

# Priority Telephony System with Pricing Alternatives

Saravut Yaipairoj      Prof. Fotios Harmantzis, Ph.D.

Stevens Institute of Technology, Castle Point on the Hudson, Hoboken, NJ 07030, USA  
{syaipair, fharmant}@stevens.edu

**Abstract.** Dynamic pricing schemes in telecommunication networks were traditionally employed to create users' incentives in such a way that the overall utilization is improved and profits are maximized. However, such schemes create frustration to users, since there is no guarantee that they would get services at the anticipated prices. In this paper, we propose a pricing scheme for priority telephony systems that provides alternatives to users. Users can choose between a) a dynamic price scheme that provides a superior quality of service or b) a fixed low price with acceptable performance degradation. Our results verify that the proposed pricing scheme improves the overall system utilization and yet guarantees users' satisfaction.

## 1. Introduction

In long-haul communications, network resources are critical commodities that require an efficient allocation mechanism to users. In recent years, researchers have been focusing on resource allocation schemes, so that network resources can be utilized efficiently and total profit from resource usage is maximized. However, it is well known that network users act independently and sometimes "selfishly", regardless of the current network traffic conditions. Therefore, even with advanced resource allocation schemes in place, it is hard to avoid congestion. As a result, congestion reduces the total system utilization. Mechanisms that give users incentives to behave in ways that improve the overall utilization and performance of the network are needed. In commercial networks, pricing had been proved an effective mean to resolve the problem of scarce resource allocation.

Network users are inherently price sensitive. Via prices, the network could send signals to the users, providing them with incentives that influence their behavior [1]. Pricing thus becomes an effective mean to perform traffic management and congestion control. Such schemes are known as *dynamic pricing schemes*. In a dynamic pricing scheme, call prices change as demand fluctuates [2]. It rises in accord with demand, deterring additional users from accessing the network or holding network resources for long periods during congestion time. Therefore, such schemes create users' incentive for efficient network utilization. In addition, during the off-peak hours, the price drops from its nominal level; this will serve as an incentive to generate more traffic to an otherwise under-utilized network [3].

However, despite the beneficiary of dynamic pricing, it has major drawbacks. Dynamic pricing schemes create frustration to users. Since price fluctuates according to demand, there is no guarantee to final charges. Users with low price expectations would risk being blocked, during congestion periods.

In this paper, we propose a middle ground where users have choices in the way they are priced: they can either accept a) a dynamic pricing scheme, where prices changes according to the system congestion levels, or b) a fixed pricing scheme, where the provider charges a low price, with users experiencing an acceptable performance degradation.

The paper is organized as follows: In Section 2, we describe the traditional dynamic pricing used in a Priority Telephony System. In Section 3, we present our priority queuing system, where the appropriate parameters are defined. Section 4 shows numerical results and how our proposed pricing scheme can improve the call admission control mechanism of the network. Discussion on the results is also taking place in that section. In Section 5, we draw the conclusion of our work.

## 2. Dynamic Pricing in Priority Telephony System

In telephony networks, whenever congestion occurs, the incoming telephone calls can either be blocked from the system or placed into a buffer (queue), waiting to be served whenever the telephone trunks are free. In the latter case, the Quality of Service (QoS) metric used for measuring the performance of the system is the delay that users experience in the queue. The shorter the time users spend in the queue, the better for them.

In queuing networks, users experience delay according to their priority agreement with the system, which can be described by a Priority Queuing Model. In a priority queuing system, users who require more attention are distinguished from those who can endure the quality of conventional services. Usually, the QoS required by priority users is higher and therefore should be served faster than the average (conventional) users. The price charged to priority users is clearly higher.

Currently, the service charge for telephone users is either fixed per call or flat. One of the advantages of these schemes is the simple billing and accounting processes [4, 5]. However, since users act independently and sometimes in a "selfish" manner, they utilize the system regardless of its traffic condition. Such pricing schemes do not provide incentives for users

to avoid congestion during peak hours and cannot react effectively to the dynamics of the network. With dynamic pricing schemes, prices change depending on the network conditions. Users who require access to the network during peak hours and are able to afford higher prices, will be admitted to the network, while users who are not able to afford such prices are blocked. However, we argue that blocked users during congestion time (even though their pricing requirements for being prioritized are not met) would result in reality to highly dissatisfied users. By using a queue to delay, instead of block, call requests during time of congestion, it is likely that users would be more satisfied and it can potentially yield to a better network operation.

### 3. Model for Priority Queue with Priority Call Admission Control

Call admission control (CAC) is widely used as an effective mean to prevent overloading in telecommunication networks. According to our model, during system congestion time, admitted calls are required to meet a certain pricing requirement. In this context, a new type of CAC is introduced here, namely, the *Priority Call Admission control* (PCAC). PCAC's main function is to control the amount of incoming calls, based on users' priorities that are regulated by pricing criteria.

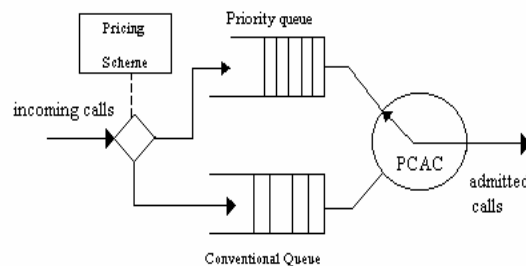


Fig. 1. Priority Queue with Pricing Alternatives

As shown in Fig 1, the system consists of two types of queues: one for the priority users and the other for the conventional ones. Both queues store incoming call requests and feed them to the telephone trunk group. The key elements in this system model are two functional blocks: the pricing block and the PCAC block. The pricing block acts as a price broadcasting point to incoming users. In this particular system, users who refuse to pay a higher price are placed in the queue for conventional users and wait until get served. Priority users who meet the higher price requirement, are placed into the priority queue, which would be served faster than the conventional one. The above procedure will only take place when the system experiences congestion. If the congestion level is not met, all calls will be placed in the conventional queue and served as soon as the system is ready.

The PCAC block can be characterized as a QoS controller that allocates system resources according to call requests coming from both queues. Since priority users require more attention than conventional users, the priority queue is served by the PCAC block in such a way that certain QoS is met. At the same time, conventional users are also served by PCAC with a QoS that is obviously inferior to the priority users. The objective of this system is to adjust system resources in such a way that we can meet the QoS constraints of both queues and maximize the number of calls being served by the system.

During peak hours, users who attempt to access the system will find themselves facing two choices: One is to accept a high price according to dynamic pricing theory, as they will enjoy the higher QoS of priority callers. The second choice is to deny the high price and be charged by a fixed low pricing scheme. As a result, in the second case, users will experience longer delays before being served, depending on the existing traffic conditions. The call procedure of the system can be described as follows:

#### Call procedure

1. Users dial in numbers and wait for a system response.
2. The current status of the system is identified. If the system is not congested, the call requests will be placed in the conventional queue waiting for available trunks.
3. If the system is congested, the system will notify users the approximated time they have to wait for service. Then, it announces the price for those users who consider priority status and ask for their choice (be prioritized or stay on the line).
4. If the answer is positive, the users' requests for call are placed in the priority queue where users are served with superior QoS.
5. For those who stay on the line, their call requests will be placed in conventional queues.

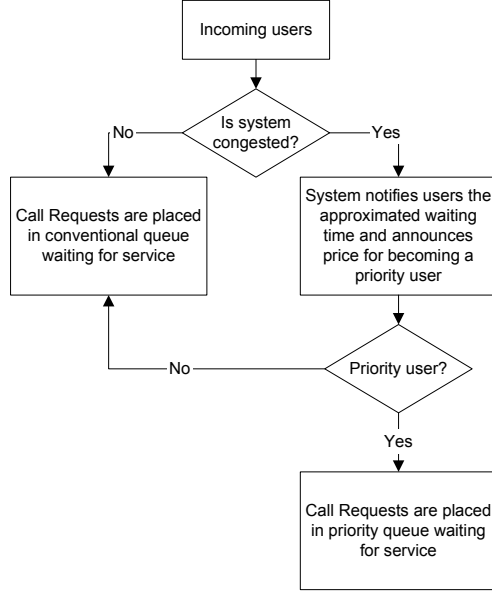


Fig. 2. Call procedure in the proposed pricing model

### 3.1. Directing Traffic

An important parameter used by the PCAC block is the *priority factor* ( $P_s$ ), i.e., the portion of total resources that is assigned to the priority queue. With a higher  $P_s$ , more resources would be dedicated to priority queues.  $P_s$  can be adjusted regularly based on the users' QoS constraints and the incoming traffic for both user types. We assume that the average holding time of priority users is shorter than that of the conventional users. Since they are willing to pay extra, it is unlikely to spend much time in the system. Both types of users share the same resources (telephone trunks in this case). Therefore, the average holding time that the system experiences from both call types would be the weighted sum of the holding time by the priority users and conventional users, i.e.,

$$T_{avg} = P_s T_{avg-p} + (1 - P_s) T_{avg-c} \quad (1)$$

where  $T_{avg-p}$  is the average holding time of priority users and  $T_{avg-c}$  is the average holding time of conventional users. The  $P_s$  parameter controls the amount of network resource assigned to priority users; the remaining resources are assigned to conventional users. With this model in mind, we can assume that the trunk group is logically divided into two groups. One group is assigned for priority users and the other group is assigned for conventional users. Both of them can be studied independently. Hence, the number of trunks assigned to priority users is  $N_p = P_s \cdot N$  whereas the number of trunks assigned to conventional users is  $N_c = (1 - P_s) \cdot N$ , with  $N$  being the total number of telephone trunks.

### 3.2. Dynamic prices

The price that is broadcasted to users when the system experiences congestion, can be derived from the demand function. The demand function describes the users' reaction to the price changes. We use the demand function that appears in [7] since it is used for different priority users, which fits our model. The demand function is as follows:

$$q = e^{-\frac{(p_h - 1)^2}{p_0}} \quad p_h \geq p_0 \quad (2)$$

where  $p_0$  is the price charged to conventional users,  $p_h$  is the price charged to priority users, and  $q$  is the percentage of priority users who are willing to pay this higher price. From (2),

$$p_h = p_0 + \frac{p_0 \sqrt{-4 \ln(q)}}{2} \quad (3)$$

The percentage of incoming users ( $q$ ) gives us information regarding the number of users who are placed in each queue. We assume that the performance of each type of queue can be considered independently. The model consists of two basic queuing models, with average holding times given by (1), and arrivals dictated by the demand function (2) at a given price. Basically, these basic queuing systems form a M/M/m system which can be studied by the Erlang-C formula, i.e.

$$C^{-1}(N, a) = B^{-1}(N, a) - B^{-1}(N - 1, a) \quad (4)$$

where  $B(N, a) = \frac{a^N}{N!} / \sum_{i=0}^N \frac{a^i}{i!}$  (Erlang-B formula)

The QoS can be characterized by the user delay experienced in the queues. More specifically, by the tail of a delay distribution, i.e.  $P[\text{user delay} > R \text{ seconds}]$  is less than a QoS requirement, e.g., 1%. Therefore, using the Erlang-C formula and the fact that we can consider both queues independently, we can identify the QoS requirement as follows:

$$P[W > t_p] = C(N_p, a_p) e^{-N_p \mu (1 - \rho_p) t_p} \quad (5)$$

$$P[W > t_c] = C(N_c, a_c) e^{-N_c \mu (1 - \rho_c) t_c} \quad (6)$$

where  $a_p = q \cdot a_{total}$ ,  $\rho_p = a_p / N_p$   
 $a_c = (1 - q) \cdot a_{total}$ ,  $\rho_c = a_c / N_c$   
 $a_{total} = \lambda(t) \cdot T_{avg}$

$C(N, a)$  is given by the erlang-C formula,  $W$  is the user delay (time in queue),  $a_c$  is the load imposed by each type of users,  $N_c$  is the number of trunks logically assigned to each type of users,  $\mu$  is average departure rate of users ( $1/T_{avg}$ ),  $\rho_c$  is the load per server for each type of users, and  $t_p$  and  $t_c$  are delay constraints for the priority and conventional queues respectively. We assume that  $t_p$  should be a lot less than  $t_c$ , when the system experiences congestion. Utilization of the overall system is given by

$$Utilization = \frac{\lambda(t) * T_{avg}}{N} \quad (7)$$

where  $\lambda(t)$  is the arrival rate in the telephone system at time  $t$ .

### 3.3. Optimal Call Arrival Rate

As the system operates, the system resources are shared in a way that the QoS requirements for each user type can be achieved. An important parameter here is the maximum number of users that the network can accommodate. The number of users need to conform to the QoS constraints of both queues. This parameter is influenced by the optimal call arrival rate ( $\lambda_{opt}$ ), which is the maximized overall arrival rate of the system. We know that  $\lambda_{opt}$  is embedded in (5) and (6). For different percentages of priority users ( $q$ ) and priority factors ( $P_s$ ), we can achieve a certain arrival rate. The  $\lambda_{opt}$  can be obtained when we find that arrival rate that maximizes the utilization of the system.

To obtain  $\lambda_{opt}$ , we need to consider equations (5) and (6). The QoS constraint in (5) and (6) can be set at a certain probability level, depending upon the user requirement.  $\lambda_{opt}$  can be found by setting probabilities in (5) and (6) as the QoS constraint (1% in this case). Here, we obtain the maximum arrival rate ( $\lambda$ ) numerically, by changing  $P_s$  and  $q$ . Therefore, for a certain value of  $q$ , we can find that  $P_s$  that yields maximum call arrival rate or  $\lambda_{opt}$ . That is

$$\lambda_{opt} \text{ for certain } (q) = f(P_s^*, q) \quad (8)$$

where  $P_s^*$  satisfies the condition  $\frac{df(P_s, q)}{dP_s} = 0$

## 4. Performance Analysis

In section 4.1, we describe the basic assumptions and parameters used in the priority queuing system. The results of our analysis are shown in section 4.2.

### 4.1 Assumptions and parameters

We assume for the sake of simplicity that the considering network is a circuit-switched telephone network. The arrivals are modeled using the Poisson law (exponentially distributed inter-arrival times). The system queues are first-come first-serve (FCFS). The parameters used throughout our analysis are as follows:

1. The number of telephone trunks equal 30. Trunks assigned to each queue are regulated by the parameter  $P_s$ .
2. The average call holding time for priority users ( $T_{avg_p}$ ) and conventional users ( $T_{avg_c}$ ) are exponentially distributed with mean 120 seconds and 300 seconds respectively.
3. Regarding the QoS parameters:
  - a. Probability of priority users kept waiting in the queue for more than 1 minutes ( $t_p$ ) is less than 1%; and
  - b. Probability of conventional users kept waiting in the queue for more than 10 minutes ( $t_c$ ) is less than 1%.
4. The normal charging rate for conventional users using the trunks ( $P_o$ ) is 8 cents per minutes. The charging rate for priority users ( $P_h$ ) depends on the demand function and it is broadcasted upon arrival.

### 4.2. Numerical results

Figure 3 shows the relationship between the arrival rate and priority factor ( $P_s$ ). For a certain percentage of priority users, there is apparently an optimal call arrival rate (the peak of the curve) that maximizes the number of calls and still maintains a QoS requirement of both queues. As  $q$  increases, the optimal call arrival rate is increased. Figure 3 also shows the improvement in call accommodation. Without our pricing scheme, the optimal arrival rate is 5.6 calls/sec. When our scheme is used, the optimal arrival rate increases to 8.5 calls/sec, with a  $q$  of 80%. We can achieve a higher optimal call arrival rate, by degrading the QoS requirements of either type of users.

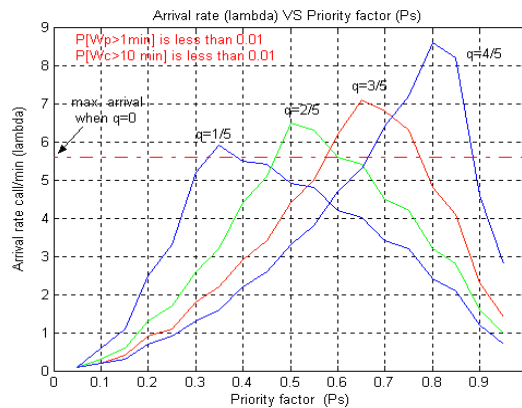


Fig. 3. Optimal arrival rate for certain percentage of priority users

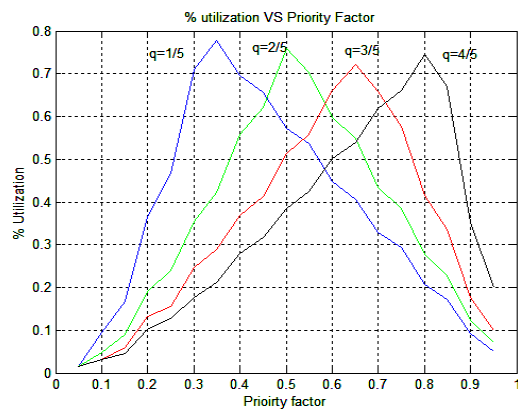


Fig.4. The total utilization VS priority factor and percentage of priority users

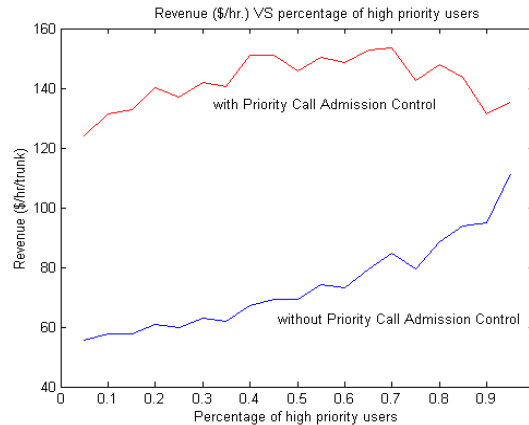


Fig.5. Revenue with Priority Call Admission Control (\$ per hour per trunk)

In terms of utilization, Figure 4 shows the utilization of the system. Apparently, the utilization of the system is roughly the same regardless of  $q$ . This is because the utilization of the system depends only on arrival rate  $\lambda(t)$  and priority factor ( $P_s$ ). Therefore, we are able to achieve high utilization of the system with minimum amount of user who is willing to be priority users. In addition, without the proposed pricing scheme, the utilization required to achieve optimal call arrival rate is 93% (from (7)). However, The system utilizes only 75% of its resources with our pricing scheme.

Figure 5 compares the revenue generated from each trunk by using the proposed scheme vs. the revenue under the traditional fixed pricing scheme, for the same amount of traffic. The revenue stream consists of the sum of the revenue created by the priority users and conventional users with their respective price factor. We observe that there is a significant revenue increase due to our way of pricing. However, the effect of the percentage of priority users ( $q$ ) to the revenue is not significant. Therefore, we can operate at a level of low percentage of priority users, and still generate higher revenue.

## 5. Conclusion

Our proposed pricing model provides incentive for users to use system resource more efficiently. Furthermore, users will be satisfied with the fact that they can choose their pricing schemes based on their expected quality of service. The proposed system is also flexible enough to adapt to the fluctuating traffic by adjusting the pricing factor and the users' QoS requirements. In addition, our pricing model is general enough and has been proposed for voice services in mobile networks [8].

## References

1. Falkner, M.: A user's perspective on Connection Admission Control: Integrating Traffic Shaping, Effective Bandwidths and Pricing. Doctoral thesis at Carleton University, Ottawa, Canada, May 12, 2000.
2. Fitkov-Norris, E.D., Khanifar, A.: Dynamic Pricing in Mobile Communication Systems. In: First International Conference on 3G Mobile Communication Technologies, 2000, 416–420
3. MacKie-Mason, J.K., Varian, H.R.: Pricing Congestible Network Resources. In: IEEE Journal on Selected Areas in Communications, Vol. 13, Issue. 7, (1995), 1141–1149
4. Viterbo, E., Chiasserini, C.F.: Dynamic Pricing for connection-oriented services in wireless networks. In: 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (2001), Vol.1, A-68 -A-72
5. Patek, S.D., Campos-Nanez, E.: Pricing of dialup services: an example of congestion-dependent pricing in the Internet. In: Proc. of the 39th IEEE Conference on Decision and Control (2000), Vol.3, 2296 -2301
6. Odlyzko, A.M.: Paris Metro Pricing for the Internet. In: Proc. ACM Conference on Electronic Commerce (1999), 140-147
7. Hou, J., Yang, J., Papavassiliou, S.: Integration of Pricing with Call Admission Control for Wireless Networks. In: 54<sup>th</sup> IEEE Vehicular Technology Conference (2001), Vol. 3, 1344 -1348
8. Yaipairoj, S., Harmantzis, F.: Dynamic Pricing with "Alternatives" for Mobile Networks. In: IEEE Wireless Communication and Networking Conference (2004)
9. Peha, J.M.: Dynamic Pricing as Congestion Control in ATM Networks. In: Global Telecommunications Conference (1997), Vol.3, 1367-137