Machine Intelligence in Allocating Bandwidth to Achieve Low-Latency Performance

Lihua Ruan and Elaine Wong

Department of Electrical and Electronic Engineering, The University of Melbourne,

VIC 3010, Australia. ewon@unimelb.edu.au

***Abstract*— In this work, we present a complete rethink of the decision-making process in allocating bandwidth in a heterogeneous Fiber-Wireless network with machine intelligence. We highlight the use of an artificial neural network (ANN) at the central office to learn the uplink latency performance using multiple network and packet features. In turn, the trained ANN enables the central office to facilitate flexible bandwidth allocations under diverse network scenarios in meeting low-latency communication demands.**

***Index Terms*—Artificial neural network; dynamic bandwidth allocation; fibre-wireless networks, machine learning; low latency; Tactile Internet.**

I. Introduction

The Tactile Internet era is igniting an explosion of real-time, remotely controlled human-to-machine (H2M) and machine-to-machine (M2M) applications [1]-[4]. To support low latency (in the order of milliseconds, ms) and highly-reliable delivery of control/sensor-oriented traffic typical of such applications, we have previously considered the delivery of traffic over converged Fiber-Wireless (FiWi) networks along with the relocation of control servers closer to the end users [5]-[6] to expedite feedback and response.

An illustration of the heterogeneous FiWi network considered in our work is shown in Fig. 1. In the FiWi network, uplink bandwidth is shared amongst many optical network units (ONUs) that support aggregated wireless local area traffic from multiple end users. The process of allocating bandwidth to and scheduling transmission from each of these end users thus influence the overall latency. In this respect, the decision making process in allocating bandwidth and scheduling transmission is critical in meeting strict latency requirements and thus warrants attention.

Dynamic bandwidth allocation (DBA) schemes in fiber access networks are commonly centralized at the central office (CO) to schedule bandwidth resources for uplink transmissions. The bandwidth allocated to individual ONUs is typically determined based on the requested bandwidth in the REPORT message sent from each ONU [7]. Efforts in reducing uplink latency have been previously reported in [7]-[10], by predicting bandwidth demand based on the information in the REPORT messages and on arrival traffic characteristics. Statistical prediction methods such as constant credit and linear credit [7], arithmetic average [8], exponential smoothing [9] and Bayesian estimation [10], have been used in DBA schemes to predict bandwidth demand and subsequently to facilitate bandwidth allocation decisions. However, the limitation of these existing algorithms lies in their use of single traffic/network features, e.g. packet arrival rate or aggregated traffic load, to predict bandwidth demand. When network traffic load, packet length, and/or network configuration such as CO-to-ONU distance vary, the effectiveness of these algorithms in predicting bandwidth demand is compromised. Research in [11] and [12] explicitly reported on the challenge in determining the appropriate bandwidth to be allocated when network/traffic parameters vary.

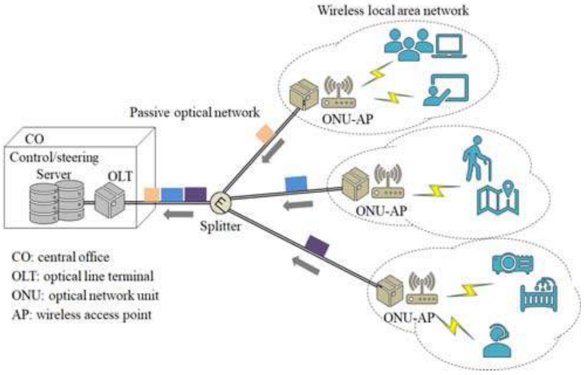

Wireless local area network
Passive optical network
ONU-AP
CO
Control/steering
Server
OLT
ONU-AP
Splitter
ONU-AP
CO: central office
OLT: optical line terminal
ONU: optical network unit
AP: wireless access point

Fig. 1. An illustration of a heterogeneous wireless local area and optical access network architecture for converged service delivery

In this work, we present a complete rethink of the decision-making process of allocating bandwidth with machine intelligence. Although machine learning (ML) techniques have been recently adopted in traffic routing, post-processing of signals, network failure prediction, the capability of machine intelligence in benefiting bandwidth resource allocation still remains an open question. For illustrative purposed, we show

978-3-903176-07-2 © 2018 IFPF

in this work, the exploitation of an artificial neural network (ANN) in (a) learning network uplink latency performance using diverse and multiple network features and in turn, (b) facilitating flexible bandwidth allocation decisions that effectively reduce the uplink latency under various network scenarios.

II. Artificial Neural Network Facilitated Dynamic Bandwidth Allocation (DBA)

A. DBA in Heterogeneous Networks

In a typical DBA algorithm, the CO grants an amount of bandwidth through a GATE message to each ONU upon receiving the REPORT messages from ONUs in the previous polling cycle(s). A polling cycle is defined as the time interval between consecutive transmissions from an ONU. Early works on predicting bandwidth demand have used a limited-service approach whereby the CO would use the requested bandwidth BW_{req} from REPORT messages to estimate the bandwidth demand, BW_{dem}. As discussed in Section I, statistical prediction methods such as constant credit and linear credit [7], arithmetic average [8], exponential smoothing [9] and Bayesian estimation [10], have been used to estimate BW_{dem}. In these early works, once BW_{dem} is obtained, the CO would subsequently grant $\min\{\mathrm{BW}_{dem}, \mathrm{BW}_{max}\}$ to the ONUs in the next polling cycle. Here, BW_{max} is the maximum bandwidth that can be allocated by the CO to the ONUs.

A major challenge of limited-service DBA algorithms lies in estimating an accurate BW_{dem} since bandwidth over-granting or likewise under-granting due to an inaccurate bandwidth prediction, may potentially increase uplink latency. Compounding the issue is that to the accuracy of $\mathrm{BW}_{\mathrm{dem}}$ depends on multiple network features, e.g. statistics of packet length, network traffic load and network configuration. It is also complex to derive BW_{dem} using conventional mathematical or analytical methods.

B. ANN Learning and Decision-Making Model

Here, we present an ANN learning and decision-making model and show how machine intelligence can be used to predict BW_{dem} with high accuracy. BW_{dem} can be resolved into two bandwidth components as shown below:

$$\mathrm{BW}_{dem} = \mathrm{BW}_{req} + \lambda T_{POLL}(\alpha S_{min} + (1-\alpha) S_{max}) \quad (1)$$

where the first term on the right hand side, BW_{req} is the requested bandwidth in the REPORT message from each ONU, and the second term on the right hand side is the predicted bandwidth. T_{POLL} is the polling cycle duration of an ONU. $S_{\max}$ and $S_{\min}$ are the maximum and minimum packet length, respectively. The parameters λ and α ($0 \le \alpha \le 1$) are arrival rate and the defined prediction coefficient, respectively. Our ANN learning and decision-making model predicts the second term on the right-hand side of (1) and hence $\mathrm{BW}_{\mathrm{dem}}$, to yield the lowest uplink latency through selection of α.

An ANN comprises an input layer, an output layer and some hidden layers in between, and learns by iteratively adjusting its weight and bias associated with the neurons in each layer to yield a desired output. An ANN learns complex nonlinear relationships between the input features, and yields a target output. A schematic of our proposed ANN learning and decision-making model is presented in the top diagram of Fig. 2. It must be noted that the average uplink latency over heterogeneous networks is impacted not only by α but also by diverse network features. Therefore, to train our ANN, we use the following key *input features*:

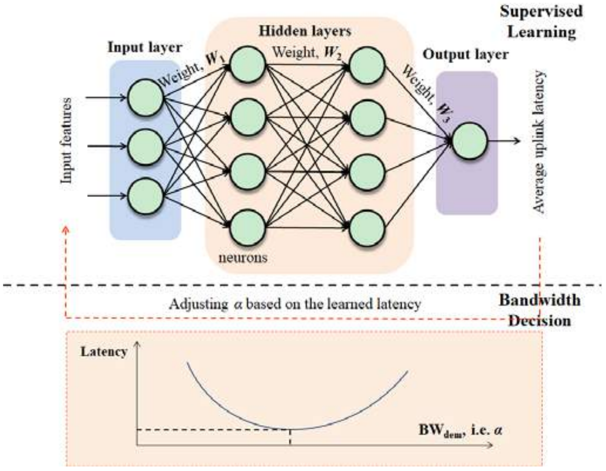


Fig. 2. An illustration of the proposed ANN learning and bandwidth decision-making model.

- $S_{\max}/S_{\min}$ — maximum/minimum packet length
- $S_{\mathrm{avg}}/S_{\mathrm{var}}$ — mean/variance packet length
- λ — packet arrival rate in the wireless local area network
- α — prediction coefficient
- N — The number of ONUs
- $D_{\max}$ — The maximum CO-to-ONU distance
- R_{PON} — Data rate of the passive optical networks (PON)
- R_{WLAN} — Data rate of the wireless local area networks (WLAN)

The target output is the average uplink latency of the network. As such, we train an ANN to learn the latency performance associated with different BW_{dem} decisions through varying α. When supervised learning is complete, the trained ANN predicts the average uplink latency for any α value that can possibly be selected (refer to bottom diagram of Fig. 2), thereby enabling α that yields minimum latency to be solved. The CO then allocates bandwidth with the BW_{dem} solution corresponding to the selected α. In the following section, we show how the supervised training can be implemented and highlight latency improvements achieved by a DBA algorithm facilitated by the trained ANN. This DBA algorithm is termed ANN-DBA for clarity.

III. Latency Performance Improvement

A. Supervised Training

We use a training set generated with varying input features to train an ANN with three hidden layers. The number of neurons of the three hidden layers is 5, 10, and 5, respectively. The target output of a training sample is the average uplink latency over a 1000-ms network running time (approximately 1000 polling cycles times), corresponding to a given input network feature in an event-driven packet-level simulation environment. With the knowledge of the dependence between uplink latency performance and the selection of α learnt by the trained ANN, the bandwidth allocation decision in (1) can be performed by finding α that minimizes latency.

For illustrative purposes, we report on the training process and decision-making outcome of a 16-ONU PON-WLAN network when λ and α change. We first use a training set containing 100 samples generated in an event-driven packet-level MATLAB simulation environment. The target output of a training sample is the average uplink latency, D_{uplink}, over a 1000-ms network running time, i.e. around 1000 polling cycles times. A network configuration comprising 16-ONUs with 10-km CO-to-ONU distance, packet lengths that are uniformly-distributed between 64 and 1518 bytes, and data rates of 1 Gbps and 100 Mbps for the optical and wireless segments respectively [7], are considered for illustrative purpose.

Another 250-sample test set was generated to validate the training outcome. The input features of the test set was fed to the trained ANN. The ANN predicted latency values were compared with the target latency values provided by the test set. Fig. 3 illustrates the prediction error arising from our use of the trained ANN, the mean square error of which is 6.6041. Next, the training set was increased from 100 to 300 samples with the training outcome validated using the same 250-sample test set. As shown in Fig. 3(b), with a MSE reduced to 2.2589 the performance of the ANN is significantly improved. As expected, the training outcome improves with an increased size of the training samples.

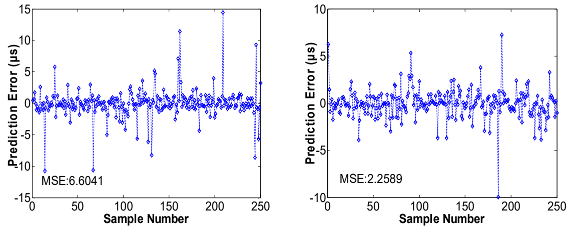
With the trained ANN, we are then able to analyze how uplink bandwidth allocation decisions, BW_{pre}, will impact uplink latency performance. Table I lists the selected α values and the corresponding minimum uplink latency as a function of traffic load. Note that the aggregated traffic load listed is normalized by $\lambda N S_{\text{avg}}/R_{\text{PON}}$. Table I highlights that after supervised training, the ANN can flexibly adjust bandwidth allocation decisions when the aggregated network load changes in the 16-ONU network.

TABLE I
OPTIMAL PREDICTION COEFFICIENT α
(16-ONU network, 10 km CO-to-ONU distance)

Traffic load	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
α	0	0	0	0.10	0.37	0.48	0.56	0.58
Latency (µs)	45.52	40.60	40.01	45.81	55.43	71.47	104.85	196.64

B. Latency Performance

The effectiveness of ANN-DBA in making flexible bandwidth allocation decisions that minimizes uplink latency, is highlighted in Fig. 4. The ANN-DBA allocates bandwidth in accordance to the decisions listed in Table I. As shown in Fig. 4, for all network loads, the uplink latency in a network using ANN-DBA (simulation) agrees with the ANN predicted latency (theory). Additionally, ANN-DBA results in latency performance improvement as compared to using the conventional limited-service DBA. A comparison of the uplink latency performance between ANN-DBA and the limited-service DBA as a function of varying CO-to-ONU distance, is shown in Fig. 3. Once again, the proposed ANN-DBA makes bandwidth allocation decisions that minimizes latency and does so irrespective of varying CO-to-ONU distances.



(a) (b)

Fig. 3. Prediction error of the trained ANN. (a) Training with 100 samples; and (b) training with 300 samples.

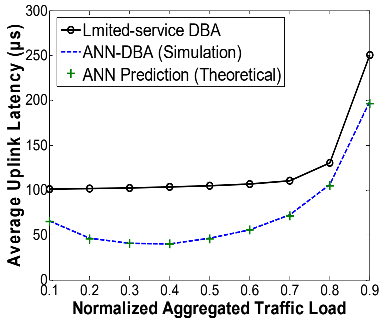


Fig. 4. Latency performance comparison as a function of traffic load.

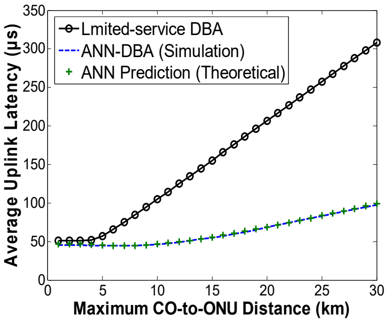


Fig. 5. Latency performance comparison as a function of CO-to-ONU distance.

Our results in Figs. 4 and 5 show that the ANN is capable of learning and predicting network latency performance with diverse network features as compared to the conventional limited-service DBA that relies on a singular traffic feature. In practice, training sets can be collected during network operation. Computation will be mainly spent in the supervised learning process, when the optimal weight matrix for each ANN layer is determined. Once training has ended, the CO needs only to store the weight matrix for each ANN layer. Mapping of input features to the target output value can be done without further computation.

IV. Conclusions

In this work, we investigated the applicability of an ANN in learning uplink latency, thereby in achieving flexible bandwidth allocation decisions that reduce latency. We highlighted the ANN's capability in predicting latency utilizing multiple network features. With the trained ANN, we showed that flexible bandwidth allocations under diverse application scenarios can be achieved and low-latency communication demands can therefore be met.

References

[1] M. Simsek, *et a*l, "5G-Enabled Tactile Internet," *IEEE J. Sel. Areas Commun.,* vol. 34, no. 3, pp. 460–473, Mar. 2016.

[2] G. P. Fettweis, "The Tactile Internet: Applications and Challenges," *IEEE Vehic. Techn. Mag.,* vol. 9, no. 1, pp. 64-70, Mar. 2014.

[3] M. Maier, *et al*, "The tactile internet: vision, recent progress, and open challenges," *IEEE Commun. Mag.,* vol. 54, no. 5, pp. 138-145, May 2016.

[4] J. Liu, et. al, "New Perspectives on Future Smart FiWi Networks: Scalability, Reliability, and Energy Efficiency," *IEEE Commun. Surveys Tuts.,* vol. 18, no. 2, pp. 1045-1072, 2nd quarter 2016.

[5] E. Wong, M. P. I. Dias, L. Ruan, "Predictive Resource Allocation for Tactile Internet Capable Passive Optical LANs," *J. Lightw. Technol.,* vol. 35, no. 13, pp. 2629-2641, Jan. 2017.

[6] S. Mondal, G. Das and E. Wong, "A Novel Cost Optimization Framework for Multi-Cloudlet Environment over Optical Access Networks," in *Proc.of. IEEE GLOBECOM*, 4-8 Dec, Singapore, 2017, pp. 1-7.

[7] G. Kramer, Ethernet Passive Optical Networks. McGraw-Hill Professional, 2005.

[8] R. Kubo, *et. al*, "Adaptive Power Saving Mechanism for 10 Gigabit Class PON Systems," *IEICE Trans. Commun.,* vol. E93–B, no. 2, 2010.

[9] M. Fiammengo, *et. al*, "Experimental Evaluation of Cyclic Sleep with Adaptable Sleep Period Length for PON," *in Proc. 37th Eur. Conf. Exhib. Opt. Commun.,* 18-22, Sep, Geneva, pp. 1-3, 2011.

[10] M. P. I. Dias, B. S. Karunaratne, E. Wong, "Bayesian Estimation and Prediction-Based Dynamic Bandwidth Allocation Algorithm for Sleep/Doze-Mode Passive Optical Networks," *J. Lightw. Technol.,* vol. 32, no. 14, pp. 2560-2568, 2014.

[11] R. Bushra, M. Hossen and M. M. Rahman, "Online Multi-thread Polling Algorithm with Predicted Window Size for DBA in Long Reach PON," in *Proc.of. ICEEICT*, 22-24, Sep, Dhaka, pp. 1-5, 2016.

[12] A. Dixit, *et. al*, "Delay models in ethernet long-reach passive optical networks," in *Proc. of. INFOCOM*, 26. Apr – 01. May, Kowloon, pp. 1239-1247, 2015.