Pricing Priority Classes in a Differentiated Services Network*

Peter Marbach
Center for Communication Systems Research
University of Cambridge
10 Downing Street
Cambridge, CB2 3DS, UK
pm259@ccsr.cam.ac.uk

Abstract

We study the role of pricing in packet-based telecommunication networks which use priorities to provide differentiated Quality-of-Service (QoS). Customers are given the freedom to choose priorities, but are charged accordingly. Using a game theoretic framework, we study the case where customers choose an allocation of priorities to packets in order to optimize their net benefit. For the single link case, we show that there always exists a (Nash) equilibrium for the corresponding non-cooperative game, which, however, is not necessarily unique. Furthermore, we show that under each equilibrium, packet loss depends only on pricing decisions and therefore pricing becomes a tool to control link congestion.

1 Introduction

There have been proposals to use priorities to provide Quality of Service (QoS) in packet-based telecommunications networks such as the Internet (see Diffserv proposal of the IETF [4] and LAN protocol of the IEEE [7]). The basic idea is the following. Consider a network which offers a finite number of different priority levels. At each node, packets are served according to a priority rule, so that packets with high priorities will experience shorter delays and are less likely to be dropped. This mechanism enables the provision of differentiated QoS: customers with tight QoS requirements (such as IP-Telephony connections) may use high priorities, whereas customers with QoS-insensitive traffic (such as Email) use low priorities.

A question that arises in this context is how the network provider can prevent customers from marking all their packets with the highest priority. One possible solution to this problem is to implement a network controller which assigns priorities to customers based on some predefined policy. This approach gives the network tight control over the use of the priorities, but is expensive to implement. In addition, it may be difficult for the the network controller to obtain the necessary information regarding the QoS requirements and traffic patterns of the individual customers to make priority assignments in an

*This research was supported by a contract with Alcatel Bell, Belgium.

efficient way. In an alternative approach, pricing could be employed to regulate the use of priorities. In this scenario, customers are free to choose the priorities they attach to their packets, but are charged accordingly (of course, the price for higher priorities will be higher). Leaving the issue of how to do billing and accounting aside, this scheme can easily be implemented. Furthermore, it gives customers the freedom to choose priorities that match best their service requirements. One would expect that customers will then try to meet their QoS requirements at the lowest possible cost and use high priorities only for packets with tight QoS requirements.

The goal of this paper is to study the priority service, combined with pricing, described above; we restrict our analysis to the the single link case. We adopt an economic framework where customers and their service requirements are characterized by an utility function. Roughly, submitting more high priority packets will increase the utility of an individual customer. However, doing so will also increase cost, i.e. the price the customer has to pay for using the link. It is then natural to assume that that customers will choose an allocation of packets to priorities in order to maximize their own net benefit, i.e. their utility minus cost. We model this situation as a non-cooperative game for which we wish to determine whether there exists a (Nash) equilibrium and how pricing decision influence the characteristic of an equilibrium. Due to space constraints, we will state results without proof.

Below we provide a few pointers to related work. The framework that we consider here is closely related to the Paris Metro Pricing (PMP) proposal [5] for providing QoS in the Internet. PMP partitions the network into several logically separated channels where channels differ in the prices paid for using them. This can be implemented by using the priority scheme, combined with pricing, described above. However, [5] does not analyze the existence of an equilibrium and the role of pricing. [2] provides simulation results to demonstrate that priorities can be an efficient tool for providing QoS in packet-based networks. In particular, [2] illustrates that using priorities, combined with pricing, achieves a higher a social welfare than a best-effort/flat-rate scheme. Gupta et. al. investigate in [3] a priority service similar to the one propose here, however they consider the situation where the network dynamically changes the prices associated with the different priority classes in order to track a socially optimal allocation. This approach is theoretically appealing, but expensive to implement. The scheme considered in this paper is simpler as prices associated with the different service classes are fixed. Park et. al. also consider in [1, 6] a service where packets get marked according to customer's QoS requirements. However, there is no price associated with the different traffic classes, i.e. the costs incurred to customers are purely performance related. They study the situation where users are free to choose their traffic class [6], as well as where users indicate their QoS requirements and a network controller assigns network resources [1].

2 Link Model

For our analysis, we assume that time is divided into slots and consider a single link with a capacity of C packets per time slot. There is no buffer available and packets that do not get transmitted in a given time slot are dropped. The link supports a finite set $I = \{1, ..., N\}$ of different priority levels, where level 1 has lowest priority - level 2 the second lowest, and so on. Customers mark each packet with a priority $i \in I$. For a given time slot, let $d = (d(1), ..., d(N)$ be the aggregated allocation, where $d(i)$, $i = 1, ..., N$, is the total number of packets submitted with priority i.

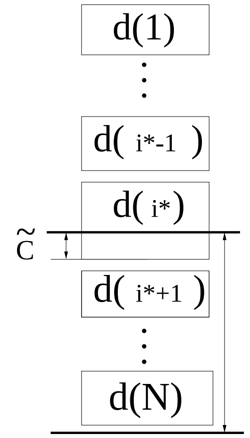

Figure 1: The link transmits all packets with a priority higher than i^*. From the packets with priority i^*, only $\tilde{C} = \left(C - \sum_{i=i^*+1}^{N} d(i)\right)$ packets are transmitted, which are chosen at random. All other packets are dropped.

The link then serves packets as follows (see Figure 1 for an illustration). All packets with priority i such that

$$\sum_{k=i}^{N} d(k) < C$$

are transmitted. For priority i^* with

$$\sum_{i=i^*+1}^{N} d(i) < C \quad \text{and} \quad \sum_{i=i^*}^{N} d(i) \geq C,$$

only

$$\left(C - \sum_{i=i^*+1}^{N} d(i)\right)$$

packets - which are chosen at random - are transmitted. All packets with priority $i < i^*$ are dropped.

The probability $P_{tr}(i,d)$ that a packet with priority i is transmitted is then given by

$$P_{tr}(i,d) = \begin{cases} 1 & \text{if } C > \sum_{k=i}^{N} d(k) \\ \frac{C - \sum_{k=i+1}^{N} d(k)}{d(i)} & \text{if } \sum_{k=i}^{N} d(k) \geq C > \sum_{k=i+1}^{N} d(k) \\ 0 & \text{otherwise} \end{cases}$$

Let $P_{tr}(d) = (P_{tr}(1,d), ..., P_{tr}(N,d)$ the corresponding transmission probability vector.

In each time slot, customers are charged a price u_i per submitted packet with priority i; we have

$$0 < u_k < u_i, \qquad \text{when } k < i, \quad i,k = 1,...,N.$$

In addition, there exists a feedback mechanism that allows customers to detect which of their packets were transmitted. Customers can then decide to retransmit dropped packets; however this will introduce additional cost. In particular, when $P_{tr}(i,d)$ is the probability that a packet with priority i is transmitted under the allocation $d = (d(1), ..., d(N))$, then the expected cost for transmitting a packet with priority i is equal to $u_i/P_{tr}(i,d)$.

The above model describes a priority service where a fixed price u_i is charged per submitted packet with priority i. However the price for actually transmitting a packet will depend on the total demand. In the following, we wish to study how customer use this service.

3 Customers with Elastic Traffic

We start out by analyzing the above priority scheme for the case where customers only care about throughput (also referred to as customers with elastic traffic), i.e. the customer's utility function depends only on the (expected) number of packets transmitted per time-slot. More precisely, when $d_r = (d_r(1), ..., d_r(N)$ is the allocation chosen by customer r, where $d_r(i)$, $i = 1, ..., N$, is the number of packets submitted with priority i, and $d_{-r} = (d_{-r}(1), ..., d_{-r}(N)$ is the aggregate allocation of all other customers, then we associate with customer r the utility U_r given by

$$U_r(d_r, P_{tr}(d_{-r} + d_r)) = \bar{U}_r(x_r),$$

where

$$x_r = \sum_{i=1}^{N} P_{tr}(i, d_{-r} + d_r) d_r(i)$$

is the expected number of packets of customer r that are transmitted and $\bar{U}_r$, is the utility of customer r under the expected throughput x_r . We assume that $\bar{U}_r$ satisfies the following assumption.

Assumption 1 *The function $\bar{U}_r$ is given by*

$$\bar{U}_r(x_r) = \int_0^{x_r} G_r(x) dx,$$

where $G_r(x) : \Re_+ \to \Re_+$ is bounded, strictly decreasing and continuously differentiable with

$$G_r(0) = u_{\max,r} > 0 \quad \text{and} \quad \lim_{x \to \infty} G_r(x) = 0.$$

Assumption 1 implies that the function $\bar{U}_r$ is strictly concave. $G_r(x)$ is the inverse of the function $D_r : \Re_+ \to \Re_+$, where $D_r(u)$ is the optimal solution the the maximization problem

$$\max_{x \geq 0} \bar{U}_r(x) - xu.$$

$D_r(u)$ can be interpreted as the demand function associated with the utility $\bar{U}_r$, i.e. $D_r(u)$ reflects the demand of customer r in number of packets per time slot when the price per submitted packet is u and each submitted packet is transmitted with probability 1. In the following we will extensively use and refer to the function $D_r(u)$.

Given an aggregated demand $d_{-r} = (d_{-r}(1), ..., d_{-r}(N)$ by customers different from r, we will assume that customer r chooses an allocation $d_r \in \Re_+^N$ which solves the following optimization problem,

$$\max_{y_r \in \Re_+^N} \left\{ U_r(y_r, P_{tr}(d_{-r} + d_r)) - \sum_i y_r(i) u_i \right\}. \tag{1}$$

3.1 Formulation as a Non-Cooperative Game

Consider the situation where all customers are simultaneously trying to optimize their net benefit by solving the maximization problem given by Eq. (1). This situation is naturally modeled as a non-cooperative game. Here, we assume that customers use local search algorithms to find an allocation which optimizes their net benefit [1]. Accordingly, we define an equilibrium allocation for this game as follows.

Definition 1 *We call an allocation* $(d_1, ..., d_R)$, $d_r \in \Re_+^N$, $r = 1, ..., R$, *an equilibrium allocation when for every user* $r = 1, ..., R$, d_r *is a locally optimal solution to the optimization problem*

$$\max_{y_r \in \Re_+^N} \left\{ U_r(y_r, P_{tr}(d_{-r} + y_r)) - \sum_{i=1}^{N} y_r(i) u_i \right\},$$

where

$$d_{-r} = \left(\sum_{l=1}^{R} d_l \right) - d_r.$$

We can think of an equilibrium allocation as a "local" Nash equilibrium where customers seek a local rather than a global optimum. We make the following assumption to exclude the trivial case where the link capacity exceeds the the demand in the lowest priority level.

Assumption 2 *We have that*

$$\sum_{r=1}^{R} D_r(u_1) > C.$$

We have then have the following result.

Proposition 1 *Let Assumption 1 and 2 hold. Then, there exists an equilibrium allocation. Furthermore, when* $(d_1, ..., d_R)$,$d_r \in \Re_+^N$, $r = 1, ..., R$, *is an equilibrium allocation, then there exists a priority* $i_0 \in \{1, ..., N\}$ *such that*

(a) $d_r(i) = 0$, *for all* $r = 1, ..., R$ *and all* $i \notin \{i_0, i_0 + 1\}$;

(b) $\frac{u_{i_0}}{u_{i_0+1}} \leq P_{tr}(i_0, d) \leq 1$;

(c) $D_r(u_r) \leq x_r \leq D_r(u_{i_0})$, *for all* $r = 1, ..., R$, *where*

$$u_r = \frac{u_{i_0}}{P_{tr}(i_0, d)} \left(1 - \frac{d_r(i_0)}{d(i_0)} \right)^{-1};$$

(d) $\sum_r x_r = C$;

where

$$d = \sum_{r=1}^{R} d_r \quad \text{and} \quad x_r = \sum_{i=1}^{N} d_r(i) P_{tr}(i, d),$$

and where we use the convention that $u_{N+1} = \max_{r=1,...,R} u_{\max,r}$, *for* $u_{\max,r}$ *as given in Assumption 1.*

---

[1] To do this, customers may probe the link to discover how to change their allocation in order to improve their net benefit

Proposition 1 states that there always exists an equilibrium allocation, which, however, is not necessarily unique. Let us briefly comment on the Properties (a)-(d) of Proposition 1,

(a) states that, under every equilibrium allocation, at most two different priorities are being used.

(b) states that when priority i is used under a given equilibrium allocation, then the probability that a submitted packet with priority i gets transmitted is lower bounded by u_i/u_{i+1}. This suggests that pricing can be used to control link congestion (see discussion in Section 5).

(c) implies that under a given equilibrium allocation, the lower bound on the expected transmission rate x_r is equal to the demand $D_r(u_r)$ for the price

$$u_r = \frac{u_{i_0}}{P_{tr}(i_0, d)} \left(1 - \frac{d_r(i_0)}{d(i_0)}\right)^{-1}.$$

Therefore, when for customer r the ratio $d_r(i_0)/d(i_0)$ is negligible small (what we will call a "small customer"), then the price u_r is (approximately) equal to $u_{i_0}/P_{tr}(i_0, d)$. Furthermore, using (b), we obtain that in this case we have

$$D_r(u_{i_0+1}) \leq x_r \leq D_r(u_{i_0}).$$

One interpretation of this result is that the priority scheme considered here protects small customers, i.e. under an equilibrium allocation, small customers are guaranteed to obtain an expected rate $x_r \in [D_r(u_{i_0+1}), D_r(u_{i_0})]$.

(d) states that under every equilibrium allocation the whole link capacity gets used.

3.2 Competitive Price-Taking Assumption

In the previous subsection we assumed that, while choosing an allocation, customers anticipate how they affect the transmission probabilities (see Eq. (1)). When there is no analytical model available, this can be done by probing the link to obtain estimates of the gradient of $P_{tr}\left(\sum_r d_r\right)$ with respect to d_r, where $d = (d_1, ..., d_R)$ is the current aggregate allocation over all customers. Unfortunately, such gradient estimates tend to have a large variance which makes this procedure computationally expensive.

As an alternative approach, customers may ignore how they influence the transmission probabilities and tune their allocation as follows;

1. For the current allocation $(d_1, ..., d_R)$, probe the network to obtain (estimates of) the transmission probabilities $P_{tr}(\sum_r d_r) = (P_{tr}(1, \sum_r d_r), ..., P_{tr}(N, \sum_r d_r))$.

2. Update the current allocation $d_r \in \Re^N_+$ towards an optimal solution of optimization problem

$$\max_{y_r \in \Re^N_+} \left\{ U_r \left(y_r, P_{tr} \left(\sum_r d_r \right) \right) - \sum_{i=1}^{N} y_r(i) u_i \right\}. \tag{2}$$

In general, it is easier to estimate the transmission probability $P_{tr}(i, \sum_r d_r))$ than to estimate the gradient of $P_{tr}(i, \sum_r d_r)$ with respect to d_r. An optimal solution to Eq. (2) can easily be computed, i.e. an optimal solution is obtained by choosing

$$d_r(i) = \begin{cases} \frac{D_r\left(u_{i^*}/P_{tr}(i^*,d)\right)}{P_{tr}(i^*,d)} & \text{if } i = i^*, \\ 0 & \text{otherwise} \end{cases}$$

where the priority i^* is such that

$$u_{i^*}/P_{tr}(i^*, d) \leq u_i/P_{tr}(i, d), \qquad \text{for all } i = 1, ...N.$$

The simplifying assumption that customers ignore how they influence the transition probabilities corresponds to the standard competitive price-taking assumption of economic theory. In this context, it can be justified when

(a) many "small" customers are sharing the link capacity so that the influence of an individual customer on the transmission probabilities is negligible.

(b) it is impossible (or too expensive) for a customer to determine how the choice of an allocation may affect the transmission probabilities.

We then define an equilibrium allocation as follows.

Definition 2 *We call an allocation* $(d_1, ..., d_R)$, $d_r \in \Re_+^N$, $r = 1, ..., R$, *an equilibrium allocation under the competitive price-taking assumption when for every user* $r = 1, ..., R$, d_r *is an optimal solution to the optimization problem*

$$\max_{y_r \in \Re_+^N} \left\{ U_r\left(y_r, P_{tr}\left(\sum_{r=1}^{R} d_r\right)\right) - \sum_{i=1}^{N} y_r(i) u_i \right\}.$$

We have the following result.

Proposition 2 *Let Assumption 1 and 2 hold. Then there exists an equilibrium allocation under the competitive price-taking assumption. Furthermore, when* $(d_1, ..., d_R)$,$d_r \in \Re_+^N$, $r = 1, ..., R$, *is an equilibrium allocation under the competitive price-taking assumption, then there exists a priority* $i_0 \in \{1, ..., N\}$ *such that*

(a) $d_r(i) = 0$, *for all* $r = 1, ..., R$ *and all* $i \notin \{i_0, i_0 + 1\}$;

(b) $\frac{u_{i_0}}{u_{i_0+1}} \leq P_{tr}(i_0, d) \leq 1$;

(c) $x_r = D_r(u^*)$, *for all* $r = 1, ..., R$, *where* $u^* = \frac{u_{i_0}}{P_{tr}(i_0,d)}$;

(d) $\sum_r x_r = C$;

where

$$d = \sum_{r=1}^{R} d_r \quad \textit{and} \quad x_r = \sum_{i=1}^{N} d_r(i) P_{tr}(i, d),$$

and where we use the convention that $u_{N+1} = \max_{r=1,...,R} u_{\max,r}$.

This result is Proposition 1 specialized to the case where

$$\frac{d_r(i_0)}{d(i_0)} = 0, \qquad \text{for all } r = 1, ..., R.$$

The price u^* in Proposition 2 can be interpreted as a "congestion price", i.e. the expected rate x_r of each customer r is equal to the demand under the congestion price u^*. The scheme is then "fair" in the sense that all customers see the same congestion price u^*.

4 Customers with QoS-Sensitive Traffic

In this section, we extend our analysis to allow customers with QoS-sensitive traffic. We model this situation by adding to the customer's utility function a penalty term which is sensitive towards link congestion.

Consider a fixed customer r and let $\bar{U}_r(x)$ be an utility function which satisfies Assumption 1. In addition, let $d_r = (d_r(1), ..., d_r(N)$ be the allocation of customer r and let $d_{-r} = (d_{-r}(1), ..., d_{-r}(N))$ be the aggregate allocation by customers different than r. We then associated with customer r the utility $U_r(d_r, P_{tr}(d_{-r} + d_r)$ given by

$$U_r\Big(d_r, P_{tr}(d_{-r} + d_r\Big) = \bar{U}_r(x_r) - \sum_{i=1}^{N} c_r\Big(P_{tr}(i, d_{-r} + d_r)\Big) d_r(i), \tag{3}$$

where

$$x_r = \sum_{r=1}^{R} P_{tr}(i, d_{-r} + d_r) d_r(i)$$

and the function $c_r : [0, \infty) \to [0, \infty)$ is non-increasing and continuous, with

$$c_r(\rho) = 0, \qquad \text{for } \rho \geq 1.$$

The function $c_r\Big(P_{tr}(i, d + d_r)$ models the QoS requirements of user r. When c_r is equal to 0 for all $\rho \geq 0$, we obtain the utility function of the previous section where customers only cared about throughput. A customer with with tight QoS requirements can be modeled by letting the function $c_r(\rho)$ increase sharply when ρ falls below 1.

We study this situation under the competitive price-taking assumption, i.e. we wish to identify an allocation $(d_1, ..., d_R)$ such that for every user $r = 1, ..., R$, d_r is an optimal solution to the optimization problem

$$\max_{y_r \in \Re_+^N} \left\{ U_r\left(y_r, P_{tr}\left(\sum_{r=1}^{R} d_r\right)\right) - \sum_{i=1}^{N} y_r(i) u_i \right\},$$

where the utility function U_r is given by Eq. (3). We have the following result.

Proposition 3 *Let Assumption 1 and 2 hold. Then there exists an equilibrium allocation under the competitive price-taking assumption. Furthermore, when $(d_1, ..., d_R)$,$d_r \in \Re_+^N$, $r = 1, ..., R$, is an equilibrium allocation under the competitive price-taking assumption, then there exists a priority $i_0 \in \{1, ..., N\}$ such that*

(a) $d_r(i) = 0$, *for all* $r = 1, ..., R$ *and all* $i \notin \{i_0, i_0 + 1\}$;

(b) $\frac{u_{i_0}}{u_{i_0+1}} \leq P_{tr}(i_0, d) \leq 1$;

(c) $D_r(u_r) \leq x_r \leq D_r(u_{i_0})$, *for all* $r = 1, ..., R$, *where*

$$u_r = \min\left\{ u_{i_0+1}, \frac{u_{i_0} + c_r\Big(P_{tr}(i_0, d)\Big)}{P_{tr}(i_0, d)} \right\};$$

(d) $\sum_r x_r = C$;

where

$$d = \sum_{r=1}^{R} d_r \quad \text{and} \quad x_r = \sum_i d_r(i) P_{tr}(i, d),$$

and where we use the convention that $u_{N+1} = \max_{r=1,...,R} u_{\max,r}$.

This result is the equivalent of Proposition 2, however the price u_r, which determines the lower bound on the expected rate x_r, now depends on customer's QoS requirements, i.e. the penalty term $c_r\Big(P_{tr}(i_0, d)\Big)$. Customers with tight QoS requirement are willing to pay a higher price to obtain a better QoS service and submit their packets with priority $i_0 + 1$. Customer with elastic traffic can tolerate packet loss and will submit their packets with priority i_0. In this sense, the priority scheme considered here enables the provision of differentiated QoS.

5 Pricing as a Tool to Control Link Congestion

Propositions 1- 3 suggest that pricing can be used to control link congestion. In particular, when for the priority i_0 of Propositions 1- 3 we have that $i_0 < N$, then the lower bound on the probability that a submitted packets gets transmitted under an equilibrium allocation is purely a function of the prices $u_1, ..., u_N$. To ensure that for every equilibrium allocation we have $i_0 < N$, by Property (d) of Propositions 1- 3, it suffices to choose the price u_N high enough so that

$$\sum_r D_r(u_N) < C.$$

This suggest to following procedure to guarantee that under every equilibrium allocation a submitted packets get transmitted with a probability greater or equal to P_0,

1. choose the price u_N so that $\sum_r D_r(u_N) < C$, and,
2. for $i = 1, ..., N$, set
$$u_i = u_N P_0^{N-i}.$$

Generally, the number of customers accessing the link and their utility function will change over time. Let $R(t)$ is the number of customers accessing the link at time $t \geq 0$ and $D_r(u, t)$, $r = 1, ..., R(t)$, are the corresponding demand functions. To ensure that the probability that a submitted packets gets transmitted never falls below P_0 under an equilibrium allocation, the prices $u_1, ..., u_N$ have to be chosen such that the following holds,

$$\max_{i \in \{1,...,N-1\}} u_i/u_{i+1} \geq P_0 \quad \text{and} \quad \sum_r D_r(u_N, t) < C, \quad \text{for all } t \geq 0.$$

As it is difficult to predict the aggregated link demand (especially over a long time horizon), one would expect that either the prices associated with the different priorities, or the link capacity C, have to be updated from time to time to track the conditions above. However, by starting with a conservative choice of u_N, i.e. by choosing u_N such that $\sum_r D_r(u_N) << C$, these updates may have to be done only infrequently (for example every month) - making the implementation of this priority scheme simpler than the one considered in [3] where prices associated with the different priorities levels are dynamically updated on a fast time-scale (in the order of seconds).

6 Future Work

Regarding future work, it would be interesting to investigate whether the existence of multiple equilibria can lead to oscillations. Related to that issues is the question whether there exists an algorithm which ensures that the customer's allocations converge to an equilibrium.

References

[1] S. Chen and K. Park, *"An Architecture for Noncooperative QoS Provision in Many-Switch Systems,"* Proc. IEEE INFOCOM '99, 1999.

[2] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, *"Pricing in Computer Networks: Motivation, Formulation, and Example,"* IEEE/ACM Transactions on Networking, 1, 1993.

[3] A. Gupta, D. O. Stahl, and A. B. Whinston, *"A Stochastic Equilibrium Model of Internet Pricing,"* Journal of Economics Dynamics and Control, 21:697–722, 1997.

[4] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, *" Assured forwarding phb group,"* IETF: RFC 2597, ftp://ftp.isi.edu/in-notes/rfc2597.txt, 1999.

[5] A. Odlyzko, *"Paris Metro Pricing for the Internet"*, http://www.research.att.com/∼amo, 1998.

[6] K. Park, M. Sitharam, and S. Chen, *"Quality of Service Provision in Noncooperative Networks: Heterogenous Preferences, Multi-Dimensional QoS Vectors, and Burstiness,"* Proc. International Conference on Information and Computation Economies, 1998.

[7] Ch. Semeria and F. Fuller, *"3com's Strategy for Delivering Differentiated Service Levels,"* http://www.3com.com/technology/tech_net/white_papers/500652.html, 1999.