Dynamic Operations of Cloud Radio Access Networks (C-RAN) for Mobile Cloud Computing Systems

Yegui Cai[*], F. Richard Yu[*], and Shengrong Bu[*†]

[*]Depart. of Systems and Computer Eng., Carleton University, Ottawa, ON, Canada

[†]Huawei Technologies Canada Co., LTD., Ottawa, ON, Canada

Email: ycai@sce.carleton.ca; richard.yu@carleton.ca; shengrbu@sce.carleton.ca

Abstract

In this paper, we jointly consider cloud radio access networks (C-RAN) and mobile cloud computing (MCC) in a holistic framework. In particular, we study how to dynamically operate C-RAN to enhance the end-to-end performance of MCC services in next generation wireless networks. An intrinsic challenge in such system is that channel state information (CSI) is outdated. With delayed CSI, only sub-optimal C-RAN operations can be made if deterministic optimization techniques are applied directly. We formulate the topology configuration and rate allocation problem with delayed CSI under a stochastic optimization framework. Such framework maximizes MCC services' sum throughput with constraints on the response latency experienced by each MCC user. We propose an optimal policy for the stochastic optimization problem, which has the merit of low computation cost. Offline and online algorithms are developed based on the optimal policy. Using simulation results, we show that, with the emergence of MCC and C-RAN technologies, the design and operation of future mobile wireless networks can be significantly affected by cloud computing, and the proposed scheme is capable of achieving substantial performance gains over existing schemes.

Index Terms

Cloud computing; cloud radio access networks; mobile cloud computing systems.

I. Introduction

Recently, as a new information technology (IT) paradigm, cloud computing has become one of the hottest topics in both academia and industry. Cloud computing is a model for enabling on-demand access to a shared pool of configurable resources (e.g., servers, storage, applications, services, etc.). The essential characteristics of cloud computing include on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service [1]. Several service models are supported, including cloud software as a service, cloud platform as a service, and cloud infrastructure as service [2]. Cloud computing has attracted significant attention, and several commercial clouds, including Amazon EC2, Microsoft Azure, and Google App Engine, have been providing services to users.

Cloud computing will have profound impacts on the design and operation of wireless networks. On one hand, with recent advances of wireless mobile communication technologies and devices, more and more end users access cloud computing systems via mobile devices, such as smart phones and tablets. The integration of cloud computing into the mobile environment enables *mobile cloud computing* (MCC), which is widely considered as a promising mobile computing paradigm with huge market [3], [4]. MCC enables offloading the computing power and data storage requirements from mobile devices into the powerful computing platforms in the cloud, bridging the gap between the increasing computing demands and the traditional mobile technologies with limited computing, storage and energy resource in mobile devices [3].

On the other hand, the powerful computing platforms in the cloud can be beneficial to radio access networks (RAN) as well (in addition to mobile end users), which leads to a novel concept of *cloud radio access networks* (C-RAN) [5]–[8]. Unlike the existing cellular networks, where computing resources for baseband processing are located at each cell site, in C-RAN, the computing resources are located in a central wireless network cloud with powerful computing platforms. This transition from distributed to centralized infrastructure for baseband processing can have significant benefits: saving the operating expenses due to centralized maintenance; improving network performance due to advanced coordinated signal processing techniques; reducing energy expenditure by exploiting the load variations [5]–[9].

Although some excellent works have been done to study cloud computing for both end users and access networks, these two important areas have traditionally been addressed separately in the literature. However, as shown in the following, it is necessary to consider these two advanced technologies together to provide better services in next generation wireless networks. Therefore, we jointly study C-RAN in MCC systems so as to improve end-to-end network performance. The motivations behind our work are based on the following observations.

- From end-to-end applications' perspective, both C-RAN and over-the-top (OTT) service provider cloud are parts of the whole system. The experience in end-to-end applications (e.g., transmission control protocol (TCP)-based applications) indicates that the optimized performance in one segment of the whole system does not guarantee the end-to-end improvement if the other segment is ignored in the optimization [10], [11].
- It is well known that, while TCP performs relatively well over wired links, its performance degrades over wireless links due to the scarce bandwidth, high bit error rate (BER) and user mobility [12]. In addition, networking has become a bottleneck that has a significant impact on the quality of cloud services [13]. Therefore, the characteristics of C-RAN should be carefully considered in MCC systems.
- Recent studies in cross-layer designs show that optimizing lower layer's performance (e.g., physical layer throughput) does not necessarily benefit quality of service (QoS) at upper layers [11]. From a user's point of view, QoS at upper layers (e.g., TCP throughput) is more important than that at other layers.

To the best of our knowledge, the study of C-RAN in MCC systems for next generation wireless networks has not been addressed in previous works. The distinct features of this work are as follows.

- We consider how to dynamically configure C-RAN to enhance MCC services' performance in a holistic framework. For C-RAN we study the topology configuration and rate allocation problem; for MCC, as a case study, we consider mobile search engine services on top of TCP connections, which has critical requirements on per user response latency. In this work, we optimize the end-to-end TCP throughput performance of MCC users in next generation cellular networks.
- Despite the potential benefits brought by C-RAN, one of the major challenges in C-RAN is that the

channel state information (CSI) is inaccurate due to the delay in obtaining and transmitting such information [7]. For instance, in LTE-Advanced, the standard interface for inter-BS communications, $X2$, is designed to allow a latency of $20ms$ for control plan messages, and the typical value for the latency is expected to be $10ms$ on average [14]. Imperfect CSI has significant impacts on not only C-RAN, but also wireless networks in general [15]–[17]. Since it is difficult to solve this problem using traditional information theoretic approach [18], we take a stochastic optimization approach, which has well-developed mechanisms to address the impacts of noisy and delayed CSI. An optimal policy is found based on the particular structure of the topology configuration and rate allocation problem.

- *Response latency* experienced by cloud users has been recognized as one of the most important performance metrics in cloud computing [19], [20]. Therefore, to improve MCC users' QoS, we model the response latency experienced by each MCC user as a constraint in our formulation.
- We investigate the trade-off between the systematic efficiency and the fairness among MCC users. A parameter is introduced to study such trade-off taking into account the delayed CSI in C-RAN. With this parameter, we re-formulate the problem to maximize the Jain's fairness index in MCC systems.
- Extensive simulations show that, with the emergence of MCC and C-RAN technologies, the design and operation of next generation wireless networks can be significantly affected by cloud computing, and the proposed scheme is capable of achieving substantial performance gains over existing schemes.

The rest of the paper is structured as follows. Section II describes the system. Section III discusses the issues caused by delayed CSI in C-RAN. We formulate the problem as a decision theoretic problem in Section IV. Fairness and efficiency trade-off is studied in Section V. Simulation results are discussed in Section VI. Finally, we conclude this study in Section VII with future work.

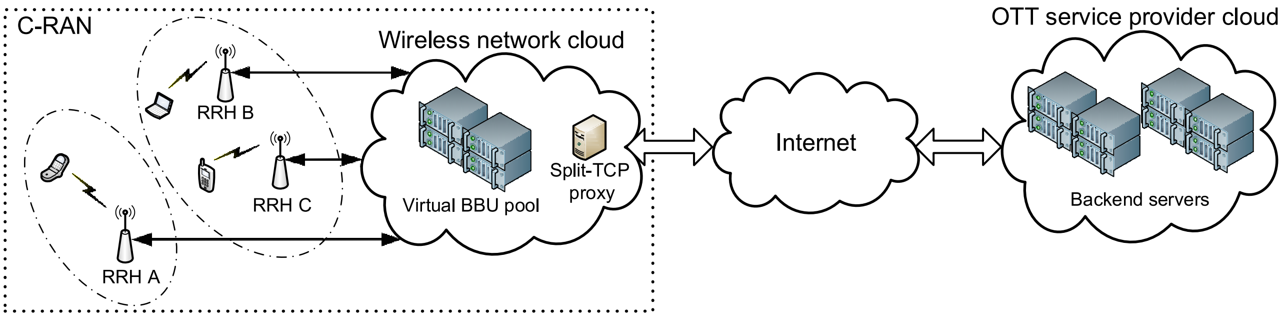

Fig. 1. A cloud radio access network in the MCC environment.

II. System Description

In this section, we first describe MCC systems and dynamical configurations of C-RAN. Then, the physical layer and link layer models for C-RAN are presented.

A. Mobile Cloud Computing with Cloud Radio Access Networks

The system we consider in this paper is shown in Fig. 1, which mainly consists of two sub-systems, i.e., C-RAN and cloud computing. The problem addressed in this work crosses the two sub-systems. The wireless communication mainly happens at the C-RAN, while the processing (e.g., data mining) for the cloud services happens at the backend servers inside OTT service provider cloud.

In C-RAN, the traditional base stations are evolved into a system consisting of remote radio heads (RRHs) distributed in different geographic locations and a baseband process unit (BBU) pool in the wireless cloud [6], [21]. The RRHs are connected to the wireless network cloud via backhaul networks. Even optical networks can provide high-capacity and low-latency connections between RRHs and BBU pools, optical networks are not always a feasible solution for backhaul networks. In contrast, microwave links and non-line-of-sight wireless links can provide more flexible deployments but introduce latency in the system [22]. Note that the implementation of the wireless network cloud varies. For example, in [23], the central processing and control unit is called *virtual base station pool*. Nevertheless, we do not have any assumption on the implementation of the wireless network cloud. In the C-RAN considered in this paper, there are B RRHs with one antenna each, which are denoted as a set $\mathcal{B}$.

Many mobile cloud services require the end-to-end reliable data transfer across the two systems. There is a *split-TCP proxy* at the edge of the wireless network cloud. The split-TCP proxy is the split point for TCP flows. Such split-TCP proxy has been widely used in cloud computing [20] and traditional cellular networks [24]. In the context of cloud computing, split-TCP is also a popular approach to provide reliable data transfer for cloud services. A client sets up a TCP connection to the nearby split-TCP proxy, then the split-TCP proxy sustains a persistent TCP connection to the data center with a very large TCP connection window [20]. In wireless networks, split-TCP proxy hides the wireless related issues from the wireline host via inserting a split point between the wireless and wired hosts. It locally acknowledges each segment and then stores and forwards the segments on the second TCP connection [24]. The split-TCP proxy can be implemented at system architecture evolution gateway (SAE-GW) in LTE systems, since the user data flows are tunneled to SAE-GW before being sent to the Internet [25].

Fig. 2 shows the logical relationship of the mobile users, wireless network cloud, and OTT service provider cloud. TCP flows carrying mobile cloud services run from the mobile device to the backend server in OTT service provider cloud. Split-TCP proxy residing at the edge of wireless network cloud splits the end-to-end connection between the mobile user and the backend server into two connections and sustains a persistent connection between itself and the backend server. Meanwhile, the wireless network cloud conducts dynamic operations on wireless networks to provide best service for the upper layer. Such dynamic operations include topology configuration and rate allocation. Topology configuration controls how the RRHs cooperate with each other. For instance, in Fig. 1, RRHs B and C form a cooperating set to serve the two MCC users together while RRH A itself is another cooperating set. After topology configuration, the wireless network cloud needs to decide the data rates that the MCC users can transmit. Inside a cooperating set, the signals are processed jointly such that there is no interference.

We denote the channel state matrix at time slot t as S^t, the topology configuration action at time slot t as Ω^t. The rate allocation vector has B elements, $\mathbf{R}^t = [R^{1,t} \dots R^{B,t}]$, where each element is the rate allocation for a user. Here, we assume that there is only one user serviced by each RRH at a time slot, which is commonly assumed in the literature due to the opportunistic scheduling operation in practice [26]. The overall action is $a^t \triangleq \{\Omega^t, \mathbf{R}^t\} \in \mathcal{A}$, where $\mathcal{A}$ is the set of actions available. In this work, we

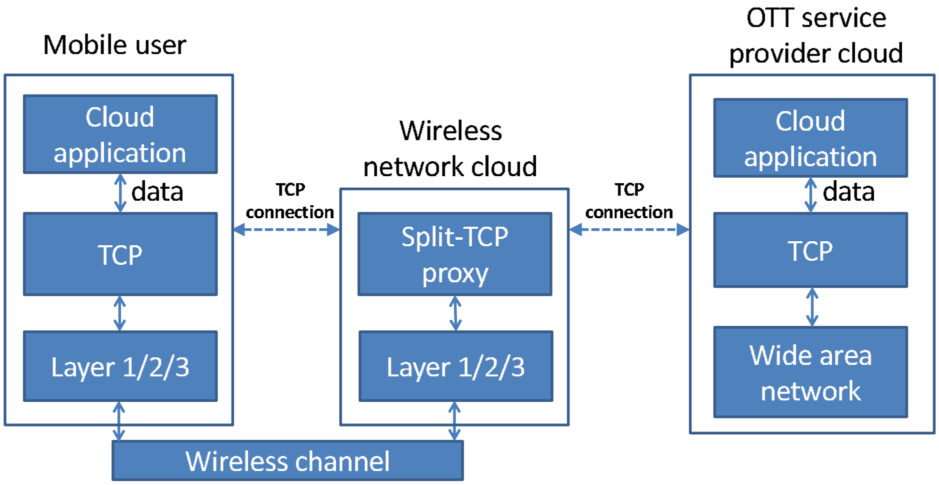

Fig. 2. Logical protocol stacks of MCC system entities: mobile device, split-TCP proxy, and cloud backend server.

assume that the power allocated to each user is the same, so that we can focus on the optimal topology configuration and rate allocation problem. On one hand, such power binary power allocation has been shown to be optimal in some scenarios [27]. Thus it is reasonable to decouple power control at the time being. On the other hand, it is straightforward to adapt this framework to support power allocation by taking the possible power allocation scheme into the action space.

B. Physical Layer and Link Layer in C-RAN

In the following we introduce the physical layer and link layer models. We attempt to make the modeling of these two layers to be as general as possible while sustaining a certain feasibility in performance analysis. This is because essentially C-RAN is supposed to be an open platform to support various technologies at lower layers [6].

Consider a cooperating set ω whose cardinality is $|\omega| = K$. Signals for mobile users served by RRHs in ω can be decoded without interfering with each other; while the mobile users served by the *non-cooperating* RRHs, $\mathcal{B} - \omega$, are interferers to ω. We number the RRHs in ω from 1 to K, and RRHs in $\mathcal{B} - \omega$ from $K + 1$ to B. Denote the complex channel gain from a mobile device served by RRH i to the antennas of all the RRHs in ω as $\mathbf{h}_i \in \mathbb{C}^{K \times 1}$, $i = 1, \ldots, K, K + 1, \ldots, B$. We assume that all mobile devices are allocated the same transmission power P, namely, the power allocation matrix of mobile devices in cooperating set ω is $\sqrt{P} \times \mathbf{I}_K$. If the complex data symbols of mobile devices served by cooperating set ω are $[x_1 \ \ldots \ x_K]$, and the data symbols of mobile devices served by the other RRHs are $[x_{K+1} \ \ldots \ x_B]$, the received signal of the antennas in a cooperating set ω is given by

$$\mathbf{y} = \sqrt{P}\sum_{l=1}^{K}\mathbf{h}_l x_l + \sqrt{P}\sum_{l'=K+1}^{B}\mathbf{h}_{l'} x_{l'} + \mathbf{n}, \tag{1}$$

where $\mathbf{n}$ is a vector of independent complex circularly symmetric additive Gaussian noise with each element $n \sim \mathcal{CN}(0, N_0)$. In the above signal level representation, the first term is the useful signal inside ω, while the second term is interference signal from $\mathcal{B} - \omega$.

As a representative signal processing technique, Minimum Mean Square Error - Successive Interference Cancelation (MMSE-SIC) receiver [28, Ch. 10.1] can achieve the multiple access channel capacity. With fixed decoding order, the capacities of users ranging from $1, 2, \ldots, K$ are given as follows.

$$\begin{array}{rcl} C_K & = & \log\left(1 + \frac{P\|\mathbf{h}_K\|^2}{N_K}\right), \\ C_{K-1} & = & \log\left(1 + P\mathbf{h}_{K-1}^T(N_{K-1}I_K + P\mathbf{h}_K\mathbf{h}_K^*)^{-1}\mathbf{h}_{K-1}\right), \\ & \cdots & \\ C_1 & = & \log\left(1 + P\mathbf{h}_1^T(N_1 I_K + \sum_{l=2}^{K} P\mathbf{h}_l\mathbf{h}_l^*)^{-1}\mathbf{h}_1\right), \end{array} \tag{2}$$

where $N_l, l = 1, 2, \ldots, K$, are the AWGN noise accounting for the receiver noise N_0 and the interference from outside ω. Specifically, the total noise at the l^{th} antenna is $N_l = N_0 + P\sum_{l'=K+1}^{B}|\mathbf{h}_{l'}|^2$. Note that any other physical layer is likewise applicable to our work.

For a particular user u, if the current channel capacity is less than the transmission rate allocated, there is an outage so that the resulting transmission rate is 0; otherwise, the resulting transmission rate is equal to the allocated rate. In slot t, the performance of a user u is partially controlled by the action a^t taken by the wireless network cloud. The probability of error without any link-layer retransmission is defined as

$$p_{1,u} = \Pr\left(r_u^t(a^t) > C_u^t(a^t)\right), \tag{3}$$

where r_u^t is the rate allocation for user u, and C_u^t is the channel capacity at time slot t, which is a random variable since the channel state is unknown. For the link layer, we use hybrid automatic repeat request (HARQ) to reduce unreliability in wireless links. HARQ combines forward error correction and ARQ to increase the communication reliability. We assume a chase combining scheme in the following.[1] The performance of such an HARQ scheme has been analyzed in [29]. It is shown that, for user u, the number of packets transmitted, denoted as a random variable N_u, follows a Gaussian distribution with mean μ_u and variance σ_u^2

$$\mu_u = \frac{1 + p_{1,u} - p_{1,u}p_{2,u}}{1 - p_{1,u}p_{2,u}}, \tag{4}$$

$$\sigma_u^2 = \frac{p_{1,u}(1 - p_{1,u} + p_{1,u}p_{2,u})}{1 - p_{1,u}p_{2,u}}, \tag{5}$$

where $p_{1,u}$ is the probability of error after decoding the information block by forward error correction, $p_{2,u}$ is the probability of error after soft combining two successive transmissions of the same information block [29]. Note that $p_{1,u}$ is essentially the outage probability without HARQ defined in (3) and that $p_{2,u}$ is usually obtained via link level simulations [29].

Therefore, if the maximum number of transmissions allowed in the link layer is ν, and if the action taken at time slot t is a^t, the packet error rate is

$$p_{e,u}(a^t) = \Pr(N_u > \nu) = Q\left(\frac{\nu - \mu_u}{\sigma_u}\right), \tag{6}$$

where $Q(\cdot)$ is the well-known Q-function. Moreover, the average transmission time of a TCP data packet over wireless links can be computed as

$$\overline{T}_{wireless,u}(a^t) = \mu_u \frac{L_{data} + L_{ack}}{r_u}, \tag{7}$$

where L_{data} and L_{ack} are the link layer frame size for a TCP data packet and a TCP acknowledgment packet, respectively. We assume the downlink data rate of the C-RAN is the same as the uplink for ease of analysis.

III. C-RAN WITH DELAYED CSI

In this section, we first introduce the channel modeling based on Finite State Markov Chains (FSMCs), which is essential in taking delayed CSI into account in our formulation. Then we discuss the delayed

[1] Other HARQ schemes, such as incremental redundancy combining, are applicable in our work as well.

CSI issue in C-RAN followed by the belief-state concept that captures the uncertainty caused by delayed CSI.

A. *Finite State Markov Chain Channel Model and Delayed CSI*

We define the vector space consisting of B^2 elements as the system state $\mathcal{S}$. Assume at time slot t, the system state S^t is $s \in \mathcal{S}$, it will jump to s' at the next time slot. With the FSMC channel modeling [30], the state-transition function A is given by

$$A(s, s') = \text{Pr}(S^{t+1} = s' | S^t = s) = \prod_{b=1,u=1}^{b=B,u=B} \text{Pr}(I_{b,u}^{t+1} | I_{b,u}^{t}), \tag{8}$$

where $I_{b,u}^{t}$ and $I_{b,u}^{t+1}$ are the current state and next state of the FSMC from a transmit antenna of mobile user u to a receive antenna of RRH b.

To see how delay comes into C-RAN, we consider a C-RAN shown in Fig. 1. The CSI is obtained via the pilot signals received at RRHs. After channel estimation, the CSI will be transmitted over backhaul networks to the wireless network cloud. At the wireless network cloud, a decision about how the RRHs cooperate and the rates at which MCC users can transmit are decided after obtaining CSI. Then, the user data is transmitted. Similar to the measurement and propagation of CSI, user signals are transmitted from MCC users to RRHs, and then are propagated over the backhaul networks. At the moment of decision making, the available CSI is delayed. We can abstract the total delay between the actual channel state at the moment of decision making and the one of observation as one single number. Then, we can map the delay in seconds into the transition steps in Markov chains. Therefore the d steps transition probability is given by multiplying matrix A by d times, A^d.

B. *Belief State with Known Delay Steps*

Given the delay in steps, we can derive the *belief state*, which is the sufficient statistic of the previous action and observation history [31]. A belief state $\mathbf{b}^t$ at time slot t is a probability distribution of the state space. Accordingly, the probability that the state at time slot t being s^t is given by the corresponding element in $\mathbf{b}^t$, denoted as $b(s^t)$. Following [31], we use *belief state* to express both the vector $\mathbf{b}^t$ and its element given a state $b(s^t)$.

With techniques such as time-stamping, we can know the number of delay steps d. With such an assumption, the observation is just the actual state delayed by d steps. Denote the observation at time t as a random variable O^t. We have $O^t = S^{t-d}, t = d+1, \ldots$. Thus we can derive the explicit relation between the current state and observation. The belief state is

$$\begin{aligned} b(s^{t+1}) =& \text{Pr}(s^{t+1}|o^{t+1}, o^t, \ldots) \\ =& \text{Pr}(s^{t+1}|s^{t+1-d}, s^{t-d}, \ldots) \\ =& \text{Pr}(s^{t+1}|s^{t+1-d}) \\ =& A^d(s^{t+1-d}, s^{t+1}). \end{aligned} \tag{9}$$

The third line is given by the first-order Markovian property assumed in the FSMC channel model, and A^d is the d-step probability transition matrix.

C. Belief State with Unknown Delay Steps

If it is difficult to get the number of delay steps, we can still compute the belief state based on Bayesian rule. We first introduce *Observation Function* $B(\cdot)$ [31]. Assume at time slot t, the observation of the system is $o \in \mathcal{O}$. [2] The observation function $B(\cdot)$ essentially depicts the probabilistic relationship between an observation $o \in \mathcal{O}$ and a state $s \in \mathcal{S}$. Formally, observation is also a function of the action taken; however, in our problem here observation is independent of the action, so it is defined as

$$B(s, o) = \text{Pr}(o|s). \tag{10}$$

State-transition function $A(\cdot)$ and observation function $B(\cdot)$ can be obtained via classical algorithms such as *Expectation Maximization* [32].

Essentially, provided a new observation at time $t+1$, o^{t+1}, the new belief should reflect the likelihood of ending up in new state s^{t+1}, the likelihood of observing o^{t+1}, and the pervious belief distribution $\mathbf{b}^t$. The rule to update the belief state according the previous belief and current observation based on the

[2] Obviously, $\mathcal{O}$ is the same as $\mathcal{S}$. We use different notations for clear presentation.

Bayesian rule [31] is

$$\begin{aligned} b(s^{t+1}) &= \Pr\left(s^{t+1}|o^{t+1},\mathbf{b}^t\right) \\ &= \frac{B(s^{t+1},o^{t+1})\sum_{s^t\in\mathcal{S}} A(s^t,s^{t+1})b(s^t)}{\sum_{s^{t+1}\in\mathcal{S}} B(s^{t+1},o^{t+1})\sum_{s^t\in\mathcal{S}} A(s^t,s^{t+1})b(s^t)}. \end{aligned} \tag{11}$$

With the belief state $\mathbf{b}^t$, we can compute the probability of error without link-layer retransmission (3) as follows.

$$p_{1,u} = \sum_{c_u(s',a^t)<r_u^t} b(s'), \tag{12}$$

where $c_u(s',a^t)$ is the capacity that user u can achieve given the action a^t and the channel state s', and r_u^t is the rate allocation.

IV. TCP Throughput over C-RAN in MCC Systems

In this section, we study the TCP throughput over C-RAN in MCC systems. Then we investigate the user response latency issue. Next, the TCP throughput maximization with response latency constraint problem is formulated as a constrained stochastic optimization problem. Finally, we derive the optimal topology configuration and rate allocation algorithm.

A. Round Trip Time and Split-TCP Throughput

Split-TCP is a popular reliable data transfer protocol for data center networks [20] and legacy cellular networks [24]. We can expect that it will play an important role in next generation MCC systems. Therefore, in this work, we adopt split-TCP as our transport layer protocol. A widely used TCP throughput model is developed in [33]. It has been used in cross-layer designs to maximize TCP throughput (for instance, [34]). In this section, we extend the existing work to take delayed CSI into account in the TCP throughput model.

We firstly discuss the round trip time (RTT). Fig. 3 shows the round trip times for mobile cloud services over C-RAN. There are two types of RTTs [20]. RTT_1 represents the RTT between clients and the split-TCP proxy at the edge of wireless network cloud. RTT_2 is the RTT between the split-TCP proxy and the backend server in the OTT service provider cloud. We will not discuss the randomness in RTT_2 because this work is focused on the effect of C-RAN on cloud service.

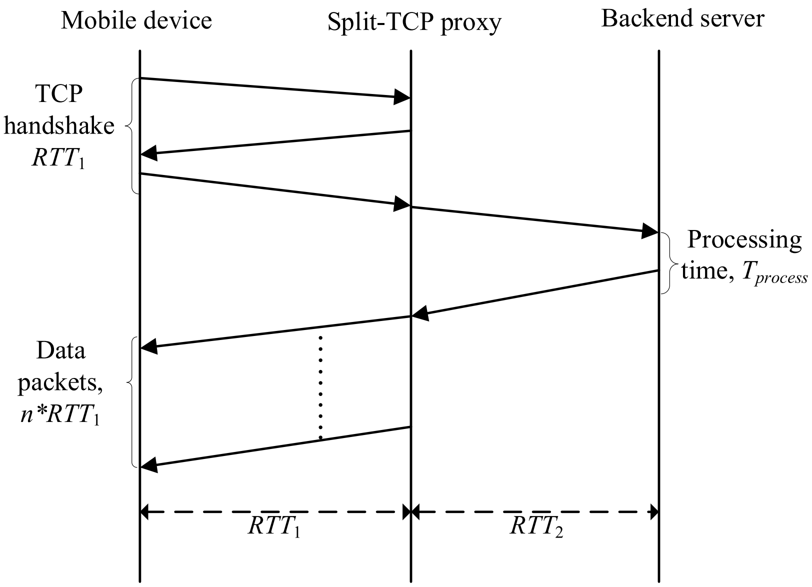

Fig. 3. Round trip times and response latency using split-TCP.

RTT_1 consists of $T_{wireless}$ and $T_{backhaul}$, which represent the round trip transmission time over the wireless and backhaul networks, respectively. Note that, due to the wireless channel fading, $T_{wireless}$ is a random variable partly controlled by the action taken by the control unit in the wireless network cloud of C-RAN. The mean value of RTT_1 is given by

$$\overline{RTT}_1(a^t) = \overline{T}_{wireless}(a^t) + T_{backhaul}, \tag{13}$$

where $\overline{T}_{wireless}(a^t)$ is defined in (7).

We assume that the maximum segment size (MSS) is set in such a way that a single segment will fit into a single link-layer frame. That is, there is only one link layer packet for a TCP segment. The queuing delay is not taken into consideration. Since we focus on the delayed CSI in this work, we can approximately take the queuing delay as a constant contribution to RTT. In terms of implementation, the queuing delay can be approximated by its statistical average. This assumption is reasonable because the time scale of queuing dynamics is generally larger than that of the wireless channel dynamics provided there is a bulk of data to transmit [35]. If the files are small, the queuing delay can be simply ignored. Therefore, we take queuing delay as a constant, which has been considered in $T_{backhaul}$.

Padhye *et al.* have developed a model for TCP connections. For user u, the average throughput can be derived as [33]

$$\overline{\eta} \approx \min \left\{ \frac{W_{max}}{\overline{RTT}}, \frac{1}{\overline{RTT}\sqrt{\frac{2n_{ack}p_e}{3}} + T_0 \min\left\{1, 3\sqrt{\frac{3n_{ack}p_e}{8}}p_e(1+32p_e^2)\right\}} \right\}, \tag{14}$$

where W_{max} is the maximum congestion window, $\overline{RTT}$ is the round trip time, n_{ack} is the number of packets acknowledged by a TCP ACK (generally 2), T_0 is the initial time-out for the TCP sender, p_e is the TCP loss probability. The accuracy of such a model has been verified against real TCP traces in [33]. Note that the throughput of a TCP connection over a radio access network, $\eta_{RAN,u}(a^t)$, is a random variable because the actual channel state S^t is unknown. $\overline{\eta}_{RAN,u}(a^t)$ is the mean value of it. For a connection in C-RAN, $\overline{RTT}$ and p_e are defined in (13) and (6), respectively.

In split-TCP, the end-to-end throughput is the minimum throughput between the two TCP connections. For a user u, denote the average throughput of the TCP connection between the mobile user and split-TCP proxy as $\overline{\eta}_{RAN,u}$, and the one between split-TCP proxy and data centers as $\overline{\eta}_{cloud}$. The overall average throughput of split-TCP for u given the action taken a^t is

$$\overline{\eta}_u(a^t) = \min\left\{\overline{\eta}_{RAN,u}(a^t), \overline{\eta}_{cloud}\right\}. \tag{15}$$

B. Per-User Response Latency

Response latency experienced by cloud users is critical for mobile cloud services [19], [20]. Since the main computation tasks are performed in data centers, MCC systems suffer from the response latency caused by processing time and communications among network entities. The processing latency is mainly caused by the hardware and the operating system, which is not the focus of our work. On the other hand, as will be shown in the following sections, the communication latency can be improved by careful design and operation of the C-RAN.

As shown in Fig. 3, the connection between the client and split-TCP proxy spends about an RTT_1 in the hand-shaking phase. $T_{process}$ is the time needed for the backend severs to process the request. The split-TCP proxy needs to wait an RTT_2 and $T_{process}$ in setting up the connection and for the backend server to compute the results and to transmit them to the split-TCP proxy. Using the same assumption as [20], it takes $n * RTT_1$ to transmit the results from the split-TCP proxy to the MCC users. Denoting the total response latency as $\tau(a^t, S^t)$, we have

$$\tau(a^t, S^t) = \ (n+1) * RTT_1(a^t, S^t) + RTT_2 + T_{process}. \tag{16}$$

A typical value of n for search engine application is 4 [20]. Recall that RTT_1 at time slot t is a random variable depending on the actual system state S^t and the action taken a^t. Accordingly the average value of total response latency $\overline{\tau}$ is given by

$$\overline{\tau}(a^t) = \ (n+1) * \overline{RTT}_1(a^t) + \overline{RTT}_2 + T_{process}, \tag{17}$$

where $\overline{RTT}_1(a^t)$ is defined in (13), $\overline{RTT}_2$ and $T_{process}$ are considered to be constant.

C. Maximizing TCP Throughput with Delayed CSI for Mobile Cloud Services

At time slot t, the system state S^t is an unobserved random variable. The wireless network cloud selects the cooperating RRHs and allocates the rate for MCC users, denoted as a^t. Denote the end-to-end throughput of a mobile user u given by (15) as a random variable $\eta_u(a^t, S^t)$, then $\sum_{u=1}^{B} \eta_u(a^t, S^t)$ is the sum throughput of the system. The number of time slots considered is h, which is called the number of horizons in Markov decision process literature. The cumulative rewards over h horizons is $\sum_{t=1}^{t=h} \sum_{u=1}^{u=B} \eta_u(a^t, S^t)$.

Accordingly, we denote the response latency defined in (16) as τ_u, and we constrain the latency to be under a threshold α. To maximize the sum TCP throughput subject to the response latency constraint, we have the following optimization problem,

$$\begin{aligned} \underset{a^t, t=1,2,\ldots,h}{\text{maximize}} \quad & \mathbb{E}\left[\frac{1}{h}\sum_{t=1}^{t=h}\sum_{u=1}^{u=B} \eta_u(a^t, S^t)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\tau_u(a^t, S^t) < \alpha\right], u = 1, \ldots, B, t = 1, \ldots, h. \end{aligned} \tag{18}$$

D. Greedy Policy

The problem in (18) is a constrained stochastic optimization problem. We first propose a greedy policy, where the expected objective function value achievable at the current time slot is maximized.

In other words, it is optimal for the stochastic optimization problem (18) when $h = 1$. Then, we will prove that the greedy policy is optimal when we consider multiple horizons. The procedures for the online and offline phases are also illustrated.

Theorem 1. *The optimal policy for the optimal topology configuration problem is given by* (19).

$$\begin{aligned} &\underset{a^t}{\textit{maximize}} \quad \mathbb{E}\left[\sum_{u=1}^{u=B} \eta_u(a^t, S^t)\right] \\ &\textit{s.t.} \qquad\quad \mathbb{E}\left[\tau_u(a^t, S^t) < \alpha\right], u = 1, \ldots, B. \end{aligned} \tag{19}$$

Proof: The optimality of the greed policy is proven via induction. Consider at horizon $h = 1$, the optimal action to take is the maximizer of $\mathbb{E}\left[\sum_{u=1}^{u=B} \eta_u(a^1)\right]$, which is obviously the action given by greedy policy to maximize the expected rewards in one step.

Assume at horizon $h, h \geq 1$, the optimal policy is the greedy policy given in (19). Then at horizon $h + 1$,

$$\frac{1}{h+1}\mathbb{E}\left[\sum_{t=0}^{t=h+1}\sum_{u=1}^{u=B} \eta_u(a^t, S^t)\right] = \frac{1}{h+1}\mathbb{E}\left[\sum_{t=0}^{t=h}\sum_{u=1}^{u=B} \eta_u(a^t, S^t)\right] + \frac{1}{h+1}\mathbb{E}\left[\sum_{u=1}^{u=B} \eta_u(a^{h+1}, S^{h+1})\right]. \tag{20}$$

So provided the hypothesis that the greedy policy maximizes the first term in the above equation, the action to take to maximize the total expected rewards is the one to maximize the second item, which is equivalent to the case with horizon 1. Therefore, the greedy policy is the optimal policy for problem (18). ■

From the channel observation and delay, we obtain the belief state, $\mathbf{b}^t$, which is the probability mass function of the current CSI. The stochastic optimization problem in (19) can be converted into a deterministic optimization problem

$$\begin{aligned} &\underset{a^t}{\text{maximize}} \quad \sum_{u=1}^{u=B} \overline{\eta}_u(a^t) \\ &\text{s.t.} \qquad\quad \overline{\tau}_u(a^t) < \alpha, u = 1, \ldots, B. \end{aligned} \tag{21}$$

Techniques to solve such integer programming problems have been well developed. For example, mediate size problems can be solved efficiently by the branch and bound method, and very large scale integer programmings can be solved by heuristics such as Genetic algorithm.

The algorithm to address the stochastic optimization (18) includes an offline and online phases. In the offline phase, for each possible observation, a belief state is computed and the integer programming (21) is solved. For the purpose of illustration, we adopt a brute-force approach in the offline phase, as shown in Algorithm 1. The optimal policy computed in the offline phase is utilized in the online phase to make the optimal decisions based on the delayed observations, as shown in Algorithm 2.

Discussion: In offline phase, the major computation cost lies on storage and manipulations of the station-transition function A. In the worst case, A is a $Q^{B^2} \times Q^{B^2}$ matrix, where Q is the channel quantization level. Fortunately, A is a sparse matrix with the majority of the non-zero elements along the diagonal. This hugely reduces the storage requirements and the manipulation complexity. In the online phase, the complexity lies on selecting the best action achieving the best objective function value, which is equivalent to sorting an entry in GP_TABLE. So the complexity is $O(|\mathcal{A}|\text{log}(|\mathcal{A}|))$ since the fundamental limit of run-time complexity for sorting algorithms in worst case is linearithmic. In contrast, the complexity of computing the optimal policy for general partial observable Markov decision processes is PSPACE-complete which is considered to be harder than NP-complete problems.

V. Fairness and Efficiency Trade-off

As in many other multi-user systems, one important performance metric for MCC systems is fairness to the MCC users. This section investigates such metric based on the widely-used Jain's fairness index. For the system we consider, Jain's index is defined as

$$J(\overline{\eta}) = \frac{\left[\sum_{u=1}^{u=B} \overline{\eta}_u\right]^2}{\sum_{u=1}^{u=B} \overline{\eta}_u^2}, \tag{22}$$

where $\overline{\eta}$ is a vector, consisting of the throughput of all MCC users. Jain's index is between $\frac{1}{B}$ and 1 when the fairness to B users ranging from the least fair to the most fair.

We adopt the definition of optimal efficiency-fairness trade-off in [36]. In particular, we consider the action a^* obtaining the optimal efficiency-fairness trade-off, if there is not another action $a \neq a^*$ that satisfies either: (a) $\overline{\eta}(a) > \overline{\eta}(a^*)$ meanwhile $J(a) \geq J(a^*)$ or (b) $\overline{\eta}(a) \geq \overline{\eta}(a^*)$ meanwhile $J(a) > J(a^*)$. In practice, it is possible that even we achieve the optimal trade-off, the fairness index is still not

Algorithm 1 Brute-force search for the greedy policy: offline phase

```
$\{$Find the feasible action at time slot $t$.$\}$
for $o \in \mathcal{O}$ do
    Given the observation $o$, compute the belief state $\mathbf{b}$ according to (9) or (11).
    for $a \in \mathcal{A}$ do
        Compute the probability of error without any link-layer retransmission based on (3), then the average transmission time over wireless channel and packet error rate for all users according to (7) and (6), respectively.
        Calculate the expected throughput for split-TCP for cloud computing according to (14) and (15), then store the value in a table $TP\_TABLE$.
        Compute the expected response latency according to (17), then store the value in a table $L\_TABLE$.
        $\{$The following step is used for the discussion on efficiency-fairness trade-off in Section V.$\}$
        Compute the Jain's index according to (22), then store the value in a table $Jain\_TABLE$.
    end for
    $\{$Brute-force search for the feasible action$\}$
    for each response latency value $\tau$ in $L\_TABLE$ do
        if $\tau < \alpha$ then
            Store the corresponding action in a table $FA\_TABLE$, they are the feasible actions.
        end if
        if $FA\_TABLE$ is not empty then
            Among the feasible actions in $FA\_TABLE$, find the action achieving the maximum throughput in $TP\_TABLE$, $a^*$. If there are more than one action achieving the same optimal throughput, an arbitrary one is taken to break the tie.
        else
            The optimization problem is infeasible, no action will be taken. [3]
        end if
        Store $a^*$ in the greedy policy table $GP\_TABLE$
    end for
end for
```

Algorithm 2 Greedy policy: online phase

```
for $t = 0, 1, \ldots$ do
    Given the observation $o^t$, lookup the optimal action in table $GP\_TABLE$. The action achieving the best objective function value is executed in the coming decision period.
end for
```

Algorithm 3 Sub-procedure to explore the efficiency-fairness trade-off

```
if $FA\_TABLE$ is not empty then
    Among the feasible actions in $FA\_TABLE$, find the maximum throughput in $TP\_TABLE$, denoted as $\overline{\eta}^*$.
else
    Find the maximum throughput in $TP\_TABLE$, denoted as $\overline{\eta}^*$.
end if
Identify the actions such that $\overline{\eta} > \overline{\eta}^* \cdot \beta$. Among them, find the one obtaining the maximum Jain fairness referring to $Jain\_TABLE$, $a^*$, which will be stored in $GP\_TABLE$.
```

satisfactory. Therefore, in such circumstance, we have to sacrifice some efficiency for the benefit of fairness.

To study the trade-off between the efficiency and fairness, we introduce a parameter β in the formulation, which is the percentage of throughput used to enhance the fairness in Jain's index. In particular, denote the maximum sum throughput we can have in (21) as $\overline{\eta}^*$, we constrain the action such that the sum throughout is not less than $(1-\beta) * \overline{\eta}^*$ while maximizing the Jain's index. Namely, we have the following problem,

$$\begin{aligned} \underset{a^t}{\text{maximize}} \quad & J(a^t) \\ \text{s.t.} \quad & \sum_{u=1}^{u=B} \overline{\eta}_u(a^t) \geq (1-\beta) * \overline{\eta}^* \\ & \overline{\tau}_u(a^t) < \alpha, u = 1, \ldots, B. \end{aligned} \tag{23}$$

In the above optimization problem, the first constraint will be pushed to meet the right hand side as close as possible, since with smaller efficiency, a higher extent of fairness can be expected. To solve

TABLE I

SIMULATION PARAMETERS

Parameter	Value
Carrier frequence	2110 Mhz
RRH antenna height	24 m
Mobile device antenna height	0.5 m
Number of multipath components	6
Sampling duration	$1/500000$
Pathloss	$30.18 + 26 * \log 10(distance)$
Receiver noise power density	-174 dBm/Hz
Mobile user transmit power	20 dBm
Per user latency constraint	350 ms

this problem, the procedure in Algorithm 1 is modified to explore the trade-off. In particular, steps from 16 to 20 in Algorithm 1 are replaced by the sub-procedure defined in Algorithm 3.

VI. SIMULATION RESULTS AND DISCUSSIONS

In this section, simulation results are presented to show the effectiveness of the proposed scheme. For the channel model and the physical system model are implemented in matlab. The link layer performance data is fed to network simulator, ns2 (version 2.34). We conduct simulations using the following settings. There are three RRHs in the C-RAN. The maximum size of a cooperating set is 2. The wireless channel is Rayleigh fading channel, and the normalized Doppler shift ranges from 0.01 to 0.06. The bandwidth is 45 KHz. The link layer allows frames to be transmitted at most 3 times. For TCP flows, the payload size is 760 bytes. W_{max} is 6 MSS. Other parameters are shown in Table I. There are two existing schemes used for comparison. In the first one, the effects of imperfect CSI in C-RAN is not considered, and the topology configuration and rate allocation decisions are made based merely on current CSI observations to maximize TCP throughput in MCC systems, which is called *Existing scheme - perfect CSI*. In the second one, TCP throughput in MCC systems is not considered, and the decisions are made to maximize the physical layer throughput based on imperfect CSI [37], which is called *Existing scheme - physical layer throughput*.

A. Performance Improvement

We measure CSI delay in C-RAN using the unit of samples, the same as in [38]. The performance metrics considered are sum TCP throughput of the MCC users and the average response latency among the MCC users. Fig. 4 and Fig. 5 show the performance of the three schemes in the low mobility scenario where the normalized Doppler shift is 0.01. Fig. 6 and Fig. 7 illustrate the results in the high mobility scenario where the normalized Doppler shift is 0.06. In the simulations, the response latency threshold α is set to be 0.35 seconds for the proposed scheme.

From these figures, we can observe that the proposed scheme outperforms the existing ones in terms of both system sum TCP throughput and response latency. In the low mobility scenario, the sum TCP throughput of both the proposed scheme and the existing scheme assuming perfect CSI in C-RAN drops slowly as the CSI delay increases. Nevertheless, the proposed scheme achieves more throughput than the existing scheme, for example, with the delay in CSI being 10 samples, by around 30%. Meanwhile, for the existing scheme assuming perfect CSI, the user response latency increases as the CSI gets more and more delayed. In the high mobility case, the proposed scheme can obtain higher throughput when the response latency is lower than the existing scheme assuming perfect CSI, as shown in Fig. 6 and Fig. 7. Note that when the delay is zero the proposed scheme still outperforms the existing schemes. That is because in the proposed scheme, we explicitly put the performance of split-TCP connections carrying MCC services into the formulation. In contrast, the existing schemes only aim to maximize physical layer throughput, which would not necessarily bring high performance for MCC.

In terms of throughput, Fig. 4 and Fig. 6 show that the performance of the existing scheme only considering physical layer throughput is the worst among the three. These two figures indicate that the existing scheme maximizing the physical layer throughput does not guarantee a higher TCP throughput. In terms of response latency, Fig. 5 and Fig. 7 show the results. In the low mobility case, as CSI delay increases, the response latency of the existing scheme maximizing physical layer throughput is getting close to that of the existing scheme assuming perfect CSI. In the high mobility case, it outperforms the existing scheme assuming perfect CSI when the CSI delay is larger than 2 samples. As shown in our previous work [37], the existing scheme maximizing the physical layer throughput has better

Thisarticlehasbeenacceptedforpublicationinafutureissueofthisjournal,buthasnotbeenfullyedited.Contentmaychangepriortofinalpublication.Citationinformation:DOI 10.1109/TVT.2015.2411739,IEEETransactionsonVehicularTechnology

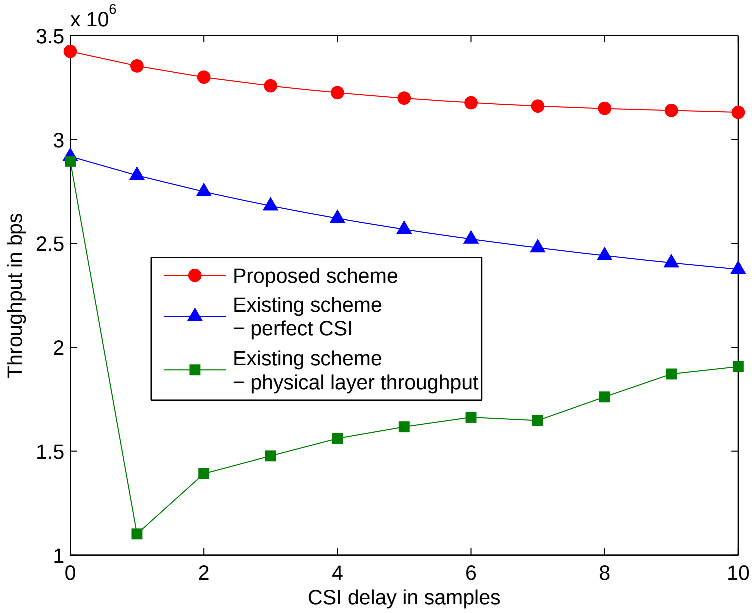

Fig. 4. The effect of delayed CSI on the end-to-end TCP throughput in the low mobility case with normalized Doppler shift 0.01.

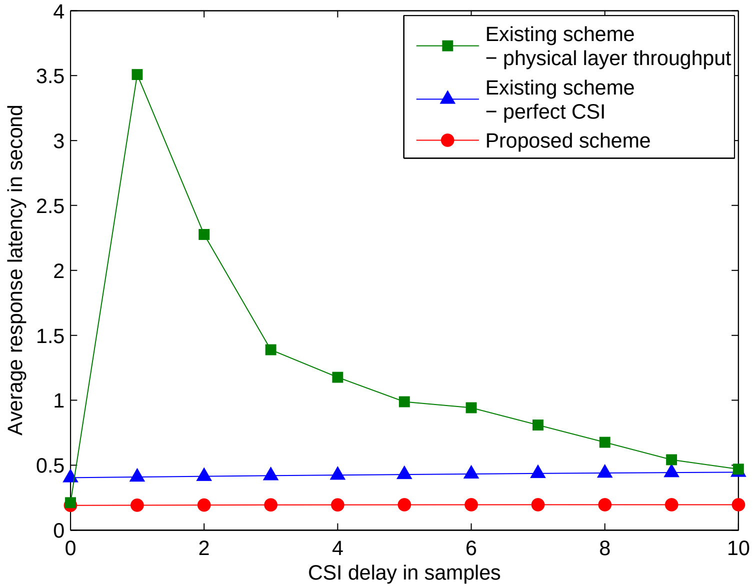

Fig. 5. The effect of delayed CSI on the average response latency in the low mobility case with normalized Doppler shift 0.01.

performance than the existing scheme assuming perfect CSI when the criterion is the sum rates of all the MCC users in the system. Furthermore, its advantage decades as CSI delay increases. However, such a scheme is not appropriate when the criterion is the sum TCP throughput of mobile cloud services.

The inherent reason is that the behavior of TCP is affected by not only the physical layer throughput but also the round trip time and the end-to-end reliability. The existing scheme maximizing the physical layer throughput only strikes a balance between the outage probability and the rate allocation to achieve maximum physical layer throughout, which might be sub-optimal for MCC systems. So, when the delay is small, e.g., 2 samples, the existing scheme maximizing physical layer throughput has the worst performance. As the delay increases, the effectiveness of such a scheme in maximizing physical layer throughput decreases. Consequently, the TCP throughput and latency get close to the one under the existing scheme assuming perfect CSI. That is the reason why we can see a spike in the low CSI delay region in these figures.

Different from these two existing schemes, the proposed one not only considers the issue caused by the delayed CSI, more importantly, it also considers the ultimate performance of split-TCP carrying mobile cloud services. Hence, the simulations results indicate that the proposed scheme is the best one to dynamically configure the C-RAN in MCC systems. Therefore, we believe that it is critical to design and operate the wireless access network in the context of mobile cloud computing, and the joint optimization can have significant advantages compared with the schemes where these two sub-systems are considered separately.

B. *Effects of RTT_2 and $\overline{\eta}_{cloud}$*

Fig. 8 shows the effect of the RTT over wireline networks, RTT_2, on the response latency. In the figure, the response latency threshold is 0.35 s. With this threshold, feasible solutions can be found for (18), such that the average response latency is bounded under the threshold. Note that if the response latency threshold is very small and/or RTT_2 is very large, it may not be possible to find feasible solutions to (18). In this situation, other mechanisms, such as admission control, should be used to limit the number of MCC users in the system.

For an MCC user, the end-to-end throughput is decided by the two connections separated by the split-TCP proxy. The effect of the throughput of the connection between the split-TCP proxy and the OTT service provider cloud, $\overline{\eta}_{cloud}$, is shown in Fig. 9. With the increase of $\overline{\eta}_{cloud}$, higher end-to-end

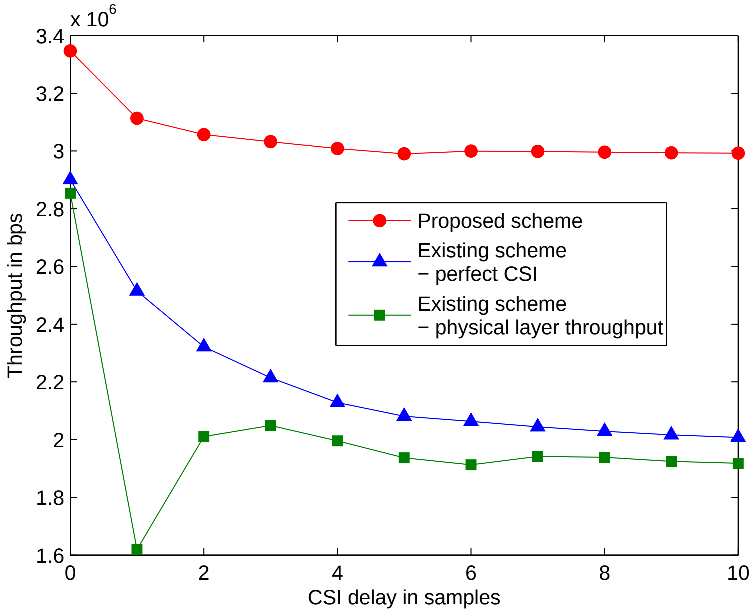

Fig. 6. The effect of delayed CSI on the end-to-end TCP throughput in the high mobility case with normalized Doppler shift 0.06.

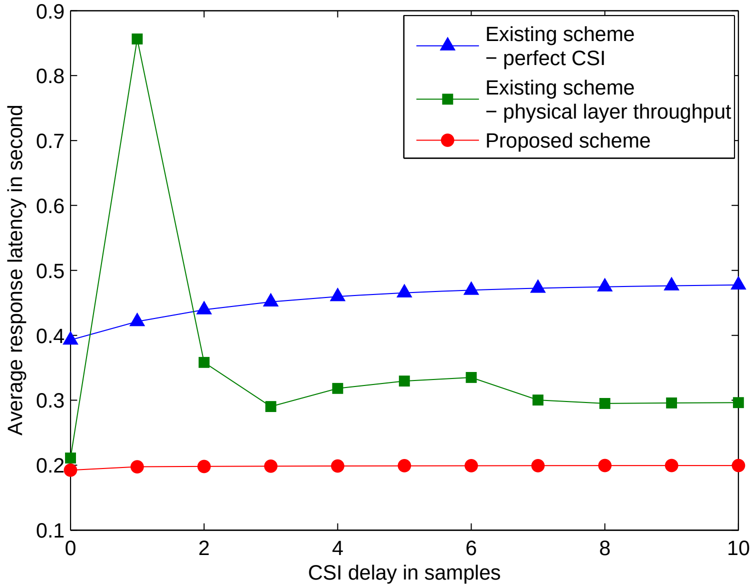

Fig. 7. The effect of delayed CSI on the average response latency in the high mobility case with normalized Doppler shift 0.06.

throughput is observed. Moreover, the end-to-end throughput hits a plateau when $\overline{\eta}_{cloud}$ is large enough. That means the bottleneck lies in the radio access network when the throughput between the split-TCP proxy and the OTT service provider cloud is sufficiently large. In this figure, we range the CSI delay

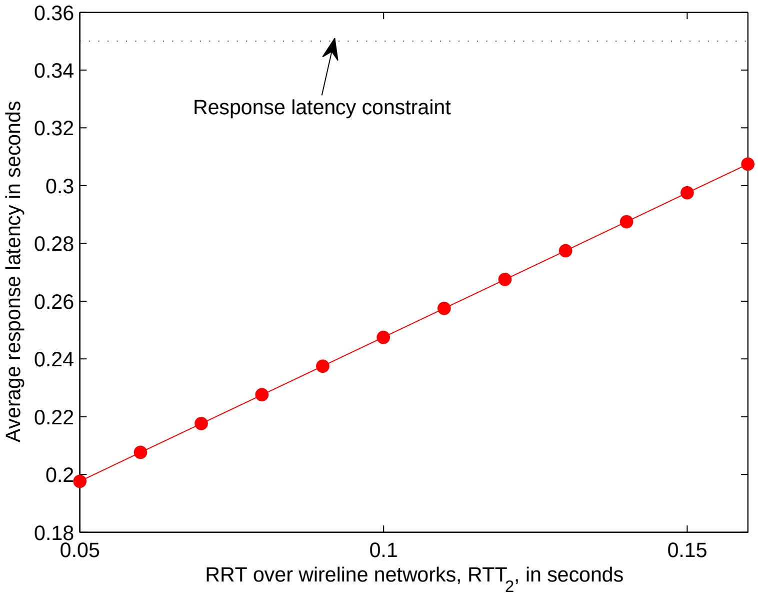

Fig. 8. The effect of response latency threshold α in the high mobility case with normalized Doppler shift 0.06.

from 0 to 8 samples. We can see that with the increase of CSI delay, the end-to-end throughput is getting smaller provided the same $\overline{\eta}_{cloud}$.

C. Efficiency-Fairness Trade-off

An interesting observation in the simulation is the effect of β (the percentage of throughput defined in Section V to enhance the fairness in Jain's index) with different settings of CSI delay. Fig. 10 and Fig. 11 show the results for low mobility and high mobility scenarios, respectively. With the increase of β, more and more efficiency, i.e., the system throughput, is traded-off for fairness. Yet, interestingly we observe that, with the increase of CSI delay, higher fairness is achieved provided the same setting of β. This implies that delay in CSI might help increase fairness, although delayed CSI is bad for system throughput.

VII. Conclusions and Future Work

In this paper, we jointly studied cloud-RAN and mobile cloud computing in next generation wireless networks. Particularly, the topology configuration and rate allocation problem in C-RAN has been

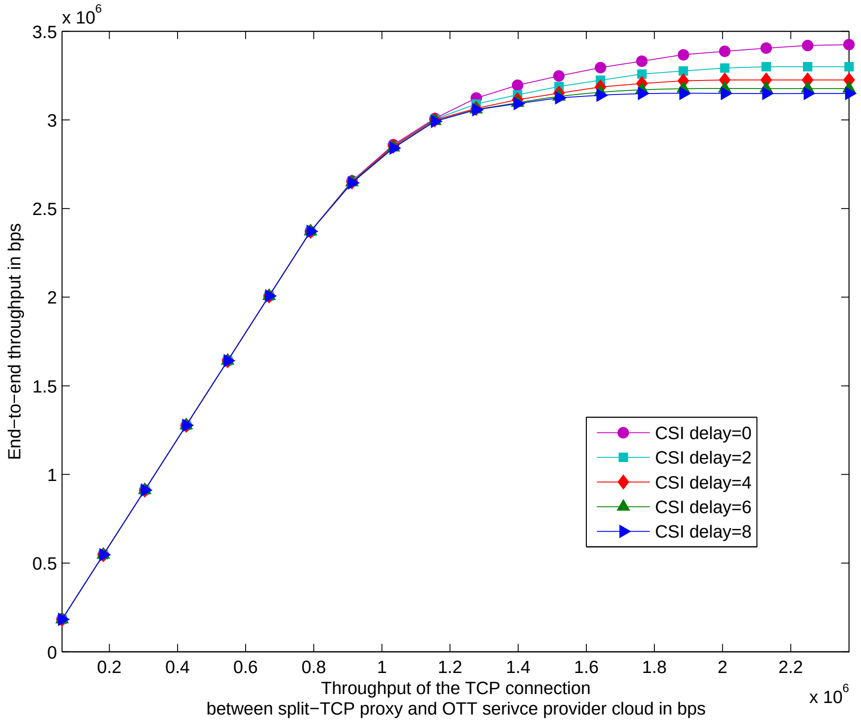

Fig. 9. The effect of $\overline{\eta}_{cloud}$ in the low mobility case with normalized Doppler shift 0.01.

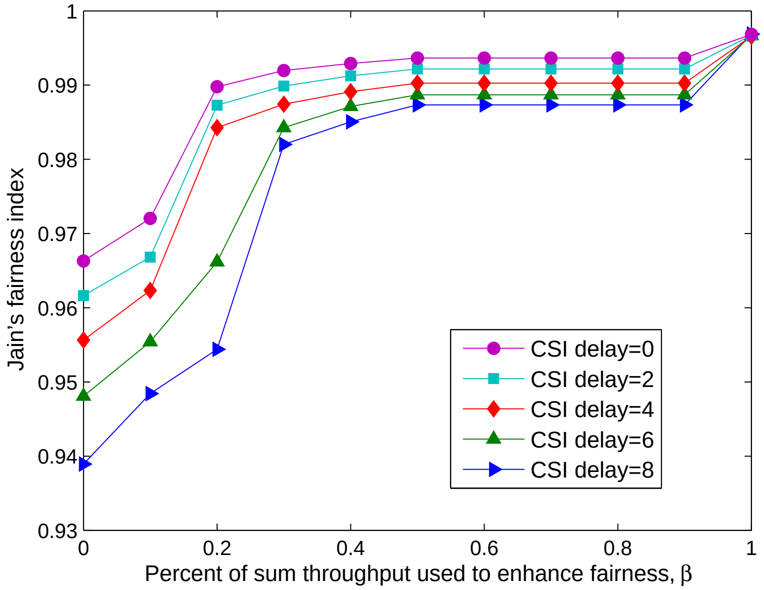

Fig. 10. Efficiency-fairness trade-off in the low mobility case with normalized Doppler shift 0.01.

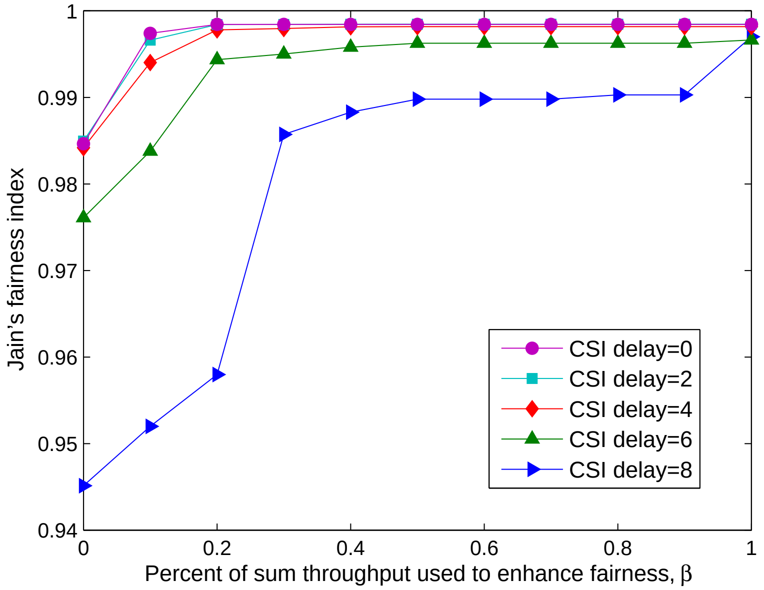

Fig. 11. Efficiency-fairness trade-off in the high mobility case with normalized Doppler shift 0.06.

investigated to improve the end-to-end TCP performance of MCC users in next generation wireless networks. We proposed a decision-theoretic approach to tackle the imperfect CSI problem in C-RAN. The response latency experienced by each MCC user was modeled as a constraint. We also studied the trade-off between the efficiency and the fairness among MCC users. Using simulation results, we showed that our proposed scheme can significantly improve the system performance in terms of throughput and response latency of MCC users. In particular, the delayed CSI in C-RAN has significant effect on the performance, and our proposed scheme is able to reduce such effect, especially in large delay and high mobility scenarios. Future work is in progress to consider wireless network virtualization [39] in the proposed framework.

Acknowledgement

We thank the reviewers for their detailed reviews and constructive comments, which have helped to improve the quality of this paper. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and in part by Huawei Technologies Canada CO., LTD.

REFERENCES

[1] G. Pallis, "Cloud computing: the new frontier of internet computing," *IEEE Internet Computing*, vol. 14, no. 5, pp. 70–73, 2010.

[2] NIST, "The NIST definition of cloud computing (draft)," Tech. Rep. Special Publication 800-145 (Draft), Jan. 2011.

[3] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless Communications and Mobile Computing*, 2011.

[4] S. Wang and S. Dey, "Adaptive mobile cloud computing to enable rich mobile multimedia applications," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 870–883, 2013.

[5] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "CloudIQ: A framework for processing base stations in a data center," in *Proc. ACM Mobicom'12*, (Istanbul, Turkey), 2012.

[6] China Mobile Research Institute, "C-RAN: the road towards green RAN," tech. rep. http://labs.chinamobile.com/, accessed: 2013-07-18.

[7] O. S.-H. Park, and Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Tech.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[8] Y. Cai, F. R. Yu, and S. Bu, "Cloud computing meets mobile wireless communications in next generation cellular networks," *IEEE Network*, vol. 28, pp. 54–59, Nov. 2014.

[9] S. Bu, F. R. Yu, Y. Cai, and P. Liu, "When the smart grid meets energy-efficient communications: Green wireless cellular networks powered by the smart grid," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 3014–3024, Aug. 2012.

[10] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Trans. Netw.*, vol. 5, pp. 756–769, Dec. 1997.

[11] C. Luo, F. R. Yu, H. Ji, and V. C. M. Leung, "Cross-layer design for TCP performance improvement in cognitive radio networks," *IEEE Trans. Veh. Tech.*, vol. 59, no. 5, pp. 2485–2495, 2010.

[12] N. H. Tran, C. S. Hong, and S. Lee, "Cross-layer design of congestion control and power control in fast-fading wireless networks," *IEEE Trans. Parallel and Dist. Systems*, vol. 24, pp. 260–274, Feb. 2013.

[13] G. Wang and T. S. E. Ng, "The impact of virtualization on network performance of Amazon EC2 data center," in *Proc. IEEE INFOCOM'10*, (San Diego, CA), Mar. 2010.

[14] 3rd Generation Partnership Project, "Reply LS to R3-070527/R1-071242 on Backhaul (X2 interface) Delay," tech. rep., 3rd Generation Partnership Project, 2007.

[15] D. J. Love, R. W. Heath, V. K. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 1341–1365, Oct. 2008.

[16] S. Bu, F. R. Yu, and H. Yanikomeroglu, "Interference-aware energy-efficient resource allocation for heterogeneous networks with incomplete channel state information," *IEEE Trans. Veh. Tech.*, online, 2015. dOI: 10.1109/TVT.2014.2325823.

[17] R. Xie, F. R. Yu, and H. Ji, "Dynamic resource allocation for heterogeneous services in cognitive radio networks with imperfect channel sensing," *IEEE Trans. Veh. Tech.*, vol. 61, pp. 770–780, Feb. 2012.

[18] A. Goldsmith, M. Effros, R. Koetter, M. Medard, A. Ozdaglar, and L. Zheng, "Beyond Shannon: the quest for fundamental performance limits of wireless ad hoc networks," *IEEE Comm. Mag.*, vol. 49, pp. 195 –205, May 2011.

[19] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, “The case for VM-based cloudlets in mobile computing,” *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, 2009.

[20] A. Pathak, Y. A. Wang, C. Huang, A. Greenberg, Y. C. Hu, R. Kern, J. Li, and K. W. Ross, “Measuring and evaluating TCP splitting for cloud services,” in *Proc. 11th Int’l Conf. Passive and Active Measurement (PAM’10)*, (Berlin, Heidelberg), pp. 41–50, Springer-Verlag, 2010.

[21] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, “Wireless network cloud: architecture and system requirements,” *IBM Journal of Research and Development*, vol. 54, no. 1, pp. 4:1–4:12, 2010.

[22] D. Bojic and N. Europe, “Advanced wireless and optical technologies for small-cell mobile backhaul with dynamic software-defined management,” *IEEE Comm. Mag.*, vol. 51, no. 9, pp. 86–93, 2013.

[23] Z. Zhu, P. Gupta, Q. Wang, S. Kalyanaraman, Y. Lin, H. Franke, and S. Sarangi, “Virtual base station pool: Towards a wireless network cloud for radio access networks,” in *Proc. 8th ACM Int’l Conf. Computing Frontiers*, (New York, NY, USA), 2011.

[24] W. Wei, C. Zhang, H. Zang, J. Kurose, and D. Towsley, “Inference and evaluation of split-connection approaches in cellular data networks,” in *Proc. Passive and Active Measurement Conference*, 2006.

[25] V. Farkas, B. Hder, and S. Novczki, “A split connection tcp proxy in LTE networks,” in *Information and Communication Technologies* (R. Szab and A. Vidcs, eds.), vol. 7479 of *Lecture Notes in Computer Science*, pp. 263–274, Springer Berlin Heidelberg, 2012.

[26] R. Xie, F. R. Yu, H. Ji, and Y. Li, “Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells,” *IEEE Trans. Wireless Commun.*, vol. 11, pp. 3910 –3920, Nov. 2012.

[27] A. Gjendemsj, D. Gesbert, G. E. Oien, and S. G. Kiani, “Binary power control for sum rate maximization over multiple interfering links,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3164–3173, 2008.

[28] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge Univ Press, 2005.

[29] M. Assaad and D. Zeghlache, “Comparison between MIMO techniques in UMTS-HSDPA system,” in *Proc. IEEE 8th Int’l Sym. Spread Spectrum Techniques and Applications*, pp. 874–878, 2004.

[30] H. S. Wang and N. Moayeri, “Finite-state Markov channel - a useful model for radio communication channels,” *IEEE Trans. Veh. Tech.*, vol. 44, pp. 163 –171, Feb. 1995.

[31] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, “Planning and acting in partially observable stochastic domains,” *Artificial Intelligence*, vol. 101, pp. 99–134, May 1998.

[32] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, pp. 257 –286, Feb. 1989.

[33] J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose, “Modeling TCP Reno performance: a simple model and its empirical validation,” *IEEE/ACM Trans. Netw.*, vol. 8, pp. 133–145, Apr. 2000.

[34] A. Toledo, X. Wang, and B. Lu, “A cross-layer TCP modelling framework for MIMO wireless systems,” *IEEE Trans. Wireless Commun.*, vol. 5, pp. 920–929, Apr. 2006.

[35] Y. Cui, Q. Huang, and V. Lau, “Queue-aware dynamic clustering and power allocation for network MIMO systems via distributed stochastic learning,” *IEEE Trans. Signal Proc.*, vol. 59, pp. 1229 –1238, Mar. 2011.

[36] A. Bin Sediq, R. H. Gohary, R. Schoenen, and H. Yanikomeroglu, “Optimal tradeoff between sum-rate efficiency and Jain’s fairness

index in resource allocation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3496–3509, 2013.

[37] Y. Cai, F. R. Yu, and G. Senarath, "Optimal clustering and rate allocation for uplink coordinated multi-point (CoMP) systems with delayed channel state information (CSI)," in *Proc. IEEE ICC'13*, (Budapest, Hungary), June 2013.

[38] K. Huang, *MIMO Networking with Imperfect Channel State Information*. Ph.D. Thesis, The University of Texas at Austin, Texas, USA, 2007.

[39] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tutorials*, 2014. DOI: 10.1109/COMST.2014.2352118.

Yegui Cai received B.Eng. degree and M.Eng. degree both in Electrical Engineering from South China University of Technology, Guangzhou, China, in 2008 and 2010, respectively. He received his PhD degree in Electrical and Computer Engineering from Carleton University, Ottawa, Canada in 2014. He is currently a Mitacs postdoc research intern in Irdeto Canada and working on protecting online advertising.

F. Richard Yu (S'00-M'04-SM'08) received the PhD degree in electrical engineering from the University of British Columbia (UBC) in 2003. From 2002 to 2004, he was with Ericsson (in Lund, Sweden), where he worked on the research and development of wireless mobile systems. From 2005 to 2006, he was with a start-up in California, USA, where he worked on the research and development in the areas of advanced wireless communication technologies and new standards. He joined Carleton School of Information Technology and the Department of Systems and Computer Engineering at Carleton University in 2007, where he is currently an Associate Professor. He received the IEEE Outstanding Leadership Award in 2013, Carleton Research Achievement Award in 2012, the Ontario Early Researcher Award (formerly Premier's Research Excellence Award) in 2011, the Excellent Contribution Award at IEEE/IFIP TrustCom 2010, the Leadership Opportunity Fund Award from Canada Foundation of Innovation in 2009 and the Best Paper Awards at IEEE ICC 2014, Globecom 2012, IEEE/IFIP TrustCom 2009 and Int'l Conference on Networking 2005. His research interests include cross-layer/cross-system design, security, green IT and QoS provisioning in wireless-based systems.

He serves on the editorial boards of several journals, including Co-Editor-in-Chief for Ad Hoc & Sensor Wireless Networks, Lead Series Editor for IEEE Transactions on Vehicular Technology, IEEE Communications Surveys & Tutorials, EURASIP Journal on Wireless Communications Networking, Wiley Journal on Security and Communication Networks, and International Journal of Wireless Communications and Networking, a Guest Editor for IEEE Transactions on Emerging Topics in Computing special issue on Advances in Mobile Cloud Computing, and a Guest Editor for IEEE Systems Journal for the special issue on Smart Grid Communications Systems. He has served on the Technical Program Committee (TPC) of numerous conferences, as the TPC Co-Chair of IEEE GreenCom'14, INFOCOM-MCV'15, Globecom'14, WiVEC'14, INFOCOM-MCC'14, Globecom'13, GreenCom'13, CCNC'13, INFOCOM-CCSES'12, ICC-GCN'12, VTC'12S, Globecom'11, INFOCOM-GCN'11, INFOCOM-CWCN'10, IEEE IWCMC'09, VTC'08F and WiN-ITS'07, as the Publication Chair of ICST QShine'10, and the Co-Chair of ICUMT-CWCN'09.

Shengrong Bu received the PhD degree in electrical and computer engineering from Carleton University in 2012, and held a research position at Huawei Technologies Canada Inc. Ottawa as a NSERC IRDF until 2014. She is now a Lecturer (Assistant Professor equivalent) with the School of Engineering at the University of Glasgow, UK. Her research interests include energy-efficient networks and systems, cyber-physical systems including the smart grid, wireless and mobile ad-hoc networks, wireless techniques for healthcare, wireless network security, cloud computing, game theory and stochastic optimization. She received the best student paper award at IEEE INDIN2005, 2012 IEEE Communications Society TAOS Technical Committees Award for Best Paper at IEEE GLOBECOM’2012, and one of the Best 50 Papers at IEEE GLOBECOM’2014.

Dr. Bu has served as an associate editor for Springer Wireless Networks, as the TPC co-chair of the IEEE ICC Workshop on Green Communications and Networks with Energy Harvesting, Smart Grids, and Renewable Energies (2015), the IEEE INFOCOM Workshop on Mobile Cloud and Virtualization (2015), and the ICCC workshop on Green Mobile Computing Networks (2013), and as a TPC member of numerous conferences. She is a reviewer for various journals, including IEEE JSAC, IEEE TWC, IEEE TPDS, IEEE TVT.