With the explosive development of mobile Internet and the emergence of multimedia-rich applications such as augmented reality, virtual reality, video streaming, online gaming, and social networks, the traffic demands in the incoming fifth-generation (5G) era could grow significantly. To accommodate this massive amount of data traffic, Internet service providers (ISPs) must deploy new infrastructures to expand the capacity of current networks, which results in high capital and maintenance costs. Furthermore, mobile Internet applications in 5G will have diverse quality-of-service (QoS) requirements in terms of latency, burst size, throughput, and packet arrival rate, and reliable QoS provisioning depends on efficient and on-demand resource allocation [1], [2]. However, it is
On-demand bandwidth slicing issues and pricing strategies have gained intensive attention from both academia and industry.

Background
Software-defined networking (SDN) provides a scalable, flexible, and programmable architecture for enabling centralized network control and management by decoupling the network control plane from the physical network topology. By integrating SDN with 5G cellular technologies, i.e., software-defined cellular networks, a series of network control and management mechanisms (e.g., traffic management, power control, and spectrum allocation) could be easily implemented with the centralized controller via standardized interfaces, without incurring a major modification to the underlying physical network components [3]–[5]. In addition, with the advanced level of physical infrastructure abstraction and virtualization, virtualized resources such as bandwidth can be split into distinct slices for dedicated usage and dynamically allocated in accordance with real-time demands. Hence, the ISP is able to satisfy the diverse QoS requirements of mobile Internet applications by exploring the on-demand bandwidth slicing capability in software-defined cellular networks.

To successfully implement smart bandwidth slicing, it is crucial to adapt the supply of bandwidth with the temporal-spatial-varying user demands. For ISPs, how to cope with the fluctuation of data traffic for reliable service delivery remains challenging. According to [6], peak-time traffic has increased more than 50% during the past few years and will likely continue to increase. During the peak time, it is difficult for the ISP to provide sufficient bandwidth resources to meet the QoS requirements of all users. Yet a large percentage of bandwidth remains underutilized during the off-peak time. Therefore, it is important to develop smart bandwidth slicing techniques for the ISP to flatten out the fluctuation of traffic demand, essential for spectrum efficiency improvement and investment expense reduction.

Pricing offers a powerful methodology for establishing user-centric control. Among numerous pricing strategies, time-dependent pricing (TDP) has become a promising solution to alleviate peak-time congestion [7]. TDP enables effective peak-time congestion management by charging users according to the amount of data as well as when these data are consumed. Therefore, TDP is more effective in flattening out peak-time traffic fluctuation compared with other pricing schemes, since the information of the temporal dimension has been incorporated for decision making.

In this article, we develop a unified pricing framework to address the on-demand bandwidth slicing problem in software-defined cellular networks from the perspective of TDP. We provide a comprehensive survey of the related works and propose a hierarchical game framework to capture the competitive interactions between ISPs and users. This article also elaborates on how to derive the optimal congestion prices for the scenarios of single-congested links and multiple-congested links based on the real-time traffic demands and the incomplete knowledge of user utility. The ISP estimates the expected traffic demands and determines the optimal congestion price to alleviate network congestion while maximizing the total revenue. Here, the price is utilized as a signal to reveal the current network congestion level, which provides an incentive mechanism to encourage delay-tolerant users to postpone their bandwidth demands form peak to off-peak time. A case study is conducted to validate the effectiveness on peak demand reduction. We conclude by identifying several potential research directions.

Related Works
On-demand bandwidth slicing issues and pricing strategies have gained intensive attention from both academia and industry. Here, we present a commentary on the development of on-demand bandwidth slicing in SDN, pricing strategy, and TDP. Some works already exist that address the bandwidth slicing problem in SDN. In [6], the authors designed a transport network architecture in the 5G-Crosshaul project to increase bandwidth efficiency. Multitenancy is enabled by an SDN-based control plane via flexible and efficient network slicing. In [8], the authors proposed the concept of bandwidth slicing in wired SDN-enabled networks, where bandwidth is virtually split into multiple slices for the dedicated use of different services. In [9], the network slicing problem was formulated as a mixed binary linear programming problem, and efficient penalty successive upper-bound minimization and penalty successive upper-bound minimization-approximation algorithms were developed to address the problem. However, the prior-mentioned works have not investigated the optimization of bandwidth slicing in SDN from a TDP perspective.

Pricing strategy provides an effective methodology to enforce user-centric control. One of the prevailing pricing models for both wired and wireless networks is the flat rate due to its simplicity [10]. The disadvantage of a flat rate is that it cannot distinguish between heavy users and light users, which forces light users to substantially subsidize heavy users. Another category of pricing widely utilized by worldwide ISPs is usage-based pricing (UBP), which is efficient to control the total data usage.
A programmable SDN controller works as the brain and enforces all control functionalities via standardized interfaces.

in a certain period of time. UBP is not suitable for peak-time congestion control since the price is not determined according to the network congestion level [11]. Game theory provides a powerful tool to investigate the interactions among decision makers with conflicting interests. An economic framework consisting of auction-based dynamic spectrum allocation and noncooperative behavior game-based service pricing was proposed in [12]. These works mainly focus on developing pricing schemes for addressing conventional resource allocation problems, and none investigate the impact of time dependency on pricing or take into account the unique characteristics of bandwidth slicing.

Another line of work is TDP, in which the price users pay for the obtained bandwidth resources in accordance with the timing and quantity of data consumed. In [13], the authors proposed a TDP framework that combines the patterns of both spatial and temporal traffic for network slicing, which achieves an average reduction of 16% in the peak-to-average ratio of the overall network. In [14], a novel off-peak charge discount service was developed to reduce peak demand, which takes into account the off-peak window size as well as the probability that users may shift their traffic demand.

In contrast to the aforementioned works, we consider both pricing and bandwidth slicing in software-defined cellular networks as well as the information asymmetry between ISPs and users. Specifically, we explore the advantages of TDP on peak-time traffic fluctuation flattening and develop a unified framework for addressing the peak-time congestion problem based on real-time traffic demands and incomplete knowledge of user utility.

Interaction Between the ISP and Users: A Hierarchical Game Formulation

A Bandwidth Slicing in Software-Defined Cellular Networks

The conceptual architecture of the software-defined cellular network is shown in Figure 1. A programmable SDN controller works as the brain and enforces all control functionalities via standardized interfaces. The distributed base stations, which are responsible for data reception and transmission, are dynamically configured by the centralized controller with global knowledge. We assume there are $L$ logical wireless links in the software-defined cellular network, and the total amount of available bandwidth for each link $l$ of $L$ is denoted as $B_l$. We adopt a time-slot model, in which the time is divided into multiple slots with an equal length of $T$. Let $S_i(t)$ denote the set of users on link $l$ in slot $t$, and $x_{req}^s$ denote the bandwidth requirement of user $s$. If the available bandwidth reserved for link $l$ is not enough to meet the bandwidth demands of all users in $S_i(t)$, i.e., $\sum_{s \in S_i(t)} x_{req}^s > B_l$, then link $l$ is denoted as a congested link, and the corresponding time slot $t$ is denoted as a peak-time slot.

The SDN controller monitors the imbalance between bandwidth supply and traffic demand on each link and will issue a congestion price to users. Therefore, the price in off-peak time is ignored for the sake of simplicity. Upon receiving the price, users who remain connected during the peak time will be charged a penalty. Users can decide whether to stay connected by comparing the delay cost with the connection cost. In this way, the congestion price provides a methodology to incentivize some delay-tolerant users to voluntarily defer their connection requests to the off-peak time, which eventually alleviates the link congestion problem. An iOS application is developed to let users monitor and react to the time-dependent congestion price. The user interfaces during the peak time are shown in Figure 2. The users who choose to defer their demands will be informed of the latest congestion price during the specified time.

Interaction Between the ISP and Users

The interaction between the ISP and users is as follows: when there is one ISP serving multiple users in the network, the ISP is able to adjust the congestion price to motivate delay-tolerant users to postpone their occupancy of the bandwidth from peak time to nonpeak time. There is a tradeoff when the ISP determines the congestion price. On one hand, if the price is set too high, the ISP is able to reduce the peak demand by increasing the penalty imposed on users, while the total revenue of the ISP will decrease since many users may choose to defer their connection requests. On the other hand, if the price is set too low, most users will choose to stay connected since the price penalty is minimal compared to the delay cost. This could result in severe congestion problems and a QoS violation. Thus, it is challenging for the ISP to determine the optimal congestion price without knowing the strategy of each user. To maximize the revenue while relieving the congestion problem, the ISP needs to design an efficient incentive mechanism not only to achieve the maximal revenue but also to motivate users to defer their connection requests for congestion avoidance.

A Hierarchical Game Formulation

Given the priority of the ISP to determine the congestion price, the competitive interaction between users and the ISP can be captured by employing a hierarchical game
formulation. The Stackelberg leader-follower game [15], in which leaders have a dominant market position compared with followers, can efficiently model the hierarchy of players. Generally, the leader determines its strategy first, and each follower then chooses its individual best response according to the leader’s strategy. The derived optimal strategies form a Nash equilibrium, which means that no player can unilaterally improve its utility by changing to another strategy.

Let $p_t$ denote the congestion price, which varies dynamically in accordance with real-time link congestion levels. The penalty imposed on user $s$ who remains connected in the peak slot $t$ is calculated as $P_{st} = p_t x_{st}^\text{con}$. In the Stackelberg game formulation, the ISP is the leader that determines the congestion price $p_t$ for each slot $t$. Accordingly, the users act as the followers do and determine whether to stay connected based on $p_t$. We define a binary variable $b_{st}$ to represent the decision of user $s$ in slot $t$. That is, $b_{st} = 1$ represents that user $s$ stays connected in slot $t$ and $b_{st} = 0$ otherwise. Intuitively, user $s$ will defer its traffic demand if the cost of service deference is less

**Figure 1** The conceptual architecture of a software-defined cellular network.

**Figure 2** (a) The user interface when reaching peak time. (b) The latest congestion price during a specified time.
We investigate the more complicated scenario where the user traffic traverses multiple-congested links and derive the corresponding optimal congestion pricing strategy.

than the connection cost, i.e., \( p_{t}x^{s}_{t}\tau \). Individual rationality is assumed for both the ISP and the users; in particular, the response of each player is optimized to maximize its own benefit. The elements of the Stackelberg game, \( G(\text{Player}, \text{Strategy}, \text{Payoff}) \), are described as follows:

- **Player**: The ISP and the multimedia streaming users act as the leader and the followers, respectively.
- **Strategy**: For the ISP, the strategy is to determine the congestion price \( p_{t} \) in each slot \( t \), which is a positive and continuous variable. For each user, the strategy is the decision whether to stay connected in slot \( t \) based on \( p_{t} \), which is represented as a binary variable \( b_{s,t} \).

**Payoff**: For the ISP, the payoff is the revenue gain; for the users, the payoff is the net benefit, which is illustrated as follows.

We model the utility of each user as a function of delay. Specifically, when the streaming service of user \( s \) is delayed by \( d_{s} \), the corresponding delay cost is defined as \( C_{s}(d_{s}) = \theta_{s}f(d_{s}) \). Here, the delay cost function \( f(d_{s}) \) is assumed as strictly increasing and convex and is represented by an exponential function \( e^{d_{s}} \). \( \theta_{s} \) is the user type, which represents the delay sensitivity of user \( s \). That is, different users may have diverse preferences toward service delay. Furthermore, user type \( \theta_{s} \) is a user-dependent factor and is the private information of user \( s \). Due to information asymmetry, the user type is not exactly known by the ISP in practical implementation. Hence, we assume that only the incomplete knowledge is known to the ISP, e.g., the distribution of \( \theta_{s} \) can be learned from historical observations by machine learning or big data techniques.

Let \( \tilde{C}_{s}(d_{s}, \tau) \) denote the marginal delay cost of user \( s \) when the traffic is further deferred by one slot \( \tau \). \( \tilde{C}_{s}(d_{s}, \tau) \) can be calculated as

\[
\tilde{C}_{s}(d_{s}, \tau) = \theta_{s}[e^{d_{s}+\tau} - e^{d_{s}}].
\]

Hence, the optimal strategy of user \( s \) can be obtained by comparing \( \tilde{C}_{s}(d_{s}, \tau) \) with \( p_{t}x^{s}_{t}\tau \). Rational users choose to stay connected when the price penalty is less than the marginal delay cost, i.e., \( \tilde{C}_{s}(d_{s}, \tau) > p_{t}x^{s}_{t}\tau \), and wait for a slot.

**Time-Dependent Congestion Pricing Strategy**

**Single-Congested Link**

To discuss how to derive the optimal time-dependent congestion price \( p_{t} \) under two different scenarios (i.e., the single-congested link and the multiple-congested link), first, we consider a simplified scenario where all users use the same single-congested link \( l \). If user \( s \) remains connected on link \( l \) in the peak-time slot \( t \), the penalty imposed on user \( s \) is \( p_{t}x^{s}_{t}\tau \), which equals the revenue gained by the ISP. With the congestion price \( p_{t} \), the probability that user \( s \) stays connected in slot \( t \) is represented as

\[
1 - \frac{x^{s}_{t}\tau p_{t}}{e^{d_{s}}(e^{\tau} - 1)}.
\]

**Figure 3** The flow charts of (a) the optimal congestion pricing algorithm and (b) the iterative bandwidth allocation algorithm.
which is a function of the probability density function of $\theta_i$. The optimal congestion price for the single-congested link can be obtained by maximizing the revenue of the ISP under the following two constraints:

1) The aggregate traffic demand on link $l$ in slot $t$ (i.e., the traffic demand when the congestion price is $p_l$) is higher than the total amount of available bandwidth, which indicates that link $l$ is a congested link and slot $t$ is a peak-time slot.

2) The available bandwidth $B_l$ of link $l$ in slot $t$ should not be lower than the expected traffic requirements of users after enforcing the congestion price $p_l$.

**Multiple-Congested Link**

We investigate the more complicated scenario where the user traffic traverses multiple-congested links and derive the corresponding optimal congestion pricing strategy. Assuming the total number of logical links is $L$, the set of congested links that the data demand of user $s$ to pass through in slot $t$ is denoted as $L(s)$. The congestion price of each link $l$ in slot $t$ is determined to maximize the revenue of the ISP.

Then, from the perspective of revenue maximization, an iterative bandwidth allocation algorithm is proposed to determine the best response of users, in which the optimal price on link $l$ in slot $t$ can be obtained by a greedy algorithm. First, we initialize the congestion price of all links as 0. Second, we find the most-loaded link $n$ in set $L$. Third, if the link $n$ is congested, then we iteratively increase the congestion price by a small perturbation $\Delta p$ and estimate the congestion relief effect on each link. The iteration process ends when all links become uncongested. The detailed flow charts of the optimal congestion pricing algorithm and the iterative bandwidth allocation algorithm are shown in Figure 3.

**Case Study**

Without a loss of generality, we assume that request arrivals follow a Poisson distribution and session durations follow an exponentially distribution with an average value of 1 h. To emulate the traffic fluctuation throughout the day, we pick the period [0:00, 8:00) (sleep hours), [8:00, 20:00) (office hours), and [20:00, 24:00) (after hours) to serve as typical off-peak, midpeak, and peak time, respectively. The request arrival rates of off-peak, midpeak, and peak time, respectively. The request arrival rates of off-peak, midpeak, and peak time, respectively.

For comparison, we consider a fixed pricing scheme. The distinctions between the two schemes are described as follows.

- **Proposed:** Congestion price for each peak slot is determined by solving the revenue optimization problem under an available bandwidth constraint.

- **Fixed:** The ISP charges a flat rate throughout the typical peak time. Specifically, we set the fixed price to 0.1 and 0.2 by considering the mean value of the proposed congestion price.

Figure 4 shows the congestion price of each time slot. It can be observed that the congestion price of the proposed scheme does not turn to zero immediately at the end of each typical peak time (e.g., 24:00). This is not surprising because the demand for the typical peak time is partially shifted to the next typical off-peak time, and the price incentive is still essential for alleviating the shifted congestion.

As shown in Figure 5, the traffic load demand (the green curve) varies as a function of time. After applying the proposed scheme, it can be observed that the expected value of the total bandwidth allocated is basically well controlled below the available bandwidth of the link (i.e., 2 Gb/s), which satisfies the capacity constraint of our revenue optimization problem. Traffic demand of typical peak time is partially shifted to the next typical off-peak time, which is evidenced by the fact that the total bandwidth allocated in the proposed scheme is even higher than the traffic load demand in the typical off-peak time (i.e., 24:00–28:00).

Figure 6 shows the revenue received from the users that choose to stay connected during peak time. We observed that, due to the stochastic properties of traffic demand and the incomplete information about the users’ preferences, implementing a pricing scheme...
that achieves real revenue optimization is unrealistic. The proposed scheme, which tries to optimize the revenue in each slot based on the traffic demand monitored at the beginning of each slot, provides a more practical solution. Although we consider revenue optimization to be the ultimate goal of ISPs, when compared to the fixed (0.2) pricing scheme, the total revenue gain of the proposed scheme is even lower. This is because the capacity constraint is merely imposed on the proposed scheme.

Selecting an appropriate congestion price is important to congestion management. As shown in Figures 4–6, a lower congestion price of the proposed scheme at the beginning of peak time (e.g., 20:00) lets users stay connected and therefore be served first to vacate more bandwidth for users who are latecomers. When traffic demand becomes very high (e.g., 24:00), a higher congestion price forces the delay-tolerant users to shift their demand from peak time to off-peak time. Hence, the ISP gets enough surplus bandwidth for the delay-sensitive users. On the other hand, a fixed congestion price (relatively high) at the beginning of peak time may force users to defer their traffic when congestion is not heavy, which makes the congestion even worse later. When congestion gets significantly heavy, the fixed congestion price (relatively low) cannot provide enough incentive to distinguish between delay-sensitive users and delay-tolerant users, resulting in a limited congestion-relief effect.

Conclusions and Open Issues
In this article, we investigated the TDP problem for on-demand bandwidth slicing in 5G software-defined cellular networks. The congestion price, which reveals the current level of the traffic load, serves as a signal to motivate delay-tolerant users to postpone their traffic demands during peak hours. Numerical results confirm the proposed scheme can yield satisfactory performance in terms of shifting the traffic demand from peak time to off-peak time and manage the congestion problems effectively.

There are several issues that require further investigation, such as user uncertainty about how many slots to wait until the next off-peak slot and the burden of decision making on whether to keep connecting in each peak slot. We believe these problems are inevitable as long as the price is set to reflect the real-time traffic load. However, these problems can be alleviated if users can set their budgets and delay tolerance in advance. If the users are not intended to monitor the time-varying congestion prices, artificial intelligence could be introduced to assist the users for decision making (i.e., to stay connected or not). Moreover, the deferred online streaming contents could be downloaded offline within the budget and delay tolerance specified by the users.

Most existing TDP schemes are developed under the assumption that there is complete knowledge of key parameters, such as channel state information and the type of user. However, this assumption is too optimistic considering the complexity of future 5G application

**Figure 5** Bandwidth allocation versus time.

**Figure 6** Revenue versus time.
scenarios and the heterogeneity of networks and their users. Furthermore, the users may deliberately hide their types and preferences or even provide misleading and deceptive information. In these cases, the legendary pricing algorithm may no longer apply, and significant research efforts are required to investigate the adverse impacts of uncertainty and develop robust schemes based on partially observed knowledge.

Acknowledgments
This work was partially supported by the Beijing Natural Science Foundation under grant 4174104, Beijing Outstanding Young Talent under grant 201600020124G081, National Natural Science Foundation of China under grant 61601181, the Fundamental Research Funds for the Central Universities under grant 2017MS13, and Japan Society for the Promotion of Science Kakenhi under grant 16K16051.

Author Information
Zhenyu Zhou (zhenyu.zhou@ncepu.edu.cn) received his M.E. and Ph.D. degrees in global information and communication studies from Waseda University, Tokyo, in 2008 and 2011, respectively. He is an associate professor at the State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, School of Electrical and Electronic Engineering, North China Electric Power University, Beijing. He received the Beijing Outstanding Young Talent Award from the Beijing Government, China, in 2016. He is an editor of IEEE Access and IEEE Communications Magazine. His research interests include green communications and smart grid technology. He is a Senior Member of the IEEE.

La Tan (13717568897@163.com) is currently working toward her M.S. degree in smart-grid information engineering at North China Electric Power University, Beijing. Her research interests include green communications and smart grid.

Bo Gu (bo.gu@cc.kogakuin.ac.jp) received his B.E. degree from Tianjin University, China, in 2004; his M.E. degree from Peking University, Beijing, in 2007; and his Ph.D. degree in global information and communication studies from Waseda University, Tokyo, in 2013. Since 2016, he has been with the Department of Information and Communications Engineering, Kogakuin University, Tokyo. His current research interests include network economics game theory and network optimization. He is a Member of the IEEE and the Institute of Electrical and Information and Communication Engineers, Japan.

Yan Zhang (yanzhang@ieee.org) received his Ph.D. degree in communication engineering from the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore. He is a full professor in the Department of Informatics, University of Oslo, Norway. His current research interests include next-generation wireless networks leading to fifth-generation as well as green and secure cyberphysical systems. He is a Senior Member of the IEEE and a member of the IEEE Communications Society, the IEEE Computer Society, the IEEE Power & Energy Society, and the IEEE Vehicular Technology Society. He is a Fellow of the Institution of Engineering and Technology.

Jun Wu (junwuhn@sjtu.edu.cn) received his Ph.D. degree in global information and communication studies from Waseda University, Japan, in 2011. He is an associate professor at the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China. He is the chair of the IEEE P21451-1-5 Standard Working Group and has been a Technical Program Committee member of more than ten international conferences. He has been a guest editor of IEEE Sensors Journal and is an associate editor of IEEE Access. He is a Member of the IEEE.

References