ON FLAT-RATE AND USAGE-BASED PRICING FOR TIERED COMMODITY INTERNET SERVICES

G. Kesidis and A. Das

CSE and EE Depts
Pennsylvania State University
kesidis@engr.psu.edu

G. de Veciana

ECE Dept
University of Texas at Austin
gustavo@ece.utexas.edu

ABSTRACT

In this note, we discuss issues pertaining to end-to-end quality-of-service management of commodity Internet applications and associated pricing incentive mechanisms. The issue of service differentiation is first studied using a simple two-class model including delay and throughput sensitive traffic. We show that by introducing service differentiation one can make more efficient use of resources, however this depends on the differences in required QoS as well as the typical capacities of the systems involved. As such, service differentiation may be more beneficial in lower capacity access networks than in high capacity core networks. We then focus on delay-sensitive and study flat-rate versus usage-based pricing under overload conditions. Our results suggest that in overload scenarios usage-based pricing is advantageous both from the system perspective, i.e., reduces degree of overload, and individual users' perspective, increases their perceived utilization.

1. INTRODUCTION

Though high-speed Internet service providers are aggressively selling bundled triple play services (voice, video, data) to broadband residential as well as wireless customers, *core* networks typically provide only single-tier bulk transport. Recently the question of network neutrality has come to the fore. The question is whether core providers should be permitted to price packets based in part on their origin or the type of application. That is, should core providers be permitted not to be "neutral" in their pricing? Currently, the access network (ISPs) are not neutral in that there is differential access bandwidth for subscribers' traffic aggregates (at differential flat rates), and "terms of use" restrictions (e.g., no third-party VoIP over wireless data networks). In the US, the Internet core is not a public utility but a federation of private networks whose owners want to conserve revenue streams associated with their "managed" telephony and video services, which are either migrating or morphing into end-to-end Internet data applications. At the same time core providers wish to improve their less profitable core data networking businesses (here then lies a Catch-22).

A typical pricing formula for data at a client-network interface or network-network interface is

$$F + C(\rho - \bar{\rho})^+ \tag{1}$$

where F \$/s is the flat rate price (depending on the maximum access bandwidth– more generally, a Service Level Agreement (SLA) profile –and, possibly, time of day, and type of day) and C \$/byte is the usage based charge for the net throughput $\rho - \bar{\rho} > 0$ over a threshold $\bar{\rho}$ byte/s (i.e., beyond which overage charges are applied). That is, the charges are solely based on aggregate traffic volume. Commodity Internet traffic are unwarranted based on these issues and not on any perceived threats to the web's "freedom" (especially considering that the highly commercialized web is itself hardly neutral).

In light of the network neutrality question and service migration,[1] ISPs and core network providers are contemplating deployment of end-to-end classes-of-service (tiered services) together with associated pricing incentives. Premium service is intended for high-volume delay-sensitive applications (such as high-definition video conferencing and distributed gaming) and Virtual Private Networks (VPNs).

A multiple class-of-service architecture is not without controversy. A single class flat rate pricing system is advocated in [17, 15]. Presently, competition is driving some growth in the core and flat rate pricing does promote growth in traffic [17]. Moreover, flat rate pricing appears to be preferred by end-users. However from the providers point of view raising (flat) prices is difficult because of competition and regulations against price fixing. Also, additional traffic may yield "proportionately" increasing revenue. Finally, end-users may dislike a "ticking clock" but, considering

G. KESIDIS IS SUPPORTED BY A CISCO URP GIFT AND BY THE NSF UNDER GRANTS CYBERTRUST 0524202 AND NETS-WN 0219747. G. DE VECIANA IS SUPPORTED BY THE NSF UNDER GRANTS CNS-0509355 AND CNS-0721532.

[1] Service migration to the commodity Internet does, however, face challenges such as a heightened threat of spam in the case of VoIP and to the integrity of advertising (and associated revenues) in the case of TV video.

long-distance telephony, one can argue that they would expect additional costs for valued premium services and may not want to look at the ticking clock while engaging in them.

In this paper, we first argue for multiple classes of service. Access to service classes needs to be regulated by a SLA involving a traffic profile that includes specification of a maximum transmission rate), and they need to be differentially priced. Premium classes-of-service (CoSs) could be charged according to tiered flat rates, possibly resulting in end-user "promotion" scheduling so that premium CoS are always fully utilized [8].

Alternatively, usage priced premium CoSs will reduce the volume of premium traffic in play as users will only assign applications to premium CoSs as needed. Usage based pricing is consistent with a *temporary* need for reserved end-to-end bandwidth for certain high-volume real-time data applications as mentioned above. But usage based pricing requires an authenticated billing system that may be costly to mount [15]. In particular, at the network periphery, usage based billing of premium applications needs to be authenticated to an "authorized" human, e.g., by visual challenge/response (as used in the web), or by prompting for a secret code not stored on the computer (e.g., using frequently refreshed RSA SecurID tokens). Also, the received QoS of applications under premium service may need to be monitored. Detailed authentication and network monitoring mechanisms may, however, have other security benefits, e.g., facilitate flow attribution and monetary deterrence for attack and nuisance/spam activity in premium service classes. Also, a billing system may be feasible to mount if associated revenues are potentially high and "scalable" accounting practices are used.

Such an end-to-end CoS system over the commodity Internet might be achieved through a middleware architecture such as MIDCOM [19, 20, 4]. This system could also coordinate existing QoS mechanisms at the client-network interface [1], inside routers (e.g., "weighted fair" queuing), and proposals for inter-domain MPLS [18, 3], the latter integrating with BGP signaling[2].

The balance of this paper is organized as follows. Using a highly idealized model, we motivate the use of classes of service in Section 2. In Section 3, tiered flat rate and usage based pricing are compared and Section 4 summarizes our observations.

2. SERVICE CLASSES FOR COMMODITY INTERNET ACCESS

Consider a two-class M/GI/1 queue supporting both delay-sensitive traffic over throughput-sensitive traffic. Suppose the delay quality of service (QoS) requirement is on a packet's mean delay *before* service (or waiting time), W,

$$\mathsf{E}[W] \quad \leq \quad \delta, \ \text{ for } \delta > 0. \tag{2}$$

Arrivals for the two classes are Poisson with intensities λ_d and λ_t respectively. An arrival for each class corresponds to a packet or burst which has an average size of s_d and s_t bits, and variance σ_d^2 and σ_t^2 respectively. Using the Pollaczek-Khintchine formula one can determine the required service capacity c bits/sec. In particular following the work of Kelly [11, 5], we can show that (2) will be satisfied if

$$\alpha_d(\delta) + \alpha_t(\delta) < c \tag{3}$$

where

$$\begin{aligned} \alpha_d(\delta) &= \lambda_d(s_d + \frac{1}{2\delta c}(s_d^2 + \sigma_d^2)) \\ \alpha_t(\delta) &= \lambda_t(s_t + \frac{1}{2\delta c}(s_t^2 + \sigma_t^2)). \end{aligned}$$

are referred to as the effective bandwidths of the two types of traffic with QoS requirement δ and a multiplexer with capacity c. Note that $\alpha_d(\cdot)$ and $\alpha_t(\cdot)$ depend on c, but one can easily determine the minimal capacity μ_1 one should provision to meet the QoS constraint (3) by solving a quadratic equation to obtain

$$\mu_1 = \frac{\rho}{2}\left(1 + \sqrt{1 + \frac{2}{\delta} \cdot \frac{s_d^2 + s_t^2 + \sigma_d^2 + \sigma_t^2}{\rho^2}}\right)$$

where

$$\rho \quad \equiv \quad \rho_d + \rho_t \ \equiv \ \lambda_d s_d + \lambda_t s_t$$

is the total load in bits/sec. The nonlinear dependence on c of $\alpha_d(\cdot)$ and $\alpha_t(\cdot)$ captures the decrease in the effective bandwidth requirement with increasing capacity, i.e., averaging out of fluctuations on bigger links. Thus the capacity required μ_1 to support these two flows under a fixed QoS constraint approaches the average load ρ as the total load ρ increases.

Using the same approach as in [5], one can show that by introducing class based differentiation at the multiplexer, in particular pre-emptive (resume) priority to delay sensitive traffic, one would only require that

$$\alpha_d(\delta) < c$$

to meet the delay sensitive traffic's QoS requirement, and

$$\alpha_d(\infty) + \alpha_t(\infty) < c$$

to meet the throughput sensitive traffic's requirement (i.e., merely that the queue be stable). Together these two constraints determine the minimal capacity μ_2 one would have

[2]Note that this augmented BGP signaling for premium classes-of-service is "secured" inter-domain through associated financial incentives and QoS monitoring by all parties involved.

to provision for a multiplexer with class based differentiation. This can be explicitly determined to be

$$\mu_2 = \max\left\{\rho_d + \rho_t,\ \frac{\rho_d}{2}\left(1+\sqrt{1+\frac{2}{\delta}\cdot\frac{s_d^2+\sigma_d^2}{\rho_d^2}}\right)\right\}.$$

Clearly $\mu_2 < \mu_1$. In particular when δ is small, the right hand term in the above maximum dominates, and we have

$$\mu_1 - \mu_2 \approx \frac{\rho_t}{2} + \frac{1}{\sqrt{2\delta}}\left(\sqrt{s_d^2+s_t^2+\sigma_d^2+\sigma_t^2} - \sqrt{s_d^2+\sigma_d^2}\right).$$

Alternatively when both ρ_t, ρ_d are large and the left hand term in the maximum dominates then we have roughly

$$\mu_1 - \mu_2 \approx \frac{1}{2\delta}\cdot\frac{s_d^2+s_t^2+\sigma_d^2+\sigma_t^2}{\rho}.$$

In summary when delay constraints are stringent, service differentiation in the multiplexer will reduce the required capacity; but due to statistical multiplexing , the savings from introducing service differentiation decreases with the total offered load. Beyond savings in the required capacity, service differentiation may facilitate traffic engineering, security, and/or reliability in the core (e.g., by protecting one class of traffic from another through segregation).

Similar results can be obtained using large buffer asymptotic effective bandwidths and virtual delay tail constraints for more general arrival processes, including the context of scheduled queues [6, 10]. A related simulation study is [22].

3. TIERED FLAT RATE VERSUS USAGE BASED PRICING

Let us now consider a queue nominally handling only delay-sensitive traffic with a *fixed* service rate μ. In the sequel we no longer use the subscripts "d" and "t" to distinguish types of traffic. However we suppose that there are a fixed number N of users and that the n^{th} delay-sensitive user has an offered load load

$$\rho_n (= \lambda_n s_n) \quad \leq \quad \rho_n^{\max},$$

where λ_n denotes the packet/burst arrival rate for the user, and s_n is the mean packet size. Each user derives utility from successfully transferring load which is a function of the form

$$U_n(\underline{\rho}) \quad \equiv \quad \begin{cases} \rho_n & \text{if } \sum_k \rho_k \leq \Lambda \\ \rho_n \exp\left(-\frac{\sum_k \rho_k}{\Lambda\beta_n}\right) & \text{else,} \end{cases}$$

where $\underline{\rho}$ is the N-vector of traffic rates ρ_n, and $\beta_n \geq 0$ is a user-dependent parameter, and Λ is an effective limit on the offered load.

The proposed "utility" function is intended to reflect increasing utility in the supported load, when the system is not overloaded, and decreases in utility with the supported load if the system is overloaded. The parameter β_n captures the extent to which user n can tolerate overloads, i.e., a violation of (2). For example, a VoIP connection may be able to tolerate fluctuating available bandwidth by adjusting the data compression rate. If $\beta_n = 0$ then the n^{th} user is intolerant of violations of (2), but if $\beta_n = \infty$ then the n^{th} user is actually transmitting a best-effort flow into the delay sensitive queue. This may temporarily or predominantly be the case when end-users, who are not actively engaged in delay-sensitive applications, employ automated "upgrade" scheduling [8] of best-effort traffic to premium service classes, which they can freely do under tiered flat rate pricing.

We compare flat rate versus usage based pricing in the regime where there is excess demand,

$$\sum_k \rho_k \quad > \quad \Lambda. \tag{4}$$

In the setting of the previous subsection,

$$\Lambda \quad = \quad \mu - \frac{1}{2\delta\mu}\sum_n (s_n^2 + \sigma_n^2)$$

where σ_n^2 denotes the variance in packet sizes for the user n and μ is the capacity allocated to the multiplexer, and δ is a QoS constraint. We also assume that the equilibria $\underline{\rho}$ reached are such that $\rho_n < \rho_n^{\max}$ for all n, i.e., they are not peak-rate limited equilibria.

3.1. Flat rate pricing

Under flat rate pricing, we assume a cost structure $M(\rho^{\max})$ for the maximum transmission rate of a user. That is, the downlink to the user n is always less than $\rho_n^{\max}$ and the overage threshold $\bar{\rho_n} = \infty$ in (1).

Under (4),

$$\frac{\partial}{\partial\rho_n} U_n(\underline{\rho}) \quad = \quad 0$$

when $\Lambda\beta_n \leq \rho_n^{\max}$ and the arrival rate by user n is *chosen* to be

$$\hat{\rho}_n \quad = \quad \beta_n\Lambda. \tag{5}$$

This requires

$$\sum_n \beta_n \quad > \quad 1 \tag{6}$$

so that, indeed, (4) holds under $\underline{\hat{\rho}}$ (note that this will always be the case if $\beta_n = \infty$ for some user n).

We assume that, for all n, $\hat{\rho}_n$ is feasible by user n, i.e., $\rho_n^{\max} > \beta_n\Lambda$ (no peak-rate limitations). Which translates to

$$\sum_k \rho_k^{\max} \quad \leq \quad \Lambda,$$

i.e., overbooking allocations in a flat rate system to attempt to use the resources inefficiently.

3.2. Usage based pricing

Under usage based pricing, users will choose their rates ρ_n to maximize their net benefit

$$U_n(\underline{\rho}) - C\rho_n$$

where $C > 0$ is the usage based charge. Herein we take the overage threshold $\bar{\rho} = 0$ and assume that the access fees F in (1) are approximately equal in both the flat rate and usage based cases.

In this case under (4), user n will choose $\rho_n = \rho_n^*$ so as to satisfy

$$\frac{\partial}{\partial \rho_n} U_n(\underline{\rho}^*) \quad = \quad C$$

which implies

$$\Lambda\beta_n - \rho_n^* \quad = \quad C\Lambda\beta_n \exp\left(\frac{\sum_k \rho_k^*}{\Lambda\beta_n}\right). \qquad (7)$$

Adding these equations for all users n and then dividing by $\Lambda \sum_k \beta_k$ gives

$$\begin{aligned} 1 - \frac{\sum_k \rho_k^*}{\Lambda \sum_k \beta_k} &= C \sum_n \exp\left(\frac{\sum_k \rho_k^*}{\Lambda\beta_n}\right) \frac{\beta_n}{\sum_k \beta_k} \quad (8) \\ &\geq C \exp\left(\frac{\sum_k \rho_k^*}{\Lambda \sum_k \beta_k^2 / \sum_k \beta_k}\right) \\ &\geq C \exp\left(\frac{\sum_k \rho_k^*}{\Lambda \sum_k \beta_k}\right) \quad (9) \end{aligned}$$

where the first inequality is Jensen's and the second is a simply due to $\sum_k \beta_k^2 \leq (\sum_k \beta_k)^2$. Thus, we can rewrite this final inequality in the form $1 - x \geq C\exp(x)$. So, necessary requirements for a solution to (8) are that the cost $C < 1$ and that

$$\Lambda \quad < \quad \sum_k \rho_k^* \; < \; \Lambda \sum_k \beta_k \; = \; \sum_k \hat{\rho}_k,$$

where the first inequality is just (4) and the second is $x < 1$. Again, (6) is clearly required.

3.3. Comparison

Comparing the total flat rate demand $\sum_n \hat{\rho}_n = \Lambda \sum_n \beta_n$ to the usage based demand $\sum_n \rho_n^*$, we conclude:

Lemma 1 *Assuming excess overall demand (4) wherein every user n does not achieve their peak demand $\rho_n^{\max}$, overall demand is lower under usage based pricing compared to flat rate pricing.*

That is, the conditions of this lemma are:

- Excessive demand (4) in the context of delay-sensitive QoS constraint (2),
- $C < 1$ for usage based pricing (i.e., sufficiently low usage based price), and
- $\hat{\rho}_n = \beta_n \Lambda < \rho_n^{\max}$ for all users n under flat rate pricing (i.e., sufficiently low flat rate price).

Considering now that the delay-sensitive queue under flat rate pricing may also be handling throughput-sensitive traffic [8], we can further argue that the β parameters are higher in the flat rate case ($\hat{\beta}_n$) than in the usage-based case (β_n^*). That is,

$$\Lambda \quad < \quad \sum_k \rho_k^* \; < \; \Lambda \sum_k \beta_k^* \; \leq \; \Lambda \sum_k \hat{\beta}_k \; = \; \sum_k \hat{\rho}_k,$$

and Lemma 1 would still continue to hold.

Note that (5) and (7) imply that $\hat{\rho}_n \geq \rho_n^*$. Thus a rephrasing of the previous display is:

$$\frac{\rho_n^*}{U_n(\underline{\rho}^*)} \; = \; \frac{\hat{\rho}_n - \rho_n^*}{C\hat{\rho}_n} \quad \leq \quad \frac{\hat{\rho}_n}{U_n(\underline{\hat{\rho}})}$$

for all end-users n. Clearly, the utility of user n under usage-based pricing, $U_n(\underline{\rho}^*)$, is higher than that under flat rate pricing, $U_n(\underline{\hat{\rho}})$, if

$$\hat{\rho}_n \quad \leq \quad \rho_n^*.$$

Also, the *net* utility of user n under usage-based pricing

$$U_n(\underline{\rho}^*) - C\rho_n^* \quad \geq \quad U_n(\underline{\hat{\rho}})$$

if

$$U_n(\underline{\hat{\rho}}) \left(\frac{1}{\hat{\rho}_n} - \frac{1}{\rho_n^*}\right) \quad \geq \quad C,$$

recalling that $\underline{\rho}^*$ depends on the usage-based charging rate C.

Related results in this framework include those of the traffic capacity of queues with rate-regulated traffic and constraints on queueing delay [13, 7, 8].

4. CONCLUSIONS

Our goal in this paper was to address some aspects of end-to-end QoS mangement of commodity Internet applications, and specifically view the relative merits of two pricing mechanisms: (tiered) flat rate versus usage based. The argument centers on two natural points. First, in supporting heterogeneous applications requiring QoS, networks can benefit substantially from putting into place mechanisms for service differentiation. Our simple model, shows what we believe to be intuitively correct, i.e., that substantial capacity savings should be possible particularly if the differences in QoS requirements are high. However this is less so if the network capacity in play is high, as would be the case in the

core vs the access network. In other words when there are substantial gains to be reaped from statistical multiplexing service differentiation is less critical. Second we compare the role of pricing mechanisms under an overload and overbooking scenario. Our premise is that the critical impact of such mechanisms is best seen in overload regime, where providers will be seeking to make the most of available resources are likely to hope to operate. Using a simple model, we show that usage based pricing, leads to a reduction in the aggregate overloads, and higher individual user utilities versus flat rate pricing.

5. REFERENCES

[1] DOCSIS Overview and 3.0 Interface, www.cablemodem.com, Jan. 2005.

[2] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow. RSVP-TE: Extensions to RSVP for LSP tunnels. IETF RFC 3209, www.ietf.org, Dec. 2001.

[3] A. Ayyangar and J.-P. Vasseur. Inter-domain GMPLS traffic engineering - RSVP-TE extensions, IETF Internet Draft, March 2006.

[4] M. Barnes (Ed.) Middlebox Communications (MIDCOM) Protocol Evaluation. IETF RFC 4097, www.itef.org, June. 2005.

[5] G. de Veciana and J. Walrand. Effective bandwidths: Call admission, traffic policing, and filtering for ATM Networks. Queueing Systems 20, pp. 37-59, 1995.

[6] G. de Veciana and G. Kesidis. Bandwidth Allocation for multiple qualities of service using generalized processor sharing. IEEE Trans. Info. Th., Vol. 42, No. 1, pp. 268-271, Jan. 1996.

[7] R.J. Gibbens, S.K. Sargood, F.P. Kelly, H. Azmoodeh, R. Macfadyen, and N. Macfadyen. An approach to service level agreements for IP networks with differentiated services. Phil. Trans. R. Soc. Lond., vol. 358, pp. 2165-2182, 2000.

[8] R. Haddad and Y. Viniotis. 3-Tier service level agreement with automatic class upgrades. IEEE GLOBECOM Workshop on Enabling the Future Service-Oriented Internet, Washington, DC, Nov. 2007.

[9] Y. Jin and G. Kesidis. Dynamics of usage-priced communication networks: the case of a single bottleneck resource. IEEE ToN, Oct. 2005.

[10] S. Jiwasurat and G. Kesidis. Hierarchical Shaped Deficit Round-Robin Scheduling. In *Proc. IEEE GLOBECOM*, 2005.

[11] F.P. Kelly. Effective bandwidths for multi-class queues. Queueing Systems 9, pp. 5-16, 1991.

[12] F.P. Kelly. Charging and rate control for elastic traffic. European Transactions on Telecommunications, Vol. 8, pp. 33-37, 1997.

[13] G. Kesidis and T. Konstantopoulos. Extremal Traffic and Worst-Case Performance for a Queue with Shaped Arrivals. In *Analysis of Communication Networks: Call Centres, Traffic and Performance*, edited by D.R. McDonald and S.R.E. Turner, Fields Institute Communications/AMS, 2000 (from Proc. Fields Institute Conference, Toronto, Nov. 1998).

[14] Y.A. Korilis, A.A. Lazar and A. Orda. Achieving Network Optima Using Stackelberg Routing Strategies. IEEE/ACM Trans. Networking, Vol. 5, No. 1, 1997.

[15] D. Levinson and A. Odlyzko. Too expensive to meter: The influence of transaction costs in transportation and communications. Preprint, Sept. 2007.

[16] R.R. Mazumdar, C.A. Courcoubetis, N. Duffield, G. Kesidis, A. Odlyzko, R. Srikant, and J. Walrand. Guest Editorial: Price-based access control and the economics of networking. IEEE JSAC, May 2006, http://www.jsac.ucsd.edu/TOC/2006/may06.html

[17] A. Odlyzko. Internet pricing and the history of communications. Computer Networks Journal, Vol. 36, No. 5/6, pp. 493-517, 2001.

[18] E. Rosen and Y. Rekhter. BGP/MPLS IP Virtual Private Networks (VPNs). IETF RFC 4364, www.ietf.org, Feb 2006.

[19] P. Srisuresh, J. Kuthan, J. Rosenberg, A. Molitor, and A. Rayhan. Middlebox Communications Architecture and Framework. IETF RFC 3303, www.ietf.org, Aug. 2002.

[20] R.P. Swale, P.A. Mart, P. Sijben, S. Brim, and M. Shore. Middlebox Communications (MIDCOM) Protocol Requirements. IETF RFC 3304, www.itef.org, Aug. 2002.

[21] R.W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ, 1989.

[22] M. Yuksel, K.K. Ramakrishnan, S. Kalyanaraman, J.D. Houle, R. Sadhvani. Value of supporting class-of-service in IP backbones. IEEE IWQoS, 2007.