Intelligent Traffic Adaptive Resource Allocation for Edge Computing-Based 5G Networks

Min Chen, *Senior Member, IEEE*, Yiming Miao, *Member, IEEE*, Hamid Gharavi, *Life Fellow, IEEE*, Long Hu, *Member, IEEE*, and Iztok Humar, *Senior Member, IEEE*

***Abstract*—The popularity of smart mobile devices has led to a tremendous increase in mobile traffic, which has put a considerable strain on the fifth generation of mobile communication networks (5G). Among the three application scenarios covered by 5G, ultra-high reliability and ultra-low latency (uRLLC) communication can best be realized with the assistance of artificial intelligence. For a combined 5G, edge computing and IoT-Cloud (a platform that integrates the Internet of Things and cloud) in particular, there remains many challenges to meet the uRLLC latency and reliability requirements despite a tremendous effort to develop smart data-driven methods. Therefore, this paper mainly focuses on artificial intelligence for controlling mobile-traffic flow. In our approach, we first develop a traffic-flow prediction algorithm that is based on long short-term memory (LSTM) with an attention mechanism to train mobile-traffic data in single-site mode. The algorithm is capable of effectively predicting the peak value of the traffic flow. For a multi-site case, we present an intelligent IoT-based mobile traffic prediction-and-control architecture capable of dynamically dispatching communication and computing resources. In our experiments, we demonstrate the effectiveness of the proposed scheme in reducing communication latency and its impact on lowering packet-loss ratio. Finally, we present future work and discuss some of the open issues.**

***Index Terms*—5G, artificial intelligence, LSTM, mobile traffic, uRLLC.**

I. Introduction

THE FIFTH generation of mobile communication networks (5G) with leveraging real-time artificial intelligence (AI) is a promising prospect towards fulfilling the high-level requirements of the heterogeneous Internet of Things (IoT). The 3rd Generation Partnership Project (3GPP) defines three emerging application scenarios: enhance mobile broadband (eMBB), massive machine-type communication (mMTC), and uRLLC [1]. While eMBB focuses on high spectral efficiency, uRLLC, has been recognized as a key trend of 5G, but faces tough challenges in order to meet the stringent requirements, such as high reliability and low latency. As uRLLC is becoming increasingly prevalent, exploiting new technologies such as AI, would be essential to tackle more complex tasks in the presence of high traffic flow. This trend is driven by the ever increasing popularity of smart devices, which are having a huge impact on the trade-offs between the services of telecommunications suppliers and users' demands [2]. For instance, user's quality of experience (QoE) is an essential prerequisite for providing intelligence and personalized services for real-time interactions. Within the integrated framework of the IoT and the cloud (IoT-Cloud), the anticipated increase in mobile-traffic flow can also have a profound effect on computation intensity, as well as dispatching pressure on the edge cloud (base-station) and remote cloud (data center) [3], [4].

In addition, the intensified mobile-traffic flow can cause a shortage of computing and networking resources in which case applications might not be able to respond to users' requests in a timely manner. Another important issue is the allocation of bandwidth resources for heterogeneous (cloud) services [5]. For instance, the proliferation of resources to support intelligent services and the gradual transition from traditional to heterogeneous IoT, creates conflicting demands for network operators and service providers. For the IoT-cloud in particular, existing data-driven methods directly unload computing tasks without being analyzed, processed, and controlled by base-stations. These can greatly undermine transmission reliability, as well as impact communication latency for delay sensitive applications [6]. Since such a high latency cannot be tolerated by uRLLC users, the main challenge is how to efficiently analyze and control the mobile-traffic-flow data in order to achieve communications with ultra-low latency and ultra-high reliability.

With the anticipated integration of 5G networks, edge computing and the IoT-Cloud, AI assisted mobile-traffic-flow, prediction, and management can offer a viable solution for future mobile-network planning and dynamic resource allocation. In view of recent advances in smart mobile devices,

Manuscript received August 21, 2019; revised October 18, 2019; accepted November 1, 2019. Date of publication November 12, 2019; date of current version June 9, 2020. This work is supported by the National Key R&D Program of China (2017YFE0123600). Prof. Long Hu's work was supported by the China National Natural Science Foundation (No. 61802139). Dr. Humar would like to acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0246). This research is partially supported by the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS). The associate editor coordinating the review of this article and approving it for publication was Y. Wu. *(Corresponding author: Long Hu.)*

M. Chen is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: minchen@ieee.org).

Y. Miao and L. Hu are with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: yimingmiao@hust.edu.cn; hulong@hust.edu.cn).

H. Gharavi is with the Department of Advanced Network Technology, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: hamid.gharavi@nist.gov).

I. Humar is with the Faculty of Electrical Engineering, University of Ljubljana, 1000 Ljubljana, Slovenia (e-mail: iztok.humar@fe.uni-lj.si).

Digital Object Identifier 10.1109/TCCN.2019.2953061

base-stations, and remote clouds, together with the latest development of computing and storage techniques, AI can play a significant role in supporting uRLLC scenarios in order to meet its strict requirements in terms of latency and reliability. Depending on advanced machine-learning (ML) methods, AI is transforming from traditional pattern recognition to the management of complex systems. In the past few decades, ML rose and fell several times as the main branch of AI. It has now reached sufficient maturity that it can be permeated into the design of many complex systems, including routing optimization of wireless communications [7]. If algorithms, such as LSTM [8], [9], are deployed at different positions in mobile networks, mobile-traffic flow can be predicted intelligently. This is because LSTM is suitable for processing and predicting important events with relatively long intervals in the time series [10].

Therefore, in our proposed frame work, an LSTM-based deep-learning algorithm is considered to predict the uRLLC mobile-traffic flow received by a single edge cloud. The predicted peak value, representing the traffic of the entire network at each time instant, is sent to a remote cloud. At the remote cloud, resources are dispatched and allocated dynamically based on traffic adaptation using a cognitive engine and an intelligent mobile traffic module to balance the network load. This contributes towards achieving high reliability and low latency of communications, hence enhancing the users' QoE.

This paper is organized as follows. In Section II, after a brief description of the three application scenarios of 5G, we present an IoT-Cloud architecture focusing on the uRLLC application scenario. Then, a description of the mobile-traffic flow is given to disclose the interactions between user devices, edge clouds, and remote clouds. Section III proposes a uRLLC mobile-traffic flow prediction algorithm that exploits an LSTM-base algorithm operating in a single-site mode. In Section IV, we present a novel intelligent mobile-traffic control architecture for a large-scale multi-site IoT-Cloud. Experiments and performance evaluations of the proposed mobile-traffic prediction and control framework are given in Section V. Section VI discusses some open issues about users' mobility prediction, strategy sharing and risk perception. Finally, Section VII summarizes the study.

II. A Heterogeneous Internet of Things Based on uRLLC

A. The Three Application Scenarios of 5G

As mentioned before, the three core services of 5G, which have been defined by 3GPP are: eMBB, mMTC and uRLLC [1]. The applications covered by these scenarios are shown in Fig. 1. The aim of eMBB is to enhance users' QoE based on existing mobile broadband business scenarios. It mainly attempts to achieve an ideal short distance and personal communication. The main use case scenarios of eMBB are smart homes, VR/AR, smart devices, smart buildings, etc. Mobile broadband businesses with large traffic flows, such as 3D/UHD video, need the support of broadband resources to guarantee smooth transmission of the mobile-data [11].

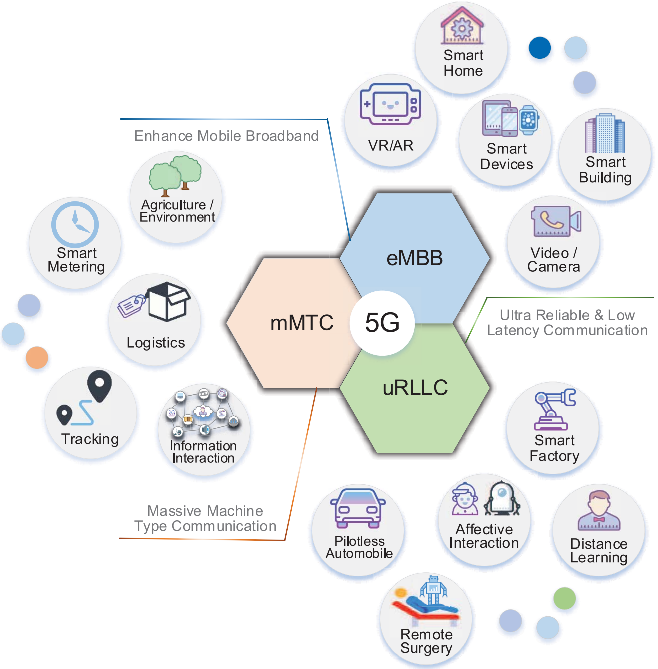


Fig. 1. Application scenarios of eMMB, mMTC and uRLLC.

The mMTC is intended to support IoT on a large scale radio coverage among humans and objects. Its main features are low cost, low energy consumption, short packets with small quantity of data, and a large number of connections. The mMTC supports applications such as the agricultural environment, smart metering, smart cities, logistics, tracking, massive information interaction, etc.

Finally, uRLLC aims to improve users' QoE [12] by establishing a highly reliable and stable communications. At the same time, uRLLC has to meet tough requirements with respect to network latency. The main application scenarios of uRLLC are smart factories, autonomous automobiles, remote surgery, affective interaction, etc. These application scenarios require reliable communications among terminal equipment, base-stations, and the cloud server, as well as intelligent computing and dynamic resource allocation.

In this paper, we mainly concern the prediction and control of mobile-traffic flow to support uRLLC services in terms of high reliability, low latency, and extremely high usability. Bear in mind that except for some regulated daily log data transmissions, uRLLC users' service requests on information transmissions are generally unpredictable and would require an advanced knowledge of traffic flow in order to efficiently execute resource-allocations and task computations. Due to the advantages of time-series prediction, LSTM is a natural approach to solving the problem of mobile-traffic-flow prediction in uRLLC scenarios. The algorithm can be deployed in edge clouds and remote clouds.

B. IoT-Cloud Architecture Based on uRLLC

Our proposed IoT-cloud architecture for uRLLC scenario is shown in Fig. 2. Further details of the architecture are described as follows:

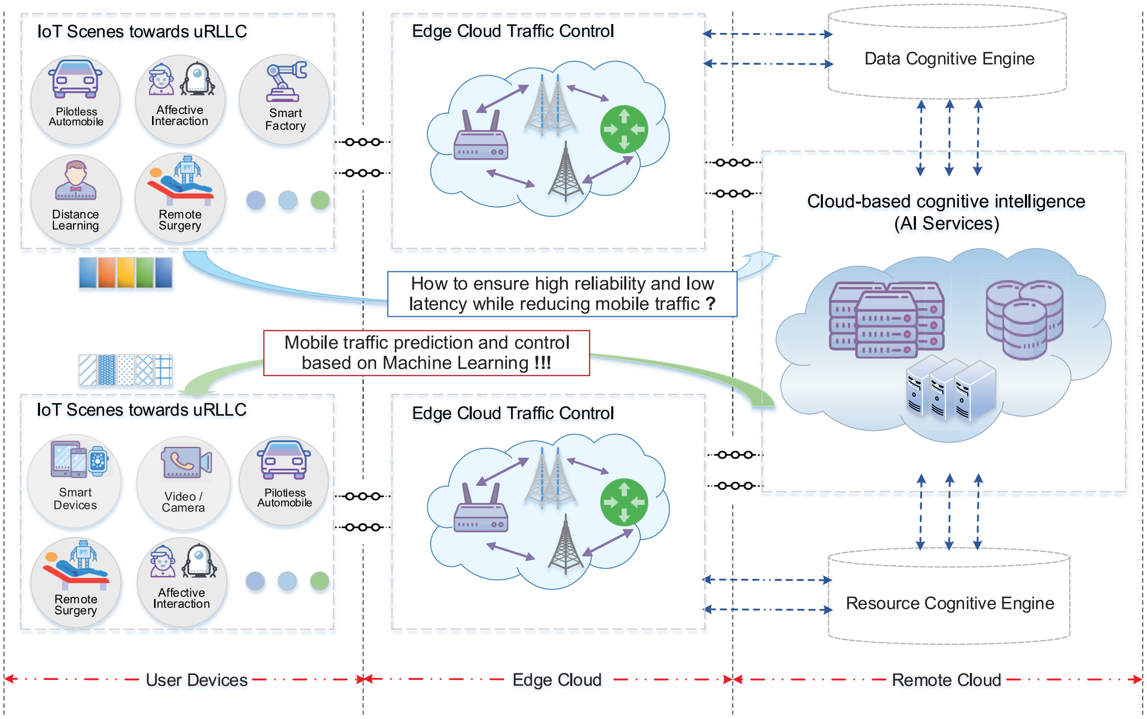

IoT Scenes towards uRLLC
Pilotless Automobile
Affective Interaction
Smart Factory
Distance Learning
Remote Surgery
Edge Cloud Traffic Control
Data Cognitive Engine
Cloud-based cognitive intelligence (AI Services)
How to ensure high reliability and low latency while reducing mobile traffic ?
Mobile traffic prediction and control based on Machine Learning !!!
IoT Scenes towards uRLLC
Smart Devices
Video / Camera
Pilotless Automobile
Remote Surgery
Affective Interaction
Edge Cloud Traffic Control
Resource Cognitive Engine
User Devices
Edge Cloud
Remote Cloud

Fig. 2. IoT-Cloud architecture based on uRLLC.

- The user-device layer mainly covers common IoT scenarios aimed at uRLLC (details in Section II-A) where each smart device requests a uRLLC-based service with an uncertain type and volume of data which would be uploaded to the edge cloud via a wireless network.
- The edge-cloud layer contains wireless connectors such as base-stations, access points, and routers. The edge cloud, which provides wireless access to each IoT device, performs lightweight caching and computing as well as unloading complex computing tasks to the remote cloud.
- The remote-cloud layer usually contains SDN [13], [14] controllers and a central node with the deployment of cognitive intelligence engine [15]. The available resources, with the help of data-cognitive engine on the cloud server, are utilized to perform mobile-traffic-flow prediction and dynamic resource allocation. The internal modules and functionality of the cognitive engine will be introduced in Section IV.

Traditional traffic prediction algorithms, such as short-term correlation prediction methods (Markov model, autoregressive model (AR) and its variants (ARMA, ARIMA)), mainly focus on modeling of stationary flow characteristics. However, these algorithms require high self-similarity of traffic time series. For example, AR algorithm adopts linear mapping method which uses p-order autoregressive model $AR(p) : V_t = c + \sum_{i=1}^{p} \phi_i V_{t-i} + \varepsilon_t$. Thereinto, the current flow value of V_t is expressed as the sum of a linear combination of one or more historical flow values of V_{t-i}, a constant term of c and a random error of ε_t. Although AR algorithm needs little data, the dependence on its own variable series easily leads to inaccurate prediction results. Therefore, the extended ARIMA algorithm adds a sliding average coefficient of q, a difference number of times of d, and a lag operator of L. $ARIMA(p,d,q)$ $(1-\sum_{i=1}^{p} \phi_i L^i)(1-L)^d V_t = (1+\sum_{i=1}^{q} \theta_i L^i)\varepsilon_t$. However, ARIMA algorithm does not consider other relevant variables in the process of modeling which is vulnerable to the impact of network dynamic changes.

In other words, existing traffic prediction and network optimization methods mainly face the following three difficulties: 1) Human intervention, traditional network optimization modelling is constructed by experts with domain knowledge. Such knowledge-driven model is costly and inefficient in implementation, 2) invalid model. More users, more access methods, more complex functions and network entities make the mathematical model and relationship functions constructed in advance unable to match the reality, 3) high complexity. From the history of prediction algorithms, we can see that researchers often add more variables to the relationship function to improve the effect of dynamic prediction, but at expense of multidimensional computing complexity.

In fact, the above problems can be solved by balancing the mobile traffic data and reducing bandwidth occupancy in the same period. Such a reduction can be achieved by efficiently predicting mobile-traffic flow and assessing the priority of the users' requests in transmission queues with dynamically allocating the resources for communication and computing. Therefore, under the proposed IoT-cloud architecture, our main objective is to develop an efficient machine-learning-based prediction and control algorithm that can ensure high reliability and low latency as required for uRLLC communications.

III. Mobile-Traffic-Flow Prediction Based on LSTM

In the uRLLC scenario, information (in terms of service requests and data communications) is produced in large quantities. This information is transmitted to the edge cloud via

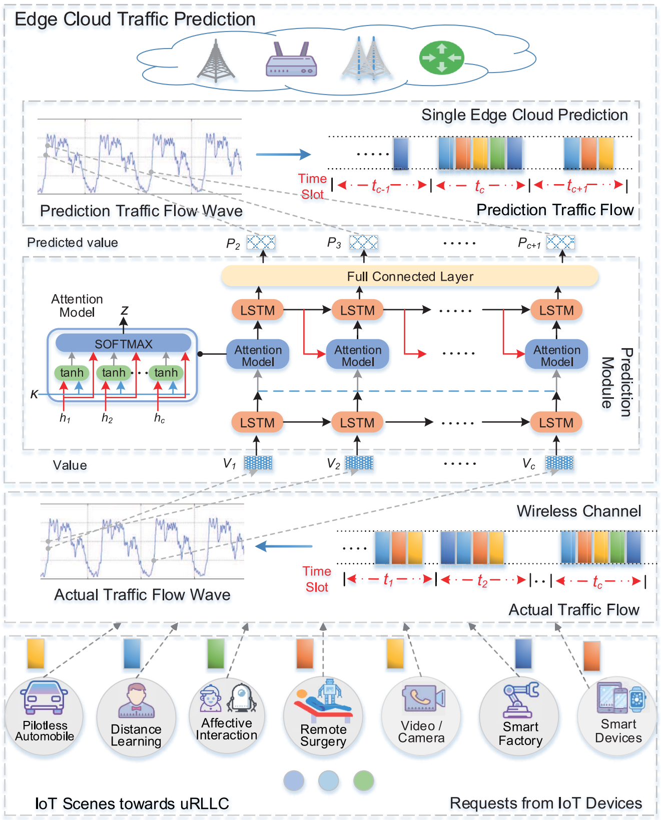


Fig. 3. uRLLC mobile-traffic-flow prediction of single-site based on LSTM.

wireless channels in the form of packets. Bear in mind that the network's bandwidth are generally limited and if the data from users is too excessive at certain time intervals, the network will block any transmissions without any alternative. This could consequently undermine the communication integrity due to packet loss. A viable approach to improve the communication performance would to develop an efficient prediction scheme that can dynamically manage the mobile-traffic flow.

We use LSTM algorithm to build a uRLLC mobile-traffic-flow prediction model of a single edge cloud [8], [9], as shown in Fig. 3. Furthermore, our proposed prediction module is made up of several stacks of attention-based LSTM layers. Each layer operates as the corresponding part of the basic LSTM cell with a attention mechanism.

A. LSTM Cell

First, within each time interval, we extract the peak value from the actual traffic-flow wave serial data, i.e., the input data set, $V = \{V_1, V_2, \dots, V_t\}$, where V_t is the peak mobile-traffic flow at the current time slot t_c. We seek to predict the peak mobile-traffic flow V_{t+1} at the next time instant t_{c+1}. The solution process of a single LSTM cell is presented in the following formula (1) [10].

$$\begin{cases} f_t = \sigma\big(W_f \cdot [h_{t-1}, V_t] + b_f\big), \\ i_t = \sigma\big(W_i \cdot [h_{t-1}, V_t] + b_i\big), \\ \tilde{C}_t = \tanh\big(W_C \cdot [h_{t-1}, V_t] + b_C\big), \\ C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t, \\ o_t = \sigma\big(W_o \cdot [h_{t-1}, V_t] + b_o\big), \\ h_t = o_t \times \tanh(C_t). \end{cases} \tag{1}$$

where, f_t, C_t, h_t, signify the forgetting gate, the input gate, and the output gate, respectively. W_f, W_C and b_f, b_C, represent the weights and the bias factors of the forgetting gate and the input gate, respectively. Furthermore, σ and $\tanh$ are activation functions. The flow feature h_t is extracted from a single LSTM model.

B. Attention Model

An attention model is widely used in natural language processing (NLP) area [16]. It is used to improve the neural network machine translation (NMT) as a sequence to sequence (encoder to decoder) model as in the case of the human brain, which pays different attentions to different events. In this section, we introduce the attention mechanism to improve the weight of the peak value in mobile traffic flow prediction. Such a mechanism aims to exploit any correlation between some key input features and output values.

Assume the output set of i^{th} LSTM layer is $h^i = \{h_1^i, h_2^i, \dots, h_t^i\}$, where $i \in \{1, 2, \dots, L\}$. Therefore, the intermediate code $\kappa = h^{i-1}$ represents the output state of a series LSTM cells in $(i-1)^{th}$ layer, where $h_t^{i-1} = F(C_{t-1}^{i-1}, h_t^{i-2}, h_{t-1}^{i-1})$. $F(\cdot)$ indicates the LSTM process calculated by formula (1). When h_{t-1}^i is fed into the attention model, the intermediate code κ is used to conduct a *tanh* process to obtain the aggregation state m_t^i, as shown below:

$$m_t^i = tanh\Big(W_{\kappa m}\kappa + W_{hm} h_{t-1}^i\Big) \tag{2}$$

where, $W_{\kappa m}$ and W_{hm} are the weights of the above two inputs. In order to get the highest value for m_t^i, we compute weights by a softmax function as:

$$softmax\Big(m_t^i\Big)_n = \frac{\exp\big(m_n^i\big)}{\sum_j \exp\Big(m_j^i\Big)} \tag{3}$$

As m_t^i becomes larger, $softmax(m_t^i)$ converges more towards $argmax(m_t^i)$. Then, we define α_t^i as the mapping of $softmax(m_t^i)$ in learning direction as shown below:

$$\alpha_t^i = \frac{\exp\big(\omega_m^T m_t^i\big)}{\sum_{t=1}^{T} \exp\big(\omega_m^T m_t^i\big)} \tag{4}$$

where, ω_m^T represents the transfer of matrix set $[W_{\kappa m}, W_{hm}]$. It should be noted that the value of α_t^i tends to increase at high peak time, while dropping at a lower peak time. Thus, the output z_t^i of the attention model can be defined as:

$$z_t^i = \sum_t \alpha_t^i h_t^i \tag{5}$$

Finally, we use the combination of the attention model output z_t^i and the LSTM cell output h_{t-1}^i from the previous time as the input data for the current time slot, i.e., $V_t^i = concat(h_{t-1}^i, z_t^i)$, where such a process will be loop execution until $i = L$.

C. Output Layer

A fully connected layer synthesizes the above-extracted features $\{h_1^L, h_2^L, \dots, h_t^L\}$ in L^{th} layer to obtain the output

Algorithm 1 Training Process of Mobile-Traffic Prediction

```
Input:
    Input Real traffic flow $V$;
    LSTM layers $L$
    Memory time slots n_input
    Number of iterations epoch
    Quantity of data input for each time batch_size.
Output:
    Data normalization
    $\widetilde{V_t} \leftarrow Rescaled(V_t) \leftarrow \frac{V_t - E_{min}}{E_{max} - E_{min}} \times (max - min) + min$
    Initial hidden-layer output $h$ and LSTM state $o$
    Create training set and reshape input data
    $input\_X \leftarrow \{[[\widetilde{V_1}, \widetilde{V_2}, \ldots, \widetilde{V}_{n\_input}],$
                $[\widetilde{V_2}, \widetilde{V_3}, \ldots, \widetilde{V}_{n\_input+1}],$
                $\ldots,$
                $[\widetilde{V}_{t-n\_input-1}, \widetilde{V}_{t-n\_input}, \ldots, \widetilde{V_{t-1}}]]\}$
    $input\_Y \leftarrow \{\widetilde{V}_{n\_input+1}, \widetilde{V}_{n\_input+2}, \ldots, \widetilde{V_t}\}$
    LSTM_Cell ← formula (1)
    Repeat
    for $i = l$ in range(L)
        for step in range(n_input × batch_size)
            $h_t, o_t \leftarrow LSTM\_Cell(input\_X_{t-1}, o_{t-1})$
            if $i \neq L$
                $z_t \leftarrow formula(2 \sim 5)$
                $input\_X_t \leftarrow V_t \leftarrow concat(h_{t-1}, z_t)$
        If $Loss(h_{t+1}, input\_Y_t) \propto 0$
            $\widetilde{P} \leftarrow \{h_2, h_3, \ldots, h_{t+1}\}$
        End if
    Until $RMSE(\widetilde{P}, V) \propto 0 \leftarrow formula(6)$
    P ← reverse_transform $(\widetilde{P})$
```

sequence $P = \{P_2, P_3, \ldots, P_{t+1}\}$, where P_{t+1} corresponds to the predicted value of the mobile-traffic flow at the next time slot, $t + 1$. By using a combination of the peak flow values from the previous time slots, we can then predict the value for every moment of mobile-traffic flow in a time span (i.e., the predicted traffic-flow wave).

To maximize the accuracy rate of the prediction (i.e., minimizing the root-mean-square error, which is used to measure the deviation between the observed value and the true value), the network should be able to accurately allocate the communication resources in order to prevent congestion according to the following formula (6).

$$RMSE(P, V) = \sqrt{\frac{1}{T} \sum_{i=t+1}^{T} (P_i - V_i)^2} \propto 0 \qquad (6)$$

Algorithm 1 shows the detailed training process of our intelligent mobile-traffic prediction using the proposed attention-based LSTM algorithm. The storage and computing capacity of edge clouds (i.e., base-stations or wireless access points) are usually insufficient, so a smaller and more accurate learning model is more suitable for high-reliability and low-latency services. Compared with the existing works, [8] uses a single LSTM cell to conduct prediction and [9] doesn't include multi-layer mechanism.

Thereinto, L can be dynamically adjusted according to the processing capacity of the edge nodes, whch makes our solution more versatile. In particular, as will be described in Section IV, our solution includes dynamic traffic control mechanism based on the prediction result which is not fully incorporated in [8] and [9]. Thus, our attention-based LSTM algorithm and IoT-Cloud architecture is more suitable for intelligent mobile traffic prediction and control.

IV. Intelligent Mobile-Traffic-Flow Control for IoT-Cloud Architecture

A single edge cloud can only serve a limited number of users (i.e., a cell of a mobile network). In large cities where thousands of users or equipment are involved in many mission critical tasks, cooperation and interaction among multi-cells, multi-edge clouds, and the remote cloud can be crucially important. For instance, a rapid increase in the number of users and equipment accessing the Internet can cause an unprecedented rise in mobile-traffic volume. This can easily affect the load-balancing, hence severely undermining the integrity of the entire network. More importantly, it goes against the goal of achieving high reliability and low latency as required by the uRLLC. Consequently, this has prompted us to further extend our intelligent resource allocation with mobile-traffic-flow prediction and control that goes beyond a single-site structure. Therefore, we propose a multi-site IoT-cloud extension for an intelligent mobile-traffic control as described below.

A. Multi-Site IoT-Cloud Architecture

Fig. 4 shows our proposed multi-site cloud structure where an IoT-Cloud module uses the same concept presented earlier in Section II-B (i.e., an IoT-Cloud architecture based on uRLLC). In a large-scale network supporting uRLLC, each cell contains intelligent equipment of an uncertain quantity. While each equipment can communicate with the edge cloud at the edge of a cell, it can also communicate with other edge clouds. Normally, each equipment tends to select the edge-cloud offering the best communication link (usually, the nearest one) within a given time frame unless the remote cloud dispatches it in advance, based on the network congestion. In this case users will lose the accuracy rate due to packet loss when communicating with others. Often, it is called network switching.

In the proposed multi-site scheme, each edge cloud predicts the mobile-traffic flow as described in Section III and reports the network status to the remote cloud. The remote cloud then invokes the mobile-traffic-flow control process by intelligently dispatching and allocating the communication and computing resources across the whole network. This enables more efficient interactions between users.

B. Cognitive Engine

We introduce a cognitive engine to implement a high-performance artificial intelligence algorithm (including the mobile-traffic-flow prediction algorithm based on our attention-based LSTM scheme as described in Section III) with the ability to store users' data in large quantities. With the help

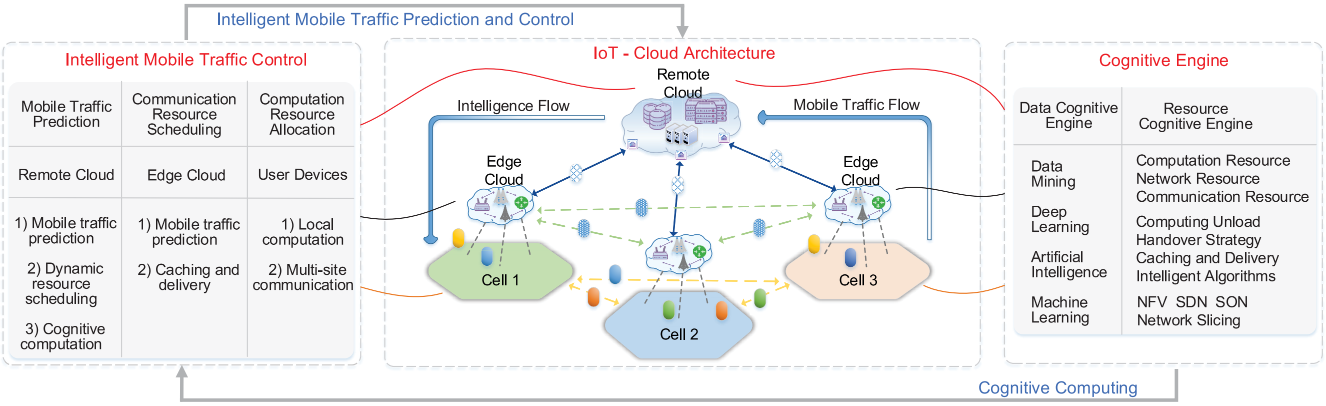


Fig. 4. Intelligent mobile-traffic-flow control for IoT-cloud architecture.

of Internet of Things traffic data in the edge clouds and remote clouds and our dynamic traffic-flow prediction, computing and data analysis can be offered with high precision. The cognitive engine can be divided into two types: resource cognitive engine and data cognitive engine. The detailed functions of the cognitive engine can be found in our previous study [20].

C. Intelligent Mobile-Traffic Control

There are three functional elements needed to meet the uRLLC requirements. These are: mobile-traffic prediction, communication resource scheduling, and computational resource allocation. To accurately execute these functional elements of the remote cloud, the edge cloud and user devices must be well synchronized. More specifically, a user device can reduce the flow of the backbone network when operating with simple local computing and offline caching. Furthermore, by participating in multi-site communications (with the help of network switching) mobile equipment can avoid local congestion. Edge clouds can cache and forward data in a lightweight manner. They can communicate with surrounding nodes, predict the mobile-traffic flow of a single cell, and notify the remote cloud in advance as an early warning in case of high-peak traffic flow. After completing complex cognitive computing, the remote cloud predicts the mobile-traffic flow on a large scale and dynamically dispatches the resources. With the help of a cognitive engine, balancing loads and stable communications of the whole network can be accomplished and the computing results returned as a feedback.

1) Edge-Cloud Selection: In order to realize an intelligent mobile-traffic control based on the traffic-prediction algorithm of Section III, we define a simple edge-cloud selection algorithm. First, we assume that the predicted traffic value for every edge cloud in the next time slot is: $Traf_i = \{Traf_1, Traf_2, \ldots, Traf_E\}$, where E represents the total number of edge clouds. This means that the predicted traffic flow of the whole network received by the cloud is: $Traf_C = \sum_{i=1}^{E} Traf_i$. Next, we assume sgn_i represents the current signal strength of a user accessing the i^{th} edge cloud. The choice of a user to access to the i^{th} edge cloud can be shown as:

$$Edge_{ac} = max\left(\left(1 - \frac{Traf_i}{Traf_C}\right) + sgn_i\right) \tag{7}$$

This means that users will choose to send requests or offload computational tasks to the edge cloud with a lower traffic flow or a better signal strength and this would be beneficial for the load balancing of the network.

2) Traffic-Adaptive Resource Allocation: In order to reduce the average delay of a single edge cloud and ensure data-transmission efficiency, we define a traffic-adaptive resource-allocation mechanism within an edge cell. We dynamically allocate a sub-carrier power to every device based on the predicted peak traffic flow [17], as shown below:

$$arg\ max \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{c_{k,n}}{N} log_2 \left[1 + \frac{p_{k,n}|h_{k,n}|^2}{N_0 B/N}\right] \tag{8}$$

To ensure fairness amongst users, we also introduce the following proportional fairness constraints:

Constraint:

$$\begin{aligned} C_1 &: \forall\ k, n, p_{k,n} \geq 0 \\ C_2 &: \forall\ k, n, c_{k,n} \in \{0, 1\} \\ C_3 &: \forall\ n, \sum_{k=1}^{K} c_{k,n} = 1 \\ C_4 &: R_1 : R_2 : \cdots : R_K = r_1 : r_2 : \cdots r_k \end{aligned} \tag{9}$$

where K is the total number of devices, N represents the total number of available sub-carriers, and B represents the total system bandwidth. Letąŕs assume the channel response of sub-carrier n for device k is $h_{k,n}$. Then, the amplitude of the channel response is $|h_{k,n}|$ and the channel-gain matrix is $H = \{|h_{k,n}|^2, k = 1, 2, \ldots, K, n = 1, 2, \ldots, N\}$. $N_0 B/N$ represents the noise power of each sub-carrier, $c_{k,n}$ is the indicative factor. Constraint C_2 represents a range of sub-carrier power allocation, i.e., $c_{k,n} = 1$ means that the sub-carrier n is allocated to the device k, while $c_{k,n} = 0$ indicates otherwise. $p_{k,n}$ is the allocated power of the sub-carrier n for the device k, which is also the power resource we need to optimize. C_4 represents the time-slot proportional fairness constraint where the data rate R_k of device k can be expressed as:

$$R_k = \sum_{n=1}^{N} c_{k,n} log_2 \left[1 + \frac{p_{k,n}|h_{k,n}|^2}{N_0 B/N}\right] \tag{10}$$

When the network loss is l in time slot t, we can easily obtain the real data rate P_k of device k, according to the following formula:

$$P_k = R_k(1-l) \tag{11}$$

The objective function of the above optimization model is the system and rate capacity, i.e., the predicted traffic flow P_t of the edge cloud in time slot t, as shown in formula (12).

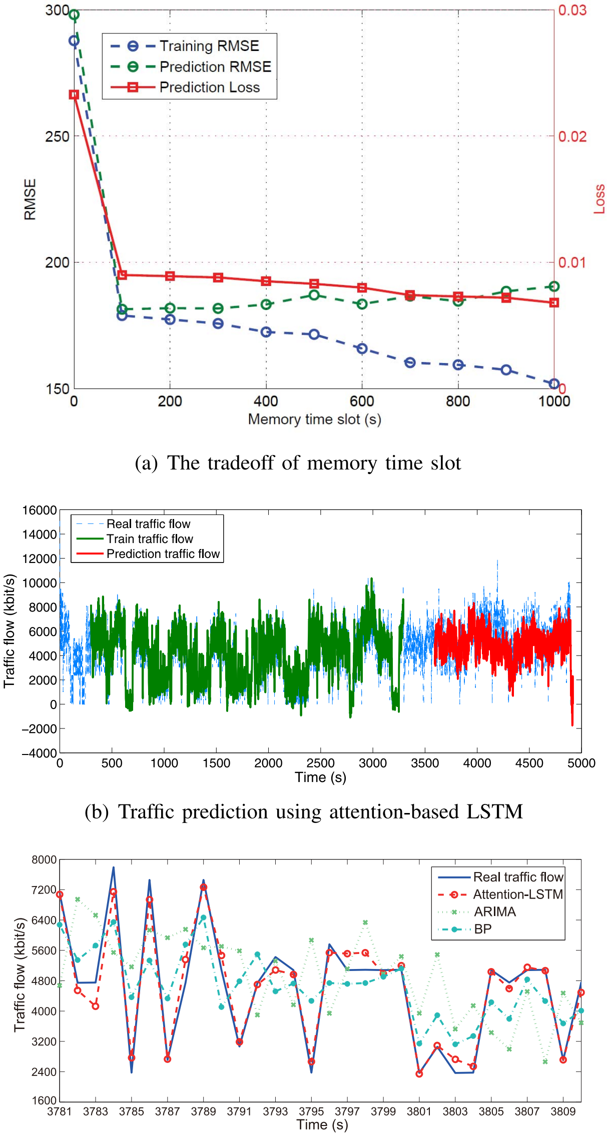
$$\begin{aligned} P_t &= \sum_{k=1}^{K} P_k = \frac{\sum_{k=1}^{K} R_k t(1-l)}{t} \\ &= \sum_{k=1}^{K}\sum_{n=1}^{N} c_{k,n} log_2 \left[1 + \frac{p_{k,n}|h_{k,n}|^2}{N_0 B/N}\right](1-l) \end{aligned} \tag{12}$$

Then, we can obtain the average delay $\overline{D_{total}}$ of the cell. $\overline{D_{total}}$ can be divided into two parts, including the transmission delay: $\overline{D_{tran}}$ and propagation delay: $\overline{D_{prop}}$, as shown in formula (13) below:

$$\begin{aligned} \overline{D_{total}} &= \overline{D_{tran}} + \overline{D_{prop}} \\ &= \frac{1}{K}\sum_{k=1}^{K} \frac{P_k t}{R_k(1-l)} + \frac{1}{K}\sum_{k=1}^{K} \frac{d_k}{P_k} \\ &= \frac{1}{K}\sum_{k=1}^{K}\left(\frac{P_k t}{R_k(1-l)} + \frac{d_k}{P_k}\right). \end{aligned} \tag{13}$$

V. Experimental Results

We trained and tested our model based on the historical mobile-traffic-flow data set cached at our local server (edge cloud), in turn based on the attention-LSTM algorithm. Then, the trained prediction model was deployed at an edge-cloud to predict the traffic flow dynamically. Our experimental set up was based on a closed room containing: a cloud server, 3 local servers, and 15 smart devices. A total of 5000 pieces of data were collected from the historical mobile traffic-flow data set, involving time, packet quantity, packet size, device log, etc. The 5000 pieces of data were divided into a training set (3350 pieces) and a testing set for the prediction (1650 pieces).



(a) The tradeoff of memory time slot

(b) Traffic prediction using attention-based LSTM

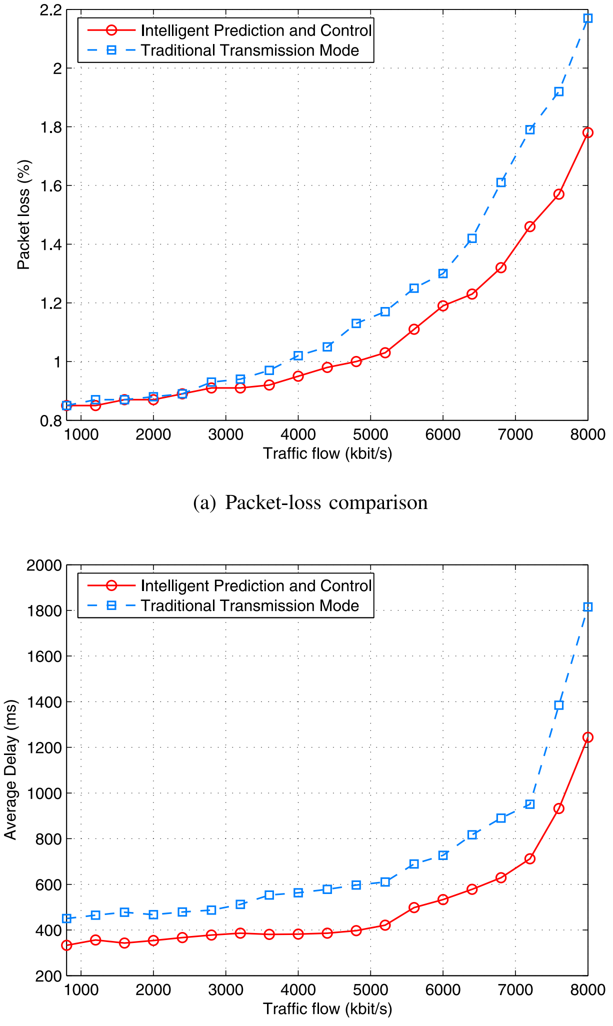
(c) Traffic prediction comparison

Fig. 5. Comparison of the predicted traffic flow and the real traffic flow.

Fig. 5(a) shows the root-mean-square error (RMSE) and loss results of the training and prediction data sets, which have been referred to as training RMSE and prediction RMSE in this figure. For both cases, the RMSE and data losses after initially falling considerably with increasing memory time slot, continue to drop at a slower pace up to 300 s. While the training RMSE continues to drop, the prediction RMSE begins to increase. This indicates that after a memory time slot exceeds 300 s, the prediction results become out-fitted (i.e., less accurate). Thus, we use 300 s as the best input memory time slot in order to build our attention-based LSTM model to predict traffic flow in our subsequent experiments. The comparison results consisting of the real, training, and prediction traffic flows are shown in Fig. 5(b). Although there is insufficient historical training data, it is clear that the trend of the peaks and valleys of the predicted mobile-traffic flow is almost identical to the actual one. Fig. 5(c) compares different traffic-flow prediction algorithms, including backpropagation (BP) [18], autoregressive integrated moving average (ARIMA) [19] and attention-based LSTM (with a memory time slot of 300 s for our approach). The experimental results show that our algorithm is capable of producing a far better traffic-prediction accuracy. This can consequently contribute to the effectiveness of the traffic-flow control, which will be examined next.

Subsequently, we deployed our mobile-traffic-flow prediction algorithm on a cloud server, and scheduled the wireless access resources based on the predicted traffic peak [20]. In these experiments, we evaluate the packet-loss rates and average transmission latency of service requests from specific equipment received by the cloud. The results are shown in Figs. 6(a) and Fig. 6(b), respectively. Due to the small packets sent by each device and the medium size processing capacity of our server, we can only compare the results based on a single user's requests under the following two modes of operation: 1) Intelligent Prediction and Control and 2) Traditional Transmission (without prediction and control). It is clear that the packet-loss rate in the traditional mode rises significantly as the mobile-traffic flow increases, which will reduce accuracy of computation and negatively influence QoE. We should point out that in our experiments, we obtain our packet-loss rates and average latency results using the least-squares method with two iterations. As indicated in Fig. 6, the proposed architecture can effectively improve both the packet-loss rate and the latency as compared with the traditional mode.

Under the proposed intelligent mobile-traffic-flow control strategy, the cloud will initiate multi-site cooperation as soon as it receives notification of a peak mobile-traffic flow, which is sent by certain edge clouds. A specific user device will be notified to access another idle edge cloud to compute the tasks-transmit service. This greatly balances network load and guarantees stable transmission of the users' requests as well as effectively controlling the latency, as shown in Fig. 6(b). However, when there is a low mobile-traffic flow, the average delay of both schemes are almost identical. This is mainly because users tend to select the best edge cloud (e.g., nearest) for communication when the channel is idle. As the mobile-traffic flow increases, the network transmission becomes saturated. Under these conditions, the latency of both schemes increases exponentially, but the proposed approach would be able to control transmission latency more effectively.



(a) Packet-loss comparison

(b) Average delay comparison

Fig. 6. Performance evolution.

VI. Open Issues

Although we have discussed the prediction algorithm and the control architecture of mobile-traffic flow targeting the uRLLC service, there are still many open issues that will need to be explored in the future, e.g.,

(1) Interactive map and prediction of the users' mobility: Radio maps are an important component of 5G. They can provide useful information in the service area about radio signal strength, channel conflict, channel interference, etc. Nodes representing edge clouds are fixed, but the mobile-network environment changes easily. This requires a continuous update of the radio maps so that users can dynamically choose the best edge cloud. Obviously, users' mobility is an important factor for predicting mobile-traffic flow, which helps the remote cloud to ascertain the peak of the group in a certain place in advance [21]. In this case, the mobile-traffic flow received by an edge cloud in densely populated areas can possibly be transferred to another edge cloud for processing. This helps to balance the network load.

(2) Edge-cloud sharing strategy: Due to the mobility of users, the popularity of content, and the high reliability of uRLLC requests, latency can be reduced to an acceptable level. Therefore, the mobile-traffic-flow prediction and control strategy realized in a single-site cell can be used repetitively and shared with other edge clouds. Learning algorithms must be deployed in each edge-cloud in order to provide a more intelligent interactions among different base-stations and to learn an appropriate resource-dispatch mode. The same edge cloud will predict the contents requested in advance or predict the mobile-traffic flow of other edge clouds.

(3) Risk perception of the Remote cloud: Users' requests or data transmissions have priority. The events with small probability, such as alarms and notices, might involve safety concerns. Therefore, the system must anticipate risks in the process of predicting and controlling mobile-traffic flow by predicting the priority of the users' requests [22]. Limited communication and computing resources will be allocated to each task. We can use an online learning algorithm (such as Q-Learning) in machine learning for remote clouds on the basis of predicting mobile-traffic flow. By training and learning historical risk data sets, shared wireless resources in networks can be dispatched in the future.

VII. Conclusion

This paper summarizes the three application scenarios of 5G, namely eMBB, mMTC and uRLLC. We have proposed an IoT-Cloud-based architecture in support of the uRLLC application scenario. In particular, we have described the issue of mobile-traffic flows and their effect on the interaction among user devices, edge clouds and remote clouds. To achieve high reliability and low latency of communications, we first use an attention-based LSTM algorithm to predict the mobile-traffic flow in a single-site mode. Then, the remote cloud collects the traffic-flow predictions of multiple sites. With the support of the cognitive engine and mobile-traffic control modules, the mobile-traffic flow for the entire network is predicted and controlled intelligently. Based on an IoT-Cloud, the performance of the proposed traffic-adaptive resource allocation algorithm is then evaluated. The experimental results verify that our mobile-traffic-flow prediction algorithm can indeed accurately predict mobile-traffic flow and is therefore capable of effectively reducing both the latency and the packetloss rate. Finally, we discuss some open issues, including: interactive maps and prediction of the users' mobility, edge-cloud sharing strategies, and the risk perception of remote clouds. This offers deeper research orientations in mobile-traffic prediction.

In the future, we will further verify our intelligent mobile-traffic-flow control architecture, based on an IoT-Cloud. Furthermore, the prediction of users' mobility, which can play a major role in mobile-traffic-flow prediction, will also be investigated. The algorithm for dispatching communication resources and computing resources, together with the corresponding allocation scheme, will be further detailed.

References

[1] J. Sachs. (2017). *5G Ultra-Reliable and Low Latency Communication*. [Online]. Available: http://cscn2017.ieee-cscn.org/files/2017/08/Janne_Peisa_Ericsson_CSCN2017.pdf

[2] *China's Telecommunications Industry in the First Half of 2018*, Ministry Inf. Ind., Beijing, China, 2018. [Online]. Available: http://www.199it.com/archives/751171.html

[3] M. Chen, Y. Qian, Y. Hao, Y. Li, and J. Song, "Data-driven computing and caching in 5G networks: Architecture and delay analysis," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 70–75, Feb. 2018.

[4] H. Wang, Y. Wu, G. Min, J. Xu, and P. Tang, "Data-driven dynamic resource scheduling for network slicing: A deep reinforcement learning approach," *Inf. Sci.*, vol. 498, pp. 106–116, Sep. 2019.

[5] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 653–662, Apr. 2015.

[6] G. Zhang, Y. Cao, L. Wang, and D. Li, "Operation cost minimization for base stations with heterogenous energy supplies and sleep-awake mode: A two-timescale approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 908–918, Dec. 2018.

[7] P. Sun, Y. Hu, J. Lan, and M. Chen, "TIDE: Time-relevant deep reinforcement learning for routing optimization," *Future Gener. Comput. Syst.*, vol. 99, pp. 401–409, Oct. 2019.

[8] Y. Zhou, Z. M. Fadlullah, B. Mao, and N. Kato, "A deep-learning-based radio resource assignment technique for 5G ultra dense networks," *IEEE Netw.*, vol. 32, no. 6, pp. 28–34, Nov./Dec. 2018.

[9] J. Feng, X. Chen, R. Gao, M. Zeng, and Y. Li, "DeepTP: An end-to-end neural network for mobile cellular traffic prediction," *IEEE Netw.*, vol. 32, no. 6, pp. 108–115, Nov./Dec. 2018.

[10] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[11] R. Lent, "Resource selection in cognitive networks with spiking neural networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 860–868, Dec. 2018.

[12] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. IEEE INFOCOM* , Honolulu, HI, USA, Apr. 2018, pp. 1970–1978.

[13] X. Cheng, Y. Wu, G. Min, and A. Y. Zomaya, "Network function virtualization in dynamic networks: A stochastic perspective," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2218–2232, Oct. 2018.

[14] W. Miao *et al.*, "Stochastic performance analysis of network function virtualization in future Internet," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 613–626, Mar. 2019.

[15] M. Chen, Y. Hao, H. Gharavi, and V. C. M. Leung, "Cognitive information measurements: A new perspective," *Inf. Sci.*, vol. 505, pp. 487–497, Dec. 2019.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2015.

[17] T. Zou, Q. Meng, and H. Cong, *Mobile Communication Technology and Application*. Beijing, China: Tsinghua Univ. Press, 2013.

[18] Z. Wang, B. Wang, C. Liu, and W.-S. Wang, "Improved BP neural network algorithm to wind power forecast," *J. Eng.*, vol. 2017, no. 13, pp. 940–943, 2017

[19] A. Biernacki, "Improving quality of adaptive video by traffic prediction with (F)ARIMA models," *J. Commun. Netw.*, vol. 19, no. 5, pp. 521–530, 2017.

[20] M. Chen, Y. Miao, X. Jian, X. Wang, and I. Humar, "Cognitive-LPWAN: Towards intelligent wireless services in hybrid low power wide area networks," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 2, pp. 407–417, Jun. 2019.

[21] M. Chen, Y. Hao, C.-F. Lai, D. Wu, Y. Li, and K. Hwang, "Opportunistic task scheduling over co-located clouds in mobile environment," *IEEE Trans. Services Comput.*, vol. 11, no. 3, pp. 549–561, May/Jun. 2018.

[22] Y.-J. Ou, X.-L. Wang, J.-F. Jiang, H.-Y. Wei, C.-L. Huang, and K.-S. Hsu, "Simulator training to drive the risk perception of the reliability and validity," in *Proc. IEEE Int. Conf. Appl. Syst. Invention (ICASI)*, Chiba, Japan, Apr. 2018, pp. 374–377.

Min Chen (SM'09) is a Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology. He has more than 300 paper publications, including over 200 SCI papers, over 80 IEEE Transactions Journal papers, 32 ESI highly cited papers, and 10 ESI hot papers. He has over 19 500 Google Scholars Citations reached with an H-index of 68 and i10-index of 218. His top paper was cited over 1800 times. His research focuses on cognitive computing, 5G Networks, and mobile edge computing. He got IEEE Communications Society Fred W. Ellersick Prize in 2017, and the IEEE Jack Neubauer Memorial Award in 2019. He was selected as a Highly Cited Researcher at 2018 (computer science category). He is a Series Editor of the IEEE Journal on Selected Areas in Communications on Network Softwarization and Enablers. He is a Technical Editor of IEEE Network, and an Editor of the IEEE Transactions on Cognitive Communications and Networking.

Yiming Miao received the B.Sc. degree from the College of Computer Science and Technology, Qinghai University, Xining, China, in 2016. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Huazhong University of Science and Technology. Her research interests include IoT sensing, healthcare big data, and emotion-aware computing.

Hamid Gharavi received the Ph.D. degree from Loughborough University, U.K., in 1980. He joined the Visual Communication Research Department, AT&T Bell Laboratories, Holmdel, NJ, USA, in 1982. He was then transferred to Bell Communications Research (Bellcore) after the AT&T-Bell divestiture, where he became a consultant on video technology and a Distinguished Member of Research Staff. In 1993, he joined Loughborough University as a Professor and the Chair of communication engineering. Since September 1998, he has been with the National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, USA. He was a Core Member of Study Group XV (Specialist Group on Coding for Visual Telephony) of the International Communications Standardization Body CCITT (ITU-T). His research interests include smart grid, wireless multimedia, mobile communications and wireless systems, mobile ad hoc networks, and visual communications. He holds eight U.S. patents and has over 150 publications related to the above areas. He received the Charles Babbage Premium Award from the Institute of Electronics and Radio Engineering in 1986, the IEEE CAS Society Darlington Best Paper Award in 1989, and the Washington Academy of Science Distinguished Career in Science Award for 2017. He served as a Distinguished Lecturer of the IEEE Communication Society. He has been a Guest Editor of a number of special issues of the *Proceedings of the IEEE*, including *Smart Grid, Sensor Networks and Applications*, *Wireless Multimedia Communications*, *Advanced Automobile Technologies*, and *Grid Resilience*. He was the TPC Co-Chair of IEEE SmartGridComm in 2010 and 2012. He served as the Editorial Board Member of the *Proceedings of the IEEE* from January 2003 to December 2008. From January 2010 to December 2013, he served as the Editor-in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is currently serving as the Editor-in-Chief for IEEE WIRELESS COMMUNICATIONS.

Long Hu received the B.S. and master's degrees from the Huazhong University of Science and Technology (HUST), and the Ph.D. degree from the School of Computer Science and Technology, HUST. His current research includes 5G mobile communication system, big data mining, marine-ship communication, Internet of Things, and multimedia transmission over wireless network. He is the Publication Chair for Fourth International Conference on Cloud Computing in 2013.

Iztok Humar received the B.Sc., M.Sc., and Ph.D. degrees in telecommunications from the Faculty of Electrical Engineering, University of Ljubljana, Slovenia, in 2000, 2003, and 2007, respectively, and the Ph.D. degree in information management from the Faculty of Economics, University of Ljubljana in 2009. He is currently an Associate Professor with the Faculty of Electrical Engineering, University of Ljubljana, where he lecturers on design, modeling and management of communication networks on graduate and postgraduate study. He was a Supervisor of many undergraduate students and some postgraduate students. His main research topics include the energy efficiency of wireless networks, cognitive computing for communications with applications and modeling of network loads and traffic for QoS/QoE. In 2010, he was a three-month Visiting Professor and a Researcher with the Huazhong University of Science and Technology, Wuhan, China. He served as the IEEE Communication Society of Slovenia Chapter Chair for ten years. He is a Senior Member of the IEEE Communication Society in 2009 and a member of the Electrical Association of Slovenia.