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

Speaker: Lin Wang

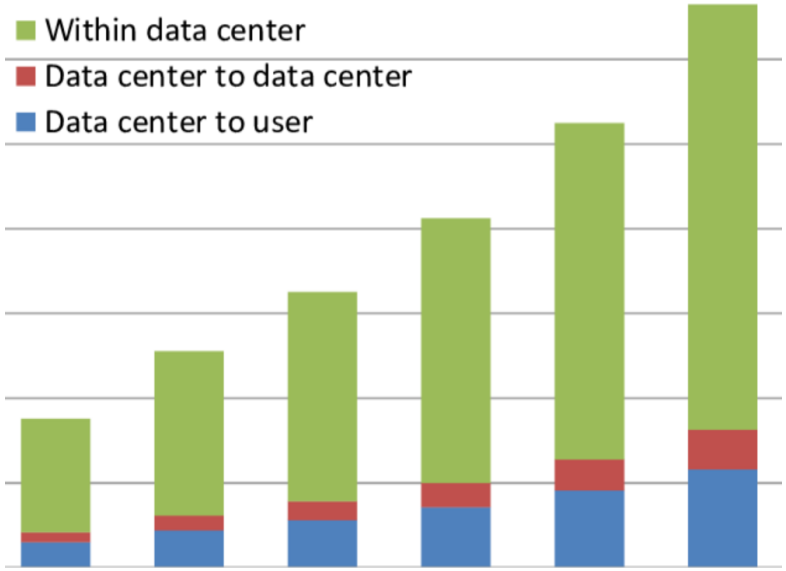
Research Advisor: Biswanath Mukherjee

UCDAVIS

A. Yu, et al. "Leveraging deep learning to achieve efficient resource allocation with traffic evaluation in datacenter optical networks." 2018 Optical Fiber Communications Conference and Exposition (OFC). IEEE, 2018.

Motivation

- **Traffic demand increasing in datacenter networks**
  - Cloud-service, parallel-computing, etc., lead to huge amount of intra datacenter traffic growth.
  - Cisco forecasts 31% increase per year of datacenter traffic by 2021



Datacenter traffic loads is growing

Introduction

- **Traditional intra-datacenter traffic detection methods**
- Support vector machine (SVM)
- Neural networks (NN)
- Decision tree
- NBD

- **Weakness**
- Simple prediction model
- Unguaranteed accuracy due to the change of bandwidth explosion and service diversity
- Unable to detect internal relations for raw data

Deep learning for traffic detection

- **Strengths**
- Discover deep connections in data
- Accurate prediction in complex networks
- Identify and characterize the complex structural characteristics in huge amounts of raw data
-

Solution for intra-datacenter networks

- **Deep neural network-based prediction strategy (DNN-PS)**

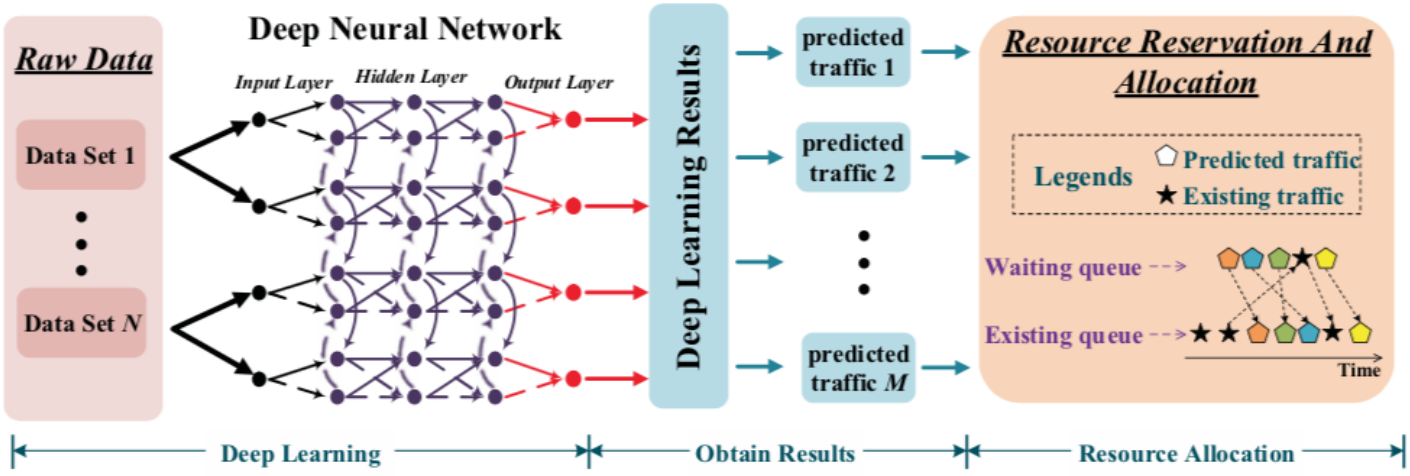


Fig.1 Logic sketch of DNN-PS.

- Build a database containing tens of millions of datacenter traffic information
- Traffic data are collected every 20s from over 200 servers
- Adjusts the weights and bias in DNN during the training process to minimize the objective function.

Solution for intra-datacenter networks

- **Deep learning-based resource allocation algorithm (DL-RA)**
- Use DDN-PS model to predict new traffic data
- Allocate resources for the new arrival traffic
- Use α to evaluate the validity of prediction results in terms of accuracy and resources

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

Solution for intra-datacenter networks

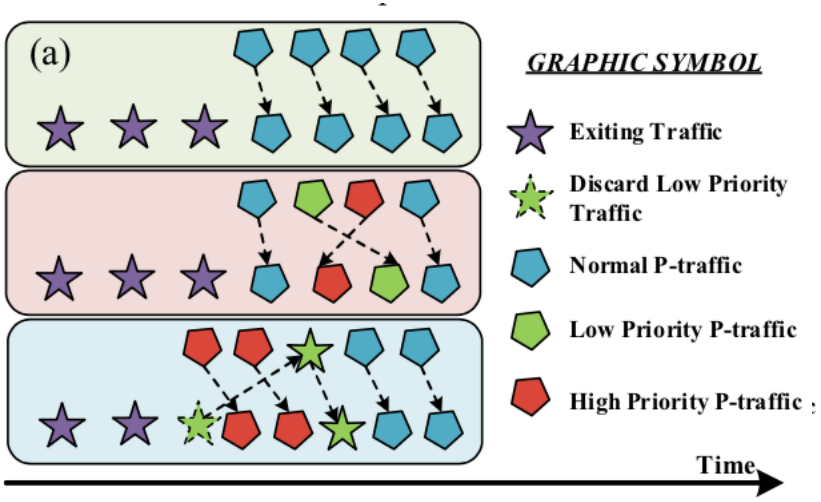
- **Deep learning-based resource allocation algorithm (DL-RA)**
- α considers both accuracy of traffic prediction and global resource utilization

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

T_{pi} : arrival time of j^{th} predicted traffic (P-traffic)
u_p : value center of T_p at different amounts of arrival time
t' : predicted arrival time
t : actual arrival time
$R_{pj}(t)$: resources required for the j^{th} arrival predicted traffic at t time
M : total amount of traffic that will arrive during time (t_0 , t_N)
N : arrival time of M_{th} traffic
β : is the adjustable weight between traffic and resource parameters with different user requirements

Solution for intra-datacenter networks

- **Deep learning-based resource allocation algorithm (DL-RA)**
- α considers both accuracy of traffic prediction and global resource utilization



Schematic of traffic queue reordering in DL-RA

Algorithms 1: DL-RA (b)

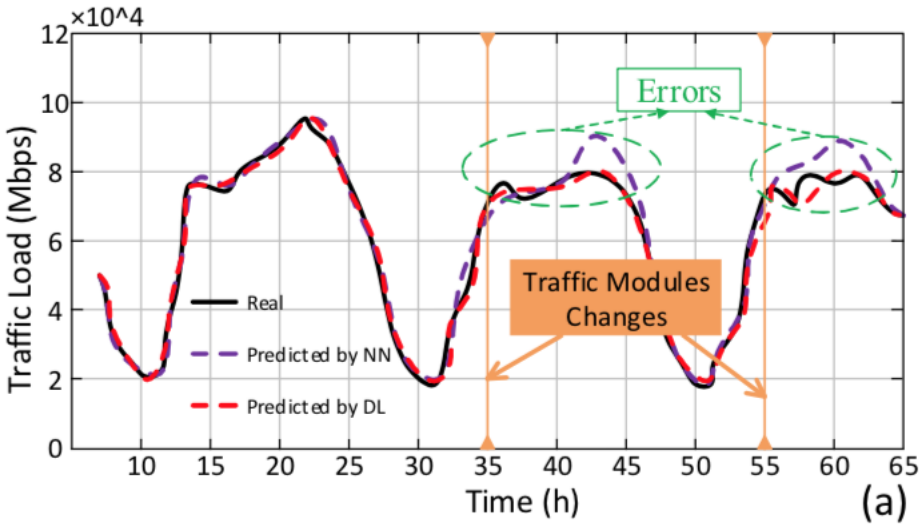
Input: training results T_{pj}; expansion parameter Priority = k

1: **for** each T_{pj}
2: **if** $R_p \le R_t$ **then**
3: **else if** there will be no low priority traffic in queue
4: **else if** k=*low* **then**
5: T will be added to traffic queue
6: **else if** $\alpha >$ *threshold* **then**
7: Allocate resources after discarding
8: **if** $R_p \le R_t$ after resource arrangement **then**
9: Allocate resources after arrangement

The pseudocode of DL-RA algorithm

Simulation results

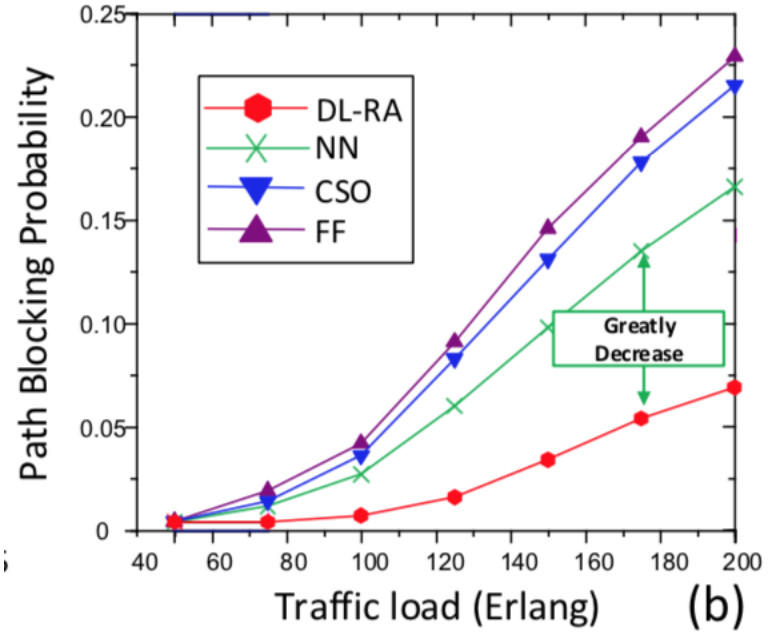
- **Deep learning-based resource allocation algorithm (DL-RA)**



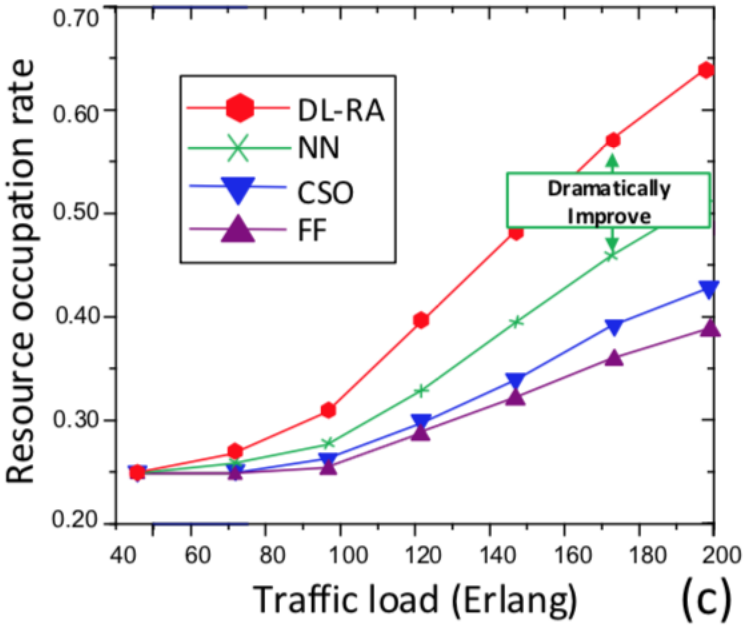
Compare real traffic load, deep learning-based prediction and neural network-based prediction.

Simulation results

- **Deep learning-based resource allocation algorithm (DL-RA)**



path blocking probability among different resource allocation strategies



resource occupation rate among different resource allocation strategies

UCDAVIS
UNIVERSITY OF CALIFORNIA

amlwang@ucdavis.edu