Resource Based Pricing Framework for Integrated Services Networks

Mostafa H. Dahshan
University of Oklahoma/Electrical and Computer Engineering, Tulsa, OK, U.S.A
Email: mdahshan@ou.edu

Pramode K. Verma
University of Oklahoma/Electrical and Computer Engineering, Tulsa, OK, U.S.A
Email: pverma@ou.edu

***Abstract*—This paper addresses the impact of Quality of Service on resource requirements in networks that implement exclusive bandwidth allocation, such as IntServ. The paper proposes a framework for pricing flows based on the impact of their reservations on the resources for which the network must be provisioned. The developed framework is analytical and is based on the economies associated with aggregating vs. segregating exclusive bandwidths that cater to customers demanding a specified Quality of Service.**

***Index Terms*—QoS Pricing, Integrated Services, Resource reservation, Economies of scale, Segregated flows, Aggregated flows.**

I. INTRODUCTION

The Internet was initially designed as a best-effort network in which all user traffic is treated equally. However, the diversity of application demands and users' willingness to pay have made it imperative to develop techniques for providing a certain level of assurance of resource availability based on application or user requirements. This has stimulated the development of Quality of Service (QoS) techniques such as the Integrated Services architecture (IntServ) [1], the Differentiated Services architecture (DiffServ) [2] and Multi-Protocol Label Switching (MPLS) [3].

While a significant research effort has been put in the design of QoS mechanisms, no such attention has been given toward understanding the cost of guaranteed resource availability and its effect on service pricing. This has resulted in pricing schemes that do not necessarily map the cost to the price of providing the service. Various pricing schemes have been proposed ranging from the simplest flat-rate pricing [4, 5] to more sophisticated techniques such as usage-based pricing [6], priority based pricing [7, 8] and congestion-based pricing [9, 10].

Rather than proposing a specific pricing scheme, the purpose of this paper is to understand the actual cost of offering guaranteed resources to specific flows under different scenarios of traffic characteristics and to provide a framework that can be deployed in designing the appropriate pricing scheme depending on the traffic pattern in the target network. As shown in [11, 12], the performance of shared communication channels is improved when traffic flows having disparate parameters (packet size and arrival rate) are segregated. Similarly, the performance for homogeneous flows is improved when these flows are aggregated. As a result, it is expected that aggregating heterogeneous flows or segregating homogeneous flows would impose some penalty. In the former case, the penalty occurs in the form of higher delay and jitter. In the latter case, the penalty occurs in the form of inefficient usage of bandwidth.

In this paper, the IntServ QoS model will be used to study the effect of bandwidth reservation. IntServ uses the RSVP protocol to allocate bandwidth for each flow upon connection setup. The delay and jitter performance are compared before and after reservation to calculate the additional cost incurred on the network. In order to obtain tractable result, the M/M/1 queueing model is used for most of the analytical work. While the M/M/1 model facilitates simple calculations, it has been shown that self-similar stochastic models provide more accurate characterization of the Internet traffic [13-16]. Some insights into the M/G/1 model are therefore provided in the appendices. The M/M/1 analyses are still necessary because they provide closed form representations that give better understanding of the cost requirements. In addition, it has been shown that the M/M/1 model is still applicable in heavy loaded networks [17].

The reminder of this paper is organized as follows: Related studies on pricing models are reviewed in Section II. The IntServ model is briefly described in Section III with emphasis on the Controlled Load (CL) service class. The economies of scale of segregation versus aggregation of flows are reviewed in Section IV. Sections V and VI provide delay and jitter analysis, respectively, for segregated versus aggregated homogeneous flows sharing a common pool of bandwidth. The results obtained in

Based on "Pricing for Quality of Service in High Speed Packet Switched Networks", by Mostafa H. Dahshan, and Pramode K. Verma which appeared in the Proceedings of the IEEE Workshop on High Performance Switching and Routing 2006, Poznan, Poland, June, 2006. © 2006 IEEE.

Sections V and VI are then used to derive the proposed pricing framework in Section VII. Finally, the summary and conclusions are presented in Section VIII.

II. Related Work

The subject of pricing of network services has been studied from different perspectives in the literature. In [5], three different network design approaches have been compared. The first approach is to separate traffic based on the application demands into two distinct networks, regular and premium. The second approach is to combine all traffic into a single network that provides a uniformly high QoS for all types of traffic with flat rate pricing. The third approach is to provide a single network that combines all types of traffic but provides differentiated QoS. The paper concluded that the uniformly high QoS, flat-rate, approach is the most efficient one. The simplicity of this approach is said to pay for its extra cost. A more recent study [4] on flat-based versus usage-based pricing suggests a similar result. This study, however, is more focused on the web hosting market, both from the user and provider prospective. Another approach, based on the auction mechanism, is proposed in [7]. In this approach, the network offers a set of priority levels. The users submit their bids on how much they are willing to pay for the desired QoS level. The network then distributes the priority levels among users proportionally to their bid values.

A commonly used pricing scheme is congestion based pricing [10]. In this scheme, the price is adjusted according to the congestion status of the network. Typically, a higher price is charged when there is more congestion. One objective of this scheme is to encourage users to reduce their transmission demands during peak periods. The interval between pricing changes is an important parameter in congestion based pricing. An interesting result of [10] is that very small pricing intervals are required to provide appropriate congestion control for self-similar traffic networks due to the high fluctuations of traffic parameters.

The increased popularity of the Internet has been largely driven by its flat rate pricing. The introduction of QoS with usage based pricing might discourage users who prefer fixed charges or those who don't need QoS. To address this problem, a two-part tariff scheme has been explored in [6]. In this scheme, users can continue to pay only flat rate for basic, best effort traffic. Additional variable charge is incurred only for using the premium QoS class of traffic. The flat rate and two-part tariff schemes are compared using simulation analysis. It is shown that flat rate pricing can be more profitable to the service provider when the flat price is set to a value that is close to the fixed part of the two-part tariff. On the other hand, the two-part tariff can be more profitable if the fixed part is constrained to be lower than the flat rate price [6].

In this paper, we provide a framework that covers a wide range of traffic scenarios. The flexibility of the two-part tariff scheme makes it more suitable to the proposed framework. The focus in this paper is on determining the criteria on which the variable part in the price is calculated.

III. The Integrated Services Architecture

The Integrated Service architecture [1] was developed by the IETF in the early 1990s to support the requirement of real-time applications. It is one of the first attempts to support QoS over the Internet [18]. IntServ is considered as a *micro-level QoS* model because it operates on a per-flow basis. It is also considered as a *hard QoS* model because it requires end-to-end reservation of resources on each router along the path prior to data transmission [19]. Additionally, each router has to keep track of active flows during the lifetime of the connections.

Application Classes

IntServ classifies applications into three main categories [20]:

- Elastic applications which can tolerate a wide range of data rates, delays and packet losses. Examples include email, File Transfer Protocol (FTP), Domain Name Service (DNS) [21].
- Tolerant real-time applications which can tolerate some packet loss and limited delay. Multimedia streaming applications and Internet gaming applications fit into this class.
- Intolerant real-time applications such as voice and video conferencing. They have stringent bandwidth, delay and jitter requirements.

Service Classes

The IntServ architecture defines two service classes to accommodate the different application classes. Besides the default best effort behavior, which is suitable for elastic applications, the following service classes are defined [20]:

- Controlled Load service (CL) class. This class emulates the behavior of a network with no or light load. It is intended to be better than best effort but can still have some occasional delay or packet loss. This service class is suitable for tolerant real-time applications.
- Guaranteed Service (GS) class. This class provides strict bounds on bandwidth and delay. It is intended for intolerant real-time applications.

The analysis of this paper addresses the effect of implementing the CL class on other flows using the same communication channel.

IV. Economies of Scale of Segregated and Aggregated Flows

The economic advantage of aggregating or segregating flows depends highly on the diversity of the parameters of those flows. In general, the economies of scale imply that when several flows share a common pool of bandwidth, the required bandwidth will be lower than the total bandwidth required to achieve the same performance if each flow is allocated an exclusive bandwidth. Previous studies have shown that this aggregation advantage diminishes as the disparity between the average packet

sizes of the individual flows increases, if the utilization of the communication channel (or the total bandwidth remains constant?) is to be kept the same [22, 23]. A more efficient approach has been proposed in [11, 12] in which flows with closer average packet size are grouped into fewer number of channels. The analysis in [11, 12] has been performed on several queueing models, including M/G/1 and G/G/1 queues. Comparable results were obtained from all queue types. Figure 1 shows the relationship between the delay and the disparity of the average packet sizes of flows.

V. DELAY ANLYSIS

Suppose homogeneous traffic served by an M/M/1 queuing system is split into n flows on a random basis. Each flow will have the same mean service time and will be identically distributed with Poisson arrivals [24] and exponential service times (Appendix A). In particular, if the distribution of the overall traffic is such that each arrival has an identical probability of falling in any one of the n flows, the utilization of all sub-channels will be the same. As discussed in Section IV, the segregation of traffic into separate channels improves the weighted mean delay when the different flows have disparate mean packet sizes. This is not the case, however, when the traffic is homogenous [25].

Table 1 summarizes the symbols used in this paper. We can formalize the split of traffic into n sub-channel as follows. Note that the suffixes 1,2,…,n represent each of the n channels into which traffic is distributed with identical probability. Thus,

$$\rho_1 = \rho_2 = \cdots = \rho_n \tag{1}$$

$$1/\mu_1 = 1/\mu_2 = \cdots = 1/\mu_n \tag{2}$$

$$\lambda_1 = \lambda_2 = \cdots = \lambda_n = \frac{\lambda}{n} \tag{3}$$

$$C_1 = C_2 = \cdots = C_n = \frac{C}{n} \tag{4}$$

The mean delay for the combined system, that includes

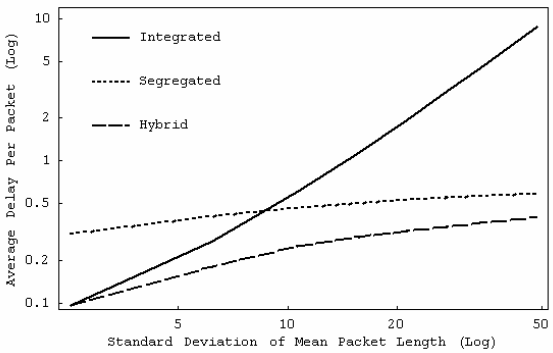

Figure 1: Delay comparison between Aggregated (Integrated), Segregated and Hybrid flows [11].

TABLE 1: SUMMARY OF SYMBOLS USED IN THIS PAPER

Symbol	Meaning
λ	Mean arrival rate
$1/\mu$	Mean packet size
C	Channel capacity
ρ	System load or utilization factor
σ_S^2	service time variance
σ_D^2	delay variance
λ_i, $1/\mu_i$, C_i, ρ_i	Suffixed parameters are the corresponding parameters for individual sub-channel i.

the waiting time and the service time, is calculated as:

$$T = \frac{\rho/\lambda}{1-\rho} \tag{5}$$

where,

$$\rho = \frac{\lambda}{\mu C} \tag{6}$$

For each sub-channel i, the mean delay is given by:

$$T_i = \frac{\rho/\lambda_i}{1-\rho} = n\frac{\rho/\lambda}{1-\rho}, \qquad i = 1,\cdots,n \tag{7}$$

Comparing (5) and (7), we can see that:

$$T_i = nT, \qquad i = 1,\cdots,n \tag{8}$$

In other words, the mean delay of n identically distributed M/M/1 systems is n times that of a single M/M/1 system serving the summation of the traffic using a bandwidth that is the sum of all the individual channels. This result is also valid for M/G/1 queues as shown in Appendix B.

VI. JITTER ANALYSIS

Jitter plays an important role in the quality of real-time traffic, such as voice or video communications. The mean opinion score (MOS) of telephony traffic, for example, is impacted by not only the fixed delay that the transmission system would impose, but also by the variability of delay, namely, jitter [26]. The jitter can be estimated by the variance of the delay [27, 28]. In this part, the variances of the mean delays for the single channel and the n sub-channels are compared.

The variance of the delay σ_D^2 for the M/M/1 queue is calculated as:

$$\sigma_D^2 = \frac{1}{\mu^2 C^2 (1-\rho)^2} \tag{9}$$

The derivation of (9) is provided in Appendix C.

Using (9), the variance of the delay for each sub-channel i is calculated as:

$$\sigma_{D_i}^2 = \frac{1}{\mu^2 (C/n)^2 (1-\rho)^2} = n^2 \frac{1}{\mu^2 C^2 (1-\rho)^2} \quad i = 1, \cdots, n \tag{10}$$

The variance of the combined system is calculated from (9).

Comparing (9) and (10) :

$$\sigma_{Di}^2 = n^2 \sigma_D^2, \quad i = 1, \cdots, n \tag{11}$$

This shows that the variance, and hence the jitter, is lower by a factor of n^2 in the fully shared system compared to that in the system with separate bandwidth reservation for each of the n flows.

VII. Compensatory Pricing for Guaranteed Quality of Service

The discussion in the previous sections shows that there is a performance penalty imposed by reserving exclusive bandwidth for each flow of traffic. Without such exclusive allocation of bandwidths for specific flows, the service provider is able to provide an average delay that is equal to $1/n$ times and a jitter that is $1/n^2$ times that of the system with equal exclusive bandwidth allocation to each flow.

Individual Resource Requirements

This part provides the derivation required in order to estimate the resources consumed in individually reserved channels against that in a shared bandwidth system.

For the combined system, (5) can be rewritten as:

$$T = \frac{1}{\mu C - \lambda} \tag{12}$$

where $\lambda=\lambda_1+\lambda_2+\ldots+\lambda_n$ and T is the mean delay per packet Suppose it is required to achieve the same value of T for an individual flow in an exclusive bandwidth-allocated system. It is required to calculate the bandwidth $C_{\{1\}}$ that satisfies:

$$\frac{1}{\mu C_{\{1\}} - \lambda/n} = \frac{1}{\mu C - \lambda} \tag{13}$$

Note that λ/n is the arrival rate of any individual flow.

Solving (13) yields:

$$C_{\{1\}} = C - \frac{\lambda}{\mu}\left(\frac{n-1}{n}\right) = C\left(1 - \rho\left(\frac{n-1}{n}\right)\right) \tag{14}$$

Dividing both sides of (14) by C yields:

$$\frac{C_{\{1\}}}{C} = 1 - \rho\left(\frac{n-1}{n}\right) \tag{15}$$

The stability of any queueing system implies that ρ is always less than 1, yielding:

$$\frac{C_{\{1\}}}{C} > 1 - \left(\frac{n-1}{n}\right) \tag{16}$$

which is equivalent to:

$$C_{\{1\}} > \frac{C}{n} \tag{17}$$

The inequality in (17) is an important result. It shows that the capacity required for each individual flow in a bandwidth-reserved system serving n flows is always more than $1/n$ times the total available capacity if it is required to achieve the same delay as for a combined, shared-bandwidth system serving the same number of flows with the same total capacity. The exact amount of capacity required can be calculated from (14).

Applying the same analysis to equate the variance (jitter) of the reserved bandwidth system with that of the bandwidth-shared system, (9) can be rewritten as:

$$\sigma_D^2 = \frac{1}{(\mu C - \lambda)^2} \tag{18}$$

The value of $C_{\{1\}}$ required to equate the variance should satisfy:

$$\frac{1}{\left(\mu C_{\{1\}} - \lambda/n\right)^2} = \frac{1}{(\mu C - \lambda)^2} \tag{19}$$

Solving (19) for $C_{\{1\}}$ yields:

$$C_{\{1\}} = C\left(1 - \rho\left(\frac{n-1}{n}\right)\right) \tag{20}$$

which is the same value obtained in (14). Equations (14) and (20) together point out the fact that the additional capacity needed to equalize the mean delays is both necessary and sufficient to equalize the variances of the delays of the two systems as well. We note that this is generally not true for systems other than M/M/n.

Figure 2 shows channel capacity required by individual flows in order to maintain the same delay in the shared-bandwidth system divided by C/n, where C is the total

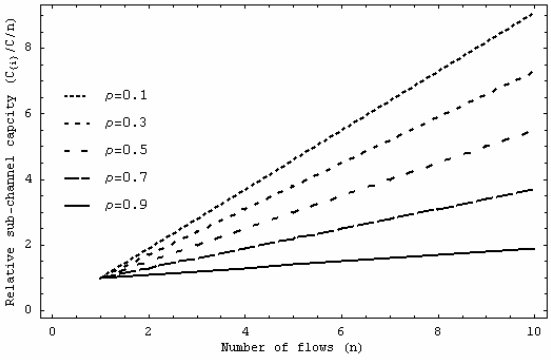

Figure 2. Capacity requirement for n sub-channels with different utilizations, relative to individual flow's share of bandwidth.

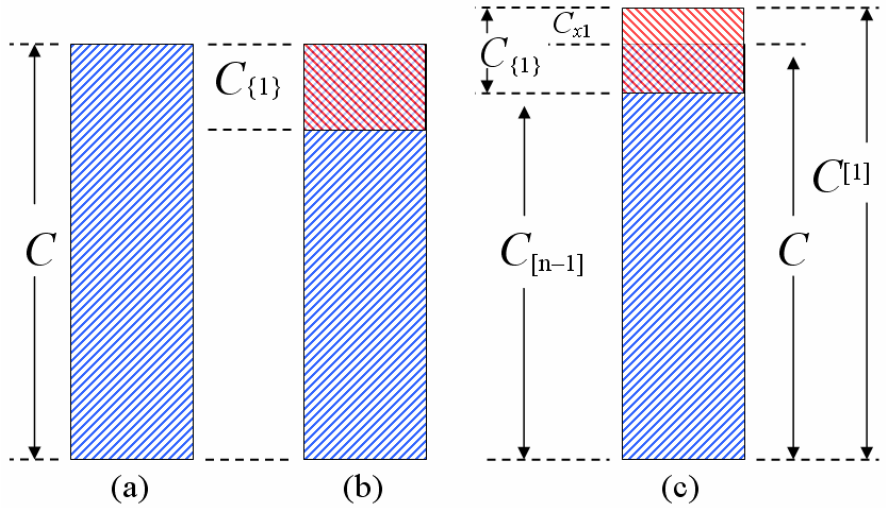

Figure 3. Communication channel (a) Channel before allocation (b) Bandwidth reserved for flow 1 (c) The system increases the total capacity to accommodate the requirements of other flows.

channel capacity for the system and n is the number of flows. The plot is done for different system loads or utilization factors. The utilization factor is for the system before allocation. It is observed from Figure 2 that the capacity required for each flow to keep the same delay before allocation is higher if the utilization of the system before allocation is lower. This is because the delay is inversely related to the utilization. Therefore, more bandwidth is needed for a flow to match a delay of a lightly loaded shared-bandwidth system. This scenario is a typical implementation of the Controlled Load (CL) service class of IntServ, which emulates the behavior of a lightly loaded best effort system, as discussed in Section III.

Additional Cost Requirements

Whenever a customer or a flow of traffic in a shared system requests an exclusive bandwidth, it is expected that the other customers or flows will be affected. It can be seen from (17) that the separated flow takes more than simply its equal share of the total bandwidth. Thus, the other flows are likely to encounter more delay. In order to maintain the delay for the other flows without excluding some of them, the provider must increase the capacity. The flow requesting exclusive bandwidth should be charged for that increment. This is the main factor in the pricing framework proposed in this paper. Figure 3 shows an overview of the proposed pricing framework with the symbols used in the framework analysis.

In the M/M/1 system, suppose flow number 1 is to be allocated exclusive bandwidth. The additional capacity requirement can be calculated as follows: First, the delay for the remaining n–1 flows is:

$$T_{[n-1]} = \frac{1}{\mu C_{[n-1]} - \lambda\left(\frac{n-1}{n}\right)} \tag{21}$$

where $C_{[n-1]}$ denotes the capacity required for the remaining n–1 flows to maintain the same delay before separating one flow. $C_{[n-1]}$ can be calculated by solving the following equation:

$$\frac{1}{\mu C_{[n-1]} - \lambda\left(\frac{n-1}{n}\right)} = \frac{1}{\mu C - \lambda} \tag{22}$$

yielding:

$$C_{[n-1]} = C - \frac{\lambda}{n\mu} = C\left(1 - \rho\left(\frac{1}{n}\right)\right) \tag{23}$$

Using (14) and (23), the new total capacity $C^{[1]}$ can be calculated as:

$$C^{[1]} = C_{[n-1]} + C_{\{1\}} \tag{24}$$

The additional capacity $C_{\{x1\}}$ is therefore equal to:

$$C_{\{x1\}} = C^{[1]} - C = C(1-\rho) \tag{25}$$

For each subsequent flow i to be allocated exclusive bandwidth, the procedure can be repeated as follows. The delay for the remaining n–i flows is:

$$T_{[n-i]} = \frac{1}{\mu C_{[n-i]} - \lambda\left(\frac{n-i}{n}\right)} \tag{26}$$

$C_{[n-i]}$ is determined by solving:

$$\frac{1}{\mu C_{[n-i]} - \lambda\left(\frac{n-i}{n}\right)} = \frac{1}{\mu C - \lambda} \tag{27}$$

yielding:

$$C_{[n-i]} = C - \frac{i\lambda}{n\mu} = C\left(1 - \rho\left(\frac{i}{n}\right)\right) \tag{28}$$

The new total capacity $C^{[i]}$ can be calculated as:

$$C^{[i]} = C_{[n-i]} + iC_{\{i\}} \quad (29)$$

where $C^{[i]}$ denotes the new required capacity of the system after separating all flows from 1 to *i*. Since all flows are homogeneous, $C_{\{i\}} = C_{\{1\}}, i = 2,\cdots,n$. Using (14) it can be stated that the additional capacity $C_{\{xi\}}$ is therefore equal to:

$$C_{\{xi\}} = C^{[i]} - C^{[i-1]} = C(1-\rho) \quad (30)$$

Thus:

$$C_{\{xi\}} = C(1-\rho), i = 1,\cdots,n-1 \quad (31)$$

i.e., for each additional flow that requests exclusive bandwidth (equal to $C_{\{1\}}$), the extra capacity that needs to be added to maintain the same delay value for the remaining flows is always the same. The additional system capacity required depends only on the system load ρ. The relation between additional capacity and system load is shown in Figure 4.

Grouping Multiple Flows

Suppose it is required to combine multiple flows in the allocated channel. This can be done, for example, if a single user is requesting exclusive bandwidth for multiple flows or if the service provider wants to provide guaranteed service on a per-class level rather than per-flow level. To maintain the same delay for the reserved channel as the shared bandwidth pool, the previous analysis can be revised as follows: Let *g* denote the number of flows to be grouped in the reserved channel, as shown in Figure 5. It is required to calculate the bandwidth $C_{\langle g\rangle}$ that satisfies:

$$\frac{1}{\mu C_{\langle g\rangle} - g\lambda/n} = \frac{1}{\mu C - \lambda} \quad (32)$$

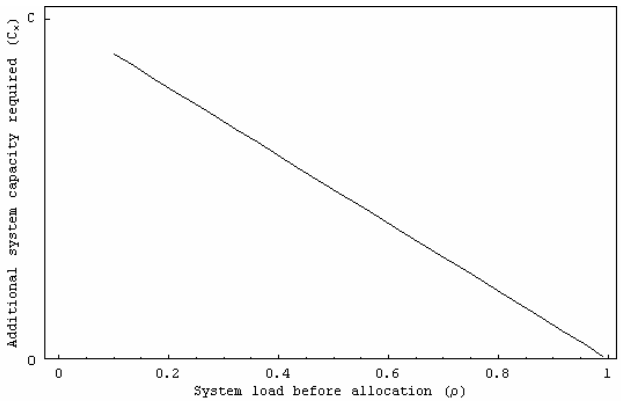

Figure 4. Additional system capacity required for different system loads.

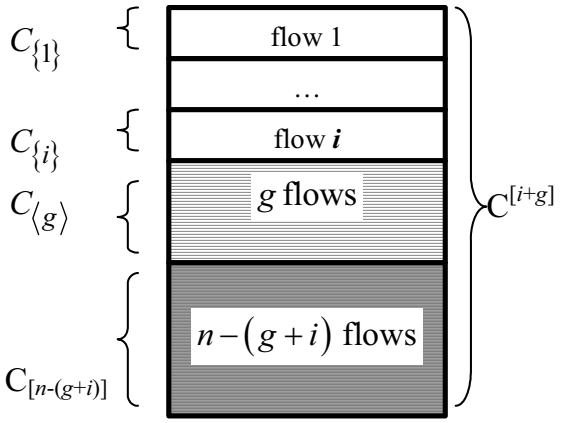

Figure 5: Grouping multiple flows

Because of the similarity of the flow parameters, the combined arrival rate of the group flow is $g\lambda/n$. Solving (32) yields:

$$C_{\langle g\rangle} = C - \frac{\lambda}{\mu}\left(\frac{n-g}{n}\right) = C\left(1-\rho\left(\frac{n-g}{n}\right)\right) \quad (33)$$

Suppose reserved channels have already been allocated for *i* flows. The additional capacity required for the grouped *g* flows can now be calculated as follows: First, the delay for the remaining $n-(i+g)$ flows is:

$$T_{[n-(i+g)]} = \frac{1}{\mu C_{[n-(i+g)]} - \lambda\left(\frac{n-(i+g)}{n}\right)} \quad (34)$$

where $C_{[n-(i+g)]}$ denotes the capacity required for the remaining $n-(i+g)$ flows to maintain the same delay before separating $(i+g)$ flows. Following similar procedure as in the previous analysis, the value of $C_{[n-(i+g)]}$ required to maintain the same delay for the remaining flows is:

$$C_{[n-(i+g)]} = C - \frac{(i+g)\lambda}{n\mu} = C\left(1-\rho\left(\frac{i+g}{n}\right)\right) \quad (35)$$

The new total capacity $C^{[i+g]}$ can be calculated as:

$$C^{[i+g]} = C_{[n-(i+g)]} + iC_{\{i\}} + C_{\langle g\rangle} \quad (36)$$

The additional capacity required $C_{\langle xg\rangle}$ is:

$$C_{\langle xg\rangle} = C^{[i+g]} - C^{[i]} \quad (37)$$

Interestingly, solving (37) yields:

$$C_{\langle xg\rangle} = C(1-\rho) \quad (38)$$

From (31) and (38), it can be seen that allocating exclusive bandwidth for group of flows incurs the same overhead as allocating exclusive bandwidth for a single flow.

Pricing Implementation

As a demonstration of the resource based pricing framework, let's consider the two-part tariff scheme discussed in Section II. In this implementation, there are two types of flows: regular flows that share the common bandwidth pool and premium flows, which request exclusive bandwidth.

The two-part tariff consists of a fixed part which both regular flows pay, and a variable part which only premium flows pay. The key factor in this implementation is to use only the extra capacity incurred by each premium flow to calculate the variable part for the price charged for that flow. Based on the analysis performed in this section, the extra bandwidth overhead $C_{\{xi\}}$ added by each flow is the factor to be used for calculating the variable part of the price. Let F denote the fixed part of the price and let R denote the tariff per unit capacity (e.g. bps). A two-part tariff implementation of the proposed pricing framework could calculate the price P_i for the flow i as:

$$P_i = F + R\, C_{\{xi\}} \tag{39}$$

In the case where the reserved bandwidth is requested for a group g of flows rather than a single flow, we have shown in (38) that the additional overhead in the analyzed queuing model is the same as the overhead incurred by reserving a channel for a single flow. Thus, if the group of flows belongs to the same user, the pricing can still be calculated from (39). On the other hand, if the flows in the group belong to different users, the variable part of the price should be divided between those users, i.e.:

$$P_g = F + R\frac{C_{\{xi\}}}{g} \tag{40}$$

Using (39) and (40), one can compute the price for requesting exclusive bandwidth. This pricing framework will maintain an acceptable level of performance for all traffic flows in the system.

VIII. CONCLUSION

This paper has presented a new framework for computing the cost of providing identical channels out of a common pool of channels shared among multiple flows. Such a scheme is implemented, for example, in the IntServ system for providing a defined QoS to a set of requesting subscribers. We have shown that exclusive allocation of bandwidth has a performance penalty on delay and jitter. We derived the additional capacity required to maintain the desired performance parameters. The presented framework concludes that flows requesting exclusive bandwidth should be charged in proportion to the overhead incurred by the system if it were to satisfy the requirements of the premium flows, while maintaining the same average delay for other flows. An implementation of the presented framework has been demonstrated using the two-part tariff pricing scheme. Using the framework developed in this paper, a service provider can develop an effective mechanism for establishing tariff for IntServ customers under a wide variety of ambient conditions.

APPENDIX A. SERVICE TIME DISTRIBUTION OF A QUEUE COMPOSED OF COMBINING INDIVIDUAL QUEUES

Suppose X is a random variable representing the packet size of the shared combined queue. X_i (i=1,…,n) represents the packet size of the individual queue i. We have:

$$\begin{aligned}\Pr[X = x] = {} & \frac{\lambda_1}{\lambda}\Pr[X_1 = x] \\ & + \frac{\lambda_2}{\lambda}\Pr[X_2 = x] \\ & + \cdots + \frac{\lambda_n}{\lambda}\Pr[X_n = x]\end{aligned} \tag{41}$$

Since X_1, X_2, …, X_n are of the same distribution and have identical mean and statistical properties. Then:

$$\Pr[X_1 = x] = \Pr[X_2 = x] = \cdots = \Pr[X_n = x] \tag{42}$$

Substituting in (41)

$$\Pr[X = x] = \left(\sum_{j=1}^{n}\frac{\lambda_j}{\lambda}\right)\Pr[X_1 = x] \tag{43}$$

From (43), it can be stated that:

$$\Pr[X = x] = \Pr[X_i = x],\ i = 1,\cdots,n \tag{44}$$

i.e., the combing multiple flows of traffic with identical packet size distributions result in a system with the same packet size distribution. Since the service time is simply the packet size divided by the channel capacity which is constant. The same can be said about the distribution of the service time.

APPENDIX B. DELAY ANALYSIS FOR M/G/1 QUEUEING SYSTEMS

The service time is determined from the packet size divided by the channel capacity. Thus, if X represents the packet size, then $S = X/C$ is the service time for the combined queue and $S_i = X_i/(C/n)$ is the service time for the individual queue. Using the proof in Appendix A, the distributions X and X_i are identical.

From the basic statistics:

$$E[aX] = a\,E[X] \tag{45}$$

$$\mathrm{var}[aX] = a^2\,\mathrm{var}[X] \tag{46}$$

Applying to the service time distributions:

$$E\left[n\frac{X}{C}\right] = n\,E\left[\frac{X}{C}\right] \tag{47}$$

$$\mathrm{var}\left[n\frac{X}{C}\right] = n^2\,\mathrm{var}\left[\frac{X}{C}\right] \tag{48}$$

Therefore:

$$\mathrm{var}[S_i] = n^2\,\mathrm{var}[S], \qquad i = 1,\cdots,n \tag{49}$$

The variance of the service time for individual queues is n^2 times greater than the combined queue when the bandwidth is equally split between them. This result is independent of the distribution of the packet size (or the service time).

Let σ_S^2 denote the variance of the service time of the combined queue and σ_{Si}^2 denote the variance of the service time of any of the individual queues *i*. Using the Pollaczek-Khintchine formula [29], the delay *T* for the combined M/G/1 queue can be calculated as:

$$T = \frac{1}{\mu C} + \frac{\rho^2 + \lambda^2\sigma_S^2}{2\lambda(1-\rho)} \tag{50}$$

For all individual queues i, the delay T_i is calculated as:

$$T_i = \frac{1}{\mu C/n} + \frac{\rho^2 + \frac{\lambda^2}{n^2}\sigma_{Si}^2}{2\frac{\lambda}{n}(1-\rho)} \tag{51}$$

From (49), we have:

$$\sigma_{Si}^2 = n^2\sigma_S^2 \tag{52}$$

Substituting in (51):

$$\begin{aligned} T_i &= n\left(\frac{1}{\mu C} + \frac{\rho^2 + \lambda^2\sigma_{Si}^2}{2\lambda(1-\rho)}\right) \\ &= nT \end{aligned} \tag{53}$$

This shows that the mean delay for the individual queues is n times greater than the mean delay for the combined queue if the total bandwidth is equally shared and all packet sizes have the same mean. This result is valid for M/G/1 queues regardless of the distribution of the packet sizes (or service times) provided that the service time variance has a finite value.

APPENDIX C. DELAY VARIANCE CALCULATION FOR M/M/1 QUEUES

Kleinrock [29] has indicated that the *k*-th moment of the system delay $D^{(k)}$ for the M/G/1 queueing system is calculated as:

$$D^{(k)} = \sum_{i=0}^{k} w^{(k-i)} b^{(i)} \tag{54}$$

where:

$$w^{(k)} = \frac{\lambda}{1-\rho}\sum_{i=1}^{k}\binom{k}{i}\frac{b^{(i+1)}}{i+1}w^{(k-i)} \tag{55}$$

The notation $b^{(i)}$ denotes the *i*-th moment of *b*, *w* denotes the waiting time and *D* denotes the system delay which equals the waiting time plus the service time.

For the M/M/1 system, the service time is exponentially distributed. Thus, the first three moments are given as

$$b^{(0)} = 1 \tag{56}$$

$$b^{(1)} = \frac{1}{\mu C} \tag{57}$$

$$b^{(2)} = \frac{2}{\mu^2 C^2} \tag{58}$$

$$b^{(3)} = \frac{6}{\mu^3 C^3} \tag{59}$$

Using these equations, the variance of the system delay can be calculated as follows:

$$w^{(1)} = \bar{w} = \frac{\lambda}{\mu^2 C^2(1-\rho)} \tag{60}$$

$$w^{(2)} = \frac{2\lambda^2}{\mu^4 C^4(1-\rho)^2} + \frac{2\lambda}{\mu^2(1-\rho)} \tag{61}$$

$$D^{(1)} = \frac{1}{\mu C} + \frac{\lambda}{\mu^2(1-\rho)} \tag{62}$$

$$D^{(2)} = \frac{2}{\mu^2 C^2} + \frac{2\lambda^2}{\mu^4 C^4(1-\rho)^2} + \frac{4\lambda}{\mu^2 C^2(1-\rho)} \tag{63}$$

$$\sigma_D^2 = D^{(2)} - \left(D^{(1)}\right)^2 = \frac{1}{\mu^2 C^2(1-\rho)^2} \tag{64}$$

REFERENCES

[1] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," RFC 1633, June 1994.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," RFC 2475, December 1998.

[3] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture," RFC 3031, January 2001.

[4] L. Zhen, L. Wynter, and C. Xia, "Usage-Based Versus Flat Pricing for E-Business Services with Differentiated QoS," in *IEEE International Conference on E-Commerce - CEC '03*, 2003, pp. 355-362.

[5] P. C. Fishburn and A. M. Odlyzko, "Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet," in *First International Conference on Information and Computation Economies*, Charleston, South Carolina, United States, 1998, pp. 128-139.

[6] S. SeungJae and M. B. H. Weiss, "Simulation Analysis of QoS Enabled Internet Pricing Strategies: Flat Rate vs. Two-Part Tariff," in *36th Annual Hawaii International Conference on System Sciences*, 2003.

[7] Z. Guanxiang, L. Yan, Y. Zongkai, and C. Wenqing, "Auction-Based Admission Control and Pricing for Priority Services," in *29th Annual IEEE International Conference on Local Computer Networks*, 2004, pp. 398-399.

[8] M. Mandjes, "Pricing Strategies under Heterogeneous Service Requirements," in *Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies - INFOCOM'03*, San Francisco, CA, USA, 2003, pp. 1210-1220.

[9] J. K. MacKie-Mason and H. R. Varian, "Pricing Congestible Network Resources," *IEEE Journal on Selected Areas in Communications,* vol. 13, pp. 1141-1149, 1995.

[10] M. Yuksel and S. Kalyanaraman, "Pricing Granularity for Congestion-Sensitive Pricing," in *Eighth IEEE International Symposium on Computers and Communications*, 2003, pp. 169-174.

[11] M. H. Dahshan and P. K. Verma, "Performance Enhancement by Segregation and Hybrid Integration in General Queueing Networks," in *International Symposium on Performance Evaluation of Computer and Telecommunication Systems – SPECTS'05*, Philadelphia, PA, USA, 2005, pp. 143-148.

[12] M. H. Dahshan and P. K. Verma, "Performance Enhancement of Heavy Tailed Queueing Systems using a Hybrid Integration Approach," in *IEEE Global Telecommunications Conference – GLOBECOM'05*, St. Louis, MO, USA, 2005.

[13] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Transactions on Networking,* vol. 2, pp. 1-15, 1994.

[14] V. Paxon and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking,* vol. 3, pp. 226-244, June 1995.

[15] M. E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking,* vol. 5, pp. 835-846, December 1997.

[16] K. Park and W. Willinger, *Self-Similar Network Traffic and Performance Evaluation*. New York, USA: John Wiley & Sons, 2000.

[17] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "Internet Traffic Tends Toward Poisson and Independent as the Load Increases," in *Nonlinear Estimation and Classification*, New York, USA, 2002.

[18] Z. Wang, *Internet QoS Architectures and Mechanisms for Quality of Service*. California, USA: Morgan Kuffmann Publishers, 2001.

[19] U. Payer, "DiffServ, IntServ, MPLS."

[20] S. Jha and M. Hassan, *Engineering Internet QoS*. London, UK: Artech House, 2002.

[21] G. Armitage, *Quality of Service in IP Networks*. Indiana, USA: Macmillan Technical Publishing, 2000.

[22] H. R. Rudin, "On Economies of Scale and Integration of Services in Certain Queued Information Transmission Systems," *IEEE Transactions on Communications,* vol. 20, pp. 991-995, October 1972.

[23] P. K. Verma and A. M. Rybzcynski, "The Economics of Segregated and Integrated Systems in Data Communication with Geometrically Distributed Message Length," *IEEE Transactions on Communications,* pp. 1844-1848, November 1974.

[24] V. S. Frost and B. Melamed, "Traffic Modeling for Telecommunications Networks," *IEEE Communications Magazine,* vol. 32, pp. 70-81, 1994.

[25] L. Kleinrock, *Communication Nets*. New York, USA: Dover Publications, 1964.

[26] B. Duysburgh, S. Vanhastel, B. De Vreese, C. Petrisor, and P. Demeester, "On the Influence of Best-Effort Network Conditions on the Perceived Speech Quality of VoIP Connections," in *Tenth International Conference on Computer Communications and Networks*, 2001, pp. 334-339.

[27] N. Davies, J. Holyer, and P. Thompson, "End-to-end Management of Mixed Applications Across Networks," in *IEEE Workshop on Internet Applications*, 1999, pp. 12-19.

[28] M. Karol, P. Krishnan, and J. J. Li, "enProtect: Enterprise-Based Network Protection and

Performance Improvement," *Avaya Labs Research - Technical Report,* August 2002.
[29] L. Kleinrock, *Queueing Systems - Volume I: Theory*. New York, USA: John Wiley & Sons, 1975.

Mostafa H. Dahshan has received his Ph.D in Electrical and Computer Engineering and M.S in Telecommunications Systems from the University of Oklahoma, USA in 2006 and 2002, respectively. He also received a B.S. degree in Computer Engineering from Cairo University, Egypt in 1999. His current research interests include Computer Networks and Quality of Service. He is currently an Information Technology Specialist and an Independent Researcher at the University of Oklahoma.

Pramode K. Verma is the director of the Telecommunications Systems Program and a Professor of Computer Engineering in the University of Oklahoma, Tulsa. He obtained his doctorate in Electrical Engineering from Concordia University in Montreal, Canada in 1970 and an MBA from the Wharton School of the University of Pennsylvania in 1984. He is the author/co-author of over 75 publications and several books in telecommunications, computer communications and related fields.

Dr. Verma has more than 20 years of leadership experience in the telecommunications industry. In his last position with Lucent Technologies as Managing Director – Business Development, Global Service Providers Business and Business Communications System, his responsibilities included creating strategic alliances and partnerships with leading organizations, and managing the associated P&L. He also held professional and management positions with Lucent Technologies – Bell Laboratories for fifteen years.