Maximum Statistical Delay-Guarantee for Next-generation C-RAN: An Effective Capacity Based Approach

Hong Ren, *Student Member, IEEE*, Nan Liu, *Member, IEEE*, Cunhua Pan, *Member, IEEE*, Maged Elkashlan, *Member, IEEE*, Arumugam Nallanathan, *Fellow, IEEE*, Xiaohu You, *Fellow, IEEE* and Lajos Hanzo, *Fellow, IEEE*

Abstract

Cloud radio access networking (C-RAN) constitutes a promising architecture for next-generation systems. Beneficial centralized signal processing techniques can be realized under the C-RAN architecture. Furthermore, given the recent rapid development of cloud computing, the C-RAN architecture is an ideal platform for supporting network function virtualization (NFV), software-defined networking (SDN) and artificial intelligence (AI). However, most of the existing contributions on C-RAN are mainly focused on the physical layer issues. The next-generation networks are expected to support compelling wireless applications satisfying diverse delay requirements, such as ultra-reliable and low-latency communications (URLLC), etc. Hence, we invoke the effective capacity theory for statistical delay-bounded QoS provision in C-RAN architectures, where the delay is taken into account. Based on the system model proposed, we conceive sophisticated power allocation schemes for maximizing the effective capacity of both single-user and multi-user scenarios. Our simulation results show that a low delay outage probability can be guaranteed by appropriately choosing the delay exponent. Furthermore, our simulation results demonstrate that the proposed algorithm significantly outperforms the existing algorithms in terms of the achievable effective capacity. Finally, some open research challenges are highlighted.

H. Ren was with the Southeast University, Nanjing, China. (e-mail:renhong@seu.edu.cn). She is now with the Queen Mary University of London, London E1 4NS, U.K. N. Liu and X. You are with the Southeast University, Nanjing, China. (e-mail:{ nanliu, xhyu}@seu.edu.cn). C. Pan, M. Elkashlan and A. Nallanathan are with the Queen Mary University of London, London E1 4NS, U.K. (Email:{c.pan, maged.elkashlan, arumugam.nallanathan }@qmul.ac.uk). L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K. (e-mail:lh@ecs.soton.ac.uk).

This work is partially supported by the National Natural Science Foundation of China under Grants 61571123 and 61521061, the National 863 Program (2014AA01A702).

L.Hanzo would like to acknowledge the financial support of the European Research Council, Advanced Fellow grant Beam-Me-Up.

Index Terms

URLLC, C-RAN, 5G, Delay, Effective Capacity.

I. Introduction

Due to the substantially increased data volumes, the fifth-generation (5G) cellular networks are expected to significantly exceed the data throughput of 4G systems [1]. Massive MIMO systems constitute a promising technique of achieving this ambitious goal by exploiting the high degrees of spatial freedom [2], and have attracted substantial research attention. However, in centralized deployments the performance of massive MIMO systems tends to be limited by the correlated fading of antennas. This issue can be dealt with by deploying a large number of geographically distributed antennas for the sake of maintaining the benefits of massive MIMO. Furthermore, both the link quality and cell coverage are dramatically improved by this distributed architecture, since the average access distance of each user is significantly reduced. This is the so-called cloud radio access network (C-RAN) concept [3], which is a promising network architecture capable of achieving the ambitious next-generation goals.

However, most of the existing literature devoted to the C-RAN concept is focused on the physical layer issues and the system performance evaluation is mainly based on the concept of classic Shannon capacity. Although this information-theoretic framework is eminently suitable for analyzing the single-user link-efficiency, it gives no cognizance to the delay from data-link layer. One of the most challenging 5G operational models is constituted by ultra-reliable and low-latency communications (URLLC) [4] conceived for supporting tactile Internet applications [5], vehicle-to-vehicle communications [6], remote control of industrial manufacturing, etc. These applications have stringent end-to-end delay requirements (say around 1 ms). Additionally, some popular multimedia services, such as seamless lip-synchronized video conferencing and interactive gaming also impose stringent delay requirements. Hence, research attention also has to be dedicated to data-link layer by considering these delay requirements. It is of paramount importance to account for the quality of service (QoS) requirements quantified in terms of delay when designing next-generation transmission schemes.

Due to the highly time-varying wireless channel conditions, it is quite a challenge to guarantee *deterministic* delay-bounded QoS requirements for these compelling applications. Fortunately, the *statistical* delay-bounded QoS theory has been proven to be a powerful tool of handling the delay requirements of near-real-time traffic. More specifically, we can control the data rate of the incoming stream for ensuring that the delay-outage probability is always below a certain threshold. For example, in the Long Term

Evolution (LTE) Advanced standard, the probability that the delay of online gaming is higher than 50 ms should be kept below 2% [7]. To facilitate the analysis of statistical delay QoS performance, Wu *et al.* introduced the important notion of *effective capacity*, which represents the maximum constant packet arrival rate that can be supported by the system, whilst satisfying a maximum delay-outage probability constraint.

The rest of this paper is organized as follows. We briefly introduce the C-RAN architecture and show that C-RANs constitute an ideal platform of supporting salient paradigms, such as network function virtualization (NFV), software-defined networking (SDN) and artificial intelligence (AI) aided system optimization. We then introduce the effective-capacity-based statistical delay-bounded QoS provision concept into the C-RAN architecture and propose a dynamic power allocation algorithm that can adapt both to the delay requirements and to the channel conditions. We provide simulation results for quantifying the benefits of our proposed algorithm and show that extremely tight delay requirements can be met by using our proposed algorithm. Finally, we conclude with some future research challenges.

II. C-RAN ARCHITECTURE

The C-RAN architecture is shown in Fig. 1, which is composed of three parts:

1) Radio remote heads (RRHs) randomly located over the coverage area;
2) Baseband unit (BBU) pool with powerful cloud computing capability in a data center;
3) High-speed low-latency fronthaul links that connect the RRHs to the CPU.

The main feature of C-RANs is that the signal processing tasks of each small cell base station (BS) are migrated to the BBU pool, which is responsible for all the baseband signal processing, such as coordinated multi-point (CoMP) transmission, centralized resource allocation, joint user scheduling, data flow control, etc. The conventional full-functionality small BSs are replaced by low-cost RRHs, which are only used for low-complexity transmission and reception. Due to its low-complexity functionality, its size is smaller than that of the conventional small-cell BSs and can be readily installed on lamp-posts and building walls, hence imposing a low maintenance cost. In Fig. 1, the C-RAN is expected to support diverse applications, such as augmented reality (AR) based tele-conferencing, drone-based parcel delivery [8], tactile Internet, vehicular communication, smart factory support, etc.

Apart from the benefits of the air interface layer, this network architecture also enjoys further benefits at the network level. For example, compelling techniques, such as network function virtualization, software-defined networking and artificial intelligence, can be realized in this centralized architecture.

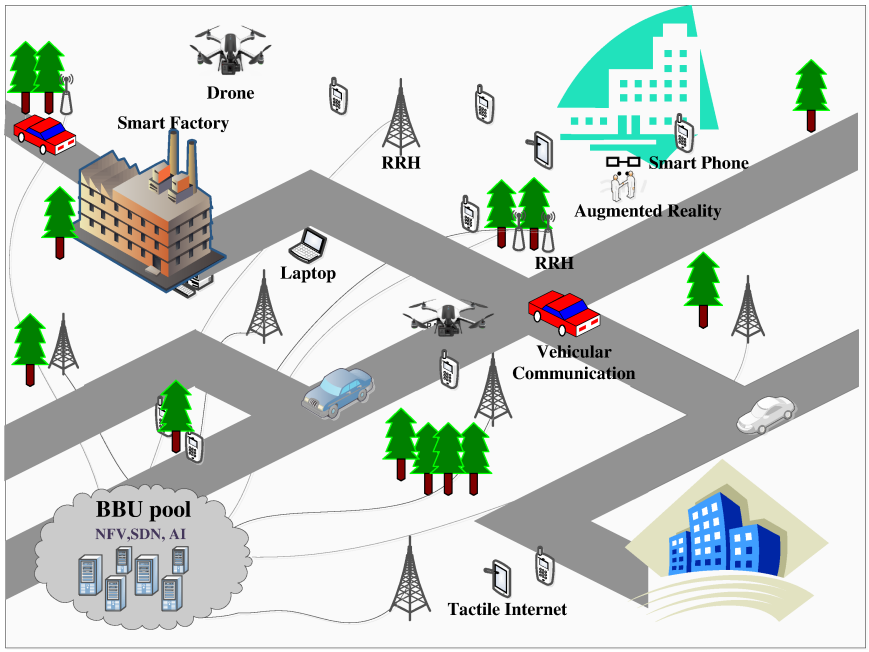

Fig. 1. Illustration of a 5G C-RAN architecture.

1) *Network Function Virtualization*: Through NFV, some network functions are separated from the conventional hardware infrastructure and can run on the cloud-computing infrastructure in the BBU pool with all the high-complexity power-thirsty signal processing tasks executed there. The main benefit of NFV is that sophisticated network functionalities can be dynamically supported depending on the near-instantaneous network state [1]. Additionally, new services can be created for discerning customers. More details about the NFV can be found in [9].
2) *Software-Defined Networking*: The SDN philosophy is at the heart of intelligent programmable networks. The key feature of SDN is that the control as well as data planes are decoupled, hence the network becomes more flexible in terms of supporting intelligent future applications. The key merit of this technology is the partitioning of network functionalities into separate software platforms, hence configuring the services by sophisticated programmable controllers. This technology is more amenable to employment in C-RANs, since the BBU pool is responsible for the whole suite of networking services. Its computing resources can be adaptively assigned and controlled through programmable controllers in the BBU pool.
3) *Artificial Intelligence*: User-centric clustering and proactive caching constitute a pair of key enabling techniques in C-RANs, which can be supported by machine learning. For user-centric clustering,

each user is cooperatively served by several of its nearby RRHs, which may indeed eliminate the cell-edge interference, provided that the near-instantaneous network conditions are known. However, this method may be unable to meet the stringent delay requirement of 5G, because excessive time is required for estimating the prevalent network state and to calculate the corresponding optimal cluster set for each user. This issue can be mitigated by using AI techniques [10]. Specifically, the BBU pool can store the users' historic data, such as their locations, the requested service, mobility pattern and speed, service demand profiles, channel characteristics, etc. By using machine learning techniques, these data can be analyzed and beneficially exploited. Then, one can predict the user's future locations, service request and even their channel information. Hence the future cluster of each user can be determined in advance, leading to low-latency predictive clustering algorithms. In C-RANs, the BBU pool is responsible for supporting the entire network. Hence, the AI-aided C-RAN is capable of forming globally optimal user-centric clusters. By contrast, the conventional cellular network is only capable of providing locally optimal solutions, since its operation is based on local information. Another promising technique in C-RANs is content caching. By caching the popular contents at the RRHs, the contents requested by the users can be directly transmitted from the nearby RRHs to the users, rather than fetching it from the core network. Hence, the access latency of the contents can be significantly reduced, therefore the fronthaul traffic is alleviated, which constitutes the bottleneck of C-RANs. The key question in cache-aided C-RAN is, which contents file should be cached in which RRH. This large-scale matching problem can also be solved by using AI techniques. For example, by analysing the users' history of requesting files from the BBU pool, machine learning is capable of calculating the file-popularity in support of this content placement problem.

Hence, the C-RAN architecture is an ideal platform of supporting the above low-delay techniques. In the following section, we introduce the effective capacity theory for statistical delay-bounded QoS provision over C-RAN.

III. The Effective Capacity Theory of Statistical Delay-Bounded QoS guarantee over C-RAN

The delay-bounded architecture of C-RANs is shown in Fig. 2. Each user's data stream is entered into its first-in-first-out (FIFO) buffer at a constant arrival rate of μ_k (measured in bit/s). At the data-link layer, the upper-layer packets are partitioned into transmission frames and then each frame will be mapped to bit-streams at the physical layer. Then, the BBU pool calculates the transmission rate

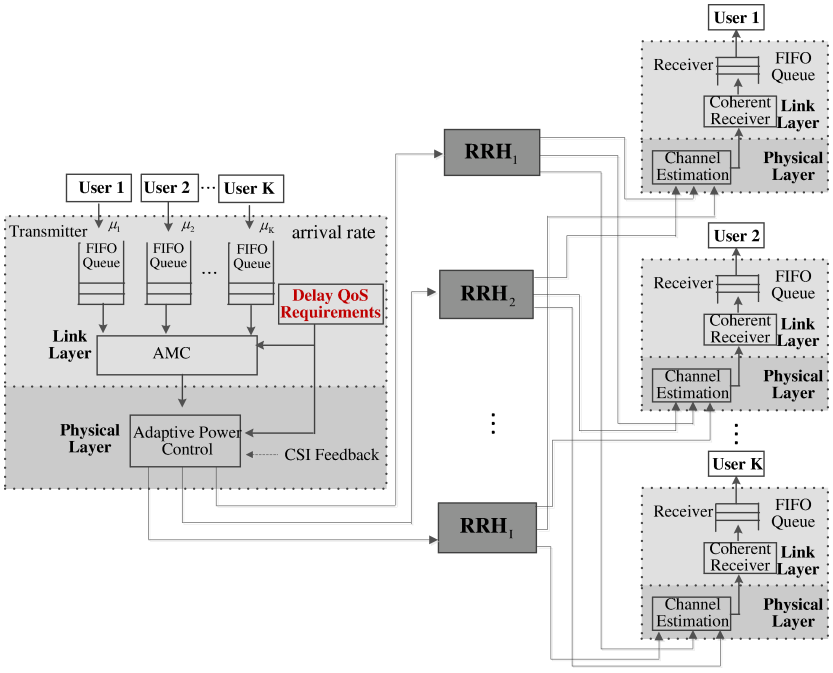

Fig. 2. The statistical QoS provisioning over the 5G C-RAN.

required and the power to be assigned to each user according to their delay requirements and to their channel state information (CSI) received via the feedback channel. Finally, the users' data streams are read out of the FIFO buffer and sent to all RRHs for transmission over the wireless channel at the service rates requested. The RRHs are assumed to be equipped with a single antenna. A block fading channel is considered, whose complex channel envelope is fixed during each transmission frame, and it is independently faded over different time frames.

We first introduce the important notion of the delay exponent θ that establishes the relationship between the maximum queue length and the buffer overflow probability, assuming that different users have different delay requirements characterized by $\theta_k, k = 1, \cdots, K$. For the C-RAN architecture of Fig. 2, the buffer overflow probability of the kth user is approximated by $e^{\theta_k Q_{\mathrm{th,k}}}$, where θ_k and $Q_{\mathrm{th},k}$ are the delay exponent and the maximum buffer length of user k. Hence, the delay exponent θ_k reflects the decay rate of the buffer overflow probability. A higher θ_k corresponds to a faster overflow decay rate, which implies that the system is capable of meeting a more stringent delay requirement for user k. By contrast, a lower θ_k leads to a slower buffer overflow decay rate, which represents a looser delay requirement for user k. In the extreme case of $\theta_k \to \infty$, the system cannot tolerate any delay, which corresponds to an extremely tight delay requirement for user k. On the other hand, when $\theta_k \to 0$, an arbitrarily long delay

can be tolerated by user k.

The probability that the delay is longer than a maximum bound of $D_{\max}$ can be approximated as [11]

$$P_{\text{delay}}^{\text{out}} = \Pr\{\text{Delay} \geq D_{\max}\} \approx \varepsilon e^{-\theta\mu D_{\max}}, \tag{1}$$

where ε is the probability that the buffer is non-empty. In general, the delay-violation probability of $P_{\text{delay}}^{\text{out}}$ has to be extremely low for the ULLRC services.

Based on the above discussions, we now introduce the important concept of effective capacity proposed by Wu *et al.* [11], which is defined as the maximum constant transmission frame arrival rate that the system can support, while satisfying a maximum delay-outage probability constraint. The effective capacity of user k is expressed as [11]

$$\text{EC}(\theta_k) = -\frac{1}{\theta_k}\log(\mathbb{E}\{e^{-\theta_k R_k}\}), \tag{2}$$

where $\mathbb{E}$ denotes the expectation operator, R_k is the instantaneous data rate of user k that is given by $R_k = T_f B \log_2\left(1+\sum_{i=1}^{I} p_{i,k}\alpha_{i,k}\right)$ with T_f, B, $p_{i,k}$ and $\alpha_{i,k}$ denoting the fixed length of each transmitted frame, the system bandwidth, the transmit power and channel gains from RRH i to user k, respectively. For simplicity, the multiuser interference is not considered here. If the delay-bound violation probability is $P_{\text{delay}}^{\text{out}}$, one should limit the incoming data rate to a maximum of $\mu_k = \text{EC}(\theta_k)$.

In conventional wireless communication systems, most of the contributions mainly focus on the ergodic capacity maximization problem, which ignores the delay requirement. By contrast, we aim for designing delay-bounded strategies to maximize the sum of the effective capacity of all users under the particular reqirement of all users. Specifically, we formulate the sum effective capacity maximization problem under the following constraints:

1) Each RRH has its individual average power constraint;
2) Each RRH is also subject to a specific peak power constraint.

The first constraint is closely related to the long-term power budget, while the second one is imposed for guaranteeing that the instantaneous power remains within the linear range of practical power amplifiers.

A. Single-user Case

We first study the single user case to glean initial insights. Due to the complex expression of the effective capacity, most of the existing contributions have been focused on the power allocation of single-transmitter scenarios, where only a single sum-power constraint is imposed. The optimal solution to this problem can be readily derived, which obeys a water-filling-like format. By contrast, in a C-RAN, all RRHs are subject to their individual power constraints, since the power cannot be shared among the

RRHs. Hence the conventional optimization method is no longer applicable and the power allocation of each RRH will no longer be the water-filling solution.

Hence we turn to convex optimization theory and derive the optimal power allocation in closed-form for the C-RAN scenario, which depends not only on the channel conditions, but also on the delay requirements. For the special case of a single RRH, the power allocation lends itself to the conventional water-filling solution. For the general case associated with multiple RRHs, the solutions reveal that the RRHs with higher channel gains have higher priorities to transmit with full power.

We can also find the closed-form solution for two extreme cases, namely when the delay exponent θ becomes zero and infinity. For the first case, the original optimization problem reduces to the conventional ergodic capacity maximization problem and its power allocation solution only depends on the channel conditions. For the latter case, the system cannot tolerate any delay and the optimal power allocation for each RRH reduces to the channel inversion associated with a fixed data rate.

B. Multiuser Case

Due to the powerful computational capability of the BBU pool, the C-RAN will serve multiple users. However, the expression of effective capacity is much more complex than that of the conventional Shannon capacity. The power control problem of the multiuser case is much more challenging to solve. To simplify the analysis, we assumed that all the RRHs transmit orthogonal signals to the different users in order to avoid the multiuser interference. Additionally, the peak power constraints are ignored for simplicity. In this case, we are able to obtain the optimal power allocation solution for each user in closed-form.

IV. Performance Evaluations

We performed simulations to evaluate the performance of our proposed power allocation scheme for a statistical delay-bounded C-RAN architecture deployed within a square area of 2 km $\times$ 2 km. We adopted the Nakagami-m block-fading subsuming the Rayleigh, Rician and the additive white Gaussian noise (AWGN) channel. The simulation results are based on the following parameters: Time frame of length $T_f = 0.04$ ms; system bandwidth of $B = 5$ MHz; the average power constraint and peak power constraint of each RRH are set to $P^{\text{avg}} = 0.5$ W and $P^{\text{peak}} = 1$ W, respectively; the Nakagami fading parameter is set to $m = 2$; the path-loss model is given by $PL_{i,k} = 148.1 + 37.6\log_{10} d_{i,k}$ (dB) [7], where $d_{i,k}$ is the distance between the ith RRH and the kth user measured in km; the noise power density is set as -174 dBm/Hz.

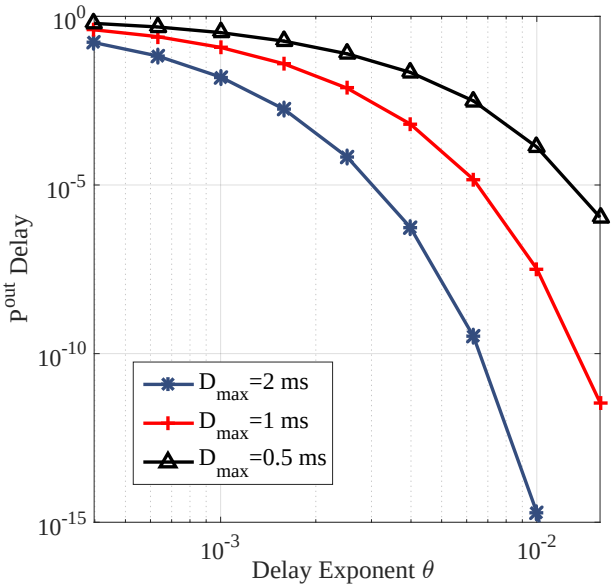

Fig. 3. Delay-outage probability versus delay exponent θ for various values of $D_{\max}$ for our proposed algorithm.

A. Single-user Case

We first consider the single-user case, where the user is located at the center of our C-RAN network. Let us assume that there are two RRHs with their coordinates randomly chosen as $[-600, 800]$ and $[900, 946]$ that is measured in meter.

Fig. 3 shows the delay-outage probability versus the delay exponent θ for our proposed power control algorithm. Three different values of the maximum delay threshold $D_{\max}$ are tested, namely, $D_{\max} = 2, 1, 0.5$ ms. The rate of incoming data streams is set as $\mu = \mathrm{EC}(\theta)$. As illustrated in Fig. 3, the delay-outage probability decreases rapidly with the delay exponent θ, since a higher θ implies a more stringent delay requirement. As expected, a higher $D_{\max}$ leads to a lower delay outage probability. When $D_{\max} = 1$ ms, the delay outage probability achieved by our proposed algorithm can be as low as 3.5×10^{-12}, when θ is chosen as $\theta = 10^{-1.8}$, which satisfies the stringent delay requirement of URLLC [4], while for the case of $D_{\max} = 2$ ms, the delay-outage probability can reach 10^{-15} when θ is set as $\theta = 10^{-2}$. Hence, the delay exponent can be adaptively set to satisfy the diverse delay requirements.

Next, we compare our algorithm to the following existing algorithms in terms of the achievable effective capacity:

1) *Nearest RRH serving algorithm*: As the terminology suggests, this algorithm assigns the nearest

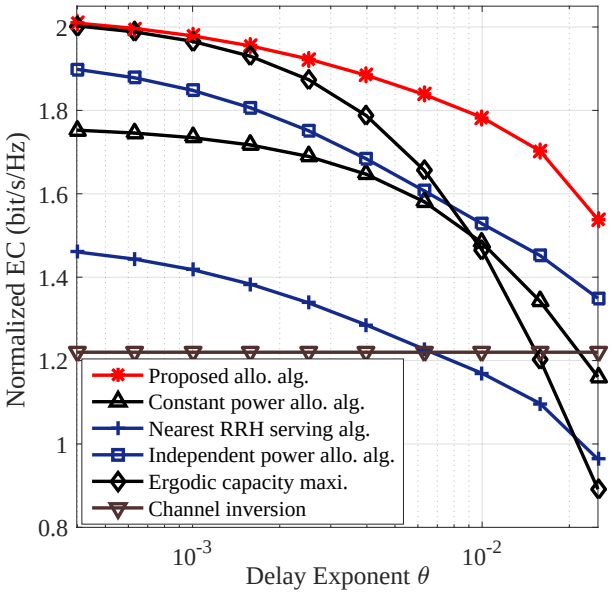

Fig. 4. Normalized EC for various algorithms vs delay exponent θ for a single user.

RRH to serve the user and the algorithm developed in [12] for simple point-to-point systems is used for solving the power allocation problem. This algorithm is provided to show the gains gleaned from cooperative transmission in C-RANs.

2) *Constant power allocation algorithm*: The transmit power of each RRH is set to its average power limit P^{avg}. This algorithm is used for showing the benefits of dynamic power allocation in the face of different channel conditions.
3) *Independent power allocation algorithm*: In this algorithm, each RRH independently optimizes its own transmission power purely based on its own channel conditions. This algorithm is provided for demonstrating the merits of optimizing the power allocation according to the joint channel conditions.
4) *Ergodic capacity maximization algorithm*: This algorithm maximizes the classic ergodic capacity for the user without incorporating the delay requirement.
5) *Channel inversion algorithm*: In this algorithm, the power allocation of each RRH is proportional to the channel inversion. This algorithm supports a constant transmission data rate.

Fig. 4 shows the normalized EC performance (which is the effective capacity divided by B and T_f) for the different algorithms versus the delay exponent θ. As illustrated in Fig. 4, the effective capacity

achieved by all algorithms (except the channel inversion algorithm) decreases with the delay exponent θ. Intuitively, a higher θ corresponds to a more stringent delay requirement and lower delay-outage probability requirement. Then, the maximum arrival rate that can be supported should be reduced for satisfying the stringent delay requirements. It is observed from this figure that our algorithm has a much better performance than the other algorithms, especially for high delay exponents. It is interesting to see that the performance of the ergodic capacity maximization algorithm approaches that of our proposed algorithm for low delay exponent θ, while it performs much worse than ours for a high θ. This can be explained as follows. When θ is small, the delay requirement is loose and then maximizing the effective capacity is approximately equivalent to maximizing the ergodic capacity, leading to similar performances for these two algorithms. However, for high θ, the delay requirement is very strict, which has to be taken into consideration when designing the transmission strategy, but this is not considered by the ergodic capacity maximization algorithm, hence resulting in a much worse performance. By using cooperative transmission among two different RRHs, the proposed algorithm has much better performance than the 'Nearest RRH serving algorithm', where only one RRH is applied for transmission. For example, when $\theta = 10^{-2}$, the performance gain is up to 0.6 bit/s/Hz. Since our proposed algorithm aims to optimize the power allocation according to the joint conditions of channel gains and delay exponents, the performance of our proposed algorithm significantly outperforms the 'Constant power allocation algorithm', where the power is kept fixed all the time. By optimizing the power allocation according to the joint channel conditions, our proposed algorithm achieves much higher normalized effective capacity than the 'Independent Power Allocation Algorithm'. As expected, the 'Channel Inversion' method has the worst performance across a wide range of θ values since it aims to provide constant data rate for various channel conditions.

B. Multiuser Case

Finally, in Fig. 5, we consider the multiuser case, where there are two users having the coordinates given by $[-100, 0]$ and $[0, 100]$, respectively. It is assumed that there are four RRHs located at $[650, 650]$, $[-650, 650]$, $[-650, -650]$, and $[650, -650]$. We compare our proposed algorithm to the ergodic capacity maximization algorithm in terms of the sum effective capacity performance. A similar performance trend has been observed to that of the single-user scenario of Fig. 4. For example, both algorithms have almost the same performance for low delay exponent θ, while our proposed algorithm outperforms the ergodic capacity maximization for high delay exponent θ and the performance gain increases with θ. In addition, we also compare the proposed algorithm with two other algorithms, namely, the 'Nearest RRH serving

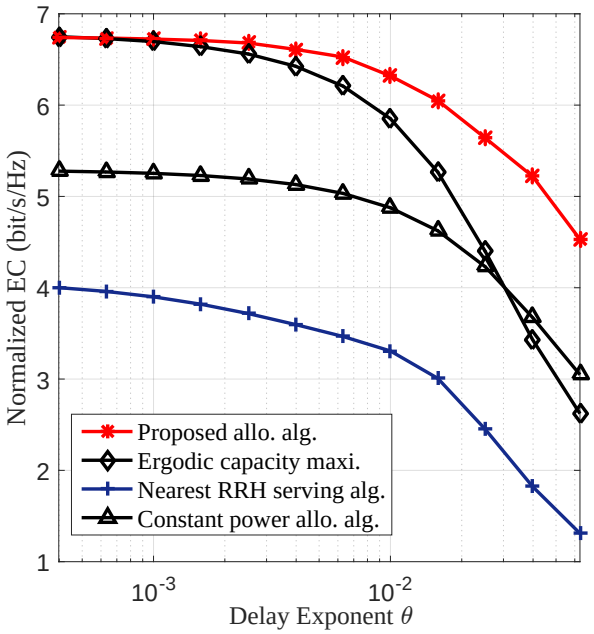

Fig. 5. The sum normalized EC vs delay exponent θ for our proposed algorithm and the ergodic capacity maximization algorithm, when supporting two users by four RRHs located at $[650, 650]$, $[-650, 650]$, $[-650, -650]$, and $[650, -650]$.

algorithm' and the 'Constant power allocation algorithm'. For the former algorithm, each user is served by its nearest RRH, while for the latter algorithm, the instaneous transmit power for each RRH is set to its average power limit P^{avg}, and the instaneous transmit power assigned by each RRH to each user is equal. It is seen from this figure that our proposed algorithm significantly outperforms these two algorithms. Specifically, the performance gain achieved by our proposed algorithm over these two algorithms are 2.8 bit/s/Hz and 1.5 bit/s/Hz, respectively, and the performance gain keeps almost fixed over all the delay exponent θ. By exploiting the multiuser diversity, the normalized EC achieved by the proposed algorithm for the two-user case is much larger than that of the single-user case.

V. Conclusions and Future Research Directions

We first highlighted the C-RAN architecture that consists of three components: the BBU pool, fronthaul links and RRHs. Centralized signal processing techniques can be relied upon by the C-RAN architecture, such as CoMP transmission, joint user scheduling and data flow control, etc. Additionally, the emerging techniques of NFV, SDN and AI can be intrinsically integrated with the C-RAN architecture. Then, we highlighted the effective capacity theory conceived for statistical delay-bounded C-RANs, where the

delay requirement was incorporated. Under the cross-layer C-RAN model, we proposed power allocation schemes for maximizing the sum effective capacity for both the single-user case and multiuser case considered. The simulation results showed that by appropriately choosing the delay exponent θ, the delay outage probability can be reduced below 10^{-9}, which is appealing for URLLC. Furthermore, the simulation results obtained also showed that our proposed algorithm significantly outperforms the existing algorithms in terms of the achievable effective capacity, especially in the case of stringent delay requirements.

However, substantial further research is required for delay-bounded C-RAN networks.

Interference Management: In this paper, we considered the idealized interference-free scenario, which typically leads to a convex optimization problem. However, when each RRH is equipped with multiple antennas, several users can be simultaneously served in the same time and frequency slot by adopting powerful beamforming techniques, which additionally improves the effective capacity performance. This kind of optimization problem becomes non-convex and hard to solve even for the simple Shannon capacity expression. The complex expression of the effective capacity makes the optimization problem challenging to solve, which needs further investigation in the future.

Limited Fronthaul Capacity: Due to their simple functionalities, RRHs can be densely deployed at low implementational cost [13]. Traditionally, the fronthaul links are usually fixed links, such as optical fibers or high-speed Ethernet. However, in densely deployed C-RANs, laying cables imposes high installation operational and maintenance costs. Hence, wireless communication links, such as millimeter wave (mmWave) transmission, are promising in this scenario. However, the available bandwidth is much lower even at mmWave frequencies than that of the fixed links. Hence, the limited fronthaul capacity should be taken into account when designing cross-layer operation.

Other Delay Sources: This paper only considered the queueing delay in the BBU pool. However, if the C-RAN is expected to cover a large area, then the propagation delay of the fronthaul links should also be taken into consideration. Furthermore, non-negligible time is required for calculating the power allocation for each user. In contrast to the LTE network, where the delays can be ignored, in URLLC the stringent delay requirements have to be carefully considered by future research. In this paper, we only focus on the delay incurred from the data-link layer. However, the delay incurred by the upper layer beyond the data-link layer should also be taken into account, such as routing and the access to a number of virtualized network functions. Furthermore, some more advanced user scheduling algorithms with low-complexity should also be developed to satisfy the stringent delay requirements.

Short Packet Transmission: In this paper, we adopted Shannon's capacity for guantifying the in-

stantaneous data rate in (2), which is accurate when the blocklength of channel codes is sufficiently large. However, in URLLC applications, short packets are preferred. Hence Shannon's capacity cannot be approached. She *et al.* mentioned this issue in [14] and introduced an approximate achievable data rate expression at a finite blocklength, which takes into account the transmission error probability. However, the resource allocation optimization problem based on this modified capacity expression does not lead to a convex optimization problem, which needs further investigation.

Energy efficiency issue: This paper focuses on the EC maximization problem. However, energy efficiency, defined as the ratio of data rate to total power consumption [15], is a key performance metric in the fifth generations (5G) cellular networks, and EE-oriented transmission design by considering the delay requirements needs further study.

REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, November 2010.

[3] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[4] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[5] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, March 2014.

[6] S. K. Datta, J. Haerri, C. Bonnet, and R. F. D. Costa, "Vehicles as connected resources: Opportunities and challenges for the future," *IEEE Vehicular Technology Magazine*, vol. 12, no. 2, pp. 26–35, June 2017.

[7] "Further advancements for E-UTRA physical layer aspects," *3GPP TR 36.814, Tech. Rep.*, 2010.

[8] J. Wang, C. Jiang, Z. Han, Y. Ren, R. G. Maunder, and L. Hanzo, "Taking drones to the next level: Cooperative distributed unmanned-aerial-vehicular networks for small and mini drones," *IEEE Vehicular Technology Magazine*, vol. 12, no. 3, pp. 73–82, 2017.

[9] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, Feb 2015.

[10] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: A survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 743–764, Secondquarter 2017.

[11] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, July 2003.

[12] W. Cheng, X. Zhang, and H. Zhang, "Statistical-QoS driven energy-efficiency optimization over green 5G mobile wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3092–3107, 2016.

[13] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint user selection and energy minimization for ultra-dense multi-channel C-RAN with incomplete CSI," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 8, pp. 1809–1824, Aug 2017.

[14] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, 2017.

[15] F. Zhou, Y. Wu, Q. Hu, Y. Wang, and W. Kai-kit, "Energy-efficient NOMA enabled heterogeneous cloud radio access networks," *arXiv preprint arXiv:1801.01996*, 2018.