Leveraging Deep Learning to Achieve Efficient Resource Allocation with Traffic Evaluation in Datacenter Optical Networks

Ao Yu[1], Hui Yang[1], Wei Bai[1], Linkuan He[1], Hongyun Xiao[2], Jie Zhang[1]
(1) State Key Laboratory of information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, China;
(2) ZTE Corporation, Shenzhen, China, 518057
Email address: yuao@bupt.edu.cn

Abstract: This paper first presents a deep learning-based resource allocation strategy supported by global evaluate factor in intra-datacenter optical networks. Numerical results show the proposed strategy improves traffic prediction accuracy and has superior performance.

OCIS codes: (060.4250) Networks; (060.4256) Networks, network optimization; (060.4510) Optical Communication

1. Introduction

Rapid growths of high-bandwidth datacenter applications, such as artificial intelligence, live video streaming and autonomous vehicles, is driving the demand for high-efficient resource utilization in datacenter optical interconnection. Many studies use traditional machine learning algorithms (e.g., support vector machine (SVM), neural networks (NN)) to achieve traffic prediction as a guidance for resource assignment [1]. The machine learning algorithms mentioned above train a simple traffic prediction model by using existing data, and then use the model to predict the future traffic flow. However, the accuracy of the predicted results based on the conventional algorithms cannot be guaranteed due to the change of the bandwidth explosion and service diversity [2]. Recently, deep learning (DL) has drawn a great deal of academic and industrial interest since it can discover deep connections in the data, which can achieve accurate prediction in the complex networks [3, 4]. Therefore, it is necessary to apply deep learning technique to spectrum resource usage in datacenter optical networks.

To improve the accuracy of the prediction, deep learning is first introduced in optical resource allocation in intra-datacenter optical networks. In this paper, we propose a deep neural network-based prediction strategy (DNN-PS) with deep learning-based traffic evaluate in datacenter optical networks. Based on DNN-PS, a deep learning-based resource allocation algorithm (DL-RA) is proposed. DL-RA can provide high-reliable resource allocation for datacenter with high-accuracy traffic prediction, which can improve the resource utilization of high-bandwidth and multi-varieties traffic in which a evaluate factor is used to keep track of anomaly candidates. The experiments and simulations demonstrate that the proposed strategy can dramatically improve the resource utilization and greatly decrease blocking probability while improving the prediction accuracy.

2. Deep Neural Network-based Prediction Strategy (DNN-PS)

Before presenting the proposed algorithm, we propose DNN-PS which is brief shown in Fig. 1. We take advantage of deep learning architectures relating to pattern recognition for the resource allocation in datacenter optical

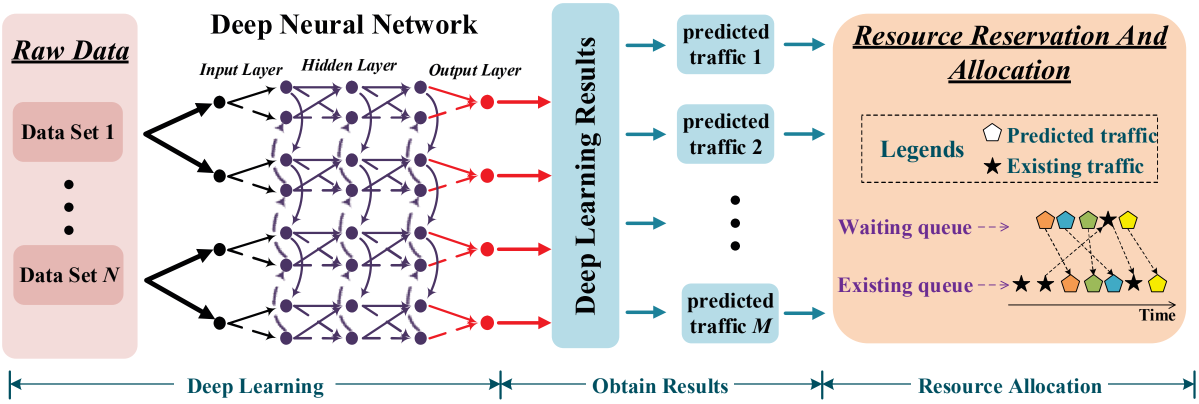


Fig.1 Logic sketch of DNN-PS.

978-1-943580-38-5 © 2018 Optical Society of America

networks. DNN are able to identify and characterize the complex structural characteristics in huge amounts of raw data, and these internal relations are not easily detected by ordinary machine learning methods. To build a more accurate prediction model, we built a database containing tens of millions of datacenter traffic information and trained it for several weeks with supervised learning techniques. The traffic data are collected every 20s from over 200 servers, which are deployed in State Key Laboratory datacenter in Beijing China. We adjusts the weights and bias in DNN during the training process to minimize the objective function. After obtaining a satisfactory prediction model, we can get the prediction results by entering new data. Then the prediction results will be imported into the DL-RA algorithm to allocate resources for the new arrival traffic. In the heart of DL-RA is to arrange the predicted traffic and exiting traffic in a high-efficiency way, in order to maximize the utilization of resources and minimize the adverse impact on traffic.

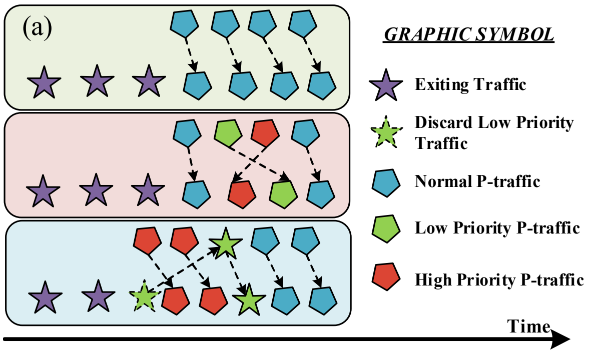
3. Deep Learning-based Resource Allocation Algorithm

Based on DNN-PS described above, we propose a deep learning-based resource allocation algorithm and a evaluate factor for the allocated results.

The algorithm is to use deep learning to estimate future traffic behavior by prediction, and to use such results to estimate the amount of future network resources that will be required by the considered traffic. In order to evaluate the validity of the prediction results from the perspective of accuracy and resources, we introduce a global evaluate factor α . The normalization factor contains the prediction and resource parameters. For prediction parameters, we characterize the degree of deviation between predicted traffic arrival time and actual arrival time by defining correlation coefficients. While for the resource parameters, we use integral to describe resource consumption in the global perspective. T_{pj} represents the arrival time of *j*th predicted traffic (P-traffic), μ_p is the value center of T_p at different amounts of arrival time, t' represents the predicted arrival time, while t represents the actual arrival time. $R_{pj}(t)$ represents the resources required for the *j*th arrival predicted traffic at t time, $R(t)$ represents the total amount of resources at t time in data center. The specific representation of resources as in our prior work [5]. In addition, M represents the total amount of traffic that will arrive during time (t_0,t_N) , in which N represents the arrival time of *M*th traffic. Therefore, the global evaluate factor α meets formula (1) as follows, where β is the adjustable weight between the traffic and resource parameters with different user requirements.

$$\alpha=\frac{E\left[T_{pj}^{2}(t\cdot t')\right]-\mu_{pj}(t)\mu_{pj}(t')}{\sqrt{D\left[T_{pj}(t)\right]}\cdot\sqrt{D\left[T_{pj}(t')\right]}}\beta+\frac{\sum_{j=1}^{M}\int_{t_0}^{t_N}R_{pj}(t)dt}{M\int_{t_0}^{t_N}R(t)dt}(1-\beta),\;\; E\left[T_{pj}^{2}(t\cdot t')\right]-\mu_{pj}(t)\mu_{pj}(t')>0 \tag{1}$$

This evaluate factor takes into account both the accuracy of traffic prediction and global resource utilization, which can illustrate the expected traffic arrival and the resources needs to be reserved.



Algorithms 1：DL-RA (b)

```
Input: training results $T_{pj}^{'}$ ;  expansion parameter Priority = k
for each $T_{pj}$
    if $R_p \le R_t$ then
    else if there will be no low priority traffic in queue
        else if $k = low$ then
            T will be added to traffic queue
            else if $\alpha > threshold$ then
                Allocate resources after discarding
                if $R_p \le R_t$ after resource arrangement then
                    Allocate resources after arrangement
```

Fig.2 (a) Schematic of traffic queue reordering in DL-RA. (b) The pseudocode of DL-RA algorithm.

For each predicted traffic T_{pj} , evaluate its priority and calculate the required resources. Controllers obtain the current resource usage status and the priority of existing traffic in datacenter. If there is not enough consecutive resources are found to be allocated when T_{pj} arrives, then find whether there is any low priority traffic in queue. If the evaluate factor α is more than the threshold, discarding the low-priority traffic based on the current distribution of resource fragments and the priority of traffic as shown in Fig. 2(a). Perform the update process when finish the above work, update the traffic queue depending on the different state of the traffic processing. Thus, the main information required for resource allocation could be obtained, and the proposed algorithm is described in detail as shown in Fig 2 (b).

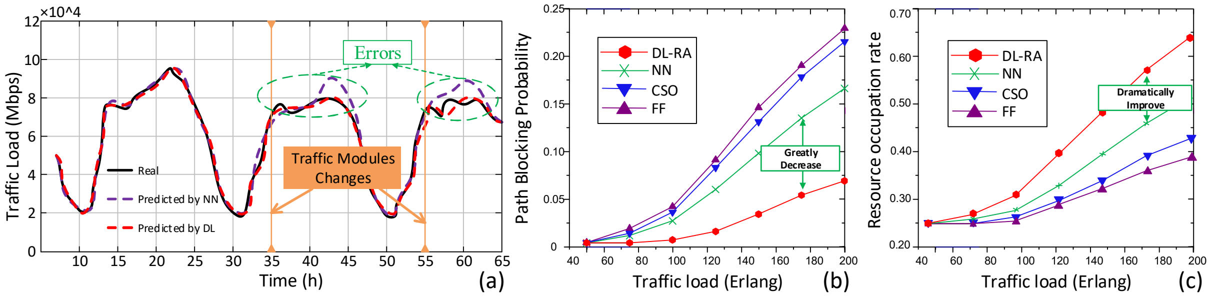


Fig.3 (a) Comparison among the real traffic load, the deep learning-based prediction and the neural network-based prediction in State Key Laboratory datacenter in Beijing China, during June 4, 5, and 6, 2017. Comparisons of (b) path blocking probability, (c) resource occupation rate among different resource allocation strategies.

4. Experimental Assessment and Results Analysis

In this section, we present results based on DNN-PS. Our goal is to demonstrate the accuracy and the network performance of DL-RA in different network environments. Our testbed is a multi-core server with 12 physical 2.90GHz CPU cores and 80GB RAM. The server runs Linux 2.6.32. We compile our code using GCC 4.6 with the -O3 option [6-8]. We run our experiments on real packet header information from the State Key Laboratory datacenter in Beijing China in June 2017. In order to fully reflect the changes in the network environment, we select the most heavy-loaded 6 hours of traces for our experiments. The traces contain 15.7 million packets that account for a total of around 7.6GB of traffic.

We compare DL-RA strategy with several state-of-the-art resource allocation techniques, including NN-RA, CSO, and FF in traffic anomalies detection. As shown in Fig. 3(a), the predicted results of deep learning are significantly better than the conventional prediction results by NN. The prediction error occurs because, the traffic is directly influenced by many non-linear factors such as the hot spot events, users' mobility pattern, etc. For this reason, many flows could not be accurately predicted. Base on the computation, the prediction accuracy of DL is about 7 percentage points higher than NN. In Fig. 3(b), we compare the resource occupation rate of four algorithms. Within our expectation, DL-RA have a better performance than NN-RA and the other two. The DL-RA strategy can allocate resources to the traffic more reasonably according to the predicted results. That is because the algorithm takes into account both the arrival time of the traffic and the resources it will need. From the results shown in Fig.3 (c), both DL-RA and NN-RA can reduce the traffic blocking probability to a large extent, and DL-RA is even performance better than NN-RA when traffic anomalies occur. The reason is that DL-RA can reserve sufficient resources in advance for predicted high priority traffic, so it reduces the blocking probability of high priority traffic under the heavy-loaded condition.

5. Conclusions

We propose DNN-PS based on deep learning for resource allocation in intra-datacenter optical network. In addition, a DL-RA algorithm is further developed to achieve efficient allocation of resources. Showcased experimentally, the proposed algorithm can effectively improve the prediction accuracy in complex network environment thereby improve the network performance.

Acknowledgment

This work has been supported in part by NSFC project (61501049), National Postdoctoral Program for Innovative Talents (BX201600021), Beijing Talents Fund (No. 2016000020124G075), China Postdoctoral Science Foundation (2017M610065) and funded by State Key Laboratory of Advanced Optical Communication Systems Networks, China, ZTE Research Fund.

6. References

[1] D. Zibar, "Machine learning techniques applied to system characterization and equalization," Proc. of OFC, Tu3K. 1 (2016).
[2] A. Aguado et al. "Towards a control plane management architecture enabling proactive network predictability" Proc. of OFC, W2A. 46 (2016)
[3] Li J et al. "Optical transport network architecture enabling ultra-low latency for communications among base stations" Proc. of OFC, Th4B. 5 (2017).
[4] A. Mayoral et al. "First experimental demonstration of distributed cloud and heterogeneous network orchestration with a common Transport API for E2E services with QoS," Proc. of OFC, Th1A. 2 (2016).
[5] H. Yang et al., "CSO: cross stratum optimization for optical as a service," IEEE Commun. Mag. **53**, 130-139 (2015).
[6] J. Li et al., "Optical transport network architecture enabling ultra-low latency for communications among base stations" Proc. of OFC, Th4B. 5 (2017).
[7] H. Zhang et al. "CODA: Toward automatically identifying and scheduling coflows in the dark" Proc. of SIGCOMM, 160-173 (2016).
[8] Q. Huang et al. "Ld-sketch: A distributed sketching design for accurate and scalable anomaly detection in network data streams" Proc. of INFOCOM, 1420-1428 (2014).