Deep Neural Network Based Dynamic Resource Reallocation of BBU Pools in 5G C-RAN ROADM Networks

Weiyang Mo[1], Craig L. Gutterman[2], Yao Li[1], Gil Zussman[2], Daniel C. Kilper[1]
[1]*College of Optical Sciences, University of Arizona, Tucson, AZ 85721, USA*
[2]*Electrical Engineering, Columbia University, New York, NY 10027, USA*
{wmo,yaoli, dkilper}@optics.arizona.edu; clg2168@columbia.edu; gil@ee.columbia.edu

Abstract: An LSTM network is developed to predict BBU pool traffic in 5G C-RAN ROADM networks. 5G throughput improvement and resource savings are observed with resource reallocation by reconfiguring the optical network 30 minutes in advance.
OCIS codes: (060.4256) Networks, network optimization; (060.0060) Fiber optics and optical communications

1. Introduction

Traditional cellular networks rely on radio-access networks (RANs) in which baseband signal processing is carried out at the location of the cellular antennas. As traffic has increased, cell sizes are decreasing, dramatically increasing the number of cell sites and their capacity. In order to improve the scalability of these large numbers of access points, separation of the Remote Radio Heads (RRHs) and the Base Band Processing Units (BBUs) has been proposed for 5G networks. In centralized or cloud-RAN (C-RAN), the BBUs will be moved to a centralized location to allow for sharing of computing resources among multiple RRHs and mobile networks. By consolidating BBUs into several common locations (called BBU pools), cost and energy can be saved by sharing power and computational resources, leading to a reduction of capital and operational expenditures [1]. The high capacity required in these C-RAN fronthaul networks motivate the use of wavelength division multiplexed (WDM) optical systems.

It has been shown that traffic in different regions of a city can have different cellular network load patterns [2]. Therefore, further efficiency improvements might be possible if traffic can be reallocated among remote BBU pools based on the traffic load, using reconfigurable optical add drop multiplexers (ROADM). Previous work on assigning traffic to different BBU pools has relied on mixed integer linear programming (MILP) solutions [3]. Given that reconfiguration of optical networks takes minutes [4], real-time reallocation requires methods that can address this significant time dependence.

In this paper, we develop a deep neural network based algorithm that can accurately predict future ROADM network resource requirements. Making accurate predictions 30 minutes in advance would allow for resource reallocation before the actual demand is needed, and therefore gives enough time for optical network reconfiguration to route the traffic through the C-RAN to a BBU pool with available computing resources. We compare the proposed resource reallocation approach with fixed resource allocation to evaluate the resource savings.

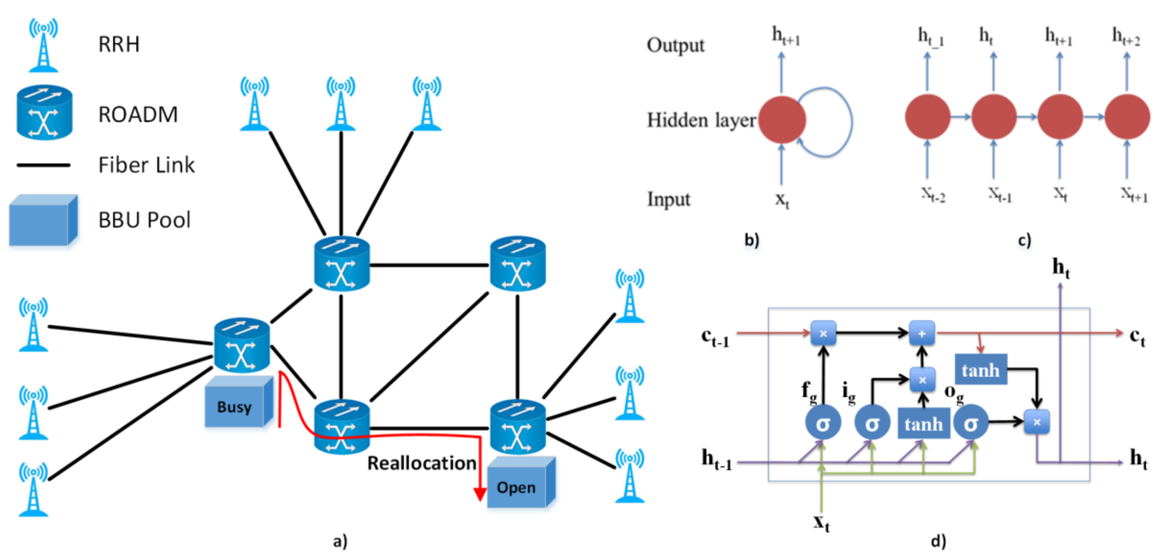

Fig.1. (a) C-RAN network architecture with the capability of resource reallocation from a busy BBU to an open BBU; (b) RNN architecture; (c) RNN unrolled through time, creating a deep neural network; (d) LSTM network.

2. Problem Statements and BBU Pool Resource Reallocation Approach

Fig. 1(a) shows a C-RAN architecture with 'fronthaul' WDM connections from an RRH to a remote BBU pool, which is typically via a common public radio interface (CPRI). Although fronthaul enables more efficient cloud based processing, the overall required transport capacity increases when the full digitized RF signal is used. Without the prediction of the traffic in advance, enough optical capacity and enough BBUs must be installed at each BBU pool to guarantee the peak processing requirement. On the other hand, the aggregated traffic at different BBU pools is different due to different cellular network load patterns, creating an opportunity to take advantage of sharing processing resources across BBU pools. To enable this sharing through ROADM optical network reconfiguration, which can take minutes, traffic pattern prediction must be made minutes in advance.

Recurrent neural networks (RNNs) can be used for predicting time series data through interconnected neurons based on previous time samples to make predictions for the next time series [5] as shown in Fig. 1(b). Unlike feedforward neural networks, RNNs do not only use information from the current input, but also use information from previous time steps. When unrolled through time, as seen in Fig. 1(c), it creates a deep neural network. A specific type of RNN architecture designed for long-term time series prediction is a Long Short-Term Memory (LSTM) network depicted in Fig. 1(d). LSTM works by storing information in the memory cell and passing it to the next time step. The inputs to the LSTM cell gates are a concatenation of the new input x_t and the previous output h_{t-1}. The LSTM cell has a forget gate f_g that allows information to be forgotten from the previous cell state c_{t-1}, an input gate i_g to add information into the new cell state c_t, and an output gate o_g for passing information from the new cell state c_t to the output h_t [6].

3. Case Study and Results

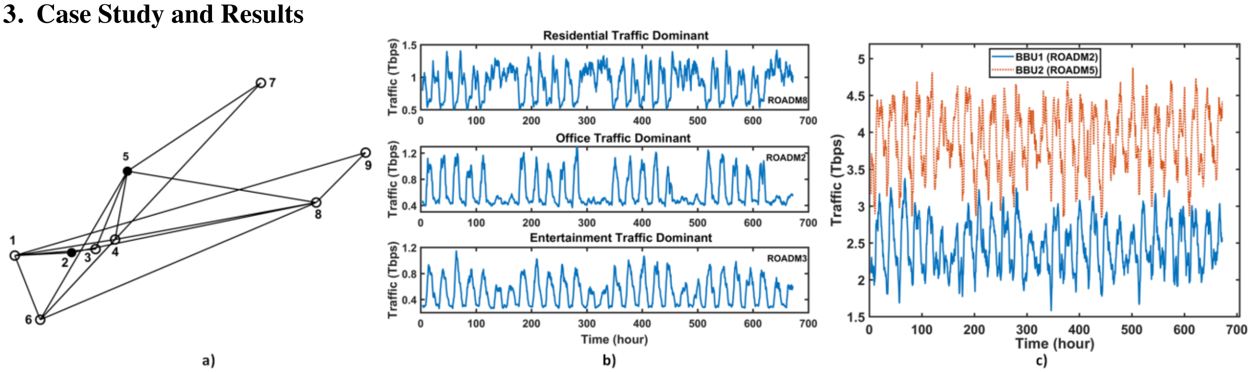

Fig.2. (a) New York City regional PoP topology; (b) Different traffic patterns (resident, office, and entertainment dominant) of different ROADMs; (c) Traffic patterns at two BBU pools.

We use discrete event simulations to evaluate the resource reallocation approach, considering the traffic rejection rate and total network throughputs as two main performance metrics. A regional New York City point of presence (PoP) network is considered, where 9 ROADM nodes cover a 400km^2 region with a 3.5 average degree, as shown in Fig. 2(a). 12060 connection requests following Poisson arrivals are generated with uniformly-distributed source and destination pairs. We assume 64 small cells (SC) are directly routed to each ROADM with a maximum fronthaul link of 23Gbps per SC (i.e., 1.472Tbps peak traffic per ROADM). Residential, office, and entertainment traffic is considered in the simulations, and the distribution of the three types for each ROADM varies based on its geographical locations [5]. As a result, different ROADMs have different time-dependent traffic patterns. Fig. 2(b) shows the traffic pattern over 28 days for ROADM 2, ROADM 5, and ROADM 8 with residential dominant, office dominant, and entertainment dominant, respectively. Each fronthaul connection is required to pass a BBU pool for data processing before dropping at the destination ROADM. First, we evaluate the BBU pool resource reallocation approach with a highly-consolidated BBU pool placement, where only ROADM 2 and ROADM 5 (solid circle) have BBU pools. For each connection, a modified k-shortest path routing (k=5) with the first-fit wavelength assignment is used, where each routing path must pass through a BBU pool. 50GHz spaced dense WDM channels over C-band links are used with grooming and 100 Gb/s PM-QPSK modulation.

In the fixed resource allocation approach, for any connection request, the shortest path is chosen and the first BBU pool along this path is chosen to process the data. Fig. 2 (c) shows the traffic pattern over 28 days for two BBU pools averaged every 30 minutes. The peak traffic of the two BBU pools is 4.7Tbps and 3.2Tbps, respectively. We then evaluate the LSTM network based resource reallocation. The time *t* of the LSTM network is in the interval of 30 minutes. Using the LSTM approach, the traffic of each BBU pool can be predicted 30 minutes in advance with high

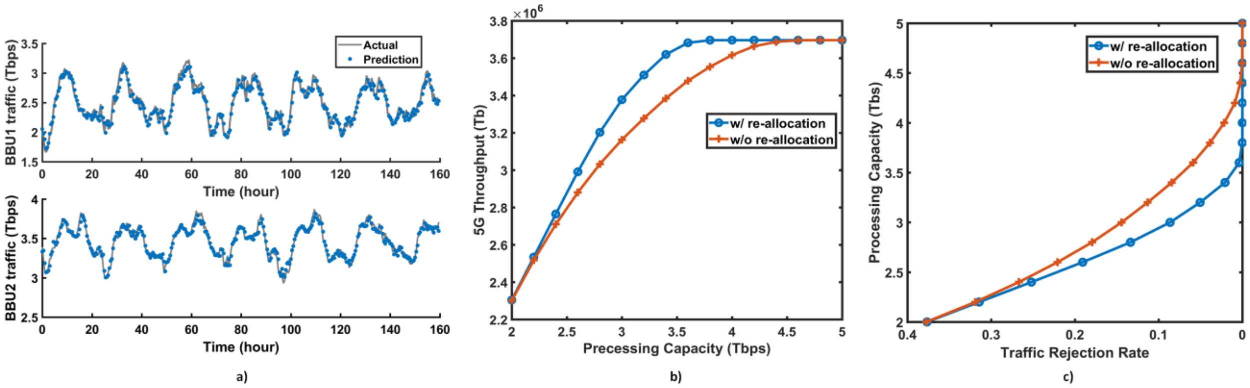

Fig.3. (a) Traffic pattern prediction on two BBU pools using LSTM; (b) Traffic throughput improvement with resource reallocation; (c) Reduced traffic rejection rate with the resource reallocation.

accuracy. An LSTM network is trained with 55% (740 samples) of the data in Fig. 2(c), and 20% (268 samples) is used for validation, and the last 25% (336 samples) is used for testing. Training is done by stochastic gradient descent where the training data is broken into batch sizes of 20 and optimized over 1000 epochs. Truncated back propagation through time (TBPTT) is used to optimize the weight parameters of the network. For the experiments, 60 time steps $X_t = \{x_{t-59}, x_{t-58}, x_{t-57}, ..., x_t\}$ are used. Fig. 3(a) shows the prediction performance of the test data. The model obtains a mean absolute error (MAE) of 84.2Gbps, mean absolute percentage error of 3.1%, a root mean square error of 96.9Gbps, and a maximum error of 319.2Gbps. With the prediction of the traffic pattern in each BBU pool, the optical network can be reconfigured 30 minutes in advance so that some resources can be reallocated from one BBU pool to another for processing, and then routed to the destination ROADM.

We study the performance of resource reallocation, by varying the peak resource processing capacity of each BBU pool from 2Tbps to 5Tbps, and compare with the case with fixed resource allocation. Fig. 3(b) shows that with a small processing capacity, resource reallocation does not give much improvement since both BBU pools are usually overloaded and there is no additional capacity for reallocation. With the increase of BBU processing capacity, resource reallocation gives a higher 5G traffic throughput, with a 7% maximum improvement at 3.5Tbps. When the BBU processing capacity further increases, the improvement decreases because there is less of a need for resource reallocation. Fig. 3(c) shows the relation between the BBU processing capacity and traffic rejection rate. It is seen that with resource reallocation, zero traffic rejection rate is achieved with 3.8Tbs capacity in each BBU pool (i.e., 7.6Tbps in total). On the other hand, 4.6Tbps processing capacity is required for each BBU pool (9.2Tbps in total) to serve all the traffic. Overall, resource reallocation leads to an 18% processing resource reduction.

4. Conclusions

We investigate an LSTM network based machine learning approach for BBU pool resource reallocations in a 5G C-RAN network using a ROADM switched optical network. The LSTM network is trained with 740 samples and predicts the future traffic pattern with a 96.9Gbps RMSE. 7% increase in network throughput and an 18% processing resource reduction is achieved by using the predicted traffic pattern to reconfigure the ROADM network 30 minutes in advance.

5. Acknowledgement

This work was supported by National Science Foundation under grant number CNS-1650669, CNS-1423105, CNS-1650685, and PFI:AIR TT 1601784.

6. References

[1] F. Musumeci, et al. "Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks." Journal of Lightwave Technology, 2016.

[2] H. Wang, et al. “Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment.” Internet Measurement Conference (IMC), 2015.

[3] A. Nag, et al. "Integrating wireless BBUs with optical OFDM flexible-grid transponders in a C-RAN architecture." Optical Fiber Communications Conference and Exhibition (OFC), 2017.

[4] L. E. Nelson, et al. "SDN-Controlled 400GbE end-to-end service using a CFP8 client over a deployed, commercial flexible ROADM system." Optical Fiber Communications Conference and Exhibition (OFC), 2017.

[5] F. A. Gers, et al. "Learning precise timing with LSTM recurrent networks." Journal of machine learning research, 2002.

[6] S. Hochreiterand and J. Schmidhuber. "Long short-term memory." Neural computation, 1997.