Achieving Low-Latency Human-to-Machine (H2M) Applications: An Understanding of H2M Traffic for AI-Facilitated Bandwidth Allocation

Lihua Ruan, *Member, IEEE*, Maluge Pubuduni Imali Dias, and Elaine Wong, *Senior Member, IEEE*

***Abstract*—Human-controlled and haptic feedback data in emerging Tactile Internet human-to-machine (H2M) applications require stringent low-latency transmission. Understanding the traffic features of the new applications is vital in innovating network control and resource allocation strategies to meet their latency demand. In this article, we present our experimental study on human control and haptic feedback traffic in H2M applications and investigate novel bandwidth allocation schemes in supporting converged H2M application delivery over access networks. We introduce our haptic experiment system, the developed H2M applications, and analyze the control and feedback traffic traces collected. Then, exploiting the correlation between real-time control and feedback reported in our analysis, we propose an artificial intelligence-facilitated low-latency bandwidth allocation (ALL) scheme for emerging H2M applications. ALL provisions priority-differentiated bandwidth allocation for aggregated H2M and conventional content-centric applications over future access networks. By using ALL, the central office preallocates bandwidth for control and its corresponding feedback traffic interactively and prioritizes their transmission over content traffic. This expedites H2M application delivery by eliminating the report-then-grant process in the existing bandwidth allocation schemes. Via extensive simulations injected with experimental traffic traces, we comprehensively investigate the latency performance of ALL and existing schemes. Our results validate the superior capability of ALL in reducing and constraining latency for H2M applications.**

***Index Terms*—Bandwidth allocation scheme, haptic communication, human-to-machine applications, low latency, tactile Internet.**

I. Introduction

A. Motivations

COMMUNICATION networks are rapidly evolving from supporting only content-centric traffic to also including machine-centric traffic through the Internet of Things (IoT). The next evolution is driven by the advent of the Tactile Internet [1], which envisions a plethora of real-time and remotely controlled human-to-machine (H2M) applications over our communication networks. The Tactile Internet is defined as an ultrareliable and ultraresponsive network for manipulating and/or perceiving both virtual and real objects in a remote environment [2]. The H2M applications with haptic capability are its featured application, whereby human operators can "feel" tactile and kinetic sensations when controlling remote objects and immersively interact with the environment [3]. Seemingly, human and machine/robot interaction mutually benefits IoT functionalities and human communication experiences.

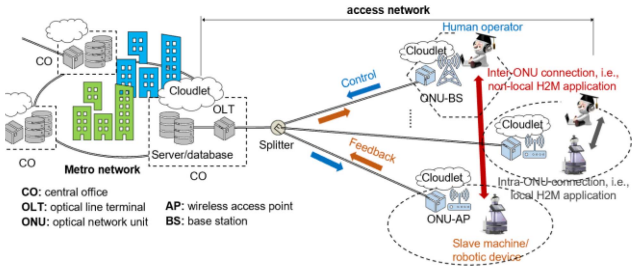

Fig. 1. H2M applications over fiber-wireless access networks.

H2M applications typically comprise: 1) a master domain with human operators and control interfaces; 2) a distant slave domain with execution machines; and 3) a network domain that communicates human-controlled and feedback packets between both master and slave domains. Unlike the existing content-centric and machine-centric applications that demand high bandwidth, stringent low latency is required in H2M applications whereby the control traffic from human operators and haptic feedback traffic from the slave machines need to adhere to an end-to-end latency of approximately 1–10 ms [4]. Lowering the network-domain latency relies on strategic infrastructure deployment and upgrading, together with efficient control and resource allocation over the network.

State-of-the-art research has emphasized the critical role of converged fiber and wireless access networks in realizing low-latency H2M applications [5]–[7]. As illustrated in Fig. 1, wireless front ends, such as cellular base stations and WiFi access points are integrated with optical network units (ONUs). Such a converged architecture benefits from the high capacity and reliability of optical networks, and the flexibility and mobility of wireless networks. Then, by strategically placing edge intelligence at the central office (CO) and/or at the wireless interface at ONUs as shown in Fig. 1,

Manuscript received March 15, 2020; revised May 17, 2020 and June 20, 2020; accepted July 1, 2020. Date of publication July 8, 2020; date of current version December 21, 2020. *(Corresponding author: Lihua Ruan.)*

The authors are with the Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, VIC 3010, Australia (e-mail: ruanl@student.unimelb.edu.au; ewon@unimelb.edu.au).

Digital Object Identifier 10.1109/JIOT.2020.3007947

data processing and exchanging can be expedited in converged access networks [8], [9]. Note that depending on the locations of the masters and slaves, H2M applications can be realized either locally (intra-ONU as shown in Fig. 1) or nonlocally (inter-ONUs via the backhaul optical access network and the CO) [6]. In this article, we focus on the latter case since such applications are more susceptible to the latency of converged access networks. In order to achieve 1–10 ms latency in inter-ONU communications, the latency attributed to the optical access network (PON) can be at most a few hundreds of μs [5]. Since multiple ONUs share the same bandwidth for uplink transmission to the CO, effective bandwidth resource allocation to ONUs is critical in reducing the latency.

B. Challenges in Bandwidth Allocation

At present, bandwidth allocation schemes in PONs are primarily designed for content-centric applications in a report-then-grant manner. ONUs report the number of packets in their buffer in a round robin. The CO then responds the transmission start time and duration of ONUs accordingly. The existing bandwidth allocation schemes can be classified into classic and predictive schemes. In the classic schemes, the CO primarily allocates bandwidth in accordance with the bandwidth requested by the ONUs. The limited service scheme is currently the widely-adopted baseline, whereby the CO grants bandwidth equal to the request but not exceeding a threshold [10]. In comparison, in the predictive schemes, the CO allocates some surplus bandwidth in addition to the request [10]. This allows ONUs to transmit arrivals without reporting, thereby reducing the latency. Credit schemes were early predictive schemes that provide a fixed amount of surplus bandwidth to ONUs [10], [11]. Furthermore, statistical predictive schemes utilize estimation algorithms to estimate the surplus bandwidth. For example, the Bayesian algorithm was utilized to estimate packet interarrival time, which can then be used to estimate bandwidth demand [12]–[14]. Various schemes as in [15]–[18] adopted short-term traffic prediction algorithms, e.g., linear regression and autoregressive moving average (ARMA), in estimating bandwidth for bursty arrivals.

Concurrently, artificial intelligence (AI) and machine learning (ML) techniques are also utilized for predictive schemes. The scheme in [19] used the *k*-nearest neighbourhood (*k*NN) to estimate arrivals of video flows. The support-vector machine was used to sifting noise data to enhance bandwidth utilization [20]. Hatem *et al.* [21] exploited a recurrent neural network and deep learning to estimate bandwidth for multiple polling cycles ahead. Our previous work investigated the use of multilayered artificial neural networks (ANNs) in improving the bandwidth estimation accuracy for bursty traffic [22], [23], [43]. Mikaeil *et al.* [24] studied the use of ANN for traffic prediction in the next-generation access networks.

Note that the design and performance of the above schemes are closely tied with the traffic characteristics. First, bandwidth estimation in these schemes relies on packet arrival statistics, such as e.g., interarrival time and the number of arrivals on record. Consequently, the performance of the existing schemes is impacted by such traffic statistics in application scenarios. For examples, credit schemes in [10] and *k*NN-based scheme in [19] are efficient for constant-flow traffic. ARMA in [18] and neural networks are typically applied to bursty arrivals. Second, bandwidth allocation in the above schemes relies on the round-robin reports from ONUs. Regardless of the estimation algorithms/techniques used, bandwidth is allocated to an ONU only upon a report is received. The reports can reflect the intra-ONU traffic statistics, but not sufficiently addressing the inter-ONU traffic association.

In light of the above, to improve the existing schemes for emerging H2M applications, understanding the H2M traffic characteristic is essential, and research on H2M traffic at present is demanded. The Tactile Internet and H2M applications are in their infancy and H2M applications have not been widely practiced over the access networks. The unique traffic characteristics of H2M applications and their aggregation in access networks, are yet to be fully understood and exploited in the existing schemes. As stringent low latency is required by H2M applications, a thorough examination on the performance of the existing schemes in supporting aggregated H2M applications is warranted. Novel solutions that can ensure latency for H2M applications in the presence of conventional applications need to be explored. These new aspects are the focus of this article.

C. Original Contributions

In this article, we develop interactive H2M applications to experimentally study the human control and haptic feedback traffic in H2M applications. In our experiments, we report the statistical and time-domain characteristics of H2M arrivals, highlighting the arrival models and the cross-correlation of control and feedback traffic. Exploiting the characteristics reported, we propose an AI-facilitated low-latency bandwidth allocation (ALL) scheme to improve the latency of access networks in supporting future aggregated H2M and content applications. Note that a preliminary study of our experiments was presented in [23], where only time-domain traffic analysis was presented. In comparison, this article details our experiments, extends comprehensive analysis on H2M traffic characteristics, and proposes the ALL scheme that: 1) facilitates interactive and predictive bandwidth allocation for H2M applications harnessing an ANN and 2) ensures low H2M latency via priority-differentiation bandwidth estimation and allocation for H2M and content traffic. With extensive simulations using experimental traffic traces, the performance of the existing schemes and the ALL in supporting aggregated H2M and content application is comprehensively compared and analyzed. The effectiveness of the ALL scheme is validated. The contributions of this work are summarized in threefold.

1) Experimental investigation on low-latency human control and haptic feedback traffic in H2M applications, and provision of insights in both statistical and time-domain characteristics of H2M traffic. Together with state-of-the-art research on H2M applications, this experimental study aims to add understandings to the H2M traffic.
2) Proposal of the ALL scheme that interactively allocates bandwidth for H2M traffic. To the best of our knowledge, the ALL scheme is the first attempt to innovate bandwidth allocation taking the unique H2M traffic

characteristics and latency demand into account for aggregated H2M and content applications in converged access networks.

3) Performance evaluation of the existing bandwidth allocation schemes and the ALL scheme, providing the first examination on existing schemes in supporting emerging H2M applications using experimental H2M traffic.

In the remainder of this article, Section II overviews the studies on H2M traffic and how such knowledge is used to improve resource allocation. The experiments and traffic analysis are in Section III. Section IV presents the ALL scheme. Results are evaluated in Section V. Finally, we summarize in Section VI.

II. Related Work on H2M Traffic

Real-time human control and haptic traffic in H2M applications are the latest traffic that needs to be supported by our communication networks. In supporting their low-latency transmission, studying the H2M traffic characteristics is vital and draws growing research interests. As detailed in [25], human control and haptic feedback in H2M applications are mainly described by the Degree of Freedom (DoF), i.e., the number of different directions that create forces, e.g., 3-DoF force along x-, y-, z-axis in a touch spot. The control/feedback packets may comprise one DoF to over hundreds of DoFs, which are sampled/generated by master control interface/slave machines [26]. The H2M communication was studied in [27]–[29] and several observations on H2M traffic were presented. Feng *et al.* [30] investigated the coding schemes and discussed a bursty profile of control and force feedback packets in their teleoperation experiment. This knowledge of burstiness was used to design a bandwidth resource reservation scheme for wireless access networks [28]. In [29], haptic traffic in several application cases, such as immersive virtual reality, teleoperation, was studied. Different haptic packet arrival models, including periodic, event based, and bursty, of these applications, were reported. Using these models, a latency optimal radio resource management was proposed to ensure latency and reliability in LTE networks. Feng *et al.* [30] modeled bursty H2M traffic in a switched Poisson process and proposed to adaptively reserve and allocate resource in cellular networks [31]. Recent research reported in [32] analyzed the statistical nature of H2M packet interarrival time in two teleoperation systems. In this study, the authors discussed different arrival models, including deterministic, Gamma, generalized Pareto (GP), and Poisson distributions. Our previous study in [13] adopted the Poisson and Pareto-distributed H2M arrival models to improve bandwidth and wavelength allocation schemes in PONs for H2M applications.

The above research implicated the burstiness of H2M traffic and described arrival models to characterize the bursty pattern. Nevertheless, compared to conventional content applications, H2M applications have not been widely utilized over access networks. Continuing experimental investigations on H2M traffic are still necessary and novel bandwidth allocation schemes to support low-latency H2M applications need to be explored. For these purposes, we present the details of our experimental analysis and the proposed ALL scheme in reducing inter-ONU H2M communication latency in PONs.

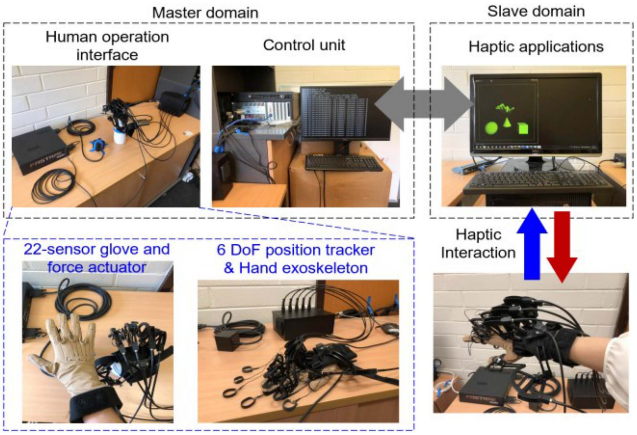

Fig. 2. Teleoperation system for haptic H2M applications.

Fig. 3. Experimental H2M (teleoperation) applications. (a) Hand movement. (b) Touching. (c) Grasping and moving.

III. Experimental H2M Applications and Traffic Analysis

A. Haptic Teleoperation System and Applications

Our experiments comprise a series of interactive H2M applications based on a commercial haptic teleoperation system [33]. As illustrated in Fig. 2, the system is an innovative touch and force feedback teleoperation platform for fingers and hands, allowing human operators to remotely touch and grasp computer-generated virtual objects. The master domain consists of a human operation interface and a control unit as shown in Fig. 2. The interface includes a 22-sensor data glove and a 6-DoF position tracker that capture human hand movements and actions. A hand exoskeleton is strapped to the glove, reacting to the haptic feedback by creating different resistive forces at fingers and joints.

In slave-domain applications, a virtual hand can be controlled to touch and/or grasp virtual objects as shown in Fig. 2. Then, the haptic feedback, indicating the reaction and friction forces of the virtual objects, is delivered back to the master-domain exoskeleton to actuate feelings of touch and force. In our study, we develop three types of H2M applications that create different haptic feedback as follows. These applications are illustrated in Fig. 3. The control and haptic feedback traffic in each application are specified in Table I.

1) *Application A:* Moving the hand in free space. The virtual hand in the virtual space moves with the human

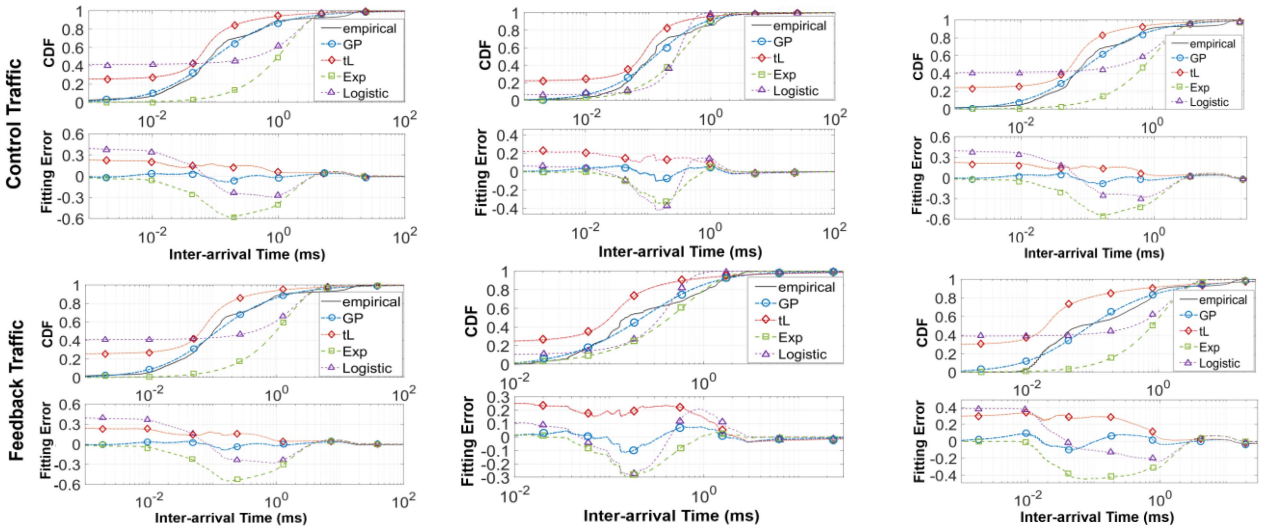

Fig. 4. Control and feedback packet interarrival time cumulative distribution functions of (a)–(c) Applications A–C.

TABLE I
EXPERIMENTAL H2M APPLICATIONS AND CONTROL/HAPTIC FEEDBACK TRAFFIC

App.	Control traffic	Feedback traffic
A	hand postures, e.g., flat palm, fist, curve figures, etc., and 6DoF position	a constant initial force in the form of tension to fingertips and joints
B	hand postures, 6DoF position, and actions of touching	stiffness of the object texture by reaction forces, i.e., pressures to fingers and joints
C	hand postures, 6DoF position, and actions of grasping and moving	stiffness of object texture by reaction force, and friction forces in the form of vibrated tensions to fingers

hand in real time as shown in Fig. 3(a). Without interactions with virtual objects, a constant default force is put to the figures.

2) *Application B:* Touching a virtual ball surface. A virtual ball is placed in the virtual space. When the hand touches the ball as shown in Fig. 3(b), reaction forces act to fingers to imitate stiffness of the object texture.
3) *Application C:* Grasping and moving objects, i.e., a ball, a box, and a cone in the virtual space. When human operators pick an object and move it in virtual space as shown in Fig. 3(c), vibrated forces are put to fingers to create the feeling of friction when moving the object.

The control traffic in Applications A–C contains similar content, i.e., hand posture and position. The haptic feedback traffic depends on the type of applications. Based on the developed Applications A–C, we collect human control and feedback traffic traces during H2M interaction. In our experiments, we allow human participants to start an application, perform operations such as grasping either a ball or a box, and move it to a random position such as in Application C, and pause/restart an application at any time. Each experiment lasts around 20 mins. Then, we analyze the statistical characteristics of the experimental control and feedback traffic. The comparative analysis with existing studies on H2M traffic and the content traffic over the Internet is presented as follows.

TABLE II
AVERAGE CDF FITTING ERROR

App.	Control traffic				Feedback traffic			
	GP	tL	Exp	Log	GP	tL	Exp	Log
A	0.0210	0.0483	0.1438	0.0961	0.0242	0.0551	0.1329	0.0875
B	0.0208	0.0452	0.1308	0.0908	0.0201	0.0505	0.1220	0.0852
C	0.0235	0.0659	0.2102	0.1309	0.0245	0.0716	0.1921	0.1244

TABLE III
FITTED PARAMETERS IN GP DISTRIBUTIONS*

App.	Control traffic	Feedback traffic
A	$\xi = 1.25253$ $\sigma = 0.0970738$	$\xi = 1.11692$ $\sigma = 0.146245$
B	$\xi = 1.26032$ $\sigma = 0.0947804$	$\xi = 1.16276$ $\sigma = 0.136011$
C	$\xi = 1.40311$ $\sigma = 0.0851359$	$\xi = 1.27666$ $\sigma = 0.123112$

* Generalized Pareto *pdf* $f_X(x) = \frac{1}{\sigma}(1+\xi\frac{(x-\mu)}{\sigma})^{-1-\frac{1}{\xi}}$
* the fitted μ value is 0 in Applications A-C

B. Statistical Analysis: Packet Interarrival Time

Packet interarrival time is one of the key features in studying network traffic. The statistics of interarrival time can be utilized in traffic classification [34] and fitting arrival models [35]. In this article, we fit control and feedback packet interarrival times into a variety of distributions in the existing network traffic studies [32], [36]. The candidate distributions include Exponential, Gamma, GP, Inverse Gaussian, Logistic, Lognormal, Nakagami, Normal, t-location (tL), and Weibull. The maximum likelihood estimation is utilized to fit the parameters in each candidate distribution. The goodness of fitting is measured by comparing the average fitting error between the fitted cumulative distribution function (CDF) and the experimental CDF. The fitting results are presented in Fig. 4 and Tables II and III.

In Fig. 4, we plot the CDFs and fitting errors of the top-four best fitted distributions, i.e., GP, tL, Logistic, and Exponential (Exp) distributions, to our experimental traces. Notably, in both control and feedback traces, the GP best fits empirical distributions. This observation is consistent in

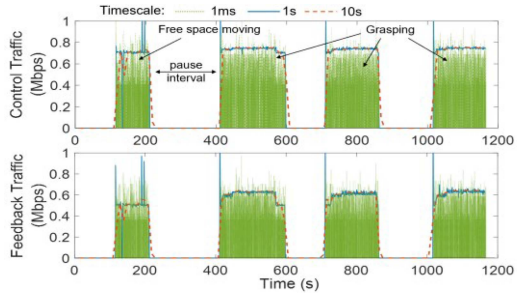

Fig. 5. Illustration of control and feedback traffic in Application C.

Applications A–C, and in both control and feedback traces in Fig. 4(a)–(c). Compared to GP distribution, the rest distributions incur higher fitting errors provided in Table II. The fitted parameters in Applications A–C are listed in Table III. It can be noted that the parameter pairs only slightly vary in different applications, but are distinct in control and feedback traces. In the existing studies, the GP-distributed H2M arrivals are also reported, such as in [29] and [32]. As such, we highlight the GP distribution as a potential model for H2M arrivals. However, it should be noted that due to the current limited knowledge on H2M traffic, more rigorous experimental studies using diverse H2M applications are still warranted to validate the observations on H2M traffic.

From Fig. 4, the CDFs of control and feedback packet interarrival time indicate intensive control and feedback packets exchange during H2M interactions. The control packets are transmitted every a few hundreds of microseconds, and the feedback packets immediately respond. As such, we further investigate the correlation in the collected traffic traces.

C. Time-Domain Analysis: Control/Feedback Correlation

In Fig. 5, we present the human control and haptic feedback traffic pattern by plotting the dynamic traffic volume (Mbps) in Application C on different time scales, i.e., 1 ms, 1 s, and 10 s. Fig. 5 shows an example experiment in which a human operator first moves the hand in virtual space, then grasps the three virtual objects, i.e., the ball, cone, and cube as depicted in Fig. 3(c). Note that after moving or grasping, the human operator pauses his action for a while. This is to distinguish traffic from different actions for the illustrative purpose. The observations of the traffic in Applications A and B are similar to those shown in Fig. 5. Therefore, we do not repeat the analysis for different experiments. As can be observed in Fig. 5, control and feedback packets are closely correlated, and packet arrivals fluctuate on the 1-ms granularity compared to the 1-s and 10-s timescales. Note that in PONs, the CO allocates uplink bandwidth to ONUs in a round-ribbon manner and the interval between consecutive transmissions of an ONU, known as a polling cycle, is in millisecond order. As such, the short-term (within ms) traffic characteristic is critical for bandwidth allocation improvement.

We analyze the correlation between control and feedback traffic on the time scale of polling cycles. Considering a 1-ms polling cycle duration and a sequence of control/feedback packets (in bytes) in N polling cycles, the Pearson cross-correlation between control and feedback traffic denoted as

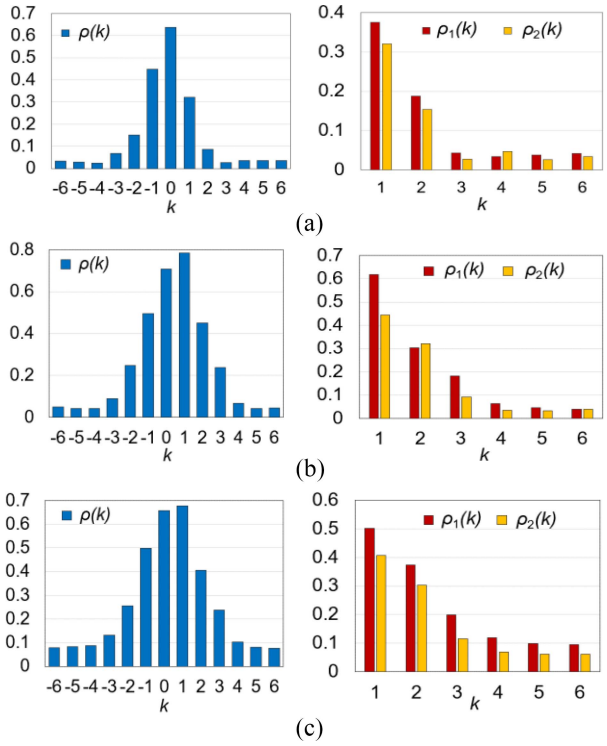

Fig. 6. Illustration of control and feedback traffic correlation. (a) Application A. (b) Application B. (c) Application C.

$\rho(k)$ can be calculated as follows:

$$\rho(k) = \frac{1}{N}\sum_{i=1}^{N}\frac{[\text{fbk}(t_i) - \mu_{\text{fbk}}]\big[\text{ctr}(t_{i-k}) - \mu_{\text{ctr}}\big]}{\sigma_{\text{ctr}}\sigma_{\text{fbk}}} \tag{1}$$

where $\text{ctr}(t_i)$ and $\text{fbk}(t_i)$ denote the total bytes of control and feedback traffic transmitted in polling cycle i, respectively. μ_{ctr}, μ_{fbk}, σ_{ctr}, and σ_{fbk} represent the means and variances of control and feedback traffic, respectively. The variable $k(k = 0, 1, 2, \cdots)$ indicates the time lag in the unit of polling cycle. Similarly, the self-correlation, i.e., autocorrelation in equivalent, of control and feedback trace, denoted as $\rho_1(k)$ and $\rho_2(k)$, respectively, can be calculated by

$$\rho_1(k) = \frac{1}{N}\sum_{i=1}^{N}\frac{[\text{ctr}(t_i) - \mu_{\text{ctr}}\big]\big[\text{ctr}(t_{i-k}) - \mu_{\text{ctr}}\big]}{\sigma^2{}_{\text{ctr}}} \tag{2}$$

$$\rho_2(k) = \frac{1}{N}\sum_{i=1}^{N}\frac{[\text{fbk}(t_i) - \mu_{\text{fbk}}\big]\big[\text{fbk}(t_{i-k}) - \mu_{\text{fbk}}\big]}{\sigma^2{}_{\text{fbk}}}. \tag{3}$$

Fig. 6 compares $\rho(k)$, $\rho_1(k)$, and $\rho_2(k)$ in Applications A–C. Note that $\rho_1(k)$ and $\rho_2(k)$ are even functions in k, and $\rho_1(0)$ and $\rho_2(0)$ equal to 1. We focus on the positive k values in analyzing the correlation. The results in Fig. 6 indicate a stronger cross-correlation, i.e., $\rho = 0.6-0.8$, between control and feedback than the self-correlation, i.e., $\rho_1 < 0.6$ and $\rho_2 < 0.4$, of the individual traces. Particularly, the high cross-correlation, i.e., $\rho > 0.6$, is shown between the bilateral traces in the same or the most recent polling cycles, i.e., at $k = 0$ and 1. This is mainly attributed to that haptic feedback responds to human control in real time. Based on the above analysis, leveraging the control and feedback cross-correlation in bandwidth prediction and allocation can potentially improve the

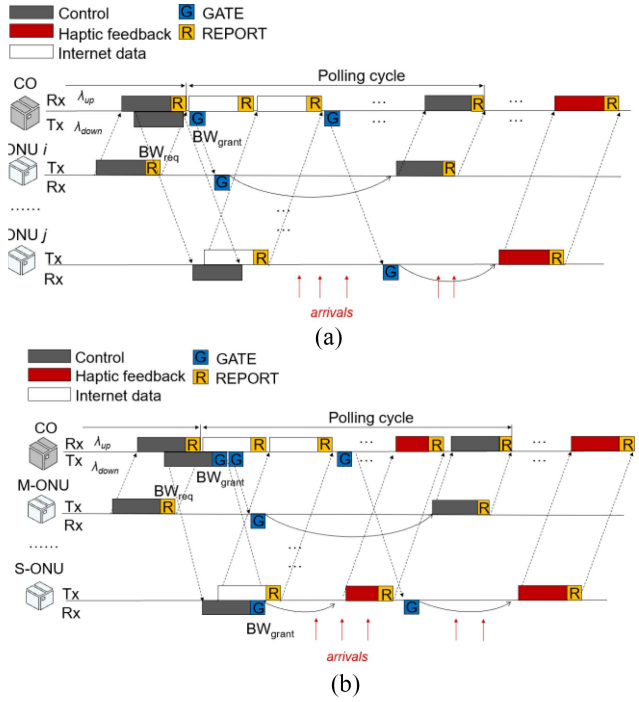

Fig. 7. Uplink bandwidth allocation over optical access networks. (a) Existing bandwidth allocation schemes. (b) Proposed ALL scheme.

network performance for inter-ONU H2M communications. As such, we propose the ALL scheme for low-latency H2M applications, detailed as follows.

IV. Bandwidth Allocation for Low-Latency H2M Applications

A. Bandwidth Allocation Schemes

Fig. 7 shows the diagram of bandwidth allocation schemes. As illustrated in Fig. 7, the CO broadcasts downlink packets to ONUs using downlink wavelength λ_{down}. In the uplink, ONUs share uplink wavelength λ_{up} for uplink transmission. Following a report–grant process as introduced earlier, an ONU requests bandwidth using a REPORT piggybacked to the uplink data. The REPORT notifies the CO of its buffer occupancy BW_{req}. Upon receiving the REPORT, the CO grants bandwidth BW_{grant}, by sending a GATE to the ONU. The GATE indicates the assigned transmission start time and duration. This process repeats for each ONU in each polling cycle.

For inter-ONU communication, the CO forwards the received data from ONU *i* to its destination ONU *j* in broadcast as shown in Fig. 7(a). Recall that in the baseline limited service scheme, the CO allocates bandwidth equal to the request, i.e., $\text{BW}_{\text{grant}} = \min\{\text{BW}_{\text{req}}, \text{BW}_{\text{max}}\}$. The BW_{max} is the maximum bandwidth that can be assigned to an ONU in each polling cycle. As such, in the baseline scheme, packets arriving in the current polling need to be reported and then transmitted in the next polling cycles. In reducing latency, the predictive schemes predict arrivals and allocate surplus bandwidth BW_{pred} such that $\text{BW}_{\text{grant}} = \min\{\text{BW}_{\text{req}} + \text{BW}_{\text{pred}}, \text{BW}_{\text{max}}\}$. This allows some arriving packets to be transmitted in the current polling cycle without needing to be reported, thereby reducing latency.

In the existing schemes, this BW_{pred} can be estimated by using statistical algorithms or ML techniques as overviewed. In this article, we consider the use of ANN in our proposed scheme for H2M bandwidth estimation, given its proven ability and the ease of use [39]–[43]. In comparison, we consider moving average-based algorithms for statistical predictive schemes. This type of algorithms remains a vital solution in estimation bandwidth for PONs [15]–[18]. It should be noted that improving bandwidth estimation is not the focus of this study. We follow the existing methods in facilitating bandwidth allocation.

Note that as shown in Fig. 7(a), the CO allocates bandwidth to ONUs based on their independent reports. Differently, in the ALL scheme, we propose to interactively allocate bandwidth exploiting the inter-ONU control and feedback traffic correlation. Furthermore, as differentiating H2M and content traffic in their bandwidth estimation and allocation is yet to be fully studied in the existing predictive schemes, we also present our design in this aspect in the ALL scheme.

B. Proposed ALL Scheme

The operation diagram of the ALL scheme is illustrated in Fig. 7(b). In the ALL scheme, the CO grants bandwidth to subsequent feedback traffic at the same time when forwarding control packets to the ONUs, thereby eliminating the bandwidth request process and effectively reducing the latency. To explain the ALL, we categorize ONUs associated with master human operators as an M-ONU, and the counterpart ONUs associated with slave machines/robots as S-ONUs as shown in Fig. 7(b). Note that an ONU can be both an M-ONU and S-ONU as it can support control and feedback traffic from multiple H2M applications. In Fig. 7(b), at the CO, the feedback bandwidth of an S-ONU is estimated based on the received control traffic from its counterpart M-ONU. Then, when the CO forwards the control, a GATE is appended to preallocate bandwidth instead of waiting for the report from the S-ONU. The ways to achieve bandwidth estimation, allocation and priority differentiation in the ALL scheme are as follows.

1) *AI for H2M Bandwidth Estimation:* We use an ANN to realize bandwidth estimation for H2M traffic, given its ability to attain accurate estimation [39]–[43]. For a pair of M-ONU and S-ONU, the control and feedback arrivals (counted in bytes) of Applications A–C at an M-ONU in the most recent four polling cycles, are used as the inputs to the ANN, and the target output is the estimated feedback bandwidth of the S-ONU in the next polling cycle. The training set contains ONUs' arrival records of 10^4 polling cycles, and the training objective is to minimize the mean-square error (MSE) of the estimation. By layerwise training [37], i.e., adding the number of neurons and layers incrementally, a 2-hidden layer ANN architecture with five and three neurons in each layer is determined. The gradient descent method is used to yield the weights and bias associated with the neurons. The final ANN architecture is shown in Fig. 8, and the attained MSE is in the order of 10^2. Our preliminary study in [23] details the ANN training process and the MSE comparison with the existing bandwidth estimation methods. In this article, we focus on innovating bandwidth allocation for H2M traffic.

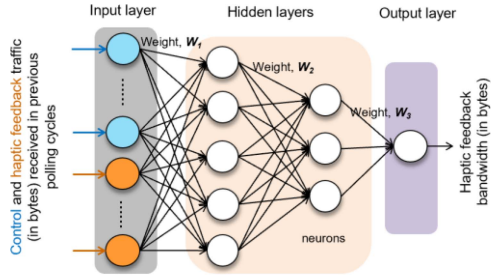

Fig. 8. Illustration of the ANN architecture.

2) *Interactive H2M Bandwidth Allocation:* The CO grants the estimated bandwidth to an S-ONU when forwarding the control traffic as shown in Fig. 7(b). This expedites the feedback delivery in response to the control. Furthermore, to reduce the waiting time of control arrivals at the ONUs, control bandwidth is allocated as in the existing predictive schemes introduced earlier. The control bandwidth estimation can be fulfilled using an ANN following the same training process above, which is not repeated justified here. Overall, for H2M applications, the CO allocates bandwidth to control traffic predictively and to the feedback interactively exploiting the correlation, thereby reducing for H2M applications.
3) *Priority-Differentiated Estimation and Allocation:* Since future networks need to support aggregation of H2M and conventional content traffic, appropriate differentiation in allocating bandwidth for H2M and content traffic, and in their transmission, is necessary. This is because H2M packets demand stringent low latency in communication. Unlike the existing predictive schemes that allocate estimated bandwidth to all aggregated arrivals to ONUs, in the ALL scheme, the CO allocates bandwidth to content traffic only equal to that in the report. Specifically, in the ALL scheme, an ONU buffers control, feedback, and content arrivals in separate queues. The number of packets in each queue, equivalently the bandwidth demand, is reported to the CO in each polling cycle. The CO allocates content bandwidth equal to that reported, grants surplus control bandwidth in addition to the request, and preallocates feedback bandwidth when forwarding control packets. In the transmission time slots of ONUs, the transmission of H2M packets in the buffer is prioritized in achieving low latency.

To show the implementation of the above three aspects, the pseudocode of the ALL scheme is provided in Algorithm 1. Upon receiving the REPORT from an ONU i in the kth polling cycle, the CO estimates control bandwidth, $\mathrm{BW}_{\mathrm{pred}}(\mathrm{ctr}, i, k+1)$, for ONU i and feedback bandwidth, $\mathrm{BW}_{\mathrm{pred}}(\mathrm{fbk}, j, k+1)$, for the counterpart ONU j (lines 2–7 in the pseudocode). Then, the CO grants $\mathrm{BW}_{\mathrm{pred}}(\mathrm{fbk}, j, k+1)$ to ONU j when forwarding the control traffic (lines 9–16). This is to expedite feedback delivery harnessing the correlation between the control and feedback. For ONU i that sends the REPORT, the total bandwidth $\mathrm{BW}_{\mathrm{grant}}(i, k+1)$, including both the requested and estimated bandwidth, is allocated (lines 17–24). In this way, priority-differentiated bandwidth estimation and allocation for content and H2M traffic are facilitated.

Algorithm 1 ALL Scheme: Pseudocode of CO Bandwidth Allocation

$t_{\mathrm{scheduled}}$ — time up to which the uplink channel has been scheduled
t_{local}— current system local time
RTT — round-trip transmission time
T_{g}, T_{process}, T_{REPORT} — guard time; processing time; REPORT
$t_{\mathrm{start}}(m, k)$ and $t_{\mathrm{end}}(m, k)$ — start and end time of granted timeslot for an ONU m in the k^{th} polling cycle
$t_{\mathrm{start\_fbk}}(m)$ and $t_{\mathrm{end\_fbk}}(m)$— start time of granted timeslot for haptic feedback traffic of an ONU m
R_{PON} — data rate in the optical access network
$\mathrm{BW}_{\mathrm{req}}(cont, i, k)$, $\mathrm{BW}_{\mathrm{req}}(ctr, i, k)$, and $\mathrm{BW}_{\mathrm{req}}(fbk, i, k)$ — content, control, and feedback bandwidth requested by ONU i in the k^{th} polling cycle
$\mathrm{BW}_{\mathrm{pred}}(ctr, i, k)$ and $\mathrm{BW}_{\mathrm{pred}}(\mathrm{fbk}, i, k)$ — estimated control and feedback bandwidth for ONU i in the k^{th} polling cycle
$\mathrm{BW}_{\mathrm{grant}}(i, k+1)$ — allocated bandwidth for ONU i

The operation at the CO upon receiving the REPORT in the k^{th} polling cycle from ONU i (repeat for all k and i):

1 {
2 // bandwidth estimation for a pair of M- and S-ONU
3 Read: $\mathrm{BW}_{\mathrm{req}}(cont, i, k)$, $\mathrm{BW}_{\mathrm{req}}(ctr, i, k)$ and $\mathrm{BW}_{\mathrm{req}}(\mathrm{fbk}, i, k)$
4 Bandwidth estimation by ANN (refer to Section IV-B:
5 AI for H2M bandwidth estimation):
6 $\mathrm{BW}_{\mathrm{pred}}(\mathrm{fbk}, j, k+1)$ for its counterpart S-ONU j
7 $\mathrm{BW}_{\mathrm{pred}}(\mathrm{ctr}, i, k+1)$ for the reporting ONU i
8 // schedule start and end time for feedback from S-ONU j
9 **if** $\mathrm{BW}_{\mathrm{pred}}(\mathrm{fbk}, j, k+1) > 0$
10 $t_{\mathrm{start\_fbf}}(j) = \min\{t_{\mathrm{local}} + RTT/2 + T_{\mathrm{process}},$
11 $t_{\mathrm{scheduled}} - RTT/2 - T_{\mathrm{process}}\}$;
12 $t_{\mathrm{end\_fbf}}(j) = t_{\mathrm{start\_fbf}}(j) + \mathrm{BW}_{\mathrm{pred}}(\mathrm{fbk}, j, k+1)/R_{\mathrm{PON}}$;
13 Send a GATE with $t_{\mathrm{start\_fbk}}(j)$ and $t_{\mathrm{end\_fbk}}(j)$ to ONU j.
14 // update scheduled time in the uplink channel
15 $t_{\mathrm{scheduled}} = t_{\mathrm{end\_fbk}}(j) + T_{\mathrm{guard}}$;
16 **End**
17 // schedule start and end time for ONU i
18 $t_{\mathrm{start}}(i, k+1) = \min\{LocalTime + RTT/2 + T_{\mathrm{process}}, t_{\mathrm{scheduled}}$
19 $-RTT/2 - T_{\mathrm{process}}\}$;
20 $\mathrm{BW}_{\mathrm{grant}}(i, k+1) = \min\{\mathrm{BW}_{\mathrm{req}}(cont, i, k) + \mathrm{BW}_{\mathrm{req}}(\mathrm{ctr}, i, k) +$
21 $\mathrm{BW}_{\mathrm{req}}(\mathrm{fbk}, i, k) + \mathrm{BW}_{\mathrm{pred}}(\mathrm{ctr}, i, k+1), \mathrm{BW}_{\mathrm{max}}$
22 $-\mathrm{BW}_{\mathrm{pred}}(\mathrm{fbk}, i, k)\}$;
23 $t_{\mathrm{end}}(i, k+1) = t_{\mathrm{start}}(i, k+1) + \mathrm{BW}_{\mathrm{grant}}(i, k+1)/R_{\mathrm{PON}} + T_{\mathrm{REPORT}}$;
24 $t_{\mathrm{scheduled}} = t_{\mathrm{end}}(i, k+1) + T_{\mathrm{guard}}$;
25 }

V. Performance Evaluation

A. Simulation Settings

To evaluate the performance of the ALL scheme, we implement packet-level simulation in MATLAB. A 1 Gb/s PON with 16 ONUs is considered. In our simulation, we pair two ONUs, e.g., ONUs i and j, and consider inter-ONU H2M communication between them. The control and feedback traces collected in Applications A–C are injected to ONUs to emulate H2M arrivals. Concurrently, noting that the characteristic of content traffic is widely investigated, such as in [10] and [38]–[40], we generate content arrivals to ONUs using synthetic traffic source following the Pareto distribution as detailed in [10]. Normalized network traffic loads from 0.1 to 1 are simulated. The maximum polling cycle duration is set as 1 ms.

We compare the average latency of H2M and content packet in: 1) the baseline scheme without priority; 2) the baseline scheme with priority, i.e., prioritizing H2M traffic; 3) the statistical predictive scheme that estimates the H2M and content bandwidth in total; 4) the statistical predictive scheme that only estimates H2M bandwidth; 5) the ML-based predictive scheme that estimates H2M bandwidth with an ANN; and 6) the proposed ALL scheme. In these schemes, the transmission of

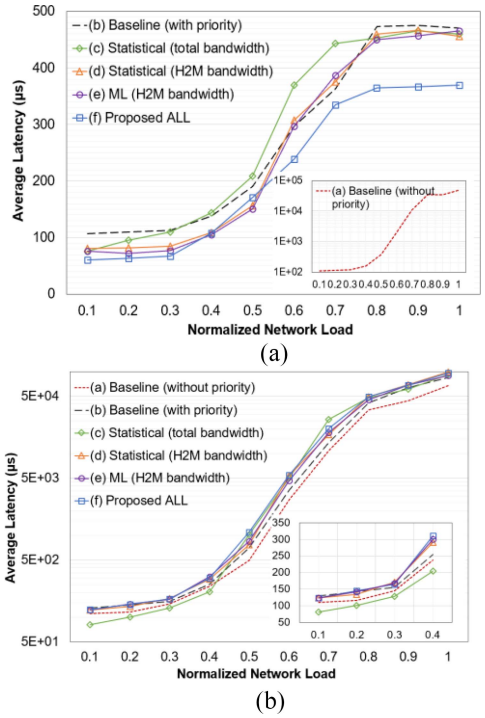

Fig. 9. Latency performance comparison. (a) H2M traffic. (b) Content traffic.

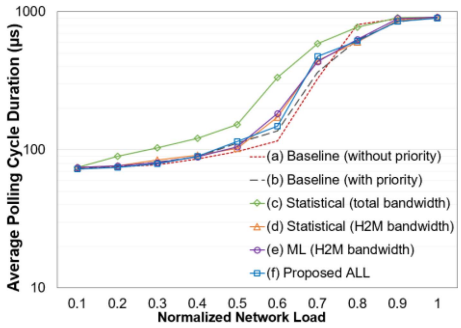

Fig. 10. Average polling cycle duration comparison.

H2M traffic is prioritized over content traffic in achieving low latency for H2M applications.

As explained in Section IV-A, we consider the use of ARMA [17] in schemes 3) and 4). The current predictive schemes typically consider bandwidth estimation for the total arrivals at ONUs. The comparison between schemes 3) and 4) is to provide insights in differentiating bandwidth estimation for H2M and content traffic. Next, the comparison between schemes 5) and 6) is to validate the effectiveness of the proposed interactive bandwidth allocation in the ALL scheme. In scheme 5), the H2M bandwidth estimation is the same as that in scheme 6), whilst the bandwidth allocation follows the report–grant process as that in schemes 1)–4).

B. Average Latency Comparisons

Fig. 9 compares the average uplink latency of H2M and content packets under these different schemes. In the baseline scheme (without priority), the latency of H2M and content traffic is equally high and increases with increasing network loads. When the network load exceeds 0.5, the average latency of both H2M and content traffic exceeds 1 ms. Prioritizing H2M traffic keeps the latency of H2M traffic below 500 μs in Fig. 9(a). This is because with priority, H2M packets can be transmitted without needing to be reported, reducing the average latency toward 0.5 polling cycle time. However, since the maximum polling cycle is 1 ms, neither the baseline nor predictive schemes 2)–4) can reduce latency from 500 μs in the high load region, i.e., network loads 0.7–1. In comparison, the ALL scheme effectively reduces this latency as it allocates feedback bandwidth in responding to the control. We plot the polling cycle durations in Fig. 10 and explained more details.

As shown in Fig. 9(a), in the light load region, i.e., network loads 0.1–0.3, adopting priority in the baseline scheme does not improve the latency performance compared to the no priority case. This is because when the content buffer is empty under light loads, the H2M packets experience the report–grant latency. As such, prioritizing H2M traffic in the baseline scheme sees benefits only when the network load is above 0.3 in Fig. 9(a). Compared to the baseline scheme 2), predictive schemes reduce the latency, depending on the network loads and the amount of surplus bandwidth estimated and allocated. This is illustrated by comparing the statistical predictive schemes 3) and 4). Fig. 9(a) shows that the H2M latency of scheme 3) is below the baseline schemes only under network loads $<$0.3, and increases dramatically with increasing network load. This is because the CO attempts to grant surplus bandwidth for both H2M and content packets arriving at an ONU. This prolongs the polling cycle compared to the rest schemes as shown in Fig. 10, and therefore causes the latency increase. On the other hand, the statistical scheme 4) reduces latency under network loads $<$0.5 in Fig. 9(a), since surplus bandwidth is granted only for H2M arrivals. This eliminates the report–grant process for H2M packets without affecting the polling cycle duration. Comparing to the statistical scheme 4), the ML scheme 5) slightly reduces the latency in the light load region, due to the use of ANN in bandwidth estimation. However, all these predictive schemes, i.e., schemes 3)–5), meet its bottleneck when network loads $>$0.5 in Fig. 9(a). This is attributed to the fact that the polling cycle duration increases steeply when the network load exceeds 0.5 as shown in Fig. 10. In this case, statistical and ML-based predictive schemes yield similar H2M latency with the baseline. In comparison, the ALL scheme reduces latency for H2M traffic under network loads 0.1–1 as shown in Fig. 9(a). This is attributed to the interactive control and feedback bandwidth allocation for H2M traffic.

The impact of prioritizing H2M traffic on the content traffic latency is presented in Fig. 9(b). Compared to the baseline scheme 2), the rest schemes lead to higher latency since content packets are deferred in the presence of H2M packets. The statistical scheme 3) is an exception in network loads 0.1–0.4. This is because surplus bandwidth is allocated for both H2M and content traffic. However, note that this latency reduction is at the cost of increasing the polling cycle and the H2M latency. When network loads $>$0.5, the baseline scheme 2) shows a lower latency over all the predictive schemes as it prevents bandwidth over granting.

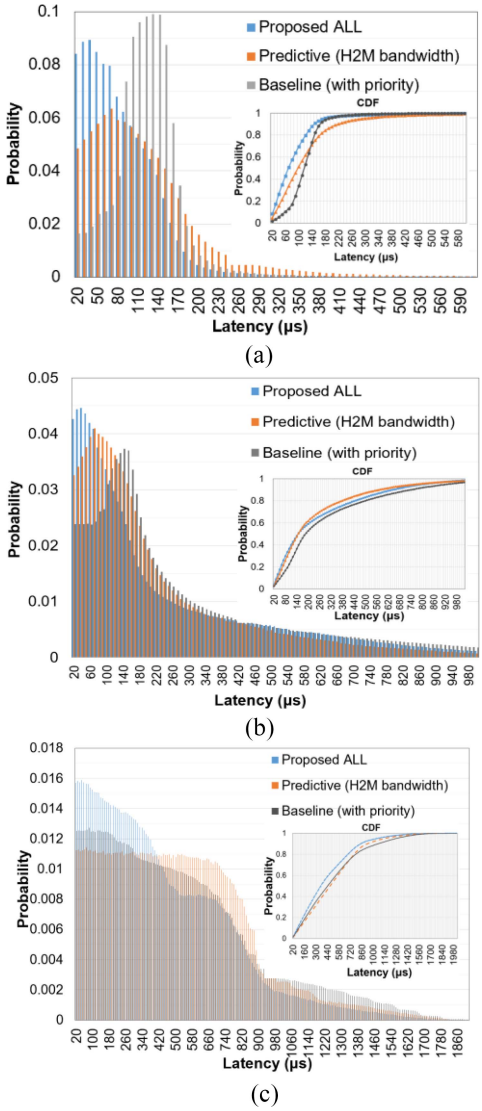

Fig. 11. Packet latency statistics. (a) 0.2 network load. (b) 0.5 network load. (c) 0.8 network load.

Overall, it can be viewed when the network load exceeds 0.5, the baseline scheme 2) could be a favorable choice than using predictive schemes 3)–5). This is because these schemes show similar H2M latency, but the latency of content traffic is lower in the baseline scheme 2). Among predictive schemes, we can see that using ML or statistical algorithms is not always the most important factor. For example, the performance of schemes 4) and 5) only differentiate in the light load region for H2M traffic. This is because the polling cycle duration approaches its maximum under high loads, limiting the effectiveness of the predictive schemes. Unwantedly, allocating surplus bandwidth for aggregated arrivals, i.e., scheme 3), increases the H2M latency. The ALL scheme effectively addresses the above issues via strategically allocating bandwidth for H2M traffic.

C. Latency Statistics of H2M Packets

Since H2M packets demand millisecond-low latency in their transmission, we investigate the capability of the existing schemes in constraining the latency of H2M packets. Based on the above evaluations, we narrow our focus to the baseline scheme (with priority), predictive scheme (H2M bandwidth), and the ALL scheme. In Fig. 11(a)–(c), we present H2M packet latency distributions and CDF under network load 0.2 (light load), 0.5 (moderate load), and 0.8 (high load), respectively.

The latency distributions under the studied schemes have distinct centers and shapes as plotted in Fig. 11. Compared to the baseline and predictive schemes, the distribution under the ALL scheme has the narrowest shape in Fig. 11(a)–(c). Referring to the CDFs in Fig. 11, the ALL scheme constrains a higher proportion of packets with latency <100 μs. Under the network load of 0.2, the peak of the distributions under baseline and predictive schemes occurs at the latency value of 1.5 and 0.5 polling-cycle times, respectively. This is consistent with our previous analysis. The H2M packets under the baseline scheme need to wait 0.5 polling cycle to be reported and another 1 polling cycle for the granted timeslot. In the predictive scheme, the latency is reduced toward 0.5 polling cycle time via allocating surplus bandwidth for H2M packets. The ALL scheme further reduces this latency since it adaptively allocates bandwidth in a polling cycle based on control and feedback correlation. More than 90% of H2M packets are transmitted within 150 μs in the ALL scheme, which is 10% and 15% higher than that in the baseline and predictive schemes. Under the network load of 0.5, similar CDFs are shown in Fig. 11(b). The predictive scheme is slightly better, constraining 3%–5% more H2M packets in 200–500-μs latency range compared to ALL scheme. Nonetheless, in both predictive and ALL schemes, more than 80% of H2M packets wait less than 500 μs in the buffer. Under the network load of 0.8 in Fig. 11(c), the latency distributions are nearly uniform within 1 ms under the baseline and predictive scheme. In comparison, the ALL scheme clearly shows a higher percentage of packets with latency <500 μs. This is again because the correlation between control and feedback is effectively exploited in allocating bandwidth for feedback. The packet latency statistics in Fig. 11 support our previous latency analysis on each scheme. The results presented in Figs. 9–11 validate the effectiveness of the ALL scheme in reducing latency for H2M applications.

VI. Conclusion

This article presented our experimental analysis on human control and haptic feedback traffic in H2M applications, and the proposed ALL scheme in supporting low-latency inter-ONU H2M communications. Using experimental traffic traces, we analyzed the statistical distributions and time correlation of control and feedback traffic. Then, we proposed the ALL scheme that allocates H2M bandwidth exploiting the traffic characteristics. Different from existing schemes, the ALL scheme expedited H2M packet delivery by estimating H2M bandwidth and adaptively allocating bandwidth to feedback corresponding to the control. Via extensive simulations, we compared the capability of existing schemes and the ALL scheme in reducing and constraining latency for H2M and content applications. The effectiveness of the ALL scheme was validated and several results were concluded: 1) the

report–grant process is the main latency cause in the baseline scheme; 2) the predictive schemes reduce the latency for H2M applications depending on bandwidth estimation methods and network loads; and 3) the ALL scheme effectively reduces latency for H2M applications compared to the existing schemes.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the Editor for their constructive feedback and suggestions that helped improve the quality of this article.

REFERENCES

[1] O. Holland *et al.*, "The IEEE 1918.1 'tactile Internet' standards working group and its standards," *Proc. IEEE*, vol. 107, no. 2, pp. 256–279, Feb. 2019.

[2] G. P. Fettweis, "The tactile Internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[3] ITU-T. (2014). *The Tactile Internet*. [Online]. Available: https://www.itu.int/dms_pub/itu-t/opb/gen/T-GEN-TWATCH-2014–1-PDF-E.pdf%

[4] M. Dohler *et al.*, "Internet of skills, where robotics meets AI, 5G and the tactile Internet," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

[5] M. Maier, M. Chowdhury, B. P. Rimal, and D. P. Van, "The tactile Internet: Vision, recent progress, and open challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 138–145, May 2016.

[6] M. Chowdhury and M. Maier, "Local and nonlocal human-to-robot task allocation in fiber-wireless multi-robot networks," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2250–2260, Sep. 2018.

[7] M. Chowdhury and M. Maier, "Collaborative computing for advanced tactile Internet human-to-robot (H2R) communications in integrated FiWi multirobot infrastructures," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2142–2158, Dec. 2017.

[8] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[9] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "Onmulti-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[10] G. Kramer, *Ethernet Passive Optical Networks*. New York, NY, USA: McGraw-Hill, 2005.

[11] C. Chen, H. Wu, and K. Ke, "Predictive credit based dynamic bandwidth allocation mechanisms in Ethernet passive optical network," in *Proc. TENCON*, 2006, pp. 1–4.

[12] M. P. I. Dias, B. S. Karunaratne, and 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, Jul. 2014.

[13] E. Wong, M. P. I. Dias, and L. Ruan, "Predictive resource allocation for tactile Internet capable passive optical LANs," *J. Lightw. Technol.*, vol. 35, no. 13, pp. 2629–2641, Jul. 1, 2017.

[14] Z. M. Fadlullah, H. Nishiyama, N. Kato, H. Ujikawa, K.-I. Suzuki, and N. Yoshimoto, "Smart FiWi networks: Challenges and solutions for QoS and green communications," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 86–91, Mar./Apr. 2013.

[15] Y. Luo and N. Ansari, "Limited sharing with traffic prediction for dynamic bandwidth allocation and QoS provisioning over EPONs," *J. Opt. Netw.*, vol. 4, pp. 561–572, Aug. 2005.

[16] H. Wang *et al.*, "LP-DWBA: A DWBA algorithm based on linear prediction in TWDM-PON," in *Proc. 14th Int. Conf. Opt. Commun. Netw. (ICOCN)*, Nanjing, China, 2015, pp. 1–3.

[17] K. Nishimoto *et al.*, "Predictive dynamic bandwidth allocation based on the correlation of the bi-directional traffic for cloud-based virtual PON-OLT," in *Proc. Int. Workshop Techn. Committee Commun. Qual. Rel.*, 2017, pp. 1–6.

[18] N. Hanaya, Y. Nakayama, M. Yoshino, K.-I. Suzuki, and R. Kubo, "Remotely controlled XG-PON DBA with linear prediction for flexible access system architecture," in *Proc. Opt. Fiber Commun. Conf.*, 2018, pp. 1–3.

[19] P. Sarigiannidis, D. Pliatsios, T. Zygiridis, and N. Kantartzis, "DAMA: A data mining forecasting DBA scheme for XG-PONs," in *Proc. 5th Int. Conf. Modern Circuits Syst. Technol. (MOCAST)*, May 2016, pp. 1–4.

[20] Y. Wu, M. Tornatore, Y. Zhao, and B. Mukherjee, "Traffic classification and sifting to improve TDM-EPON fronthaul upstream efficiency," *J. Opt. Commun. Netw.*, vol. 10, no. 8, pp. 15–26, 2018.

[21] J. A. Hatem, A. R. Dhaini, and S. Elbassuoni, "Deep learning-based dynamic bandwidth allocation for future optical access networks," *IEEE Access*, vol. 7, pp. 97307–97318, 2019.

[22] L. Ruan, M. P. I. Dias, and E. Wong, "Machine learning-based bandwidth prediction for low-latency H2M applications," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3743–3752, Apr. 2019.

[23] L. Ruan, M. P. I. Dias, M. Maier, and E. Wong, "Understanding the traffic causality for low-latency human-to-machine applications," *IEEE Netw. Lett.*, vol. 1, no. 3, pp. 128–113, Oct.–Dec. 2019.

[24] A. M. Mikaeil, W. Hu, and S. B. Hussain, "Traffic-estimation-based low-latency XGS-PON mobile front-haul for small-cell C-RAN based on an adaptive learning neural network." *Appl. Sci.*, vol. 8, no. 7, p. 1097, 2018.

[25] C. Pacchierotti, S. Sinclair, M. Solazzi, A. Frisoli, V. Hayward, and D. Prattichizzo, "Wearable haptic systems for the fingertip and the hand: Taxonomy, review, and perspectives," *IEEE Trans. Haptics*, vol. 10, no. 4, pp. 580–600, Oct.–Dec. 2017.

[26] H. Culbertson, S. B. Schorr, and A. M. Okamura, "Haptics: The present and future of artificial touch sensation," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 1, pp. 385–409, Feb. 2018.

[27] J. Kammerl, I. Vittorias, V. Nitsch, E. Steinbach, and S. Hirche, "Perception-based data reduction for haptic force-feedback signals using velocity-adaptive deadbands," *Presence*, vol. 19, no. 5, pp. 450–462, 2010.

[28] M. Condoluci, T. Mahmoodi, E. Steinbach, and M. Dohler, "Soft resource reservation for low-delayed teleoperation over mobile networks," *IEEE Access*, vol. 5, pp. 10445–10455, 2017.

[29] K. S. Kim *et al.*, "Ultrareliable and low-latency communication techniques for tactile Internet services," *Proc. IEEE*, vol. 107, no. 2, pp. 376–393, Feb. 2019.

[30] D. Feng *et al.*, "Toward ultrareliable low-latency communications: Typical scenarios, possible solutions, and open issues," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 94–102, Jun. 2019.

[31] Z. Hou, C. She, Y. Li, T. Q. S. Quek, and B. Vucetic, "Burstiness-aware bandwidth reservation for ultra-reliable and low-latency communications in tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2401–2410, Nov. 2018.

[32] M. Maier and A. Ebrahimzadeh, "Towards immersive tactile Internet experiences: Low-latency FiWi enhanced mobile networks with edge intelligence," *J. Opt. Commun. Netw.*, vol. 11, no. 4, pp. B10–B25, 2019.

[33] CyberGlove Systems Inc. (2017). *CyberGrasp*. [Online]. Available: http://www.cyberglovesystems.com/cybergrasp/

[34] T. T. T. Nguyen and G. Armitage, "A survey of techniques for Internet traffic classification using machine learning," *IEEE Commun. Surveys Tuts.*, vol. 10, no. 4, pp. 56–76, 4th Quart., 2008.

[35] P. Assimakopoulos *et al.*, "Statistical distribution of packet inter-arrival rates in an Ethernet Fronthaul," in *Proc. IEEE ICC Workshops*, 2016, pp. 140–144.

[36] C. Majumdar, M. Lopez-Benitez, and S. N. Merchant, "Accurate modelling of IoT data traffic based on weighted sum of distributions," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[37] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. NIPS*, 2007, pp. 153–160.

[38] W. H. Tranter *et al.*, "On the self-similar nature of Ethernet traffic (extended version)," in *Proc. IEEE Commun. Netw. Res.*, 2007, pp. 517–531.

[39] R. Fontugne *et al.*, "Scaling in Internet traffic: A 14 year and 3 day longitudinal study, with multiscale analyses and random projections," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2152–2165, Aug. 2017.

[40] P. Cortez, M. Rio, M. Rocha, and P. Sousa, "Multi-scale Internet traffic forecasting using neural networks and time series methods," *Expert Syst.*, vol. 29, no. 2, pp. 143–155, 2012.

[41] W. Wang *et al.*, "A network traffic flow prediction with deep learning approach for large-scale metropolitan area network," in *Proc. IEEE/IFIP Netw. Oper. Manag. Symp. (NOMS)*, 2018, pp. 1–9.

[42] A. M. Mikaeil, W. Hu, and S. B. Hussain, "A low-latency traffic estimation based TDM-PON mobile front-haul for small cell cloud-RAN employing feed-forward artificial neural network," in *Proc. Int. Conf. Transp. Opt. Netw.*, 2018, pp. 1–4.

[43] L. Ruan, M. P. I. Dias, and E. Wong, "Enhancing latency performance through intelligent bandwidth allocation decisions: A survey and comparative study of machine learning techniques," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 12, no. 4, pp. B20–B32, Apr. 2020.