Cloud Radio Access Networks (C-RAN) in 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*—Cloud computing will have profound impacts on wireless networks. On one hand, the integration of cloud computing into the mobile environment enables mobile cloud computing (MCC) systems; on the other hand, the powerful computing platforms in the cloud for radio access networks lead to a novel concept of cloud radio access networks (C-RAN). In this paper, we study the topology configuration and rate allocation problem in C-RAN with the objective of optimizing the end-to-end performance of MCC users in next generation wireless networks. An intrinsic issue related to such system is that only sub-optimal decisions can be made due to the fact that the channel state information is outdated. We employ a decision-theoretic framework to tackle this issue, and maximize the system throughput with constraints on the response latency experienced by each MCC user. 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 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]. Cloud computing has attracted significant attention, and several commercial clouds, including Amazon EC2, Microsoft Azue, 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 [2]. 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 [2].

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) [3]. 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 [3].

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. 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.

In this paper, we study the topology configuration and rate allocation problem in C-RAN with the objective of optimizing 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. We take a decision-theoretic approach, which has well-developed mechanisms to address the impacts of noisy and delayed CSI. *Response latency* experienced by cloud users has been recognized as one of the most important performance metrics in cloud computing [4], [5]. Therefore, to improve MCC users' QoS, we model the response latency experienced by each MCC user as a constraint in our formulation.

The rest of the paper is structured as follows. Section II describes the system. We formulate the problem as a decision theoretic problem in Section III. Simulation results are discussed in Section IV. Finally, we conclude this study

978-1-4799-3088-3/14/$31.00 ©2014 IEEE

in Section V with future work.

II. System Description

A. Mobile Cloud Computing with C-RAN

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 the cloud. For the C-RAN, there are B base stations (BSs) with one antenna each, which are denoted as a set $\mathcal{B}$. Compared with traditional base stations, the BSs here are simplified because most of the signal processing and decision making happen at the *wireless network cloud* [6]. The BSs are connected to the wireless network cloud via backhaul networks. Note that the implementation of the wireless network cloud varies. For example, in [7], the central processing and control unit is called *virtual base station pool*.

Many mobile cloud services require the end-to-end reliable data transfer, TCP, 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 traditional cellular networks [8]. 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 [8].

We denote the channel state matrix at time slot t as S^t. Denote 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. The overall action is $a^t \triangleq \{\Omega^t, \mathbf{R}^t\} \in \mathcal{A}$, where $\mathcal{A}$ is the set of actions available.

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 [3].

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

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

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 $\mathcal{B} - \omega$. Given the channel matrix, we can compute the data rates. E.g., with Minimum Mean Square Error - Successive Interference Cancelation (MMSE-SIC) receiver and fixed decoding order, the capacities of users ranging from $1, 2, \dots, K$ are

$$\begin{aligned} C_K &= \log\left(1 + \tfrac{P\|\mathbf{h}_K\|^2}{N_K}\right), \\ C_{K-1} &= \log\left(1 + P\mathbf{h}_{K-1}^T A^{-1}\mathbf{h}_{K-1}\right), \\ &\dots \\ C_1 &= \log\left(1 + P\mathbf{h}_1^T(N_1 I_K + \textstyle\sum_{l=2}^{K} P\mathbf{h}_l\mathbf{h}_l^*)^{-1}\mathbf{h}_1\right), \end{aligned} \qquad (2)$$

where $A = (N_{K-1}I_K + P\mathbf{h}_K\mathbf{h}_K^*)$, $N_l, l = 1, 2, \dots, K$, 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. Note that MMSE-SIC receiver has been recognized as a sound technique for uplink CoMP [9, Chapter 6.1.2].

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

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

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). We assume a chase combining scheme in the following. The performance of such an HARQ scheme has been analyzed in [10]. 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

$$\mu_u = \frac{1 + p_{1,u} - p_{1,u}p_{2,u}}{1 - p_{1,u}p_{2,u}}, \qquad (4)$$

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

where $p_{1,u}$ is the probability of error after decoding the information block by forward error correction, $p_{2,u}$ is the

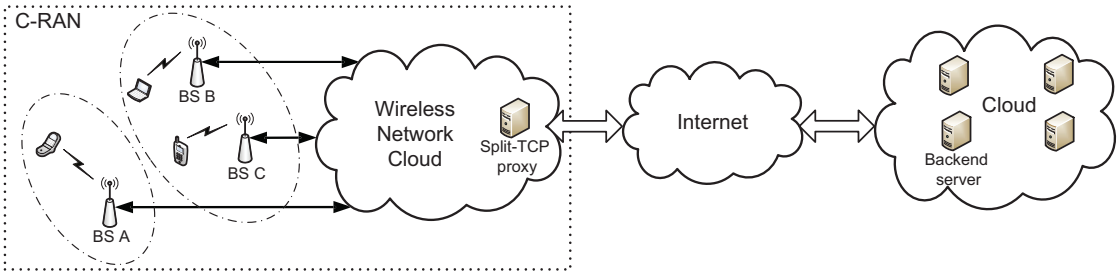

Fig. 1. A cloud radio access network in the MCC environment.

probability of error after soft combining two successive transmissions of the same information block [10]. 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 [10].

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.

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

1) 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 [11]–[13], the state-transition function A is given by

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

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 BS 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 BSs. 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 BSs 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 BSs, then are propagated over the backhaul networks. At the moment of decision making, the available CSI is outdated. 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. Such mapping is essential in modeling the delay under the Markov chain framework. Specifically, the d steps transition probability is given by A^d.

2) Belief State: Given the delay in steps, we can derive the *belief state*, which is the sufficient statistic of the previous action and observation history [14]. 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 [14], 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

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

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.

With the belief state $\mathbf{b}^t$, we can compute the probability of

error without link-layer retransmission (3) as follows.

$$p_{1,u} = \sum_{c_u(s',a^t)<r_u^t} b(s'), \tag{10}$$

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.

III. 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 discuss a topology configuration and rate allocation algorithm.

A. Round Trip Time and Split-TCP Throughput

Split-TCP has become the dominant reliable data transfer protocol for data center networks and legacy cellular networks. 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 [15]. It has been used in cross-layer designs to maximize TCP throughput (for instance, [16]). 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). There are two types of RTTs [5]. 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 cloud. We will not discuss the randomness in RTT_2 because this work is focused on the effect of C-RAN on cloud service.

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

$$\overline{RTT}_1(a^t) = \overline{T}_{wireless}(a^t) + T_{backhaul}, \tag{11}$$

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

Padhye *et al.* have developed a model for TCP connections. For user u, the average throughput can be derived as [15]

$$\overline{\eta} \approx \min\left\{\frac{W_{max}}{\overline{RTT}}, \frac{1}{\overline{RTT}\sqrt{\frac{2bp_e}{3}} + T_0 \min\{1, 3V\}}\right\}, \tag{12}$$

where $V = \sqrt{\frac{3bp_e}{8}}p_e(1 + 32p_e^2)$, 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 [15]. 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 (11) 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

$$\overline{\eta}_u(a^t) = \min\left\{\overline{\eta}_{RAN,u}(a^t), \overline{\eta}_{cloud}\right\}. \tag{13}$$

B. Per-User Response Latency

The communication latency can be improved by careful design and operation of the C-RAN. 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 [5], 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

$$\tau(a^t, S^t) = (n+1)*RTT_1(a^t, S^t) + RTT_2 + T_{process}. \tag{14}$$

A typical value of n for search engine application is 4 [5]. 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

$$\overline{\tau}(a^t) = (n+1) * \overline{RTT}_1(a^t) + \overline{RTT}_2 + T_{process}, \tag{15}$$

where $\overline{RTT}_1(a^t)$ is defined in (11), $\overline{RTT}_2$ 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 BSs and allocates the rate for MCC users, denoted as a^t. Denote the end-to-end throughput of a mobile user u given by (13) 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 [14]. 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 (14) 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,

$$\begin{aligned} \underset{a^t, t=1,2,\ldots,h}{\text{maximize}} \quad & \mathbb{E}\left[\frac{1}{h}\sum_{t=1}^{t=h}\sum_{u=1}^{u=B}\eta_u(a^t, S^t)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\tau_u(a^t, S^t) < \alpha\right], u = 1, \ldots, B, t = 1, \ldots, h. \end{aligned} \tag{16}$$

D. Greedy Policy

The problem in (16) is a constrained stochastic optimization problem. We propose a greedy policy,

$$\begin{aligned} \underset{a^t}{\text{maximize}} \quad & \mathbb{E}\left[\sum_{u=1}^{u=B}\eta_u(a^t, S^t)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\tau_u(a^t, S^t) < \alpha\right], u = 1, \ldots, B. \end{aligned} \tag{17}$$

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 (17) can be converted into a deterministic optimization problem

$$\begin{aligned} \underset{a^t}{\text{maximize}} \quad & \sum_{u=1}^{u=B}\overline{\eta}_u(a^t) \\ \text{s.t.} \quad & \overline{\tau}_u(a^t) < \alpha, u = 1, \ldots, B. \end{aligned} \tag{18}$$

The algorithm to address the stochastic optimization (16) includes an offline and an online phases. In the offline phase, for each possible observation, a belief state is computed and the integer programming (18) is solved. And the actions to take are stored in a table. In the online phase, the action to take is looked up in the table according to realtime observation.

IV. Simulation Results and Discussions

We conduct simulations using the following settings. There are three BSs in the C-RAN. The maximum size of a co-operating 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. 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 - observed 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 [17], which is called *Existing scheme - physical layer throughput*.

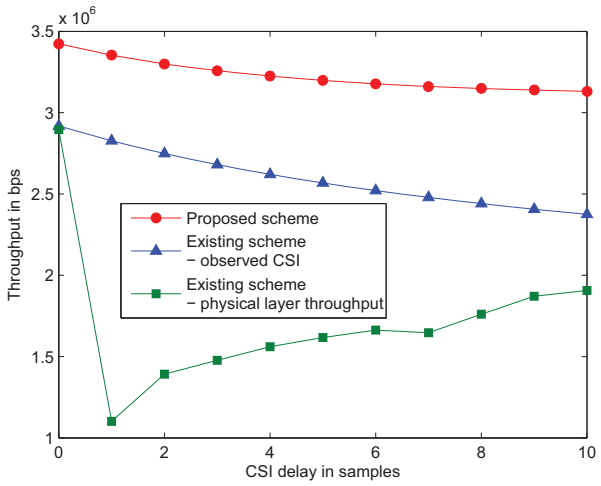

Fig. 2. The effect of delayed CSI on the end-to-end TCP throughput in the low mobility case with normalized Doppler shift 0.01.

A. Performance Improvement

We measure the CSI delay in C-RAN using the unit of samples. The performance metrics considered are sum TCP throughput of the MCC users and the average response latency among the MCC users. Fig. 2 and Fig. 3 show the performance of the three schemes in the low mobility scenario. 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 the 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 outdated.

In terms of throughput, Fig. 2 shows that the performance of the existing scheme only considering physical layer throughput is the worst among the three. It indicates that the existing scheme maximizing the physical layer throughput does not guarantee a higher TCP throughput. In terms of response latency, Fig. 3 shows that as the response latency of the existing scheme maximizing physical layer throughput is getting close to that of the existing scheme assuming perfect CSI. As shown in our previous work [17], the existing scheme maximizing the physical layer throughput has better 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 the delay in CSI increases. However, such a scheme is not appropriate when the criterion is the sum TCP throughput of mobile cloud services. The inherent

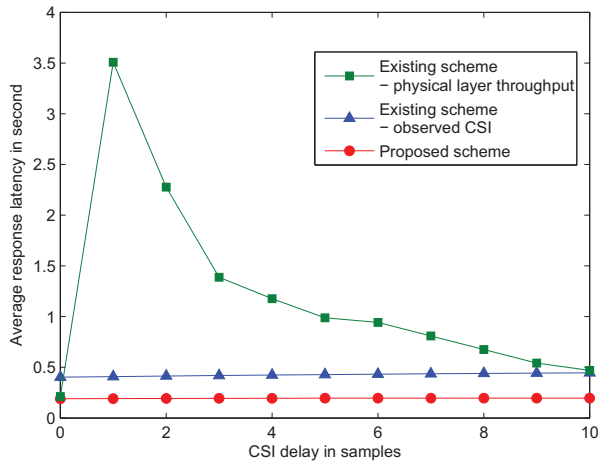

Fig. 3. The effect of delayed CSI on the average response latency in the low mobility case with normalized Doppler shift 0.01.

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 outdated 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 subsystems are considered separately.

V. Conclusions and Future Work

In this paper, the topology configuration and rate allocation problem in C-RAN has been 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. In the future, we will consider energy efficiency issues [18], [19] and wireless network virtualization in the proposed framework.

Acknowledgment

This work was supported in part by Huawei Technologies Canada and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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

[2] 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.

[3] China Mobile Research Institute, *C-RAN: The Road Towards Green RAN*. Research Report, http://labs.chinamobile.com/, accessed: 2013-07-18.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] P. Marsch and G. P. Fettweis, *Coordinated Multi-Point in mobile communications: from theory to practice*. Cambridge University Press, Jul. 2011.

[10] 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.

[11] 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.

[12] Y. Wei, F. R. Yu, and M. Song, “Distributed optimal relay selection in wireless cooperative networks with finite-state Markov channels,” *IEEE Trans. Veh. Tech.*, vol. 59, pp. 2149 –2158, June 2010.

[13] L. Zhu, F. R. Yu, B. Ning, and T. Tang, “Handoff performance improvements in MIMO-enabled communication-based train control systems,” *IEEE Trans. Intelligent Transportation Systems*, vol. 13, pp. 582 –593, June 2012.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] F. R. Yu, X. Zhang, and V. C. M. Leung, *Green Communications and Networking*. New York: CRC Press, 2012.

[19] 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.