Congestion Pricing in Wireless Cellular Networks

Bader Al-Manthari, *Student Member, IEEE,* Nidal Nasser, *Member, IEEE,* and Hossam Hassanein, *Senior Member, IEEE*

***Abstract*—While the demand for wireless cellular services continues to increase, radio resources remain scarce. As a result, network operators have to competently manage these resources in order to increase the efficiency of their Wireless Cellular Networks (WCN) and meet the Quality of Service (QoS) of different users. A key component of Radio Resource Management (RRM) is congestion control. Congestion can severely degrade the performance of WCN and affect the satisfaction of the users and the obtained revenues. Several congestion control techniques have been proposed for WCN. These techniques, however, do not provide incentives to the users to use the wireless network rationally, and hence they cannot solve the problem of congestion. Recently, there has been some research on providing monetary incentives to the users through congestion pricing to use the wireless network rationally and efficiently. Congestion pricing is a promising solution that can help alleviate the problem of congestion and generate higher revenues for network operators. This paper surveys recent research work on congestion pricing in WCN. It also provides detailed discussions and comparisons of the surveyed work as well as open problems and possible future research directions in the area.**

***Index Terms*—Wireless cellular networks, congestion pricing, call admission control, power control, auction-based congestion pricing.**

I. Introduction

THE RAPID proliferation of Wireless Cellular Networks (WCN), as well as the emergence of new wireless "content rich" multimedia applications, with diverse Quality of Service (QoS) requirements, present many challenges to wireless network operators due to the limited bandwidth. As a result, network operators have to manage their shared resources in the most efficient way. Specifically, such resources must be managed in a way that satisfies the QoS requirements of the supported services, distributes the network radio resources in an efficient and fair way, provides incentives to the users to use the network services and maximizes the obtained revenues. To achieve these objectives, wireless network operators utilize Radio Resource Management (RRM), which is a set of algorithms that control the usage of radio resources. Congestion control is one of the key components of RRM. Congestion occurs when the demand for bandwidth exceeds the supply. This has recently arisen due to the rapid increase in demand for wireless services especially multimedia services that require high bandwidth allocations. Greed also is another source of network congestion. When a resource is publicly shared, as is the case with the wireless spectrum in Wireless Fidelity (Wi-Fi) networks or when price is affordable, a greedy user may continuously transmit, rendering the system unusable for others. This phenomenon is known in economics as the "tragedy of the commons" [1] (Check the appendix for a complete list of glossary). If a user transmits when the network is congested, the QoS of other users in the network such as packet delay and packet loss may become severely degraded. In economics terms, this is known as congestion externality, which is closely related to the tragedy of the commons problem [2]. Therefore, without proper congestion control mechanisms, the QoS requirements of different users may never be guaranteed.

Any congestion control mechanism should promote efficient use of the shared wireless resources by penalizing greedy users to avoid the tragedy of the commons problem. It should also provide incentives to users to use the network services and maximize economic efficiency in a way that the users who value the network resources the most are the ones who actually get these resources. There are a number of ways to deal with congestion in WCN. One such way is to keep the demand for bandwidth at a low level by imposing high fees on customers or expanding the wireless resources. Imposing high fees may of course discourage users from using the network services to the extent of underutilizing the network, and hence affecting the obtained revenues. It is also very costly to expand bandwidth in wireless networks and in many cases this might not even be feasible due to physical or regulatory restrictions. Another approach to deal with congestion is to impose a penalty on greedy users. For example, the wireless network operator may drop the packets of greedy users or place a bandwidth, volume caps or time limits on their calls to alleviate congestion. Despite their effectiveness, penalty schemes fail to achieve economic efficiency, since they provide no guarantees that the users who value the shared wireless resources the most are the ones who actually get them.

Recently, there has been some research on using economic-based approaches to deal with congestion. The most common approach is congestion pricing. Congestion pricing is based on the idea of dynamically increasing or decreasing the prices of network services (or resources) according to the congestion level in the network [2]. Hence, congestion pricing is also referred as dynamic pricing. Congestion pricing solutions in general assume that users are price-sensitive, which is normally the case with most users. This paper reviews and dis-

Manuscript received 5 July 2008; revised 28 January 2009, 18 September 2009, and 25 March 2010 . This work was supported by the IEEE.

B. Al-Manthari is with the Center of Information Security, Information Technology Authority, Azaibah, Sultanate Of Oman P.O.Box: 1807 P.C: 130 (e-mail: bader.almanthari@ita.gov.om).

N. Nasser is with the Department of Computing & Information Science University of Guelph, Guelph, ON, N1G 2W1 Canada (e-mail: nasser@cis.uoguelph.ca).

H. Hassanein is with the Telecommunications Research Laboratory (TRL), School of Computing (SC) Queen's University, Kingston, ON, K7L 3N6 Canada (e-mail: hossam@cs.queensu.ca).

Digital Object Identifier 10.1109/SURV.2011.090710.00042

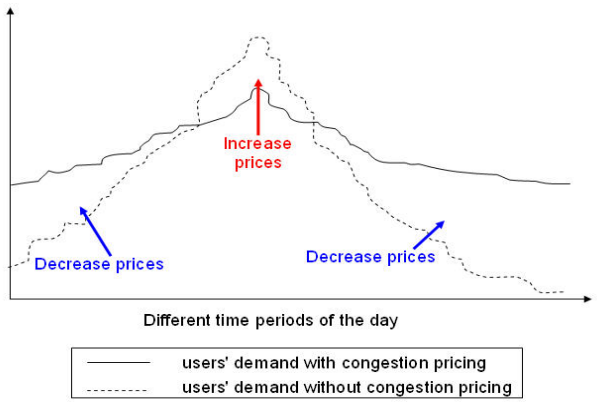

Fig. 1. Effect of Congestion Pricing on Shifting the Users' Demand

cusses in depth several recent research studies on congestion pricing in WCN. The paper is structured as follows. Section II provides a brief overview of congestion pricing. Sections III and IV present recent work on admission-level and power-level congestion pricing, respectively, which are the main two types of congestion pricing in WCN. Section V states our conclusions and outlines open problems and possible future research directions in the area.

II. Congestion Pricing

With congestion pricing, the charges of network services are dynamically determined according to the network load [2], where they are increased when the network load is high to reduce the demand for network services and they are decreased when the network load is low to increase the demand for network services as shown in Figure 1 [2]. Congestion pricing can, therefore, competently promote rational and efficient usage of the shared wireless resources by influencing users' behaviors. Congestion pricing is, therefore, a promising solution to traffic control problems, which can help alleviate the problem of congestion. In addition, it can enhance economic efficiency, since it ensures that the wireless resources are given to those who value them the most. Furthermore, congestion pricing is more cost-effective than bandwidth expansion and it can generate higher revenues.

We remark that congestion pricing (and pricing in general) is a complex and multifaceted problem since there are many economical, social and technical issues that need to be taken into consideration [3] and [4]. Examples of economical issues include incentives to use the network services and knowledge of users demand functions and their sensitivity to prices. Social issues include social welfare[1], social fairness[2] and user acceptance of congestion pricing. Technical issues include compatibility with existing and future standards, complexity, determination of pricing interval and billing mechanisms.

In general, the design of any congestion pricing scheme depends primarily on two crucial components:

1) User demand behavior: any congestion pricing scheme must take into account the demand behaviors of users. Different users react differently to prices because some of them are more sensitive to prices than others. This is known in economics as the *price elasticity of demand*, which measures the responsiveness of a change in demand for a good or service to a change in price [5]. Different pricing schemes use different demand models. For example, some of them use exponential functions to represent the users' demands for wireless services whereas others use utility functions to represent the users' preferences and/or their *Willingness to Pay* (WTP) for a certain service.
2) Price function: the price function determines how the price of a certain service is computed for a unit of time, bandwidth and/or power. Different schemes use different price functions depending on many factors including objectives of congestion pricing, characterizations of resource usage, causes of congestion, assumptions about users' behaviors, etc.

Various congestion pricing schemes have been proposed for wired networks, particularly the Internet, such as smart-market congestion pricing [6], progressive congestion second pricing [7] and proportional fair congestion pricing [8], [9] and [10]. These schemes differ in many aspects including their price functions, demand functions, etc. Details on pricing schemes, including congestion pricing, in wired networks can be found in [3] and [11]. Congestion pricing in WCN is more involved than wired networks due to limited bandwidth, interference and user mobility. In this paper, we survey the schemes that have been proposed for WCN. We classify congestion pricing schemes in WCN into two categories namely, admission-level and power-level congestion pricing. This classification is based on the level at which users are charged. It should be noted that the congestion charge[3] is not the only component of the total charge for a certain service. Along with a congestion charge, the charge of a service may include an access charge, a usage charge and/or a QoS charge [12].

III. Admission-Level Congestion Pricing

In admission-level congestion pricing, the price for a unit of time or bandwidth is determined when the user initiates a call request and before the call is admitted to the system. The price in this case is fixed for the call duration. This price is dynamically determined according to the network load. There are two types of calls at admission level, new and handoff calls. A new call occurs when a user requests a new connection, while a handoff call occurs when an active user moves from one cell to another. It should be noted that since calls are charged at admission level, handoff calls are not affected by congestion prices since they were already charged at the cell where the calls have been initiated. Therefore, most of the schemes surveyed in this section, consider only the effect of congestion pricing on new calls.

[1]Social welfare is the aggregate utility of the people

[2]Social fairness refers to the state of economy where the majority of people are able to buy certain products regardless of their incomes. In the context of this paper, it refers to the ability to buy or use network services

[3]Charge is defined as the amount that is billed for a service, whereas price is the amount of money associated with a unit of service. That is, price is used to compute the charge [2]

The rationale behind using congestion pricing at the admission level is to control call request arrivals to the system through monetary incentives in order to maintain the connection-level QoS at a desired threshold. The main QoS metrics at the connection level are new call blocking and handoff call dropping probabilities. The new call blocking probability is the probability that a new call is rejected. Whereas, the handoff call dropping probability is the probability that a handoff call is dropped. The general procedure for admission-level congestion pricing is as follows. When a user makes a new call request, the base station or the Radio Network Controller (RNC) computes the price for a unit of time or bandwidth and announces this price to the user as shown in Figure 2.(a). If the user accepts the price, he can then establish the call. Otherwise, he can retry later when the price is lower. Different schemes might use variations of this general procedure. For example, in case of congestion pricing based on bidding, when the user sends a new call request, he attaches a bid with his request. The base station or the RNC then determines whether this bid is accepted or not. If the call request is of type handoff, then the base station or the RNC only checks if there are enough available resources to accept the call or reject it without computing a new price for the call as shown in Figure 2.(b). We classify admission-level congestion pricing schemes into two types: Direct and Indirect Call Admission Control.

A. Direct Call Admission Control

Wireless networks operators employ Call Admission Control (CAC), which is a resource provisioning strategy that is used to limit the number of connections in the network in order to improve its performance and meet the Quality of Service (QoS) of ongoing users' connections. CAC, by itself cannot avoid congestion because it does not provide incentives to the users to use the shared wireless network resources rationally and efficiently. Therefore, the new call blocking and handoff call dropping probabilities can reach high levels during congestion periods. To provide such incentives and reduce these probabilities, CAC is integrated with congestion pricing. Hence, the term Direct CAC.

Integrated Congestion Pricing with CAC

In [13], Hou *et al*. propose a scheme that integrates congestion pricing with CAC to simultaneously address the problem of congestion and maximize total user utilities (i.e., social welfare). In this scheme, the utility of a user is defined as a function of the call rejecting probabilities P_b , where P_b is defined as a weighted sum of the new call blocking probability and handoff call dropping probability. The authors in [13] prove that for a given wireless network, there exists an optimal new call arrival rate λ_n^* that maximizes the total utility of users. When $\lambda_n > \lambda_n^*$, the network becomes congested as a large number of new and handoff calls are blocked, and hence users receive lower QoS than their requirements. Therefore, it is in best interest of the wireless network operator to keep the new arrival rate λ_n less than or equal to λ_n^*, which is the objective of congestion pricing.

Figure 3 shows a schematic representation of the proposed CAC scheme in [13]. The system consists of two blocks, the CAC block and the pricing block. The CAC block is responsible for accepting or rejecting calls based on the amount of available resources. The pricing block works as follows: when $\lambda_n < \lambda_n^*$ a fixed price is charged to every user and when $\lambda_n > \lambda_n^*$ a congestion price will be charged to users who still want to use the network services. If users do not accept the offered congestion price, they can make their calls later when the network load decreases and the price is lower. These users, denoted by hold-off users in Figure 3, generate retry traffic with arrival rate λ_r. The scheme uses the demand function proposed in [14] to model the users' behaviors towards price changes as follows

$$D(cp(t)) = \exp\Big[-\Big(\frac{cp(t)}{fp} - 1\Big)^2\Big], \quad cp(t) \geq fp \qquad (1)$$

where $D(cp(t))$ is the percentage of users that will accept the congestion price, fp is the fixed price charged when $\lambda_n \leq \lambda_n^*$ and $cp(t)$ is the price charged during congestion periods. Using this demand function and through some mathematical manipulations, $cp(t)$ is calculated such that the arrival rates $\lambda_n + \lambda_r$ (i.e., of new and retry users) are equal to the optimal arrival rate λ_n^*. When compared to conventional systems, where pricing is not taken into consideration, the scheme in [13] shows considerable improvements in terms of congestion prevention, achieved total user utility and obtained revenue. In addition, the scheme assumes a realistic queuing model where it considers new users as well as hold-off users. However, the pricing scheme is designed to prevent the system from over loading by keeping $\lambda_n \leq \lambda_n^*$, but it fails to prevent it from being under utilized. This is because when λ_n is much less than λ_n^*, users are charged a fixed price and they are not given any incentives to increase their usage of the network, which results in resource wastage, and hence revenue loss. Furthermore, the price setting totally depends on the demand function and the assumption that the user's utility is a function of the parameter P_b. Therefore, if users' real demands or utilities are different from the assumed ones, the congestion price will not be able to maximize the social welfare of the system, and hence the performance of the system may be greatly affected. Moreover, the assumed demand function does not take into account the price elasticity of demands of different users, which should be considered in any realistic pricing scheme.

Congestion Pricing with Alternatives

S. Yaipairoj *et. al*. [15] propose a CAC with a congestion pricing scheme, which divides users into two types, priority users and conventional users. Priority users pay a higher price in order to be served immediately or relatively sooner, whereas conventional users pay a fixed price but are delayed for a longer time. The scheme model, as illustrated in Figure 4, consists of two queues, one for priority users and the other for conventional users. The scheme consists of two important functional blocks, a pricing block, which is responsible for price broadcasting to incoming users and a Priority Call Admission Control (PCAC) block, which is responsible for admitting users from both queues. The scheme works as follows. When the system is not congested, all incoming users are charged a fixed price and are placed in the conventional queue since there are enough resources to serve them all.

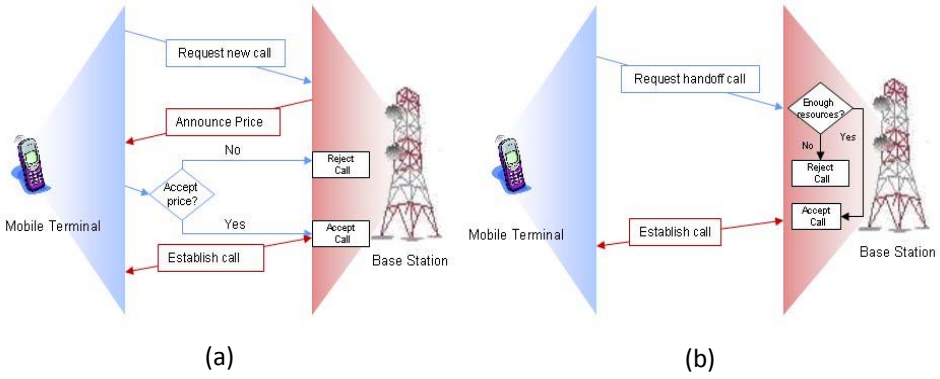

Fig. 2. Admission-Level Congestion Pricing Procedure for (a) New Call (b) Handoff Call

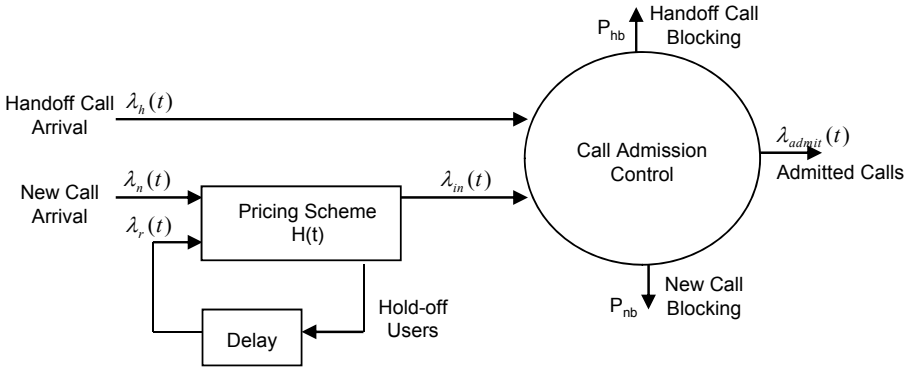

Fig. 3. Schematic Representation of the CAC Scheme in [13]

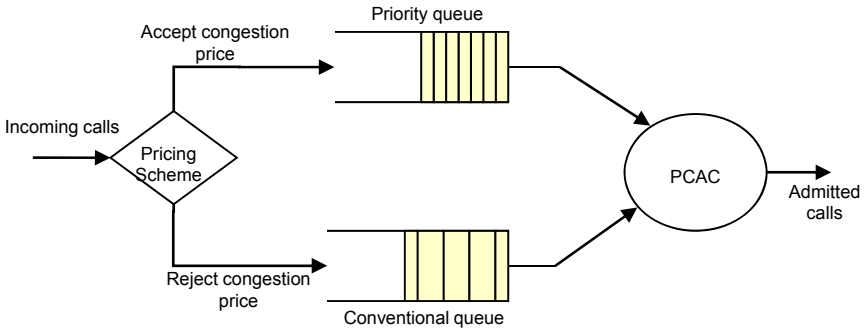

Fig. 4. Schematic Representation of the CAC Scheme in [15]

However, during congestion periods, a higher congestion price is calculated and announced to incoming users. If users accept the congestion price, they are placed in the priority queue where they are served faster; otherwise, they will be placed in the conventional queue. The authors in [15] define equations for calculating the probability that the user's delay of each type of queue is larger than a certain threshold. Therefore, the scheme can utilize these equations to guarantee that the user's call would not exceed a certain amount of time in the corresponding queue. To distinguish between priority and conventional users, the PCAC block uses a *priority factor* (P_s), which is the probability that the priority queue is served over the conventional queue. Therefore, P_s greatly impacts the performance of the system and the obtained revenues. The value of P_s is determined with the help of congestion pricing such that the maximum number of users that the network can accommodate, and yet conform to the QoS constraints of users. Numerical results [15] show that there is a significant revenue increase using this scheme with pricing over a CAC scheme that does not utilize congestion pricing. The overall system utilization is also improved.

Although the scheme considers the delay the users experience in the two queues, it does not take into account the new call blocking and handoff call dropping probabilities. This may not be practical since wireless network operators have a limit on the number of calls they can block, which is usually

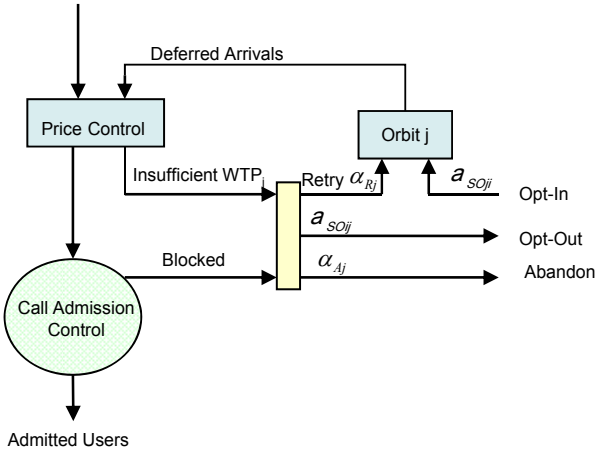

Fig. 5. Schematic Representation of the CAC Scheme in [16]

determined by regulations. Furthermore, there should be some investigations on the amount of time users can tolerate waiting in the queues prior to admission before they cancel their call requests after which they are considered blocked.

CAC-Based Congestion Pricing and Dynamic Bandwidth Reservation for Handoff Calls

S.L. Hew [16] proposes a congestion pricing scheme that is integrated with CAC in order to reduce congestion and maximize revenues in multi-service WCN. The behaviors of the users of service class i towards prices is modeled by their Willingness to Pay (WTP), which is represented by a Weibull distribution [17] with mean φ_i and shape β_i. The Weibull distribution is used because it is versatile and can take up the characteristics of other types of distributions depending on the value of β_i. In fact, exponential demand functions such as the one used in [13] and [15] are special cases of the Weibull distribution [16]. Using the concept of WTP, a new user of service class i will decide to make a connection request if his WTP is larger than or equal to the expected cost of the service, i.e., $\varphi_i \geq cp_i(t)\dot{b}_i/u_i$, where $cp_i(t)$ is the price of the service of class i at time t, b_i is the units of the required bandwidth and u_i is the average call duration. Assuming k classes of service, new users of service class i who are rejected by the CAC or have insufficient WTP, will either retry later with probability α_{R_j}, opt-out to another service $j \neq i$ with probability $\alpha_{SO_{ij}}$ or abandon the system with probability $\alpha_{A_{ij}}$. New users of service class i who retry later and users of service class $j \neq i$ who opt-in service class i are said to be in *orbit*, as shown in Figure 5.

By varying the price, the arrival rate of a certain service can be encouraged or discouraged in order to reserve some bandwidth for arriving handoff or higher revenue-generating users. The system is described by a Markov Chain [18] and the problem of CAC and congestion pricing is formulated as a Markov Decision problem [19] in order to exercise CAC and congestion pricing such that congestion is reduced and the long-run expected revenues are maximized. Numerical results show that this scheme increases revenues and reduces congestion compared to other schemes that do not employ congestion pricing.

The most distinctive feature of this scheme compared to other ones is the realistic queuing model that considers the effects of prices on arrivals, retrials and substitutions among services. However, this increases its complexity, and hence the complexity of the solution. Another important feature is that the scheme does not reserve a fixed number of channels (or bandwidth) for handoff users as is the case with all schemes presented in this section. Instead, the reservation of bandwidth for handoff users is determined in real time based on the expected revenues that the system would get from handoff users. However, similar to the scheme in [15], this scheme fails to consider the new call blocking and handoff call dropping probabilities, which may limit its practicality.

B. Indirect Call Admission Control

The pricing schemes surveyed in this section are designed to regulate the usage of the wireless network resources without being integrated with CAC. Unlike Direct CAC schemes, the schemes in this section do not employ explicit CAC but rather they implicitly limit the number of connections at admission level during congestion periods. Hence, we call them Indirect CAC schemes.

Congestion Pricing for Differentiated Mobile services

In reference [20] a dynamic pricing scheme for differentiated cellular mobile services is proposed. The scheme classifies users into k QoS classes, where each class i $(1 \leq i \leq k)$ is defined by the new call admission probability δ_i (i.e., the probability that a call is admitted) and $\delta_1 > \delta_2 > \cdots > \delta_k$. The main feature of the scheme is that the price of the call from the i^{th} QoS class is varied according to the traffic load in m discrete steps namely, $cp_{i1}, cp_{i2}, \ldots, cp_{im}$ within a range $cp_i^{\min}$ and $cp_i^{\max}$ in order to regulate users' demands. This is done such that the revenues of the wireless network operator are maximized.

The effect of the price on the demand is modeled as follows [21]:

$$D_i(t) = a_i(t)e^{-b_i(t)cp_i(t)}, \tag{2}$$

where at any given time t, $D_i(t)$ is the quantity of resources demanded by the users in the i^{th} QoS class, $cp_i(t)$ is the price of a call from the i^{th} QoS class, $a_i(t)$ is the demand shift constant of i^{th} QoS class and $b_i(t)$ is the price elasticity of demand of the i^{th} QoS class. $a_i(t)$ and $b_i(t)$, which can assume different values for different times of the day and QoS classes can be determined by market studies on real demand behaviors for the different users.

Since prices are chosen from within a predetermined set, users relatively know how much they are expected to be charged. However, determining the range of the price set (i.e., $cp_i^{\min}$ and $cp_i^{\max}$), its cardinality (i.e., m) and the different prices (i.e., $cp_{i1}, cp_{i2}, \ldots, cp_{im}$) for each QoS class is very difficult. For example, choosing a high value of $cp_i^{\max}$ may lead to revenue loss as more users are discouraged from using the wireless services. In addition, if m is large, the pricing problem in [20] becomes much harder to solve. Therefore, these values should be carefully chosen.

Auction-Based Congestion Pricing with Reservation Price

S. Yaipairoj *et al.* [22] propose a congestion pricing scheme for wireless data services in General Packet Radio Service (GPRS) system - a 2.5G WCN [23]. The authors in [22] argue

that estimating the users' demands can be too complex and time consuming because users' demands can be a function of QoS, price structure and the applications the users are running, which vary significantly among different markets. The argument in [22] is that without accurate estimation of the users' demands, auctions can be used to price wireless services. The scheme is conceptually similar to smart-market pricing [6] in which users submit bids along with their service requests. The main difference between the smart-market approach and this scheme is that in the former, bids are submitted with every packet, whereas in the scheme in [22], bids are submitted only at the beginning of the call. The auction method assumed in the scheme is the multi-unit second-price sealed bid or multi-unit Vickery auction [24]. In multi-unit Vickery auction, users are allowed to bid for more than one unit of the same item and the K highest bidders are admitted into the system where they pay the value of the highest losing bid. In this scheme, the submitted bids are required to be greater than or equal to a certain value denoted by the *reserve price*. The authors in [22] calculate the optimal reserve price that maximizes the excepted revenues of the wireless network operator. This price is to be used when the network is not congested. During congestion periods, indicated by increased mean system delay, the optimal reserve price is increased to discourage excessive network usage. The new reserve price is calculated so that it achieves the maximum delay improvement with the minimum loss of revenue. Numerical results show that by increasing the auction's reserve price, the network congestion is significantly reduced.

The most important feature that distinguishes the scheme in [22] is that it does not require any estimate of user demand, since this information is conveyed by the users through their submitted bids. In addition, with Vickery auction users are less likely to shade their bids, a phenomenon by which users bid below their true valuations of the items being auctioned to avoid subsequent loss of wining when bidding high prices in auctions where users pay the highest bids [2]. As a result, the largest overestimation of an item's value ends up winning the auction. Moreover, the scheme is designed for GPRS where the QoS support in this system is very limited. The authors in [22] claim that emerging technologies such as Enhanced Data rates for GSM Evolution (EDGE) [23] - a GPRS-based system, and Universal Mobile Telecommunication System (UMTS) [25] - a 3G WCN, are potential candidates for deployment of their proposed scheme. However, the deployment of the scheme on EDGE and UMTS, requires new formulas for the mean system delay to be derived as these systems differ from GPRS in a number of aspects, including the supported data rate, modulation and channel coding algorithms [23] and [25].

Auction-Based Congestion Pricing for Differentiated Wireless Services

S. Mandal *et al*. [26] and [27] propose a congestion pricing scheme for differentiated wireless services. The proposed scheme is similar to the one in [22] in that it is based on bidding and that it aims at maximizing the revenues of the wireless network operator in addition to controlling congestion. The difference with the scheme in [22] is that the former does not use the reserve price. Two multi-unit auctions mechanisms are proposed, namely uniform pricing auction and discriminatory pricing auction. In uniform pricing auction [2], the highest L bidders are chosen and the bidders pay the same price, which is the clearing price at which the demand exhausts the available resources (i.e., the price of the lowest wining bidder). In discriminatory pricing auction [2], the highest L bidders are chosen such that the demand exhausts the supply and each chosen bidder is charged his own bid. The proposed scheme assumes k QoS classes where each class i ($1 \leq i \leq k$) is assumed to be defined by new call admission probability δ_i. To maximize the revenue, the scheme admits the highest bidders according to the pricing auction being used such that the QoS of different classes is met.

Simulation results show that uniform and discriminatory pricing auctions outperform flat pricing in terms of revenues. The major disadvantage of discriminatory pricing auction is that it suffers from bid shading since users may bid below their true valuations of the auctioned items to avoid paying high prices when wining the auction. Similar to Vickery auction [24], users with uniform pricing auction are less likely to shade their bids and hence, uniform pricing auction may result in higher revenues than discriminatory pricing auction [2]. Users, however, will bid on average higher prices with uniform pricing than with discriminatory pricing because the former eliminates the winner's curse. As aforementioned, the scheme in [26] and [27] does not use a reserve price. Nevertheless, it is expected that this scheme could help control congestion because as the arrival rate to the wireless system increases, the number of bidders for resources increases and as a result, the expected bids' values go up. This makes the scheme easy to implement. However, it also makes it less responsive to congestion since without the reserve price, it takes the users some time to figure out that the right values that they should bid. In addition, without the reserve price, the earned revenues depend totally on the amount of submitted bids, which if they happen to be very low, will result in great revenue loss to the wireless network operator. This problem becomes more aggravated if users cooperate to submit low bids, which is a highly unlikely but possible scenario.

Congestion Pricing for Connection-Oriented Services

E. Viterbo *et al*. [28] propose a linear optimal congestion pricing scheme for connection-oriented services in 3G WCN in order to maximize the revenues of network operators. The authors argue that linear prices are less complex to compute and are more understood by users. The user's behavior is modeled by the following demand function [29]:

$$D(cp(t), Q) = e^{-\alpha cp(t) + \beta Q}, \tag{3}$$

where $cp(t)$ is the congestion price at time t, Q is a quality of service metric, which is defined to be the call admission probability, and both α and β determine the rate of change of $D(cp(t), Q)$ as a result of the change in the congestion price and/or the quality of service metric. Let N be the maximum number of available communication channels. The authors in [28] describe the system evolution as a set of states using a Markov chain, where each state $i = 0, , N$ represents the number of active calls in the system (i.e., each call uses one channel). A pricing vector $\mathbf{cp} = (cp_0, cp_1, \ldots, cp_{N-1})$ is used to represent the cost of using the service where cp_i represents the price of a call when the system is in state i (i.e., there are

i active calls). The main objective of this scheme is to find $\mathbf{cp}$ such that it maximizes the wireless network operator's revenues. Numerical results show that when compared to flat rate pricing, the pricing scheme in [28] achieves higher revenues and lower call blocking probability.

The most distinctive feature of this pricing scheme is that it models the user's call duration as a decreasing function of the congestion price. Such modeling is logically important because even though the users know how much they will be priced at the beginning of their calls; they usually tend to decrease their calls durations as the prices go higher. However, this depends on many factors in addition to the price such as the users' WTP, their budgets, etc. Unfortunately, those factors are not considered in the proposed model. Another important aspect of the proposed scheme is the use of linear pricing. Linear pricing may not yield the maximum revenues because there could be another non-linear pricing vector $\mathbf{cp}$ that achieves higher revenues. However, as argued by the authors in [28], linear pricing schemes are more understood by users, which might increase their acceptance of congestion pricing. This pricing scheme is similar to the scheme in [20] in which users can anticipate how much they will be priced at, or at least they can anticipate the range that the prices will fall in. However, the set of prices in [20] are predetermined whereas in this scheme they are optimally computed.

Table I summarizes the most distinctive features and the limitations of the schemes surveyed in this section.

TABLE I
SUMMARY OF FEATURES AND LIMITATIONS OF ADMISSION-LEVEL CONGESTION PRICING SCHEMES

Scheme Reference	Most Distinctive Features	Limitations
[13]	Realistic queuing model	No prevention of system's underutilization. No consideration of effect of price on call duration. Limited QoS support. Total dependence on assumptions about utility function.
[15]	Alternatives to the users to choose between dynamic and fixed prices	No consideration of effect of price on call duration. No support for connection-level QoS. Limited QoS support.
[16]	Realistic queuing model. General distribution function for user demand. Good QoS support. Dynamic bandwidth reservation for handoff calls.	Complex solution. No consideration of effect of price on call duration. No support for connection-level QoS
[20]	Price determination from within a known set.	Difficulty in determining the price set, its cardinality and prices within it. Difficulty in determining the price set, its cardinality and prices within it. No consideration of effect of price on call duration. Limited QoS support
[22]	No estimate of user demand is required.	Effect of price on call duration is not considered. No QoS support.
[26], [27]	Simple. No estimate of user demand is required	Less responsiveness to congestion. Effect of price on call duration is not considered. Limited QoS support. Revenue loss if users cooperate.
[28]	Consideration of effect of price on call duration. Use of simple linear pricing	Limited QoS support.

IV. POWER-LEVEL CONGESTION PRICING

In power-level congestion pricing, users are dynamically charged according to their power consumption in order to regulate their power usage in the network and to help control congestion. This means that power-level congestion pricing is exercised after the call is admitted where the price varies during the call according to the user's power usage. Power-level congestion pricing schemes are specific to Code Division Multiple Access-based (CDMA) networks [30] such as UMTS [25] because these networks suffer from interference caused by the power transmitted by the users to the base station (uplink) or vice versa (downlink), which degrades their QoS. The rationale behind using congestion pricing at the power level is twofold. The first is providing monetary incentives to the users to use the power rationally and efficiently in order to mitigate interference, and hence alleviate congestion. The second is to save power, and thus increase battery life of users' mobile terminals. Therefore, congestion pricing at the power level serves as a power control function. To determine the amount of power needed in the uplink or downlink, a measure of the user's signal quality is needed. Three alternate measures for the signal quality are used [31], namely:

1) Received Signal-to-Interference Ratio (SIR): the ratio of power of desired signal to the power of all other interfering signals. High SIR values mean good signal quality, which allows higher data rate transmission.
2) Received Signal-to-Interference-Plus Noise Ratio (SINR): similar to SIR expect that it takes into account physical noises such as the thermal noise and background noise. SINR is considered more accurate than SIR. However, since the physical noises are independent of the system under evaluation, SIR is used for evaluating different power control schemes
3) Bit-Energy-to-Noise-Density Ratio (E_b/N_0): obtained by multiplying SIR or SINR by the processing gain (i.e., the physical bandwidth divided by the actual data rate). This measure is usually used to obtain the Bit Error Rate (BER), which is the number of erroneous bits received divided by the total number of bits transmitted. The BER is a very important physical QoS metric because high BER values correspond to resource wastage due to the large number of erroneous bits that need to be retransmitted.

Since the channel characteristics in the uplink are different from those in the downlink, congestion pricing in the uplink is different from the downlink. Therefore, we classify the schemes surveyed in this section into uplink and downlink power-control congestion pricing. The general procedure for power-level congestion pricing is as follows. In the uplink, the base station or the RNC dynamically determines the price per unit of power in the uplink and announces this information to the users as shown in Figure 6(a). The users then decide the amount of power they will transmit at in the uplink according to the announced price. In the downlink, the base station announces the price per unit of power in the downlink. The

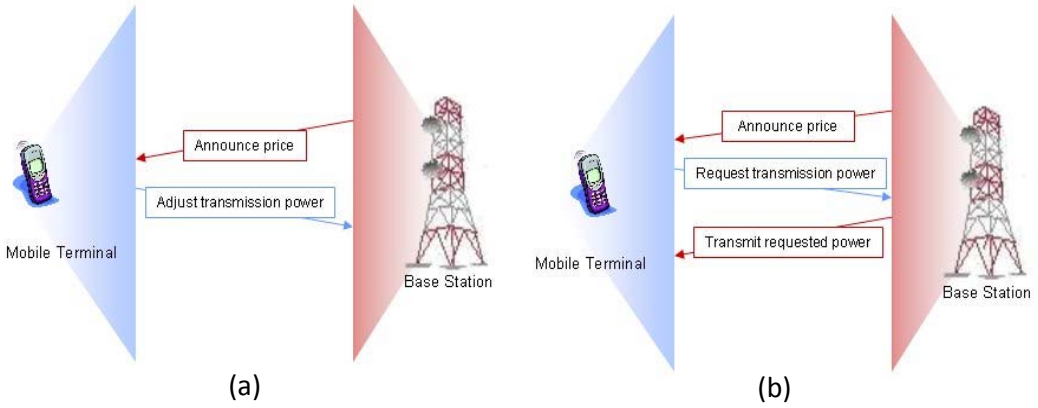

Fig. 6. Power-Level Congestion Pricing Procedure for (a) Uplink (b) Downlink

users then decide the amount of power in the downlink and send this information to the base station, which will transmit to the users according to their requested power as shown in Figure 6(b). Different schemes might use different variations of this procedure. For example, some schemes are distributed and allow the congestion price to be computed at the user side, as well as at the base station or RNC.

A. Uplink Power Control Congestion Pricing

The uplink in CDMA-based networks is interference-limited [32], which implies that users cannot increase their power without bound due to the interference that they would cause to other users in the network. Therefore, the schemes surveyed in this section aim at reducing interference in the uplink by controlling the transmitted power by the users through providing monetary incentives to them.

Utility-Based Congestion for WCN

To sustain high uplink data rates, users need to transmit at high power levels in order to keep the received SIR at the base station at high levels. However, as aforementioned, transmitting at high power levels causes interference to other users, which affects their attainable data rates. S.W. Han *et. al*. [33] propose a distributed congestion pricing scheme for optimal uplink power allocation in CDMA-based networks. They model the users' behaviors by a utility function of the received SIR, which is in turn a function of the transmitted power (i.e., the users are happier with high SIR since it translates to high data rates). In their scheme, each user decides on a transmission power as to maximize his utility minus the cost of power (i.e., his net utility). The price is updated in steps based on the user's transmission power. The authors then prove that their scheme converges to an optimal power allocation that maximizes the social welfare.

The scheme in [33] reduces the signaling by the users and the base station because the price and power computation is done at the user's side. This increases the computation at the user side, which might affect his battery life. In addition, the user's utility is assumed to be a function of the received SIR, which is a concave function. We remark, however, that concave functions are incapable of representing the satisfactions of users with minimum data requirements, which are usually represented by sigmoid utility functions [34] and [35]. Furthermore, users are charged differently in this scheme based on their locations where users with good signal qualities are charged less than the others. This is more network efficient, since users contributing more to interference are charged higher than others. Such a pricing scheme, however, is unfair to users, since for the same the data rates, users with different signal qualities will be charged differently. Therefore, a balanced solution between network efficiency and fairness would be more appropriate.

Uplink Resource Control for Elastic Traffic

V. A. Siris [32] proposes a distributed resource control scheme based on congestion pricing for the uplink and downlink in CDMA-based networks. It is mathematically shown that the resource usage for each user i in the uplink in CDMA-based networks is an increasing function of the product of two parameters, the transmission rate r_i and the signal quality, expressed in terms of the target Bit-Energy-to-Noise-Density Ratio, γ_i (i.e., $E_b/N_0 = \gamma_i$). Therefore, in this scheme, users are charged proportionally to their uplink resource usage (i.e., $r_i \cdot \gamma_i$). The work in [32] considers the case of elastic (best-effort) traffic where users value only the average throughput of successful data transmission, which is shown to be a function of r_i and γ_i. Similar to [33], the users' behaviors are modeled by a utility function and the scheme aims at maximizing social welfare. However, the user's utility is assumed to be a function of his average throughput, which is the most important measure of user satisfaction for best-effort traffic. In this scheme, each user decides on a target E_b/N_0 (i.e., γ_i) and a transmission power as to maximize his utility minus their cost. It should be noted that the users can maintain a target E_b/N_0 by gradually increasing their power until the desired E_b/N_0 is achieved. The author in [32] shows that the optimal values of r_i and γ_i can be computed in a distributed manner where each user only has to compute his optimal transmission rate whereas the optimal γ_i is computed at the base station.

The most distinctive feature of this scheme is that users are charged based on their transmission rate r_i and their target Bit-Energy-to-Noise-Density Ratio, γ_i, irrespective of the amount of power used in obtaining γ_i. This means that the price is independent of user's location. Whereas in the scheme [33], the price depends on the power level, and hence users who are close to the base station will potentially enjoy lower prices,

since they require lower power to transmit than those users who are far away from the base station. However, this comes at the expense of reduced network efficiency since, for a same data rate, users causing high interference are charged the same amount as those users causing low interference. Another feature of this scheme is that it can support other cost functions in addition to the resource usage such as the cost of battery power of the user and the cost of congestion in wired network as explained in [32]. Therefore, the scheme can be used as a basis for the integration of congestion control mechanisms in wired and wireless networks.

B. Downlink Power Control Congestion Pricing

Unlike the uplink, the downlink in CDMA-based networks is power limited [32], which means that there is a limit on the total transmission power by the base station. This is because the base station causes interference to adjacent cells, which degrades the QoS of the downlink for users in those cells. Therefore, the goal of the schemes surveyed in this section is to regulate the allocation of the downlink power by using congestion pricing.

Downlink Resource Control for Elastic Traffic

V. A. Siris [32] proposes a downlink resource control scheme for CDMA-based networks, which is similar to the uplink version in the same paper. It is mathematically shown that the transmission power from the base station to user i characterizes resource usage. Therefore, users are charged proportionally to their power usage. In this scheme, each user decides on a downlink transmission power as to maximize his utility minus the cost of power, where the user utility is assumed to be a function of his average throughput. The author in [32] also shows that the optimal values for the transmission rate r_i and the target Bit-Energy-to-Noise-Density Ratio γ_i can be computed in a distributed manner.

This scheme shares the same features as its uplink counterpart. However, unlike the uplink version, users are charged according to the amount of transmitted power. Hence, for the same average throughput, users close to the base station will enjoy potentially lower prices than those who are far away from it, which increases the efficiency of the network at the expense of unfairness. In addition, the distributed algorithm provided in [32] to compute the optimal values of r_i and γ_i involves more computations at the user's side than the uplink one, which reduces the signaling between the users and the base station but might affect their battery life.

Downlink Power Allocation for Multi-Call Wireless Systems

J.W. Lee *et al.* [34] and [35] propose a downlink power allocation scheme for wireless systems supporting multiple classes of service. The scheme uses a utility function to represent user satisfaction of the received QoS. The utility of user i is assumed to be a function of the received SIR. Unlike other schemes, the authors in [34] and [35] argue that natural utility functions for users are typically non-concave and, therefore, they allow the utility function of each user with respect to power allocation to be one of three types: sigmoidal-like, strictly concave or a strictly convex function. The objective of the scheme is to obtain a power allocation that maximizes the social welfare. Since the scheme assumes a general utility function, obtaining a global optimal power allocation is difficult. Therefore, the authors propose a distributed algorithm based on congestion pricing to approximate the performance of the global solution and show that its asymptotic performance converges to the global optimum. The algorithm consists of two stages, mobile selection and power allocation. In the mobile selection stage, mobile users to which the power is allocated are selected (only a subset of users are allocated power since there is a limit on the total transmitted power). Once the users are selected, the power is optimally allocated to the selected users in the power allocation stage. In this stage, each selected user will choose a transmission power to maximize his utility minus the cost of power. Based on the power request of each user, the base station will dynamically adjust the price of power usage ($cp(t)$) such that it obtains a good approximation of the global power allocation.

The most distinctive feature of this scheme is the use of a general utility function, which is not restricted to be concave. This is more practical and allows for supporting multi-class services in a unified way, since different services might have different utility functions. Note that the utility function assumed in [32] is a special case of the utility function assumed in [34] and [35]. Similar to the scheme in [32], the price depends on the amount of transmitted power. Mobile selection and power allocation require a lot of signaling between the users and the base station, which might increase interference in the system, as well as increase resource wastage, since such signaling consumes power. In addition, the high signaling between the users and the base station might decrease the battery life of the users.

Downlink Power Allocation Using Power and Code Congestion Pricing

P. Liu *et al.* propose a downlink power control scheme for voice users in CDMA-based networks in [36] and [37]. They state that CDMA-based networks are limited by available resources, which constitute transmission power and codes where the number of available codes is limited by bandwidth, whereas the amount of transmitted power is limited by physical constraint or by the interference caused to neighboring cells. Therefore, pricing should include both transmitted power and codes. The users are represented by a utility function, which reflects their WTP for a certain QoS. Since only voice users are considered, QoS is determined by the received SINR. If the received SINR is above a certain threshold γ^*, then the QoS is acceptable and the user is indifferent to any additional increase in the received SINR. Conversely, if the received SINR is less than γ^*, then the QoS is unacceptable and the user derives zero utility. To model this, a step utility function of the received SINR $U_i(\gamma_i)$ is used, where $U_i(\gamma_i) = U_i$, $\gamma_i \geq \gamma^*$ and $U_i(\gamma_i) = 0\, \gamma_i < \gamma^*$. In the scheme, the base station announces a price per unit of power $cp_p(t)$ and a price per code $cp_c(t)$. Hence, the total charge for service incurred by user i is $cp_c(t) + cp_p(t) \cdot p_i$, where p_i is the requested power. The user then responds by requesting a transmission power that maximizes his net utility. The power radiated by the base station causes interference to adjacent cells, which degrades the QoS of the users in those cells. This is accounted for by assuming that the base station makes a transfer payment to the

network given by $\beta \cdot P$, where β is a constant and P is the total transmitted power by the base station. The objective of the base station is then to find $cp_p(t)$ and $cp_c(t)$ to maximize social welfare or revenue minus the transfer payment. The authors in [36] and [37] show that the optimal prices (i.e., $cp_p(t)$ and $cp_c(t)$) that maximize social welfare can be found and provide a closed-form solution for them. However, when maximizing the revenues, finding the optimal prices becomes more involved since as the prices increase, the number of active users decreases, but the revenue per user increases. Optimization becomes difficult in this case since revenue as a function of price is an irregular-surfaced function that has many jumps corresponding to the specific prices at which users become activated or deactivated [36] and [37]. In this case, to find the optimal prices, the authors in [36] and [37] use a large system model in which they assume that the number of users and the number of codes tend to infinity while keeping the load constant. This model is used to avoid the associated analytical problems with a finite system and find the optimal prices when the objective function is maximizing revenues. Please refer to [36] and [37] for more details on this model. Extensive experiments on the scheme are done in [37] and [38] where the effect of code and power prices on the network load are shown.

A very distinctive feature of the scheme in [36] and [37] is the identification of resource usage by both power and codes. Unlike other schemes, which only consider the amount of transmitted power, this is more practical since a network with limited power and codes may have high demands on codes and less demand on power and vice versa. Another important feature of the scheme is the consideration of the interference caused by the cell on its adjacent cells, which help reduce interference, and hence congestion in the whole network. However, similar to other schemes, this scheme suffers from unfairness since users are charged based on their signal quality.

Table II summarizes the most distinctive features and the limitations of the surveyed schemes in this section.

TABLE II
SUMMARY OF FEATURES AND LIMITATIONS OF ADMISSION-LEVEL CONGESTION PRICING SCHEMES

Scheme Reference	Most Distinctive Features	Limitations
[33]	Simple Low signaling overhead High network efficiency	High computations at the users' side. Limited utility function. Unfairness. No consideration for effect of price on call duration. Limited QoS support.
[32] (Uplink)	Fairness Ability to accommodate other costs such as the battery cost and the cost of congestion in wired networks High network efficiency Ability to be used as a basis of integration of congestion in wired and wireless networks	No consideration for effect of price on call duration. Low network efficiency. Unfairness. Limited QoS support.
[32] (Downlink)	Ability to accommodate other costs such as the battery cost and the cost of congestion in wired networks Ability to be used as a basis of integration of congestion in wired and wireless networks High network efficiency	High computations at the user's side. Unfairness. No consideration for effect of price on call duration. Limited QoS support.
[34], [35]	General utility function Support for multi-service classes in a unified way High network efficiency	High signaling overhead. Unfairness. No consideration for effect of price on call duration. Limited QoS support.
[36] [37]	Identification of resource usage by both power and codes Accounts for interference caused by the base station on adjacent cells High network efficiency	Unfairness. Effect of price on call duration is not considered. Limited QoS support.

V. CONCLUSIONS, OPEN ISSUES AND FUTURE RESEARCH OPPORTUNITIES

Wireless cellular services have encountered an enormous demand in the past few years. This trend is expected to continue in the future due to the rapid support of new wireless multimedia services that were previously available to wireline users. In spite of such high growth in demand for wireless cellular services, radio resources remain scarce. Therefore, these resources must be managed in the most efficient way in order to competently maximize the efficiency of the wireless network and meet the requirements of both network operators and users. Congestion control is one of the components of Radio Resource Management (RRM) techniques without which wireless networks cannot work efficiently. Several congestion control mechanisms have been proposed in the literature. None, however, can effectively alleviate congestion especially when the demand for a certain service is high. This is because such techniques do not provide incentives to the users to use the network rationally and efficiently. Recently, there has been some research on providing monetary incentives to the users to use the network wisely through congestion pricing. Congestion pricing is an effective congestion control mechanism since it provides control signals to the users to decrease their demand for a certain service when the network is congested. It can also generate higher revenues to the service provider, which maybe used to fund capacity expansion.

In this paper, we surveyed some recent congestion pricing techniques for Wireless Cellular Networks (WCN). We classified them into two categories depending on the level at which they are executed namely, admission-level and power-level congestion pricing. In admission-level congestion pricing, the price per bit or unit of time is dynamically determined at the beginning of the call according to the network's load and is fixed during the call lifetime. This type of congestion pricing has the advantage that users know how much they will be priced. Therefore, they can make judicious decisions whether to accept the price, and hence make the call, or not depending on their budgets. However, pricing at the admission level makes the network less reactive to congestion because the system can never anticipate how much traffic the data users will generate. Hence, it is possible that the network reaches a congestion state and stays in such a state for a long period even if the prices of the offered services are raised because the users who are overloading the network are already admitted and are not affected by the congestion prices. In power-level congestion pricing the price is dynamically determined during the call according to the network's load

TABLE III
COMPARISON BETWEEN ADMISSION LEVEL AND POWER LEVEL CONGESTION PRICING

Comparison Criterion	Admission-Level Congestion Pricing	Power-Level Congestion Pricing
Price Computation Interval	Every call	Every transmission power unit or decision epoch
Price Duration	Fixed during the call	Varies during the call
Effect of congestion price on handoff calls	No	Yes
Flexibility to respond to congestion	Medium	High
User acceptance	Maybe more acceptable	Requires change in user perception
Billing and Accounting overhead	Medium	High
Price Coverage	Both uplink and downlink	Price for uplink is different from price for downlink
Channel quality conditions of the users	Not considered	Considered
Type of network	Different types of WCN in general	Specific to CDMA-based networks

and the user's power usage. This means that unlike the case in admission-level congestion pricing, handoff calls in power-level congestion pricing are affected by the congestion prices. Since congestion usually occurs in short time periods, power-level congestion pricing is, therefore, more flexible in responding to congestion than admission-level congestion pricing. However, short-term price fluctuations may not be desirable from the user's perspective. This is because users do not like the uncertainty regarding the costs of their calls and, therefore, they may prefer admission-level over power-level congestion pricing. In addition, the short-term price fluctuations in power-level congestion pricing incur significant overhead, since the network operator needs to keep detailed billing and accounting records for every call in which the price changes during it. Admission-level congestion pricing may incur some billing and accounting overhead since the price changes from one call to another but due to the fixed price during the call, this overhead is lower than that of power-level congestion pricing. Furthermore, at the power level, congestion pricing for the uplink is different from that for the downlink due to the different channel characteristics in each direction whereas congestion pricing at admission level includes both the uplink and downlink. Moreover, unlike the pricing schemes at the call admission level, power-level congestion pricing schemes consider the channel quality conditions of the users because power control needs this information for its functionality. Additionally, congestion pricing schemes at power level are specific to CDMA-based networks whereas such schemes at call admission level are usually general and can be implemented on a wide range of different WCN. Table III provides a general comparison between congestion pricing at admission level and power level.

In comparing individual schemes, we define the following criteria (refer to Table IV):

- Objectives: every pricing scheme has a goal besides controlling congestion. For example, some of the schemes surveyed in this paper try to maximize social welfare or revenues.
- Centralized/decentralized: some of the congestion pricing schemes are centralized (i.e., prices are centrally computed at the base station) while others are decentralized (i.e., prices are computed at the base station as well as at the user's side).
- Type of supported QoS: different schemes support different QoS. For example, some schemes are concerned only with maximizing average throughput while others support multiple classes of service.
- Model of user behavior: different schemes have different models for user behaviors. For example, some of them use a demand function and others use a utility function.
- Price initiator: the price in the surveyed schemes is either initiated by the base station/RNC or the users themselves (in case of bidding).
- Congestion measure: different schemes use different congestion measures such as system delay or arrival rate.

To conclude, we highlight some open problems that can be derived from our study of the work described in this paper.

1) Comprehensive QoS support: QoS support is still an open problem. Even though most of admission-level congestion pricing schemes consider the connection-level QoS requirements of the calls, which include new call blocking and handoff call dropping probabilities, they still lack the support for packet-level QoS such as packet delay, packet loss, etc. Power-level congestion pricing schemes also lack the support of packet-level QoS. To provide end-to-end QoS for the users, both connection- and packet-level QoS should be considered in the congestion pricing scheme.
2) User mobility: regretfully, all existing schemes do not consider the effect of congestion prices on user mobility. This may have a great impact on the performance and planning of the network because some cells might become congested in the long-run due to user mobility as a result of high congestion prices in neighboring cells. We are currently investigating the adaptation of existing trip distribution models in order capture the effect of prices on user mobility. Trip distribution is a model of the number of trips that occur between an origin zone and a destination zone. A popular trip distribution model is the gravity model [39], which determines the most probable distribution of the number of trips between two regions depending on the attractiveness of the destination. This model could be used to study the user mobility between two cells depending on their attractiveness (price wise).
3) Call duration: since congestion pricing is based on the assumption that users are price sensitive, it is natural to assume that they will respond to high congestion prices either by postponing their calls until congestion is relieved or by lowering their call durations. Therefore, the effect of price on call duration should be scrupulously investigated.
4) Social fairness: congestion pricing may raise the prices

TABLE IV
COMPARISON BETWEEN DIFFERENT CONGESTION PRICING SCHEMES

Scheme Reference	Objectives	Centralized / Decentralized	Supported QoS	User Behavior	Price Initiator	Congestion Measure
[13]	Maximize social welfare	Centr.	call rejecting probabilities	Exponential demand function [14]	Base station/ RNC	Call arrivals
[15]	Maximize number of admitted users	Centr.	Delay in queues before admission	Exponential demand function [14]	Base station/ RNC	Call arrivals
[16]	Maximize revenues	Centr.	Different bandwidth requirements	WTP through Weibull distribution [17]	Base station/ RNC	Call arrivals/ bandwidth
[20]	Maximize revenues	Centr.	Different new admission probabilities	Exponential demand function [21]	Base station/ RNC	Call arrivals/ network resources
[22]	Maximize revenues	Centr.	None	None	User	System delay
[26], [27]	Maximize revenues	Centr.	Different new admission probabilities	None	User	Call arrivals/ network resources
[28]	Maximize revenues	Centr.	Call admission probability	Exponential demand function [29]	Base station/ RNC	Call arrivals/ communication channels
[33]	Maximize social welfare	Decentr.	Data rates	Utility function of received SIR	Base station/ RNC/user	Uplink interference
[32] (Uplink)	Maximize social welfare	Decentr.	Average throughput	Utility function of average throughput	Base station/ RNC/user	Uplink interference
[32] (Downlink)	Maximize social welfare	Decentr.	Average throughput	Utility function of average throughput	Base station/ RNC/user	Downlink power
[34], [35]	Maximize social welfare	Decentr.	Different data rates requirements	Utility function of received SIR	Base station/ RNC/user	Downlink power
[36], [37]	Maximize social welfare/ maximize revenues	Decentr.	SINR	Step utility function of received SINR	Base station/ RNC/user	Downlink codes and power

of wireless resources (i.e., bandwidth) to very high levels especially during congestion periods. Such prices may not be affordable by many users. Congestion pricing may, therefore, be viewed as promoting social unfairness as fewer people in this case can afford to make connection requests. Hence, social fairness should be carefully taken into consideration when designing congestion pricing schemes.

5) Network stability: congestion pricing can affect the stability of the network. When the network load is low, prices are decreased, which encourages users to demand more resources, thereby increasing the network load and causing congestion. Therefore, the network state changes quickly from underutilization to congestion and vice versa, which causes instability in the offered QoS. This problem is even more aggravated in power-level congestion pricing since prices change at a smaller time intervals than those at admission level. Such instability complicates the tasks of network planning and optimization and QoS provisioning. Hence, solutions need to be designed to efficiently limit the effects of congestion pricing on network instability.
6) Channel quality condition: unlike power-level congestion pricing schemes, none of the surveyed schemes at admission level consider the effect of the channel quality conditions on the price. As aforementioned, the amount of resources (e.g. power, time slots, etc) that are needed to sustain a certain service to the users (e.g. minimum bandwidth guarantees) depend on their channel quality conditions [40]. Hence, this should be taken into consideration when charging users in order to improve the efficiency of the network. However, considering the channel quality conditions of the users as an additional dimension to the congestion pricing problem at admission level is not an easy task. This is because at admission level, the congestion price is fixed during the user's call but the channel quality condition of the user continuously changes depending on his mobility and geographical location. One possible solution that we are currently investigating is on predicting the average channel quality condition of the user at the beginning of his call (i.e., after the user makes a call request) and including the prediction in the price. If the user's actual channel quality during the call is better than the predicted one, then the price of his call will be less than the announced one and, therefore, the user will save money. On the other hand, if the user's actual channel quality during the call is worse than the predicted one, then the user is charged the announced price. In this case, the user is never charged more than the announced price. Existing channel prediction schemes such as the ones in [41]–[43] can be used in our investigation.
7) Demand function: congestion pricing depends, among other factors, on modeling the users' demands in computing prices. Therefore, inaccurate demand modeling can lead to undesirable network performance that not only would affect its functionality but also affect the obtained revenues and the satisfaction of the users. For instance, underestimating the reduction of demand for a certain increase in price would lead to revenue loss and user unsatisfaction. Therefore, it is crucial to model the users' demand behaviors accurately and design congestion pricing schemes that can accommodate different demand models.

8) Next generation WCN: congestion pricing for next generation wireless cellular-based networks such as High Speed Downlink Packet Access (HSDPA) [44], High Speed Uplink Packet Access (HSUPA) [45], Worldwide Interoperability for Microwave Access (WiMAX) [46] and UMTS Long Term Evolution (LTE) [47] is still an interesting open problem. RRM in general, and congestion control in particular, are even more critical in these networks due to the variety of services that they offer and the increased support for bandwidth-intensive multimedia applications such as video on demand and high quality online gaming.
9) Congestion pricing standards and protocols: in addition to the above-mentioned problems, the support of congestion pricing requires new standards and protocols as current wireless standards are not designed with congestion pricing in mind. Moreover, new mobile services and software applications need to be developed for the user's mobile terminals. Such applications and services are required to offer necessary services such as price updates and information gathering to help anticipate the user's utility function. To the best of our knowledge; no work has been done in these two areas.

APPENDIX A
GLOSSARY

- *Auction*: The public sale of a property to the most eligible bidder(s) as determined by the auction method.
- *Bid*: The price the user offers in an auction.
- *Charge*: The amount that is billed for a service.
- *Congestion externality*: The degradation of quality of service that occurs to other users when a certain user transmits when the network is congested.
- *Handoff call*: An active call that moved from one cell to another and is requesting the service of its base station.
- *Handoff call dropping probability*: The probability of dropping a handoff call.
- *Multi-unit discriminatory pricing auction*: An auction method in which the highest bidders win and each bidder pays his own bid.
- *Multi-unit uniform pricing auction*: An auction method in which the highest bidders win and they all pay the clearing price at which the demand exceeds the supply.
- *Multi-unit Vickery auction*: An auction method in which the highest bidders win and they all pay the price of the highest losing bid.
- *New call*: A new call that is requesting to access the network.
- *New call admission probability*: The probability of admitting a new call.
- *New call blocking probability*: The probability of rejecting a new call.
- *Price*: The amount of money associated with a unit of service.
- *Price elasticity of demand*: The change in demand for a certain product or service due to a change in its price.
- *Radio Resource Management*: A set of algorithms to control the usage of radio resources.
- *Social fairness*: The state of economy where the majority of people are able to buy certain products regardless of their incomes. In the context of this paper, it refers to the ability to buy or use network services.
- *Social welfare*: Aggregate utility of people.
- *Tragedy of the commons*: The phenomenon by which greedy users use more than their fair share of a common property to the point of damaging or destroying it.
- *Trip distribution*: A model of the number of trips that occur between an origin zone and a destination zone.
- *User's Willingness to Pay*: The amount of money the users are willing to pay for a certain product or service.
- *Bid shading*: A phenomenon that occurs when users bid below their true valuations of the items being auctioned to avoid subsequent loss of wining when bidding high prices in auctions where users pay the highest bids such as discriminatory auctions.

REFERENCES

[1] G. Hardin, "Tragedy of the commons," *Science Mag*., vol. 162, no. 3859, pp. 1243–1248, June 1968.
[2] C. Courcoubetis and R. Weber, *Pricing Communication Networks: Economic, Technology and Modeling*. John Wiley & Sons, May 2003.
[3] M. Falkner, M. Devetsikiotis, and I. Lambadaris, "An overview of pricing concept for broadband ip networks," *IEEE Commun. Surveys & Tutorials*, vol. 3, no. 2, pp. 2–13, April 2000.
[4] R.-F. Liao, R. Wouhaybi, and A. Campbell, "Wireless incentive engineering," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 10, pp. 1764 – 1779, dec. 2003.
[5] H. Varian, *Intermediate Microeconomics: A Modern Approach*, 7th ed. W.W. Norton & Company, December 2005.
[6] J. MacKie-Mason and H. Varian, "Pricing congestible network resources," *IEEE J Sel. Areas Commun.*, vol. 13, no. 7, pp. 1141 –1149, sep 1995.
[7] A. Lazar and N. Semret, "Auctions for network resource sharing," Columbia University, New York, U.S.A., CTR Technical Report 468-97-02, February 1997.
[8] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *J. Automatica*, vol. 35, no. 12, pp. 1969–1985, December 1999.
[9] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, January 1997.
[10] A. K. M. F. Kelly and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness, and stability," *J. Operational Research Society*, vol. 49, no. 3, pp. 237–252, March 1998.
[11] L. A. DaSilva, "Pricing for qos-enabled networks: A survey," *IEEE Commun. Surveys & Tutorials*, vol. 3, no. 2, pp. 2 –8, April 2000.
[12] J. Walrand and P. Varaiya, *High-Performance Communication Networks*, 2nd ed. Morgan Kaufmann, January 2000.
[13] J. Hou, J. Yang, and S. Papavassiliou, "Integration of pricing with call admission control to meet qos requirements in cellular networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 13, no. 9, pp. 898 – 910, sep 2002.
[14] P. Fishburn and A. Odlyzko, "Dynamic behavior of differential pricing and quality of service options for the internet," in *Proc. ACM International Conference on Information and Computation Economies (ICE)*, Charleston, U.S.A, October 1998, pp. 128–139.
[15] S. Yaipairoj and F. Harmantzis, "Congestion pricing with alternatives for mobile networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 4, Atlanta, U.S.A, March 2004, pp. 671–676.
[16] S. Hew and L. B. White, "Optimal integrated call admission control and congestion pricing with handoffs and price-affected arrivals," in *Proc. Asian-Pacific Conference on Communications (APCC)*, Perth, Australia, October 2005, pp. 396–400.
[17] B. Dodson, *The Weibull Analysis Handbook*, 2nd ed. ASQ Quality Press, April 2006.
[18] P. Bremaud, *Markov Chains*, 1st ed. Springer, January 2001.
[19] D. White, *Markov Decision Processes*, 1st ed. John Wiley & Sons, January 2001.

[20] S. Mandal, D. Saha, and A. Mahanti, "A technique to support dynamic pricing strategy for differentiated cellular mobile services," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 6, December 2005, pp. 3388–3392.
[21] E. D. Fitkov-Norris and A. Khanifar, "Congestion pricing in cellular networks, a mobility model with a provider-oriented approach," in *Proc. IEEE International Conference on 3G Mobile Communication Technologies (3G)*, London, UK, March 2001, pp. 63–67.
[22] S. Yaipairoj and F. Harmantzis, "Auction-based congestion pricing for wireless data services," in *Proc. IEEE International Conference on Communications (ICC)*, Istanbul, Turkey, June 2006, pp. 1059–1064.
[23] A. Z. Dodd, *The Essential Guide to Telecommunications*, 4th ed. Prentice Hall PTR, June 2005.
[24] W. Vickery, "Counter speculation, auctions, and competitive sealed tenders," *J. Finance*, vol. 16, no. 1, pp. 8–37, 1961.
[25] H. Kaaranen, A. Athtiainen, L. Laitinen, S. Naghian, and V. Niemi, *UMTS Networks: Architecture, Mobility and Services*, 2nd ed. John Wiley & Sons, April 2005.
[26] D. S. S. Mandal and M. Chatterjee, "Pricing wireless network services using smart market models," in *Proc. IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, U.S.A., January 2006, pp. 574–578.
[27] S. Mandal, D. Saha, and M. Chatterjee, "Dynamic price discovering models for differentiated wireless services," *J. Communications*, vol. 1, no. 5, pp. 50–56, August 2006.
[28] E. Viterbo and C. Chiasserini, "Dynamic pricing for connection-oriented services in wireless networks," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, San Diego, September 2001, pp. 68–72.
[29] L. Taylor, *Telecommunications Demand in Theory and Practice*, 1st ed. Kluwer Academic Publishers, January 1994.
[30] F. A. Y. Park, *Enhanced Radio Access Technologies for Next Generation Mobile Communication*, 1st ed. Springer, February 2007.
[31] A. Goldsmith, *Wireless Communications*, 1st ed. Cambridge University Press, August 2005.
[32] V. Siris, "Resource control for elastic traffic in cdma networks," in *Proc. ACM International Conference on Mobile Computing and Networking (MOBICOM)*, Atlanta, U.S.A., September 2002, pp. 193–204.
[33] S. Han and Y. Han, "A simple congestion pricing in wireless communication," in *Proc. IEEE Vehicular Technology Conference (VTC)*, Dallas, U.S.A, September 2005, pp. 795–798.
[34] J. Lee, R. Mazumdar, and N. Shroff, "Downlink power allocation for multi-class cdma wireless networks," in *Proc. IEEE Joint Conference of Computer and Communications Societies (INFOCOM)*, New York, U.S.A, June 2002, pp. 1480–1489.
[35] ——, "Downlink power allocation for multi-class cdma wireless systems," *IEEE/ACM Trans. Netw.*, vol. 13, no. 4, pp. 854–867, August 2005.
[36] P. Liu, M. Honig, and S. Jordan, "Forward-link cdma resource allocation based on pricing," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 3, Chicago, U.S.A., September 2000, pp. 1410–1414.
[37] P. Liu, P. Zhang, S. Jordan, and M. Honig, "Single-cell forward link power allocation using pricing in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 2, pp. 533 – 543, march 2004.
[38] P. L. P. Zhang, S. Jordan and M. Honig, "Power control of voice users using pricing in wireless networks," in *Proc. of SPIE ITCOM Conference on Modeling and Design of Wireless Networks*, Denver, U.S.A., August 2001, pp. 155–165.
[39] A. Wilson, "A statistical theory of spatial distribution models," *J. Transportation Research*, vol. 1, pp. 253–269, 1967.
[40] B. Al-Manthari, H. Hassanein, and N. Nasser, "Packet scheduling in 3.5G high-speed downlink packet access networks: breadth and depth," *IEEE Network*, vol. 21, no. 1, pp. 41 –46, jan.-feb. 2007.
[41] A. Duel-Hallen, S. Hu, and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Processing Mag.*, vol. 17, no. 3, pp. 62 –75, may 2000.
[42] R. Vaughan, P. Teal, and R. Raich, "Short-term mobile channel prediction using discrete scatterer propagation model and subspace signal processing algorithms," in *Proc. IEEE Vehicular Technology Conference-Fall (VTC)*, September 2000.
[43] T. Ekman, "Prediction of mobile radio channels, modeling and design," Ph.D. Dissertation, Uppsala University, Sweden, 2002.
[44] "High speed downlink packet access (HSDPA), overall description," March 2003, 3GPP TS 25.308, Release 5.
[45] "FDD enhanced uplink, overall description," December 2005, 3GPP TS25.309, Release 6.5.0.
[46] "Ieee 802.16-2005e standard for local and metropolitan area networks: Air interface for fixed broadband wireless access systems - amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands," *IEEE Std 802.16f-2005 (Amendment to IEEE Std 802.16-2004)*, December 2005, IEEE 802.16 Working Group.
[47] "Evolved universal terrestrial radio access (utra) and universal terrestrial radio access network (utran) radio interface protocol aspects," November 2005, 3GPP, TS25.813.

Bader Al-Manthari (M06) received his B.Sc with Honors, M.Sc. and Ph.D. from Queen's University, Kingston, Canada in 2004, 2005, and 2009, respectively. He is currently working as an information security specialist at the Center of Information Security, Information Technology Authority in the Sultanate of Oman. His research interests include information security, economic-based radio resource management in next generation wireless cellular networks, wireless ad hoc and sensor networks, performance evaluation of communication protocols and schemes.

Nidal Nasser (M00) received his B.Sc. and M.Sc. degrees with Honors in Computer Engineering from Kuwait University, State of Kuwait, in 1996 and 1999, respectively. He completed his Ph.D. in the School of Computing at Queen's University, Kingston, Ontario, Canada, in 2004. He is currently an Associate Professor in the Department of Computing and Information Science at University of Guelph, Guelph, Ontario, Canada. He has authored several journal publications, refereed conference publications and seven book chapters. He has also given tutorials in major international conferences. He is an associate editor of the Journal of Computer Systems, Networks, and Communications, Wiley's International Journal of Wireless Communications and Mobile Computing and Wiley's Security and Communication Networks Journal. He has been a member of the technical program and organizing committees of several international IEEE conferences and workshops. Dr. Nasser is a member of several IEEE technical committees. He received Fund for Scholarly and Professional Development Award in 2004 from Queen's University. He received the Best Research Paper Award at the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'08) and at the International Wireless Communications and Mobile Computing Conference (IWCMC'09).

Hossam Hassanein is with the School of Computing at Queen's University working in the areas of broadband, wireless and variable topology networks architecture, protocols, control and performance evaluation. Dr. Hassanein obtained his Ph.D. in Computing Science from the University of Alberta in 1990. He is the founder and director of the Telecommunication Research (TR) Lab http://www.cs.queensu.ca/ trl in the School of Computing at Queen's. Dr. Hassanein has more than 350 publications in reputable journals, conferences and workshops in the areas of computer networks and performance evaluation. He has delivered several plenary talks and tutorials at key international venues, including Unconventional Computing 2007, IEEE ICC 2008, IEEE CCNC 2009, IEEE GCC 2009, IEEE GIIS 2009, ASM MSWIM 2009 and IEEE Globecom 2009. Dr. Hassanein has organized and served on the program committee of numerous international conferences and workshops. He also serves on the editorial board of a number of International Journals. He is a senior member of the IEEE, and is currently chair of the IEEE Communication Society Technical Committee on Ad hoc and Sensor Networks (TC AHSN). Dr. Hassanein is the recipient of Communications and Information Technology Ontario (CITO) Champions of Innovation Research award in 2003. He received several best paper awards, including at IEEE Wireless Communications and Network (2007), IEEE Global Communication Conference (2007), IEEE International Symposium on Computers and Communications (2009), IEEE Local Computer Networks Conference (2009) and ACM Wireless Communication and Mobile Computing (2010). Dr. Hassanein is an IEEE Communications Society Distinguished Lecturer.