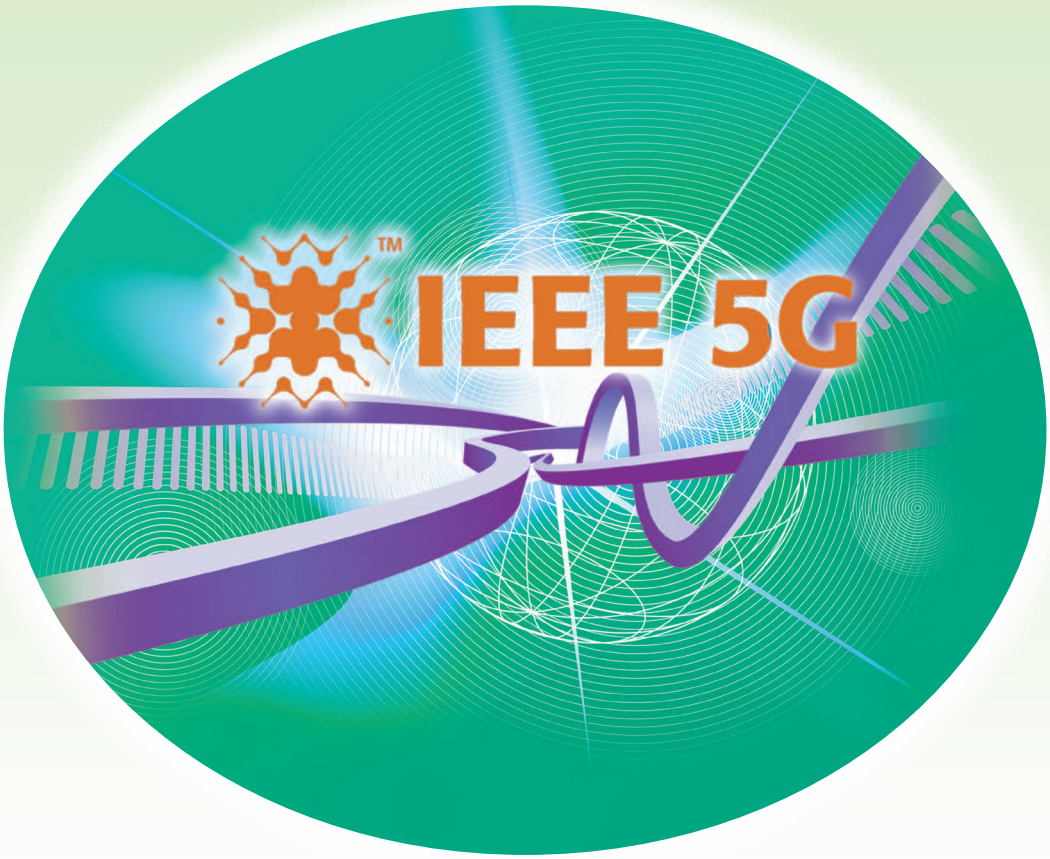
Low-Latency C-RAN



A Next-Generation Wireless Approach

Hong Ren, Nan Liu, Cunhua Pan, Maged Elkashlan,
Arumugam Nallanathan, Xiaohu You, and Lajos Hanzo

The cloud radio access network (C-RAN) constitutes a promising architecture for next-generation systems. Beneficial centralized signal processing techniques can be realized under the C-RAN framework. Furthermore, given the recent rapid development of cloud computing, this architecture is an ideal platform for supporting network function virtualization (NFV), software-defined networking (SDN), and artificial intelligence (AI).

Digital Object Identifier 10.1109/MVT.2018.2811244
Date of publication: 20 April 2018

However, most of the existing contributions in C-RAN are mainly focused on the physical-layer issues. The next-generation networks are expected to support challenging wireless applications that have diverse delay requirements, such as ultrareliable and low-latency communications (URLLC). Hence, we invoke the effective capacity (EC) theory for statistical delay-bounded quality of service (QoS) provision in C-RAN architectures, where the delay is taken into account. Based on the system model we propose, we conceive sophisticated power allocation schemes for maximizing the EC of

both single-user and multiuser scenarios. Our simulation results show that a low delay-outage probability can be guaranteed by appropriately choosing the delay exponent. Furthermore, the results demonstrate that the proposed algorithm significantly outperforms the existing algorithms in terms of the achievable EC. Finally, we highlight some open research challenges.

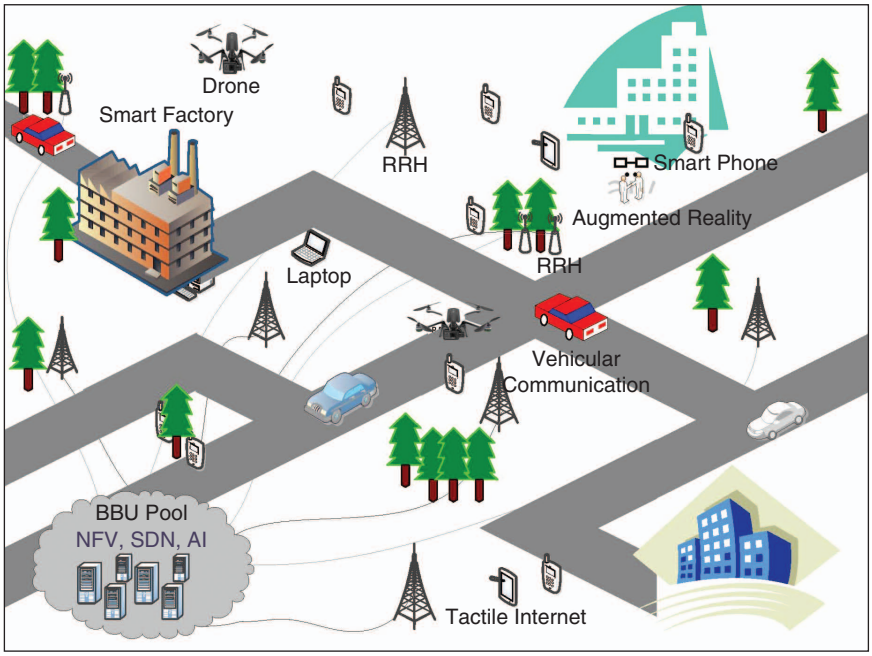


FIGURE 1 An illustration of a 5G C-RAN architecture.

A Quest for Low Latency

Fifth-generation (5G) cellular networks, because of the substantially increased data volumes they allow, are expected to significantly exceed the data throughput of fourth-generation systems [1]. Massive multiple-input, multiple-output (mMIMO) systems constitute a promising technique for 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 mMIMO 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 systems' benefits. 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 C-RAN concept [3], which offers 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 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 does not recognize the delay from the data-link layer. One of the most challenging 5G operational models is URLLC [4], conceived for supporting tactile Internet applications [5], vehicle-to-vehicle communications [6], remote control of industrial manufacturing, and so forth. These applications have stringent end-to-end delay requirements (approximately 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 the data-link layer by considering these delay requirements. It is of paramount importance to account for the QoS requirements quantified in terms of delay when designing next-generation transmission schemes.

Because of the highly time-varying wireless channel conditions, it is a considerable challenge to guarantee deterministic delay-bounded QoS requirements for these important applications. Fortunately, the statistical delay-bounded QoS theory has proven to be a powerful tool for handling the delay requirements of near-real-time traffic. More specifically, we can control the data rate of the incoming stream to ensure 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 in 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 EC, which represents the maximum constant packet arrival rate that can be supported by the system while satisfying a maximum delay-outage probability constraint.

C-RAN Architecture

The C-RAN architecture is shown in Figure 1 and composed of three parts:

- remote radio heads (RRHs) randomly located over the coverage area
- a baseband unit (BBU) pool, with a powerful cloud computing capability in a data center
- high-speed, low-latency fronthaul links that connect the RRHs to the central processing unit.

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 of the baseband

C-RANs are expected to support diverse applications, such as augmented reality-based teleconferencing.

signal processing, such as coordinated multipoint (CoMP) transmission, centralized resource allocation, joint user scheduling, and data-flow control. The conventional full-functionality small BSs are replaced by low-cost RRHs, which are used only for low-complexity transmission and reception. Because of its low-complexity functionality, the RRH is smaller than a conventional small-cell BS and can be readily installed on lampposts and building walls, hence imposing a low maintenance cost. In Figure 1, we can see that C-RANs are expected to support diverse applications, such as augmented reality-based teleconferencing, drone-based parcel delivery [8], tactile Internet, vehicular communication, and smart factory support.

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 NFV, SDN, and AI, can be realized in this centralized architecture.

- *NFV*: 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 NFV can be found in [9].
- *SDN*: The SDN philosophy is at the heart of intelligent programmable networks. The key feature of SDN is that the control and data planes are decoupled, so the network becomes more flexible in terms of supporting intelligent future applications. The key merit of this technology is the partitioning of network functionalities onto separate software platforms, 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.
- *AI*: Usercentric clustering and proactive caching constitute a pair of key enabling techniques in C-RANs, which can be supported by machine learning. For usercentric clustering, each user is cooperatively served by several of its nearby RRHs, which may eliminate cell-edge interference, provided that the near-instantaneous network conditions are known. However, this method may be unable to meet 5G's stringent delay requirement, because excessive time is required to estimate 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' historical data, such as their locations, the requested service, mobility pattern and speed, service demand profiles, and channel characteristics. By using machine-learning techniques, these data can be analyzed and beneficially exploited. Then one can predict a user's future locations, service request, and even 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 usercentric clusters. By contrast, the conventional cellular network is capable only 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. The access latency of the contents can thus be significantly reduced, alleviating the fronthaul traffic that constitutes C-RANs' bottleneck. The key question in cache-aided C-RANs is deciding 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 analyzing 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 for supporting the above low-delay techniques. In the following section, we introduce the EC theory for statistical delay-bounded QoS provision over C-RANs.

The Theory of the Statistical Delay-Bounded QoS

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

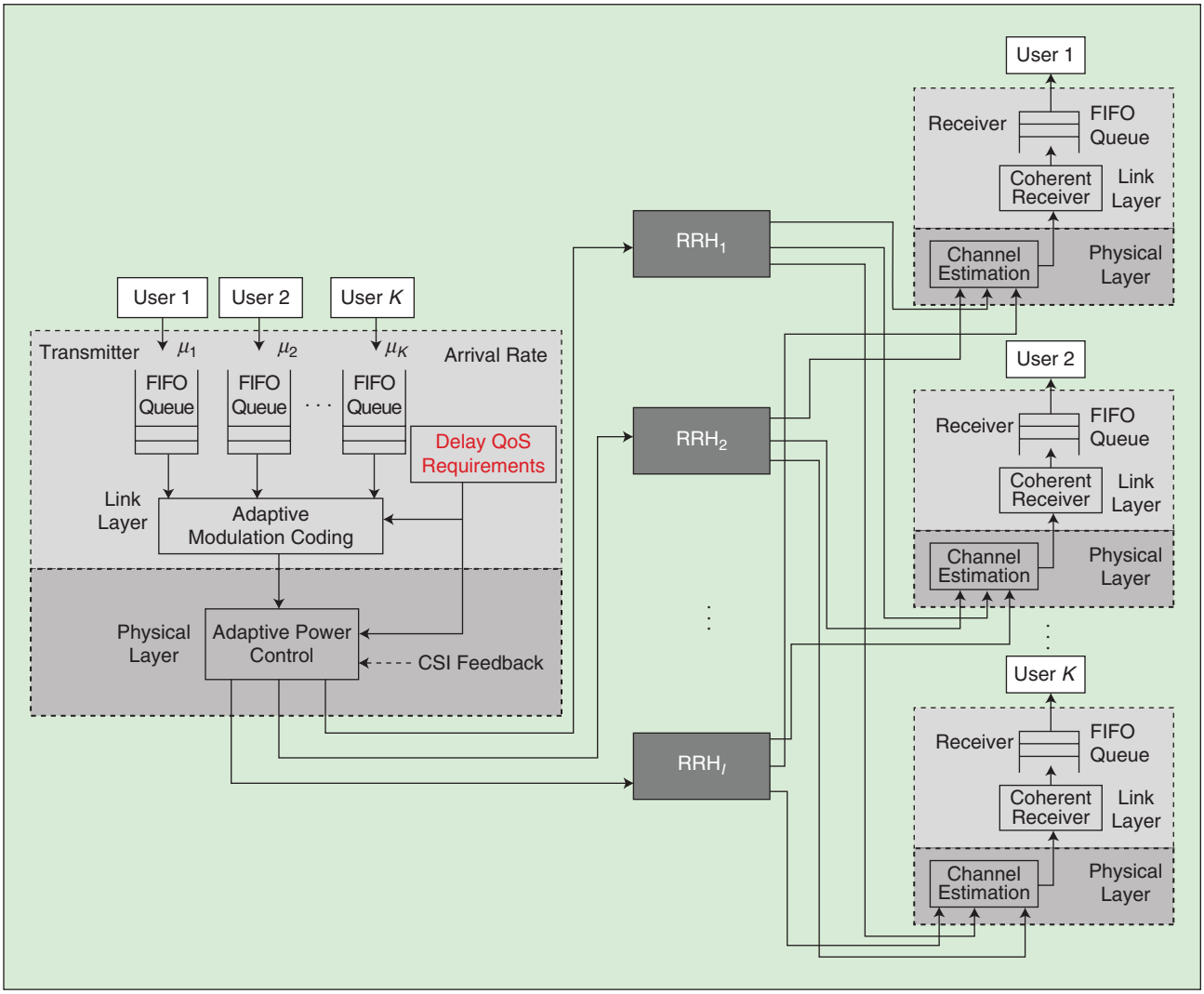


FIGURE 2 The statistical QoS provisioning over the 5G C-RAN.

to be equipped with a single antenna. We consider a block-fading channel, 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 $\theta_k, k = 1, \ldots, K$. For the C-RAN architecture of Figure 2, the buffer overflow probability of the kth user is approximated by $e^{\theta_k Q_{\text{th},k}}$, where θ_k and $Q_{\text{th},k}$ are the delay exponent and the maximum buffer length of user k, respectively. 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 [11] as

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

where ε is the probability that the buffer is not empty. In general, the delay-violation probability of $P_{\text{delay}}^{\text{out}}$ has to be extremely low for URLLC services.

EC proposed by Wu et al. [11] is defined as the maximum constant transmission frame arrival rate that the system can support while satisfying a maximum delay-outage probability constraint. The EC of user k is expressed [11] as

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

WE AIM TO DESIGN DELAY-BOUNDED STRATEGIES TO MAXIMIZE THE SUM OF THE EC OF ALL USERS UNDER THEIR PARTICULAR REQUIREMENTS.

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 the 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 $\mu_k = \text{EC}(\theta_k)$.

Among the studies on conventional wireless communication systems, most focus mainly on the ergodic capacity maximization problem, which ignores the delay requirement. By contrast, we aim to design delay-bounded strategies to maximize the sum of the EC of all users under their particular requirements. Specifically, we formulate the sum EC maximization problem under the following constraints:

- Each RRH has its individual average power constraint.
- 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 is imposed for guaranteeing that the instantaneous power remains within the linear range of practical power amplifiers.

Single-User Case

We first study the single-user case to glean initial insights. Because of the complex expression of the EC, most existing contributions have 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 devices. Hence, the conventional optimization method is no longer applicable, and the power allocation of each RRH will no longer be determined by a water-filling solution.

Therefore, 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 the delay requirements as well. 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 those with higher channel gains have higher priorities to transmit with full power.

We can also find the closed-form solution for the two extreme cases, i.e., 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 depends only on the channel conditions. For the infinity 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.

Multiuser Case

Because of the powerful computational capability of the BBU pool, the C-RAN will serve multiple users. However, the expression of EC 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 to avoid the multiuser interference. Additionally, we ignored the peak power constraints for simplicity. In this case, we were able to obtain the optimal power allocation solution for each user in closed form.

Performance Evaluations

We carried out 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 × 2 km. We adopted the Nakagami-m block-fading channel, subsuming the Rayleigh, Rician, and additive white Gaussian noise channels. The simulation results are based on the following parameters:

- a time frame of length $T_f = 0.04$ ms
- a system bandwidth of $B = 5$ MHz
- the average power constraint and peak power constraint of each RRH set to $P^{\text{avg}} = 0.5$ W and $P^{\text{peak}} = 1$ W, respectively
- the Nakagami fading parameter set to $m = 2$
- the path-loss model 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 kilometers
- the noise power density set as –174 dBm/Hz.

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] m and [–900, 946] m.

Figure 3 shows the delay-outage probability versus the delay exponent θ for our proposed power control algorithm. We tested three different values of the maximum delay threshold $D_{\max}$, i.e., $D_{\max} = 2, 1,$ and 0.5 ms. The rate of incoming data streams is set as $\mu = \text{EC}(\theta)$. As illustrated in Figure 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 URLLC's stringent delay requirement [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 EC:

- *Nearest RRH serving algorithm*: As the terminology suggests, this algorithm assigns the nearest RRH to serve the user. The technique 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.
- *Constant power allocation algorithm*: The transmit power of each RRH is set to its average power limit P^{avg}. This method is used for showing the benefits of dynamic power allocation in the face of different channel conditions.
- *Independent power allocation algorithm*: In this approach, each RRH independently optimizes its own transmission power based purely on its own channel conditions. This method is provided for demonstrating the merits of optimizing the power allocation according to the joint channel conditions.
- *Ergodic capacity maximization algorithm*: This algorithm maximizes the classic ergodic capacity for the user without incorporating the delay requirement.
- *Channel inversion algorithm*: Here, the power allocation of each RRH is proportional to the channel inversion. This algorithm supports a constant transmission data rate.

Figure 4 shows the normalized EC performance (which is the EC divided by B and T_f) for the different algorithms versus the delay exponent θ. As illustrated in Figure 4, the EC achieved by all of the algorithms (except the channel inversion algorithm) decreases with the delay exponent θ. Intuitively, a higher θ corresponds to a more stringent delay requirement and a lower delay-outage probability requirement. Then, the maximum arrival rate that can be supported should be reduced for satisfying the stringent delay requirements. We observe from this figure that our algorithm has a much better performance than the others, 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 EC is approximately equivalent to maximizing the ergodic capacity, leading to similar performance for these two algorithms. However, for high θ, the delay requirement is very strict, which has to be taken into consideration

THE DELAY EXPONENT CAN BE ADAPTIVELY SET TO SATISFY THE DIVERSE DELAY REQUIREMENTS.

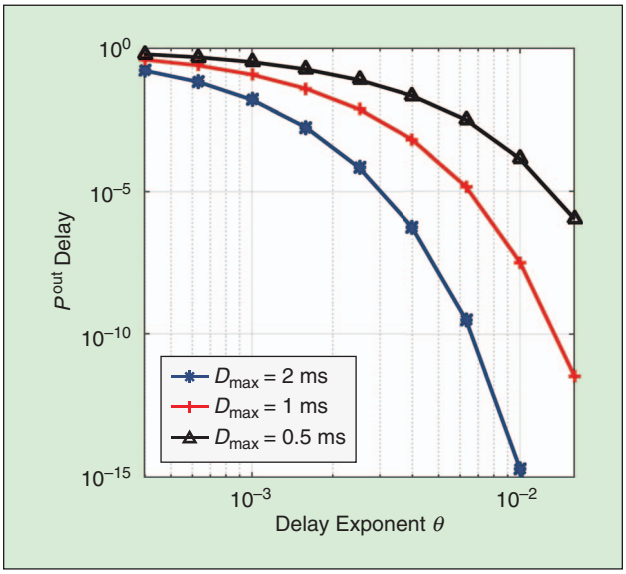


FIGURE 3 The delay-outage probability versus the delay exponent θ for various values of $D_{\max}$ for our proposed algorithm.

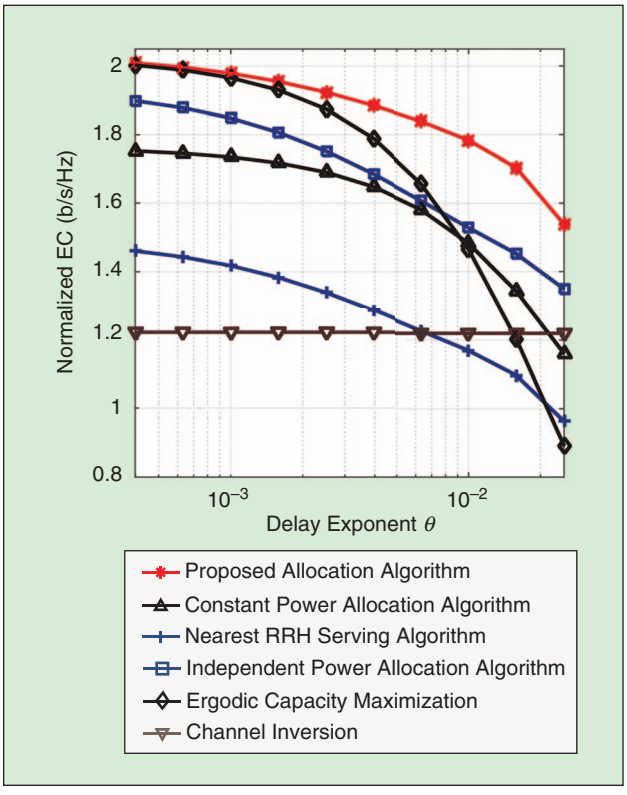


FIGURE 4 The normalized EC for various algorithms versus the delay exponent θ for a single user.

THE SIMULATION RESULTS SHOWED THAT, BY APPROPRIATELY CHOOSING THE DELAY EXPONENT θ, THE DELAY-OUTAGE PROBABILITY CAN BE REDUCED BELOW 10^{-9}.

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 provides 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 b/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 EC 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 a constant data rate for various channel conditions.

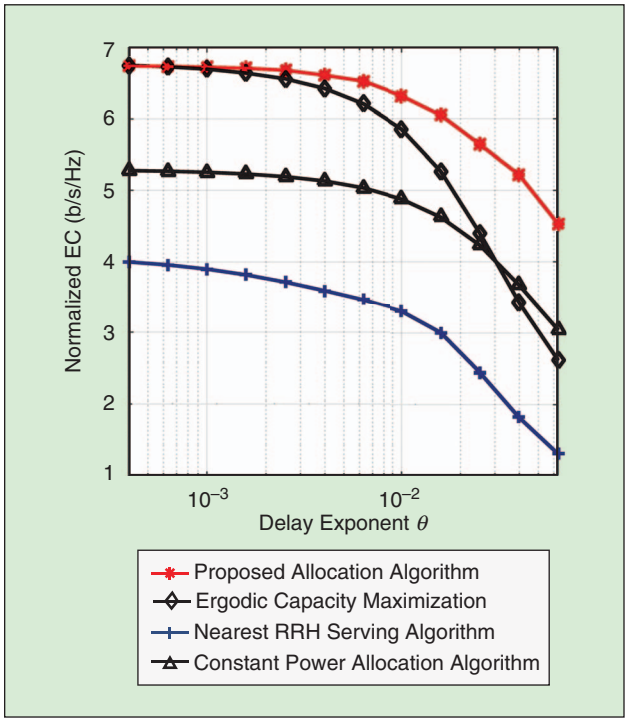


FIGURE 5 The sum normalized EC versus the delay exponent θ for our proposed algorithm and the ergodic capacity maximization algorithm, when supporting two users by four RRHs located at [650, 650], [−650,650], [−650, −650], and [650, −650].

Multiuser Case

Finally, in Figure 5, we consider the multiuser case, where there are two users having the coordinates given by [−100,0] and [0,100], respectively. We assume 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 EC performance. We observe a performance trend similar to that of the single-user scenario of Figure 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 compare our proposed algorithm with two others, i.e., the nearest RRH serving algorithm and the constant power allocation algorithm. For the former, each user is served by its nearest RRH, while for the latter algorithm, the instantaneous transmit power for each RRH is set to its average power limit P^{avg}, and each RRH assigns equal instantaneous transmit power to each user. We see from this figure that our proposed algorithm significantly outperforms these two others, achieving a performance gain over them of 2.8 b/s/Hz and 1.5 b/s/Hz, respectively—and the performance gain remains nearly fixed over all of the delay exponent θ. By exploiting the multiuser diversity, the normalized EC attained by the proposed algorithm for the two-user case is much larger than that of the single-user case.

Conclusions

We first highlighted the C-RAN architecture that consists of three components: the BBU pool, fronthaul links, and RRHs. The C-RAN architecture can rely on centralized signal processing techniques, such as CoMP transmission, joint user scheduling, and data-flow control. Additionally, the emerging techniques of NFV, SDN, and AI can be intrinsically integrated with the C-RAN architecture. Then we highlighted the EC 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 EC for both the single-user case and multiuser case that we 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 EC, especially in the case of stringent delay requirements.

However, substantial further research is required on delay-bounded C-RAN networks in the following areas:

- *Interference management*: In this article, 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 EC performance. This kind of optimization problem becomes nonconvex and hard to solve even for the simple Shannon capacity expression. The complex expression of the EC makes the optimization problem challenging to solve, which needs further investigation.

- *Limited fronthaul capacity*: Because of 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 (mm-wave) transmission, are promising in this scenario. However, the available bandwidth is much lower even at mm-wave frequencies than with fixed links. Hence, the limited fronthaul capacity should be taken into account when designing cross-layer operation.
- *Other delay sources*: This article considered only 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, nonnegligible 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 article, we focused only 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 article, we adopted Shannon's capacity for quantifying the instantaneous 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 article focuses on the EC maximization problem. However, energy efficiency, defined as the ratio of the data rate to total power consumption [15], is a key performance metric in 5G cellular networks. So, energy efficiency-oriented transmission design, considering the delay requirements, needs further study.

SOME MORE-ADVANCED USER SCHEDULING ALGORITHMS WITH LOW COMPLEXITY SHOULD ALSO BE DEVELOPED TO SATISFY THE STRINGENT DELAY REQUIREMENTS.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under grants 61571123, 61521061, and 61701198; the Nature Science Foundation of Jiangsu Province under grant BK20170557; and the Research Fund of the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China (grant 2018A03). Lajos Hanzo would like to acknowledge the financial support of the European Research Council, Advanced Fellow grant Beam-Me-Up.

Author Information

Hong Ren (renhong@seu.edu.cn) received her B.S. degree in electrical engineering from Southwest Jiaotong University, Chengdu, China, in 2011, and her M.S. and Ph.D. degrees in electrical engineering from Southeast University, Nanjing, China, in 2014 and 2018, respectively. She is currently a postdoctoral scholar at the School of Electronic Engineering and Computer Science, Queen Mary University of London. Her research interests lie in the areas of communication and signal processing, including green communication systems, cooperative transmission, and cross-layer transmission optimization. She is a Student Member of the IEEE.

Nan Liu (nanliu@seu.edu.cn) received her B.Eng. degree in electrical engineering from Beijing University of Posts and Telecommunications in 2001 and her Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, in 2007. In 2009, she became a professor at the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing, China. Her research interests are in network information theory for wireless networks, self-organizing network algorithms for next-generation cellular networks, and energy-efficient communications. She is a Member of the IEEE.

Cunhua Pan (c.pan@qmul.ac.uk) received his B.S. and Ph.D. degrees in wireless communications from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2010 and 2015, respectively. He is currently a postdoctoral researcher with Queen Mary

University of London. His research interests mainly include ultradense cloud radio access networking, unmanned aerial vehicles, the Internet of Things, nonorthogonal multiple access, and mobile edge computing. He serves as the Student Travel Grant chair for the IEEE International Conference on Communications (ICC) 2019 and has served as a Technical Program Committee member for many conferences, such as the ICC and the IEEE Global Communications Conference. He is a Member of the IEEE.

Maged Elkashlan (maged.elkashlan@qmul.ac.uk) received his Ph.D. degree in electrical engineering from the University of British Columbia, Vancouver, Canada, 2006. In 2011, he joined the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests fall into the broad areas of communication theory and statistical signal processing. He currently serves as editor of *IEEE Transactions on Wireless Communications* and *IEEE Transactions on Vehicular Technology*. He received the Best Paper Awards at the IEEE International Conference on Communications in 2014 and 2016, the International Conference on Communications and Networking in China in 2014, and the IEEE Vehicular Technology Conference in 2013. He is a Member of the IEEE.

Arumugam Nallanathan (a.nallanathan@qmul.ac.uk) received his B.Sc. degree (honors) from the University of Peradeniya, Sri Lanka, in 1991, his C.P.G.S. degree from the University of Cambridge, United Kingdom, in 1994, and his Ph.D. degree from the University of Hong Kong, China, in 2000, all in electrical engineering. He has been a professor of wireless communications and head of the Communication Systems Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017. His research interests include fifth-generation wireless systems, the Internet of Things, and molecular communications. He was a corecipient of the Best Paper Award presented at the IEEE Global Communications Conference 2017 and IEEE International Conference on Communications 2016. He is an editor of *IEEE Transactions on Communications*. He was selected as a Web of Science (Institute for Scientific Information) Highly Cited Researcher in 2016. He is an IEEE Distinguished Lecturer and a Fellow of the IEEE.

Xiaohu You (xhyu@seu.edu.cn) received his master's and Ph.D. degrees in electrical engineering from Southeast University, Nanjing, China, in 1985 and 1988, respectively. Since 1990, he has been working with National Mobile Communications Research Laboratory at Southeast University, where he has held the rank of director and professor. He has contributed over 100 IEEE journal papers and two books in the areas of adaptive signal processing and neural networks and their applications to communication systems. Since 2013, he has been the principal investigator for the China National 863 5G Project. He is secretary general of the FuTURE Forum, vice chair of the China 2020 International Mobile Telecommunications Promotion Group, and vice chair of the China National Mega Project on New Generation Mobile Network. He was the recipient of the National First Class Invention Prize in 2011. He is a Fellow of the IEEE.

Lajos Hanzo (lh@ecs.soton.ac.uk) received his M.S. degree in electronics in 1976, his Ph.D. degree in 1983, and his D.Sc. degree in 2004. Since 1986, he has been with the School of Electronics and Computer Science, University of Southampton, United Kingdom, where he is the chair in telecommunications. He has successfully supervised 111 Ph.D. students. He has coauthored 18 John Wiley/IEEE Press books on mobile radio communications and published more than 1,700 research contributions in IEEE *Xplore*. In 2009, he received an honorary doctorate from the Technical University of Budapest and in 2015 from the University of Edinburgh. He has more than 33,000 citations and an h-index of 73. He is a Fellow of the IEEE, the Royal Academy of Engineering, the Institution of Engineering and Technology, and the European Association for Signal Processing.

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 J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 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," *Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 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 J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[5] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 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 Veh. Technol. Mag.*, vol. 12, no. 2, pp. 26–35, June 2017.

[7] "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project, Sophia Antipolis, France, Tech. Rep. 36.814, 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 Veh. Technol. Mag.*, 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 Commun. Mag.*, 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," *Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 743–764, 2017.

[11] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, 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 J. Sel. Areas Commun.*, 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 Commun. Mag.*, vol. 55, no. 6, pp. 72–78, 2017.

[15] F. Zhou, Y. Wu, Q. Hu, Y. Wang, and K.-K. Wong, "Energy-efficient NOMA enabled heterogeneous cloud radio access networks," arXiv preprint, arXiv:1801.01996v1 [cs.NI], 2018.

VT