

Service Differentiation for Delay-Sensitive Applications: An Optimisation-Based Approach

Peter Key^a, Laurent Massoulié^a and Jonathan Shapiro^b

^a Microsoft Research Limited, 7 JJ Thomson Avenue, Cambridge, UK.

^b Department of Computer Science, University of Massachusetts at Amherst, MA, USA

E-mail: peterkey@microsoft.com, lmassoul@microsoft.com, jshapiro@cs.umass.edu

Abstract— This paper deals with the performance of delay-sensitive applications running over a network that offers multiple classes of service. We first discuss the gain in utilisation allowed by the introduction of several classes of service by comparing schedulable regions for different schedulers. We then discuss what feedback information should be sent to traffic sources from different classes. We establish a connection between the *sample-path shadow price* rationale for feedback synthesis and the *rare perturbation analysis* technique for gradient estimation in discrete event systems theory. We propose several marking schemes, for simple priority or earliest-deadline-first-based differentiation. The interaction of these marking algorithms with simple congestion control algorithms is studied via simulations.

I. INTRODUCTION

With a single best-effort service class it is possible to offer a simple form of service differentiation [2], whereby users who pay a higher price achieve a proportionally higher transmission rate. However, applications have very different sensitivities to delay, and a single service class is enough only if it provides queueing delays that are acceptable to the application with the most stringent requirements.

The question of whether delay-sensitive applications require service differentiation was addressed by Bajaj et al. [3], who found the answer depends on the adaptive behaviour of the applications themselves and also on the burstiness of traffic. We study rate-adaptive control rather than the delay-adaptive behaviour considered in [3], but reach similar conclusions.

Alvarez and Hajek [4] also consider the necessity of multiple classes in the case of a mixed population of rate-sensitive and loss-sensitive users within the pricing framework of Gibbens and Kelly [2], which we also adopt here. They conclude that two classes are indeed required to satisfy the requirements of both user types, but that the benefit of service differentiation is only realized when the price per packet for each class of service is set correctly. The authors show the existence of the correct prices empirically but provide no way for the network to find them. Although we replace the loss-sensitive users with delay-sensitive ones, our research complements [4] by providing a mechanism for adaptively setting prices.

We first discuss (Section II) the gain in utilisation offered by the introduction of two classes of service at a single resource. We show that two differentiation schemes—priority scheduling and earliest-deadline-first—yield very similar performance and

that both outperform a single-class FIFO scheduler. Moreover, we confirm the findings in [3] that the performance gain depends on traffic burstiness.

We then discuss what feedback information should be sent to traffic sources from different classes. In the context of the optimization-based congestion control framework we use, this problem reduces to one of correctly setting prices. We exhibit a connection between the *sample-path shadow price* rationale for feedback synthesis and the *rare perturbation analysis* technique for gradient estimation in discrete event systems theory (Section 3), then propose marking schemes for simple priority-based differentiation with a measure of cost based on loss or delay, and also for earliest-deadline-first-based differentiation with loss-based cost (Section 4). The interaction of these marking algorithms with simple congestion control algorithms is studied via simulations (Section 5).

II. SCHEDULABLE REGIONS

In this section we discuss the potential performance benefits of having multiple service classes by considering the following simple scenario. Packets of two types reach a single resource; type 1 packets must be served by time d_1 , and type 2 packets must be served by time d_2 , with $d_2 > d_1$. We want to compare the pairs of achievable rates (ρ_1, ρ_2) that can be sustained by various differentiation mechanisms. We assume Poisson arrivals with rates λ_1, λ_2 in each class, unit capacity $c = 1$, and fixed packet sizes σ . In this model the burstiness is captured by the parameter σ^1 . We assume an infinite buffer, so that packets don't get dropped. Now define the schedulable region as the set of parameters (ρ_1, ρ_2) (where $\rho_i := \lambda_i \sigma$) such that the probability of a class i packet experiencing a delay larger than d_i is less than some fixed value θ , e.g. $\theta = 1\%$. Our aim is to compare schedulable regions for single class and FIFO scheduling, and two classes with either EDF or priority scheduling. We apply recent results in heavy traffic approximation theory to obtain tractable formulas.

For FIFO, the total arrival rate $\lambda = \lambda_1 + \lambda_2$ must be such that the probability of experiencing a delay larger than d_1 is smaller than θ . Let W denote the stationary workload in the queue. Then we need to ensure that $\mathbf{P}(W + \sigma > d_1) < \theta$. In heavy traffic, the distribution of W is close to that of $\lambda \sigma^2 / (2(1 - \rho))X$,

¹An alternative approach considered in [1] consists in describing traffic burstiness by (σ, ρ) -constraints. Schedulable regions can then be determined by using available bounds on the worst case delays.

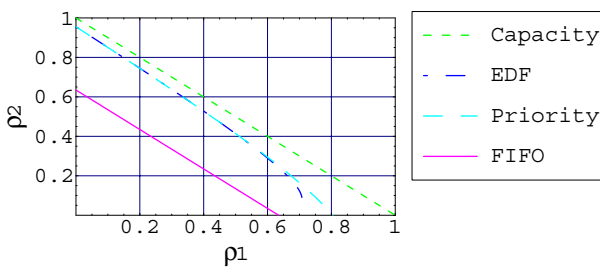


Fig. 1. Schedulable regions for $d_1 = 5$, $d_2 = 50$, $\theta = 0.01$, $\sigma = 1$.

where X is an exponential random variable with unit mean (see e.g. [5]). Hence, we must ensure that

$$\rho_1 + \rho_2 \leq \frac{2(d_1 - \sigma)}{-\sigma \log(\theta) + 2(d_1 - \sigma)}. \quad (1)$$

For preemptive priority scheduling, class 1 packets do not see class 2 packets. Applying the same heavy traffic approximation for the stationary workload due to class 1 packets only shows that we must ensure that

$$\rho_1 \leq \frac{2(d_1 - \sigma)}{-\sigma \log(\theta) + 2(d_1 - \sigma)}. \quad (2)$$

For class 2 packets, martingale calculus can be used to establish the following convergence in distribution,

$$(1 - \rho)T \xrightarrow{D} \frac{\lambda \sigma^2}{2(1 - \rho_1)} \text{Exp}(1).$$

where T is the sojourn time of a class 2 packet. Using the limiting distribution as an approximation, the condition $\mathbf{P}(T > d_2) \leq \theta$ reads:

$$\rho_2 \leq \frac{2(1 - \rho_1)(d_2 - \sigma)}{-\sigma \log(\theta) + 2(1 - \rho_1)(d_2 - \sigma)} - \rho_1. \quad (3)$$

For preemptive resume EDF scheduling where all customers are accepted, by applying recent heavy traffic results [6] we find [1] that the rates need to satisfy

$$\frac{2(1 - \rho)(d_1 - \sigma + (d_2 - d_1)(1 - \rho_1/\rho))}{\sigma \rho} \geq -\log \theta. \quad (4)$$

Figure 1 illustrates the different schedulable regions, obtained from (1) for FIFO, (2) and (3) for Priority, and (4) for EDF.

The comparison of the schedulable regions based on this and similar plots suggests that in the presence of burstiness, service differentiation can provide significant advantages. In most cases the EDF offers a larger gain than Priority, however the two are almost indistinguishable, which argues in favour of implementing the simpler Priority scheme.

III. RARE PERTURBATION ANALYSIS (RPA) BASIS FOR SAMPLE PATH SHADOW PRICES

We now place ourselves in the congestion pricing framework described by Kelly et al. [2], [7], which requires the network to send feedback signals (or charges) to users, perhaps conveyed

by a packet marking strategy, that reflect the marginal cost of congestion in the network. We consider both delay-based and loss-based congestion costs. In each case, one has a measure of cost incurred by a packet (sojourn time in the queue for delay, 1 if the packet is lost, zero otherwise for loss), denoted by Z_n for the n^{th} packet.

The theory of congestion pricing is usually formulated in terms of fluid models. If $C(x)$ is the cost per time unit incurred when the bandwidth consumed at the resource is x , then the charge per unit time per unit bandwidth is $C'(x)$. On the other hand, the sample path shadow price (s.p.s.p.) of a packet is defined (see [8]) as the difference between the actual cost and that which would occur if this packet had not been submitted.

In Rare Perturbation Analysis (RPA, see [9]), the goal is to estimate the gradient of a performance measure with respect to parameters based on a single simulated sample path. In this section, we relate sample path shadow prices to RPA estimates of the derivative of a cost function with respect to the aggregate rates of one or more classes of service.

Consider an M/G/1 queue that supports multiple classes of service, indexed by $i \in \{1, \dots, I\}$. Arrivals for each class are assumed to be Poisson, with the vector $\mathbf{x} = (x_i)$ describing their respective rates. Define the cost function $C(\mathbf{x})$ as the average cost incurred to packets per time unit. Thus $C(\mathbf{x}) = \sum_i x_i J(\mathbf{x})$, where $J(\mathbf{x})$ is the average per-packet cost. Assume that packets $1, \dots, N$ constitute a busy period for the queue under consideration. Let $t(n) \in \{1, \dots, I\}$ denote the class of packet n . For our two cost structures, the average per-packet cost $J(\mathbf{x})$ can be expressed according to the well-known cycle formula as

$$J(\mathbf{x}) = \frac{\mathbf{E} \sum_{n=1}^N Z_n}{\mathbf{E}(N)}. \quad (5)$$

RPA can be applied to obtain estimates of the derivatives of both the numerator and denominator in this expression (see [1]). This yields the expression for $x_i \partial_{x_i} J(\mathbf{x})$:

$$\frac{\mathbf{E} \left[\sum_{m=1}^N 1_{t(m)=i} \sum_{n=1}^N (Z_n - Z_n^{(-m)}) \right]}{\mathbf{E}[N]} - \frac{\mathbf{E}[\sum_{n=1}^N Z_n]}{\mathbf{E}[N]} \times \frac{\mathbf{E}[N_i]}{\mathbf{E}[N]},$$

where N_i is the number of type i packets in the busy period, and $Z_n^{(-m)}$ is the cost to packet n had packet m been removed. This in turn implies that

$$\partial_{x_i} C(\mathbf{x}) = \frac{\mathbf{E} \sum_{m=1}^N 1_{t(m)=i} \sum_{n=1}^N (Z_n - Z_n^{(-m)})}{\mathbf{E}[N_i]}. \quad (6)$$

Since packet m has an impact only on packets belonging to the same busy period, the sample path shadow price for packet m is exactly $\sum_{n=1}^N Z_n - Z_n^{(-m)}$. Applying the cycle formula, it is then seen that the corresponding average packet price for packets of a given type, say type i , does coincide with that derived from the RPA analysis as in (6).

IV. PROPOSED MARKING SCHEMES

A. Single Class of Service

We consider a single FIFO queue, with either loss or delay-based cost.

1) *Loss-Based Cost*: For loss-based cost, the corresponding cost Z_n to packet n is 1 if packet n is lost and zero otherwise. As argued in [8], the sample path shadow price is 1 if in the nominal trajectory, one of the packets $m, m+1, \dots, N$ is lost, and zero otherwise. Hence all packets until the last lost packet in the busy period, included, are marked (i.e. have a unit s.p.s.p.).

Because this scheme is not implementable, Gibbens and Kelly [8] suggest using an approximate marking scheme, marking all packets from the *first* lost packet until the end of the busy period. This scheme marks the correct number of packets per busy period on average, a fact that derives from the stochastic reversibility of sample paths of the M/D/1/C queue.

One drawback of this approach is that it does not charge the ‘right packets’. This suggests the alternative where a packet is charged by its expected s.p.s.p. given the information available. Assuming for simplicity exponentially distributed service times with mean μ^{-1} , an incoming packet finding $n-1$ packets in the queue, would then be charged by the amount $(\rho^{-n}-1)/(\rho^{-C-1}-1)$, where $\rho = x/\mu$, and x is the packet arrival rate.

2) *Delay-Based Cost*: Here, the cost Z_n to packet n is simply its sojourn time in the queue. An adaptation of the time-reversal approach of Gibbens and Kelly [8] consists in charging a packet n according to the cumulated impact of previously arrived packets on its delay. This scheme puts the same cumulative charge on any busy period as the exact non-causal scheme. However it requires state variables to be kept for each packet having previously arrived in the current busy period, which makes it too complex to implement.

A more tractable scheme is obtained when one makes the approximation that a packet imposes a delay equal to its service time on all later-arriving packets in the busy period. This amounts to marking packet n by its waiting time Z_n , plus the length of the current busy period when it enters service. The amount by which this overestimates the correct packet price on average can be computed exactly for an M/D/1 queue, where the resulting average price is never more than twice the correct price, and approaches the correct price as load increases to 1.

Yet another approach which does not introduce extra charges, and also does not require to charge the ‘wrong packets’, consists in charging each packet according to the conditional expectation of its s.p.s.p., given the state of the queue found upon arrival. In the case of the M/D/1 queue with constant service times σ , the resulting expected s.p.s.p. $p(w)$ of a virtual packet entering at time zero the queue and finding a workload of w is seen to be

$$p(w) = w + \sigma + \sigma \mathbf{E}(N(0, T(w))) + K,$$

where $T(w)$ is the time at which the busy period started at 0 with workload w ends, $N(0, T(w))$ is the corresponding number of packet arrivals and the constant K does not depend on w . Indeed, $w + \sigma$ is the sojourn time of the packet arriving at time 0, all packets arriving in the interval $(0, T(w))$ are delayed by an amount σ , while the workload at time $T(w)$ is exactly σ hence the impact on later arriving packets does not depend on w . Applying the integration formula (8.3.3), p. 49 in [10] yields the expression $w\rho/(1-\rho)$ for the middle-term in the right-hand side of this expression. The constant K is then determined by

using the expression $\partial_x(x\mathbf{E}(W + \sigma))$ for the average s.p.s.p., yielding

$$p(w) = \frac{w}{1-\rho} + \sigma + \frac{x\sigma^2}{2(1-\rho)}. \quad (7)$$

We note that the constant term in the above is exactly the expectation of the stationary workload. This suggests the simpler marking scheme, according to which a packet finding an amount of w is charged $w[1 + 1/(1-\rho)]$. This collects the correct amount on average, and under-charges slightly packets finding a near-empty system. Note that the expected length of a busy period is given by $\sigma/(1-\rho)$, so that the term $1/(1-\rho)$ in this scheme can thus be estimated as \hat{B}/σ , where \hat{B} is a sample mean of the observed busy periods.

B. Sample Path Shadow Prices for Two Classes of Service

1) Delay-Based Shadow Price with Priority Scheduling:

We assume a preemptive priority discipline. Low priority packets thus have no impact on high priority ones, and we charge each low priority packet by the number of low priority packets served since the start of the current busy period.

The charge for a high priority packet has two components. The first is the cost imposed on subsequent high priority packets within the same high priority busy period. For this as in the single class case, charge is the number of high priority packets in the same high priority busy period served prior to itself. The second component is the cost imposed on low priority packets, and reversal cannot be applied as we don’t want low priority packets to pay for the harm done to them by high priority packets. The high priority packet delays all those low priority packets currently in the queue plus those that will arrive before the end of the busy period. As an approximation, we charge the length of low priority queue at the time the high priority packet enters service, multiplied by a factor $1 \leq \alpha \leq 2$.²

As in the single class case, alternative schemes which do not introduce over-charging can be designed, based on formulas available for mean delays in M/G/1 priority queues.

2) *Loss-Based Cost with Priority Scheduling*: Gibbens and Kelly [2] consider a queue in which the packet dropping policy is governed by two parameters, B_1 and B_2 so as to enforce the following constraints:

$$q_1 \leq B_1 \quad (8)$$

$$q_1 + q_2 \leq B_2, \quad (9)$$

where q_1 and q_2 are the buffer occupancies of high priority and low priority packets, respectively.

Gibbens and Kelly propose treating such a queue as a pair of virtual single class resources. The first virtual resource is a single class queue with buffer capacity B_2 used by all packets, while the second is a single class queue with buffer capacity B_1 used only by high priority packets. Sample path shadow prices are computed independently for each virtual resource using the single-class loss-based marking scheme [8] discussed above. Following the first loss due to constraint (8), the queue

²In simulations, we observe that α has very little effect on the performance measures of interest. For the experimental results presented later, $\alpha = 1$.

marks all high priority packets until the high priority queue becomes idle, whereas following the first loss due to constraint (9), the queue marks all packets until the end of the busy period. Observe that a high priority packet has two chances to be marked, but carries only a single marking bit.

As we show in [1] this scheme, modulo time reversal, computes the correct shadow price for loss.

3) Loss-Based cost with Earliest Deadline First Scheduling:

Under EDF, a two class system is implemented by granting high priority arrivals a short deadline d_1 and low priority arrivals a deadline $d_2 \gg d_1$. We assume preemptive scheduling to simplify the analysis.

The contribution of earlier arriving packets to the loss of a high priority packet n at time T_n can be understood in terms of the process $f(s)$, defined as the value of the earliest deadline present in the system at time s .³ The arrival of n occurs in what we will call a *local busy period* starting at time s^* . High priority packets arriving before the start of this local busy period have no effect on the loss of n . A low priority packet may contribute to the loss of n if it has arrived within the local busy period, but prior to some critical time t such that its deadline at T_n is closer than that of the arriving packet.

In [1], we characterize the values of s^* and t precisely and show how they can be estimated as the busy period evolves. Appealing to the time reversal argument, we propose a causal marking scheme that, following a high priority loss, marks packets for the same duration as intervals $[s^*, t]$ and $[s^*, T_n]$ for low and high priority packets respectively.⁴

V. SIMULATION RESULTS

We ran simulations to compare the performance of the three proposed mechanisms with that of a single-class FIFO scheduler. The objectives of these experiments are twofold. First, we would like to observe the predicted efficiency gain due to the introduction of a low delay service class. Second, we would like to determine whether our marking schemes provide the correct congestion feedback to users of both classes while treating the low priority class fairly.

We performed the experiment proposed by Gibbens and Kelly [2]. The queue operates in discrete time, serves one packet per time slot, and is shared by a population of N users. In each simulation run, each user is designated as either high or low priority and generates packets of the appropriate type. Users send packets in accordance with the *elastic-user* algorithm of [8], where user i alters the sending rate x_i to try and match the rate at which marks are received with weight $w_i = iw_0$ for some common scale parameter w_0 . In the single-class case with a population of elastic-users sharing a single link, user i will achieve throughput proportional to its relative weight $w_i / \sum_{j=1}^N w_j$. Associated with the high and low service classes, respectively, are delay bounds d_1 and d_2 , with $d_1 < d_2$. Packets served past their delay bounds are discarded by the users and thus considered lost.

³If n is low priority, all previous packets in the current busy period contribute to its loss.

⁴Following a low priority loss, packets are marked for the duration of the busy period.

For the initial run, all users are designated as low priority and the system simulated for enough time steps to observe convergence of the sending rates. A low priority user is then selected at random to promote to the high priority class and the simulation is run again. Users are repeatedly promoted and the system simulated until all users generate high priority packets.

A. Schedulable Regions

Our first result is an experimental validation of the analytical results of Section II. In the experiment we record the *goodput*—the throughput of packets that respect delay bounds $d_1 = 5$ and $d_2 = 50$ —for each user. Plotting aggregated goodput for high priority traffic against that for low priority traffic, as in Fig. 2, defines the boundary of the schedulable region. Comparing the resulting plots for EDF and loss-based priority schedulers with those of a single class FIFO scheduler, we observe the qualitative behaviour predicted in Section II.⁵

B. Class performance: Incentive compatibility and fairness

Performance of the system is evaluated by the average waiting time for each class and the average transmission rate (normalized by willingness-to-pay) for each class. By plotting these performance metrics for each class against the proportion of high priority demand, we can get some idea of whether the pricing mechanisms we propose provide the correct incentives to users of each class and whether members of the low priority class are treated fairly. We make the assumption, similar to Alvarez and Hajek [4] and Hurley et al. [11], that users of the low priority class are delay tolerant, but seek to maximize throughput whereas high priority users seek low delay, perhaps at the expense of lower throughput. Our criteria for incentive compatibility and fairness as the proportion of high priority demand increases are that (a) low priority users get throughput at least as good as when no high priority traffic is present and thus have no incentive to switch classes, (b) high priority users experience lower delay than low priority users, and (c) the delay for low priority users is finite.

Fig. 3 shows the normalized throughputs for high and low priority classes.⁶ These plots show that all three disciplines satisfy the throughput fairness criterion. Under heavy high priority demand, the low priority class can actually receive substantially better throughput than in the single class case.⁷ We also observe that, for moderate amounts of high priority demand (< 0.8) the high priority class achieves a throughput comparable to the low priority class, although this was not an explicit design goal.

We also observe that under all three disciplines, high priority class traffic sees a substantially lower delay than low priority traffic. Furthermore, the average delay seen by the low priority traffic is well below its delay bound of 50ms, a property guaranteed only by the EDF scheduler, but that appears to be satisfied

⁵Note that we do not provide a plot for the delay-based pricing mechanism in this section since this mechanism cannot be tuned to provide deterministic delay bounds for either class with the elastic-user strategy.

⁶Observe that while the sum of throughputs clearly must be limited by the total capacity of the link, no such restriction applies to the sum of *normalized* throughputs shown here.

⁷The single class case is shown by the endpoints of each plot—the throughput of the low (resp. high) priority class when the proportion of high (resp. low) priority traffic is zero.

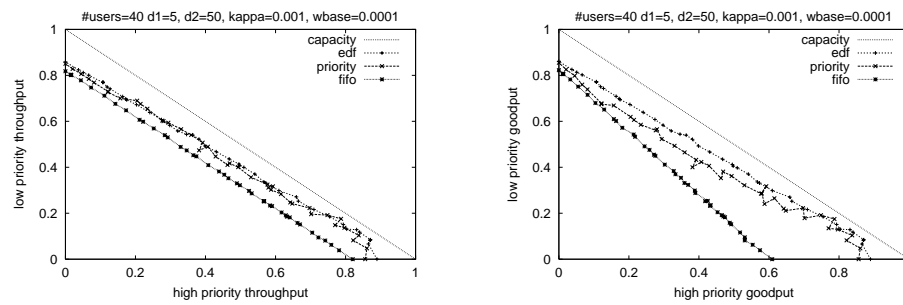


Fig. 2. Aggregate throughput (left) and goodput (right) for high vs. low priority traffic. The topmost diagonal line represents the available capacity in the system.

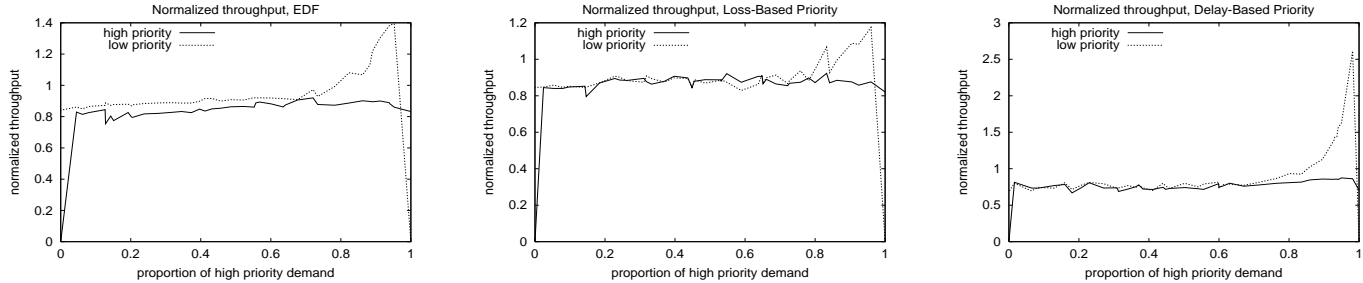


Fig. 3. Average normalized throughput for each class, plotted for all three service disciplines.

by the loss-based priority scheduler. These results are presented and discussed further in [1].

VI. CONCLUSIONS

When users are sensitive to delay and there is but a single class of service and FIFO scheduling, delay requirements for such users can only be met if the network operates in lightly loaded region. The introduction of scheduling allows much more efficient use of the network. We have quantified the benefits analytically for unresponsive traffic and by simulation for responsive users.

For responsive users, we have shown how to calculate the correct ‘shadow-price’ or feedback signals that reflect congestion costs, allowing decentralized optimization of social welfare. We have derived sample-path shadow prices for loss-based and delay-based priority queues, and loss-based prices for EDF and used these as a basis for practical marking schemes. These were then used with a simple user-adaptation algorithm in simulations.

Our results show that such marking schemes are effective for controlling delay to sensitive users, with tangible gains over a single-class FIFO, and also provide a way of implementing differentiation for proposals such as ABE [11]. In the simulations, delay-based marking for priority ran at a low utilisation, suggesting some further tuning is required. EDF performed slightly better than priority with loss-based marking, but the latter is simpler to implement. This resonates with the analytic heavy-traffic results which also showed that little was gained by using EDF rather than priorities for unresponsive traffic.

The difficult question remains of where, if anywhere, such service differentiation should, in fact, be deployed in the network. This decision ought to depend on whether actual bursti-

ness of future network traffic makes the over-provisioning of capacity economically unattractive to network operators. The availability of efficient and controllable alternatives to over-provisioning, such as those shown here, allows this choice to be grounded in economic concerns, rather than technological ones.

REFERENCES

- [1] P. Key, L. Massoulié, and J. Shapiro, “Service differentiation for delay-sensitive applications: An optimisation-based approach,” Tech. Rep., Microsoft Research, 2001, MSR-TR 2001-115.
- [2] R. J. Gibbens and F. P. Kelly, “Resource pricing and the evolution of congestion control,” *Automatica*, 1999.
- [3] S. Bajaj, L. Breslau, and S. Shenker, “Is service priority useful in networks?,” in *Proceedings of ACM Sigmetrics ’98*, June 1998, pp. 66–77.
- [4] Juan Alvarez and Bruce Hajek, “On using marks for pricing in multiclass packet networks to provide multidimensional QoS,” Submitted, 2001.
- [5] M. Reiman, “Open queueing networks in heavy traffic,” *Mathematics of Operations Research*, vol. 9, no. 3, pp. 441–458, 1984.
- [6] B. Doytchinov, J. Lehoczy, and S. Shreve, “Real-time queues in heavy traffic with earliest-deadline-first queue discipline,” *To appear in Annals of Applied Probability*, 2000.
- [7] F. Kelly, A. Maulloo, and D. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research Society*, vol. 49, 1998.
- [8] R. J. Gibbens and F. P. Kelly, “A note on packet marking at priority queues,” to appear in *IEEE Transactions on Automatic Control*, 2000.
- [9] Felisa J. Vazquez-Abad, *Discrete Event Systems, Analysis and Control* (R. Boel and G. Stremersch, ed.), chapter A Course on Sensitivity Analysis for Gradient Estimation of DES Performance Measures, Kluwer Academic Publishers, 2000.
- [10] F. Baccelli and P. Brémaud, *Elements of Queueing Theory*, Springer Verlag, 1994.
- [11] J.-Y. Le Boudec, P. Hurley, M. Kara and P. Thiran, “ABE: Providing a low-delay service within best-effort,” *IEEE Network, Special issue on Control of Best-Effort Traffic*, vol. 15, no. 3, 2001.