

Charging the Internet Without Bandwidth Reservation: An Overview and Bibliography of Mathematical Approaches

BRUNO TUFFIN

IRISA-INRIA

Campus Universitaire de Beaulieu

35042 Rennes Cedex, France

E-mail: btuffin@irisa.fr

Pricing is one of the biggest challenges facing the next generation of the Internet. Even if flat rate pricing is one of the main reasons for the success of the Internet, the only way to prevent network congestion and to differentiate services is to adopt usage-based pricing schemes. We review in this paper, from a mathematical modeling point of view, the pricing schemes *without resource reservation* that have been developed in the literature. Indeed, an advantage of the absence of reservation in the Internet is that network management is cheap. Even if accounting and billing will increase this cost, we believe that pricing without resource reservation is the lesser of two evils when costly bandwidth reservation procedures are applied.

Keywords: fairness, internet economics, optimization, pricing, service differentiation

1. INTRODUCTION

The Internet is experiencing tremendous traffic growth. A consequence is that real users complain that large data transfers take too long and that they have no way to improve this situation by themselves (by paying more, for instance). To cope with this congestion, it is possible to develop link capacities, but many authors consider that this is not a viable solution as the network must respond to increasing demand (and experience has shown that the demand for bandwidth has always been ahead of the supply), especially now that the Internet is becoming a commercial network. Furthermore, incentives to achieve fair utilization between customers is not included in the current Internet (see, for instance, [1, 2]). For these reasons, it is suggested that the current flat rate fees, where customers pay a subscription fee and obtain unlimited usage, be replaced by usage-based fees [3]. Also, the future Internet will supply different kinds of services, such as video, voice, email, ftp, telnet, and html among others. Each of these applications requires a different quality of service (QoS): for example, video needs very small delays and packet losses, voice requires small delays but can afford some cell losses, email can afford delay (within a given bound), while ftp needs a good average throughput, and telnet benefits more from short round trip times. Some pricing incentives should exist so that each user does not always need to choose the best QoS for his application, and so that the final result is fair utilization of bandwidth. On another hand, we need to be aware of the

Received February 5, 2002; accepted March 6, 2003.

Communicated by Chu-Sing Yang.

trade-off between engineering efficiency and economic efficiency; indeed, measurement, for example, helps improve the management of the network but is costly.

In [4], J. Roberts classifies pricing schemes in three categories, flat rate pricing, congestion pricing, and transaction pricing, and studies their impact on QoS (see also other introductory or overview papers [3, 5-8] and [9], where an interesting time-scale methodology and classification is presented). Another classification approach separates schemes into edge pricing schemes, where the charge is set only at the edge of the network, and node-per-node pricing schemes. Our paper differs from the previous ones in that mathematical models are displayed (when available). We classify the suggestions for future Internet pricing in eight different families as follows, most of them being sub-categories of congestion pricing in [4].

1. As already explained, a first group of people (see, for instance, [10, 11]) are those arguing that even if the number of customers (and their demands) is growing quickly, the network capacity is also adapting itself to the demand. Furthermore, if the system has survived so far and has known such success, why should we introduce a costly billing model?
2. For a second group of people, incentive pricing will be necessary to regulate these various levels of quality of service, and some services must be *guaranteed*. As in ATM networks, charging models for guaranteed services, such as voice or video, should be related to connection acceptance control (CAC) [12-16], resource reservation and effective bandwidth theory [17]. On the Internet, resource reservation may be done using RSVP [18]. In [19] and [20], reservation is used and only non-guaranteed services are accomplished using best effort techniques adapted to the user's willingness to pay. In [21], CAC and bandwidth reservation are applied to loss networks; a nice characteristic of this approach is that arrival rates for each class of service depend on the connection fee of the class. A dynamic programming method is used to obtain optimal and quasi-optimal prices, and it is shown that time-of-day pricing efficiently approximates congestion pricing. These results are extended in [22] to general loss networks (non-exponential holding times), and to the case where the system has prior knowledge of connection times, at their arrival. In [23], the pricing of elastic traffic flows is related to routing.
3. Another alternative has been suggested by A. Odlyzko in [24]. The proposal is called Paris Metro Pricing (PMP) since it is analogous to the Paris Metro System. The network is decomposed into several separate networks, and each network, working like the current Internet, has a different connection fee so that we expect that the most expensive networks will not be less congested. Thus, no QoS is guaranteed, but the model can be easily implemented without huge overhead.
4. The Cumulus Pricing Scheme (CPS) [25, 26] is also a simple possibility. A contract is negotiated between the ISP and the user. During periods of time, the utilisation is measured, and (positive or negative) cumulus points are awarded, depending on whether the contract is satisfied or not. At a given time, extra-fees can be charged.
5. Another group suggests using priority pricing, without reservation of resources (see [27-32] and the references therein). Each class is assigned a priority number and is served according to this policy at each node of the network. Priority pricing schemes are divided into two sub-classes:

- (a) The first one is posted priority pricing, where each priority class price is established in advance. In [27], each customer is assigned a quota for high priority packets (following his contract), and if his quota is exceeded, he is charged a penalty the following month. In [28], a priority flag is assigned to each packet according to the type of service, but so is a reject flag for services which can bear some losses. In [33-35], a discrete time model is described, where the time is divided into time slots. Optimal prices are computed in order to maximize the network benefits.
 - (b) The second sub-class is non-posted priority pricing, where the price of the packet class depends on the traffic level. In [30], an adaptive priority pricing scheme that depends on the context (similar to the principles given in [36]) is used. In [32, 37], an optimal incentive-compatible pricing scheme for the M/M/1 multi-class queue is studied (note that the result can be easily extended to the M/G/1 queue).
6. Bidding for priority has also been proposed in [38, 39]. The user makes a bid for each packet, and only bids greater than some cutoff values are admitted. In [40-42], auctions for packets are replaced by auctions for bandwidth during specific intervals of time to reduce the management overhead. Efficiency, stability and fairness issues are solved not only in the case of one node, but also in the case of interconnected networks.
 7. Another scheme is the *expected capacity* theory developed by Clark [43], where packets are flagged *in* or *out* and are served without priority except in the case of congestion, where *out* packets receive a congestion pushback.
 8. A last group of pricing schemes employs charging for elastic traffic based on transfer rates (see also [44]). In Kelly et al.'s work [1, 45], the user decides on his payment and receives as the transmission rate what the network allocates to him. In Low et al.'s work [2, 46-49], the user decides his own rate and pay for it according to the price computed by the network. A variant of Kelly et al.'s work has been given by La and Anantharam in [50, 51], where the flow rates are actually controlled by the window-based algorithm of TCP connections.

A vast literature has focused in recent years on the future Internet and on the integration of different services [24, 27-30, 43, 52-55] as well as on the fairness issue (see, for instance, [56-60]). Which of the different charging groups will be implemented and how prominent they will be in practice can only be guessed. As stated in [30], we believe that the arguments in favor of simply overprovisioning the capacity of the Internet is dangerous in the current situation. Moreover, capacity reservation for some types of services is expensive to implement. We are, therefore, betting that the next pricing scheme will be pricing without bandwidth reservation.

The aim of this paper is to review current works (and when possible, the mathematical models) on pricing without bandwidth reservation theory (from 3 to 8 in the previous classification).

2. PARIS METRO PRICING [24]

The proposal in [24] is to partition a network into several logically separate networks (or classes), each having a fixed fraction of the capacity of the entire network. All

networks would route packets according to the current TCP and UDP. There is no formal guarantee of QoS, but by charging different rates for different classes (served in the same way), it is supposed that the most expensive classes will be less congested as a result of self-regulation and will then deliver better QoS. The name given to this model, *Paris Metro Pricing* (PMP), is based on the Paris Metro of about 20 years ago, where two classes of cars existed in trains, but with exactly the same quality of seats. As tickets prices were different, the cars for the most expensive class were less congested leading to better perceived QoS.

The advantage of PMP pointed out by Odlyzko is that, even if no QoS is guaranteed, it would permit dispensing with measures, such as RSVP and their complexity, and would retain the simpler and cheaper current model of the Internet.

It is suggested that only a few (3 or 4) subnetworks would be implemented to minimize losses due to not aggregating all the traffic. PMP charges would be assessed on each packet, and would probably consist of a fixed charge per packet and a fee depending on the size of the packet.

Recently, in [61], Gibbens et al. studied PMP in the case of two Internet service providers (ISP) competing to maximize their profits. In their paper, a user joins the network i which maximizes his utility $U(\theta, i) = V - \theta Q^i/C^i - p^i$, where V is the positive valuation of the user, θ is his preference for lack of congestion (θ is assumed to follow a uniform distribution in $[0, 1]$), Q^i/C^i is the mass of users divided by the capacity (at network i), i.e., the measure of congestion, and p^i is the price per unit time charged by network i . Network i then tries to maximize its benefit $p^i Q^i$. It is shown that, at the stable point, the ISP will not provide multiple services. By then, they state that PMP may not survive under competition (at least if the system follows the given assumptions).

3. THE CUMULUS PRICING SCHEME (CPS) [25, 26]

Like PMP, CPS is interesting due to its implementation simplicity. In this scheme, the user negotiates with the ISP a given level of utilization or a given QoS during a period of time. Say, for instance, that the contract is based on the number of packets sent. If this number is $V(t)$ at time t and is measured between period $[t_{i-1}, t_i]$, the resulting over or under-utilization is

$$\Delta_i = \int_{t_{i-1}}^{t_i} V(t)dt - x(t_i - t_{i-1})$$

with respect to the expected mean use x per unit of time. Define the thresholds θ_n ($n = -N, \dots, N$) such that $\theta_i < \theta_j$ if $i < j$, and let $\theta_0 = 0$. Also, let $\theta_{(N+1)} = -\infty$, and let $\theta_{(N+1)} = \infty$. c_i cumulus points (positive or negative) are assigned by the ISP to the user during period $[t_{i-1}, t_i]$ if $\theta_{c_i} \leq \Delta_i < \theta_{c_i+1}$.

Let $\Lambda_n = \sum_{i=1}^n c_i$ be the sum of the cumulus points assigned to the user during $[0, t_n]$. The ISP reacts and renegotiates the contract if $|\Lambda_n| \geq \Theta$.

The tariff function $p(x)$ per unit at service level x has to be determined (the total charge is $c(x) = xp(x)$). For convenience, $p(x)$ will also be used for extra-fees: if the *observed* service level is x_1 , the penalty charge is

$$\Psi(x, x_1) = c(x_1) - (c(x) + c(x_1 - x)).$$

The following requirements are inserted in order to obtain a fair scheme:

1. $p(x) > 0$ is monotonically decreasing; $c(x)$ is monotonically increasing.
2. $\Psi(x, x_1) < 0$ if $x \neq x_1$, and $\Psi(x, x_1) = 0$ if $x = x_1$, so that, due to the penalty charge, the user has an incentive to indicate his true service level requirement. $\Psi(x, x + \delta)$ is decreasing in δ .
3. $|\Psi(x, x_1)| < |\Psi(\beta x, \beta x_1)| \leq \beta |\Psi(x, x_1)|$ for $\beta > 1$, meaning that the penalty is higher for high bandwidths, but smaller proportionally for the expected ones.

For instance, $p(x) = C/\sqrt{x}$ fulfills these requirements. It is suggested that no more than 3-5 thresholds be used. Moreover, the number of assigned cumulus points should be “independent” of the measurement technique for determining x_1 . Assuming that the stochastic process $V(t)$ is in equilibrium, and performing N independent measurements during each interval $[t_{i-1}, t_i]$, a confidence interval of $E(V)$ can be obtained using Student distribution, at confidence level $1 - \alpha$, by a standard Monte Carlo method. Let $\varepsilon_{\alpha, N}$ be the half width of the interval. Then taking $\theta_{i+1} - \theta_i > 2\varepsilon_{\alpha, N}$ will ensure that, with probability at least $1 - \alpha$, the number of assigned cumulus points is not sensitive to the measurement technique.

4. POSTED PRIORITY PRICING

4.1 Work by Bohn et al. [27]

In their work, Bohn et al. use the 3-bit precedence field in the protocol header to introduce priorities (from 0, the lowest, to 7, the highest) in the traffic as was imagined (but not publicized) in the mid-80s when the NSFNET backbone was highly congested. This scheme was proposed in [27] as an interim solution before the Internet was redesigned to incorporate protocols with bandwidth reservation, but it is worth studying.

Internet Service Providers negotiate with users some soft quotas for the total volume of traffic by specific IP Precedence levels: a quota system is introduced to discourage users from setting high precedence values throughout their traffic. Another solution is to buy a total quota which is a weighted sum of the priority values in its packets per unit of time. They suggest the formula

$$Q = \sum_{i=2}^6 x_i \alpha^{i-2},$$

where Q is the total quota used by the customer, x_i is the number of packets sent with priority i during the metered period, and α is a parameter greater than 1 (they propose $\alpha = 2$). Priority levels 0 and 1 are not considered in the formula because they are free, and priority level 7 is reserved for network management.

This scheme is not directly related to pricing, but a pricing scheme can be devised by the ISP. It can also be seen as a charging scheme somewhere between the previous CPS and the next posted priority pricing.

4.2 Work by Cocchi et al. [28, 29]

This work also used the 3-bit precedence field in the protocol header to introduce priorities. The model is as follows. Let s_i denote a characterization of the network service received by the i th user ($1 \leq i \leq n$), and let $V_i(s_i)$ denote the i th user's level of satisfaction, expressed in money, with a given network service s_i (we will give some examples later). If the user is charged an amount c_i for that service, the overall level of satisfaction is $U_i = V_i(s_i) - c_i$. Each user sends a request σ_i (not necessarily involving a call set-up). Let $\underline{\sigma} = (\sigma_1, \dots, \sigma_n)$, and let $s_i(\underline{\sigma})$ be the resulting network service. Define

$$\underline{\sigma}^{max} = \operatorname{argmax}_{\underline{\sigma}} \sum_{i=1}^n V_i(s_i(\underline{\sigma})) \quad \text{and} \quad V_{max} = \sum_{i=1}^n V_i(s_i(\underline{\sigma}^{max}))$$

as, respectively, the vector maximizing the total satisfaction and the maximum total satisfaction. As each user is acting selfishly, i.e., is trying to maximize his own satisfaction $U_i(s_i(\underline{\sigma})) = V_i(s_i(\underline{\sigma})) - c_i(\underline{\sigma})$, the system needs to be in *Nash equilibrium*. Formally, $\underline{\sigma}$ is a Nash equilibrium if for all i and all $\tilde{\sigma}_i$, $U_i(\underline{\sigma}) \geq U_i(\underline{\sigma} | \tilde{\sigma}_i)$, where $(\underline{\sigma} | \tilde{\sigma}_i)$ is the vector where the i th coordinate of $\underline{\sigma}$ is replaced by $\tilde{\sigma}_i$. This means that user i can not alone increase his level of satisfaction. A pricing scheme is then said to be *acceptable* if $\underline{\sigma}^{max}$ is the unique Nash equilibrium scheme. It can be easily seen that without a pricing scheme, i.e., $c_i(\underline{\sigma}) = 0$, the Nash equilibrium is unlikely to be achieved.

The scheme is then illustrated by means of examples. In [29], a simple two-class model is simulated on two different network topologies. The two different classes have different service priorities at each switch (or node) of the network. Per-byte pricing is used with a higher price for the highest priority. The applications considered are e-mail, FTP, Telnet and Voice. The different functions, V_i , following the required QoS, are

$$\begin{aligned} V_{\text{email}} &= -0.1 \text{ (avg. message delay (sec))} \\ &\quad \text{-(\% of messages not delivered in loose delay of 5 minutes),} \\ V_{\text{FTP}} &= 100 \text{ (average normalized throughput),} \\ V_{\text{Telnet}} &= \text{-(avg. packet round trip time (ms))/10,} \\ V_{\text{Voice}} &= \text{-(\% of packets not obeying the tight delay of 100ms)-d/100,} \end{aligned}$$

where d is the average one-way delay of voice packets (in ms). The requests σ_i are merely the priority settings on the packets. In the implementation, each particular application is assumed to use the same priority settings. The range of acceptable prices is given according to the topology of the network, but some exist for a wide range of network conditions.

In [28], the same kind of example is used, but in addition to the two service priorities, there is a blocking priority, resulting in 4 different classes. This situation is interesting, for some applications require small delays but can afford losses or, conversely, require no or very few losses but can afford delays. We then have four prices per byte p_{ij} , $0 \leq i, j \leq 1$, where the first bit i means that the service priority flag is on or off and j gives the status of the no-drop flag.

4.3 Work by Honig and Steiglitz [31]

In this model, K users are assumed to compete for a resource (possibly at the gateway of a network or directly at a switch, for instance) for the same type of traffic, meaning the same type of QoS. User k wishes to send packets at rate λ_k , so that the total rate is $\Lambda = \sum_{\text{active } k} \lambda_k$. The QoS perception is given by a function $D(\Lambda)$. In [31], the delay represents the QoS, but other measures can be considered. A utility function $u_k(\delta)$ is associated with user k , depending on the observed QoS δ . If the price per packet is P , user k transmits his packets if and only if $u_k(\delta) \geq P$. In equilibrium, the QoS announced by the network must be what the user observes; that is, the following fixed-point equation must be satisfied:

$$D\left(\sum_{k: u_k(\delta) \geq P} \lambda_k\right) = \delta.$$

Under some assumptions (u_k monotonically decreasing and with limit 0 at $+\infty$ and D strictly positive, finite, continuous, and monotonically increasing), it can be proved that there is a unique equilibrium for each price P . The idea is then to choose the price P that maximizes the revenue $R = P\Lambda$. Some examples are provided.

As extensions, multiple priorities and time of day pricing are discussed.

4.4 Work by Marbach [33-35]

This work is devoted to DiffServ, where packet classes are served according to a given priority. Prices per sent-packet are static. Indeed, it is argued that, by charging for all *submitted* packets, users have an incentive to reduce their rates during periods of congestion, as they pay for lost packets.

The mathematical model considers a single link and is as follows. Time is discretized, divided into slots. During each slot, the link has the ability to serve C packets. It is assumed that packets not served when the slot is lost. There are N different (and ordered) priority classes, where 1 is the lowest priority. R users are supposed to compete for link access. Let u_i be the price charged for a class- i packet submitted for access (of course, $u_i < u_j$ if $i < j$), and let $d_r(i)$ be the number of class- i packets that user r submits in a given time slot. User r 's whole allocation is given by the vector $d_r = (d_r(1), \dots, d_r(N))$. The number of submitted class- i packets is $d(i) = \sum_{r=1}^R d_r(i)$, and $d = (d(1), \dots, d(N))$ is the aggregated allocation.

Let i^* be the priority class such that $\sum_{i=i^*+1}^N d(i) < C$ and $\sum_{i=i^*}^N d(i) \geq C$. Packets with priority $i > i^*$ are served (say, with probability $P_{tr}(i, d) = 1$), those with priority $i < i^*$ are lost (say, with probability $P_{tr}(i, d) = 0$) and those of class i^* are served with probability

$$P_{tr}(i^*, d) = \frac{(C - \sum_{i=i^*+1}^N d(i))}{d(i^*)}.$$

User r 's throughput is then

$$x_r = \sum_{i=1}^N d_r(i) P_r(i, d).$$

A utility function $U_r(x_r)$ is associated with user r . U_r is assumed to be increasing, bounded, strictly concave and twice differentiable. Users are assumed to play a non-cooperative game, where user r chooses allocation d_r^* such that

$$d_r^* = \operatorname{argmax}_{d_r} \left(U_r(x_r) - \sum_{i=1}^N d_r(i) u_i \right).$$

In equilibrium, this happens for all users. If we suppose without loss of generality that the total demand at price u_1 , $D(u_1)$, exceeds C , and that $D(u_i) > 0 \forall i$, then

- there exists an equilibrium. If there is a class i_0 such that $D(u_{i_0}) > C > D(u_{i_0+1})$, then the equilibrium is unique.
- $d_r^*(i) = 0 \forall i \notin \{i_0, i_0 + 1\}$; $P_r(i_0, d^*) \geq u_{i_0}/u_{i_0+1}$, where $u_{N+1} = \max_r U_r'(0)$; and $x_r^* = D_r(u^*)$ with $u^* = u_{i_0}/P_r(i_0, d^*)$.

In [35], the game is played dynamically. A gradient algorithm is used to prevent oscillations. In [33], the model is extended to bursty traffic.

5. NON-POSTED PRIORITY PRICING

5.1 Works by Mendelson and Whang [32], and by Ha [37]

The works described here were not dedicated to Internet management. However, even if some points are not related to our concern, they are worth studying.

Mendelson and Whang consider a pricing scheme for a multi-class M/M/1 queue (which can be easily extended to a M/G/1 queue if all job classes have the same coefficient of variation). Arrivals of class- i jobs ($1 \leq i \leq R$) to the system reflect the aggregation of infinitesimal users' job flow. The arrival rate is λ_i . The value function of class- i jobs $V_i(\lambda_i)$, representing the gross value gained by class- i users per unit of time, is assumed to be differentiable, nondecreasing and concave on λ_i . λ_i and the "full price" z are related in the following way: $\lambda_i = D_i(z) = (1 - F_i(z))\Lambda_i$, where Λ_i is the maximum potential arrival rate of class i and $F_i(\cdot)$ is the distribution function of the service valuation. Inverting this function, we have $V_i'(\lambda_i) = D_i^{-1}(\lambda_i)$. Let $\underline{\lambda} = (\lambda_1, \dots, \lambda_R)$. The total expected value function is

$$V(\underline{\lambda}) = \sum_{i=1}^R V_i(\lambda_i).$$

Each class- i job is characterized by a delay cost of v_i per unit of time. Class- i jobs are assumed to be served following an exponential distribution with mean c_i , and the priority policy of the server is assumed to be non-preemptive. It is also assumed, without loss of generality, that the classes are ordered from highest to lowest priority so that the expected average delay cost per unit time is minimized, i.e.,

$$\frac{v_1}{c_1} \geq \frac{v_2}{c_2} \geq \dots \geq \frac{v_R}{c_R}.$$

The idea is to maximize the expected net value of the jobs processed by the system, i.e. to find

$$\max_{\underline{\lambda}} \left\{ V(\underline{\lambda}) - \sum_{i=1}^R v_i L_i(\underline{\lambda}) \right\}, \quad (1)$$

where L_i is the mean number of class- i jobs in the system in steady-state. The administrator sets the price vector

$$\underline{p} = (p_1, \dots, p_R),$$

where p_i is the price charged to a class- i job. If the class- i demand relationship (which is set such that, at equilibrium, the marginal value will be the same for joining and or not joining the system) is $V_i'(\lambda) = p_i + v_i W_i(\lambda)$, then it is proved in [32] that the optimal price per class- i job is given by

$$p_i^* = \sum_{j=1}^R u_j \lambda_j^* \frac{D W_j(\underline{\lambda}^*)}{D \lambda_i},$$

where W_j is the expected delay of a class- j job and $\underline{\lambda}^*$ maximizes (1).

In the homogeneous case, i.e., $c_1 = \dots = c_R$, we have $L_i(\underline{\lambda}) = \frac{\lambda_i S_R}{S_{i-1} S_i} + \lambda_i$ and $W_i(\underline{\lambda}) = \frac{S_R}{S_{i-1} S_i} + 1$, where $S_0 = 0$, $S_i = \sum_{j=1}^i \lambda_j$ and $\bar{S}_i = 1 - S_i$. We can then explicitly get the optimal prices:

$$p_i^* = \sum_{k=1}^R \frac{\lambda_k^* v_k}{\bar{S}_k \bar{S}_{k-1}} + \sum_{k=1}^R \frac{\lambda_k^* v_k W_k^q + \lambda_{k+1}^* v_{k+1} W_{k+1}^q}{\bar{S}_k},$$

where $W_k^q = W_k - 1$ is the expected waiting time of a class- k job in the queue and $\lambda_{R+1}^* = v_{R+1} = W_{R+1}^q = 0$.

The problem here is that the prices are determined on a centralized basis, which is practically irrelevant. More specifically, both the users and system administrator know $(v_i, V_i, c_i) \forall_i$, but only the users know their *real* class membership. To cope with this problem, Mendelson and Whang consider priority-dependent pricing schemes. The idea is to obtain a *Nash equilibrium*, i.e., a situation where no user, by unilaterally changing his own request, can increase his own net value. This property is decomposed in *incentive-compatibility*, which means that it is in all users' interest to classify their jobs according to their correct priority classes, and *optimality*, which means that the resulting arrival rates maximize the expected net value of the system as a whole. Optimality and incentive-compatibility are obtained when the optimal prices are used in the homogeneous case and when a class- i user decides not to enter the system if

$$\min_{i \leq j \leq R} \{0, p_j + v_i W_j(\underline{\lambda}(\underline{p})) - V_i'(\lambda_i)\} = 0$$

and to join the system otherwise.

Unfortunately, incentive-compatibility is not valid in the heterogeneous case. The previous posted charging mechanism should take into account additional information, such as the actual processing time of the job. We then have a priority and time-dependent pricing scheme. If we have

$$p_i(t) = A_i t + (1/2) B t^2$$

with

$$B = \sum_{k=1}^R \frac{v_k \lambda_k^*}{\bar{S}_{k-1} \bar{S}_k}$$

and

$$A_i = \frac{a_i}{\bar{S}_{i-1} \bar{S}_i^2} + \sum_{k=i+1}^R a_k \left(\frac{1}{\bar{S}_{k-1}^2 \bar{S}_k} + \frac{1}{\bar{S}_{k-1} \bar{S}_k^2} \right),$$

where $a_i = v_i \lambda_i^* \sum_{k=1}^i c_k^2 \lambda_k^*$, then the pricing scheme is optimal and incentive-compatible. Note that it consists of a basic charge (corresponding to the lowest priority charge) and a priority surcharge (proportional to the processing time).

In [37], A.Y. Ha extends the previous work to the case where service requirements are controllable by customers. Then, each customer decides whether to request service from the facility and, if it is desirable, determines his service requirement. Also investigated is the case of the $M/G/s$ processor sharing queue, for which the optimal prices are found to be two-parts linear in time in the system. The first-come-first served $M/G/1$ queue is also studied, and a quadratic price is also obtained.

5.2 Work by Gupta et al.

In [30, 62], Gupta, Stahl and Whinston also develop a priority pricing scheme. They first argue that the posted priority pricing scheme of Bohn et al. may lack an incentive to provide multiple precedence networks (i.e., the providers may not be appropriately rewarded), and they point out that we must look at the context in which the applications are used, not just categorize them.

In [30], a four priority class model is introduced, where the highest priority is for real-time services with no tolerance of lost packets, the second class is for real-time services that are relatively tolerant to lost packets and the two lowest priority classes for two levels of best effort service (to provide a finer division of delay requirements). In [62], the number of classes is kept general.

The price at a particular server for a particular class is represented by the following system of equations:

$$r_{mk}(q) = \sum_l]D\Omega_l / D_{X_{mkq}}] \sum_i \sum_j \delta_{ij} x_{ijlm}, \quad (2)$$

where

- $r_{mk}(q)$ is the price of a job of size q at server m for priority class k ,
- X_{mkq} is the arrival rate of job of size q at server m in priority class k ,
- Ω_l is a continuously differentiable, strictly increasing function of the arrival rate X_{mkq} and capacity v_m which provides the waiting time at a server m for priority class l ,
- δ_j is the delay cost parameter of consumer i for service j ,
- x_{ijm} is the flow rate of service j for consumer i with priority k at server m .

$[D\Omega_l/DX_{mkq}]$ is the derivative of the waiting time, and $\sum_i \sum_j \delta_j x_{ijm}$ is the accumulated delay cost of the system. This kind of priority pricing prevents the situation where the “highest priority can preempt all the available capacity” (as noted in [43]) in the case of posted priority of previous subsections.

In [62], a general mathematical model is introduced, and it is shown that this choice maximizes a system-wide welfare stochastic allocation function.

The prices are computed using the following iterative equation:

$$r_{mk}^{t+1} = \alpha \hat{r}_{mk}^{t+1} + (1 - \alpha) r_{mk}^t,$$

where

- α is a real number between 0 and 1; the authors suggest taking $\alpha = 0.1$;
- \hat{r}_{mk}^{t+1} is the estimated new price at time $t + 1$ using Eq. (2);
- r_{mk}^t is the implemented price during the time interval $(t, t + 1)$.

Many experiments were performed using a simulation platform.

6. SMART MARKET: AUCTION IN THE NETWORK

6.1 Smart Market of McKie-Mason and Varian [38, 39]

In their paper on the history of the Internet, cost and pricing [38], McKie-Mason and Varian argue that posted priority pricing as described in section 4 is not a good solution. Indeed, if the network is at capacity, some users with high willingness-to-pay may be unable to access the network. Pricing based on the time of day attempts to achieve this goal but does not efficiently allocate the available bandwidth.

McKie-Mason and Varian suggest the use of a “smart market,” which is actually a variation of the Vickrey auction. Each packet is given a bid representing the user’s willingness to pay. The packets are given a priority at each node of the network according to this bid. Using the Vickrey auction, if the network is not congested, the price is zero whereas if there is congestion, the charge is based on the willingness-to-pay of the lowest priority packet admitted.

Unfortunately, the smart market concept is not an ideal solution. As noted in [38], the current TCP/IP version would not support a smart market. Moreover, it requires the use of complicated systems to conduct auctions for individual packets. The model was more an incentive for further research than a solution.

6.2 Progressive Second Price (PSP) Auction

In [40-42, 63, 64], costly auctions for individual packets are replaced with auctions for bandwidth during specific intervals of time. A good analysis of this scheme based on game theory is provided, including fairness properties. As stated in [41], “in market-based approaches, no precise model need be assumed [...], the seller does not require a priori demand information.” The behavior of the system is then essentially real-time, and not model-based.

To briefly explain how this auction works, consider a single resource of capacity Q and I players competing for it. Player i 's bid is $s_i = (q_i, p_i)$, where q_i is the capacity the player i is looking for and p_i is the unit price he is proposing. A bid profile is $s = (s_1, \dots, s_I)$. Let $s_{-i} = (s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_I)$ be the profile where player i 's bid is excluded from the game. For $y \geq 0$, define

$$\underline{Q}_i(y; s_{-i}) = \left[Q - \sum_{p_k \geq y, k \neq i} q_k \right]^+.$$

The progressive second price allocation rule gives to player i a bandwidth of

$$a_i(s) = \min(q_i, \underline{Q}_i(p_i; s_{-i}))$$

and sets the total cost to be

$$c_i(s) = \sum_{j \neq i} p_j [a_j(0; s_{-i}) - a_j(s_i; s_{-i})].$$

Thus, the highest bids are allocated the desired quantity, and the cost is given by the declared willingness to pay (bids) of the users who are excluded by i 's presence.

Assume that player i attempts to maximize his utility $u_i(s) = \theta_i(a_i(s)) - c_i(s)$, where θ_i is the valuation function that player i gives to his allocation. Under some smoothness assumptions on θ_i and with a bid fee ε each time a player submits a bid, it is stated that if for all i player i bids $(v_i, w_i = \theta'_i(v_i))$ with

$$v_i = \left[\sup \left\{ z : z \leq \underline{Q}_i(\theta'_i(v_i), s_{-i}) \text{ and } \sum_{j \neq i} p_j [a_j(0; s_{-i}) - a_j(z; s_{-i})] \leq b_i \right\} - \varepsilon / \theta'_i(0) \right]^+,$$

where $\underline{Q}_i(y; s_{-i}) = [Q - \sum_{p_k \geq y, k \neq i} q_k]^+$ and b_i is the budget constraint, then convergence, efficiency and fairness problems are solved (the property in [64] of the equal-bid case (when the total required bandwidth at this unit price is not available) does not occur in the PSP scheme).

The game is extended in [41, 63] to networked auctions, and the same properties are obtained. In this networked game, players can be raw bandwidth sellers, end-users, or service providers buying and selling bandwidth to each other. Each player acts in the single node case, that is, tries to optimize his utility $u_i = \theta_i \circ e_i(a) - \sum_j c_i^j$, where e_i is a function called the *expected bottleneck* depending on the type of player and c_i^j is the total cost charged to player i by seller j .

In [65], simultaneous multi-unit descending-price auctions (or Dutch auctions) with different decreasing speeds are used. Indeed, the authors argue that, among other drawbacks, in the PSP auction, each player splits equally his bid among links, which might not be correct (depending on the congestion levels). The mechanism allows each user to buy the same quantity of bandwidth in all the links. According to experimental results, social welfare is improved with respect to PSP.

In [66], two new auction schemes are designed: the delta auction, which allows bids to take place continuously in order to prevent additive setup delays (at each node), and the Connection-Holder-is-Preferred-Scheme (CHiPS), based on the RSVP protocol, where holders of already running connections are preferred and are given a second chance if their actual bids are exceeded by new ones.

7. EXPECTED CAPACITY [43]

In [43], Clark also discusses how to charge the Internet. Like many authors, he examines more issues than he solves. One of his points about priority pricing is the following: “the effect of priority queuing is to build up a queue of lower-priority packets which will cause packets in this class to be preferentially dropped due to queue overflow. While dropped packets will be retransmitted, the rate adaptation of TCP translates these losses into a reduction in sending packets for these flows of packets.” Moreover he says that there is no obvious way to relate a particular priority to a particular achieved service. He then introduces his notion of *expected capacity*. The mechanism works as follows. At the network access, packets are flagged (*in* or *out*) depending on whether the incoming stream is inside or outside of the profile of the expected capacity (without any traffic shaping). When there is a point of congestion, *out*-tagged packets receive a congestion pushback notification (dropping or explicit congestion notification (ECN)). During periods of congestion, each sender executes a TCP algorithm which receives a congestion indication when it exceeds its expected capacity and starts to send packets that are flagged *out*.

As noted in [43], this scheme can also be implemented in a heterogeneous network of multi-provider Internets, where cooperating groups of providers make contracts to carry each other's traffic; when too many packets are marked according to the contract, they can be shifted out, or they can be charged according to some formula. Some dynamic tagging can also be implemented as is done in the case of smart markets by McKie-Mason and Varian.

Unfortunately, some problems need to be solved before this scheme can be implemented efficiently. First, depending on the application, the customer can be the sender or the receiver. The scheme previously described works if the customer is the sender. If he is the receiver, there is a need to design a complex protocol by means of which the sender is informed of the expected capacity contract, which can be also quite complex (to maintain flexibility of contracts). Second, what about multicast when each receiver has a different expected capacity?

8. CHARGING FOR ELASTIC TRAFFIC BASED ON THE TRANSFER RATE [1, 2, 45, 46, 48]

8.1 Work by Kelly et al.

The model presented here make it possible to combine different elastic traffic [1, 36, 45], where the rates are proportional to the willingness of each user to pay. The model is as follows. Consider a set of J resources with a capacity of C_j for resource j . A route r is a non-empty subset of J , and R is the set of possible routes. Let $A_{jr} = 1$ if $j \in r$ and 0 otherwise, and define A as $A = (A_{jr})$. If each route is associated with a user r , let $U_r(x_r)$ be the utility function of the user when the flow rate is x_r for user r . U_r is assumed to be an increasing, strictly concave and continuously differentiable function. Let $U = (U_r(\cdot), r \in R)$ and $C = (C_j, j \in J)$. From the system point of view, the idea is to maximize

$$\sum_{r \in R} U_r(x_r) \quad (3)$$

subject to $Ax \leq C$ and $x \geq 0$. From the user point of view, the idea is to maximize

$$U_r\left(\frac{w_r}{\lambda_r}\right) - w_r \quad (4)$$

over $w_r \geq 0$; here, the flow rate is $x_r = w_r \lambda_r$, where w_r is the amount that user r is willing to pay per unit of time and λ_r is the charge per unit of flow and unit of time for user r . Assume that the network knows $w = (w_r, r \in R)$ and attempts to maximize

$$\sum_{r \in R} w_r \log x_r \quad (5)$$

subject to $Ax \leq C$ and $x \geq 0$. This last assumption is very convenient because it makes it possible to compute optimal flow rates very easily. Indeed, it is shown in [1, 36, 45] that there always exist vectors λ , w and x satisfying $w_r = x_r / \lambda_r \forall r \in R$ such that w_r maximizes (4), x maximizes (5) and x is the unique solution maximizing (3).

It is also shown that the vector of rates x per unit charge is *proportionally fair*; that is, if $x \geq 0$ and $Ax \leq C$, and for any other feasible vector x^* , the aggregate proportional change is zero or negative:

$$\sum_{r \in R} w_r \frac{x_r^* - x_r}{x_r} \leq 0.$$

Even if solving this problem is mathematically tractable, the maximization of (5) needs to be done on a centralized basis, which is undesirable. In the following, how to proceed on a decentralized basis is explained. Consider the system of differential equations

$$\frac{d}{dt} x_r(t) = K_r \left(w_r(t) - x_r(t) \sum_{j \in r} \mu_j(t) \right), \quad (6)$$

where

$$\mu_j(t) = p_j \left(\sum_{s: j \in s} x_s(t) \right)$$

is the shadow price per unit flow through j and $p_j(t)$ is the derivative of the rate at which a cost is incurred at resource j when the load through it is y . The motivation behind these equations is as follows. If resource j generates a continuous stream of feedback signal at rate $yp_j(y)$ when the total flow through resource j is y ; then that resource j sends a proportion x_r/y of these feedback signals to a user r with a flow of rate x_r through resource j ; and that user r views each feedback signal as a congestion indication requiring some reduction of flow x_r . It is, then, a flow-control algorithm. It is shown using Lyapunov functions that the system of differential equations has a unique value x such that $x_r = w_r/\sum_{j \in r} \mu_j$ arbitrarily closely approximates the optimization of problem (5). Some stochastic perturbations of Eq. (6) are also analyzed in [45].

Eq. (6) shares several characteristics with TCP but also presents several differences as pointed out in [1]. In TCP, congestion is indicated by dropped or marked packets. There are, then, two multiplicative effects. In addition, it is shown that multiple TCP can be modeled by the system of differential equations

$$\frac{d}{dt} x_r(t) = \frac{m_r}{T_r} \left(\frac{m_r}{T_r^2} + \frac{x_r(t)^2}{2m_r} \right) \sum_{j \in r} \mu_j(t) \quad (7)$$

and can be viewed as acting as if the utility function of user r is

$$\frac{\sqrt{2}m_r}{T_r} \arctan \left(\frac{x_r T_r}{\sqrt{2}m_r} \right),$$

where T_r is the round trip time for the connection of user r and m_r is a parameter which would *inter alia* be multiplied by m , the rate of additive increase, and make $1 - 1/2m$ the multiplicative decrease factor in Jacobson's TCP algorithm. The stable point is then such that $\forall r$

$$x_r = \frac{m_r}{T_r} \left(\frac{2(1 - p_r)}{p_r} \right)^{1/2},$$

where $p_r = \sum_{j \in r} p_j$. Note that this conclusion cannot be reached when users or the network have routing choices.

Each customer can use intelligent agents [67] in order to optimize his willingness to pay according to the network congestion status.

In [50, 51], the necessary feedback to the users who adjust their rates is based on window-based congestion control, which is practically easy when connections use TCP. The method is proved to give optimal values. It is shown that the solution solves the same problem than the one of Kelly et al. Other implementations of the scheme are presented in [68-71] which give some scenarios and algorithms for user adaptation and network feedback signals for flow control.

8.2 Work by Low et al.

In [2, 46, 48] Low et al. study the same kind of problem than Kelly et al. investigated (mainly for ABR in ATM networks rather than for TCP on the Internet), and they obtain very similar solutions. The main difference is that in Low et al.'s work, users decide on their rates and pay, whereas in Kelly et al.'s work, users decide on their payments and receive what the network allocates. In addition, they use a decentralized algorithm to set prices according to changing network conditions. As in the previous subsection, we have a set L of unidirectional links of capacities c_l , $l \in L$, a set S of sources characterized by utility function $U_s(x_s)$ concave, with an increasing transmission rate x_s . The system is willing to maximize

$$\sum_{s \in S} U_s(x_s)$$

over x_s subject to capacity constraints. The problem is also decomposed, and the following synchronous algorithm is used in [48]

1. Each link receives the rates $x_s(t)$ if s 's route is through link l .
2. Each link l calculates its price $p_l(t+1)$ for a unit of bandwidth (in order to optimize the benefits obtained) using the gradient projection algorithm

$$p_l(t+1) = [p_l(t) + \gamma(x^l(t) - c_l)]^+, \quad (8)$$

where γ is a stepsize.

3. Each link communicates $p_l(t+1)$ to each source whose route is through link l .

Then, the algorithm for each source is as follows:

1. Each source is fed back the price $p^s = \sum_{l \in L(s)} p_l$ where $L(s)$ is the set of links that s uses.
2. The source then chooses then its transmission rate x_s (in an interval (m_s, M_s)) which maximizes its benefit:

$$U_s(x_s) - p^s(t) x_s.$$

3. These rates $x_s(t+1)$ are sent to the links which again calculate new prices, and so on.

The algorithm approaches a price vector $(p_l^* \mid l \in L)$ that aligns individual and system optimality with fairness properties. In [46], the gradient projection method is replaced with the Newton method, which typically converges much faster. Eq. (8) is, then, replaced with

$$p_l(t+1) = [p_l(t) + \gamma H_{ll}^{-1}(t)(x^l(t) - c_l)]^+, \quad (9)$$

where H is a Hessian matrix (see [46] for details). In [47], the equation is replaced with

$$p_l(t+1) = [p_l(t) + \gamma(\alpha_l b_l(t) + x^l(t) - c_l)]^+, \quad (10)$$

where α_l is a constant and $b_l(t)$ is the buffer backlog at link l .

The model is extended to the asynchronous case, where the updates at the sources and the links are not synchronized, which better resembles reality of large networks. The communication between sources and links is also greatly simplified as follows. In [2], the links estimate source rates using local information without omitting the optimality property. In [47, 72], communication from links to sources is accomplished using the proposed ECN (Explicit Congestion Notification) bit in the IP header. These modifications lead to a flow control scheme called REM (Random Early Marking), a variant of RED, and a stochastic version of the previous algorithm: link l marks an arriving packet with probability $m_l(t) = 1 - \Phi^{-p^s(t)}$ (with $\Phi > 1$). This leads to $m^s(t) = 1 - \Phi^{-p^s(t)}$. Inverting this equation, $p^s(t)$ is estimated by $\hat{p}^s(t) = -\log_\Phi(1 - \hat{m}^s(t))$ where $\hat{m}^s(t)$ is the fraction of marked packets (known by usual acknowledgement). The stability, performance and robustness of this version of the algorithm is studied in [73] using a continuous time version of the dynamics.

9. CONCLUSIONS

In this paper, we have surveyed usage-based pricing schemes without bandwidth reservation, with an emphasis on mathematical models. All these schemes have their own advantages, ranging from implementation simplicity to fairness. An interesting problem would be to compare (mathematically and in practice) their respective costs and benefits on a simple network in order to see one which is likely to perform the best.

Other issues are also worth studying. First, an interesting area of research is the pricing of Weighted Fair Queuing schemes. According to Clark [43], this mechanism would only achieve local equality inside one switch. This raises several questions. For example, in the multicast case, what does congestion along one path as to do in response to congestion along another? This also shows that the multicast case [74-76] needs more attention, which it received in [77], where pricing adaptation based on transfer rates is applied to multicast flows and fairness properties are obtained. Furthermore, as pointed out in [78], the optimality paradigm is not a panacea; more attention needs to be paid to architectures and structures.

REFERENCES

1. F. P. Kelly, "Mathematical modelling of the internet," in *Proceedings of the Fourth International Congress on Industrial and Applied Mathematics*, 2000, pp. 105-116.
2. S. H. Low, "Optimization flow control with on-line measurement or multiple paths," in *Proceedings of the 16th International Teletraffic Congress*, 1999, pp. 237-249.
3. P. Dolan, "Internet pricing is the end of the world wide wait in view?" *Communications & Strategies*, Vol. 37, 2000, pp. 15-46.
4. J. W. Roberts, "Quality of service guarantees and charging in multiservice networks," *IEICE Transactions on Communications*, Vol. E81, 1998, pp. 824-831.
5. L. A. DaSilva, "Pricing of QoS-enabled networks: a survey," *IEEE Communications Surveys & Tutorials*, Vol. 3, 2000.
6. M. Falkner, M. Devetsikiotis, and I. Lambadaris, "An overview of pricing concepts for broadband IP networks," *IEEE Communications Surveys & Tutorials*, Vol. 3, 2000.

7. T. Henderson, J. Crowcroft, and S. Bhatti, "Congestion pricing, paying your way in communication networks," *IEEE Internet Computing*, Vol. 5, 2001, pp. 77-81.
8. B. Stiller, P. Reichl, and S. Leinen, "Pricing and cost recovery for internet services: practical review, classification, and application of relevant models," *Netnomics*, Vol. 2, 2000, pp. 149-171.
9. P. Reichl and B. Stiller, "Nil nove sub sole: why internet charging schemes look like as they do," in *Proceedings of the 4th Berlin Internet Economic Workshop*, 2001.
10. L. Anania and R. J. Solomon, "Flat – The minimalist price," Lee W. McKnight and Joseph P. Bailey, ed., *Internet Economics*, MIT Press, 1997, pp. 91-118.
11. A. M. Odlyzko, "The current state and likely evolution of the internet," in *Proceedings of Globecom '99*, 1999, pp. 1869-1875.
12. Z. Fan, "Pricing and provisioning for guaranteed internet services," P. Lorenz, ed., *International Conference on Networking 2001*, LNCS, Springer-Verlag, Vol. 2093, 2001, pp. 55-64.
13. R. J. Gibbens and F. P. Kelly, "Measurement-based connection admission control," in *Proceedings of the 15th International Teletraffic Congress*, 1997, pp. 879-888.
14. R. J. Gibbens and F. P. Kelly, "Distributed connection acceptance control for a connectionless network," in *Proceedings of the 16th International Teletraffic Congress*, 1999, pp. 941-952.
15. F. P. Kelly, P. B. Key, and S. Zachary, "Distributed acceptance control," *IEEE Journal on Selected Areas in Communications*, Vol. 18, 2000, pp. 2617-2628.
16. D. Songhurst, ed., *Charging Communication Networks: From Theory to Practice*, Elsevier, Amsterdam, 1999.
17. F. P. Kelly, "Note on effective bandwidths," F. P. Kelly, S. Zachary, and I. B. Ziedins, ed., *Stochastic Networks: Theory and Applications*, Royal Statistical Society Lecture Notes Series, Oxford University Press, 1996, Vol. 4, pp. 141-168.
18. L. S. Zhang, D. Estrin, S. Shenker, and D. Zappala, "RSVP: a resource ReSerVation protocol," *IEEE Network Magazine*, 1993, pp. 8-18.
19. Q. Wang, J. M. Peha, and M. A. Sirbu, "Optimal pricing for integrated services networks," Lee W. McKnight and Joseph P. Bailey, ed., *Internet Economics*, MIT Press, 1997, pp. 353-376.
20. 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 Differential Services*.
21. I. Ch. Paschalidis and J. N. Tsitsiklis, "Congestion-dependent pricing of network services," *IEEE/ACM Transactions on Networking*, Vol. 8, 2000, pp. 71-184.
22. X. Lin and N. B. Shroff, "Pricing-based control of large networks," S. Palazzo, ed., *Evolutionary Trends of the Internet, 2001 Tyrrhenian International Workshop on Digital Communications (IWDC)*, LNCS, Springer-Verlag, Vol. 2170, 2001, pp. 212-231.
23. D. Mitra, K. G. Ramakrishnan, and Q. Wang, "Combined economic modeling and traffic engineering: joint optimization of pricing and routing in multi-service networks," in *Proceedings of the 17th International Teletraffic Congress*, 2001, pp. 73-85.
24. A. Odlyzko, "Paris metro pricing for the internet," in *ACM Conference on Electronic Commerce (EC '99)*, 1999, pp. 140-147.

25. P. Reichl, P. Flury, J. Gerke, and B. Stiller, "How to overcome the feasibility problem for tariffing internet services: the cumulus pricing scheme," in *Proceedings of IEEE International Conference on Communications 2001*, Vol. 7, 2001, pp. 2079-2083.
26. P. Reichl and B. Stiller, "Edge pricing in space and time: theoretical and practical aspects of the cumulus pricing scheme," in *Proceedings of the 17th International Teletraffic Congress*, 2001.
27. R. Bohn, H. W. Braun, K. C. Claffy, and S. Wolff, "Mitigating the coming internet crunch: multiple service levels via precedence," *Journal of High Speed Networks*, Vol. 3, 1994, pp. 335-349.
28. R. Cocchi, D. Estrin, S. Shenker, and L. Zhang, "A study of priority pricing in multiple service class networks," in *Proceedings of SIGCOMM '91*, 1991, pp. 123-130.
29. R. Cocchi, D. Estrin, S. Shenker, and L. Zhang, "Pricing in computer networks: motivation, formulation and example," *IEEE/ACM Transactions on Networking*, Vol. 1, 1993, pp. 614-627.
30. A. Gupta, D. O. Stahl, and A. B. Whinston, "Priority pricing of integrated services networks," L. W. McKnight and J. P. Bailey, ed., *Internet Economics*, MIT Press, 1997, pp. 323-352.
31. M. L. Honig and K. Steiglitz, "Usage-based pricing of packet data generated by a heterogeneous user population," in *Proceedings of IEEE INFOCOM 95*, 1995, pp. 867-874.
32. H. Mendelson and S. Whang, "Optimal incentive-compatible priority pricing for the M/M/1 queue," *Operations Research*, Vol. 38, 1990, pp. 870-883.
33. P. Marbach, "Pricing differentiated services networks: bursty traffic," in *Proceedings of IEEE INFOCOM 2001*, 2001, pp. 650-658.
34. P. Marbach, "The role of pricing in differentiated services networks," Technical Report CSRG-421, Dept. of Computer Science, University of Toronto, 2001.
35. P. Marbach, "Differentiated services networks: pricing and software agents," Technical Report CSRG-422, Dept. of Computer Science, University of Toronto, 2001.
36. F. P. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, 1997, Vol. 8, pp. 33-37.
37. A. Y. Ha, "Optimal pricing that coordinates queues with customer-chosen service requirements," *Management Science*, Vol. 47, 2001, pp. 915-930.
38. J. K. McKie-Mason and H. R. Varian, "Some economics of the internet," Technical report, No. 9401, University of Michigan, November 1993; <http://wueconb.wustl.edu:8089/eps/comp/papers:9401/9401001.pdf>.
39. J. K. McKie-Mason and H. R. Varian, "Pricing congestible network resources," *IEEE Journal on Selected Areas in Communications*, Vol. 13, 1995, pp. 1141-1149.
40. A. A. Lazar and N. Semret, "Design and analysis of the progressive second price auction for network bandwidth sharing," to appear in *Telecommunication Systems*, Vol. 13, 2001; <http://comet.columbia.edu/~nemo/telecomsys.pdf>.
41. N. Semret, "Market mechanisms for network resource sharing," PhD thesis, Dept. of Computer Science, Columbia University, 1999.
42. N. Semret, R. R.-F. Liao, A. T. Campbell, and A. A. Lazar, "Market pricing of differentiated internet services," in *Proceedings of the 7th International Workshop on Quality of Service*, 1999, pp. 184-193.

43. D. D. Clark, "Internet cost allocation and pricing," in L. W. McKnight and J. P. Bailey, ed., *Internet Economics*, MIT Press, 1977, pp. 215-252.
44. H. Yaïche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Transactions on Networking*, Vol. 8, 2000, pp. 667-678.
45. F. P. Kelly, A. K. Mauloo, and D. K. H. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, Vol. 49, 1998, pp. 237-252.
46. S. Athuraliya and S. H. Low, "Optimization flow control with Newton-like algorithm," *Telecommunication Systems*, Vol. 13, 2000, pp. 345-358.
47. S. Athuraliya and S. H. Low, "Optimization flow control, II: implementation," Technical Report, Net Lab., California Institute of Technology, 2000.
48. S. H. Low and D. E. Lapsley, "Optimization flow control, I: basic algorithm and convergence," *IEEE/ACM Transactions on Networking*, Vol. 7, 1999, pp. 861-874.
49. S. H. Low, F. Paganini, and J. C. Doyle, "Internet congestion control," *IEEE Control Systems Magazine*, Vol. 22, 2002, pp. 28-43.
50. R. J. La and V. Anantharam, "Charge-sensitive TCP and rate control in the internet," in *Proceedings of IEEE INFOCOM 2000*, 2000, pp. 1166-1175.
51. R. J. La and V. Anantharam, "Window-based control with heterogeneous users," in *Proceedings of IEEE INFOCOM 2001*, 2001, pp. 1320-1329.
52. R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, Vol. 35, 1999, pp. 1969-1985.
53. D. Hazlett, "An interim economic solution to internet congestion," *Social Science Computer Review*, Vol. 15, 1997, pp. 181-189.
54. F. P. Kelly, "Models for a self-managed internet," *Philosophical Transactions of the Royal Society*, Vol. A358, 2000, pp. 2335-2348.
55. S. Shenker, "Service models and pricing policies for an integrated services internet," in *Performance of Public Access to the Internet*, 1993, pp. 315-337.
56. T. Donald and L. Massoulié, "Impact of fairness on internet performance," in *Proceedings of ACM Sigmetrics 2001*, 2001, pp. 82-91.
57. L. Massoulié and J. Roberts, "Arguments in favour of admission control for TCP flows," in *Proceedings of the 16th International Teletraffic Congress*, 1999, pp. 33-44.
58. L. Massoulié and J. Roberts, "Bandwidth sharing: objectives and algorithms," in *Proceedings of IEEE INFOCOM '99*, 1999, pp. 1395-1403.
59. J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," in *Proceedings of SPIE '98*, 1998, pp. 55-63.
60. J. W. Roberts and L. Massoulié, "Bandwidth sharing and admission control for elastic traffic," *Telecommunication Systems*, Vol. 15, 2000, pp. 185-201.
61. R. Gibbens, R. Mason, and R. Steinberg, "Internet service classes under competition," *IEEE Journal on Selected Areas in Communications*, Vol. 18, 2000, pp. 2490-2498.
62. A. Gupta, D. O. Stahl, and A. B. Whinston, "A stochastic equilibrium model of internet pricing," *Journal of Economic Dynamics and Control*, Vol. 21, 1997, pp. 697-722.
63. N. Semret, R. R.-F. Liao, A. T. Campbell, and A. A. Lazar, "Pricing, provisioning

- and peering: dynamic markets for differentiated internet services and implications for network interconnections,” *IEEE Journal on Selected Areas in Communications*, Vol. 18, 2000, pp. 2499-2513.
64. B. Tuffin, “Revisited progressive second price auction for charging telecommunication networks,” *Telecommunication Systems*, Vol. 20, 2002, pp. 255-263.
 65. C. Courcoubetis, M. P. Dramitinos, and G. D. Stamoulis, “An auction mechanism for bandwidth allocation over paths,” in *Proceedings of the 17th International Teletraffic Conference*, 2001, pp. 1163-1174.
 66. P. Reichl, B. Stiller, and S. Leinen, “Auction models for multiprovider internet connections,” in *Proceedings of Messung, Modellierung und Bewertung MMB '99*, 1999.
 67. C. Courcoubetis, G. D. Stamoulis, C. Manolakis, and F. P. Kelly, “An intelligent agent for optimizing QoS-for-money in priced ABR connections,” *Telecommunications Systems, Special Issue on Internet Economics*, to appear.
 68. A. Ganesh, K. Laevens, and R. Steinberg, “Dynamics of congestion pricing,” Technical Report No. 70, Microsoft Research Limited, Cambridge, U.K., 2000.
 69. P. Key and D. R. McAuley, “Differential QoS and pricing in networks: where flow-control meets game theory,” in *IEE Proceedings*, 1999, pp. 39-43.
 70. P. Key and L. Massoulié, “User policies in a network implementing congestion pricing,” Technical Report, Microsoft Research Limited, Cambridge, U.K., 1999.
 71. K. Laevens, P. Key, and D. McAuley, “An ecn-based end-to-end congestion-control framework: experiments and evaluation,” Technical Report 104, Microsoft Research Limited, Cambridge, UK, 2000.
 72. D. E. Lapsley and S. H. Low, “Random early marking: an optimisation approach to internet congestion control,” in *Proceedings of IEEE International Conference on Networks '99*, 1999, pp. 67-74.
 73. F. Paganini, “Flow control via pricing: a feedback perspective,” in *Proceedings of the 2000 Allerton Conference*, 2000.
 74. A. Basu and S. J. Golestani, “Estimation of receiver round trip times in multicast communications,” Technical Report, Bell Laboratories; <http://www.belllabs.com/user/golestani/rtt.ps>.
 75. S. J. Golestani and S. Bhattacharyya, “A class of end-to-end congestion control algorithms for the internet,” in *Proceedings of International Conference on Network Protocol '98*, 1998, pp. 137-150.
 76. S. J. Golestani and K. K. Sabnani, “Fundamental observations on multicast congestion control in the internet,” in *Proceedings of IEEE INFOCOM '99*, 1999, pp. 990-1000.
 77. E. E. Graves, R. Srikant, and D. Towsley, “Decentralized computation of weighted max-min fair bandwidth allocation in networks with multicast flows,” in S. Palazzo, ed., *Evolutionary Trends of the Internet, 2001 Tyrrhenia International Workshop on Digital Communications*, LNCS, Springer-Verlag, Vol. 2170, 2001, pp. 326-342.
 78. S. Shenker, D. Clark, D. Estrin, and S. Herzog, “Pricing in computer networks: reshaping the research agenda,” *Computer Communication Review*, Vol. 26, 1996, pp. 19-43.



Bruno Tuffin (IRISA/INRIA) received his PhD degree in applied mathematics from Rennes 1 University in 1997. Since, he has been with INRIA-Rennes, France. His research interests include developing Monte Carlo and quasi-Monte Carlo simulation techniques for the performance evaluation of computer and telecommunication systems, and more recently developing Internet active measurement techniques and new pricing schemes. On this last topic, he is the coordinator of the INRIA's cooperative research action PRINXNET (see <http://www.irisa.fr/armor/Armor-Ext/RA/prixnet/ARC.htm>).