An MDP-based Dynamic Pricing Scheme for Revenue Maximizing in Wireless Networks

Asim Rasheed
INMC, School of EECS
Seoul National University
Email: asim@netlab.snu.ac.kr

Sung-Guk Yoon
INMC, School of EECS
Seoul National University
Email: sgyoon@netlab.snu.ac.kr

Saewoong Bahk
INMC, School of EECS
Seoul National University
Email: sbahk@snu.ac.kr

***Abstract*—Future communication networks are envisioned to be based upon a new communication paradigm which is user centric, i.e., users have freedom to choose a service provider to maximize their utilities. Pricing schemes adopted by service providers will impact the decision of users in selecting networks. In this paper, we formulate the problem as a Markov Decision Process (MDP) framework and apply Q-learning to design an optimal pricing policy that aims to maximize the long term revenue of the service provider. By applying dynamic pricing, each service provider operates an optimal policy to maximize its revenue. Simulation results show that our proposed framework is successful in maximizing the service provider's revenue while supporting an appropriate level of user satisfaction in terms of price and call level quality of service.**

Keywords-revenue maximization; MDP; dynamic pricing; Q-learning; service provider

I. Introduction

Network heterogeneity is a common feature of 4G where multiple access networks will coexist to support ubiquitous wireless services. These heterogeneous wireless networks typically differ in terms of coverage, data rate, latency and loss rate [1]. Moreover with the advent of sophisticated mobile devices and terminals, mobile users could subscribe service providers for a short duration against a current scenario of contractual agreement. This scenario is referred to as Always Best Connected (ABC) [2]. In such highly competitive environments, pricing will be one of the most important issues. Pricing schemes adopted by service providers will significantly impact the decision of each mobile user in selecting a network. Through an appropriate pricing policy, the service provider can maximize its revenue, while mobile users achieve high satisfaction from the chosen service providers.

In the current scenario, service providers follow a static charging model, i.e., a flat charging model based upon per minute duration or for a specific subscription period during which users can have limited or unlimited usage of services. This charging model is independent of the current state of the network and does not consider any engineering issues related with the pricing mechanism that results in revenue loss. However, in such competing environments, service providers have the potential to increase their revenue using customized prices designed according to users' satisfaction and personal benefit. Therefore, a dynamic pricing scheme plays a very important role in achieving the goal of revenue maximization.

Most of the previous studies consider the pricing and revenue maximization problem from two perspectives: competitive and non-competitive [3]. In a competitive pricing model, competition between service providers or between a service provider and users is established and a game theoretic approach is used to model this type of competition. This approach usually runs in either a cooperative or non-cooperative manner.

This paper considers a dynamic pricing problem for revenue maximization under the framework of Markov Decision Process (MDP). We apply reinforcement learning to learn the system gradually from user's behavior, and propose an optimal pricing policy that dynamically adjusts the offered price of the service provider. A solution that achieves the satisfaction of all entities is desirable.

The contributions of this paper are three-fold: i) it considers dynamic pricing for revenue maximization, ii) it investigates the problem under the framework of MDP and applies reinforcement learning, iii) it determines an optimal policy to adjust a service provider's offered price dynamically to increase the total revenue for the service provider.

The rest of this paper is organized as follows. Section II presents the related work. The system model and assumptions are provided in Section III. We formulate the problem in Section IV, and introduce Q-learning algorithm in Section V. Simulation results for performance comparison are provided in Section VI. Future work and concluding remarks are given in Section VII.

II. Related Work

Under assumptions of competitive wireless networks, there are several recent studies [1], [4]–[7]. Most of the previous works focused on determining the existence of unique equilibrium solution, e.g., Nash or Stackelberg equilibrium. The authors in [4] formulated an evolutionary stochastic game and determined the Nash equilibrium (price). Under the same

This work was supported by the IT R&D program of MKE/KEIT. [KI001809, Intelligent Wireless Communication Systems in 3 Dimensional Environment]

This research work is sponsored by "Higher Education Commission (HEC), Govt. Of Pakistan" under the scholarship program titled: MS Level Training in Korean Universities/Industry.

assumptions, the study of Lin and Das in [5] showed that either a dominant strategy for service provider or Nash equilibrium exists. Their proposed framework significantly increases service provider's revenue. They pointed that pricing must be included and should be dynamic in nature, but they did not consider the pricing effect.

Under the same assumptions, the authors in [7] developed a new admission control and a scalable pricing policy using non-cooperative game theory by incorporating sigmoid utility[1]. Backward induction was used to solve the game and Nash equilibrium was achieved within the sub game. However, they did not propose changing in current pricing policy and also pointed that the improvement in the current pricing scheme is essential. We also use the same user satisfaction sigmoid utility proposed in [7]. However, our objective is different and we formulate the problem differently using an MDP framework and Q-learning is applied to determine an optimal policy that aim at maximizing the network revenue by incorporating the dynamic pricing.

Dynamic pricing under framework of game theory was studied in [1] and the references there in. However, all these studies were based on determining Nash equilibrium (price). To guarantee call level and packet level quality of service (QoS), many QoS metrics were proposed in [8] under analytical framework. MDP and Q-learning framework is widely used in wireless networks to study call admission control, wireless resource management, cell configuration, handoff under heterogeneous network and spectrum access schemes [9]–[12].

III. System Model and Assumptions

We consider a wireless cellular scenario where multiple base stations (BSs) operate in a common service area and manage their independent pricing policy. These BSs can belong to single or different service providers as shown in Fig. 1. We consider a distributed approach and focus our study for a single BS case. Service providers charge each ongoing call with price $p_b \in P$ based on the current condition of its network. Because of the lack of information about the performance and payoff a service provider obtain at certain time, service providers have to learn gradually and change its decision. A price announced by a BS at any time greatly impacts the users to initiate connections. We assume that users' arrival rate is a function of price offered by service provider and normal price. If the price offered by the service provider is higher than the normal price, the connections arrival rate will significantly decrease and if the price is less than the normal price, the connection arrival rate will increase significantly. Therefore, user's connection rate can be modeled as [4].

$$\bar{\lambda} = \lambda \left(exp \left(-(\frac{p_b}{P_0} - 1) \right) \right)^2 . \quad (1)$$

where p_b is the price announced by service provider $b \in B$ and P_0 is the normal price (average offered price in the common service area). Service providers adopt dynamic pricing mechanism for a possible control on QoS.

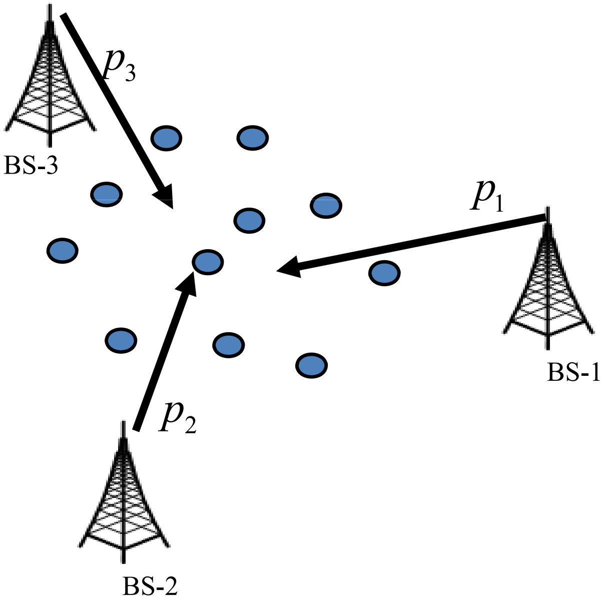

Fig. 1. Different BSs belong to different service providers or single service provider has overlaid coverage area and manages its prices p_b.

¿From the users point of view, who are under a common service of multiple BSs, users attachment with a service provider is per connection duration only and users are certain price and QoS sensitive. We model user's sensitivity in terms of user's satisfaction, which describes that how sensitive the users are to changes in price and QoS. Degradation in the QoS will cause the users to churn from serving service provider. These users's sensitivity is modeled in section IV.

IV. Problem Formulation

We formulate the dynamic pricing and revenue maximization problem as a Markov Decision Process (MDP) [13]. The notable solution variants are value iteration, policy iteration and linear programming. However, these classical and model based solutions suffer from "curse of dimensionality" and modeling as they require prior knowledge of state transition probabilities. These solutions are ineffective in large scale and complex problems. To tackle these problems, we use reinforcement learning approach which is even suitable when the state space become computationally intractable. So, we propose Q-Learning based solution to determine an optimal pricing policy for the revenue maximization in wireless networks.

A. Problem Formulation as a MDP

A finite state system in a generic network is considered where the system changes its state with each arrival and departure of on-going calls. BS periodically adjusts its offered price to adopt changes in network conditions. We assume that each state is a Markovian state that is independent of the past

[1] A user satisfaction function used to model user satisfaction in terms of price and packet blocking probability.

history of the system and retains all the relevant information of the system, and that an action is chosen based on the current state of the network. With each state change the network earns a reward. The objective is to optimize the action (offered price) that maximize the long term reward of the network. The detail problem formulation is as follows.

•State-space: The network is represented by a discrete time system with a finite number of feasible states identified by

$$S = \sum_{c} n_c(t)\delta_c \leq C. \tag{2}$$

where n_c is the number of connections of class-c at time t, δ_c is bandwidth requirement of class-c user, and C is the capacity of system. Connections initiation requests follow Poisson distribution with mean arrival rate λ from (1) and call holding time is exponentially distributed with mean $1/\mu$.

•Action space: The feasible pricing range offered by the network is modeled as a feasible action space. At each decision epoch the state of the wireless network will change and according to its new state the network announces a new price among feasible offered price set. The action space in a state s is

$$A^s = P = [p_1, p_2, \cdots p_n]. \tag{3}$$

•Decision epoch: In wireless network, a new terminal initiate or terminate a connection the state of the network changes accordingly, the natural decision epoch is a call arrival or departure instance. Therefore, the service provider adjusts the price at each decision epoch.

•Reward function: Because our proposed scheme is user centric, we propose an accumulated reward as the sum of user satisfactions. If a function contributes positively [negatively] towards user satisfaction, it will be added [subtracted] for the accumulated reward function. Let $r(s,a)$ denote the immediate reward the system get if $a \in P$ is chosen in a system state s, and we have

$$r(s,a) = \begin{cases} \sum_{u=1,2\cdots U} y_u(u_u^+ - u_u^-) & \text{, system stable} \\ 0 & \text{, otherwise.} \end{cases} \tag{4}$$

We assume that the system is stable under a scheduling and call admission control. In (4), u_u^+ $[u_u^-]$ contribute positively [negatively] towards the user satisfaction and y_u is the weight associated with each utility function. As the system behavior is the same for all ongoing connections, any unstable situation means all ongoing connections experience the same situation and that turn the immediate reward into zero. For example, a threshold based QoS user satisfaction function contributes negative towards accumulated reward. When the QoS offered by the network exceeds the predefined threshold, all the users will churn from that network and the resulting reward will be zero.

The sigmoid function has been widely used to capture user satisfaction [1], [5], [7] and the references there in. We use the same sigmoid utility proposed in [7]. The positive and negative user satisfactions as functions of offered price p is given by

$$u_1^+(p) = \begin{cases} \frac{1}{1+e^{-L_c(P_0^c - p)}} & \text{, } p \leq P_0^c \\ 0 & \text{, otherwise} \end{cases} \tag{5}$$

and

$$u_1^-(p) = \begin{cases} \frac{1}{1+e^{L_c(P_0^c - p)}} & \text{, } p > P_0^c \\ 0 & \text{, otherwise,} \end{cases} \tag{6}$$

respectively.

In the function, L_c is a constant which represents the steepness of these functions for class-c users, P_0^c is the normal price and u_1^+ $[u_1^-]$ shows this function contributes positively [negatively] towards accumulated reward of the system.

To guarantee the packet level QoS for ongoing connections, we model user churning behavior through a packet delay. We assume that each ongoing call has certain QoS threshold requirement represented by d_{max}^c for class-c user. A degradation in the offered QoS will increase the churning rate from the serving service provider. The packet delay is calculated simply by M/M/1 queuing analysis [14]. So, we have the second user satisfaction function as

$$u_2^-(d) = \frac{1}{1 + e^{K_c(d_{max}^c - d)}}. \tag{7}$$

As a consequence, the immediate reward $r(s,a)$ obtained in a state $s \in S$ with an action $a \in P$ is

$$r(s,a)_{a \in P} = [y_1 u_1^+(p) - \{y_1 u_1^-(p) + y_2 u_2^-(d)\}], \tag{8}$$

where y_1 and y_2 is the weight associated with user satisfaction functions.

V. Q-Learning

Q-Learning [15], [16] is a reinforcement learning technique for solving MDPs when the state transition probabilities are not known. This technique works by directly learning MDPs' action value function by interacting with control environment. In an MDP, the value function is a utility that gives the expected utility of taking an action in given state. Accordingly, if the value function is learned, the optimal policy is simply the set of actions that maximizes the function at each state. Q-learning provides the following update rule to successively approximate the value function $Q(s,a)$ referred as Q-function. The update is taken by

$$Q_{t+1}(s_t, a_t) = (1-\alpha)Q_t(s_t, a_t) + \alpha[r(s_t, a_t) + \gamma V^*(s_{t+1})], \tag{9}$$

where $\alpha \in [0,1)$ is the learning rate showing what extent the newly acquired information will override the old values and $\gamma \in (0,1]$ is the discount factor that accounts for importance of future reward. $r(s_{t+1})$ is the reward received at time step t and $V^*(s_{t+1})$ is the value function that maximizes the Q-function at state s_{t+1} over all actions a. That is,

$$V^*(s_{t+1}) = \max_{a_t \in P} Q_{t+1}(s_t, a_t). \tag{10}$$

Since our objective is to determine an optimal policy that maximize the long term revenue of service provider, the optimal policy is given by

$$Q_t^*(s_t, a) = \max_{a \in P} Q_t(s_t, a). \tag{11}$$

TABLE I
SIMULATION SETTING

Parameters	value
Number of mobile users	100
Connection arrival rate λ in (1)	1-18/min
Mean connection holding time μ	1min
Weighting factor y_1	10
Weighting factor y_2	30
Max tolerable delay d_{max}	0.05sec
Normal price P_0^c	1.6
Static charging price	1.3
Offered Price p	1.0-2.0units/min
Learning rate α	0.6
Discount factor γ	0.9

The optimal policy is the action with the best Q-value in each state. Since a greedy policy causes the system to converge to a locally optimal solution, where the behavior of the greedy scheme can be sub-optimal because its decisions are based on maximization of local reward. Therefore, it is necessary to visit all the sets of possible actions for all states to find the globally optimum solution. This is the "exploration / exploitation dilemma" [10]. An action of state is selected from the feasible action set using an exploitation and exploration policy.

VI. SIMULATION RESULTS

In this section, simulation results are presented to illustrate the performance of our proposed scheme.

We consider a single hot spot cell with varied traffic load. The traffic load varies according to (1). The mobile users are uniformly distributed and the new connection arrival rate follows Poisson distribution and connection duration is exponentially distributed with mean 1 min. We assume that users generate persistent data traffic and always have data to send for the connection period. Data packets generated by each user are aggregated at the system and the system is able to support aggregated generated traffic (i.e., stable) under a scheduling and call admission control scheme. Only single class real time traffic with a delay constraint 0.05 sec is considered. The other parameter settings are described in Table I. The simulation is run under varied connection arrival rate and the results are averaged over 20 runs. For the purpose of comparison, we consider a greedy scheme that always considers the maximum possible reward at any decision epoch and does not consider the long term effect of the policy for revenue maximization.

First, we vary the connection arrival rate and the result is shown in Fig. 2. The decision is made at each decision epoch, i.e., whenever the state of the network change on connection arrival and departure instance. Therefore, the price offered at each decision epoch is averaged over all states for each arrival rate. Mean price offered by the service provider decreases [increases] with the decrease [increase] of the arrival rate. When the connection arrival rate is small, the service provider should set low price to attract more users for initiating connections. However, for moderately large connections arrival rate the service provider can set high price to maximize its revenue. Another reason to set high price is to restrict higher arrival rate, and it grantees QoS for ongoing connections.

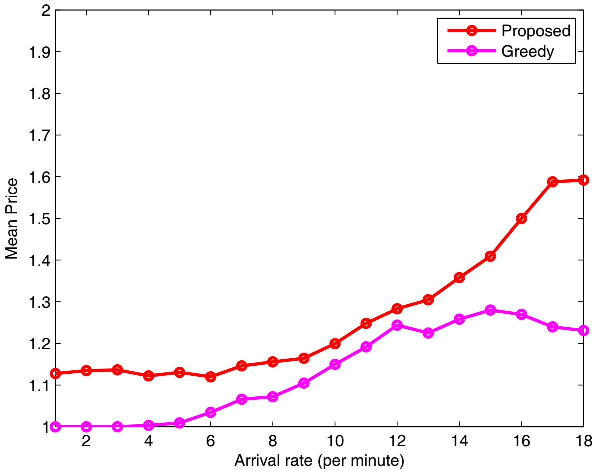

Fig. 2. Mean price according to arrival rate.

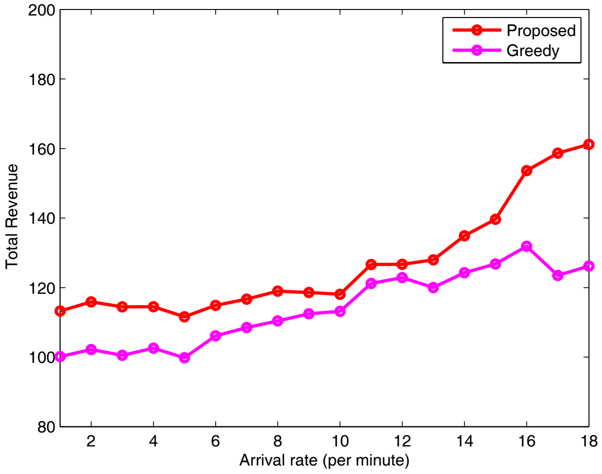

Fig. 3. Total network revenue according to arrival rate.

Fig. 3 shows the revenue of the service provider. Our proposed scheme dramatically improves the revenue compared to the greedy scheme. Since the service provider choose a higher price for higher arrival rate, the revenue is increased. The comparison is based on the total revenue generated from all connections admitted into the network, while Fig. 4 shows the service provider's average revenue earned per user. Because total revenue is increased, average revenue earned per user is also increased.

VII. CONCLUSION AND FUTURE WORK

In this paper, we studied the problem of dynamic pricing and revenue maximization under the framework of Markov Decision Process (MDP). Because of highly competitive environments, static pricing schemes do not work properly especially with high traffic load. We used the reinforcement learning to

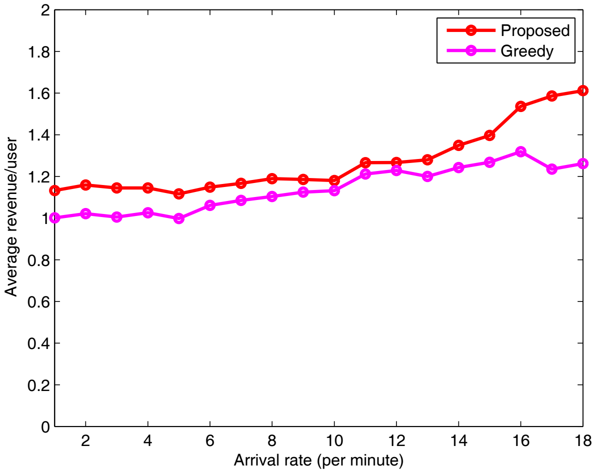

Fig. 4. Average revenue per user according to arrival rate.

learn the system gradually and Q-learning algorithm to design an optimal policy that maximizes the total network revenue under a dynamic pricing scheme. Simulation results show how the charged mean price of the service provider varies according to the arrival rate, and confirm that the total network revenue obtained by the service provider is higher under the dynamic pricing scheme.

As a part of the future work, we would like to extend our research to consider the stochastic game modeling and to study its behavior with correlated-Q learning and Nash equilibrium under competitive environments.

References

[1] S. Sengupta, S. Anand, M. Chatterjee, and R. Chandramouli, “Dynamic Pricing for Service Provisioning and Network Selection in Heterogeneous Networks,” *Physical Communication*, vol. 2, issues 1-2, pp. 138-150, Mar./Jun. 2009.

[2] E. Gustafsson and A. Jonsson, “Always Best Connected,” *IEEE Wireless Communications*, vol. 10, issue 1, pp. 49 - 55, Feb. 2003.

[3] D. Niyato and E. Hossain, “Competitive Pricing in Heterogeneous Wireless Access Networks: Issues and Approaches,” *IEEE Network*, vol. 22, no. 6, pp. 4-11, Nov.-Dec. 2008.

[4] D. Niyato and E. Hossain, “Modeling User Churning Behavior in Wireless Networks Using Evolutionary Game Theory,” in Proc. *IEEE WCNC*, Las Vegas, USA, 31 Mar.-Apr. 2008.

[5] H. Lin, M. Chatterjee, S. K. Das, and K. Basu, “ARC: An Integrated Admission and Rate Control Framework for Competitive Wireless CDMA Data Networks Using Noncooperative Games,” *IEEE Transactions on Mobile Computing*, vol. 4, no. 3, pp. 243-258, May/Jun. 2005.

[6] S. K. Das, H. Lin, and M. Chatterjee, “An Econometric Model for Resource Management in Competitive Wireless Data Networks,” *IEEE Network*, vol. 18, no. 6, pp. 20-26, Nov./Dec. 2004.

[7] A. N. Rouskas, A. A. Kikilis, and S. S. Ratsiatos, “A Game Theoretical Formulation of Integrated Admission and Pricing in Wireless Networks,” *European Journal of Operational Research* , vol 191, no. 3, pp. 1175-1188, 2008.

[8] D. Niyato and E. Hossain, “A Novel Analytical Framework for Integrated Cross-layer Study of Call-level and Packet-level QoS in Wireless Mobile Multimedia Networks,”*IEEE Transactions on Mobile Computing*, vol. 6, no. 3, pp. 322-335, Mar. 2007.

[9] A. Pietrabissa, “An Alternative LP Formulation of the Admission Control Problem in Multiclass Networks,” *IEEE Transactions on Automatic Control*, vol. 53, issue. 3, pp. 839-845, Apr. 2008.

[10] C. Laio, F. Yu, V. Leung, and C. Chang, “A Novel Dynamic Cell Configuration Scheme in Next-Generation Situation-aware CDMA Networks,” *IEEE JSAC*, vol 24, no. 1, pp. 1825- 1829, Jan. 2006.

[11] Husheng Li, ”Multi-agent Q-learning of Channel Selection in Multi-user Cognitive Radio Systems: A Two by Two case,” in Proc. *IEEE SMC*, San Antonio, USA, Oct. 2009.

[12] S. N. Enrique, Y. X Lin, and W. S. W. Vincent,“An MDP-based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks,” *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1243-1254, Mar. 2008.

[13] M. L. Putterman, *Markove Decision Process: Discrete Stochastic Dynamic Programming*, New York: Wiley, 1994.

[14] D. P. Bertsekas, *Data Networks*, NJ:Prentice-Hall, 1987.

[15] V. Bui and W. Zhu, “A Game Theoretic Framework for Multipath Optimal Data Transfer in Multiuser Overlay Networks,” in Proc. *IEEE ICC*, Beijing, China, May 2008.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, USA: The MIT Press, 1999.