Cloud Computing Meets Mobile Wireless Communications in Next Generation Cellular Networks

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 next generation cellular networks, cloud computing will have profound impacts on mobile wireless communications. 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 cellular networks. We use a decision theoretical approach to tackle the delayed channel state information problem in C-RAN. Simulation results show that 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; next generation cellular networks.

I. INTRODUCTION

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, broadband 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. 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 next generation mobile wireless cellular networks. On the one hand, with advances of wireless mobile communication technologies and devices, more and more end users access cloud computing systems via mobile devices, such as smartphones 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 computing technologies with limited computing, storage and energy resource in mobile devices.

On the other hand, the powerful cloud computing technology 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], [4]. 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 computation resource pool. This transition from distributed to centralized infrastructure for baseband processing can have significant benefits: saving the operating expenses due to centralized maintenance; enabling better load balancing; improving network performance due to advanced coordinated signal processing techniques; reducing energy expenditure by exploiting the load variations.

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 joint study of C-RAN in MCC for next generation cellular 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. Imperfect CSI has significant impacts on not only C-RAN, but also wireless networks in general. Since it is difficult to solve this problem using traditional information theoretic approach [5], we take a decision-theoretic approach, which has well-developed mechanisms to address the impacts of noisy and delayed CSI [6]. An optimal policy can be found based on the particular structure of the topology configuration and rate allocation problem. Simulation results show that 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.

The rest of the paper is structured as follows. Section II introduces the system. Section III discusses the issues caused by delayed CSI in C-RAN. We formulate the problem as a stochastic optimization problem in Section IV. Simulation results are discussed in Section V. Finally, we conclude this study in Section VI with future work.

II. MOBILE CLOUD COMPUTING WITH C-RAN FOR NEXT GENERATION CELLULAR NETWORKS

Fig. 1 shows a next generation cellular network with MCC and C-RAN technologies. Compared with traditional base stations (BSs), the BSs in C-RAN are simplified because most of the signal processing and decision making happen at the wireless network cloud. The BSs are connected to the wireless network cloud via backhaul networks. To provision better end-to-end data transfers, a popular solution is split-TCP. In terms of architecture, we can have a split-TCP proxy at the edge of the wireless network cloud. 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 [7]. The split-TCP proxy is the split point for TCP flows. In MCC systems, split-TCP proxy hides the wireless related issues from the backend servers. It acknowledges each segment and then stores and forwards the segment on the second TCP connection. Thanks to the adoption of split-TCP proxy, the application providers’ cloud does not need to change its policy for MCC users. The backend servers inside the cloud will be able to provide services to mobile users using general service models.

Fig. 2 shows the logical relationship of mobile devices, the wireless network cloud, and backend servers. The TCP flows carrying mobile cloud services run from the mobile devices to the backend servers in the cloud. Split-TCP proxy residing inside the wireless network cloud splits the end-to-end connection between
Fig. 1. A cloud radio access network in the MCC environment.

Fig. 2. Logical protocol stacks of network entities: data center, split-TCP proxy, and mobile user.

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 BSs cooperate with each other. For instance, in Fig. 1, BSs B and C form a cluster to serve the two mobile users together while BS A itself is another cluster. After clustering, the wireless network cloud needs to decide the data rates the mobile users can transmit. Inside a cluster, the signals are processed jointly such that there is no inference. The operations in C-RAN require the knowledge of the wireless channel, namely, CSI.
III. C-RAN WITH DELAYED CHANNEL STATE INFORMATION

In general, the backhaul networks in C-RAN are capacity limited and introduce delay in transmitting CSI. Such limitation introduces a delay between the observed channel state and the actual channel. To see how delay comes into C-RAN, we consider a C-RAN shown in Fig. 1 and its sequential operations shown in Fig. 3. 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 mobile users can transmit are decided after obtaining CSI. Then, the user data is transmitted. The same as the measurement and propagation of CSI, the user signals are transmitted from mobile 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. With the widely-used Markov chain channel model, we can introduce the delay issue into the channel modeling. We map the delay in seconds into the transition steps in Markov chains. Given the delay in steps, we can derive the belief state, which is the sufficient statistic of the previous action and observation history. A belief state $b^t$ at time slot $t$ is a probability distribution of the state space. The probability that the state...
at time slot $t$ is $s^t$ given by the corresponding element in $b^t$, denoted as $b(s^t)$. 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. Then we can derive the belief state from the delay and the transition probability matrix. 

IV. TCP THROUGHPUT OVER C-RAN IN MCC SYSTEMS

In this section, we study the TCP throughput over C-RAN in MCC systems and investigate the user response latency issue. The TCP throughput maximization with response latency constraint problem is formulated as a constrained stochastic optimization problem. To solve the stochastic optimization problem, we can first prove that the greedy policy (i.e., maximizing the instantaneous throughput) is optimal. Then we derive the belief state, which is the probability mass function of the actual system states. With the belief state, the stochastic optimization problem for instantaneous throughput maximization becomes a deterministic discrete optimization problem.

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 MCC next generation cellular networks. Therefore, in this work we adopt split-TCP as our underline transport layer protocol. A widely used TCP throughput model is developed in [9]. It has been used in cross-layer designs to maximize TCP throughput (for instance, [10]). In this section we extend the previous work to take delayed CSI into account in the TCP throughput model.

We firstly discuss RTT. Fig. 4 shows the round trip times for mobile cloud services over C-RAN. There are two types of RTTs. $RTT_1$ 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. $RTT_1$ is affected by the C-RAN operations. $RTT_1$ consists of $T_{\text{wireless}}$ and $T_{\text{backhaul}}$, which represent the round trip transmission time over the wireless and backhaul networks, respectively. Due to the wireless channel fading, $T_{\text{wireless}}$ is a random variable partly controlled by the action taken by the control unit in the wireless network cloud of C-RAN. While $T_{\text{backhaul}}$ can be considered as a constant, the mean value of

1If we do not know the delay, belief states can be derived based on Bayesian rule [8].
\( T_{\text{wireless}} \) can be derived from the link layer and physical layer models. The value of \( RTT_2 \) can be measured via techniques such as time-stamping.

Without loss of generality, 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. If a single TCP message cannot be loaded in to a frame, the link layer frame size can become part of the decision [11]. The queueing delay is not taken into consideration. Since we focus on the delayed CSI in this work, we can approximately take the queueing delay as a constant contribution to RTT. In terms of implementation, the queueing delay can be approximated by its statistical average. This assumption is reasonable because the time scale of queueing dynamics is generally larger than that of the wireless channel dynamics provided there is a bulk of data to transmit. If the files are small, the queueing delay can be simply ignored. Therefore, we take queueing delay as a constant, which has been considered in \( T_{\text{backhaul}} \).

As for the TCP packet error probability over wireless links, it is decided by the physical layer error probability, i.e., the outage probability, and the link layer protocol HARQ. A closed form expression is available in the literature provided a particular HARQ scheme is used.

With the average RTT and parameters including the maximum congestion window, the number of packets acknowledged by a TCP ACK (generally 2), the initial time-out for the TCP sender, and the TCP loss
probability, we can compute the throughput for a TCP connection based on the model developed in [9].
The accuracy of such a model has been verified against real TCP traces in [9]. In split-TCP, the end-to-end
throughput is the minimum throughput between the two TCP connections.

B. Per-user Response Latency

Response latency is critical for mobile cloud services [12]. Since the main computation tasks are per-
formed 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.

The authors of [12] propose a simple model for response latency in a cloud system using split-TCP. As
shown in Fig. 4, it takes about an $RTT_1$ in the hand-shaking phase for the connection between the client
and split-TCP proxy. The split-TCP proxy needs to wait an $RTT_2$ between data transmissions to the client.
Recall that, in MCC with C-RAN, $RTT_1$ is a random variable partially controlled by the action taken. So
the total response latency is a linear combination of the two types of RTTs and the processing time in the
backend servers. The coefficients of the combination is up to a particular cloud service [12].

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

Due to the outdated CSI, the performance of the mobile cloud services over split-TCP is a random variable
affected by the decisions in C-RAN. We take a decision-theoretic approach, which has well-developed
mechanisms to address the impacts of noisy and delayed CSI.

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 mobile users, denoted as $a^t$. The end-to-end throughput of
a mobile user $u$ is denoted as a random variable $\eta_u(a^t, S^t)$, and $\sum_{u=1}^{B} \eta_u(a^t, S^t)$ is the sum throughput of
the system. Denote the number of time slots considered as $h$, which is called the number of horizons in
Markov decision process literature [8]. The cumulative rewards over $h$ horizons is $\sum_{t=1}^{h} \sum_{u=1}^{B} \eta_u(a^t, S^t)$.

Accordingly, we denote the response latency defined in Subsection IV-B as $\tau_u$, and we constrain the
latency to be under a threshold $\alpha$. To maximize the sum TCP throughput and to constrain the response
latency, we have the following optimization problem,
\[
\begin{align*}
\text{maximize} & \quad \mathbb{E} \left[ \frac{1}{h} \sum_{t=1}^{t=h} \sum_{u=1}^{u=B} \eta_u(a^t, S^t) \right] \\
\text{subject to} & \quad \mathbb{E} \left[ \tau_u(a^t, S^t) < \alpha \right], u = 1, \ldots, B, t = 1, \ldots, h.
\end{align*}
\tag{1}
\]

The problem in (1) is a constrained stochastic optimization problem. We first propose a greedy policy, where the expected objective function value achievable in the current time slot is maximized. In other words, it is optimal for the stochastic optimization problem (1) when \( h = 1 \). We can prove that the greedy policy is optimal when we consider multiple horizons. The outline of the proof is as follows. Consider at horizon \( h = 1 \), the optimal action to take is the maximizer of \( \mathbb{E} \left[ \sum_{u=1}^{u=B} \eta_u(a^1, S^1) \right] \), 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. At horizon \( h + 1 \) case, provided the hypothesis that the greedy policy maximizes the first \( h \) steps, the action to maximize the total expected rewards is the one to maximize the \( h + 1^{th} \) step reward, which is equivalent to the case with horizon 1. Therefore, the greedy policy is the optimal policy for problem (1).

From the channel observation and delay, we can obtain the belief state, \( b^t \), which is the probability mass function of the current CSI. The stochastic optimization problem in the greedy policy can be transformed to a deterministic optimization problem since the mean value of the throughput and the response latency can be computed from the belief states. Moreover, the deterministic optimization problem is an integer programming problem since the actions are discrete. Techniques to solve such integer programming problems have been well developed. For example, medium size problems can be solved efficiently by the branch and bound method, and very large scale integer programming can be solved by heuristics such as Genetic algorithm.

V. SIMULATION RESULTS AND DISCUSSIONS

The performance of the proposed scheme is illustrated via computer simulations based on NS2. We conduct simulations using the following settings. The carrier frequency is 2.1GHz. There are three BSs in the C-RAN. The maximum size of a cooperating cluster is 2. The wireless channel is Rayleigh fading channel, the normalized Doppler shift of which 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. We consider a search engine application in the mobile cloud system \(^2\). Response latency is

\(^2\)Other MCC applications, such as mobile gaming and online social networks, can also be considered under our framework.
especially important for search engine services accessed from mobile devices. The setting of the response latency constraint $\alpha$ follows the one suggested in [12], which is set to be 0.35 seconds in the proposed scheme.

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 [13], which is called Existing scheme - Physical layer throughput.

### A. Performance Improvement

We measure the CSI delay in C-RAN using the unit of samples. Fig. 5 and Fig. 6 show the throughput and response latency in the low mobility and the high mobility scenarios, respectively.

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 drop slowly as the delay increases. Nevertheless, the proposed scheme achieves more throughput than the existing scheme, say, with the delay in CSI being 10 samples, by around 30%. Meanwhile, for the existing scheme assuming perfect CSI, the user response time increases as the CSI gets more and more outdated. A higher performance gain in the proposed scheme can be seen in the high mobility case shown in Fig. 6.

In terms of throughput, the performance of the existing scheme only considering physical layer throughput is the worst among the three. In terms of response latency, in low mobility case, as CSI delay increases, the latency is getting close to that of the existing scheme assuming perfect CSI; in high mobility case, it outperforms the existing scheme assuming perfect CSI when the delay is larger than 2 samples. As shown in our previous work [13], 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 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 reason is that the behavior of TCP is not only controlled by the physical layer throughput but also
the round trip time and the end to end reliability. The policy maximizing the physical layer throughput strikes a balance between the outage probability and the rate allocation to achieve maximum physical layer throughput, which is suboptimal for mobile cloud services. So when the delay is small, e.g., 2 samples, the maximizing physical layer throughput scheme 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 observe a spike in the low CSI delay region in these figures. From these results, we can see that it is critical to design and operate C-RAN that can adapt to the behavior of end-to-end TCP carrying mobile cloud services in next generation cellular networks.

Note that the RTT over wireline networks, $RTT_2$, can have effects on both user response latency and system throughput. With very large $RTT_2$ or very small response latency threshold, it may not be possible to find feasible solutions to (1). In this situation, other mechanisms, such as admission control, should be used to limit the number of MCC users in the system.
VI. CONCLUSIONS AND FUTURE WORK

In this paper, we jointly studied cloud-RAN and mobile cloud computing in next generation cellular networks. Particularly, 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 cellular 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. 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 in the proposed framework.

ACKNOWLEDGEMENT

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


---

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. Since 2011, he has been a PhD student in the Department of Systems and Computer Engineering at Carleton University, Ottawa, Canada. His main research interest is in computer networks.
F. Richard Yu is currently an Associate Professor at Carleton University, Canada. 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 Globecom 2012, IEEE/IFIP TrustCom 2009 and Int’l Conference on Networking 2005. His research interests include cross-layer design, security, green IT and QoS provisioning in wireless networks. He serves on the editorial boards of several journals, including IEEE Transactions on Vehicular Technology, IEEE Communications Surveys and Tutorials. He has served on the Technical Program Committee (TPC) of numerous conferences, as the TPC Co-Chair of IEEE Globecom’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, before joining Huawei Technologies Canada Inc. Ottawa as a NSERC IRDF. 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 and 2012 IEEE Communications Society TAOS Technical Committee’s Award for Best Paper at IEEE Globecom’2012. She has served as an associate editor for Springer Wireless Networks, on the TPC of conferences such as CNSR 2011, INFOCOM 2011 - GCN Workshop, GHTCE 2012, ICC’12 WS - GCN, INFOCOM’12 WS - CCSES, ICC 2013 Workshops - GMCN 2013, GC13 CogRN, IEEE GreenCom 2013 - CN, GNDS-2013, CCNC 2013, GC2014, and VTC-fall 2014, and as the TPC co-chair of the ICC workshop on Green Mobile Computing Networks. She is a reviewer for various journals, including IEEE JSAC, IEEE TWC, IEEE TPDS, IEEE Network Magazine.