{\theta_i^v}(s_i^t)$. The advantage indicates how much better than expected of taking a_i^t. Then, the gradients for the policy and value DNNs can be calculated with the following loss functions,

$$L^p(\theta_i^p) = -\sum_t \delta_i^t \log f_{\theta_i^p}(s_i^t, a_i^t) - \alpha \sum_t \sum_{a \in A_i} f_{\theta_i^p}(s_i^t, a) \log f_{\theta_i^p}(s_i^t, a), \qquad L^v(\theta_i^v) = \sum_t \left(f_{\theta_i^v}(s_i^t) - R_i^t \right)^2. \tag{1}$$

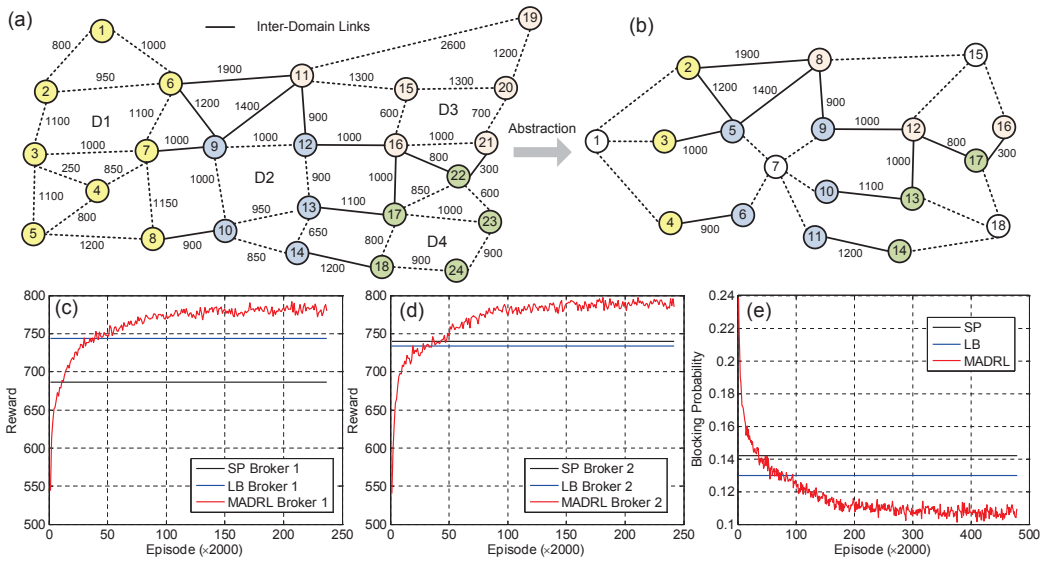

Fig. 2. (a) 4-domain substrate topology, (b) virtual topology abstracted by brokers, (c)-(d) comparison between MADRL and baselines: (c) reward of broker 1, (d) reward of broker 2 and (e) overall blocking probability of inter-domain requests.

The rationale of Eq. 1 is that by minimizing $L^p(\theta_i^p)$, we reinforce (i.e., increase the probabilities of) actions with larger advantages and maximize the entropy of the policies as well to encourage exploration. Finally, the agent applies the gradients to update θ_i^p and θ_i^v and empty the experience buffer for the next episode.

4. Performance Evaluation

We evaluate the performance of the proposed cognitive inter-domain networking framework with a 4-domain substrate topology (Fig. 2(a)). We implement two brokers, which abstract the substrate topology as the virtual topology shown in Fig. 2(b). Each link can accommodate 100 FS's. Each domain edge node has 10 optical-electrical-optical converters. The dynamic lightpath requests are generated according to a Poisson process, with the bandwidth demands uniformly distributed within $[2,15]$ FS's. We assume that requests originating from nodes $[3,4,5,7,8,11,12,15,17,18,23,24]$ are serviced by broker 1, while the rest ones are handled by broker 2. A fully-connected DNN with 8 hidden layers (128 neurons each) and activation function of *ELU* is used for both DRL agents. M, γ, α and learning rate are set as 30, 0.95, 0.1 and 10^{-5}, respectively. We chose the *shortest path routing* (SP) and *load balanced routing* (LB) algorithms as the baselines, with each broker selecting paths with the least hops (SP) or the largest amounts of available FS's (LB). The DMs apply the k-shortest path routing and first-fit spectrum assignment scheme for intra-domain provisioning.

Figs. 2(c) and (d) show the evolutions of the brokers' rewards during training, where each data point is an average of the rewards from 2000 episodes. We can see that at the beginning, the rewards of both brokers are much lower than those of the baselines due to the random policies they apply (the DNNs are randomly initiated). However, with MADRL, the agents quickly learn correct policies and their rewards increase logarithmically. After training through 10^5 episodes, MADRL can beat both of the baselines. Then, the performance of MADRL keeps improving steadily, indicating that the agents are continuously learning stronger policies. Eventually, with MADRL, brokers 1 and 2 can achieve $\sim 5.1\%$ and $\sim 6.7\%$ higher rewards respectively, comparing with applying the best baseline schemes. Fig. 2(e) plots the evolution of the blocking probability of overall inter-domain lightpath requests, showing a similar trend with those of the evolutions of rewards. After 8×10^5 episodes, MADRL achieves $\sim 17.7\%$ blocking reduction compared with the LB scheme. Overall, the results demonstrate that MADRL enables brokers to adapt to each other intelligently, leading to a point where interests of brokers and DMs are jointly optimized.

5. Conclusion

This paper demonstrates that MADRL very effectively facilitates enhanced performance of multi-broker orchestrated MD-ONs by enabling brokers to learn correct provisioning policies from dynamic inter-domain operations.

References

[1] Z. Zhu *et al.*, *J. Lightw. Technol.*, vol. 33, pp. 1508-1514, Apr. 2015.
[2] X. Chen *et al.*, *IEEE Commun. Mag.*, vol. 56, pp. 152-158, Aug. 2018.
[3] X. Chen *et al.*, *J. Lightw. Technol.*, vol. 34, pp. 3867-3876, Aug. 2016.
[4] L. Sun *et al.*, *J. Lightw. Technol.*, vol. 35, pp. 3722-3733, Sept. 2017.
[5] V. Mnih *et al.*, *Nature*, vol. 518, pp. 529-533, 2015.
[6] Z. Xu *et al.*, arXiv preprint arXiv:1801.05757, 2018.
[7] R. Lowe *et al.*, in Proc. of *NIPS*, pp. 6379-6390, 2017.
[8] V. Mnih *et al.*, in Proc. of *Int. Conf. Mach. Learn.*, vol. 48, pp. 1928-1937, 2016.

Acknowledgements: This work was supported in part by DOE DE-SC0016700