Heterogeneous Statistical QoS Provisioning Over 5G Mobile Wireless Networks

Xi Zhang, Wenchi Cheng, and Hailin Zhang

Abstract

As a critical step towards the next new era of mobile wireless networks, recently 5G mobile wireless networks have received significant research attention and efforts from both academia and industry. The 5G mobile wireless networks are expected to provide different delay-bounded QoS guarantees for a wide spectrum of services, applications, and users with extremely diverse requirements. Since the time-sensitive services in 5G multimedia wireless networks may vary dramatically in both a large range from milliseconds to a few seconds and diversity from uniform/constant delay-bound to different/variable delay-bound guarantees among different wireless links, the delay-bound QoS requirements for different types of services promote the newly emerging heterogeneous statistical delay-bounded QoS provisioning over 5G mobile wireless networks, which, however, imposes many new challenging issues not encountered before in 4G wireless networks. To overcome these new challenges, in this article we propose a novel heterogeneous statistical QoS provisioning architecture for 5G mobile wireless networks. First, we develop and analyze the new heterogeneous statistical QoS system model by applying and extending the effective capacity theory. Then, through the wireless coupling channels, we apply our proposed heterogeneous statistical QoS architecture to efficiently implement the following powerful 5G-candidate wireless techniques: 1) device-to-device networks; 2) full-duplex networks; and 3) cognitive radio networks, respectively, for providing heterogeneous statistical delay-bounded QoS guarantees. Finally, using the simulation experiments we show that our proposed architecture and schemes significantly outperform the existing traditional statistical delay-bounded QoS provisioning schemes in terms of satisfying the heterogeneous delay-bounded QoS requirements while maximizing the aggregate system throughput over 5G mobile wireless networks.

As the fourth-generation (4G) wireless communications and networks are becoming more and more mature and widely implemented in the mobile wireless industrial and commercial products, the fifth-generation (5G) mobile wireless communication technologies are rapidly emerging in research fields. The emerging 5G wireless networks aims at ensuring that various contemporary wireless applications can be timely and satisfactorily served at any time and any place, and in any way [1–3]. One of the most important services in 5G wireless applications is the bandwidth-intentive and time-sensitive multimedia, even including 3D immersive media, transmissions that were previously confined to wired networks such as the Internet, but are now making forays into mobile devices and wireless networks. Video traffic constitutes a significant 51 percent of the mobile traffic volume and is expected to increase to 67 percent by 2017 [4, 5]. Full high definition (FHD) video is also being increasingly shared through popular media such as YouTube, and ultra high definition (UHD) and 3D video content will eventually take over in the not so distant future.

To support these highly bandwidth-intentive and time-sensitive multimedia services for the emerging 5G wireless networks, for the last several years the telecommunications academia and industry have made a great deal of effort/processes in investigating various advanced wireless techniques such as device-to-device (D2D) communications [6], wireless full-duplex (FD) communications [7, 8], advanced cognitive radio (CR) communications [8], and so on, as well as quality of service (QoS) provisioning techniques. The key design issue for multimedia wireless services is how to efficiently guarantee timely multimedia data transmissions within specified delay bounds. Because of the highly varying wireless channels, the deterministic delay-bounded QoS requirements for high-volume multimedia wireless traffic are usually hard to guarantee. Alternatively, the statistical delay-boundeded QoS provision-

Xi Zhang is with Texas A&M University.

Wenchi Cheng and Hailin Zhang are with Xidian University.

This work was supported in part by the U.S. National Science Foundation under Grants ECCS-1408601 and CNS-1205726, the U.S. National Science Foundation CAREER Award under Grant ECCS-0348694, the National Natural Science Foundation of China (No. 61401330 & No. 61371127 & No. 62993610), the 111 Project of China (B08038), the Important National Science and Technology Specific Projects (2012ZX03003012), the Fundamental Research Funds for the Central Universities (No. 72136037 & No. 7214603701), and the Program for New Century Excellent Talents in University (NCET-12-0918).

0890-8044/14/$25.00 © 2014 IEEE

ing theory has been proposed and shown to be a powerful technique to characterize and implement the delay-bounded QoS guarantee for wireless real-time traffic [9, 10]. The traditional statistical delay-bounded QoS guarantee technique, called homogeneous statistical delay-bounded QoS provisioning, typically assumes that the QoS provisioning for each link can be individually processed.

However, examining the new existing and expected wireless communication techniques for 5G wireless networks, we observe that more than one type of traffic corresponding to different delay-bounded QoS constraints typically coexist in the 5G mobile wireless networks. Therefore, 5G mobile wireless networks need to provide different delay-bounded QoS guarantees for diverse types of services, applications, and users with extremely diverging requirements in different 5G candidate-technique-based networks. Since the time sensitivities of different sets of services vary from 1 ms to a few seconds across different wireless links for different mobile users, the delay-bounded QoS guarantees for different types of services in 5G mobile wireless networks demand new heterogeneous statistical delay-bounded QoS provisioning architectures, frameworks, schemes, and algorithms, thus imposing many new design problems not encountered before in 4G wireless networks.

To overcome the above-mentioned challenges, in this article we propose the novel heterogeneous statistical QoS provisioning architecture over 5G mobile wireless networks. We identify and analyze three promising 5G candidate-techniques-based networks:

- D2D technique-based networks, where the D2D communication and the cellular communication are performed at the same time
- FD technique-based networks, where two wireless FD terminals send their own data to each other using the *same* frequency-time channel simultaneously
- CR technique-based networks where the secondary transmitter sends its own data to the secondary receiver when the primary/licensed transmitter and receiver are not using the licensed channel

We build up our 5G-candidate techniques-based heterogeneous QoS provisioning system model and characterize the common features of having two coexisting wireless communications pairs over the three 5G technique-based wireless networks by employing the wireless coupling channels communications model. Using these system models, we develop the heterogeneous statistical delay-bounded QoS provisioning architecture for the generic 5G candidate-techniques-based networks and analyze the heterogeneous statistical delay-bounded QoS provisioning for the D2D technique-based networks, the FD technique-based networks, and the CR technique-based networks, respectively. We also develop the heterogeneous QoS-based power allocation schemes to maximize the aggregate effective capacity of the 5G-candidate technique-based networks, which are evaluated by the simulation results.

The rest of this article is organized as follows. We establish the system model for the heterogeneous statistical delay-bounded QoS provisioning architecture over 5G wireless networks, which consists of the cloud radio access network (Cloud-RAN) infrastructure [11], the three 5G candidate wireless-techniques-based networks, and the wireless coupling channels model. We extend the effective-capacity-based homogeneous statistical delay-bounded QoS provisioning existing in 4G wireless networks into the heterogeneous statistical delay-bounded QoS provisioning for 5G wireless networks. We apply our proposed architecture to implement the heterogeneous statistical delay-bounded QoS provisioning for D2D-5G networks, FD-5G networks, and CR-5G networks, respectively. We simulate and evaluate our proposed heterogeneous QoS architecture-based power allocations schemes in terms of the aggregate effective capacities for the three types of 5G candidate-techniques-based wireless networks. We then conclude this article.

The System Model for Heterogeneous QoS Provisioning Architecture Over 5G Wireless Networks

Figure 1 depictures the architecture model for the 5G mobile wireless networks, which is applied and integrated with the following three promising 5G-candidate wireless communications techniques:

- D2D wireless communication technique
- FD wireless communication technique
- CR wireless communication technique

As shown in Fig. 1, the 5G wireless network consists of a number of 5G candidate-techniques-based networks, each of which is implemented with one type of 5G candidate technique specified by the performance requirements such as throughput, delay bound, and spectrum/energy efficiency. All 5G candidate-techniques-based networks are connected to the *big data center/server* facilitated by the *super base statio*n system with distributed massive multiple-input multiple-output (MIMO) of the Cloud-RAN infrastructure [1]. Also, as illustrated in Fig. 1, the three 5G candidate- techniques-based networks are implemented by the three highly demanded and promising 5G wireless communication techniques, including the D2D wireless communication, wireless FD communication, and CR wireless communication techniques, respectively. We call 5G wireless networks using the D2D wireless communication [2], FD wireless communication technique [1, 7], and CR wireless communication technique [8], respectively, D2D-5G networks, FD-5G networks, and CR-5G networks, respectively, which are further elaborated on in the following.

•**D2D-5G networks:** As depicted in Fig. 1, D2D communications enables the exchange of data directly between two mobile users without the use of a base station (BS) or the core network other than for assistance of setting up direct connections. D2D communications is beneficial in increasing area spectrum efficiency and cellular coverage while decreasing end-to-end delay, cellular interference, and power consumption. Thus, D2D is a very promising candidate technique for 5G wireless networks. In D2D-based networks, D2D devices implement the direct half-duplex communication (D2D communication) with each other, bypassing the BS, while cellular devices need to communicate through the BS. A pair of D2D communication parties forms a D2D pair, where the D2D communications is conducted. In contrast, the cellular communication goes through the BS. Under our proposed D2D and cellular communication model as shown in Fig. 1, D2D-5G consist of a D2D pair, a cellular device, and a BS.

•**FD-5G networks:** As illustrated in Fig. 1, an FD BS transmits and receives from different mobile users simultaneously using the same frequency channel. Thus, FD communication can potentially double the spectrum efficiency of wireless networks while improving the energy efficiency of wireless communications, making FD another very strong candidate technique for 5G wireless networks. In FD-based wireless networks, the mobile device performs the wireless FD communication with the BS, which enables transmitting and receiving simultaneously by using the same frequency-time channel and thus can significantly increase the spectrum efficiency of the

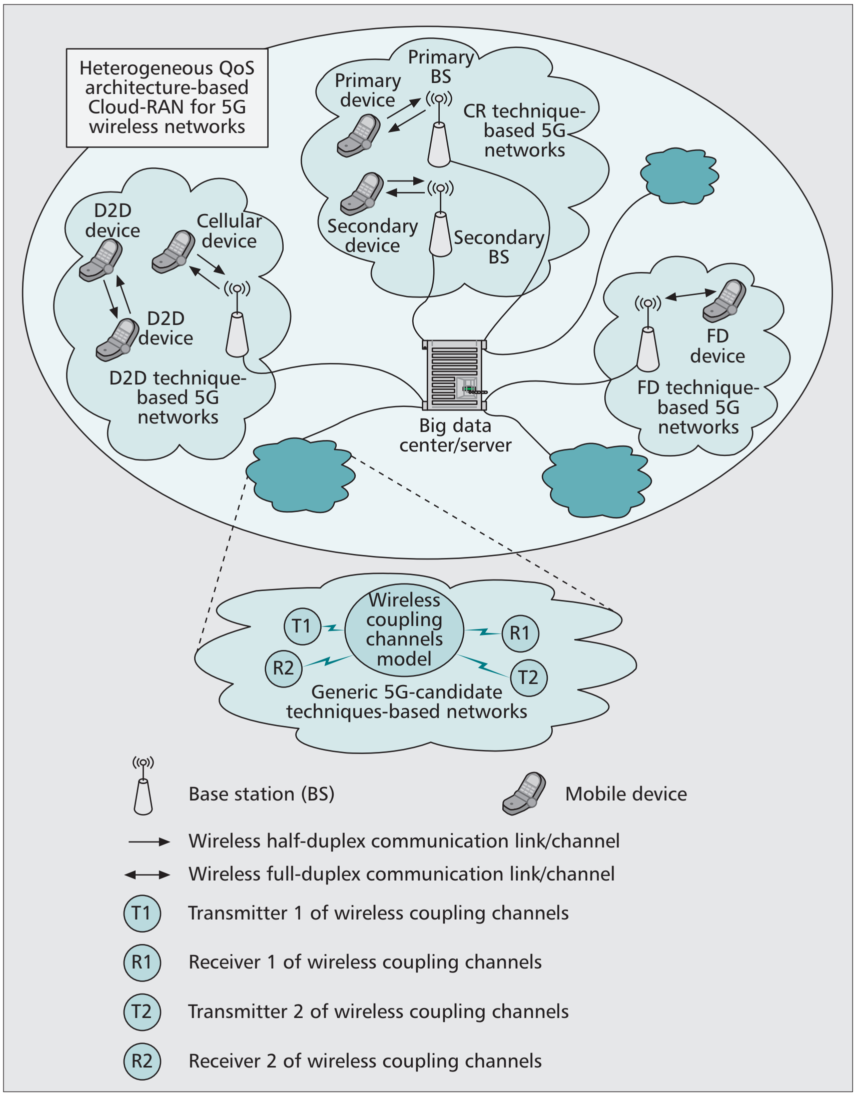
Heterogeneous QoS architecture-based Cloud-RAN for 5G wireless networks
Primary BS
Primary device
CR technique-based 5G networks
Secondary device
Secondary BS
D2D device
Cellular device
D2D device
D2D technique-based 5G networks
FD device
FD technique-based 5G networks
Big data center/server
Wireless coupling channels model
T1
R1
R2
T2
Generic 5G-candidate techniques-based networks
Base station (BS)
Mobile device
Wireless half-duplex communication link/channel
Wireless full-duplex communication link/channel
T1 Transmitter 1 of wireless coupling channels
R1 Receiver 1 of wireless coupling channels
T2 Transmitter 2 of wireless coupling channels
R2 Receiver 2 of wireless coupling channels

Figure 1. The heterogenous statistical QoS system architecture model for 5G mobile wireless networks supported by the Cloud-RAN infrastructure, which is centered with the big data center/server facilitated by the super *base station* (BS) systems equipped with distributed massive MIMO.

wireless networks. The key technique to successfully implementing wireless FD communication is how to cancel/mitigate the self-interference, which is the interference at the receiver caused by the local transmitter.

•**CR-5G networks:** As shown in Fig. 1, CR communications allows secondary mobile users to share spectrum bands with licensed users in either an underlay (interference-tolerant) or overlay (interference-free/minimization) basis with the primary users receiving higher priority. CR can also significantly increase spectrum efficiency and energy efficiency, thus making CR a very powerful 5G candidate technique. Within CR-based networks, primary communication is defined as the communication between the primary device and the primary BS. Secondary communication is defined as the communication between the secondary device and the secondary BS. The primary device has higher priority to implement the primary communication with the primary BS than the secondary device because the channel is licensed to the primary device and primary BS. To increase the spectrum efficiency of CR-based networks, the secondary device communicates with the secondary BS using the channel licensed to the primary devices and the BS while keeping the interference to the primary communication below the tolerable interference noise floor/threshold.

Having observed the common feature of having the two coexisting wireless communications pairs over the three types of 5G wireless networks as discussed in the above, we can characterize this feature by defining the *wireless coupling channels model*, which consists of two transmitters (T1 and T2) and two receivers (R1 and R2) sharing the same frequency band and time slot, as shown in Fig. 1. Typically, the delay-bounded QoS requirement for the communication channel from T1 to R1 differs from the other different delay-bounded QoS requirements for the channel from T2 to R2, forming the generic heterogeneous delay-bounded QoS-based wireless coupling channels model, which can be applied into the three 5G technique-based networks as follows. For D2D-5G networks, a T1 and R1 communication pair denotes the D2D communication pair while T2 and R2 denote the other coexisting communication pair between the cellular device and the BS, respectively, with different delay-bounded QoS constraints for the two communication pairs. For FD-5G, T1 and R2 denote the communication pair between the transmitter and the receiver, respectively, of the FD device, while T2 and R1 denote the other coexisting communication pair between the transmitter and the receiver, respectively, of the BS, where the delay bound QoS requirements are different between the two communication pairs. For CR-5G networks, T1 and R1 denote the communication pair between the primary device and the primary BS, respectively, while T2 and R2 denote the other communication pair between the secondary device and the secondary BS, respectively, with two different delay-bounded QoS constraints for the primary and secondary communications pairs.

For one wireless coupling channel of the D2D-5G networks and the CR-5G networks, there are four mobile devices (T1 and R2 are two different mobile devices, while T2 and R1 are also two different mobile devices). However, for one wireless coupling channel of the FD-5G networks, there are only two mobile devices (T1 and R2 are the same device, while T2 and R1 are also the same device).

The Effective Capacity Theory for Heterogeneous Statistical Delay-Bounded QoS Provisioning Over 5G Mobile Wireless Networks

Using the system model for heterogeneous QoS provisioning over 5G wireless networks, we develop the effective capacity theory for heterogenous delay-bounded QoS provisioning over 5G mobile wireless networks.

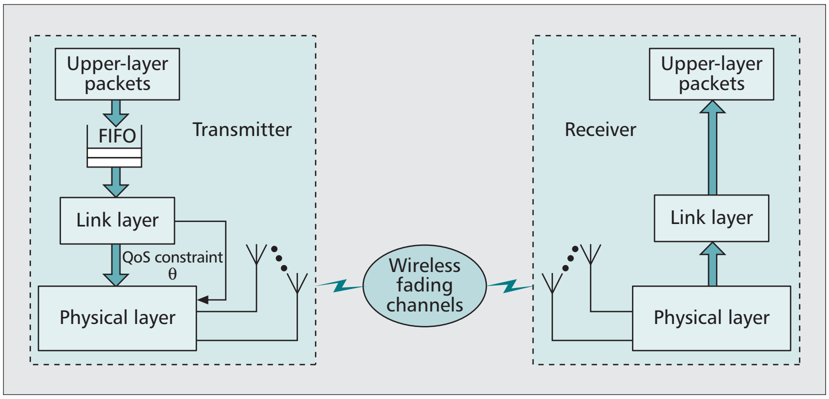

Figure 2. The homogeneous statistical QoS provisioning over the traditional wireless channel in 4G wireless networks.

Statistical Delay-Bounded QoS Provisioning over the Traditional 4G Wireless Networks

We introduce the fundamentals of homogeneous statistical delay-bounded QoS provisioning for 4G wireless networks, which is based on the *effective capacity theory* [9, 10, 12]. Inspired by the principle of effective bandwidth, the authors in [9] defined the effective capacity as the maximum constant arrival rate that can be supported by the service rate to guarantee the specified statistical delay-bounded requirement over a wireless channel. The specified statistical delay-bounded requirement, called *QoS exponent* and denoted by θ, is a positive real-valued number that builds up the relationship between the queue length threshold and the probability of queue length exceeding the given threshold. Figure 2 shows the homogeneous statistical delay-bounded QoS provisioning framework over the traditional wireless link. The upper-layer packets are first buffered in first-in first-out (FIFO) queues to be transmitted to their destinations. Then, at the link layer, the packets are divided into frames and then split into bitstreams at the physical (PHY) layer. Based on the service-determined QoS exponent θ corresponding to the real-time traffic of this link and the channel state information (CSI) fed back from the corresponding receiver, we need to develop the homogeneous delay-bounded QoS-driven strategies to optimize the system performance under the given QoS exponent θ. For the traditional wireless link (Fig. 2), the probability of queue length exceeding the given threshold is given by $e^{-\theta Q_{th}}$, where Q_{th} is the queue length threshold. The QoS exponent measures the exponential decay rate of the delay-bounded QoS violation probabilities. A larger θ corresponds to a faster decay rate, which implies that the system can provide a more stringent QoS requirement. A smaller θ leads to a slower decay rate, which indicates a looser QoS requirement. Asymptotically, when θ → ∞, it is implied that the system cannot tolerate any delay, which corresponds to the very stringent statistical delay-bounded QoS constraint. On the other hand, when θ → ∞, the system can tolerate an arbitrarily long delay, which corresponds to the very loose statistical delay-bounded QoS constraint. Then we can derive the analytical expression of the *effective capacity*, denoted by C, as follows [10, 12]:

$$C = -\frac{1}{\theta}\log\left(\mathbb{E}\left\{e^{-\theta TR}\right\}\right) \qquad (1)$$

where R is the instantaneous transmission rate of one time frame, T is the fixed length of each time frame, and $\mathbb{E}\{\cdot\}$ denotes the expectation.

Although homogeneous statistical delay-bounded QoS provisioning can guarantee the QoS requirements of the 4G wireless networks, it has inherent deficiencies to be implemented in 5G wireless networks. Because more than one type of traffic corresponding to different delay-bounded QoS constraints typically coexist in the 5G mobile wireless networks, 5G mobile wireless networks need to provide different delay-bounded QoS guarantees for diverse types of services, applications, and users with extremely divergent requirements in different 5G candidate-techniques-based networks. However, homogeneous statistical delay-bounded QoS provisioning can only guarantee one type of traffic, which violates the diverse delay-bounded QoS requirements of 5G wireless networks.

Heterogeneous Statistical Delay-Bounded QoS Provisioning through the Wireless Coupling Channels Model in 5G Wireless Networks

Unlike 4G wireless networks, where the homogeneous statistical delay-bounded QoS provisioning can guarantee the QoS requirements of the wireless networks, 5G wireless networks need more sophisticated statistical delay-bounded QoS provisioning strategy because different links have different delay-bounded QoS requirements, and thus, the diverse statistical delay-bounded QoS provisioning requirements need to be jointly considered for *all links* over 5G candidate-techniques-based networks. We call this type of wireless traffic delay-bounded QoS requirements *heterogeneous* statistical delay-bounded QoS provisioning.

Figure 3 shows an example of the heterogeneous statistical delay-bounded QoS provisioning framework over the wireless coupling channel for the generic 5G candidate-techniques-based networks. As illustrated in Fig. 3, transmitter 1 (T1) and transmitter 2 (T2) send their signals to receiver 1 (R1) and receiver 2 (R2), respectively, using the same/shared frequency-time channel. The QoS constraints θ_1 and θ_2, corresponding to T1 and T2, respectively, need to be jointly guaranteed at the same time. Because θ_1 and θ_2 typically have different values, we need to jointly provide the heterogeneous statistical delay-bounded QoS provisioning for the links from T1 to R1 and from T2 to R2 simultaneously. The upper-layer packets processes follow the same procedure as in the homogeneous statistical delay-bounded QoS provisioning framework. Based on the service-determined QoS constraints (θ_i, i = 1, 2) corresponding to the real-time multimedia traffic of two channels and CSI fed back from the receivers to the corresponding transmitters, we need to develop heterogeneous delay-bounded QoS-driven strategies to optimize system performance in terms of effective capacity under the given heterogeneous QoS exponents θ_1 and θ_2.

To quantitatively characterize the heterogeneous statistical delay-bounded QoS provisioning, we introduce and define the *aggregate effective capacity*, denoted by $C_a(\theta_1, \theta_2, R_1, R_2)$, as the sum of the effective capacities corresponding to the links over the wireless coupling channels from T1 to R1 and from T2 to R2, where R_1 and R_2 denote the instantaneous transmission rates corresponding to the wireless coupling channels from T1 to R1 and from T2 to R2, respectively. Then, extending the derivation for Eq. 1, we can derive the generic analytical function for the aggregate effective capacity $C_a(\theta_1, \theta_2, R_1, R_2)$ over the wireless coupling channel as follows:

$$C_a(\theta_1,\theta_2,R_1,R_2) = -\frac{1}{\theta_1}\log\left(\mathbb{E}\left\{e^{-\theta_1 TR_1}\right\}\right) - \frac{1}{\theta_2}\log\left(\mathbb{E}\left\{e^{-\theta_2 TR_2}\right\}\right) \qquad (2)$$

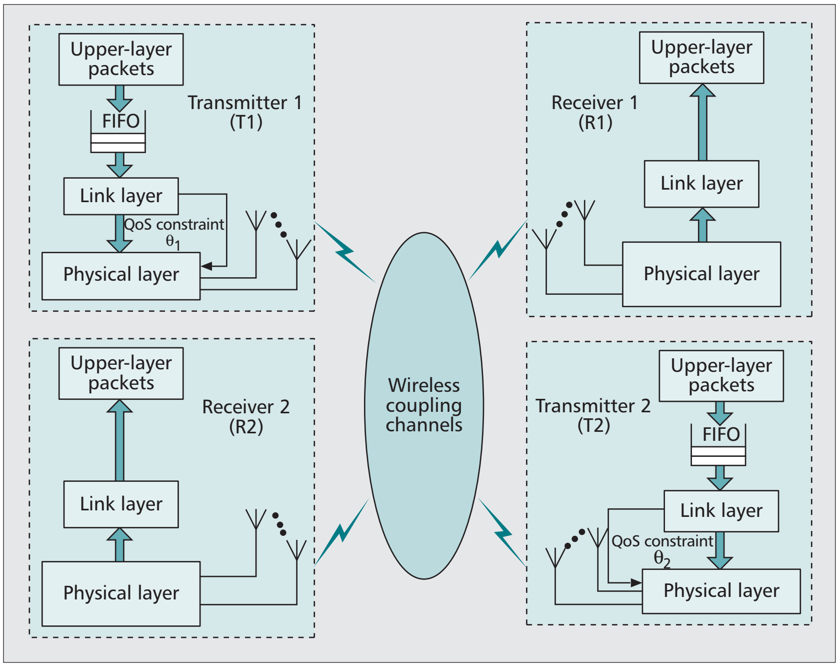

Figure 3. The heterogeneous statistical QoS provisioning over the "wireless coupling channels model" in all 5G candidate-technique-based networks.

For heterogeneous statistical delay-bounded QoS provisioning over 5G wireless networks, our objective is to maximize the aggregate effective capacity of each type of 5G candidate-technique-based networks. However, the heterogeneous statistical delay-bounded QoS provisioning imposes the new challenges that the resource allocation for each link depends not only on its own corresponding delay-bounded QoS requirement, but also on the delay-bounded QoS requirements corresponding to all the other links in the 5G wireless networks.

If the delay-bounded QoS provisioning for each link can be individually processed, this type of wireless network scenario is suitable to be characterized by the homogeneous statistical delay-bounded QoS provisioning scheme. If more than one type of traffic corresponding to different delay-bounded QoS constraints typically coexist for each link, this type of wireless network scenario can be better characterized by the heterogeneous statistical delay-bounded QoS provisioning scheme.

Using derivations similar to those used for the homogeneous QoS-driven power allocations over the corresponding traditional wireless networks [10], we can derive the statistical delay-bounded QoS-based power allocation strategies for the three types of 5G candidate-technique-based networks, called the heterogeneous QoS-based power allocation schemes, which can efficiently maximize the aggregate effective capacities for the three types of 5G candidate-technique-based networks, respectively, as detailed in the next section.

Applications of Heterogeneous Statistical Delay-Bounded QoS Provisioning Architecture for Implementing Three 5G Candidate-Techniques-Based Wireless Networks

In this section, we show how our proposed heterogeneous statistical delay-bounded QoS provisioning architecture can support the implementations for the following three powerful 5G candidate wireless techniques:

- D2D-based networks [13]
- FD-based networks [14]
- CR-based networks [15]

The D2D-5G, FD-5G, and CR-5G networks can coexist in 5G mobile wireless networks. The D2D, FD, and CR techniques can be jointly implemented for 5G wireless networks. In our model in this article, these three networks are optimized separately to simply illustrate the framework of heterogenous delay-bounded QoS provisioning. However, these networks can be jointly optimized for further performance improvement.

Heterogeneous Statistical Delay-Bounded QoS Provisioning for D2D-Based 5G Mobile Wireless Networks

For D2D-5G networks, the delay-bounded QoS constraints (θ_i, $i = 1, 2$) for cellular communication and D2D communication typically differ from each other. Although two D2D devices communicate with each other bypassing the BS, both of them are still controlled by the BS. There are two fundamental modes for D2D-5G networks: the co-channel and orthogonal-channel modes. The co-channel mode is defined as that where the D2D and cellular communication use the entire frequency-time resource of the D2D-5G networks. On the other hand, the orthogonal-channel mode is defined as that where D2D communication uses one part of the entire frequency-time resource, and cellular communication uses the other part, which is orthogonal to the first part of the frequency-time resource in D2D-5G networks.

Regardless of whether D2D-5G networks uses the co-channel or orthogonal-channel mode, the delay-bounded QoS requirements θ_1 and θ_2 for D2D communication and cellular communication are typically different from each other and also need to be jointly guaranteed at the same time, which defines the *heterogeneous statistical delay-bounded QoS provisioning framework for D2D-5G networks*. Then the aggregate effective capacity of D2D-5G networks is the sum of the effective capacities for D2D communication and that for cellular communication over wireless coupling channels. Thus, we can derive the aggregate effective capacities for D2D-based 5G wireless networks under the co-channel and orthogonal-channel modes as $C_a(\theta_1, \theta_2, R_{d1}, R_{d2})$ and $C_a(\theta_1, \theta_2, \alpha R_{d1}, (1 - \alpha)R_{d2})$, respectively, where the generic aggregate effective capacity function $C_a(\cdot, \cdot, \cdot, \cdot)$ is specified by Eq. 2, R_{d1} and R_{d2} denote the instantaneous transmission rates of the D2D and cellular communication, respectively, for D2D-based 5G wireless networks, and α and $(1 - \alpha)$ represent the percents of time lengths of one time frame used for D2D and cellular communication, respectively, when D2D-5G networks employ the orthogonal-channel mode. For D2D-5G networks under heterogeneous delay-bounded QoS constraints, the design objective is to maximize the aggregate effective capacities for the co-channel mode and orthogonal-channel mode, respectively. Clearly, the aggregate effective capacities jointly depend on both the QoS exponents θ_1 and θ_2 for the D2D and cellular communication. Thus, we can derive the optimal power allocation strategies to maximize the aggregate effective capacities by jointly taking into account θ_1 and θ_2 over the wireless coupling channels at the same time in D2D-5G networks.

Heterogeneous Statistical Delay-Bounded QoS Provisioning for FD-Based 5G Mobile Wireless Networks

For FD 5G networks, the FD device sends its own data to the BS while the BS transmits its own data to the FD device using the same frequency-time resource over the wireless coupling channels simultaneously. Because both the FD device and the BS use the wireless FD transmission mode over the wireless coupling channels, the different delay-bounded QoS constraints (θ_1 and θ_2) for the two channels from the FD device to the BS and from the BS to the FD device (along the wireless coupling channels) need to be jointly guaranteed at the same time, which defines the *heterogeneous statistical delay-bounded QoS provisioning framework for FD 5G networks*. The aggregate effective capacity of FD 5G networks is the sum of the effective capacities that correspond to the channel from the FD device to the BS and the channel from the BS to the FD device, respectively, over the wireless coupling channels. Thus, we can derive the aggregate effective capacity of FD 5G networks as $C_a(\theta_1, \theta_2, R_{f1}, R_{f2})$, where the generic aggregate effective capacity function $C_a(\cdot, \cdot, \cdot, \cdot)$ is specified by Eq. 2, and R_{f1} and R_{f2} denote the instantaneous wireless FD transmission rates over the wireless coupling channels from the FD device to the BS and the BS to the FD device, respectively. For FD 5G networks, under the heterogeneous delay-bounded QoS constraints (θ_1 and θ_2), our objective is to maximize the aggregate effective capacity, which jointly depends on both the QoS exponents θ_1 corresponding to the channel from the FD device to the BS and θ_2 corresponding to the channel from the BS to the FD device. We can further derive the optimal power allocation strategies to maximize the aggregate effective capacity under the two QoS exponents (θ_1 and θ_2) for FD-based 5G wireless networks.

The self-interference severely impacts the statistical delay-bounded QoS guarantees. First, the self-interference decreases the received SNR of the mobile devices, thus decreasing the effective capacity of FD 5G wireless networks. Second, the self-interference increases the error probability for channel estimations, thus also decreasing the effective capacity of FD 5G wireless networks.

Heterogeneous Statistical Delay-Bounded QoS Provisioning for CR-Based 5G Mobile Wireless Networks

For CR-based 5G wireless networks, we mainly focus on the interference-tolerant mode of CR wireless networks, where the communications between the primary device and primary BS and between the secondary device and secondary BS are concurrent over the wireless coupling channels as long as the interference caused by the secondary communication is below the given interference-noise floor/threshold. Clearly, the delay-bounded QoS requirements θ_1 and θ_2 for the primary and secondary communications over the wireless coupling channels typically differ from each other and thus need to be jointly guaranteed at the same time, which defines the *heterogeneous statistical delay-bounded QoS provisioning framework for CR-5G networks*. Then the aggregate effective capacity of CR-5G networks is the sum of the effective capacities corresponding to the primary and secondary communications over the wireless coupling channels. Thus, we can derive the aggregate effective capacity of CR-5G networks as $C_a(\theta_1, \theta_2, R_{c1}, R_{c2})$, where the generic aggregate effective capacity function $C_a(\cdot, \cdot, \cdot, \cdot)$ is specified by Eq. 2, and R_{c1} and R_{c2} are the instantaneous transmission rates for the primary and secondary communications, respectively. Under the heterogeneous delay-bounded QoS provisioning framework for CR-5G wireless networks, our objective is to maximize the aggregate effective capacity, which depends not only on the QoS exponent θ_1 over the primary channel, but also on the QoS exponent θ_2 over the secondary channel. Thus, we can derive the optimal power policies that aim at maximizing the aggregate effective capacity by jointly taking into account the two QoS exponents θ_1 and θ_2 at the same time for the CR technique-based 5G wireless networks. Unlike in D2D- and FD-5G wireless networks, where the two coupling channels receive the same priority, the primary communications channel needs to be given higher priority than the secondary communications channel, which also needs to be taken into account when determining the values of θ_1 and θ_2 in CR-5G networks.

The homogeneous statistical delay-bounded QoS provisioning scheme can only be used for CR-5G wireless networks when the delay-bounded QoS requirements for the primary and secondary devices are equal. If the delay-bounded QoS requirements for the primary and secondary devices are different, the heterogeneous statistical delay-bounded QoS provisioning scheme can support delay-bounded QoS guarantees, while the homogeneous statistical delay-bounded QoS provisioning scheme cannot guarantee the delay-bounded QoS for CR-5G wireless networks.

Performance Evaluations

We conduct simulation experiments to validate and evaluate our proposed heterogeneous statistical delay-bounded QoS provisioning architecture and the related resource allocation schemes for D2D-5G, FD-5G, and CR-5G networks, respectively. We use the Nakagami-*m* channel model, which is very generic and often best fits land-mobile and indoor-mobile multiple propagations. We set the bandwidth for the 5G networks $B = 100$ kHz and the fading parameter m of Nakagami-m distribution with $m = 2$.

For the D2D technique-based networks [2], the average power degradation of each channel is determined by $\overline{\gamma} = K (d_0/d)^{\eta}$, where d is the transmission distance, d_0 is the reference distance, K is a unitless constant corresponding to the antenna characteristics, and η is the path loss exponent. In our simulations, we set $d_0 = 1$ m, $\eta = 3$, and $T = 2$ ms. Furthermore, we choose K such that $\overline{\gamma} = 0$ dB at $d = 100$ m. Also, we set the average power constraint for the D2D technique-based networks as $\overline{P} = 1$ W. The coordinates of the BS, the cellular device, and two D2D devices are randomly chosen as (0, 0), (37.13, 27.85), (–20.76, –41.50), and (–5.76, –56.50), respectively, in a 2D space. Figure 4 plots the aggregate effective capacities using heterogeneous statistical delay-bounded QoS provisioning and homogeneous statistical delay-bounded QoS provisioning for the D2D-5G networks under the co-channel mode and orthogonal-channel mode, respectively, where the QoS exponent of the cellular communication (θ_2) is set as 10^{-2}, while the QoS exponent (θ_1) of the D2D communication varies from $10^{-4.5}$ to $10^{-1.5}$. As illustrated in Fig. 4, for both the co-channel and orthogonal-channel modes, the heterogeneous statistical delay-bounded QoS provisioning scheme can achieve significantly larger aggregate effective capacity than that for the homogeneous statistical delay-bounded QoS provisioning scheme. The homogeneous statistical delay-bounded QoS provisioning scheme can obtain the same aggregate effective capacity as that for the heterogeneous statistical delay-bounded QoS provisioning scheme only when the delay-bounded QoS requirement (θ_1) for the D2D communication is equal to the delay-bounded QoS requirement (θ_2) for cellular communication at $\theta_1 = \theta_2 = 10^{-2}$. This is expected because the heterogeneous statistical delay-bound-

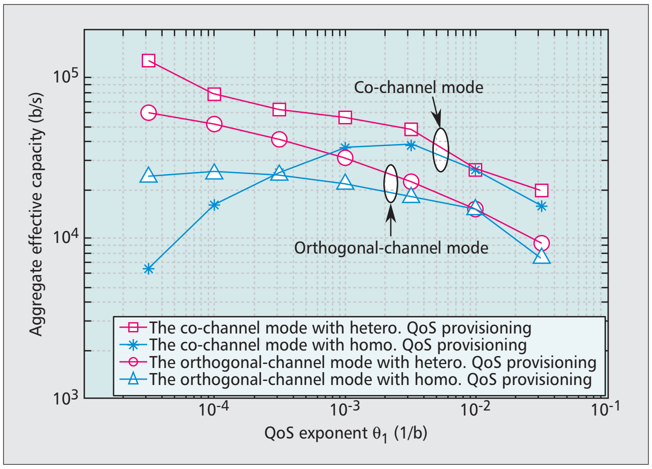

Figure 4. The aggregate effective capacities comparison between heterogeneous statistical delay-bounded QoS provisioning and homogeneous statistical delay-bounded QoS provisioning for D2D-5G mobile wireless networks under the co-channel mode and orthogonal-channel mode, respectively.

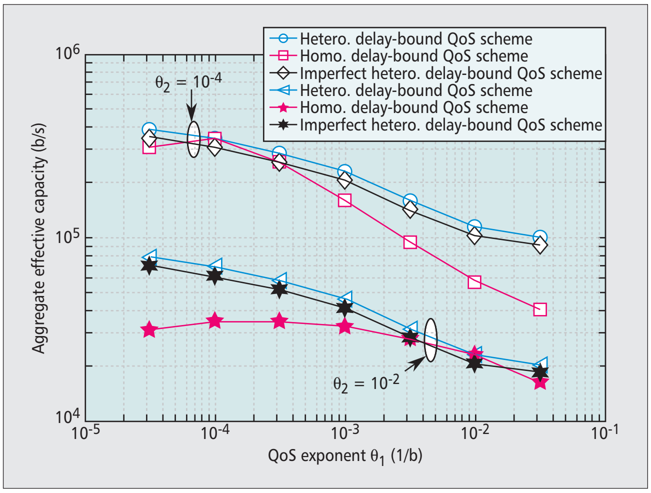

Figure 5. The aggregate effective capacities comparison between heterogeneous statistical delay-bounded QoS provisioning and homogeneous statistical delay-bounded QoS provisioning for FD-5G mobile wireless networks.

ed QoS provisioning scheme in fact reduces to the homogeneous statistical delay-bounded QoS provisioning scheme when both the D2D communication and cellular communication have exactly the same delay-bounded QoS requirements (i.e., $\theta_1 = \theta_2$), verifying the correctness of our system model and analyses in this aspect.

For the FD technique-based networks [1, 7], we set the average received average signal-to-noise ratio (SNR) as –3 dB, the average transmit power constraint as 1 W, and the time frame length T as 2 ms. Without loss of generality, we set the self-interference mitigation factors[1] for both the FD device and the BS as 0.95. Figure 5 illustrates the aggregate effective capacities comparison between heterogeneous statistical delay-bounded QoS provisioning and homogeneous statistical delay-bounded QoS provisioning for FD-5G, where we consider two cases:

- The required QoS exponent (θ_1) for the FD device varies from $10^{-4.5}$ to $10^{-1.5}$, while the required QoS exponent (θ_2) for the BS is $10^{-4.2}$).
- The required QoS exponent (θ_1) for the FD device varies from $10^{-4.5}$ to $10^{-1.5}$, while the required QoS exponent (θ_2) for the BS is 10^{-2}.

As can be observed from Fig. 5, when $\theta_1 \neq \theta_2$, the heterogeneous statistical delay-bounded QoS provisioning scheme always yields the larger aggregate effective capacity as compared to the homogeneous statistical delay-bounded QoS provisioning scheme for the FD-5G networks. The homogeneous statistical delay-bounded QoS provisioning scheme can achieve the same aggregate effective capacity as that for the heterogeneous statistical delay-bounded QoS provisioning scheme when the required QoS exponents θ_1 and θ_2 are the same (Fig. 5) because under this condition the heterogeneous statistical delay-bounded QoS provisioning scheme actually already reduces to the homogeneous statistical delay-bounded QoS provisioning scheme for FD 5G wireless networks. This observation also verifies the correctness of our system model and analyses in this aspect. The imperfect information exchange of the QoS exponents decreases the obtained aggregate effective capacity in approximately uniform fashion (Fig. 5).

For CR-5G networks [8], we set the required primary traffic load as 100 kb/s, the average SNR of each channel is equal to 10 dB, and the time frame length T is 10 ms. We set both the average power constraints for the primary and secondary devices as 1 W. Figure 6 plots the aggregate effective capacities using heterogeneous statistical delay-bounded QoS provisioning and homogeneous statistical delay-bounded QoS provisioning, respectively, for CR-5G networks, where the primary communication's QoS exponent θ_1 is set to 10^{-2} while letting the secondary communication's QoS exponent θ_2 vary from $10^{-4.5}$ to $10^{-1.5}$. For comparison purposes, Fig. 6 also plots the required effective capacity for the primary communication, which is always equal to 10^3 b/s. From Fig. 6, we can observe that the achieved aggregate effective capacity using the heterogeneous statistical delay-bounded QoS provisioning scheme is significantly larger than that using the homogeneous statistical delay-bounded QoS provisioning scheme. The achieved aggregate effective capacity using the homogeneous statistical delay-bounded QoS provisioning scheme is equal to that using the heterogeneous statistical delay-bounded QoS provisioning scheme only at $\theta_2 = 10^{-2}$ when the QoS exponents for the primary and secondary communications are the same at $\theta_1 = \theta_2 = 10^{-2}$ (Fig. 6). This is expected because when $\theta_1 = \theta_2$, the heterogeneous statistical delay-bounded QoS provisioning scheme in fact reduces to the homogeneous statistical delay-bounded QoS provisioning scheme for CR-5G wireless networks, showing that the homogeneous statistical delay-bounded QoS provisioning scheme is a special case of the heterogeneous statistical delay-bounded QoS provisioning scheme.

Conclusions

To overcome the new challenges imposed by efficiently supporting bandwidth-intensive and time-sensitive multimedia services over the emerging 5G mobile wireless networks, we propose the novel heterogeneous statistical delay-bounded QoS provisioning architecture for 5G mobile wireless networks that integrates three promising 5G candidate techniques. Under the proposed architecture, we develop the wireless coupling channel, joint heterogeneous QoS-exponents

[1] The self-interference mitigation factor, denoted by $\kappa = (SNR_s/SNR_r)$, is defined as the impact of self-interference on the local received SNR, where $\kappa \in (0, 1]$, the symbol SNR_s denotes the received SNR using the self-interference mitigation techniques, and the symbol SNR_r represents the received SNR without taking into account the self-interference impact caused by wireless FD transmission.

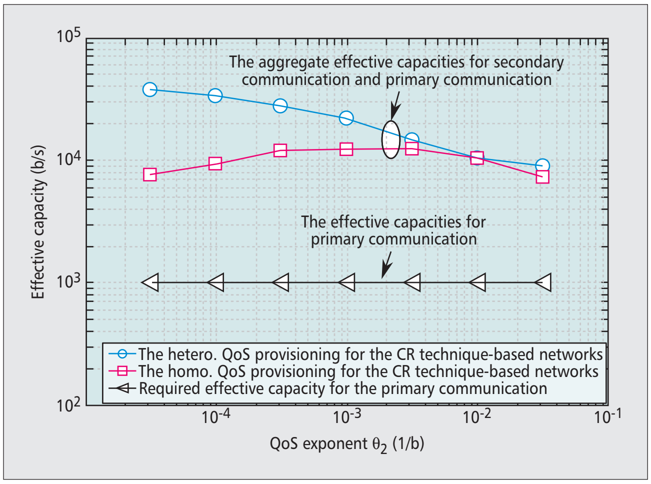

Figure 6. The comparison of effective capacities corresponding to the secondary communication using heterogeneous statistical delay-bound QoS provisioning and homogeneous statistical delay-bound QoS provisioning for CR-5G mobile wireless networks.

optimizations, and system integration models that can support the implementations of D2D-, FD-, and CR-based 5G networks, respectively. Using the developed system models, we derive the heterogeneous QoS provisioning architecture-based power allocation schemes that can maximize the aggregate effective capacities for the three 5G candidate-techniques-based wireless networks, respectively. Using the simulation experiments, we evaluate the achieved aggregate effective capacities of our proposed heterogeneous delay-bounded QoS architecture for D2D-5G, FD-5G, and CR-5G networks, respectively. The obtained simulation results validate our proposed architecture and also show that our proposed heterogeneous statistical delay-bounded QoS provisioning architecture and its corresponding power allocation schemes significantly outperform the existing traditional homogeneous statistical delay-bounded QoS provisioning schemes in terms of satisfying the heterogeneous delay-bounded QoS requirements while maximizing the aggregate system throughput over 5G mobile wireless networks.

References

[1] I. Chih-Lin *et al.*, "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.
[2] B. Bangerter *et al.*, "Networks and Devices for the 5G Era," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 90–96.
[3] "5G: A Technology Vision," tech. rep., Huawei, Mar. 2014.
[4] "Assessment of the Global Mobile Broadband Deployments and Forecasts for International Mobile Telecommunications," tech. rep., ITU-R M.2243.
[5] White paper, "Cisco Visual Networking Index: Forecast and Methodology," Cisco VNI Report, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white paper c11-481360.pdf, May 2013.
[6] N. Bhushan *et al.*, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 82–89.
[7] S. Hong *et al.*, "Applications of Self-Interference Cancellation in 5G and Beyond," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 114–21.
[8] C.-X. Wang *et al.*, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 122–30.
[9] X. Xhang *et al.*, "Cross-Layer-Based Modeling for Quality of Service Guarantees in Mobile Wireless Networks," *IEEE Commun. Mag.*, vol. 44, no. 1, Jan. 20063, pp. 100–06.
[10] J. Tang and X. Zhang, "Quality-of-Service Driven Power and Rate Adaptation over Wireless Links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, Aug. 2007, pp. 3058–68.
[11] White paper, "C-RAN: The Road Towards Green RAN," C. M. R. Institute, http://labs.chinamobile.com/cran/wpcontent/uploads/CRAN white paper v2 5 EN.pdf, Oct. 2011.
[12] W. Cheng, X. Zhang, and H. Zhang, "QoS-Aware Power Allocations for Maximizing Effective Capacity over Virtual-MIMO Wireless Networks," *IEEE JSAC*, vol. 31, no. 10, 2013, pp. 2043–57.
[13] W. Cheng, X. Zhang, and H. Zhang, "Optimal Power Allocation for Full-Duplex D2D Communications over Wireless Cellular Networks," *IEEE GLOBECOM*, Austin, TX, Dec. 2014.
[14] W. Cheng, X. Zhang, and H. Zhang, "Optimal Dynamic Power Control for Full-Duplex Bidirectional-Channel Based Wireless Networks," *IEEE INFOCOM*, Turin, Italy, Apr. 2013.
[15] X. Zhang and H. Su, "Opportunistic Spectrum Sharing Schemes for CDMA-based Uplink MAC in Cognitive Radio Networks," *IEEE JSAC*, vol. 29, no. 4, Apr. 2011, pp. 716–30.

Biographies

Xi Zhang [S'89, SM'98] received his B.S. and M.S. degrees from Xidian University, Xi'an, China, an M.S. degree from Lehigh University, Bethlehem, Pennsylvania, all in electrical engineering and computer science, and his Ph.D. degree in electrical engineering and computer science (electrical engineering-systems) from the University of Michigan, Ann Arbor. He is currently a professor and the founding director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station. He was a research fellow with the School of Electrical Engineering, University of Technology, Sydney, Australia, and the Department of Electrical and Computer Engineering, James Cook University, Australia. He was with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hill, New Jersey, and AT&T Laboratories Research, Florham Park, New Jersey, in 1997. He has published more than 270 research papers on wireless networks and communications systems, network protocol design and modeling, statistical communications, random signal processing, information theory, and control theory and systems. He received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He is an IEEE Distinguished Lecturer of both IEEE Communications Society and IEEE Vehicular Technology Society. He received Best Paper Awards at IEEE GLOBECOM 2007, IEEE GLOBECOM 2009, and IEEE WCNC 2010, respectively. He also received a TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering at Texas A&M University, College Station, in 2006. He is serving or has served as an Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Wireless Communications*, and *IEEE Transactions on Vehicular Technology*, twice as a GuestEditor for *IEEE Journal on Selected Areas in Communications* for two Special Issues on Broadband Wireless Communications for High Speed Vehicles and Wireless Video Transmissions, an Associate Editor for *IEEE Communications Letters*, a Guest Editor for *IEEE Communications Magazine* and *IEEE Wireless Communications*, an Editor for Wiley's *Journal on Wireless Communications and Mobile Computing*, *Journal of Computer Systems, Networking, and Communications*, and Wiley's *Journal on Security and Communications Networks*, and an Area Editor for Elsevier's *Journal on Computer Communications*, among many others. He is serving or has served as TPC Chair for IEEE GLOBECOM 2011, TPC Vice-Chair IEEE INFOCOM 2010, TPC Area Chair for IEEE INFOCOM 2012, Panel/Demo/Poster Chair for ACM MobiCom 2011, General Vice-Chair for IEEE WCNC 2013, ACM MobiCom 2011, and TPC/General Chair for numerous other IEEE/ACM conferences, symposia, and workshops.

Wenchi Cheng [M'14] received his B.S. and Ph.D. degrees in telecommunication engineering from Xidian University, China, in 2008 and 2014, respectively. He joined the Department of Telecommunication Engineering, Xidian University, in 2013 as an assistant professor. He worked as a visiting Ph.D. student under the supervision of Prof. Xi Zhang at the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University from 2010 to 2011. His research interests focus on 5G wireless networks, wireless full-duplex transmission, statistical QoS provisioning, cognitive radio techniques, and energy-efficient wireless networks. He has published multiple papers in *IEEE Journal on Selected Areas in Communications*, IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, and so on. He is serving as a TPC member for IEEE ICC 2015.

Hailin Zhang [M'98] received his B.S. and M.S. degrees from Northwestern Polytechnic University, Xi'an, China, in 1985 and 1988, respectively, and the Ph.D. from Xidian University, Xi'an, China, in 1991. In 1991, he joined the School of Telecommunications Engineering, Xidian University, where he is a senior professor and Dean of this school. He is also currently the director of the Key Laboratory in Wireless Communications sponsored by China Ministry of Information Technology, a key member of the State Key Laboratory of Integrated Services Networks, one of the state government specially compensated scientists and engineers, a field leader in telecommunications and information systems in Xidian University, and an associate director for the National 111 Project. His current research interests include key transmission technologies and standards on broadband wireless communications for 5G wireless access systems. He has published more than 100 papers in journals and conferences.