# Price Discovery in Dynamic Capacity Contracting Framework

Gurjeet S. Arora, Murat Yuksel, Shivkumar Kalyanaraman, Thiagarajan Ravichandran, Aparna Gupta
Rensselaer Polytechnic Institute, Troy, NY
garora@emc.com, yuksem@rpi.edu, shivkuma@ecse.rpi.edu, ravit@rpi.edu, guptaa@rpi.edu

*Abstract*—**Congestion sensitive pricing for providing better than best effort service has received significant attention in the last decade. One obstacle thus far has been the fact that prices for premium network services are not known at the time of service contracts but are determined by the congestion conditions during the period of contract. In this paper we identify a robust parameter for capturing congestion conditions in a edge to edge framework. Further, we propose a adaptive pricing scheme based on the pricing parameter identified, that can be used for providing prices at the time of contracting. The pricing scheme operates in cycles of queue drain and queue build up trying to discover the optimal price with changing network conditions. We evaluate the performance of the pricing function proposed in the Dynamic Capacity Contracting framework. This pricing function offers parameters that can be tuned (tunables) so that the edge queues are controlled in limits that ensure that service level agreements are met. Selection of these tunable parameters for a given Customer-ISP setting is based on trade-offs between performance metrics and relative importance attached to them. The scheme is purely adaptive in the sense that it does not assume any specific user behavior. It is readily deployable as it needs $q_{i-1}$ and $C_i$ to determine the price in period i for premium services (both parameters are available at ingresses in edge to edge framework. We evaluate the conditions under which the pricing scheme ensures stability. A user model is then proposed and the interaction of the pricing scheme with a user model is examined. We then observe the response of the system to step increase in demand (*congestion*) and a sustained increase in demand. We conclude with some observations on how allocation of scarce premium service is carried out between users having different utility functions.**

*Keywords*— **Network Pricing, Congestion-Sensitive Pricing, Congestion Control, Quality-of-Service**

## I. Introduction

A method for network resource allocation and congestion control that has become increasingly popular is congestion sensitive pricing. The idea behind congestion sensitive pricing is that network performance cannot be solely derived from congestion control protocols[1]. During periods of resource contention or congestion epochs, there is a need for distinguishing one packet from the other based on the utility functions attached to it (these utility functions are typically a combination of user preferences and application requirements). Thus during congestion, increasing prices of network services makes for a allocation that is closer to pareto efficient (i.e allocates resources to users such that the overall user surplus is close to the highest possible from a given capacity) and gives the users the right incentives to adjust their demands to alleviate congestion[2].

MacKie Mason and Varian [3] propose a pricing scheme in which they propose that in the event of congestion, usage of each unit of network resource should be priced at the total marginal congestion cost that it imposes on other users. This price is called Shadow price. The idea behind shadow pricing is to emphasize the fact that each user should face the same price for usage of network resources : the sum of the marginal congestion costs that each user imposes on other user. Shadow prices are closely linked with user utility functions which are difficult to determine and change frequently. Gibbens and Kelly [2] intoduce a approximate packet marking scheme in which the number of packets marked at a congested network resource are used to convey the shadow price to users and provide them with correct incentives to use the network efficiently. These approximate shadow prices determined by counting the number of marked packets actually determine price for the network resources that have already been used (contacted for). It is however, desirable that users have priori knowledge of the price that they will pay for usage of network services. In this paper we identify an important parameter for pricing premium services in the Dynamic Capacity Contracting framework [4] employing Edge to Edge traffic control. Based on the observations made, we propose a pricing scheme to adaptively price network resource at the beginning of a contract period based on congestion conditions in the past period.

In section II we describe a likely scenario for evolution of bandwidth markets. This is followed by a brief overview of Dynamic Capacity Contracting and Edge to Edge Control in section III and IV respectively. In section V, we take a look at the problem of pareto efficient network resource allocation in the face of congestion while maintaining high utilizations using Dyanamic Capacity contracting in a Edge to Edge framework. In section VI, we show the the queue at Edge can be used as a effective parameter for pricing network services as it represents the cumulative difference between demand and capacity from the last time the queue was empty. In section VII, we propose and evaluate stability conditions for a pricing scheme that offers the tunables to adjust to different ISP and user requirements. In section VIII, we show how the various parameters presented can be tuned. This follows, experimental results to show how the scheme re-acts to congestions of different magnitude and how the parameters of pricing function can be tuned to get desired queue lengths and other operating conditions (Section IX). Premium service allocation done between two users for different base demand (demand at price=0) and reservation price is evaluated in Section X. Conclusions and directions for further work are presented in section XI.

## II. Likely Evolution of Bandwidth Markets

This section presents a likely scenario for evolution of Bandwidth Markets. The players that engage in transactions involving bandwidth can be grouped in the following three categories : Capacity providers, Resellers (Internet service providers and other bandwidth service providers) and End consumers.

The transactions between the Capacity providers and Resellers are best handled in a exchange kind of environment [5]. This is because it offers the best efficincy by providing the bandwidth capacity providers and resellers a variety of options (access to a large number of capacity providers and resellers). Another big advantage of exchanges is tranparent pricing which encourages healthy competition. This kind of pricing where the transactions are for a large amount of bandwidth which can be resold by the reseller in packages providing a variety of services to end users is what we call *bandwidth pricing*.

Resellers create value by catering to the variety of needs of end-users (their diverse applications which vary in terms of their delay and jitter sensitivities). Bandwidth among its other uses, can be used for internet services as well. We call the domain of pricing bandwidth for providing internet services to end users *internet pricing*. In this paper we focus on the problem of allocating premium internet services provided by dynamic capacity contracting in edge to edge congestion control framework.

## III. Dynamic Capacity Contracting (DCC)

Dynamic Capacity Contracting uses short term contracts for premium network services. Short term contracts are used so that prices can be revised depending on the changes in congestion condition of the network which happen on a short timescale. DCC is a pragmatic (implemantable in Diff-serv architecture) intermediate approach between Clark's expected capacity model and the smart market model [6].

In this paper we propose a pricing fuction that uses the event of edge queue exceeding certain threshold (to mark the beginning of a congestion epoch) to increase price for premiun network services provided by the DCC model. The price is determined adaptively from one contact period to another as the congestion conditions of the network change. The pricing algorithm is readily deployable in edge to edge framework as the parameters it needs are readily available in edge to edge framework at the beginning of a contracting period.

## IV. Edge to Edge Congestion Control framework

The basic architecture of *edge to edge traffic control* works at the network layer and involves pushing congestion back from the interior of a network, distributing it across edge nodes where the smaller congestion problems can be handled with flexible and cheaper methods [7]. DCC pricing scheme in the framework being examined in this paper provide the correct incentives to users sending packets to the edge to adjust there demand. Thus, the queues at the edge are kept at levels consistent with the service level agreements for the premium service. Another, advan-

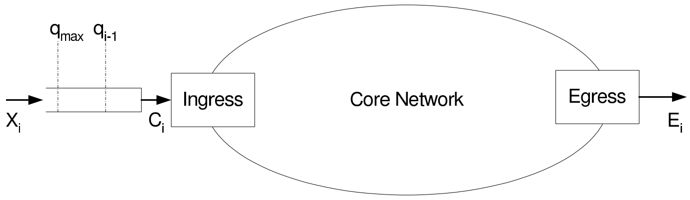


Fig. 1. Edge to edge framework at the beginning of period i.

tage that accrues from using this edge to edge control scheme to implement Dynamic Capacity Contracting is that the core routers are not assumed to have capabilities to forward packets differentially. The network delays can be determined with much greater accuracies as the congestion is pushed to the sophisticated edge routers. Thisserves as a building block for the premium service provided by DCC (which ensures assured forwarding and bounded delays).

The basic idea behind edge to edge congestion control is that the rate of tarffic inflow at the egress, during congestion epochs is kept slightly lower than the rate of departure of packets measured at the egress (thus the internal queues that have built up are drained).

In this paper, we will be using the edge to edge congestion control scheme described in [7]. The parameter used to measure the severity of congestion in the pricing scheme is the queue length at the ingress. We use the rate limit at the ingress at the beginning of a period as the network capacity available in a period in the simulations and analysis presented. The parameters relevant to our analysis in edge to edge framework are shown in Figure 1. $q_{i-1}$ is the length of edge queue at the end of period $i-1$ and at the beginning of period $i$. $C_i$ is the rate limit in period $i$ calculated at the ingress based on edge to edge control algorithms used. $E_i$ is the rate at which packets leave the egress. This will be slightly higher than $C_i$ during congestion epochs. $X_i$ is the total demand during period $i$.

## V. Issues in Congestion Sensitive Pricing

In this section, we first formulate the utility maximization problem for a user consuming premium network service provided using DCC. This premium DCC service is provided using edge-to-edge congestion control in diffserv framework.

In DCC each user contracts for network resources at the beginning of a period at a price that is declared by the service provider at the beginnign of a period. We now formulate the objective function for a individual user who wishes to maximize his benifits from the network services in a period $i$.

Suppose that there are a total of $M$ users. The utility function of a user $j$ (where $1 \leq j \leq M$) in period $i$ is given by $U_{ij}$. Let the number of packets sent by user $j$ in period $i$ be $x_{ij}$. Then the single period utility maximization problem for user $j$ in period $i$ can be stated as :

$$\max_{x_{ij}} U_{ij}(x_{ij}, q_{i-1}) - p_i.x_{ij}$$

subject to:

$$x_{ij} \geq 0 \tag{1}$$

This leads to the first order condition :

$$\frac{dU_{ij}(x_{ij}, q_{i-1})}{dx_{ij}} = p_i \tag{2}$$

The expression on the left in the above equation is also called the reservation price. Thus, we see that in the process of maximizing its individual benifits from the network each user continues to use the network resource until its reservation price equals the actual price.

In order to make the goals of individual users concur with the premium service level agreements guaranteed by DCC. The price in period $i$ should be set such that :

$$\sum_{j=1}^{M} x^{\star}_{ij} = C_i \tag{3}$$

where, $x^{\star}_{ij}$ is the solution of Problem 1 that satisfy Equation 2.

The solution obtained from Equation 3 is the ideal solution to the problem of allocating network resources as it not only ensures that users with highest utilities get the premium network services but also assures that the network resources are fully utilized.

However, it is hard to determine the user utility functions and they change very often. The problem is that if the price advertized (determined by an alternate means) is higher than the optimum price that satisfies Equation 3, the contracted capacity will be less than capacity available. This will decrease the resource utilization and we will be wasting the resouce that we wanted to allocate efficiently. If the price is lower than the one that satisfies Equation 3, total user demand will exceed capacity and it will be hard to maintain premium service conditions.

The approach that we take in this paper is to have a queue at the edge and use a adaptive pricing function to control the length of the queue. Clearly, the longer the permissible queue, the higher the average utilization but the greater the delay that each packet encounters. Thus there is a trade-off, some people would like to have the best service even if that means lower utilizations. Other ISPs would like to maintain higher utilizations (bigger customer base) even if that implies that the delays that packets encounter increase slightly (proportional to the average edge queue length).

In this paper, we propose a adaptive scheme that tries to keep queues at the edge under control so that service level agreement conditions are met. Further, our pricing function provides tunables that can be used to select a desired level of utilization and service quality, depending on the customer and ISP metrics and the trade-offs they are ready to make. In the next section, we look at a pricing parameter that we use in our adaptive pricing scheme.

## VI. Selection of system parameter for pricing in DCC framework

The state of the network system at the beginning of contacting period $i$ using edge to edge control is completely captured by specifying the queue length carried over from previous period $q_{i-1}$ and capacity $C_i$ from DCC perspective. $q_{i-1}$ represents the cumulative difference in demand for premium services and capacity since the last period in which queue length was equal to zero. Thus, it is natural to expect $q_{i-1}$ in any pricing parameter chosen for congestion sensitive pricing.

To investigate the behavior of the system further, we set-up a single period optimization problem and solved it using optimizer functions built in Matlab. We then investigated the relationship between the optimal prices obtained for a period $i$ and the state in period $i-1$ (which is known when price for period $i$ is to be determined.

The objective in the problem formulated is to maximize the utilization of network resources in a period $i$, given the distribution of the total budget of all users (all users are assumed to be in the same class in the sense that they have the same reservation price, for tractability reasons. The total budget of all users was used in the formulation as it can be estimated with better confidence as compared to individual user budgets. ), distribution of available network resources available for contracting in period $i$, maximum permissible queue length (expected to be determined by service level agreement specifications) and queue length in period $i-1$.

The single period optimization problem solved using fmincon (optimizer in Matlab for multi-dimensional constrained optimization) can be formulated as :

$$\min_{p_i} \sum_{k=1}^{1000} 1 - [B_k(1-\frac{p_i}{\hat{p}})]\frac{1}{p_i C_k}$$

subject to:

$$q_i = q_{i-1} + B_k(1-\frac{p_i}{\hat{p}})\frac{1}{p_i} - C_k \le q_{max} \forall k \in [1, 1000] \tag{4}$$

where,
Number of observations (to take a sufficiently large sample from Budget and Capacity distributions), $k$ $= 1000$. Total budget of all users in observation $k$, $B_k \in U[20, 50]$. Capacity in observation $k$, $C_k \in N[\mu = 98, \sigma = 2]$ truncated in the range $[96, 100]$. Maximum permissible queue length, $q_{max} = 50$.

User adaptation to changing demand is captured by changing the budget according to the following simple linear rule:

$$B_{k,spent} = B_k(1-\frac{p_i}{\hat{p}}) \tag{5}$$

The primary motive of the above numerical optimization exercise is to observe the behavior of optimal price if all the needed information (like budget and user adaptation model) was available. Thus, any reasonable model for user adaptation is good for the task at hand. After optimization, we try to see how well a job could we have done at predicting the optimal price from the information that we know at the beginning of a contracting period (like $q_{i-1}$ and $C_i$) without using the parameters that we assume to formulate the problem.

We present the results of optimization for different values of $q_{i-1}$.

Regression results of best fit linear model between $q_{i-1}/C_{i,mean}$ (independent variable) and $p_i^\star$ obtained are as under :

The regression equation is :

$$p_i^\star = 0.281 + 0.223\frac{q_{i-1}}{C_{i,mean}} \tag{6}$$

TABLE I
OPTIMIZATION RESULTS : $p_i^\star$ FOR VARIOUS $q_{i-1}$.

| $q_{i-1}$ | $q_{i,mean}$ | $p_i^\star$ | $q_{i-1}/C_{i,mean}$ |
|---|---|---|---|
| 0 | 13.11 | 0.286 | 0.00 |
| 5 | 14.28 | 0.296 | 0.51 |
| 10 | 14.15 | 0.304 | 0.10 |
| 15 | 15.02 | 0.314 | 0.15 |
| 20 | 16.73 | 0.324 | 0.20 |
| 25 | 16.97 | 0.332 | 0.26 |
| 30 | 17.72 | 0.346 | 0.31 |
| 35 | 17.52 | 0.360 | 0.36 |
| 40 | 18.50 | 0.371 | 0.41 |
| 45 | 19.92 | 0.388 | 0.46 |
| 50 | 20.16 | 0.401 | 0.51 |

The small value of coefficient of $q_{i-1}$ should not be confused as a week relationship. We shall use a F-test to see if the relationship holds at 5% type I error.

TABLE II
REGRESSION ANALYSIS : $p_i^\star$ VERSUS $\frac{q_{i-1}}{C_{i,mean}}$

| Predictor | Coefficient | SE Coefficient | T | P |
|---|---|---|---|---|
| Constant | 0.281 | 0.0022 | 125.54 | 0 |
| $\frac{q_{i-1}}{C_{i,mean}}$ | 0.223 | 0.0074 | 30.10 | 0 |

### A. F Tests to Validate the Association Between $p_i^\star$ and $\frac{q_{i-1}}{C_{i,mean}}$

Let us call the coefficient of $\frac{q_{i-1}}{C_{i,mean}}$ in the above linear relationship, $\beta$. Then, the null hypothesis and the alternate hypothesis for the test are :

$$H_0 : \beta = 0 \qquad H_1 : \beta \neq 0 \tag{7}$$

We limit the probability of Type $I$ error (rejecting a null hypothesis when it is true), $\alpha$ at 0.05. Number of observations, n for our simulation is 11. Using the decision rule for checking above hypothesis, based on F statistic, we find that:

$$F^\star = \frac{MSR}{MSE} = 905.92$$

[From Table III] and, from tables for F-statistic:

$$F(1-\alpha; 1, n-2) = F(0.95; 1, 9) = 5.12$$

We conclude $H_1$ as:

$$F^\star = 905.92 > F(1-\alpha; 1, n-2) = 5.12$$

The fact that there is a strong association between $\frac{q_{i-1}}{C_{i,mean}}$ and $p_i^\star$ implies that any adaptive scheme should use the predictor variable $\frac{q_{i-1}}{C_{i,mean}}$ to determine $p_i$. Notice, that the predictor variable does not require any knowledge of total user budgets which were assumed to be known during the simulation. Further, notice that the predictor variable for a period can be estimated at the beginning of the period as queue length at the end of previous period is known and the estimated capacity available in a peiod is known at the beginning of the period in the edge to edge framework based on the round trip statistics that the ingress maintains. In the next section we will propose a adaptive pricing function with tunables that enable the provider to choose a suitable trade off between packet delays because of edge queues and network resource utilization. Another trade-off that is made in the process of choosing the tunables is system responsiveness versus price variations.

## VII. ADAPTIVE PRICING ALGORITHM WITH TUNABLES

Typically an ISP providing premium services using DCC on Edge to Edge framework would like to control the edge queue in a range $[q_l, q_h]$. The precise selection of this range would be determined by the trade-off that the ISP and the users are ready to make between delays and utilization (service level agreement). For instance, higher levels of network resource utilization and lower prices are possible if higher edge queues are maintainedd. This of course, comes at the cost of higher delays because of the edge queue.

It is important that the pricing function does not allow the queues to grow far beyond the higher threshold, $q_h$. For, this reason and to be able to control rising demands in the periods of congestion by providing the right pricing feedback, we increase price multiplicatively proportional to the difference between current queue length and the higher queue threshold, $q_h$. When the queue length falls below the lower threshold, we start decreasing the price but this time with a constant factor (additively) so that queue length can rise back in the desirable range.

The pricing fuction based on the above observations and the pricing parameter identified in the previous section can be specified as:

TABLE III
ANALYSIS OF VARIANCE: $p_i^\star$ VERSUS $\frac{q_{i-1}}{C_{i,mean}}$

| Source | Degrees of Freedom | Sum of Squares | Mean Sum of Squares | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 0.0143 | 0.0143 | 905.92 | 0.00 |
| Residual Error | 9 | 0.0001 | 0.0000 | | |
| Total | 10 | 0.0144 | | | |

`If` $q_i - q_h > 0$ `then`

$$p_{i+1} = p_i + \alpha_1 \frac{(q_i - q_h)}{C_i} \tag{8}$$

`If` $q_i - q_l < 0$ `then`

$$p_{i+1} = p_i - \alpha_2 \tag{9}$$

`If` $q_i \in [q_l, q_h]$ `then`

$$p_{i+1} = p_i \tag{10}$$

Now, we find the condition for the tunables that ensure that the system remains stable i.e. edge queues do not grow unboundedly.

Let us assume that the highest reservation price that a user has is $\hat{p}_{max}$. Let $x_{ij}$ be the capacity contracted by user j in period i. From these statements it follows that:
At $p_i = \hat{p}_{max} : \sum_1^M x_{ij} = 0$ (by definition of $\hat{p}$)
The above statement implies that under above condition, the entire capacity is being used to drain the queue.
For $\forall\ \hat{p}_{max} > 0, \exists \alpha_1$
s.t. $\hat{p}_{max} \leq \alpha_1 [q_{max} - q_h]$
where, $q_{max}$ is the capacity of the edge queue (its total size, not $q_h$).

The above statement, gives us the condition for ensuring stability:

$$\frac{\hat{p}_{max}}{q_{max} - q_h} \leq \alpha_1 \tag{11}$$

Thus we conclude that as long as $\alpha_1$ satisfies the above condition, the pricing algorithm will guarantee stability. However, a preferred manner of guaranteeing a similar effect is to provision for very long edge queue though chances of them being fully used are rare.

In the next section we develop a user model for analysis of the adaptive algorithm and explore some properties of the tunables in the algorithm of this section and what parameters of the system they help control.

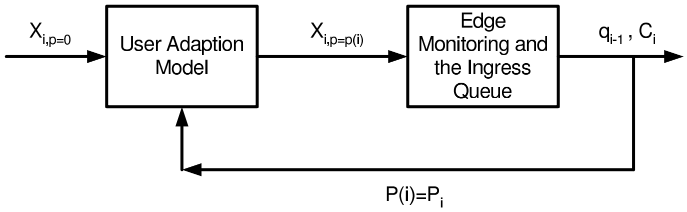


Fig. 2. General model of user adaptation to price

## VIII. A SIMPLE USER MODEL AND ITS INTERACTION WITH THE PRICING SCHEME

In general, the user model of adaptation to prices can be depicted as in Figure VIII.

In the figure, $X_{i,p}$ is the user demand for network services when price for them is p per unit. This is determined by the user adaptation model which takes the demand when price is zero, $X_{i,p=0}$ and the current price for network services $p_i$ as inputs. The later of course is determined by the pricing algorithm defined in the previous section based on the system state variable at the beginning of period i. The state of the system at the beginning of a contracting period can be completely defined by the queue carried over from the last period $q_{i-1}$ and capacity in period i, $C_i$ (Figure 1).

Notice that the user model that we propose here will be used for evaluating the performance of the pricing algorithm proposed in this paper and the impact that pricing algorithm tunables have on the system performance only. The adaptive pricing algorithm is not dependent on the specific form of user adaptation curves.

Let, $\hat{p}$ be the price at which the user becomes indifferent to using the service or keeping the money, also called the reservation price. We know that at the reservation price the user demand for network service is zero. Using this fact, we use this simple but reasonable user adaptation function for evaluation of the pricing algorithm.

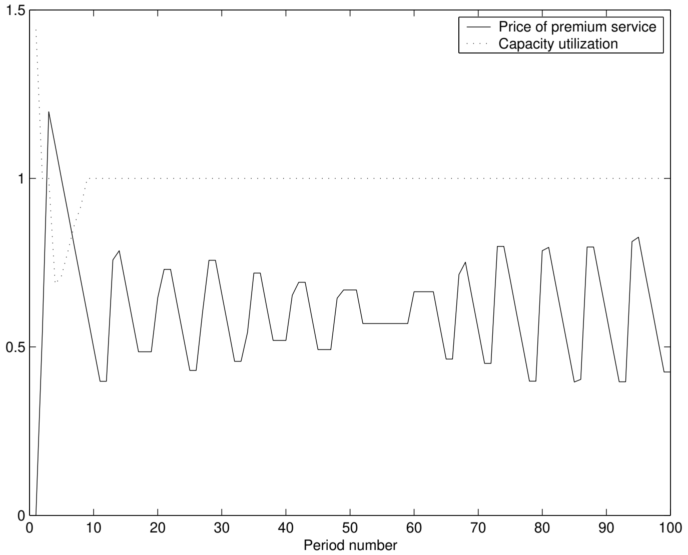


Fig. 3. Price and Utilization for system parameters $(\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p} = 2)$

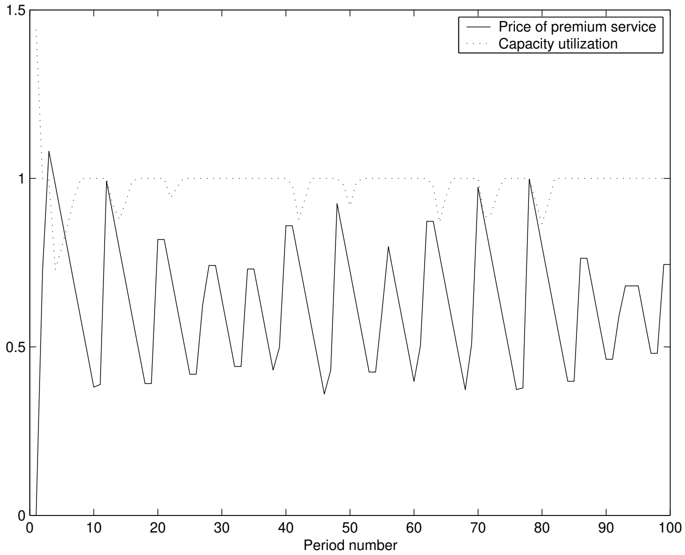


Fig. 4. Price and Utilization for system parameters $(\alpha_1 = 4, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p} = 2)$

$$X_{i,p=p_i} = X_{p=0}\frac{\hat{p} - p_j}{\hat{p}} \tag{12}$$

Now, we use the above user adaptation model to show the impact that the various tunables offered by the pricing algorithm have on the system. We use the system performance with system parameters : $(\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p} = 2)$ as reference for performance evaluation (Figure 3). Unless otherwise specified $X_{p=0} = 140$.

## A. *Effect of* $\alpha_1$

By increasing $\alpha_1$ the response to congestion ($q_i \geq q_h$) can be made more aggressive. This also implies that on an average the edge queue length will be smaller and utilization will be lower for high $\alpha_1$. Comparing Figure **??** with 4 we see that an increase in $\alpha_1$ from 3 to 4 results in faster variation in price. This is because the system has been made more responsive to queue length increases. High $\alpha_1$ is a good idea if load on system can ramp up rapidly as the system is very responsive, but this comes at the cost of increased price changes and a possible lower utilization. Thus, keeping $\alpha_1$ as low as possible satisfying the service level agreements and keeping in mind the expected variability in the load is recommended ($\alpha_1$ should be high enough to give the users enough incentive to restrain their demand when edge queue starts building up).

## B. *Effect of* $\alpha_2$

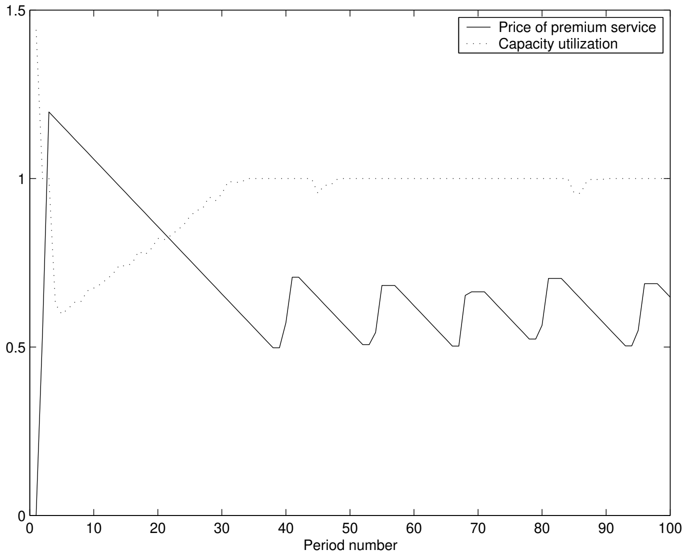


Fig. 5. Price and Utilization for system parameters $(\alpha_1 = 3, \alpha_2 = 0.02, q_h = 25, q_l = 15, \hat{p} = 2)$

By decreasing $\alpha_2$, the range in which price varies between *queue drain* (demand control) and *queue build up* (demand probe) cycles can be decreased. This of course comes at the cost of compromising the capability to slash prices quckly if demand becomes very less. This is seen in Figure 5 from period 4 to 37 when utilization drops significantly and prices are slashed at a rate slower than that in Figure 3.

By increasing $\alpha_2$, we see that the system becomes more responsive to demand changes. This also causes a increased fluctuation in the prices. (Figure 6)

It is important to ralize that, demand variation, quality of service requirements etc are characteristics of the users. The pricing scheme proposed here provides the ISP with the tunables to balance user requirements and their goals, like high utilization. These user preferences are going to differ from ISP to ISP (depending on their client base) and thus there will be need for selecting different values of the tunables. A way of looking at this is : each value set of the tunables represents some trade-offs, the precise trade-offs made would depend on the ISP and user requirements (the weightage given to different user and ISP performance metrics).

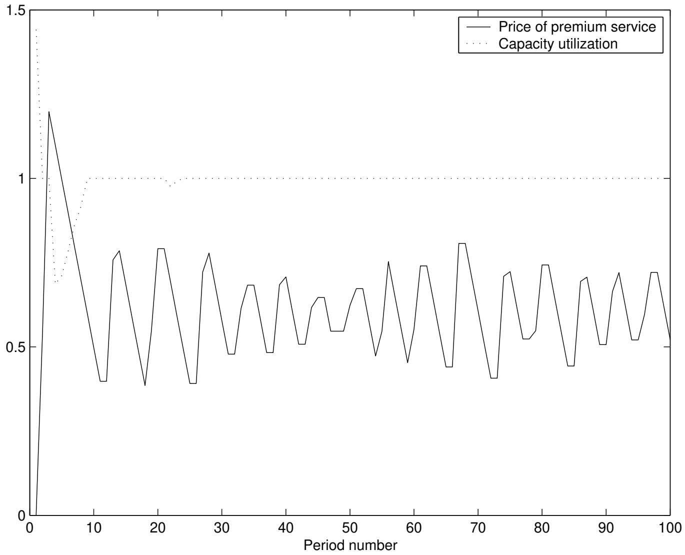


Fig. 7. Price and Utilization for system parameters $(\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 20, \hat{p} = 2)$

### C. *Effect of* $q_l$

Having a higher $q_l$ would increase the utilization of network resources in the face of demand variability. This will happen, at the cost of maintaining higher queue lengths at the edge. Thus, the trade-off here is between, very good service (low $q_l$) and higher utilization (higher $q_l$). The precise value chosen for $q_l$ would depend on the preferences of the participants(users and ISPs). For system parameters of Figure **??**, the average length of the queue on the edge is19.22 as opposed to 17.52 for Figure **??**.

### D. *Effect of* $q_h$

Increasing $q_h$ results in increase in the average queue length at the edge as the queue drain or demand control part of the cycle now starts later than a setting with lower $q_h$. There is an increase in utilization at the cost of the quality of service provided. Notice that this action is equivalent to resetting the definition of congestion i.e. actions taken to control congestion are now triggered at a different indicator level.

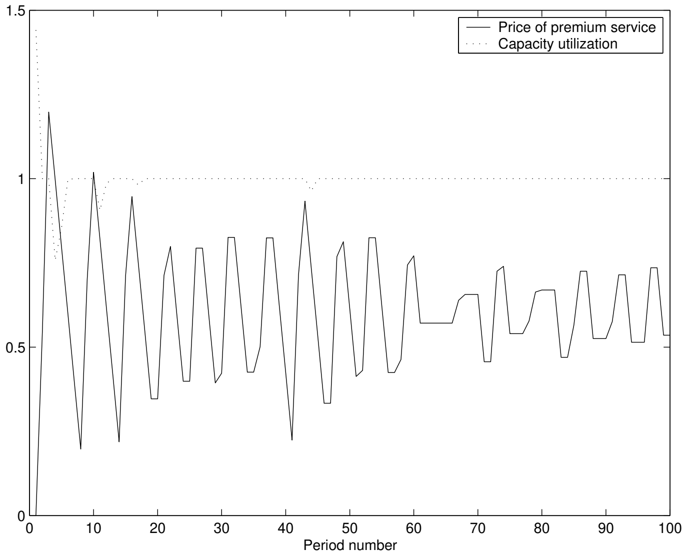


Fig. 6. Price and Utilization for parameters $(\alpha_1 = 3, \alpha_2 = 0.2, q_h = 25, q_l = 15, \hat{p} = 2)$

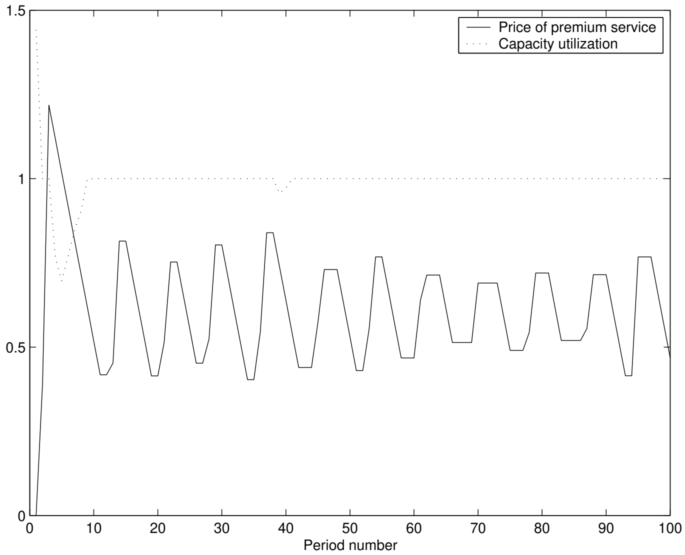


Fig. 8. Price and Utilization for system parameters $(\alpha_1 = 3, \alpha_2 = 0.1, q_h = 30, q_l = 15, \hat{p} = 2)$

### E. *Effect of distance between $q_l$ and $q_h$*

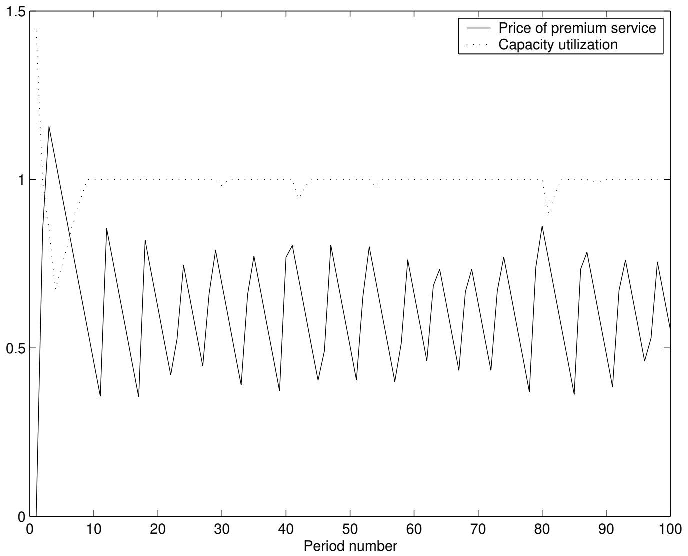


Fig. 9. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 15, q_l = 15, \hat{p} = 2$)

In Figure 9, we set $q_l$ and $q_h$ equal to each other. As is evident this causes very frequent variations in price and the average magnitude of these variations is higher than that of Figure 10.

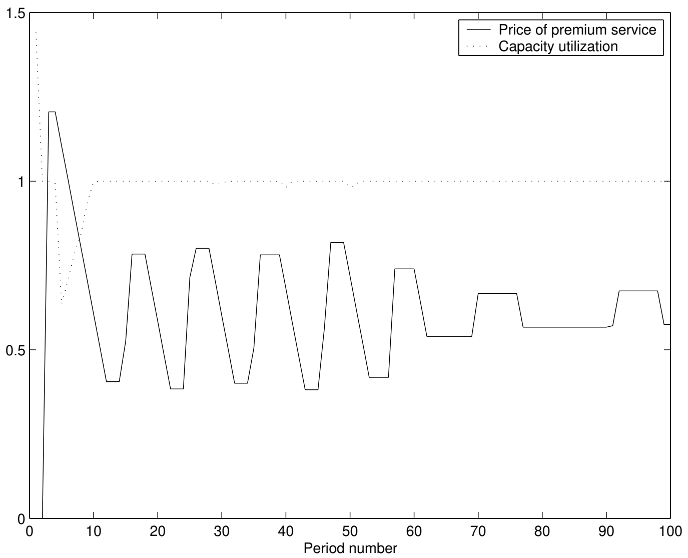


Fig. 10. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 45, q_l = 15, \hat{p} = 2$)

$q_l$ cannot be too low to avoid underflows and thus the resulting lower utilization levels. $q_h$ cannot be too high as service level agreements need to be met. Thus, there are limitations to how much we can gain on this price smoothing. Moreover, this price smoothing comes from accepting large variabilities in queue length which would imply large variability in service quality. Thus, the need to choose a distance between $q_l$ and $q_h$ that represents the needs of users and objectives of the ISP (a trade-off that meets the demands of participants the best).

## IX. Evaluating the algorithm robustness under different loads

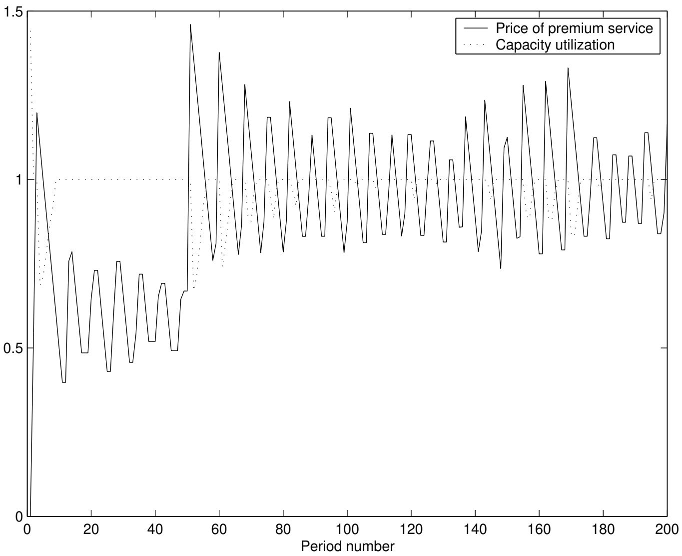


Fig. 11. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p} = 2, X_{p=0} = 140, \Delta X_{p=0} = 50u_t(50)$)

In Figure 11, we subject the system to an extra user load equivalent to increment of demand at price zero, $X_{i,p=0}$ by 50 units in period numbers [50,200]. The extra load can be specified as :

$$\Delta X_{p=0} = 50u_t(50) \tag{13}$$

where,
$u_t(50) = 1$ if $t \geq 50$
$u_t(50) = 0$, otherwise.

In Figure 11 we see that increasing the demand for premium service is responded to by increasing the price so that users have the correct incentives for regulating their demands. The average utilization in Figure 11 is 0.980 as compared to 0.994 for Figure **??**. The scheme reacts in a desirable manner, keeping utilization high while allocating resources efficiently.

In Figure 12, we increase the demand at price zero from 140 to 190 in period $i \in [50, 100]$. We see that the prices, go up during the period of higher demand

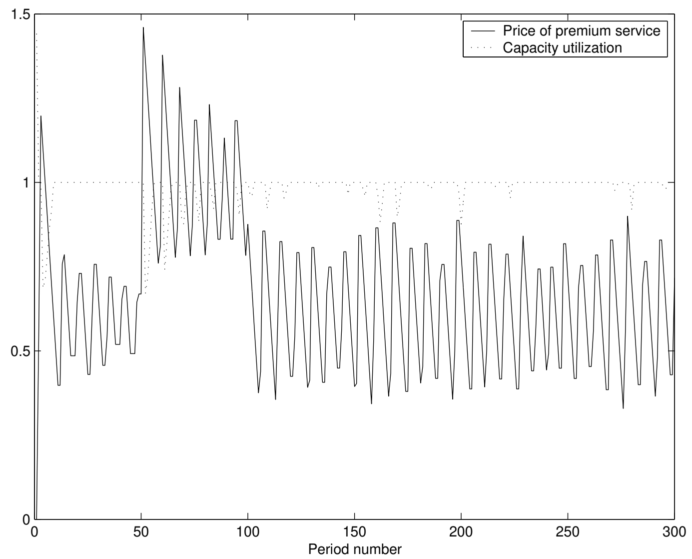


Fig. 12. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p} = 2, X_{p=0} = 140, \Delta X_{p=0} = 50[u_t(50) - u_t(100)]$

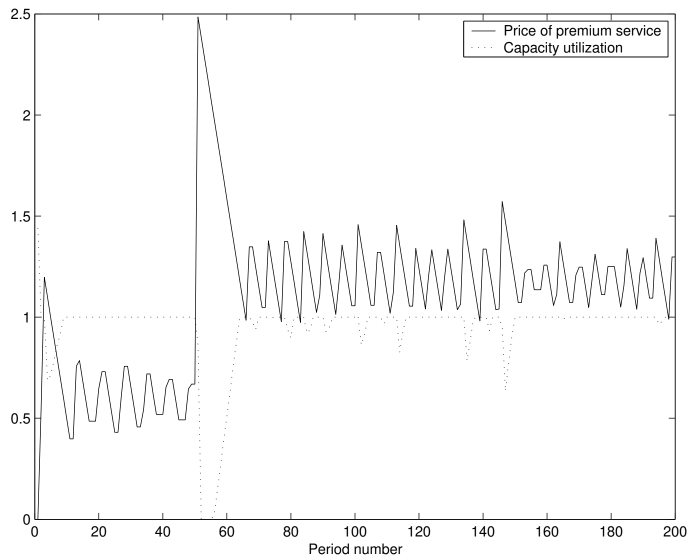


Fig. 13. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p} = 2$), $X_{p=0} = 140$, $\Delta X_{p=0} = 100 u_t(50)$

and come down when the demand subsides. Hight level of utilizations are maintained throughout except for during transition because of the unanticipated change in demand involved.

In Figure 13, the system is subjected with an increase in demand similar in form to that of Figure **??**. The only difference is that this time the demand disturbance is higher in magnitude (double). As the system was not ready (tuned) for such a spike in demand, we see that there is a short period of low utilization when the extra demand is introduced.

In Figure 14, we look at the same demand spike as in Figure 15. The only difference is, this time, one of the tunable parameters, $\alpha_2$ has been tuned to permit for more aggressive demand probing. This results in much lower penalties in terms of lower utilization during transition. The average utilization over 200 periods for Figure 13 is 0.941 and that for Figure 14 is 0.965.

In Figure 15 we subject the changed (tuned) system of Figure 14 with a step increase in demand equal to 200 starting at period number 50. This creates a longer time period of low utilization because of the huge magnitude of the increase in demand (about 150% increase in demand).

In summary, we see that there is a range in which the tunables work fine. For instance, the tunables in these simulations could handle demand spike of 50 units without a lot of deterioration of performance in the transition phase. Demand spikes significantly more than this level resulted in lower utilizations in the transient phase. Nevertheless, in the steady state, satisfactory levels of utilization were achieved irrespective of the magnitude of demand spike.

## X. Evaluating the Algorithm Performance when Users or Applications Have Different Demands and Reservation Prices

In this section, we will simulate different user scenarios and see how the pricing function impacts resource allocation between user classes having different demands and utility functions.

In Figure 16, we see that User 1, who has higher reservation price and demand is allocated a higher fraction of network resources. The average of resources allocated to User 1 is 58.07 as compared to 38.06 for User 2. The ratio of these allocations is 1.78, which is higher than the ratio of the reservation prices of the two users $\frac{2}{1.5} = 1.33$.

In Figure 17, we examine the effect of decreasing the reservation price of User 2 on the allocation to two users. The average allocations are found to be 62.88 and 34.32 respectively. The ratio of the allocations for this case is 1.83 which is higher than that of 16, which is to be expected.

In Figure 18, we examine the effect of change in base demand to the scenario of Figure 16. The av-

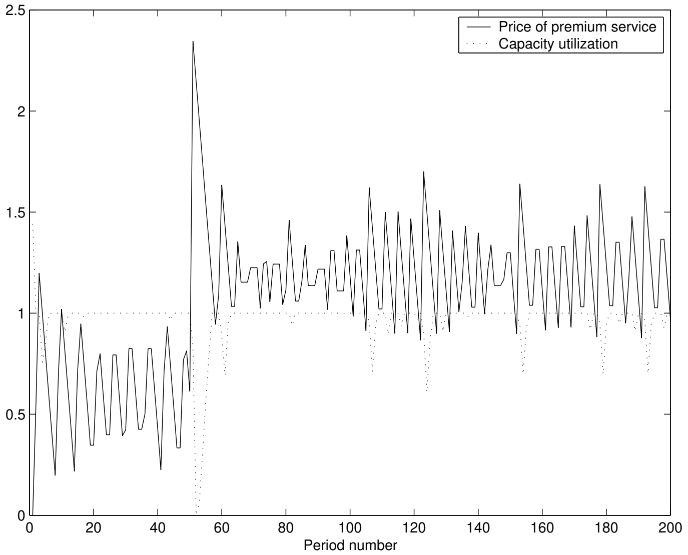


Fig. 14. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.2, q_h = 25, q_l = 15, \hat{p} = 2), X_{p=0} = 140, \Delta X_{p=0} = 100u_t(50)$

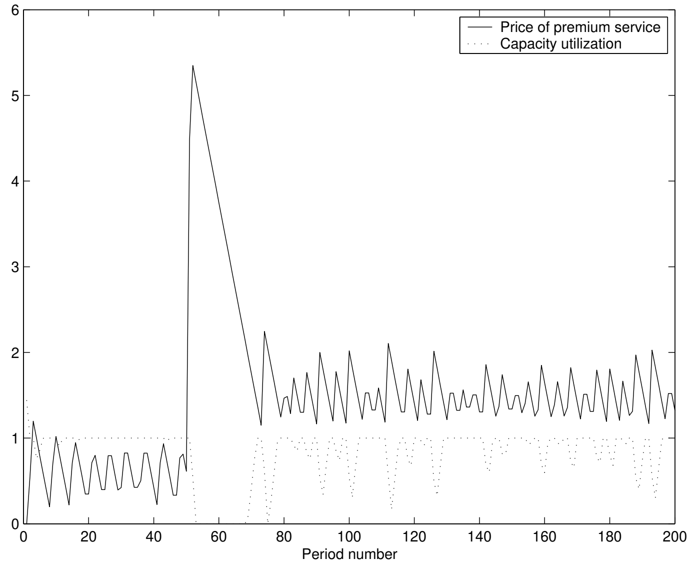


Fig. 15. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.2, q_h = 25, q_l = 15, \hat{p} = 2), X_{p=0} = 140, \Delta X_{p=0} = 200u_t(50)$

erage allocations over the 100 contract periods observed is 74.31 for User 1 and 19.70 for User 2. This indicates that the base demand (demand when p=0) plays a significant role in capacity allocation which is desirable.

In Figure 19, we examine the scenario of Figure 18 when reservation prices of the two users are changed. The average allocation to the users now becomes 80.77 for User 1 and 14.10 for User 2.

Thus, we see that the scheme can allocate resources to a wide variety of users based on their reservation price and base demand.

Notice that User 2 is able to increase its resource share by increasing its reservation price from 1 (Figure 19) to 1.5 (Figure 18).

## XI. Conclusion

In this paper, we identify a parameter for congestion sensitive pricing of premium services provided using Dynamic Capacity Contracting in Edge to Edge framework. This parameter, together with the requirements of a typical ISP is used to formulate an algorithm for determining prices at the beginning of a contract period in Dynamic Capacity Contracting. The algorithm is easily deployable in edge to edge framework as it needs $q_{i-1}$ and $C_i$ for determining the price in period $i$. Both these parameters are easily available in edge to edge control environment. The algorithm is adaptive in its nature and has tunable parameters that can be set to provide a range of better than best effort services by making suitable trade-offs. The three important trade-offs are : high utilization versus smaller edge queues (better service); price variability versus system responsiveness and price variability varsus Quality of Service variation. We also observe that the allocations made under different user utilization and base demand scenarios are well-behaved. Further, work would involve exploring algorithms for changing the tunables of the pricing scheme (which uses the edge queue) by capturing the information about change in base demands. An interesting parameter that can be used for changing the tunables is the rate of change of edge queue as it is a good measure of change in load being offered to the network.

## References

[1] David D. Clark and David L. Tennenhouse, "Architectural considerations for a new generation of protocols," in *Proceedings of SIGCOMM*, 1990.

[2] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *Journal of Operations Research Society*, vol. 49, pp. 237–252, 1998.

[3] J. K. MacKie-Mason and H. R. Varian, *Pricing the Internet*, Kahin, Brian and Keller, James, 1993.

[4] R. Singh, M. Yuksel, S. Kalyanaraman, and T. Ravichandran, "A comparative evaluation of Internet pricing models: Smart market and dynamic capacity contracting," in *Proceedings of Workshop on Information Technologies and Systems (WITS)*, 2000.

[5] S. S. Chiu and J. P. Crametz, "Taking advantage of the emerg-

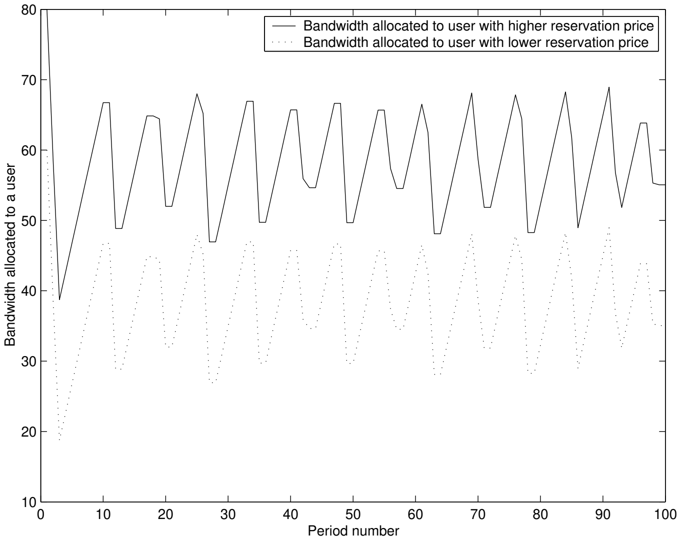


Fig. 16. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p}_1 = 2, X_{1,p=0} = 80, \hat{p}_2 = 1.5, X_{2,p=0} = 60$)

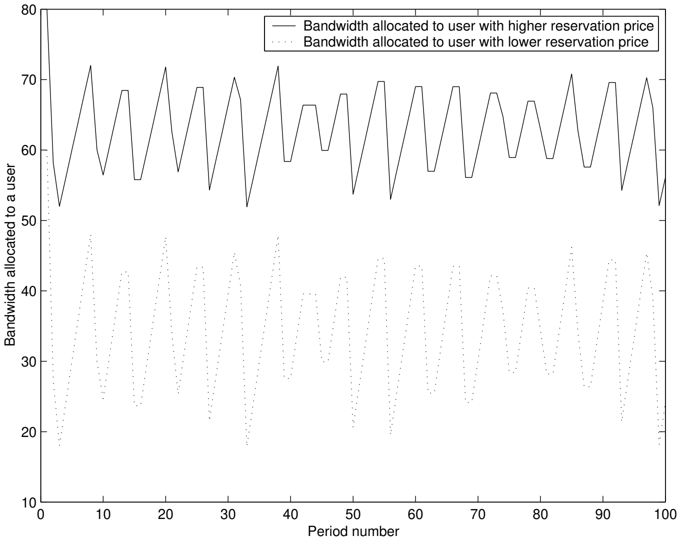


Fig. 17. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p}_1 = 2, X_{1,p=0} = 80, \hat{p}_2 = 1, X_{2,p=0} = 60$)

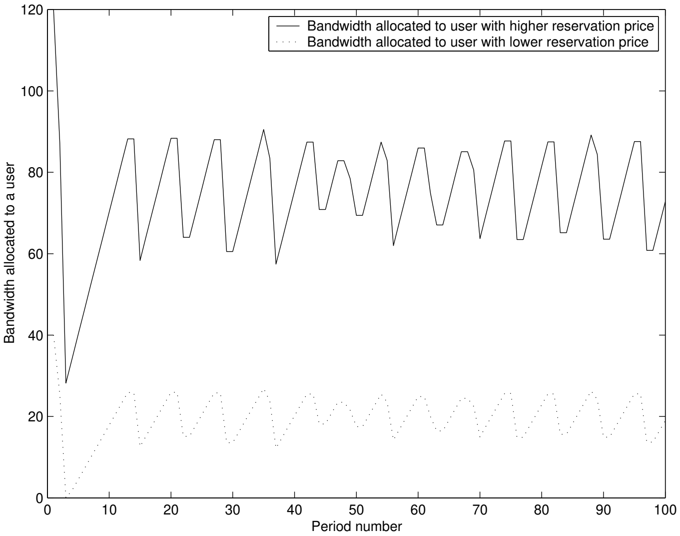


Fig. 18. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p}_1 = 2, X_{1,p=0} = 120, \hat{p}_2 = 1.5, X_{2,p=0} = 40$)

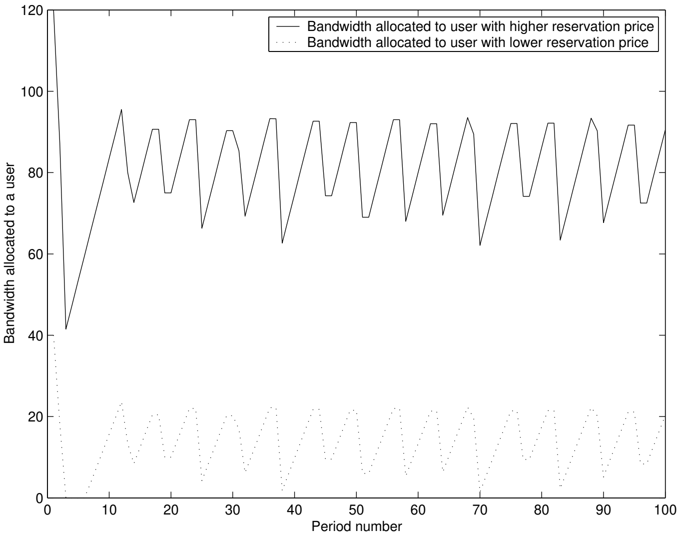


Fig. 19. Price and Utilization for system parameters ($\alpha_1 = 3, \alpha_2 = 0.1, q_h = 25, q_l = 15, \hat{p}_1 = 2, X_{1,p=0} = 120, \hat{p}_2 = 1, X_{2,p=0} = 40$)

ing bandwidth market," Tech. Rep., RateXchange Laboratories, 1999.

[6] M. Yuksel and S. Kalyanaraman, "Simulating the Smart Market pricing scheme on Differentiated Services architecture," in *Proceedings of Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS) part of Western Multi-Conference (WMC)*, 2001.

[7] D. Harrison, S. Kalyanaraman, and S. Ramakrishnan, "Overlay bandwidth services: Basic framework and edge-to-edge closed-loop building block," Poster in the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM), 2001.

[8] S. Blake et. al, "An architecture for Differentiated Services," *IETF RFC 2475*, December 1998.

[9] Bou and Sasse, "Why value is everything?:," in *Proceedings of IEEE/IFIP International Workshop on Quality of Service (IWQoS)*, 2001, pp. 184–193.

[10] A. C. Chiang, *Fundamental Methods of Mathematical Economics*, McGraw Hill, Inc., 1984.

[11] D. M. Chiu, "Some observations on fairness of bandwidth sharing," Tech. Rep. TR-99-80, Sun Microsystems Laboratories, September 1999.

[12] D. Clark, *Internet cost allocation and pricing*, Eds McKnight and Bailey, MIT Press, 1997.

[13] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, "Pricing in computer networks: Motivation, formulation and example," *IEEE/ACM Transactions on Networking*, vol. 1, December 1993.

[14] R. J. Edell and P. P. Varaiya, "Providing Internet access: What we learnt from the INDEX trial," Tech. Rep. 99-010W, University of

California, Berkeley, 1999.

[15] A. Gupta, D. O. Stahl, and A. B. Whinston, *Priority pricing of Integrated Services networks*, Eds McKnight and Bailey, MIT Press, 1997.

[16] S. Kunniyur and R. Srikant, "End-to-end congestion contro: Utility functions, random losses and ecn marks," in *Proceedings of Conference on Computer Communications (INFOCOM)*, 2000.

[17] S. H. Low and D. E. Lapsley, "Optimization flow control – I: Basic algorithm and convergence," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–875, 1999.

[18] J. K. MacKie-Mason and H. R. Varian, "Pricing the congestible network resources," *IEEE Journal on Selected Areas of Communications*, vol. 13, pp. 1141–1149, 1995.

[19] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, October 2000.

[20] "UCB/LBLN/VINT network simulator - ns (version 2)," http://www-mash.cs.berkeley.edu/ns, 1997.

[21] A. M. Odlyzko, "A modest proposal for preventing Internet congestion," Tech. Rep., AT & T Research Lab, 1997.

[22] A. Orda and N. Shimkin, "Incentive pricing in multi-class communication networks," in *Proceedings of Conference on Computer Communications (INFOCOM)*, 1997.

[23] N. Semret, R. R.-F. Liao, A. T. Campbell, and A. A. Lazar, "Pricing, provisioning and peering: Dynamic markets for differentiated Internet services and implications for network interconnections," *IEEE Journal on Selected Areas of Communications – to be published*, 2001.

[24] N. Semret, R. R.-F. Liao, A. T. Campbell, and A. A. Lazar, "Market pricing of differentiated Internet services," in *Proceedings of IEEE/IFIP International Workshop on Quality of Service (IWQoS)*, 1999, pp. 184–193.

[25] X. Wang and H. Schulzrinne, "Pricing network resources for adaptive applications in a Differentiated Services network," in *Proceedings of Conference on Computer Communications (INFOCOM)*, 2001.

[26] X. Wang and H. Schulzrinne, "RNAP: A resource negotiation and pricing protocol," in *International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, 1999, pp. 77–93.

[27] M. Yuksel, S. Kalyanaraman, and B. Sikdar, "Effect of pricing intervals on the congestion-sensitivity of network service prices," in *Submitted to IWQoS*, 2002.