Pricing Network Resources: A New Perspective

Joy Kuri and Sharmili Roy
Centre for Electronics Design and Technology
Indian Institute of Science
kuri@cedt.iisc.ernet.in sharmili@cedt.iisc.ernet.in

Abstract—We consider I users sending elastic traffic into a Service Provider's (SP's) network. Each user has a contract to transmit data at a nominal rate. However, a user is free to transmit at a higher rate if she wishes.

For each user, we assume a realistic traffic model. A user alternates between two phases over time. In the *limited data* phase, a user's transmission rate is capped at some value below her contracted rate. This happens because the user does not have enough data to send. In the *unlimited data* phase, a user has enough data to sustain any rate.

At time t_1, a (possibly empty) set S_1 of users is in the limited data phase. This leaves unutilized resources in the network that ought to be exploited by others. At time $t_2 > t_1$, users in S_1 are back in the unlimited data phase. Now it should be possible for these users to reclaim their contracted rates at the expense of "overusers."

We show that under certain conditions, a very simple pricing scheme can ensure fair and efficient operation in the above sense. A SP needs just one price for all the users.

Index Terms—ricing, Congestion, Fairnessricing, Congestion, FairnessP

I. Introduction and Related Work

Pricing has been suggested as a mechanism to control congestion and ensure fair and efficient operation of networks. In much of the published literature ([3] [2] [6]), elastic traffic is considered and the context of operation is as follows. Each user has a utility function that quantifies the benefit that she derives from the network. The utility is a concave increasing function of the rate at which user can send data through the network. System objective is maximization of the sum of all users' utility functions. The problem is to find vector of users' rates such that system objective is realized.

The resulting constrained optimization problem can be solved in a centralized manner if all the utility functions were known. In [3], Kelly proposed a decentralized method to arrive at the system-optimal rates. In this method, the network declares *prices*, and each user individually solves the problem of maximizing her *net benefit* or *net utility*, which is her utility minus the total cost paid to the network. He shows that there exists a vector of prices such that the vector of individually optimal rates arrived at by the users is, indeed, the system-optimal rate vector.

Several authors have pointed out that it is enough to work under the assumption of *direct revelation*, in which the users' utility functions are revealed to a central controlling authority ([4], [1]). Accordingly, we assume in this paper that users' utility functions are known to the central authority (the SP) which can then use this knowledge to design appropriate prices. It will turn out that the complete utility function need not be known; actually, far less knowledge suffices — only the derivative of the utility function at a single point is enough. Further, [5] mentions that prices are used in two kinds of problems, either to promote fair and efficient resource sharing or to maximize the revenue earned by the central authority. As in [5], our objective in this paper is fair and efficient sharing of a network.

The published literature, however, tacitly assumes an *infinite data model*. Every source is assumed to have an infinite backlog of data. So she can send traffic at *any* rate (obtained from the solution to the individual optimization problem) continuously. In practice, of course, sources will occasionally run out of data. We consider a *finite* data model, in which, occasionally, a source does not have enough data to send. Thus, a source may not be able to sustain a data rate that is suitable for fair and efficient operation of the system.

Further, in practice, users have contracts with the SP that specify the rates at which they can send traffic into the network. We call a user who is sending traffic at a rate lower than her contracted rate as "underuser" and a user with plenty of data available as "overuser" because she can send data at a rate higher than her contracted rate.

We also believe that one must have congestion-dependent and user-dependent pricing. If the network is not congested, then the price should remain low, so that users with excess data can utilize the network. But when the network becomes congested, the price should not increase equally for all users; rather, those users who have exceeded their contracted rates and have caused congestion should be charged heavily, while those who are compliant should be charged at no more than their nominal rates. Even though our framework allows different prices for different users, our analysis shows that under some easily satisfyable condition, a *single* price for all users suffices.

We are interested in ensuring that network operation is characterized by the following.

- When some users are underusers because of limited available data, it should be possible for others to increase their rates beyond contracted so as to utilize the slack. Does there exist a pricing scheme such that users with plenty of data available are *encouraged* to become overusers?
- Later, when underusers wish to increase their rates because they have more data to send, they should have the incentive to do so and overusers should be encouraged to back down. Does their exist a pricing scheme such that this happens?

II. Model

A. Utility function

We consider a network which is shared among I users, where I is a given and fixed integer. We use a fluid model for traffic. User i injects fluid at rate λ_i into the network. We emphasize that this is a variable, because the user may not be able to generate traffic at a constant rate throughout.

User i has a contract to send traffic at rate γ_i, and the price charged by the SP is π_i; this is the *total price*, not the price per unit flow. When $\lambda_i > \gamma_i$, we call $(\lambda_i - \gamma_i)$ the "excess rate." The *utility* of user i is a concave strictly increasing function of the rate of user i traffic *actually carried* from the source to destination node. If λ_i is the flow i injects at source node and some part of i's traffic is dropped, then $\mu_i \leq \lambda_i$ is the amount carried to the destination node. The utility for user i is $U_i(\mu_i[\lambda_i])$. Since carried traffic (μ_i) depends upon the injected traffic (λ_i), we denote μ_i by $\mu_i[\lambda_i]$ to emphasize this dependence.

B. Multiplicative congestion-penalty function

The pricing scheme is characterized by the following features. *Underusers* are charged less than their contracted price. This is motivated by the goal of usage-based pricing. If a user is sending at a rate which is a fraction f of her contracted value, the charge is correspondingly a fraction f of her contracted charge. However, *overusers* charges depend upon whether the network is congested or not.

When the network is not congested, overusers are charged their contracted prices. Because as long as there is no congestion, users should be permitted to go above their contracted rates at no extra cost.

However, when the network is congested, overusers are charged heavily, because the overusers are themselves responsible for the congestion. If overuser i is sending at a rate $\lambda_i > \gamma_i$, then the price charged is $\pi_i P_i(\lambda_i - \gamma_i)$, where $P_i(.)$ is a multiplicative congestion-penalty function, and it takes the excess rate $(\lambda_i - \gamma_i)$ as its argument. $P_i(\lambda_i - \gamma_i)$ is a convex increasing function of excess rate. Clearly we should set $P_i(0) = 1$. In Figure 1, we give a schematic representation of the pricing scheme.

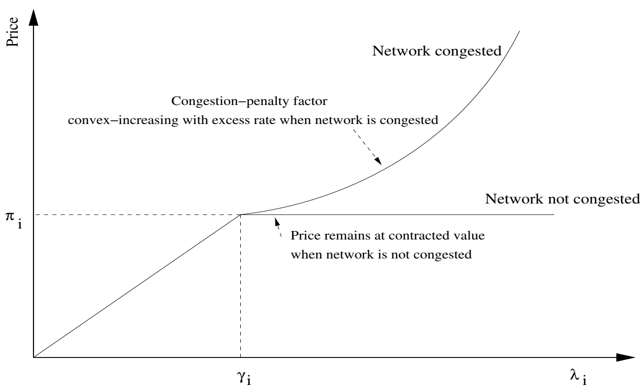

Fig. 1. Variation of price with user transmission rate and network congestion

C. Disutility function

Further, when user i is unable to send traffic at her contracted rate γ_i, we consider a "disutility" for user i. This measures the amount of dissatisfaction that user i suffers from at not being able to generate sufficient traffic. $(\gamma_i - \lambda_i)$ is referred to as the "shortfall" of user i, and the disutility function is a convex increasing function of shortfall. Further, the disutility function is defined to be zero when $\lambda_i \geq \gamma_i$.

One of our objectives is to design a scheme in which underusers have the incentive to increase their data rates when they have sufficient data to send. The disutility function is crucial in encouraging an underuser to increase rate when other users have not left any spare capacity in the network.

Furthermore, it can be shown mathematically that if we do not include a disutility function, we get different prices for different users which need to be continuously modified as the network evolves.

D. Net utility function

The "net utility" of user i is defined to be her utility minus disutility (which may be zero) minus the total price paid.

III. Analysis

In this section, we look to set prices π_i, $1 \leq i \leq I$, so that desirable behavior is enforced. Note that if these prices are too low, users cannot be prevented from transmitting above their contracted rates even if there is congestion in the network. Also, prices cannot be too high because then users will not have incentive to increase flows even to their contracted rates. Clearly, there is an upper and a lower bound on these prices. We would like to compute these bounds. We focus on user i and assume that the transmission rates λ_j of the other users $j, 1 \leq j \leq I, j \neq i$, are given and fixed.

Let us consider a network which has N nodes and L links. There is a node link incidence matrix A and a set of flow vectors, say $y^{(i)}$'s, each associated with a user i. An element $y^{(i)}(l)$, $1 \leq l \leq L$ is user i's traffic carried on link l. Let us call the source and destination nodes of a user as s_i and z_i respectively. $d^{(i)}$ is a vector of size $N \times 1$ for user i such that an element $d^{(i)}(n)$, $1 \leq n \leq N$, gives i's traffic dropped at node n. The flow conservation equation for user i is

$$Ay^{(i)} + d^{(i)} = v^{(i)} \tag{1}$$

where $v^{(i)}(s_i) = \lambda_i$, $v^{(i)}(z_i) = -\mu_i[\lambda_i]$ and all other elements are 0. $\mu_i[.]$ denotes the function which gives the traffic delivered to destination node. Link capacity constraints are

$$\sum_{i=\{1,2,\ldots,I\}} y^{(i)}(l) \leq C_l, \qquad \forall \;\; l = \{1, 2, \ldots, L\} \tag{2}$$

where C_l is the capacity of l^{th} link.

One question which may arise at this point is; given the injected traffic of all users (λ_i's), how do we know their flow vectors ($y^{(i)}$'s), vector of drops ($d^{(i)}$'s) and the carried traffic ($\mu_i[\lambda_i]$'s)? There can be multiple $y^{(i)}$'s, $d^{(i)}$'s and $\mu_i[\lambda_i]$'s which satisfy the flow conservation and capacity constraints. To get unique values, the network can perform optimizations like maximization of sum of utility functions, maximization of sum of carried traffics, maximization of the net revenue earned by the SP etc subject to Equations 1 & 2.

Now suppose λ_j, $j \in \mathcal{I} \setminus \{i\}$, are given and fixed. We define a congestion threshold ct_i for user i as the maximum injected traffic λ_i till which i's traffic is not dropped in the network. From user i's perspective, either ***(a)*** the network gets congested after the user is above her contracted rate, $ct_i > \gamma_i$, or ***(b)*** there is congestion in the network as soon as i tries to exceed contracted rate, $ct_i = \gamma_i$ or ***(c)*** even before user reaches contracted rate, there is congestion, $ct_i < \gamma_i$.

Due to space constraints we are unable to provide the proofs of the following results. In all cases, the proofs are based on the definition of net utility for that case, elementary calculus and simple algebra.

A. Congestion threshold above contracted rate

In this section, we consider a system where $ct_i > \gamma_i$. This means that even if user i transmits at her contracted rate, there will be some spare capacity left in the network. In this situation, it is desirable to let user i transmit at ct_i, so that the network is fully utilized, but without causing congestion. The net utility function for user i is denoted as NU_i. Following are the three different cases which can arise depending upon the flow rate of user i.

- When $\lambda_i < \gamma_i$,

$$NU_i \;\; = \;\; U_i(\lambda_i) - D_i(\gamma_i - \lambda_i) - \left(\frac{\lambda_i}{\gamma_i}\right)\pi_i$$

- When $\gamma_i \leq \lambda_i \leq ct_i$,

$$NU_i \;\; = \;\; U_i(\lambda_i) - \pi_i$$

- **When $\lambda_i > ct_i$,**

$$NU_i \quad = \quad U_i(\mu_i[\lambda_i]) - \pi_i P_i(\lambda_i - \gamma_i)$$

If $\frac{dNU_i}{d\lambda_i} > 0$, then i has incentive to increase her rate beyond λ_i. Corresponding conclusions apply when the derivative at λ_i is negative or zero.

Lemma 3.1: **When the total traffic from users $j \in \{1, 2, .. ., I\}$, $j \neq i$, is such that even if user i transmits at γ_i, there is some spare capacity left in the network, NU_i is maximized at the point $\lambda_i^* = ct_i$ if and only if the price π_i satisfies**

$$\frac{U_i'(\mu_i[ct_i + \delta^*])\mu_i'[ct_i + \delta^*]}{P_i'(ct_i + \delta^* - \gamma_i)} \leq \pi_i \leq \gamma_i(U_i'(\gamma_i) + \lim_{x \downarrow 0} D_i'(x)) \quad (3)$$

where $\delta^* = \arg\max_{\delta > 0}\left(\frac{U_i'(\mu_i[ct_i+\delta])\mu_i'[ct_i+\delta]}{P_i'(ct_i+\delta-\gamma_i)}\right)$. $\mu_i[ct_i + \delta^*]$ is user i's flow received at the destination node when $(ct_i + \delta^*)$ is the flow injected at the source node. Further, NU_i is an increasing function of λ_i to the left of λ_i^* and a decreasing function of λ_i to the right of λ_i^*.

B. Congestion threshold equal to contracted rate

Similar to the above case here also we can write the net utility of user i for different values of λ_i.

Lemma 3.2: **When the total traffic from users $j \in \{1, 2, .. ., I\}$, $j \neq i$, is such that, when user i transmits at γ_i the network is full, NU_i is maximized at the point $\lambda_i^* = \gamma_i$ if and only if price π_i satisfies**

$$\frac{U_i'(\mu_i[\gamma_i + \delta^*])\mu_i'[\gamma_i + \delta^*]}{P_i'(\delta^*)} \leq \pi_i \leq \gamma_i(U_i'(\gamma_i) + \lim_{x \downarrow 0} D_i'(x)) \quad (4)$$

where $\delta^* = \arg\max_{\delta > 0}\left(\frac{U_i'(\mu_i[\gamma_i+\delta])\mu_i'[\gamma_i+\delta]}{P_i'(\delta)}\right)$. $\mu_i[\gamma_i + \delta^*]$ is the amount of user i's flow received at the destination node when $(\gamma_i + \delta^*)$ is the flow injected at the source node. Further, NU_i is an increasing function of λ_i to the left of λ_i^* and a decreasing function of λ_i to the right of λ_i^*.

C. Congestion threshold below contracted rate

Although $ct_i < \gamma_i$, i should increase her rate to the "rightful" share γ_i while other users back down.

Lemma 3.3: **When the total traffic from users $j \in \{1, 2, .. ., I\}$, $j \neq i$, is such that even if user i transmits at a rate less than γ_i, the network is congested, NU_i is maximized at the point $\lambda_i^* = \gamma_i$ if and only if price π_i satisfies**

$$\frac{U_i'(\mu_i[\gamma_i + \delta^*])\mu_i'[\gamma_i + \delta^*]}{P_i'(\delta^*)} \leq \pi_i \leq \min(\gamma_i\eta_i, \gamma_i\zeta_i) \quad (5)$$

where $\eta_i = U_i'(ct_i) + D_i'(\gamma_i - ct_i)$ and $\zeta_i = U_i'(\mu_i[ct_i + \delta^+])\mu_i'[ct_i + \delta^+] + D_i'(\gamma_i - ct_i - \delta^+)$. $\mu_i[\gamma_i + \delta^*]$ is amount of user i's flow received at the destination node when $(\gamma_i + \delta^*)$ is the flow injected at source node. δ^+ is defined as $\arg\min_{0 \leq \delta \leq (\gamma_i - ct_i)} \gamma_i\left(U_i'(\mu_i[ct_i+\delta])\mu_i'[ct_i+\delta] + D_i'(\gamma_i - ct_i - \delta)\right)$ and $\delta^* = \arg\max_{\delta > 0}\left(\frac{U_i'(\mu_i[\gamma_i+\delta])\mu_i'[\gamma_i+\delta]}{P_i'(\delta)}\right)$. Further, NU_i is an increasing function of λ_i to the left of λ_i^* and a decreasing funct in of λ_i to the right of λ_i^*.

Remark: **The lower bounds in Equations 3, 4 and 5 are the ratio of user i's marginal utility and marginal penalty incurred at the carried traffic (just above λ_i^*) multiplied by the marginal carried traffic. If marginal utility or marginal carried traffic in going above λ_i^* is high and marginal penalty is low, then the user is already inclined to increase her rate. To prevent this from happening, a high price must be set. Hence, in this situation, the lower bound on π_i is large. On the other hand, if marginal penalty is high and marginal utility and marginal carried traffic are low, the user is already disinclined to increase her rate. Even a small price is sufficient in this case, and the lower bound on π_i is small.**

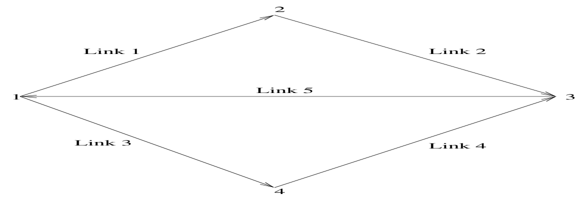

Fig. 2. Network considered for the experiment.

Now let us focus our attention on the upper bounds. Treating $\left(\frac{\pi_i}{\gamma_i}\right)$ as the effective "price per unit bandwidth", we see that the upper bound says that: Price per unit bandwidth should be less than the sum of user i's marginal utility when transmitting at the contracted rate and the marginal disutility at a positive shortfall. If the unit price exceeds the sum above, there is no motivation for the user to increase the flow up to even the contracted value.

Now we would compute the intersection of the three ranges.

The significance of this is that it may be possible to choose a value of π_i that will work (*i.e.,* lead to desirable behavior) irrespective of whether the congestion threshold is above, below or equal to contracted rate.

Lemma 3.4: **We can set $\pi_i = \frac{U_i'(0)}{\lim_{x \downarrow 0} P_i'(x)}$ irrespective of the congestion state of the network, if we choose congestion penalty function such that**

$$\lim_{x \downarrow 0} P_i'(x) \geq \frac{U_i'(0)}{\gamma_i \lim_{x \downarrow 0} D_i'(x)} \quad (6)$$

Thus, it is possible to set π_i appropriately without knowing ct_i and the SP can pick one price which lies within the intersection of the three ranges. Next we explore whether there is one price which lies within the intersection of the three ranges of *all* users sharing the network.

Lemma 3.5: **If we choose congestion penalty functions such that**

$$\lim_{x \downarrow 0} P_i'(x) \geq \frac{\max_k U_k'(0)}{\min_k \gamma_k \min_k \lim_{x \downarrow 0} D_k'(x)}, \forall i \in \{1, 2, \ldots, I\} \quad (7)$$

where the maximization and minimization are done over $k = \{1, 2, ..., I\}$, one single value of price for all users, irrespective of the network behavior, is sufficient to enforce desirable system operation. That value of price is

$$\pi_i = \frac{\max_k U_k'(0)}{\min_k \lim_{x \downarrow 0} P_k'(x)}, \qquad \forall \;\; i \in \{1, 2, \ldots, I\} \quad (8)$$

The proofs are based only on simple algebra. Conclusion is although all users are only interested in maximizing their net benefits, one price — the same for all — compels them to behave in the system optimal manner also.

IV. Simulation

In this section we present the results of a MATLAB simulation. In this experiment, 4 users share the network shown in Figure 2. Each user is associated with one flow. All the assumptions are tabulated in Table I. Note that all congestion penalty functions obey Equation 7 and hence a contracted price of 1 unit is applicable to all users.

The SP conveys the contracted price, penalty function and whether the network is congested or not to the users. The SP computes $y^{(i)}$'s, $d^{(i)}$'s and the $\mu_i[\lambda_i]$'s by solving maximmization of sum of weighted carried traffic of all users subject to flow

TABLE I
ASSUMPTIONS OF THE EXPERIMENT

	s_i	z_i	γ_i	$U(x)$	$D(x)$	$P(x)$
User 1	1	3	1.0	x	e^x	$1 + x^2 + 2x$
User 2	1	4	1.5	$log(1 + x)$	x	e^x
User 3	3	2	2.0	x	$x^2 + 2x$	$1 + x$
User 4	3	4	1.8	$1 - e^{-x}$	$(1 + x)^3$	e^x

conservation and capacity constraints. In our example all users have same weight.

Now if at a later point of time, the set of underusers changes — because some underusers now have enough data or a new group of users are now data-limited — the SP re-routes the flows by re-solving the maximization problem in order to maximally utilize the network.

At the start of the simulation, we assume that the initial data rates are $\lambda = [2.0\ 1.0\ 1.8\ 1.5]$. At this point we assume that all users have unlimited data to sustain any rate. At each iteration, all users perturb their present data rates and compute new net utilities at these perturbed rates. If the net utility increases on increasing rate, users increase their flow rates in the next iteration; similarly, if the net utility increases on decreasing rate, users decrease rates in the next iteration. Here we assume the step-size to be 0.01 for all.

We find that within 100 iterations, all users converge to their respective contracted rates in the process of maximizing their net utilities. This is because the links are provisioned such that the network has just enough capacity to accommodate all users at their respective γ_i's (Figure 3).

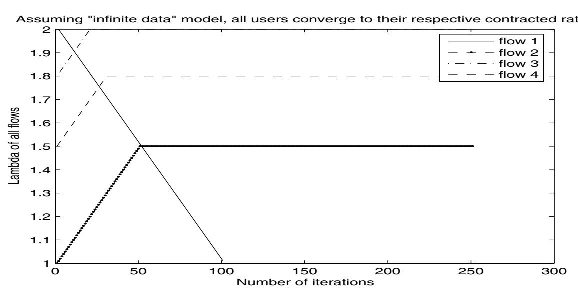

Fig. 3. First part of the experiment.

After 250 iterations, we cap the flow rates of users 1 and 4 to 0.1 and 0.2 respectively and let users 2 and 3 vary rates to increase their net utilities as described earlier. Figure 4 shows that by the 400^{th} iteration, both users 2 and 3 increase flow to a value beyond contracted.

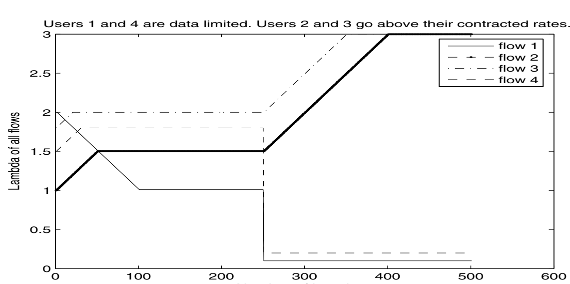

Fig. 4. Second part of the experiment.

From 500^{th} iteration onwards, we again let all users vary their rates. The pricing scheme compels overusers to back down, so that users 1 and 4 regain their shares (Figure 5).

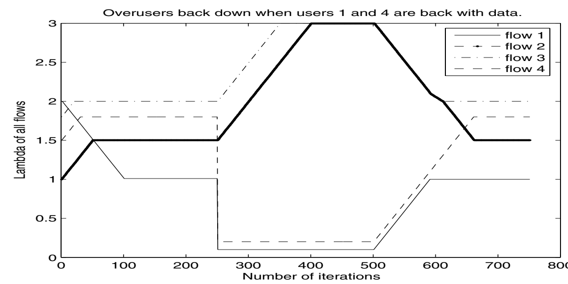

Fig. 5. Concluding part of the experiment

This experiment illustrates that it *is* possible to assign prices once and for all at the beginning, such that fairness and efficiency is maintained.

V. STABILITY

When no user is constrained by limited amounts of available data, what is the rate vector that the collection of users converges to? If the increase in rate of any user causes a non-zero drop in the network, *i.e.*, due to increase in λ_i for any i, there is a $1 \leq k \leq N$ and a $1 \leq j \leq I$, such that $d^{(j)}(k) > 0$, we say that the network has no spare capacity. When prices are set according to Equation 8, $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_I^*)^t$ is a Nash equilibrium if, $\lambda_i^* \geq \gamma_i$ for all $1 \leq i \leq I$ and there is no spare capacity in the network. This can be seen as follows. Considering any $1 \leq i \leq I$, we find that the user is either at contracted rate or already an overuser and the network is fully utilized. Therefore, Lemma 3.1 or 3.2 applies. This means that user i has no incentive to change her rate from λ_i^*, and this conclusion applies to all the users.

VI. CONCLUSION

We considered sources that could occasionally be constrained by limited amounts of available data. We were interested in a pricing scheme that would ensure fair and efficient sharing of the network resources.

We introduced the idea of disutility for underusers and noted that the disutility term encourages underusers to increase their rates whenever they have sufficient data. We showed that under certain conditions *one* price motivates all users to behave in fair and efficient manner.

We recognize that we have not explicitly considered the problem of revenue maximization for the SP. While our goals of fair and efficient sharing of the network are natural, we would like to consider explicit revenue maximization as well.

REFERENCES

[1] A. Delenda, "Mécanismes dènchere pour le partage de ressources télécom," France Telecom – R & D, Tech. Rep. Tech Rep 7831, 2002.
[2] F. Kelly, A. K. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar 1998.
[3] F. Kelly, "Charging and rate control for elastic traffic (corrected version)," *European Trans on Telecom*, vol. 8, no. 1, pp. 33–37, Jan 1997.
[4] A. A. Lazar and N. Semret, "Design and analysis of the progressive second bid auction for network bandwidth sharing," *Telecom Systems – Special Issue on Network Economics*, 1999.
[5] P. Maillé and B. Tuffin, "Multi-bid auctions for bandwidth allocation in communication networks," in *Proc of IEEE Infocom*, April 2004.
[6] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans on Networking*, vol. 8, no. 5, pp. 556–567, Oct 2000.