Congestion Pricing by Priority Auction

Guanxiang Zhang, Jianhua He, Yajie Ma, Wenqing Cheng, Zongkai Yang

Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, P.R.China, 430074

ABSTRACT

This paper analyzes a communication network facing users with a continuous distribution of delay cost per unit time. Priority queueing is often used as a way to provide differential services for users with different delay sensitivities. Delay is a key dimension of network service quality, so priority is a valuable resource which is limited and should to be optimally allocated. We investigate the allocation of priority in queues via a simple bidding mechanism. In our mechanism, arriving users can decide not to enter the network at all or submit an announced delay sensitive value. User entering the network obtains priority over all users who make lower bids, and is charged by a payment function which is designed following an exclusion compensation principle. The payment function is proved to be incentive compatible, so the equilibrium bidding behavior leads to the implementation of "$c\mu$-rule". Social warfare or revenue maximizing by appropriately setting the reserve payment is also analyzed.

Keywords: Priority, Pricing, Auction, Payment function

1. INTRODUCTION

Traditional applications, such as web browsing, file transfer, remote terminal and electronic mail, etc, do not impose severe requirements on the network. They can tolerate relatively large packet delays. New Internet applications, such as real-time applications such as interactive voice and video are more delay-sensitive. Heterogeneity of the delay requirements makes it necessary that different users are handled differently. The emergence of time-critical applications on the Internet is one of the primary reasons for customer-oriented service differentiation. On the other hand in order to survive in the highly competitive Internet services market, the network service providers will have to provide customized network services. Internet is becoming more and more a multi-service network. Clearly, any successful solution to supporting multiple services cannot rely on technical solutions only but also has to take into account the economic aspects. Corresponding to the best effort service, today the most common charging method in Internet is based on the flat-rate model. Given the differentiation of network services, the flat-rate pricing model which is commonly applied to charge users for the access service to the Internet Service Providers (ISPs) will become inadequate. Without an appropriate pricing scheme, any service differentiation is useless. If there were no price difference between the priority classes, all users would prefer the best one.[1][2]

This work was supported by the National Science Foundation of China (NSFC) under grant number 60202005

Network Architectures, Management, and Applications II, edited by S. J. Ben Yoo,
Gee-Kung Chang, Guangcheng Li, Kwok-wai Cheung, Proc. of SPIE Vol. 5626
(SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.577065

Priority queueing is often used as a way to provide differential services for users with different delay sensitivities. When a capacity-constrained network service provider faces delay sensitive customers, delay is a key dimension of network service quality, so priority is a valuable resource which is limited and should be optimally allocated. Pricing can play an important role in the allocation of service capacity and the appropriate determination of priority [3]. Many studies of price and service differentiation in priority queueing systems analyze the centralized pricing, where the provider sets an incentive compatible price-service menu for finite classes of users. But in settings with many or continuum customers whose delay sensitivities are not known to the provider, it may be beneficial to use auctions, where the provider allocates priorities and charges corresponding payments based on customers' bids [4].

In this paper, we consider a priority queueing system with many infinitesimal users. The main QoS parameter that users care about is delay. Assume there is a continuous distribution of users' delay cost per unit time. We analyze the allocation of priority in queues via a simple bidding mechanism. In our model, the stochastically arriving users are privately informed about their own marginal costs of delay which is observed neither by the provider nor by the other customers, and arriving customers cannot observe the system state. Arriving users can decide not to enter the network at all or submit an announced delay sensitive value. When a user enters the network, he obtains priority over all users (waiting in the queue or arriving while he is waiting) who make lower bids, and is charged by a payment function which is designed following an exclusion compensation principle.

Consider the participation constraint, under some conditions users with the highest delay cost values will decide not to enter the network, because the service value is not enough to cover its total cost. For the users entering the network, we know that it is optimal (minimizing the total delay cost per unite time of users) to schedule them by the so called "$c\mu$-rule" which provides a higher priority to those users who have a higher marginal delay cost. This rule is implemented by the payment function which is proved to be incentive compatible: users entering the network will submit their true delay sensitive values. So a user with higher marginal cost submits a higher bid, and higher priority services are allocated to the users who are more sensitive to delay. As a result the equilibrium bidding behavior leads to the implementation of "$c\mu$-rule". When allow the network imposing a uniform reverse payment on users who enter the network, Social warfare or revenue maximizing can be realized by appropriately setting the reserve payment.

A number of authors have addressed related issues. Auction motivates lots of research interest in network resources allocation and congestion pricing, and several schemes were proposed such as Smart-market Pricing [5], Progressive Second Price auction (PSP) [6] and Smart Pay Admission Control (SPAC) [7] mechanism. But these schemes mainly focus on the allocation of limited network resource to users not the priority in which agents' jobs are processed. Mendelson and Whang [3] analyze the M/M/1 non-preemptive priority queue with multiple user classes, deriving an incentive compatible priority pricing scheme. Mandjes [1] analyzes the incentive compatible problem of data users and voice users in computer network when network provides a service with only two priorities. Kleinrock [8] was the first one to study the allocation of priorities based on payments, he derived steady-state expected waiting times (which depend on the bribes) and studied the resulting queue discipline for various payment functions. Liu [9] revisits Kleinrock's model, but assumes that customers make payments in order to minimize its total cost; he derives a bidding equilibrium, and shows that a higher marginal cost leads to a higher bid. Afèche and Mendelson [4] analyze the priority auctions with a generalized delay cost structure. Kittsteiner and Moldovan [10] analyze the bidding strategy that depends on processing time. Our analysis is mainly based on Liu' work, in our model the user' bid behavior is simplified: users only need to decide whether to enter the network and announce their delay values if entering, the payment calculation is left to the network, and we focus on the optimal setting of reserve payments to maximize the social warfare or revenue, while Liu are more care about optimally setting service speed.

The remainder of this paper is organized as follows: In Section 2, we describe the basic model of the priority services, a payment function is given and proved to be incentive compatible. In section 3 we consider incentive compatible payment function when a reverse payment is allowed. We analyze how users entering the network can be regulated by reverse payment, and we analyze the problem of social warfare or revenue maximizing by appropriately setting the reserve payment. Finally, in Section 4, we conclude the paper with a summary.

2. THE BASIC MODEL

We consider a capacity-constrained network service provider, modeled as an M/M/1 queueing system that serves users with differential delay sensitivities. The network service provider distributes a network service which has constant value *V*. Service time obeys an exponential distribution with unit mean. Users which are infinitesimal relative to the market size arrive according to a Poisson process. λ is assumed to be the mean arrival rate or market size of users which arrive at the network according to a Poisson process.

Users differ in their delay sensitivities c, the marginal delay cost per unit time. Assume that there is a cumulative distribution of delay sensitivities represented by $A(c)$, $c\in[0, c_{max}]$. It's assumed that $A(c)$ is common knowledge, and is continuously differentiable.

The provider allocates priorities of queues via a simple direct-revelation bidding mechanism. When a user comes to the queue he can choose either of two strategies: (1) not to enter the queue at all or (2) submit an announced delay sensitive value $\hat{c}$ and pay a charge which is decided by a payment function $p(\hat{c})$ (non-revisable and non-refundable) to the network. A user who submits $\hat{c}$ gets priority over all those with strictly lower values $\hat{c}' < \hat{c}$, and equal bidders are served FIFO. Suppose $\bar{c} \le c_{\max}$ is the maximization delay value of users who choose to enter the queue. Following Naor [11], the utility of a user with true (marginal delay cost) value *c*, announced value $\hat{c}$ and experiences a delay *w* is:

$$U(c,\hat{c}) = V - (p(\hat{c}) + cw) \tag{1}$$

From [8] and [9], when users are ranked by their true delay sensitive values (when each user announces $\hat{c}=c$ in our model), the aggregate delay cost of users entering the queue is minimized, and the mean delay of user with delay value *c* is given by

$$w(c) = \frac{1}{(1-\lambda A(\bar{c}) + \lambda A(c))^2}\,. \tag{2}$$

Proposition 1. Under above assumptions, the bidding mechanism is incentive compatible and quasi-optimal when the payment function is given by

$$p(\hat{c}) = \int_0^{\hat{c}} - xw'(x)dx\,. \tag{3}$$

Proof: From (3) we have

$$p'(\hat{c}) = \hat{c}w'(\hat{c})\,.$$

So

$$\begin{aligned} U'(\hat{c}) &= -p'(\hat{c}) - cw'(\hat{c}) \\ &= (\hat{c} - c)w'(\hat{c}) \end{aligned}\,.$$

From (2), $w'(\hat{c})$ <0, when $\hat{c}$ >*c,* $U'(\hat{c})$ 〈0, while when $\hat{c}$ <*c,* $U'(\hat{c})$ >0, so when $\hat{c}$ =*c,* user obtains maximizing utility. So users bidding strategies is incentive compatible:

$$U(c,c) \ge U(c,\hat{c}),\ \forall\, \hat{c} > 0.$$

Because users bidding strategies is incentive compatible, and users are ranked by the announcing values, the aggregate delay cost of users entering the queue is minimized by the bidding mechanism. Thus the mechanism is quasi-optimal. This completes the proof.

Now the utility of a user with delay value *c* in the incentive compatible mechanism is:

$$U(c) = V - (p(c) + cw(c)) \tag{4}$$

Besides incentive compatible attribute, participation constrain is also important for mechanism design. Because the bidding mechanism is incentive compatible, we have

$$\begin{aligned} U'(c) &= -p'(c) - w(c) - cw'(c) \\ &= cw'(c) - w(c) - cw'(c) \\ &= -w(c) \end{aligned}$$

From (2) $w(c)>0$, we have $U^{'}(c)<0$, so $U(c)$ is a strictly decreasing function over $[0, c_{max}]$, we have mentioned that $\bar{c} \le c_{max}$ is the maximizing value of users who choose to enter the queue. Assume $\dot{c}$ is the unique solution of

$$V - (p(c) + cw(c)) = 0\,.$$

$\dot{c}$ is determined by the mean arrival rate λ and cumulative distribution of *c*. and we have $\bar{c}$ = *min* (c_{max}, $\dot{c}$). The participation constrain of the bidding mechanism is $c \le \bar{c}$. Users with true value $c \le \bar{c}$ enter the queue and announce $\hat{c}$ =*c,* while other users choose not to enter the queue. Thus all users will get nonnegative expected utility.

By (3), we know that the marginal increase of the payment of a user (with value c) is equal to the resulting decrease of his delay cost $-cw^{'}(c)$. Each user's payment equals his priority externality, the marginal net value losses he inflicts on all lower-priority customers. So the intuition behind the payment function is an exclusion compensation principle. A user with higher delay sensitivity will submit higher delay sensitive value, and will be scheduled by a higher priority, so more users' mean delay will be inflicted by his enter; as a result he should be charged a higher payment. This pricing principal is analogous to the Vickrey-Clarke-Groves mechanism, which is widely used in the allocation of interdependent resources.

3. OPTIMIZATION OF SETTING RESERVE PAYMENT

In the basic model of section 2, marginal delay value of users entering the network is determined by the mean arrival rate λ and cumulative distribution of *c*. A user with unit delay value *c*=0 will not be charged any payment. Now we assume that the network imposes a uniform reverse payment $\underline{p}$ on all users who enter the network (Auction model of section 2, can be treated as a special case with $\underline{p}$ =0). As same as section2, when a user comes to the network he can decide not to enter the network at all or submit an announced delay sensitive value $\hat{c}$ and pay a charge which is decided by a payment function $p(\hat{c})$. Users are also ranked by their announced delay sensitive value.

Proposition 2. When a reverse payment is charged, the bidding mechanism is incentive compatible and quasi-optimal when the payment function is given by

$$p(\hat{c}) = \int_0^{\hat{c}} - xw'(x)dx + \underline{p}\,. \tag{5}$$

Prove of Proposition 2 is similar to Proposition 1, we omit it here.

Consider the participation constrain $U(\dot{c})$=0, and $\bar{c}$ = *min* (c_{max}, $\dot{c}$). Now network service provider can regulate the marginal delay value of users by the choice of variable $\underline{p}$. When network services provider sets a higher reverse payment, marginal delay value will be also higher. So a higher reverse payment leads to fewer users entering the network. On the contrary lower reverse payment leads to more users entering the network. Network service provider can optimally set the reverse payment to maximize the revenue or the social warfare.

To simple the analyzing of optimal problems of revenue maximizing or social warfare maximizing, we should get more explicit expression of *w*(*c*), *p*(*c*), and $\bar{c}$. We assume *c* follows a uninformed distribution over [*0*, c_{max}]:

$$A(c) = Ac, c \in [0, c_{\max}] \tag{6}$$

So Ac_{max}=1. And from (2) we have

$$w(c) = \frac{1}{(1 - \lambda A\bar{c} + \lambda Ac)^2} \tag{7}$$

From (5) we have

$$\begin{aligned} p(c) &= \int_0^{c} - xw'(x)dx + \underline{p} \\ &= \frac{1}{\lambda A(1-\lambda A\bar{c})} - \frac{1}{\lambda A(1-\lambda A\bar{c}+\lambda Ac)} - \frac{c}{(1-\lambda A\bar{c}+\lambda Ac)^2} + \underline{p} \end{aligned} \tag{8}$$

For *c*=$\bar{c}$, we have

$$p(\bar{c}) = \frac{1}{\lambda A(1-\lambda A\bar{c})} - \frac{1}{\lambda A} - \bar{c} + \underline{p}$$

For $\bar{c} < c_{max}$, $U(\bar{c})$=0, so

$$p(\bar{c}) = V - \bar{c}$$

From the last two equations we have:

$$\bar{c} = \frac{V - \underline{p}}{1 + \lambda A(V - \underline{p})} \tag{9}$$

It can be see that marginal delay value of users entering the network is determined by the mean arrival rate λ and cumulative distribution A, and the reserve payment $\underline{p}$. When λ and A are fixed, from (9) $\bar{c}$ is a decreasing function of $\underline{p}$. The reverse function of (9) is given by

$$\underline{p}(\bar{c}) = V - \frac{\bar{c}}{1 - \lambda A\bar{c}}\,. \tag{10}$$

Define the minimizing value of $\underline{p}$ as $\underline{p}_{\min}$ at which exactly all users enter the network. When $\underline{p} > \underline{p}_{\min}$ only some users enter the network. By substituting $\bar{c} = c_{max}$ and $Ac_{max}=1$ to (10) we have

$$\underline{p}_{\min} = V - \frac{c_{\max}}{1-\lambda}\,.$$

Considering the constraint $\underline{p} \geq 0$, so when $\underline{p}_{\min} < 0$ and $\underline{p} > 0$ or when $\underline{p}_{\min} > 0$ and $\underline{p} > \underline{p}_{\min}$, $\bar{c}$ is determined by (9). While when $\underline{p}_{\min} > 0$ and $0 \leq \underline{p} \leq \underline{p}_{\min}$, $\bar{c} = c_{max}$.

3.1 Revenue maximization

Revenue of network service provider consists of the payments obtained from those users who enter the network:

$$\Pi = \lambda \int_0^{\bar{c}} Ap(c)dc\,. \tag{11}$$

When $\underline{p}_{\min} < 0$ and $0 \leq \underline{p} \leq V$, $\bar{c}$ is determined by (9), together with (8) we can derive the expression of Π as a function of $\underline{p}$.

$$\Pi(\underline{p}) = V - \underline{p} + \frac{(1 + \underline{p}\lambda A)(V - \underline{p})}{1 + \lambda A(V - \underline{p})} - \frac{2\ln(1 + \lambda A(V - \underline{p}))}{\lambda A} \tag{12}$$

From (12) we can derive the first-order and second-order derivatives of $\Pi(\underline{p})$:

$$\Pi'(\underline{p}) = \frac{\lambda A(V - 2\underline{p})}{(1 + \lambda A(V - \underline{p}))^2} \tag{13}$$

$$\Pi''(\underline{p}) = \frac{-2\lambda A(1 + \lambda A\underline{p})}{(1 + \lambda A(V - \underline{p}))^3} \tag{14}$$

Because $\Pi''(\underline{p}) < 0$ so Π is a strictly concave function of $\underline{p}$. And by first order condition $\Pi'(\underline{p}) = 0$, the network service provider gets a maximum when he sets $\underline{p} = V/2$.

When $\underline{p}_{\min}$ >0, the curve of $\Pi(\underline{p})$ have two sections: 0≤ $\underline{p} \le \underline{p}_{\min}$ and $\underline{p}_{\min} < \underline{p}$ ≤V. When 0≤ $\underline{p} \le \underline{p}_{\min}$, from (8), (11) , $\bar{c}$ =c_{ma} and Ac_{max}=1, we obtain

$$\Pi(\underline{p}) = \frac{c_{\max}}{1-\lambda} + c_{\max} + \frac{2c_{\max}\ln(1-\lambda)}{\lambda} + \lambda \underline{p} \qquad 15)$$

From (15), we can derive the first-order of $\Pi(\underline{p})$:

$$\Pi'(\underline{p}) = \lambda$$

So when 0≤ $\underline{p} \le \underline{p}_{\min}$, revenue of network service provider can be increased by raising $\underline{p}$ till it reach $\underline{p}_{\min}$. While when $\underline{p}_{\min} < \underline{p}$ ≤V the expression of Π is given by (12) as same as the situation of $\underline{p}_{\min}$ <0 and 0< $\underline{p}$ <V. From(15),(12), Π is a linear function over 0≤ $\underline{p} \le \underline{p}_{\min}$, a concave function over $\underline{p}_{\min} \le \underline{p}$ ≤V, and Π is continuous at $\underline{p}_{\min}$. So when $\underline{p}_{\min}$ ≤V/2, the network service provider gets a maximum when he sets $\underline{p}$ =V/2, while when $\underline{p}_{\min}$ >V/2 the network service provider gets a maximum when he sets $\underline{p} = \underline{p}_{\min}$.

3.2 Social warfare maximization

Social warfare consists of the aggregate net value of users who enter the network.

$$\Gamma = \lambda \int_0^{\bar{c}} (V - cw(c))Adc \qquad (16)$$

By substituting (7) to (16) we can derive the expression of Γ as a function of $\bar{c}$.

$$\Gamma(\bar{c}) = \lambda A\bar{c}V + \bar{c} + \frac{\ln(1-\lambda A\bar{c})}{\lambda A} \qquad (17)$$

From (12) we can derive the first-order and second-order derivatives of Γ :

$$\Gamma'(\bar{c}) = \lambda AV + 1 - \frac{1}{1-\lambda A\bar{c}} \qquad (18)$$

$$\Gamma''(\bar{c}) = \frac{-\lambda A}{(1-\lambda A\bar{c})^2} \qquad (19)$$

From (19), we have $\Gamma''(\bar{c})$ <0, Γ is a strictly concave function of $\bar{c}$.

Consider the situation: $\underline{p}_{\min}$ <0 and 0≤ $\underline{p}$ ≤V. From (9) we obtain the value of $\bar{c}$ corresponding to $\underline{p}$ =0

$$\bar{c}^0 = \frac{V}{1+\lambda AV}$$

From (18), we have $\Gamma'(0) = \lambda AV > 0$. So the maximum of $\Gamma(\bar{c})$ may obtain at $\bar{c}^*$=min($\bar{c}^0$, $\bar{c}^1$), where $\bar{c}^1$ is obtained by the first-order condition. And from (10) we can obtain the corresponding reserve payment maximizing the social warfare.

Consider the situation of $\underline{p}_{\min}$ >0, the maximum of $\Gamma(\bar{c})$ also obtain at c^*=min(c_{max}, $\bar{c}^1$). When $\bar{c}^*$ = c_{max} the corresponding reserve payment could be any value between [0, $\underline{p}_{\min}$], while when $\bar{c}^*$ = $\bar{c}^1$ we can obtain the corresponding reserve payment maximizing the social warfare from (10).

4. CONCLUSIONS

This paper analyzes a communication network facing users with a continuous distribution of delay cost per unit time. We solve the problem of how to allocate priorities in queues via a simple bidding mechanism. In our model, arriving users can decide not to enter the network at all or submit an announced delay sensitive value. A user obtains priority over all users who make lower bids, and is charged by a payment function which is designed following an exclusion compensation principle. The payment function is proved to be incentive compatible, so the equilibrium bidding behavior leads to the implementation of “$c\mu$-rule”. Social optimization or revenue maximizing by appropriately setting the reserve payment is also analyzed.

REFERENCE

1. M.R.H. Mandjes, “Pricing strategies under heterogeneous service requirements,” *Proceedings Infocom 2003,* San Francisco USA.
2. J. Altmann, H. Daanen, H. Oliver, and A Sánchez-Beato Suárez, "How to Market-Manage a QoS Network (zip ps)," IEEE Infocom2002, New York, USA, June 2002.
3. H. MENDELSON and S. WHANG, “Optimal incentive-compatible priority pricing for the M/M/1 queue,” Operations Research, Vol. 38, pp.870-883, 1990.
4. Philipp Afeche and Haim Mendelson, “Pricing and Priority Auctions in Queueing Systems with a Generalized Delay Cost Structure,” Management Science 50(7), pp. 869–882, 2004.
5. J. MacKie-Mason, H. Varian: Pricing the Internet; In Public Access to the Internet, B. Kahn, J. Keller (eds.), Prentice Hall, Englewood Cliffs, New Jersey, U.S.A., 1995.
6. Lazar, Semret, “Design and Analysis of the Progressive Second Price Auction for Network Bandwidth Sharing”, Telecommunications Systems, Special Issue on Network Economics, 1999.
7. Jun Shu, Pravin Varaiya , “Pricing Network Services”, proceeding of INFOCOM 2003, San Francisco USA.
8. Kleinrock, L, “Optimum bribing for queue position,” Oper. Res. 15 304–318, 1967
9. Lui, F. T, “ An equilibrium queuing model of bribery,” J. Political, Econom. 93 760–781, 1985.
10. Kittsteiner and Moldovanu, “Priority Auctions and Queue Disciplines that Depend on Processing Time,” to appear in Management Science 2004.
11. Naor, P. “On the regulation of queue size by levying tolls.” Econometrica 37 15–24. 1969.