Pricing Granularity for Congestion-Sensitive Pricing *

Murat Yuksel[1] and Shivkumar Kalyanaraman[2]

[1] CS Department, Rensselaer Polytechnic Institute,
110 8th Street, Troy, NY, 12180, USA.
Phone: (518) 276-8289, Fax: (518) 276-2433
yuksem@cs.rpi.edu
[2] ECSE Department, Rensselaer Polytechnic Institute,
110 8th Street, Troy, NY, 12180, USA.
shivkuma@ecse.rpi.edu

An earlier version was submitted to International Workshop on Internet Charging and QoS Technologies (ICQT) 2002.

Abstract. One of the key issues for implementing congestion pricing is the pricing granularity (i.e. pricing interval or time-scale). The Internet traffic is highly variant and hard to control without a mechanism that operates on very low time-scales, i.e. on the order of round-trip-times (RTTs). However, pricing naturally operates on very large time-scales because of human involvement. Moreover, structure of wide-area networks does not allow frequent price updates for many reasons, such as RTTs are very large for some cases. In this paper, we investigate the issue of pricing granularity, identify problems, and propose solutions.

We first focus on how much level of control over congestion can be achieved by congestion pricing. To represent the level of control over congestion, we use correlation between prices and congestion measures. We develop analytical models for the correlation. In order to validate the correlation model, we develop packet-based simulation of our congestion pricing scheme Dynamic Capacity Contracting. We then present the fit between simulation results of the pricing scheme and the correlation model. The correlation model reveals that the correlation degrades at most inversely proportional to an increase in the pricing interval. It also reveals that the correlation degrades with an increase in mean or variance of the traffic.

Secondly, we discuss implications of the correlation model. According to the model and simulation results, we find that control of congestion by pricing degrades significantly as pricing granularity increases. We experimentally show that congestion control achieved by pricing vanishes at a pricing granularity of 40 RTTs even for a very low-variance traffic. Then, for this granularity problem, we propose pricing architectures which can allow both tight control on congestion and human involvement in pricing at the same time.

Keywords. Network Pricing, Congestion Pricing, Congestion Control, QoS

* This work is sponsored by NSF under contract number ANI9819112.

1 Introduction

One proposed method for controlling congestion in wide area networks is to apply *congestion-sensitive pricing* [1, 2], which is a form of dynamic pricing. Many proposals have been made to implement dynamic pricing over wide area networks and the Internet [3–14]. Most of these schemes aimed to employ congestion pricing. The main idea of congestion-sensitive pricing is to update price of the network service dynamically over time such that it increases during congestion epochs and causes users to reduce their demand. So, implementation of congestion-sensitive pricing protocols (or any other dynamic pricing protocol) makes it necessary to change the price after some time interval, what we call *pricing interval*.

Clark's Expected Capacity [3] scheme proposes long-term contracts as the pricing intervals. Kelly's packet marking scheme [5] proposes shadow prices to be fed back from network routers which has to happen over some time interval. MacKie-Mason and Varian's Smart Market scheme [6] proposes price updates at interior routers which cannot happen continuously and have to happen over some time interval. Wang and Schulzrinne's RNAP [8] framework proposes to update the price at each service level agreement which has to happen over some time interval. Hence, congestion-sensitive pricing can only be implemented by updating prices over some time interval, i.e. pricing interval.

It has been realized that there are numerous implementation problems for dynamic or congestion-sensitive pricing schemes, which can be traced into pricing intervals. We can list some of the important ones as follows:

- *Users do not like price fluctuations:* Currently, most ISPs employ flat-rate pricing which makes individual users happy. Naturally, most users do not want to have a network service with a price changing dynamically. In [15], Edell and Varaiya proved that there is a certain level of desire for quality-of-service. However, in [16] and [17], Odlyzko provides evidence that most users want simple pricing plans and they easily get irritated by complex pricing plans with frequent price changes. So, it is important that price updates should happen as less as possible. In other words, users like a service with *larger* pricing intervals.
- *Control of congestion degrades with larger pricing intervals:* Congestion level of the network changes dynamically over time. So, the more frequent the price is updated, the better the congestion control. From the provider's side, it is easier to achieve better congestion control with *smaller* pricing intervals.
- *Users want prior pricing:* It is also desired by the users that price of the service must be communicated to them before it is charged. This makes it necessary to inform the users of the network service before applying any price update. So, the provider has to handle the overhead of that price communication. The important thing is to keep this overhead as less as possible, which can be done with *larger* pricing intervals.

Hence, length of pricing intervals is a key issue for the implementation of congestion-sensitive and adaptive pricing protocols. In this paper, we focus on modeling and analysis of pricing intervals to come up with a maximum value for it such that the level of congestion control remains in an acceptable range. Beyond this range, pricing could be used to regulate demand, but it becomes less useful as a tool for congestion management.

The rest of the paper is organized as follows: In Section 2, we first explore steady-state dynamics of congestion-sensitive pricing with a detailed look at the behavior of prices and congestion relative to each other. We then develop and discuss an approximate analytical model for the correlation of prices and congestion measures in Section 3. In Section 4, we validate the model by simulation experiments and present the results. Finally, in Section 5 we discuss the implications of the work and possible future work.

2 Dynamics of Congestion-Sensitive Pricing

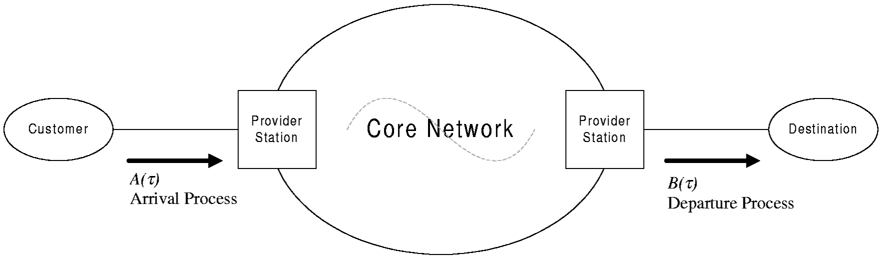


Fig. 1. A sample customer-provider network.

This section explains the behavior of congestion-sensitive prices and congestion measures relative to each other in a steady-state system. A sample scenario is described in Figure 1. The provider employs a pricing interval of T to implement congestion-sensitive pricing for its service. The customer uses that service to send traffic to the destination through the provider's network. The provider observes the congestion level, c, in the network core and adjusts its advertised price, p, according to it. Note that c and p are in fact functions of time (i.e. $c(t)$ and $p(t)$ where t is time), but we use c and p throughout the paper for simplicity of notation. It is a realistic assumption to say that the provider can observe the network core over small time intervals, i.e. a few round-trip-times (RTTs). To understand effect of pricing interval to the dynamics of congestion-sensitive pricing, we look at the relationship between c and p over time.

Assuming that we have continuous knowledge of congestion level, c, we can represent the dynamics of congestion-sensitive pricing as in Figure 2. Figure 2 represents the relationship between c and p for two different pricing interval

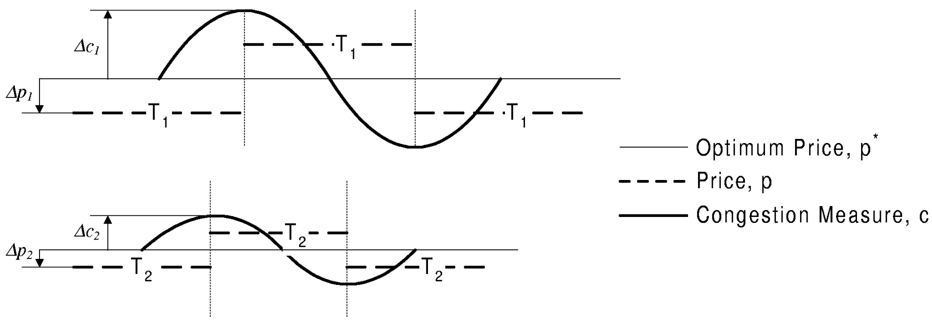


Fig. 2. Congestion measure relative to congestion-sensitive prices in a steady-state network being priced.

lengths, $T_1 > T_2$. For both lengths, the steady-state behavior of congestion-sensitive pricing is represented. The advertised price, p, varies around an optimum price, p^*.

When the provider sees that the congestion level has been decreasing, it decreases the advertised price such that the network resources are not under-utilized. Then the customer starts sending more traffic in response to the decrease in price, and congestion level in the core starts increasing accordingly. The congestion level continues to increase until the price is increased by the provider at the beginning of the next pricing interval. When the provider increases price because of the increased congestion in the last pricing interval, the customer starts sending less traffic than before. Then congestion level starts decreasing. This behavior continues on in steady-state. This explains how congestion-sensitive prices can control the congestion in a network. The important difference is that with a larger pricing interval the congestion level oscillates larger as represented in Figure 2.

Another important characteristic of congestion-sensitive pricing is that the price must be oscillating around an optimum price, p^*, to guarantee both congestion control and high utilization of network resources. In other words, the average of advertised prices must be equal to the optimum price value. Notice that the customer will send less traffic which will under-utilize network resources when $p > p^*$, and the customer will send excessive traffic than the network can handle which will cause uncontrolled congestion when $p < p^*$. So the provider needs to satisfy the condition that the average of advertised prices equals to the optimum price.

The important issue to realize is that congestion control becomes better if the similarity between the advertised price and congestion level is higher. Because of the above explained implementation constraints, the advertised price cannot be updated continuously. This results in dissimilarity between the price and congestion level. Intuitively, if the correlation between the advertised prices and the congestion measures is higher, fidelity of control over congestion becomes higher. Again by intuition, the correlation becomes smaller if the pricing interval is larger.

Another important issue is the *price oscillation* caused by the discontinuous price updates. As the pricing intervals get larger, the oscillation in price also gets larger. This in effect leads to oscillation in user demand (i.e. traffic) correspondingly. So, larger oscillations in price are expected to cause larger oscillation and *higher variance* in incoming traffic. Then, more oscillated traffic causes more oscillated congestion level. This behavior is represented in Figure 2 with the case that $\triangle c_1 > \triangle c_2$ and $\triangle p_1 > \triangle p_2$.

In the next section, we will develop an approximate model of correlation between the advertised prices and congestion measures analytically and find the largest value for the pricing interval such that the system functions in a desired range of service.

3 Analytical Model for Correlation of Prices and Congestion Measures

3.1 Assumptions and Model Development

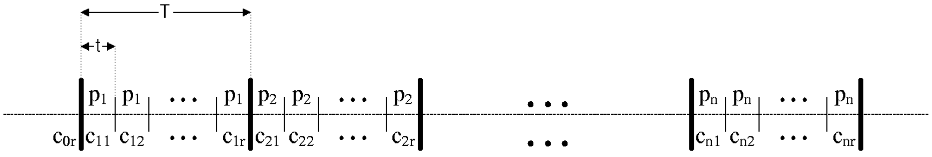

T
t
p1 p1 ... p1 p2 p2 ... p2
...
pn pn ... pn
C0r C11 C12 ... C1r C21 C22 ... C2r
...
Cn1 Cn2 ... Cnr

Fig. 3. Prices and congestion measures for subsequent observation intervals.

Assume the length of pricing interval stays fixed at T over n intervals. Also assume the provider can observe the congestion level at a smaller time scale with fixed observation intervals, t. Assume that $T = rt$ holds, where r is the number of observations the provider makes in a single pricing interval. Assume that the queue backlog in the network core is an exact measure of congestion. [18]

We assume that the customer has a fixed budget for network service and he/she sends traffic according to a counting process, which is a continuous time stationary stochastic process $A(\tau)$,$\tau \geq 0$ with first and second moments of λ_1 and λ_2 respectively. In reality, λ_1 is not fixed, because the customer responds to price changes by changing its λ_1. However, since we assume steady-state and fixed budget for the customer, it is reasonable to say that the customer will send at a constant rate over a large number of pricing intervals. Let m_{ij} be the number of packet arrivals from the customer during the j th observation interval of ith pricing interval, where $i = 1..n$ and $j = 1..r$. So the total number of packet arrivals during the ith pricing interval is

$$m_i = \sum_{s=1}^{r} m_{is} \tag{1}$$

Also assume that the packets leave after the network service according to a counting process, which is a continuous time stationary stochastic process $B(\tau)$,$\tau \geq 0$ with first and second moments of μ_1 and μ_2 respectively. Let k_{ij} be the number of packet departures during the jth observation interval of ith pricing interval, where $i = 1..n$ and $j = 1..r$. So the total number of packet departures during the ith pricing interval is

$$k_i = \sum_{s=1}^{r} k_{is} \tag{2}$$

Assuming that no drop happens in the network core, the first moments of the two processes are equal in steady-state, i.e. $\lambda_1 = \mu_1$, but the second moments are not.

As represented in Figure 3, let p_i be the advertised price and c_{ij} is the congestion measure (queue backlog) at the end of the jth observation in the ith pricing interval. In our model we need a generic way of representing the relationship between prices and congestion. We assumed that the congestion-sensitive pricing algorithm calculates the price for the ith pricing interval according to the following formula[3]

$$p_i = a(t, r) \;\; c_{(i-1)r} \tag{3}$$

where $a(t, r)$, *pricing factor*, is a function of pricing interval and observation interval defined by the congestion pricing algorithm. We assume that $a(t, r)$ is only effected by the interval lengths, not by the congestion measures. Notice that this assumption does not rule out the effect of congestion measures on price, but it splits the effect of congestion measures and interval lengths to price. We will use a instead of $a(t, r)$ for notation simplicity.

Within this context, the following equations hold:

$$c_{ij} = c_{0r} + \sum_{u=1}^{i-1} (m_u - k_u) + \sum_{s=1}^{j} (m_{is} - k_{is}) \tag{4}$$

$$c_{ir} = c_{0r} + \sum_{j=1}^{i} (m_j - k_j) \tag{5}$$

where $i \geq 1$. Reasoning behind (4) and (5) is that the queue backlog (which is the congestion measure) at the end of an interval is equal to the number of packet arrivals minus the number of packet departures during that interval.

[3] Note that this is a simplifying formula for tractability, and cannot capture all aspects of congestion pricing.

Let the average price be $\overline{p}$ and the average queue backlog be $\overline{c}$. By assuming that the system is in steady-state we can conclude that the following equation is satisfied

$$\overline{p} = a\overline{c} \qquad (6)$$

Since the system is assumed to be in steady-state, we can assume the initial (right before the first pricing interval) congestion measure equals to the average queue backlog, i.e.

$$c_{0r} = \overline{c} \qquad (7)$$

We want to approximate the model of correlation between p and c according to the above assumptions. We can write the formula for correlation between p and c over n pricing intervals as

$$Corr_n = \frac{E_n[(c-\overline{c})(p-\overline{p})|m,k]}{E_n[(c-\overline{c})^2|m,k]E_n[(p-\overline{p})^2|m,k]} \qquad (8)$$

assuming that total of m packet arrivals and k packet departures happen during the n rounds.

We can calculate the numerator term in (8) as follows:

$$E_n[(c-\overline{c})(p-\overline{p})|m,k] = \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}(p_i-\overline{p})(c_{ij}-\overline{c}) \qquad (9)$$

By applying (3), (6) and (7) into (9) we can get

$$E_n[(c-\overline{c})(p-\overline{p})|m,k] = \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}(ac_{(i-1)r} - ac_{0r})(c_{ij}-c_{0r}) \qquad (10)$$

Then by applying (4) and (5) into (10), we get the following

$$E_n[(c-\overline{c})(p-\overline{p})|m,k] = \frac{a}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(c_{0r} + \sum_{\theta=1}^{i-1}(m_\theta - k_\theta) - c_{0r}\right)\left(\sum_{u=1}^{i-1}(m_u-k_u) + \sum_{s=1}^{j}(m_{is}-k_{is})\right) \qquad (11)$$

After going through the derivation, we can put (11) into the following form

$$E_n[(c-\overline{c})(p-\overline{p})|m,k] = \frac{a}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(H_1 + \sum_{\theta=1}^{i-1}(m_\theta-k_\theta)\sum_{s=1}^{j}(m_{is}-k_{is})\right) \qquad (12)$$

where $H_1 = \sum_u (m_u-k_u)^2 + \sum_u \sum_{v\neq u} 2(m_u-k_u)(m_v-k_v)$, $u = 1..i-1$ and $v = 1..i-1$.

We can calculate the variance of congestion measures similarly as follows:

$$E_n[(c-\overline{c})^2|m,k] = \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}(c_{ij}-\overline{c})^2 \qquad (13)$$

By applying (4) and (7) into (13) we can get

$$E_n[(c-\overline{c})^2|m,k] = \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(\sum_{u=1}^{i-1}(m_u - k_u) + \sum_{s=1}^{j}(m_{is} - k_{is})\right)^2 \tag{14}$$

After going through the derivation, we can put (14) into the following form

$$E_n[(c-\overline{c})^2|m,k] = \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(H_1 + H_2 + 2\sum_{u=1}^{i-1}(m_u - k_u)\sum_{s=1}^{j}(m_{is} - k_{is})\right) \tag{15}$$

where $H_2 = \sum_s (m_{is} - k_{is})^2 + \sum_s \sum_{z \neq s} 2(m_{is} - k_{is})(m_{iz} - k_{iz})$, $s = 1..j$, $z = 1..j$.

We finally can calculate the variance of price as follows:

$$E_n[(p-\overline{p})^2|m,k] = \frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r}(p_i - \overline{p})^2 \tag{16}$$

By using (3), (5) and (6) into (16) we can get the following

$$E_n[(p-\overline{p})^2|m,k] = \frac{a^2}{n}\sum_{i=2}^{n}\left(\sum_{j=1}^{i-1}(m_j - k_j)\right)^2 \tag{17}$$

Similarly after going through derivation, we can put (17) into the following form

$$E_n[(p-\overline{p})^2|m,k] = \frac{a^2}{n}\sum_{i=2}^{n}H_1 \tag{18}$$

Now we can relax the condition on m and k by summing out conditional probabilities on (12), (15), and (18). Specifically, we need to apply the operation

$$E_n[x] = \sum_{m_{ij}=0}^{\infty}\sum_{k_{ij}=0}^{\infty} E_n[x|m,k]P_{m_{ij};k_{ij}} \tag{19}$$

for all $i = 1..n$ and $j = 1..r$, where $P_{m_{ij};k_{ij}}$ is $P\{A(t) = m_{ij}; B(t) = k_{ij}\}$. This operation is non-trivial because of the dependency between the processes $A(\tau)$ and $B(\tau)$, and it is not possible to reach a closed-form solution without simplifying assumptions. After this point, we develop two *approximate* models by making simplifying assumptions.

Model-I Although the arrival and departure processes are correlated, there might also be cases where the correlation is negligible. For example, if the distance between arrival and departure points is more, then the lag between the arrival and departure processes also becomes more which lowers the correlation between them. So, for simplicity, we assume *independence* between the arrival and departure processes and derive an *approximate* model. The independence assumption makes it very easy to relax the condition on m and k, since the joint probability of having $A(t) = m_{ij}$ and $B(t) = k_{ij}$ becomes product of probability of the two events. After the relaxation, we then substitute $\mu_1 = \lambda_1$ because of the steady-state condition, and get the followings:

$$E_n[(c - \overline{c})(p - \overline{p})] = \frac{atr}{2}(n-1)(\lambda_2 + \mu_2 - 2tr\lambda_1^2) \tag{20}$$

$$E_n[(c - \overline{c})^2] = \frac{t}{2}(\lambda_2 + \mu_2)(rn+1) - t^2\lambda_1^2(1 + r - r^2 + r^2 n) \tag{21}$$

$$E_n[(p - \overline{p})^2] = \frac{a^2 tr}{2}(n-1)(\lambda_2 + \mu_2 - 2tr\lambda_1^2) \tag{22}$$

Let σ_A^2 be the variance of the arrival process and σ_B^2 be the variance of the departure process. By substituting (20), (22), and (21) into (8) we get the correlation model for the first n rounds as follows:

$$Corr_n = \frac{1}{at(\frac{\sigma_A^2 + \sigma_B^2}{2} + \lambda_1^2)(rn+1) - a(t\lambda_1)^2(1 + r - r^2 + r^2 n)} \tag{23}$$

Model-II To make a more realistic model, we try to develop a model where the arrival and departure processes are not considered independent. We consider the system as an $M/M/1$ queueing system with a service rate of μ. Notice that μ is different from the parameters μ_1 and μ_2 which are first and second moments of $B(\tau)$. We now try to derive the joint probability as follows:

$$P_{m_{ij};k_{ij}} = P_{m_{ij}} * P_{k_{ij}|m_{ij}} \tag{24}$$

where $P_{m_{ij}} = P\{A(t) = m_{ij}\}$ and $P_{k_{ij}|m_{ij}} = P\{B(t) = k_{ij}|A(t) = m_{ij}\}$. Notice that $P_{m_{ij}}$ is probability of having m_{ij} events for the Poisson distribution with mean $\lambda_1 t$. However, it is not that easy to calculate $P_{k_{ij}|m_{ij}}$, since probability of having k_{ij} departures depends not only on the number of arrivals m_{ij} but also the number already available in the system which is $c_{i(j-1)}$. Let N be the random variable that represents the number available in the system, then we can rewrite $P_{k_{ij}|m_{ij}}$ as follows:

$$P_{k_{ij}|m_{ij}} = \sum_{c_{i(j-1)=k_{ij}-m_{ij}}}^{\infty} P_{k_{ij}|m_{ij};c_{i(j-1)}} * P_{c_{i(j-1)}} \tag{25}$$

where $P_{c_{i(j-1)}} = P\{N = c_{i(j-1)}\}$. Observe that the minimum value of $c_{i(j-1)}$ can be $k_{ij} - mij$, because the condition $k_{ij} \leq m_{ij} + c_{i(j-1)}$ must be satisfied for all time intervals. In (25), $P_{c_{i(j-1)}}$ is known for a steady-state $M/M/1$ system. Let $\rho = \lambda_1/\mu$, then $P_{c_{i(j-1)}} = (1-\rho)\rho^{c_{i(j-1)}}$. [19] However, calculation of $P_{k_{ij}|m_{ij};c_{i(j-1)}}$ is not simple, because the m_{ij} arrivals may arrive such that there is none waiting for the service. Fortunately, this is a very rare case for a loaded system. So, we can formulate $P_{k_{ij}|m_{ij};c_{i(j-1)}}$ for the usual case as if all the m_{ij}

arrivals happened at the beginning of the interval t. Within this context, we now derive $P_{k_{ij}|m_{ij};c_{i(j-1)}}$.

Let $E(\mu)$ be an Exponential random variable with mean $1/\mu$, and $E_r(k,\mu)$ be an Erlangian random variable with mean k/μ. Then, we can formulate the probability of having $k > 0$ departures in time t as follows:

$$P_{k>0\ in\ t} = \int_0^t P\{E_r(k,\mu) < x\}\left[1 - P\{E(\mu) < t - x\}\right] dx \tag{26}$$

Now, we can formulate the CDF of $P_{k_{ij}|m_{ij};c_{i(j-1)}}$ as follows:

$$P\{B(t) \leq k_{ij}|m_{ij};c_{i(j-1)}\} = P_{0\ in\ t} + \sum_{k=1}^{k_{ij}} P_{k>0\ in\ t} \tag{27}$$

Notice that $P_{0\ in\ t} = 1 - P[E(\mu) < t]$. We used Maple to derive the CDF formula in (27), and got the following result:

$$P\{B(t) \leq k_{ij}|m_{ij};c_{i(j-1)}\} = e^{-\mu t} + \frac{1}{\mu}\left(k_{ij} - e^{-\mu t}\sum_{i=1}^{k_{ij}}\sum_{j=0}^{i}\frac{(\mu t)^j}{j!}\right) \tag{28}$$

By using the CDF formula in (28) in Maple, we then find pmf as:

$$\begin{aligned} P_{k_{ij}|m_{ij};c_{i(j-1)}} &= P\{B(t) \leq k_{ij}|m_{ij};c_{i(j-1)}\} - P\{B(t) \leq k_{ij} - 1|m_{ij};c_{i(j-1)}\} \\ &= \frac{1}{\mu}\left(1 - e^{-\mu t}\sum_{i=0}^{k_{ij}}\frac{(\mu t)^i}{i!}\right) \end{aligned} \tag{29}$$

Afterwards, we apply the operation in (25), i.e.:

$$P_{k_{ij}|m_{ij}} = \sum_{c_{i(j-1)}=k_{ij}-m_{ij}}^{\infty} \frac{1}{\mu}\left(1 - e^{-\mu t}\sum_{i=0}^{k_{ij}}\frac{(\mu t)^i}{i!}\right) * \left(1 - \frac{\lambda_1}{\mu}\right)\left(\frac{\lambda_1}{\mu}\right)^{c_{ij}} \tag{30}$$

Again by using Maple, we finally derive $P_{k_{ij}|m_{ij}}$ as:

$$P_{k_{ij}|m_{ij}} = \frac{1}{\mu}\left(\frac{\lambda_1}{\mu}\right)^{(k_{ij}-m_{ij})}\left[1 - e^{-\mu t}\sum_{i=0}^{k_{ij}}\frac{(\mu t)^i}{i!}\right] \tag{31}$$

Even though we have found a nice solution to $P_{k_{ij}|m_{ij}}$ in (31), it does not allow us to get a closed-form model for the correlation after the relaxation operation in (19). In order to get a closed-form correlation model, we approximated the term with summation in (31). Notice that the term with summation is equivalent to ratio of two Gamma [20] functions, i.e.:

$$e^{-\mu t}\sum_{i=0}^{k_{ij}}\frac{(\mu t)^i}{i!} = \frac{\Gamma(k_{ij}+1,\mu t)}{\Gamma(k_{ij}+1)}$$

In Appendix, we approximated the ratio $\Gamma(x,y)/\Gamma(x)$ and used that method to approximate the term with summation in (31). After the approximation, we did get a closed-form correlation model. But, it is not possible to provide it in hardcopy format[4] because it is a very large expression. However, we will provide numerical results of the model later in Section 4.

3.2 Model Discussion

Since Model-II is a very large expression, we only discuss Model-I. Assuming that the other factors stay fixed, the correlation model in (23) implies three important results:

1. *The correlation degrades at most inversely proportional to an increase in pricing intervals (T):* For the smallest n value (i.e. 1), denominator of (23) will have $r+1$ as a factor which implies linear decrease in the correlation value while the pricing interval, $T = rt$, increases linearly. Notice that its effect will be less when n is larger.
2. *Increase in traffic variances (σ_A^2 and σ_B^2) degrades the correlation:* From (23), we can observe that the correlation decreases when the variance of the incoming or outgoing traffic increases.
3. *Increase in traffic mean (λ_1) degrades the correlation:* Again from (23), we can see that the correlation decreases while the mean of the incoming traffic increases.

These above results imply that lower pricing intervals must be employed when variance and/or mean of the traffic starts increasing. We validate these three results in Section 4 by experiments. Note that the model reveals non-intuitive effect of traffic mean on the correlation. Also, observe that the model incorporates not only the effect of pricing intervals on the correlation, but also the effects of statistical parameters (e.g. traffic mean and variance).

As previously mentioned, the correlation between prices and congestion measures is a representation of the achieved control over congestion. Congestion-sensitive pricing protocols can use such a model to maintain the control at a predefined level by solving the inequality $Corr_n \geq Corr_{min}$ for r, which defines the length of the pricing interval. If feedback from the other end (i.e. egress node in DiffServ [21] terminology) is provided, then such a model can be implemented in real-time. σ_B^2 can be calculated by using the feedbacks from the other end, and σ_A^2 and λ_1 can be calculated by observing the incoming traffic.

4 Experimental Results and Model Validation

4.1 Experimental Configuration

We use Dynamic Capacity Contracting (DCC) [22] as the congestion pricing protocol in our simulations. DCC provides a contracting framework over DiffServ

[4] It is available at *http://networks.ecse.rpi.edu/˜ yuksem/intervals/the_model.mws.*

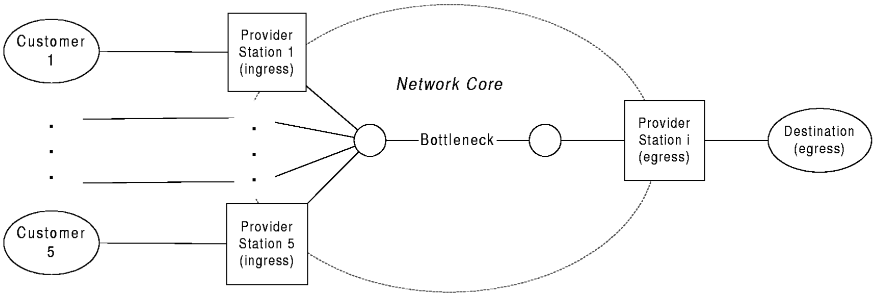

Customer 1
Provider Station 1 (ingress)
Network Core
Bottleneck
Provider Station i (egress)
Destination (egress)
Customer 5
Provider Station 5 (ingress)

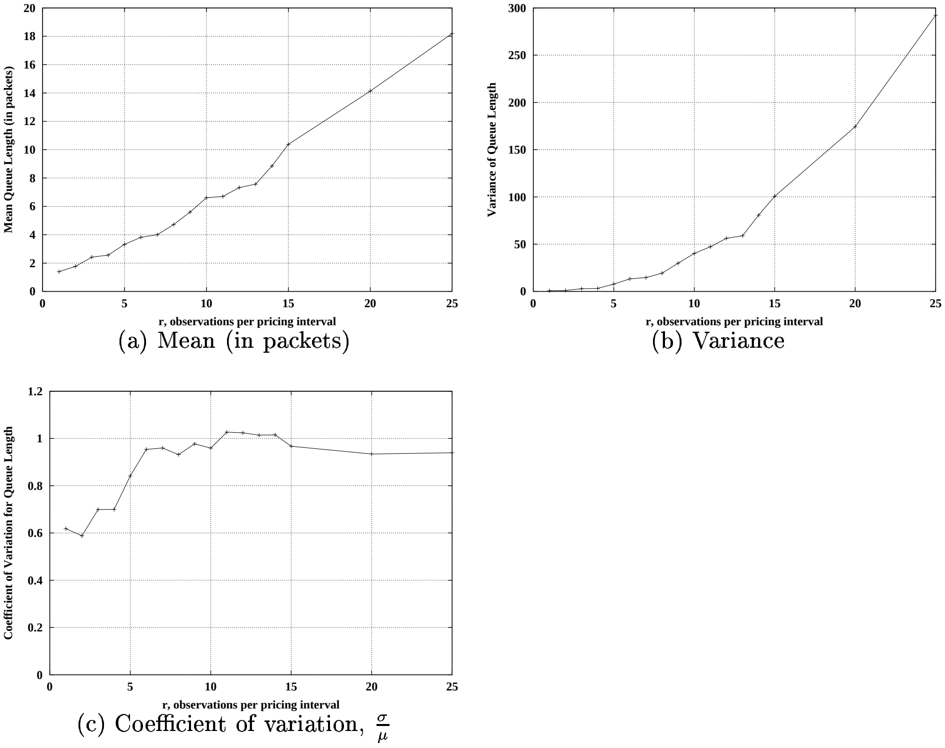
Fig. 4. Topology of the experimental network.

[21] architecture. The provider places its stations at edges of the DiffServ domain. The customers can get network service through these stations by making ***short-term contracts*** with them. The stations provide a variety of short-term contracts and customers select the contracts based on their utility. During the contracts, the station observes congestion in network core. The station uses that congestion information to update the price at the beginning of each contract. The short-term contracts corresponds to the pricing intervals in our modeling.

Figure 4 represents the topology of network in our experiments in *ns* [23]. There are 5 customers trying to send traffic to the same destination over the same bottleneck with a capacity of 1Mbps. Customers have equal budgets and their total budget is 150 units. We observe the bottleneck queue length and use it as congestion measure. The observation interval is fixed at $t = 80ms$ and RTT for a customer is $20ms$. We increase the pricing interval by incrementing the number of observations (i.e. r) per contract. We run several simulations and calculate correlation between the advertised prices and the observed bottleneck queue lengths during the simulations.

Customers send their traffic with a fixed variance but changing mean according to the advertised prices for the contracts. We assume that the customers have fixed budgets per contract with additional leftover from the previous contract. The customers adjust their sending rate according to the ratio B/p where B is the customer's budget and p is the advertised price for the contract. So, customers increase or decrease their sending rate right before the contract starts accordingly. Notice that since the customers' budget is fixed, the sending rate of the customers is actually fixed on long run, which fits to the fixed average incoming traffic rate (λ_1) assumption in the model.

Customers send their traffic with mean changing according to the advertised prices for the contracts. We assume that the customers have fixed budgets per contract with additional leftover from the previous contract. The customers adjust their sending rate according to the ratio B/p where B is the customer's



(a) Mean (in packets)

(b) Variance

(c) Coefficient of variation, $\frac{\sigma}{\mu}$

Fig. 5. Statistics of bottleneck queue length.

budget and p is the advertised price for the contract [5]. Notice that since the customers' budget is fixed, the *average* sending rate of the customers is actually *fixed on long run*, which fits to the fixed average incoming traffic rate assumption in the model.

4.2 Results

In this section, we present several simulation results for validation of the model and the three results it implies.

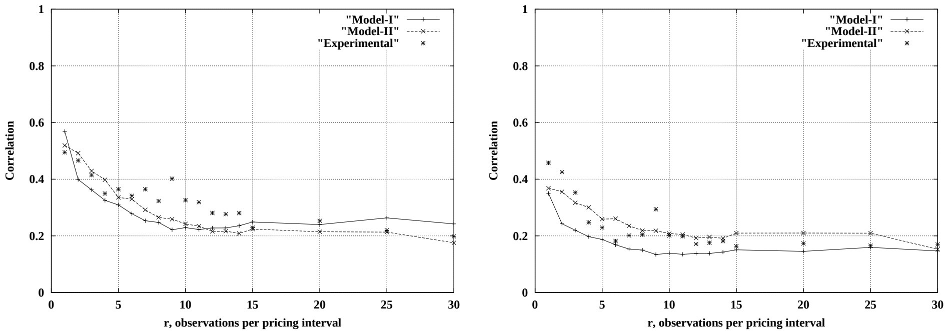
Figures 5-a and 5-b show mean and variance of the bottleneck queue length respectively. We observe steady increase in mean and variance of bottleneck queue as the pricing interval increases. Furthermore, Figure 5-c shows the change in the coefficient of variation for the bottleneck queue length as the pricing interval increases. Note that an increase in the coefficient of variation means a decrease in the level of control. We observe that coefficient of variation increases as the pricing interval increases until $10r$, and stays fixed there after. This is because the congestion pricing protocol looses control over congestion after a certain length of pricing interval, which is $10r$ in this particular experiment. These results in Figures 5-a to 5-c validate our claim about the degradation of control when pricing interval increases. Furthermore, they also show that dynamic pricing does not help congestion control when the pricing interval is longer than a certain length.

To validate the model, we present the fit between our correlation models and experimental results obtained from simulations. Figures 6-a and 6-b represent the correlations obtained by inserting appropriate parameter values to the model and corresponding experimental correlations, respectively for $n = 15$ and $n = 25$. We observe that Model-II fits better than Model-I, since Model-II considers the dependency between arrival and departure processes. Notice that the model is dependent on the experimental results because of the parameters for incoming and outgoing traffic variances (i.e. σ_A^2 and σ_B^2), pricing factor (i.e. a), and mean of the incoming traffic (i.e. λ_1). We first calculate the parameters σ_A^2, σ_B^2, a (ratio of average price by average bottleneck queue length) and λ_1 from the experimental results, and then use them in the model.

We now validate the three results implied in Section 3.2. Figures 6-a and 6-b show that the correlation decreases slower than $1/r$ when r increases linearly. This validates the first result. Figure 7-b represents the effect of change in the variance of incoming and outgoing traffic (i.e. σ_A^2 and σ_B^2) on the correlation. The horizontal axis shows the increase in variances of both the incoming and outgoing traffic. The results in Figure 7-b obviously show that an increase in traffic variances causes decrease in the correlation. This validates the second result. Finally for validation of the third result, Figure 7-a represents the effect of change in the mean of the incoming traffic (i.e. λ_1) on the correlation. We can see that increase in λ_1 causes decrease in the correlation. Another important

[5] Note that $x = B/p$ maximizes surplus for a customer with utility $u(x) = B\ log(x)$.

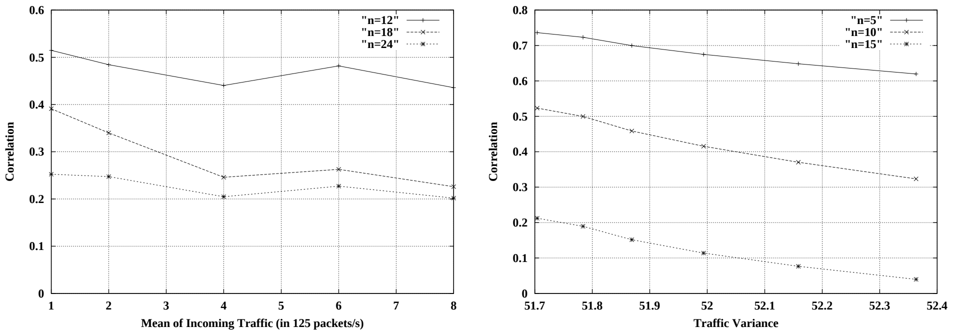
realization is that the correlation is more sensitive to variance changes than mean changes as it can be seen by comparing Figures 7-a and 7-b.



(a) For 15 pricing intervals

(b) For 25 pricing intervals

Fig. 6. Fitting analytical model to experimental results.



(a) Effect of traffic mean

(b) Effect of traffic variance

Fig. 7. Effect of traffic patterns to the correlation (for $T = 800ms$ and $r = 10$).

Before concluding this section, we would like to stress on the relationship between the correlation and the level of congestion control. As we previously stated, Figures 6-a and 6-b show the effect of increasing pricing intervals on the correlation for different values of n. We can see that the correlation value stays almost fixed after the pricing interval reaches to $10r$. Also, Figure 5-c shows the coefficient of variation for the bottleneck queue length. Remember that coefficient of variation for the queue length represents the level of congestion control being achieved. We observe in Figure 5-c that it reaches to its maximum value (approximately 1) when the pricing interval reaches to $10r$, which is the same point where the correlation starts staying fixed in Figures 6-a and 6-b. So, by comparing Figure 5-c with Figures 6-a and 6-b, we can observe that the correlation decreases when the level of congestion control decreases, and also it stays fixed when the level of congestion control stays fixed. This shows that the correlation can be used as a metric to represent the level of congestion control.

5 Summary

We investigated steady-state dynamics of congestion-sensitive pricing in a customer-provider network. With the idea that correlation between prices and congestion measures is a measurement for level of congestion control, we modeled the correlation. We found that the correlation decreases at most inversely proportional to an increase in pricing interval. We also found that the correlation is inversely effected by the mean and variance of the incoming traffic. This implies that congestion-sensitive pricing schemes need to employ very small pricing intervals to maintain high level of congestion control for current Internet traffic with high variance [24].

From the model and also from the simulation experiments we observed that the correlation between prices and congestion measures drops to very small values when pricing interval reaches to 40 RTTs even for a low variance incoming traffic. Currently, we usually have very small RTTs (measured by milliseconds) in the Internet. This shows that pricing intervals should be 2-3 seconds for most cases in the Internet, which is not possible to deploy over low speed modems. This result itself means that deployment of congestion-sensitive pricing over the Internet is highly challenging. As the link speeds are getting higher and RTTs are getting smaller, it becomes harder to deploy congestion-sensitive prices.

The results obviously show that there will be need for intermediate middle-ware components (i.e. intermediaries) between individual users and ISPs, when ISPs deploy congestion-sensitive pricing for their service. These middle-ware components will be expected to lower price fluctuations such that price changes will be possible implement over low speed modems. This scenario suggests that congestion-sensitive prices can be implemented among ISPs to control congestion, but there has to be middle-ware components which can handle the transition of the congestion-sensitive prices to the individual customers in a smooth way. Alternatively, instead of using congestion-sensitive pricing directly for the purpose of congestion control, it can be used to improve fairness of an underlying congestion control mechanism. This way it will be possible to control congestion at small time-scale, while maintaining human involvement to pricing at large time-scale. We believe that the second approach is more realistic way of implementing congestion-sensitive pricing over the Internet.

Another key implementation problem for congestion pricing is that current Internet access is point-to-anywhere. It is not possible to obtain information about the exit points of the traffic. However, it is not possible to determine congestion information and prices without coordinating entry and exit points of the traffic. So, this particular aspect implies that it is highly challenging to implement congestion pricing at individual user to ISP level. But, if an ISP has enough control over the entry and exit points, then it is possible. Alternatively, if ISPs of the current Internet collaborate on providing information about the entry and exit points to each other, then again it will be possible.

Future work should include complex modeling of the dynamics of congestion-sensitive pricing by relaxing some of the assumptions. For example, a model without fixed arrival rate assumption would represent the behavior of the system more appropriately. Also, better budget models are needed in the model.

Another important issue to explore is how much congestion control can be achieved with exactly what level of correlation between prices and congestion measures. In this particular modeling work we assumed that the correlation value is a direct representation of the level of congestion control that was achieved. Although we supported this idea by providing the match between the correlation and the coefficient of variation in Section 4.2, this issue needs more investigation.

References

1. 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.
2. 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.
3. D. Clark, *Internet cost allocation and pricing*, Eds McKnight and Bailey, MIT Press, 1997.
4. A. Gupta, D. O. Stahl, and A. B. Whinston, *Priority pricing of Integrated Services networks*, Eds McKnight and Bailey, MIT Press, 1997.
5. 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.
6. J. K. MacKie-Mason and H. R. Varian, *Pricing the Internet*, Kahin, Brian and Keller, James, 1993.
7. J. K. MacKie-Mason, L. Murphy, and J. Murphy, *Responsive pricing in the Internet*, Eds McKnight and Bailey, MIT Press, 1997.
8. 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.
9. 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.
10. X. Wang and H. Schulzrinne, "An integrated resource negotiation, pricing, and QoS adaptation framework for multimedia applications," *IEEE Journal on Selected Areas of Communications*, vol. 18, no. 12, pp. 2514–2529, 2000.
11. 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*, vol. 18, no. 12, pp. 2499–2513, 2000.

12. 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 Qualityof Service (IWQoS)*, 1999, pp. 184–193.
13. A. Orda and N. Shimkin, “Incentive pricing in multi-class communication networks,” in *Proceedings of Conference on Computer Communications (INFOCOM)*, 1997.
14. 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.
15. 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.
16. A. M. Odlyzko, “The economics of the Internet: Utility, utilization, pricing, and quality of service,” Tech. Rep., AT & T Research Lab, 1998.
17. A. M. Odlyzko, “Internet pricing and history of communications,” Tech. Rep., AT & T Research Lab, 2000.
18. 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.
19. L. Kleinrock, *Queueing Systems, Volume I: Theory*, John Wiley and Sons, 1975.
20. D. G. Childers, *Probability and Random Processes*, McGraw Hill, Inc., 1997.
21. S. Blake et. al, “An architecture for Differentiated Services,” *IETF RFC 2475*, December 1998.
22. 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.
23. “UCB/LBLN/VINT network simulator - ns (version 2),” http://www-mash.cs.berkeley.edu/ns, 1997.
24. M. E. Crovella and A. Bestavros, “Self-similarity in World Wide Web traffic: Evidence and possible causes formulation and example,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, December 1997.
25. J. L. Devore, *Probability and Statistics for Engineeting and the Sciences*, Brooks/Cole Publishing Company, 1995.

6 APPENDIX: Approximating Ratios of Complete or Incomplete Continuous Gamma Functions

6.1 The Gamma Function and Problem Definition

Gamma function has two versions: *complete*, *incomplete* [25]. Complete and incomplete continuous Gamma functions are respectively as follows:

$$\Gamma(x) = \int_{t=0}^{\infty} e^{-t} t^{x-1} dt \tag{32}$$

$$\Gamma(x, y) = \int_{t=y}^{\infty} e^{-t} t^{x-1} dt \tag{33}$$

Discrete version of the complete Gamma function is a simple factorial:

$$\Gamma(x) = (x-1)! \tag{34}$$

Let f be the function being integrated in the continuous Gamma functions, i.e.:

$$f(t,x) = e^{-t}t^{x-1} \tag{35}$$

Figure 8 shows plot of the function $f(t,x)$ for various values of x. Notice that the Gamma function is nothing but the area under the curve of $f(t,x)$. Figures 9 illustrates the difference between complete and incomplete Gamma functions in terms of area under the curve of $f(t,x)$. The area $A+B$ corresponds to the complete Gamma function $\Gamma(x)$, and A corresponds to the incomplete Gamma function $\Gamma(x,y)$.

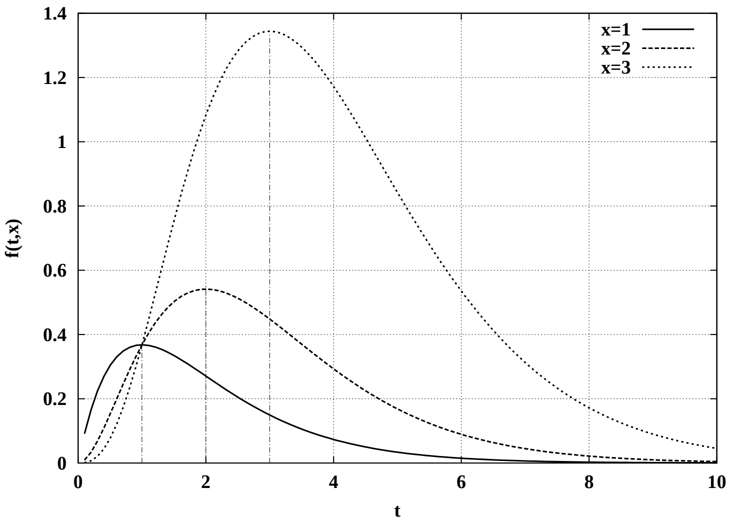


Fig. 8. The function $f(t,x)$ for various values of x.

Given the above information, we want to approximate ratio:

$$\frac{\Gamma(x,y)}{\Gamma(x)} = \frac{A}{A+B} \tag{36}$$

6.2 Approximation Methodology

The intuition behind our approximations is the similarity of shape of $f(t,x)$ to *triangle*. Observe from Figure 8 that as the parameter x gets larger the shape of $f(t,x)$ more triangular. We use this similarity in approximating the ratio in (36).

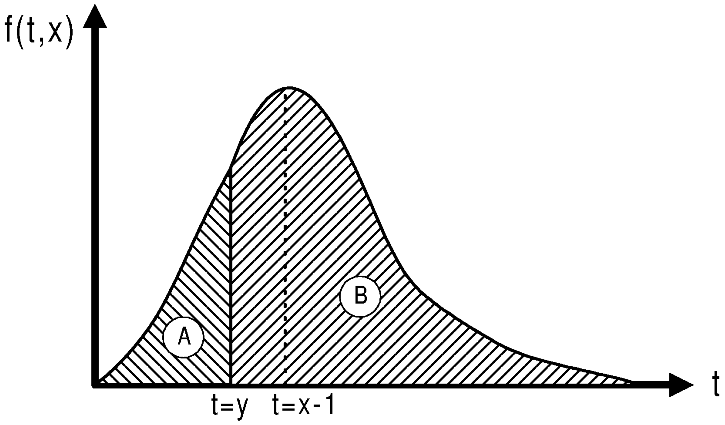


Fig. 9. Visualization of complete and incomplete Gamma functions: The area B is $\Gamma(x, y)$, and the area $A + B$ is $\Gamma(x)$.

Figure 10-a shows an example triangle being matched to the $f(t, x)$ function. In that example we approximate the ratio of (36) as:

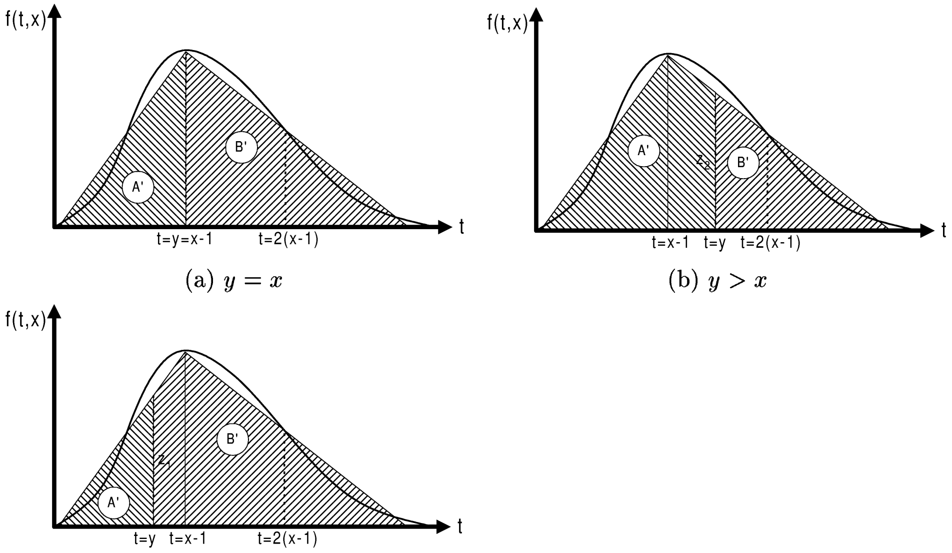
$$R = \frac{\Gamma(x,y)}{\Gamma(x)} = \frac{A}{A+B} \cong \frac{A'}{A'+B'}$$

Notice that the function $f(t, x)$'s maxima is the point at $t = x - 1$, i.e. $f(x-1, x)$. Just to ease notation, let $t_m = x - 1$ and $g(t) = f(t, x)$. Also, let's call the smaller piece of the triangle between $t = 0$ and $t = x - 1$ as *left-piece triangle*, and the other piece of it as *right-piece triangle*. So, the left-piece triangle will have coordinates: $(0, 0)$, $(0, t_m)$, $(t_m, g(t_m))$. For the right-piece triangle, we can consider various coordinates depending how well we want to approximate. Actually, the problem is to identify where should the hypotenus of the right-piece triangle intersect with $f(t, x)$. Since the shape of $f(t, x)$ gets similar to an equi-sided triangle as x gets larger, we choose to select this intersection point at $t = 2t_m = 2(x - 1)$, which will resemble it more to an equi-sided triangle. With this consideration, we can calculate the coordinates of the right-piece triangle by simple geometry rules: $(0, t_m)$, $(t_m, g(t_m))$, $(t_m(g(t_m) - g(2t_m))/(g(t_m) - g(2t_m)), 0)$.

Since we know the function $f(t, x)$, we now can calculate areas A' and B'. However, this is dependent on whether y resides on the left of the right of $t_m = x - 1$. So, we need to consider three cases:

Case I: $y = x - 1$ This case is shown in Figure 10-a. Calculations of the triangular areas in the figure will be as follows:

$$A' = \frac{t_m g(t_m)}{2}$$



(a) $y = x$

(b) $y > x$

(c) $y < x$

Fig. 10. Three possible cases for approximation of ratio $\Gamma(x, y)/\Gamma(x)$.

$$B' = \frac{\left(t_m + \frac{t_m g(2t_m)}{g(t_m) - g(2t_m)}\right) g(t_m)}{2}$$

So, the ratio R for this case will be:

$$R_1 = \frac{g(t_m) - g(2t_m)}{3g(t_m) - 2g(2t_m)}$$

Case II: $y < x - 1$ This case is shown in Figure 10-b. Calculations of the triangular areas in the figure will be as follows:

$$A' = \frac{yz_1}{2}$$

$$B' = \frac{\left(2t_m + \frac{t_m g(2t_m)}{g(t_m) - g(2t_m)}\right) g(t_m)}{2} - \frac{yz_1}{2}$$

where $z_1 = yg(t_m)/t_m$. So, the ratio R for this case will be:

$$R_2 = \frac{y^2}{t_m^2} \frac{g(t_m) - g(2t_m)}{2g(t_m) - g(2t_m)}$$

Case III: $y > x - 1$ This case is shown in Figure 10-c. Calculations of the triangular areas in the figure will be as follows:

$$A' = \frac{\left(2t_m + \frac{t_m g(2t_m)}{g(t_m)-g(2t_m)}\right) g(t_m)}{2} - \frac{yz_2}{2}$$

$$B' = \frac{yz_2}{2}$$

where $z_2 = g(2t_m)$. So, the ratio R for this case will be:

$$R_3 = \frac{2t_m g(t_m)^2 - (t_m - y)g(t_m)g(2t_m) + yg(2t_m)^2}{t_m g(t_m)(2g(t_m) - g(2t_m))}$$

Integration of All Cases In order to calculate the ratio $R = \Gamma(x, y)/\Gamma(x)$, we need to know if y is equal to, less than, or greater than $x - 1$ as presented in the previous sections corresponding to each case.

We can put together an integrated formula for R by considering probability of each case happening. Let p_1 be the probability of being y equal to x-1 (i.e. Case I), p_2 be the probability of being y less than x-1 (i.e. Case II). Then, an integrated formula for R will be:

$$R = p_1 R_1 + p_2 R_2 + (1 - p_1 - p_2)R_3 \tag{37}$$

Since p_1 and p_2 will depend on distribution of y, the integrated approximation of R will change significantly based on that distribution. In modeling of the correlation between prices and congestion measures, in Section 3.1, we used the integrated formula by calculating the probabilities p_1 and p_2 based on the Possion distribution of the traffic.

Also note that in this particular appendix we only provided methodology for approximating the ratio $\Gamma(x, y)/\Gamma(x)$. It is possible to use the ideas in this appendix for approximating other possible ratios of Gamma functions, such as $\Gamma(x, y_1)/\Gamma(x, y_2)$, $\Gamma(y)/\Gamma(x)$.