Incentive Pricing in Multi-Class Systems

Ariel Orda and Nahum Shimkin

Department of Electrical Engineering
Technion, Israel Institute of Technology, Haifa 32000, Israel
E-mail: {ariel;shimkin}@ee.technion.ac.il

In modern communication networks which offer multiple classes of services, an appropriate assignment of service classes to users (or applications) will have a key influence on the performance profile. Differentiated pricing is an important tool for guiding the user's choice. We consider a basic model for a multi-class system that offers multi-class services to multiple types of traffic, and propose a pricing framework which is based on the concept of *nominal* traffic assignment. Users (or their associated traffic) are categorized into a finite number of *traffic types*, which are distinct in their performance utilities at the different service classes. The system administrator specifies a required traffic assignment, which associates with each traffic type a *nominal* service class. Users, on the other hand, choose service classes so as to optimize their own performance. *Optimal prices* should provide incentives for the users to assign each traffic type to its nominal service class. Our goal is to implement a simple pricing scheme that provides such incentives.

We establish necessary and sufficient conditions for the existence of optimal prices and provide an algorithm for their computation. We indicate that optimal prices can tolerate fluctuations in the various parameters. Next, we propose a distributed algorithm, which can be used by the system to compute optimal prices even when it does not have sufficient knowledge of traffic characteristics. We then generalize our analysis to an extended model, which explicitly includes congestion effects.
Keywords: Network pricing, multi-class systems, network control, network design and management, game theory, broadband networking.

1. Introduction

Emerging broadband networks will be multi-class systems that offer several classes of service and support multiple types of traffic [1]. Users of such systems will be able to choose the service class for their traffic, thus playing an active role in the determination of system flows. The above scenario could lead to a deficient usage of resources. Users are concerned with the performance of their own traffic, rather than with the overall system performance. Thus, a user could be tempted to choose the best class of service for its traffic, even if a lower class of service would suffice. The network, on the other hand, needs its resources to be used in the most efficient way, meaning that, from a system point of view, premium service should be consumed only by those who really need it.

Our study is motivated by the fundamental distinction between *traffic types* and *service classes*: the traffic type is a property that characterizes the traffic requirements as perceived by a user (or, equivalently, the utility it perceives by meeting certain traffic requirements), and is typically known only to the user; the service class is a parameter assigned to the traffic stream (e.g., to each connection), which identifies the grade of service to be granted by the network. Consequently, the network relies on the user to choose the service class for its traffic. The basic idea of incentive pricing is to assign different prices to the various service classes, thereby providing the incentive for users to choose their service classes in a way which is appropriate from the network point of view.

A fundamental standpoint in our study is that such an appropriate assignment of traffic to service classes is specified by the network by associating with each traffic type a *nominal* service class. The choice of nominal service classes should be done in a way that efficiently combines both traffic performance requirements and network resource utilization. Often, the proper choice presents itself naturally, and is inherent in the design of the network service classes. As an example, consider a network that offers services according to the *Int-Serv* profiles proposed for the Internet, namely guaranteed-service [2], controlled-load [3] and best-effort. To keep the example simple, suppose that the network offers one class of guaranteed service (GS), two classes of controlled service (CL1 and CL2), and one class of best-effort service (BE1), each having the following properties.

- **GS:** Hard bound D on end-to-end delay, no packet losses (except for line bit errors).
- **CL1:** For traffic with maximal average bandwidth B_1, upper bounds on the fraction of the traffic that might exceed a specified end-to-end delay value $\bar{D}_1$ and a specified packet loss value $\bar{P}$; we assume that $\bar{D}_1 \approx D$.
- **CL2:** Similar to $CL1$, but with parameters $B_2 < B_1$ and $\bar{D}_2 < \bar{D}_1$.
- **BE:** No QoS guarantees.

Suppose that the network is to support the following types of traffic: “traditional”

TCP/IP applications, such as *e-mail* and *ftp*; interactive moderate-bandwidth applications, such as *audio conferencing*; large-bandwidth applications for entertainment, such as *video on demand*; and time-critical applications, such as *tele-surgery*. Clearly, all applications would benefit from being served by GS, and would rank BE as their last preference; the second preference depends on the application, e.g., it is $CL1$ for *video on demand* but $CL2$ for *audio conferencing*. On the other hand, the highly preferred service classes, GS in particular, demand considerably larger amounts of resources. Therefore, the network should be engineered to offer such classes of service to the traffic types that truly need them. In the above example, a proper matching between service classes and traffic types, *i.e.*, the *nominal assignment*, would be: GS to *tele-surgery*, $CL1$ to *video on demand*, $CL2$ to *audio conferencing* and BE to traditional applications. Accordingly, our pricing mechanism should provide the incentives to make these choices, e.g.: the price for $CL1$ should be smaller enough than GS but not too high compared to BE, so that *video on demand* applications would choose it, in view of their performance requirements.

In this study, we investigate pricing mechanisms that provide the required incentives for achieving the nominal assignment. An important requirement of network pricing, just as any other network control scheme, is simplicity of implementation. Specifically, it should allow an easy determination of the required parameters for a given system, in terms of the requirements it imposes on empirical measurements of traffic and user parameters. In addition, the pricing scheme should be simple in terms of the on-line computational requirements. Accordingly, we carry our investigation on a model that is simple yet captures the essence of the traffic assignment problem, and attempt to provide the required incentives through simple, fixed pricing schemes. We point out that the pricing schemes that we present are not intended to respond to dynamic, short-term fluctuations of the traffic conditions. Rather, they provide incentives for a balanced "long-term" allocation of traffic types among available service classes. Standard means of mitigating short-term congestion can and should be superimposed on this pricing scheme; these include admission and flow control, and even congestion-dependent pricing. In effect, our pricing scheme would present such short-term schemes with an improved traffic profile. For example, admission control would not end up rejecting "proper" connections from a service class due to previous admissions of "improper" connections.

The incorporation of pricing in modern networking has been gaining increasing interest in the past few years [4]–[18]. Some of these studies, e.g. [4,9], concentrated on best-effort networks such as the Internet, while others, e.g. [5,6], concentrated on ATM-based networks. Several studies proposed pricing mechanisms that are load-sensitive, often based on some bidding scheme such as the "smart market" approach [9]. Such schemes typically focus on the access to a *single-class* network, which offers the same type of service to all accepted packets. Multi-class services have been typically tackled by pricing schemes that maximize some prescribed social function (e.g., [5,11]). Pricing, as a mechanism for inducing efficient usage of system resources by self-optimizing users, has been the subject of several studies in the area of queuing systems, e.g.. [19]–[22]; in particular, [21] used differential pricing in a priority queuing system to induce a traffic allocation that minimizes a prescribed social cost. We note that none of these studies considered "nominal" allocations as the target of the pricing schemes.

While this work is motivated by networking applications, our model may also be applied to other multi-server, multi-user systems, provided that the proposed solution concept is meaningful there. Our pricing model may be compared and contrasted with the classical competitive market theory, and specifically the theory of general equilibrium (see, e.g., [23]). In the classical model, consumers optimize utility functions, subject to budgetary constraints; in our model, on the other hand, prices form part of the (dis)utility functions, as this formulation enables us to address directly the issue of traffic assignment. More importantly, our model requires matching a specific customer (traffic type) with a specific supplier (service class); this type of constraints is not covered by the competitive market model.

We proceed to outline our findings. Our basic model characterizes the relative utility of each traffic type at a given service class through a single figure of merit, or performance (dis)utility parameter, which is additive to the price. In fact, we allow interval uncertainty in these parameters, to account for variations within a traffic type and for changing network conditions, including congestion effects. This implicit consideration of congestion is expected to suffice in a properly designed network, which operates near nominal conditions. Moreover, such consideration fits well a "rigid" interpretation of an advertised grade of service. An extended model, which incorporates explicitly the effects of congestion, is considered in the second part of the paper.

For the basic model, we establish necessary and sufficient conditions for the existence of *optimal* prices, which lead each user to allocate its traffic according to the nominal assignment. We specify these prices and provide an algorithm for their computation. We then turn our attention to the practical problem of establishing the required prices when the network does not have sufficient knowledge on the traffic characteristics, or needs to adapt to changes. For such scenarios, we suggest a distributed, iterative algorithm, whereby each service class updates its price solely by measuring its current traffic. We show that the algorithm converges to optimal prices, thereby inducing the nominal traffic allocation. Next, we consider the congestion model. Here, a distinction must be made between different kinds of user optimization criteria, based on the amount of flow that they control. In this article we focus on the representative case of *individual optimization*, where each user controls only a small part of the total traffic. We also overview the effect of other kinds of users, which are considered in detail in [24]. These include the extreme case of *social optimization*, in which a single user is responsible for the whole traffic, and the typical intermediate case of *type optimization*, where each traffic type constitutes a different user.

The paper is structured as follows. In section 2 we formally define the basic model. The existence and structure of the corresponding optimal price assignments is presented in Section 3, while the iterative algorithm is formulated and analyzed in Section 4. Section 5 deals with the congestion model. Finally, in Section 6 we discuss these results and suggest several directions for further research. Due to space limits, some of the proofs are omitted from this version, and can be found in [24].

2. Model Formulation

We formulate in this section the basic model that will be considered in this paper. The proposed model is intended to capture the preferences of the users, in terms of the suitability of available service classes for their performance requirements, and the impact of pricing on these preferences.

The basic model does not include explicit dependence of performance on traffic conditions; however, it does allow for interval uncertainty and variability of these parameters. The former case will be treated separately in section 5.

Consider a communication network which offers several classes of service, denoted by the set $\mathcal{K} = \{1, 2, \ldots, K\}$. We assume that traffic in this system is divided into a finite number of distinct traffic types, denoted $\mathcal{I} = \{1, 2, \ldots, I\}$.

Traffic types differ in their service requirements; the requirements of each type are captured in our model through the relative level of satisfaction, or utility, from the performance that this traffic type will experience at each service class. This measure will be quantified below by specifying *performance disutility parameters* for each traffic type.

Traffic is composed of data units, or flow units, which may be packets, sessions, etc., depending on the application. The flow units of all traffic types are assumed to be similar, and, in particular, they cannot be distinguished from each other by the network. The arrival process is assumed to be ergodic, with rate Λ^i for type $i \in \mathcal{I}$. The users of the system are free to allocate any part of their traffic to any service class.

A basic premise of our formulation is that the "proper" traffic distribution is specified by a unique allocation of each traffic type to one of the service classes. We refer to this allocation as the *nominal class allocation*, and denote by $k(i)$ the nominal service class for traffic type i. The nominal allocation reflects the intent, or design, of the network service provider. While we do not specify at this stage a specific mechanism through which this allocation is chosen, it is obvious that a proper allocation should take into account the suitability of the service class for the traffic type, and a balanced utilization of network resources. In fact, the specification of service class characteristics may often be done with specific traffic types and applications in mind.

To be more precise, we introduce some additional notation. Denote by λ_k^i the rate of type i traffic which is directed to service class k, and let $\lambda_k = \sum_{i\in\mathcal{I}} \lambda_k^i$ denote the total rate of traffic allocated to service class k. The vector $\boldsymbol{\lambda}^i = \{\lambda_k^i\,,\, k \in \mathcal{K}\}$ denotes the *traffic flow of type* i, and $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^i\,,\, i \in \mathcal{I}\}$ is the *(global) traffic flow*. The set of *feasible flows* for traffic type i is the set of allocations $\boldsymbol{\lambda}^i$ which are non-negative and flow conserving, namely $\mathcal{L}^i = \{\boldsymbol{\lambda}^i : \lambda_k^i \geq 0,\, \sum_k \lambda_k^i = \Lambda^i\}$. Similarly, $\mathcal{L} = \{\boldsymbol{\lambda} : \boldsymbol{\lambda}^i \in \mathcal{L}^i\}$ is the set of *feasible traffic flows.*

The *nominal flow allocation* $\boldsymbol{\lambda}^*$ may now be defined as

$$\lambda_k^{i*} = \begin{cases} \Lambda^i \; k = k(i) \\ \;0 \;\; k \neq k(i) \end{cases} . \tag{1}$$

Here $k(i) \in \mathcal{K}$ is the nominal service class for traffic type i, as specified by the nominal class allocation $\{k(i),\, i \in \mathcal{I}\}$.

Pricing is assumed to depend on the service class, and is proportional to traffic volume. Thus, each flow unit may be charged a pre-specified price, which may depend on its service class (but not on its traffic type attribute, which, as mentioned above, is indistinguishable by the network). The price per unit flow in service class k is denoted by p_k, and $\mathbf{p} = (p_1, \ldots, p_K)$ is the *price vector.* As mentioned, in this paper we explore the use of this basic differential pricing mechanism as a means to regulate the traffic distribution between the different service classes, so that it will coincide with the nominal flow allocation.

To model the service requirements and preferences of the different traffic types, we associate with each traffic type i a set of *performance disutility parameters* c_k^i, one for each service class k. This parameter measures the subjective disutility (negative level of satisfaction) per unit flow of traffic type i with respect to the performance offered by service class k. We assume that performance disutility is additive to the prices p_k, so that the total disutility (or cost) per unit flow of traffic type i at service class k equals $c_k^i + p_k$.

In order to accommodate variability within a service class, which may account for model inaccuracies and changing network conditions, we allow for interval uncertainly in the performance disutility parameters. Thus, we only specify lower and upper bounds for these parameters, so that each parameter c_k^i is assumed to reside in an interval $[\underline{c}_k^i, \bar{c}_k^i]$. We do not make any additional restriction regarding the distribution of the actual traffic parameters within the specified interval; e.g., the parameter c_k^i of a data unit of type i may be near the upper bound $\bar{c}_k^i$ for each flow unit, or it may be near the lower bound, or it may be evenly distributed in this interval.

Assuming self-optimizing users, all traffic of type i, with performance disutility parameter c_k^i, will be applied to the service class k that minimizes the per-unit cost $c_k^i + p_k$. A traffic flow which obeys this rule is called a *user-optimal flow* for the price vector $\mathbf{p}$ (more precise definitions will be given in the next section). In what follows, we seek to characterize those price vectors $\mathbf{p}$ which cause the user-optimal allocation to coincide with the nominal traffic allocation.

Remark: *Elastic demand.* Our model apparently assumes that the amount of traffic which enters the system is fixed. However, the important case of cost-dependent demand may be incorporated simply by appending a "null" service class (say, $k = 0$) to the actual ones. The disutility parameters c_0^i associated with this class should reflect the dissatisfaction experienced by type-i traffic for not getting service, and the service price p_0 should naturally be set to 0. The latter does not interfere with the results in the following sections, since all price vectors are specified in relative terms (i.e., up to an additive bias, which may be used to get $p_0=0$).

3. Optimal Pricing

This section is devoted to the characterization of optimal prices. In accordance with the objectives outlined in the previous section, we say that a price vector is optimal if it induces nominal flows for all traffic types. We start by stating precise definitions. We then consider the issue of existence of optimal prices, and state explicit conditions on the nominal flow assignment which guarantee existence of optimal prices. Computation of optimal prices is considered next, followed by an indication of certain relations between optimal prices and network design.

3.1. Definitions

Given a price vector $\mathbf{p}$, a flow allocation $\boldsymbol{\lambda} \in \mathcal{L}$ is considered *user-optimal* if every flow unit is allocated to the service class which presents a minimal cost for that flow, according to its type. To express that in mathematical terms, consider first the special case with no uncertainly in the performance disutility parameters. That is, $\underline{c}_k^i = \bar{c}_k^i = c_k^i$. Then $\boldsymbol{\lambda}$ would be user-optimal if every traffic type i chooses the service class k with minimal cost $c_k^i + p_k$. Equivalently,

$$\lambda_k^i > 0 \text{ implies } c_k^i + p_k \leq c_l^i + p_l \quad \text{for all } l \neq k,, \ l, k \in \mathcal{K} \ . \tag{2}$$

In the general case with interval uncertainty, since we make no assumptions on the distribution of the actual parameters c_k^i within their intervals $[\underline{c}_k^i, \bar{c}_k^i]$, we only require the above relation to hold for some possible value of the parameters. Using the appropriate upper and lower bounds in the above inequality, we arrive at the following definition.

Definition 1. Given a price vector $\mathbf{p}$, a flow allocation $\boldsymbol{\lambda} \in \mathcal{L}$ is *user-optimal* if, for every traffic type i,

$$\lambda_k^i > 0 \text{ implies } \underline{c}_k^i + p_k \leq \bar{c}_l^i + p_l \quad \text{for all } l \neq k. \tag{3}$$

It is readily verified that (3) is satisfied if and only if $\boldsymbol{\lambda}^i = \{\lambda_k^i,\ k \in \mathcal{K}\}$ is a minimizer over $\mathcal{L}^i$ of $\sum_k (c_k^i + p_k)\lambda_k^i$, for *some* set of parameters $\{c_k^i \in [\underline{c}_k^i, \bar{c}_k^i]\}$. Since this minimization is independent for each traffic type i, we can unify these minimization problems and conclude that an allocation $\boldsymbol{\lambda}$ is user-optimal if and only if it is a solution of the following "social" minimization problem, for some set of parameters $\{c_k^i \in [\underline{c}_k^i, \bar{c}_k^i]\}$:

$$\min\{\sum_{i,k}(c_k^i + p_k)\lambda_k^i \,:\, \boldsymbol{\lambda} \in \mathcal{L}\}\,. \tag{4}$$

This equivalent characterization will be useful in the sequel. We note that this equivalence of individual and social optimization is a consequence of the assumed congestion-independent cost structure, which eliminated cost externalities between users. Note further that this "social" cost reflects the average users' cost, but need not coincide with the system objectives.

Obviously, a user-optimal flow allocation always exists. For uniqueness, it is required that the same service class k will be uniquely optimal for type i traffic, for all possible values of the parameters. It easily follows that $\boldsymbol{\lambda} \in \mathcal{L}$ is a *unique* user-optimal flow allocation if, for every traffic type i,

$$\lambda_k^i > 0 \text{ implies } \bar{c}_k^i + p_k < \underline{c}_l^i + p_l \quad \text{for all } l \neq k. \tag{5}$$

We may now define the class of price vectors in which we are interested:

Definition 2. A price vector $\mathbf{p}$ is said to be *optimal* if it induces a unique user-optimal flow allocation which equals the nominal flow allocation.

Any flow allocation which is the unique user-optimal allocation for *some* price vector will be termed *price-inducible.*

3.2. Existence of Optimal Prices

Existence of an optimal price vector is not guaranteed. Intuitively, some consistency should be maintained between the user service preferences (expressed by the disutility parameters c_i^k) and the nominal flow allocation, specified by the set $\{k(i),\ i \in \mathcal{I}\}$. To illustrate, consider the case of two traffic types and two service classes. Assume that $c_1^1 < c_2^1$ and $c_2^2 < c_1^2$, namely, with no external pricing, traffic type 1 prefers service class 1, and traffic type 2 prefers service class 2. Assume further that the nominal class allocation was chosen as $k(1) = 2$ and $k(2) = 1$ (traffic type 1 to service class 2, and type 2 to class 1). It is easily seen that no pricing vector $\mathbf{p} = (p_1, p_2)$ can induce this allocation. Indeed, fixing p_2, if p_1 is chosen high enough so that type 1 traffic prefers service class 2, then so will type 2 traffic. The price-inducible flow allocations in this system are:

$\{k(1) = 1, k(2) = 2\}$ (with null pricing), $k(1) = k(2) = 1$ (set p_1 low enough), and $k(1) = k(2) = 2$ (set p_1 high enough).

Existence of an optimal price vector is resolved in the following two propositions, which give alternative forms of the uniqueness conditions. The first is stated in terms of cyclic permutations of traffic between the service classes, and the second in terms of a related social optimization problem.

Loosely speaking, condition (7) of the next theorem means that there should be no circular deviation out of the nominal flow which results in a decrease in the social performance cost. To present the required conditions, define the constants

$$\alpha(k,l) \triangleq \min_{i:k(i)=k} (\underline{c}_l^i - \bar{c}_k^i) \,. \tag{6}$$

If the minimization domain is empty, i.e., no i exists with $k(i) = k$, define $\alpha(k,l) \triangleq \infty$. The minimized term may be interpreted as the (minimal) difference in disutility for traffic type i between service classes k and l; the minimum is taken over all traffic types which are allocated to class k as their nominal service class. Furthermore, define a *cycle* C as a cyclic sequence of service class pairs:

$$C = \{(k_1,k_2),(k_2,k_3),\ldots,(k_{m-1},k_m),(k_m,k_1)\}\,, \quad m \geq 2.$$

A cycle is simple if it does not contain any sub-cycle, namely, if the participating classes $\{k_1,\ldots,k_m\}$ are distinct. Obviously, the number of simple cycles is finite.

Theorem 3. An optimal price vector exists if and only if

$$\sum_{(k,l)\in C} \alpha(k,l) > 0 \tag{7}$$

for every simple cycle C. Service classes which are empty under the nominal allocation need not be included in these cycles.

The starting point for the proof of this result is the following basic observation.

Lemma 4. A price vector $\mathbf{p}$ is optimal if and only if

$$p(k,l) := p_k - p_l < \alpha(k,l) \text{ for all } k,\ l \neq k\,. \tag{8}$$

Proofs of this lemma and Theorem 3 appear in the Appendix.

The following result presents an alternative existence condition, for the special case where there is no uncertainty in the parameters.

Proposition 5. Assume that the parameters $\{c_k^i\}$ are precisely specified ($\underline{c}_k^i = \bar{c}_k^i$). Let λ_k^* denote the total flow allocated to service class k under the nominal allocation $\boldsymbol{\lambda}^*$. Consider the following constrained minimization problem:

$$\min_{\boldsymbol{\lambda}}\{\sum_{i,k} c_k^i \lambda_k^i \ : \ \boldsymbol{\lambda} \in \mathcal{L}, \ \sum_i \lambda_k^i = \lambda_k^*\}\,. \tag{9}$$

An optimal price vector exists if and only if this problem has a unique solution which coincides with the nominal flow allocation.

Proof. The proof here is simple upon noting that, by collecting terms, the characterization of user optimality in (4) may be equivalently written as

$$\min\{\sum_{i,k} c_k^i \lambda_k^i + \sum_k p_k \lambda_k \ : \ \boldsymbol{\lambda} \in \mathcal{L}\}\,. \tag{10}$$

Now the latter is just the Lagrangian formulation of the constrained optimization problem (9), with p_k the Lagrange multiplier corresponding to the constraint $\lambda_k := \sum_i \lambda_k^i = \lambda_k^*$. It then follows that $\boldsymbol{\lambda}^*$ uniquely solves (9) if and only if there exists a vector $\mathbf{p}$ so that $\boldsymbol{\lambda}^*$ uniquely solves (10). $\square$

The necessity of the last condition is easily understood in light of the characterization (10) of user-optimal behavior. Obviously, any solution of (10), for given $\mathbf{p}$, must minimize the first sum subject to the second being kept constant, hence in particular when the link flows $\{\lambda_k\}$ are fixed. The reverse (and more interesting) implication means that the nominal allocation needs only be optimal subject to fixed total link flows, and then pricing can be used to regulate those flows.

It is not clear how the latter result can be extended to the case of uncertain parameters. The problem is of course that we require the *same* price vector to induce the nominal allocation for all possible values of the parameters.

It should also be mentioned that the last proposition is close in spirit to the (first and second) welfare theorems of general equilibrium theory in microeconomics (see, e.g., [23]), which state that, in a competitive market, any Walrasian equilibrium is Pareto efficient, and vice-versa. In particular,[1] the last proposition may be deduced from general results pertaining to exchange economies with indivisible commodities, cf. [25].

[1] We would like to thank one of the referees for bringing these relations to our attention.

We conclude with a brief discussion on the computation of optimal prices. The set of optimal price vectors is defined by condition (8), which is a set of linear constraints. Hence, this set is an open polyhedron. An optimal price vector may be obtained by one of the standard methods for obtaining a feasible point in a set of linear inequalities. A particular (finite) algorithm, implied by the proof of theorem 3, is presented in [24].

In general, there is a whole set of price vectors which are optimal according to our definitions, and one may use additional considerations in order want to choose a particular member of this set. If these considerations can be expressed through a cost function, this yields a minimization problem subject to the linear constraints (8), for which efficient algorithms exist. A possible criterion may be to choose a "centrally located" element of the optimal set (in some appropriate sense), in order to guarantee some robustness with respect to model variations. We shall not pursue this approach to robust design here, but rather point out to the explicit robustness which is inherent in the uncertain disutility formulation.

3.3. Pricing and Network Design

Till now we have not considered a specific way through which the nominal allocation is chosen. Here we briefly consider the case where the nominal allocation is chosen as an outcome of a network design problem. In particular, consider the following problem:

$$\min_{\boldsymbol{\lambda}}\{\sum_{i,k} c_k^i \lambda_k^i + \sum_k g_k \lambda_k\}\,,$$

In the minimized cost, the first term stands for the user disutility (with c_k^i the same disutility parameters as considered in our model so far, assumed to be given). The second term, with $g_k > 0$, measures the cost, for the service provider, of setting up and maintaining the service class k to support a traffic volume λ_k.

The solution of this design problem may be specified as the nominal flow allocation. Let us assume, for simplicity of exposition, that this problem has a unique solution.[2] The following simple result, which is a direct consequence of the characterization (10) of user-optimality, shows how the optimal price vector may be chosen in this case.

[2] This will be the case if $\arg\min_k\{c_k^i\}$ is a singleton for each i; this obviously holds when all the c_k^i parameters are distinct, which holds generically in the parameter space.

Proposition 6. Let the nominal allocation be chosen as the solution of the above network design problem (assumed to be unique). Then an optimal price vector, which makes this allocation user-optimal, is given by $p_k = g_k$, $k \in \mathcal{K}$.

An alternative network design criterion, which directly takes into account the revenues obtained by the service provider through pricing, is the following one:

$$\min_{\boldsymbol{\lambda}}\{\sum_{i,k} c_k^i \lambda_k^i + \sum_k g_k \lambda_k - \sum_k p_k \lambda_k\} \, .$$

This cost function is appropriate when prices have an absolute meaning, which prevents addition of an arbitrary bias to all elements of the price vector. This is the situation in the elastic demand case, discussed at the end of the previous section. With this cost function, an identical result to the last proposition is obtained, except that now the price vector should be chosen as $p_k = g_k/2$.

4. Iterative Pricing

In the previous section we addressed the issue of price determination from an "open loop" design viewpoint - the prices are computed using a model of the system. Here we look at this problem from an operational viewpoint. Assuming that the system is in operation, using a fixed price vector, it is expected that some deviation from the expected performance will be observed, or perhaps develop over time. One option, then, is to update our model of the system based on the new information, and compute a new price vector according to this model. The second possibility is to update the price vector as a direct function of the observed performance parameters. This second option is explored in this section.

Consistent with our basic assumption, whereby different traffic types may not be discernible by the system, we assume that the observed quantities are the total flow at each service class (but not the distribution according to traffic types). Thus, we observe the vector $\{\lambda_k, k \in \mathcal{K}\}$, to which we refer as the *total flow*.

As a preliminary result, we show first that, under user-optimality, the total flow uniquely determines the flow allocation vector (i.e., the distribution of user types among the service classes may be determined even if we observe only the total flow at each service class). This implies, in particular, that just by looking at the total flow we can determine whether a given (user-optimal) flow allocation is the nominal one. Thus motivated, we propose an iterative algorithm for updating the price vector according to the observed total flow, and analyze its convergence to the optimal prices.

Proposition 7. Let $\boldsymbol{\lambda}$ be a unique user-optimal allocation induced by some price-vector $\mathbf{p}$, and let $\hat{\boldsymbol{\lambda}}$ be a unique user-optimal allocation corresponding to $\hat{\mathbf{p}}$. If the total flow is the same in both cases ($\lambda_k = \hat{\lambda}_k,\ k \in \mathcal{K}$), then $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\lambda}}$ coincide.

Proof. See [24]. □

Let $\lambda^o = \{\lambda_k^o\}$ denote a nominal total-flow vector, which we seek to enforce through an appropriate price vector $\mathbf{p}^o$. The following iterative update law is suggested for determining a price vector $\mathbf{p}^o$ that will lead to the nominal total flow allocation λ^o.

Iterative Procedure: For $t = 0, 1, 2, \ldots$:

$$p_k(t+1) = p_k(t) + d_k(\lambda_k(t) - \lambda_k^o)\,, \quad k \in \mathcal{K}\,, \tag{11}$$

where each d_k is a pre-specified positive constant.

The interpretation is as follows: at each step t we employ the price vector $p(t)$, and observe the resulting flow $\lambda(t) = \{\lambda_k(t)\}$. If it deviates from the required (nominal) one, we update the price in proportion to that deviation, and with the proper sign – if the flow is too high the price for that class is increased, and vice-versa. We note that prices may become negative in the iterative schemes; see the remark at the end of this section.

Our proposed iterative process is mathematically akin to a (discrete-time version of) the classical Walrasian tatonnement process, which offers a mechanism for pre-trade price adjustment that converges (under suitable assumptions) to the market equilibrium (see [23]). Stability results for this process are surveyed in [26]. Nonetheless, our case is not covered by these results. A specific difficulty is that our model includes "dead zones", where local changes in the price vector do not affect the flow, and which leads in general to non-uniqueness of the optimal price vector. A convergence proof, appropriate for this case, is thus provided in the following.

Note that the proposed procedure is distributed, in the sense that only local flow information is used at each service class to update the price for that class.

The following assumptions will be employed for convergence analysis of this scheme.

Assumptions I:

(I1) For each price vector $\mathbf{p}$, the user-optimal flow allocation is unique. The total flow vector $\{\lambda_k\}$ corresponding to that user-optimal flow will be denoted by $\lambda^*(\mathbf{p})$.

(I2) $\lambda(t) = \lambda^*(\mathbf{p}(t))$ for every $t \geq 1$.

(I3) If the price p_k changes by Δp_k, the corresponding change $\Delta\lambda_k^*$ in $\lambda_k^*(\mathbf{p})$ is bounded by

$$|\Delta\lambda_k^*| \leq \alpha_k |\Delta p_k| ,$$

where α_k is a finite constant. Note that this may be written as $|\frac{\partial\lambda_k^*(\mathbf{p})}{\partial p_k}| \leq \alpha_k$ wherever $\lambda_k^*(\cdot)$ is differentiable.

(I4) The nominal total flow vector λ^o is user-optimal for some price vector $\mathbf{p}^o$, i.e., $\lambda^o = \lambda^*(\mathbf{p}^o)$. (Note that $\mathbf{p}^o$ need not be unique.)

Assumption (I2) means that a new price change is performed only after the flow vector has settled on its user-optimal value. Assumption (I4) merely requires the existence of an optimal price vector. The main technical assumption is (I3), which requires that a small change in price to result in a small change in flow; note that the uniqueness requirement (I1) is actually included in (I3) (take $\Delta p_k = 0$).

The continuity requirement (I3) may be justified on practical grounds, since we do expect the users' reaction to small price changes to be proportionally small. In terms of our mathematical model, this requirement means that we cannot allow a non-negligible proportion of users to have exactly the same disutility parameter, but rather require each parameter (c_k^i) to be continuously distributed (with no point masses) within its uncertainty interval $[\underline{c}_k^i, \bar{c}_k^i]$. Note that by "distribution" here we refer to the empirical (non-probabilistic) distribution which specifies the number (or fraction) of users in any range of possible parameters (in the same way that the height distribution in a given population may be specified). This distribution is (absolutely) continuous if the cumulative distribution function has no jumps (and can be described by a finite density).

To clarify this discussion, we consider a simple example.

Example 8. Consider two service classes, $k = 1$ and $k = 2$, and a single traffic type with disutility parameters c_1 and c_2 (the superscripts denoting the traffic type are dropped for convenience). Assume first that there is no uncertainty in these parameters, namely $c_k = \underline{c}_k = \bar{c}_k$. This means that all members of this traffic type have exactly the same parameters. Choosing $p_1 = c_2 - c_1 + p_2$ will then result in user indifference between the two traffic types, and the user-optimal allocation will not be unique. Moreover, this is a discontinuity point since modifying p_1 by $-\epsilon$ or $+\epsilon$ results in a shift of the whole flow from service class 1 to 2.

Consider next the same situation except that the disutility parameters are continuously distributed in their uncertainty intervals. Assume, for concreteness, that c_1 in known to lie in [2,3] and c_2 in [8,9], and further that their distributions within these intervals are uniform and independent (this means, e.g., that 1% of the traffic has c_1 in $[a,\, a+.01]$ for $2 \leq a \leq 2.99$). Then, it is easily seen that both uniqueness and continuity of the user-optimal response are guaranteed. Indeed, denoting the flow requirement of this type by λ, the optimal flow assignment will be $[\lambda_1, \lambda - \lambda_1]$, with $\lambda_1 = 0$ for $p_1 > 7$, $\lambda_1 = \lambda$ for $p_1 < 5$, and $\lambda_1 = \lambda(7 - p_1)/2$ for $5 \leq p_1 \leq 7$.

We stress again that the main argument for the continuity assumption is its practical plausibility. What we have claimed here is that this property can be comfortably accommodated within our model.

Theorem 9. Consider the algorithm (11), with the coefficients d_k chosen so that $d_k \alpha_k < 1$. Then, for any initial condition $\lambda(0)$,

$$\lim_{t \to \infty} \lambda(t) = \lambda^o\,,$$

and $\mathbf{p}(t)$ converges to a constant and optimal price vector.

The remainder of this section will be devoted to the proof of this theorem, which proceeds through the following lemmas.

Lemma 10. Given two price vectors $\mathbf{p}$ and $\hat{\mathbf{p}}$, the difference $\lambda^*(\hat{\mathbf{p}}) - \lambda^*(\mathbf{p})$ may be expressed as $\lambda^*(\hat{\mathbf{p}}) - \lambda^*(\mathbf{p}) = -M \cdot (\hat{\mathbf{p}} - \mathbf{p})$, where $M = M(\mathbf{p}, \hat{\mathbf{p}}) = \{m_{kl}\}$ is a matrix which satisfies the following properties:
(P1) $0 \leq m_{kk} \leq \alpha_k$, with α_k as in assumption (I3).
(P2) $m_{kl} \leq 0$ for $k \neq l$

(P3) $\sum_k m_{kl} = 0$ (zero column-sums).
(P4) $\sum_l m_{kl} = 0$ (zero row-sums).

The proof appears in the Appendix.

Using vector notation, the price update law (11) may be written as:

$$\mathbf{p}(t+1) = \mathbf{p}(t) + D(\lambda(t) - \lambda^o)\,,$$

where D is the diagonal matrix $D = \mathrm{diag}\{d_1, \ldots, d_K\}$. Let $\mathbf{p}^o$ denote a price vector which satisfies $\lambda^o = \lambda^*(\mathbf{p}^o)$. While this equality may be satisfied by many price vectors, $\mathbf{p}^o$ denotes hereafter a fixed choice of one such vector. Recalling that $\lambda(t) = \lambda^*(\mathbf{p}(t))$, and using the last lemma to evaluate $\lambda(t) - \lambda^o = \lambda^*(\mathbf{p}(t)) - \lambda^*(\mathbf{p}^o)$, we obtain

$$\mathbf{p}(t+1) = \mathbf{p}(t) - DM(t)(\mathbf{p}(t) - \mathbf{p}^o)\,,$$

where $M(t) := M(\mathbf{p}(t), \mathbf{p}(t+1))$ is a matrix that satisfies properties (P1–P4). We have thus established the following relation:

Lemma 11. Let $\mathbf{p}^o$ be a price vector which satisfies $\lambda^o = \lambda^*(\mathbf{p}^o)$, and denote $y(t) = \mathbf{p}(t) - \mathbf{p}^o$. Then

$$y(t+1) = (I - DM(t))y(t)\,, \tag{12}$$

where I is the unit matrix, $D = \mathrm{diag}\{d_1, \ldots, d_k\}$, and $M(t)$ is a square matrix which satisfies properties (P1–P4).

The main point in the remaining proof will be to show that the relative price vector $y(t)$ (hence the price vector $\mathbf{p}(t) = y(t) + \mathbf{p}^o$) converges to a fixed vector y_∞. We emphasize that we do not expect that this convergence will necessarily be to the zero vector, since in general the equality $\lambda^o = \lambda^*(\mathbf{p})$ may be satisfied by many price vectors (i.e., $\mathbf{p}^o$ is not unique).

Let $A = \{a_{kl}\}$ denote the square matrix $A = (I - DM)$, with D as defined in the previous lemma, and M any matrix which satisfies properties (P1–P4). It is easily verified that the following corresponding properties hold for A, with $g_k := 1/d_k$ and $\beta_0 = min_k(1 - d_k\alpha_k)$.
(Q1) $a_{kk} \geq \beta_0 > 0$.
(Q2) $a_{kl} \geq 0$ for all k, l.
(Q3) $\sum_k g_k a_{kl} = g_l$ (normalized column-sums).
(Q4) $\sum_l a_{kl} = 1$ (unit row-sums).

Equation (12) may now be written as $y(t+1) = A(t)y(t)$, where $A(t)$ satisfies properties (Q1–Q4). The following convergence result may be established for this system:

Lemma 12. Let

$$y(t+1) = A(t)y(t)\,, \quad t = 0, 1, \ldots, \tag{13}$$

where every $A(t)$ satisfies conditions (Q1–Q4) above, and $y(0)$ is arbitrary. Then $y(t)$ converges to a finite constant vector y_∞.

The proof appears in the Appendix.

We may now conclude the proof of Theorem 9. The last lemma has established that $\mathbf{p}(t)$ converges to a finite limit. Recalling (11), where $d_k > 0$, it follows immediately that $\lambda_k(t)$ converges to λ_k^o for every k, as claimed. Thus, the limit prices make the users abide by the nominal flow. □

To conclude, we have established the asymptotic convergence of the proposed iterative procedure to the optimal pricing vector. More complicated schemes, which rely on additional information regarding the system, may be useful in terms of convergence speed, and are left for further study.

Remark. The update law (11) can lead to negative prices for some of the service classes, which may not be directly implementable. However, this may be remedied in a natural way. First, we note that prices are relative, so that a uniform bias may be added. As even simpler option is to modify the algorithm by truncating any negative price to zero and proceeding as usual. It may be shown that convergence will not be affected – see the corresponding remark at the end of the Appendix. Finally, we note that while emergence of negative prices is mathematically feasible, we do not expect this to happen under normal operation where the iterative scheme will be used mainly for price calibration. In this context a negative price should indicate an outstanding circumstance which may indeed call for action outside the scope of the algorithm.

5. Congestion Model

Till now we have not included in our model an explicit representation of the effects of possible congestion. If the impact of congestion is limited and reasonably small, then it may be handled within the uncertainty model, discussed previously: the lower bound would correspond to the ideal, uncongested state, whereas the upper bound would correspond to the worst (congested) state. However, if the impact of congestion is major, it may be required to model it explicitly and analyze its effect. This is the subject of the current section.

We thus consider a modification of our former model. In the new model, the subjective disutility parameter c_k^i is a *function* of the load (total flow) at the service classes, $c_k^i(\lambda)$, which grows as congestion builds up. Here $\lambda = \{\lambda_k\}$ is the total flow vector, where λ_k is the total flow at service class k.

A major difference between this and the previous (uncongested) model is that now each part of the traffic may affect the performance of the other parts. Therefore, the concept of "user optimality" now depends on the amount of flow that each user controls. In this article, we focus on the typical case of *individual optimization*, where each user controls an infinitesimally small part of the total traffic. This is the case, for example, when each individual call is optimized independently. At the end of the section we also briefly discuss the cases of *social optimization*, where a single user is responsible for the whole traffic, and *type optimization*, where each traffic type constitutes a different user, which are explored in detail in [24].

In all but the social optimization case, the inter-dependence of user choices creates a game problem. Accordingly, user-optimality will be defined below by resorting to the Nash equilibrium solution [27].

We begin by considering general congestion disutilities, which extend the basic model of Section 2. We then specialize the general results within the framework of a particular model, as defined in subsection 5.2.

5.1. Congestion-Dependent Disutilities

As mentioned above, $c_k^i = c_k^i(\lambda)$ is a function of the total traffic flow vector λ. We assume that, for all $i \in \mathcal{I}$ and $k \in \mathcal{K}$, c_k^i is continuous, strictly increasing in λ_k and nondecreasing in λ_l for all $l \in \mathcal{K}$. When no further assumptions are made, we shall term such functions as *general congestion disutility parameters.*

The inherent complication of general congestion disutility parameters is that, in absence of an appropriate form of convexity, there may exist multiple (Nash-) equilibrium points, which may be unconnected and result in different performance for the users. As we shall see, it is fairly easy to identify price vectors under which the nominal allocation is an equilibrium. However, the co-existence of other equilibria cannot be prevented in general.

We proceed with a formal definition of individual optimality, which can be viewed as a Nash equilibrium for infinitesimal users (or Wardrop equilibrium). The relevant local results are indicated. We then formulate appropriate convexity-type conditions which guarantee the uniqueness of the equilibrium solution.

Definition 13. Given a price vector $\mathbf{p}$, the flow allocation $\boldsymbol{\lambda} \in \mathcal{L}$ is *individually optimal* if, for every traffic type i,

$$\lambda_k^i > 0 \text{ implies } c_k^i(\lambda) + p_k \leq c_l^i(\lambda) + p_l \quad \text{for all } l \neq k. \tag{14}$$

It is *strictly* individually optimal if the above inequalities are strict.

A price vector is *locally optimal* if the nominal traffic allocation is individually optimal with respect to it. Observe that the total flow vector λ corresponding to the nominal allocation is known, hence so are the values of the associated disutility parameters $c_k^i(\lambda)$. It is then trivial to show that a complete characterization of locally optimal prices may be obtained exactly as in the uncongested case, i.e. when c_k^i are constants, simply be replacing these constants by their nominal values. This result, while local in nature, has practical significance. Indeed, frequently, the user is not likely to consider drastic alterations of the nominal allocation. Rather, starting from this assignment, it would consider (or try) small deviations. In this case, the local result is sufficient to maintain the nominal allocation.

In order to obtain a global result, paralleling the one obtained for the uncongested case, further assumptions are required. For example, the following *diagonal monotonicity* condition is sufficient for guaranteeing a unique individually optimal flow allocation for each given price vector.

Proposition 14. Suppose that, for every pair of feasible traffic flows $\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}} \in \mathcal{L}$, with $\lambda \neq \hat{\lambda}$ (unequal total flows), there exist positive numbers $\{r^1, r^2, \ldots, r^I\}$ for which the following relation holds:

$$\sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} r^i (\hat{\lambda}_k^i - \lambda_k^i)(c_k^i(\hat{\lambda}) - c_k^i(\lambda)) > 0. \tag{15}$$

(i) If $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\lambda}}$ are individually optimal allocations, their corresponding total flow vectors $\lambda = \{\lambda_k\}$ and $\hat{\lambda} = \{\hat{\lambda}_k\}$ are identical.

(ii) If an individually optimal allocation is strict, then it is unique, i.e., no other individually optimal allocation exists.

Proof. See [24]. □

A similar result, but with $r^i \equiv 1$, was stated in [28], in the context of transportation networks. The essence of Condition (15) is that each $c^i_k(\hat{\lambda})$ is increasing in the corresponding variable λ^i_k, and the effect of this variable is dominant with respect to other variables.

A particular case for which this condition is satisfied is studied in the next subsection. As another example, in [24] we show that (15) holds in a network with a two-priority server.

5.2. Proportional Congestion Disutilities

Suppose that, for each service class k, there is a function T_k and constants $c^1, c^2, \ldots, c^I$, such that $c^i_k(\lambda) = c^i \cdot T_k$. We assume that T_k depends on the flow at only service class k, i.e., $T_k = T_k(\lambda_k)$, and that the number of service classes matches the number of traffic types, i.e., $K = I = N$. The classes of service are ordered by descending quality, i.e., service class 1 offers the best service quality and service class N offers the worst quality, in a sense defined below (Assumption T4). Similarly, traffic types are ordered by descending quality of service requirements, meaning that $c^i > c^{i+1}$ for all $1 \leq i < N$. Service classes are matched to the traffic types accordingly, so that the i-th service class is the nominal one for the i-th traffic type.

We make the following technical assumptions on the functions T_k:

(T1) $T_k : [0, \infty) \to (0, \infty]$.

(T2) $T_k(\lambda_k)$ is positive, strictly increasing and convex.

(T3) $T_k(\lambda_k)$ is continuously differentiable.

In addition, we make the following assumption, which differentiates among service classes:

(T4) $T_k(\Lambda^k) < T_{k+1}(\Lambda^{k+1})$ for all $1 \leq k < N$.

The last assumption requires that, at nominal flow allocation, the quality of service is indeed descending in the class number.

As before, a flow unit pays a price p_k for getting service from the k-th server. Thus, a flow unit belonging to the i-th traffic type, sent by the user to the k-th server, incurs a total cost of

$$c^i \cdot T_k(\lambda_k) + p_k,$$

when the total traffic sent to that server is λ_k.

Note that the diagonal monotonicity condition (15) in Proposition 14, hence the uniqueness results of that proposition, hold for the present cost structure. Indeed, choosing $r^i = \frac{1}{c^i}$, the left-hand side of (15) becomes

$$\sum_k (\hat{\lambda}_k - \lambda_k)(T_k(\hat{\lambda}_k) - T_k(\lambda_k))$$

which is positive for any $\lambda \neq \hat{\lambda}$, by Assumption (T2).

Definition 15. A price vector $\mathbf{p}$ is *globally optimal* with respect to individually optimal allocations if, given this price vector, the nominal flow allocation is a unique and strict individually optimal flow allocation.

We are now ready to present a complete characterization of optimal price vectors.

Theorem 16. Assume strict individually optimal allocations.

(i) A price vector $\mathbf{p}$ is optimal if and only if it satisfies

$$c^{k+1}(\tau_{k+1} - \tau_k) < p_k - p_{k+1} < c^k(\tau_{k+1} - \tau_k)\,, \quad k = 1, \ldots, N-1\,, \tag{16}$$

where $\tau_k = T_k(\Lambda^k)$.

(ii) An optimal price vector always exists.

(iii) Any optimal price vector is class-monotonic, namely $p_k > p_{k+1}$ for $k \in \mathcal{K}$.

Proof. See [24]. □

Condition (16) has the following simple interpretation. The right inequality ensures that a flow unit of class k, nominally being served in service class k, will not find it advantageous to move to the next lower service class $k+1$. This is achieved by choosing $(p_k - p_{k+1})$ *small* enough. The left inequality, on the other hand, ensures that a flow unit of class $k+1$ will not find it advantageous to move to the next higher service class k. This is achieved by choosing $(p_k - p_{k+1})$ *large* enough.

Note that condition (16) only deals explicitly with the benefit of shifting flow to an *adjacent* service class. The point of the above theorem is that, under our class and type ordering assumptions, this is enough to ensure that no flow unit will find it advantageous to shift to *any* other service class rather than its nominal one.

Finally, note that condition (16) guarantees a range of optimal price vectors, i.e., it provides a robust solution, in the sense discussed at the end of Section 3.2.

5.2.1. Extensions: Social and Type Optimality

So far in our consideration of congestion effects, we have assumed that users are of infinitesimal size. In other situations the network may support large users that control a non-negligible part of the traffic. It is obviously of interest to examine the possible consequences of this situation.

An analysis of "large user" behavior can be found in [24], for the proportional congestion costs case. We briefly review some of these results for two kinds of user behavior, namely *social* and *type* optimality.

Social optimality applies to the extreme case when all traffic is controlled by a single self-optimizing user. The network employs pricing in order to make this user allocate its traffic according to the nominal allocation. It turns out that the user cost function is not convex in general, and may exhibit multiple local minima. (A notable exception is when all the disutility parameters c^i are equal, and the problem reduces to the standard single-class routing problem. This case is obviously of no interest here.) For this reason, price optimality results are of local nature here.

It is shown in [24] that there always exists a price vector under which the nominal flow allocation is *locally socially optimal*, i.e. a local solution of the user optimization problem. The set of price vectors with this property is characterized by conditions similar to (16), modified to account for cost externalities of shifted traffic. For global results, additional network supervision is required, and some initial steps in this direction are mentioned there.

We turn now to the intermediate case of *type optimality*, where each traffic type is governed by a distinct user. In other words, it is assumed that the different traffic types are non-cooperative, perhaps ignorant of each other, and each seeks to optimize its own type-specific cost function. Interaction between different traffic types exists, however, through the sharing of service resources. The resulting multi-agent decision problem faced by the users (for given prices) corresponds to a non-cooperative game, and the type-optimal flow allocation is thus the Nash equilibrium solution [27] of this game. It is shown in [24] that a *unique* equilibrium exists for any price vector. However, the existence of an optimal price vector (for which the equilibrium solution coincides with the nominal allocation) is not guaranteed, but rather depends on the model parameters. Certain congestion-dependent pricing schemes are considered for this case and shown to guarantee local results.

It is worthy to note that individual optimization (involving infinitesimal users) presented the easiest scenario for our pricing scheme. The introduction of large self-optimizing users in a congested network brought about several problems, ranging from multiple locally-optimal flows to non-existence of optimal fixed prices, which require further consideration.

6. Conclusions

This paper proposes an approach for incentive pricing in multi-class systems. It consists of two phases: in the first phase, which is typically interconnected with the network design, distinct traffic types are identified, and a nominal service class is assigned to each traffic type. In the second phase, an appropriate pricing vector is specified, which induces the nominal traffic allocation by providing the required incentives to the users.

The basic premise which underlies incentive pricing is that the choice of service class, among several available options, is (and should) be left to the user. Differential pricing is a key tool through which the network can induce efficient use of network resources, while keeping high performance service available for those who require it. The pricing scheme proposed in this paper is simple and is based on per-unit pricing at each service class.

User performance preferences are captured by disutility parameters, allowed to vary in given intervals, which depend on the traffic type and service class. Prices are levied on a per-unit basis at each service class. For this fundamental model, a complete characterization of pricing vectors and their computation is presented, as well as an iterative algorithm for on-line price adjustment.

In a congestion-prone system, an explicit consideration of congestion and its effect on user preferences may be required. In this case the identity of the user, in terms of the amount of traffic at its disposal, is a key factor. We have focused here on the case of many small users, and outlined other fundamental cases, namely the extreme case of a single user and an intermediate case where each user controls a single traffic type. The results highlight the differences between these cases, as well as the difficulties that emerge in the congestion-prone model, which are closely related to the inherent non-uniqueness of the equilibrium conditions in these systems.

Some important issues present themselves as possible subjects for further research. Additional study of congestion effects is called for. Aspects of network topology and their interaction with pricing schemes and congestion need also be explored. More generally, the pricing mechanism studied in this paper is the simplest, fixed, per-unit pricing. More involved pricing schemes have been suggested in the literature, including schemes which are meant to mitigate congestion by using congestion-adaptable prices [9].[3] The incorporation of such pricing mechanisms in our framework for *multi-class* networks seems worthy of further investigation. From a practical standpoint, it is clear that many implementation issues are yet to be considered. These include engineering issues, such as how to apply our findings to specific network architectures, as well as market investigation issues, which are needed in order to quantify user preferences and requirements.

APPENDIX

Proof of Lemma 4. By definition of the nominal allocation, it is uniquely user-optimal if and only if

$$c^i_{k(i)} + p_{k(i)} < c^i_l + p_l \quad \text{for all } l \neq k(i),\ i \in \mathcal{I}.$$

This may be written as

$$p_{k(i)} - p_l < c^i_l - c^i_{k(i)}\,, \quad \text{for all } l \neq k(i),\ i \in \mathcal{I}.$$

Using the definition of $\alpha(k,l)$, we obtain (8). □

[3] Flow dependent pricing has been touched upon very briefly at the end of the previous section.

Proof of Theorem 3. To begin, we clarify the issue of empty service-classes, namely those service classes to which no flow is assigned under the nominal allocation. By definition, if k is an empty service class we have $\alpha(k,l) = \infty$, so that condition (7) holds trivially whenever a cycle includes an empty class. Moreover, keeping a service class empty may always by achieved by assigning to it a high enough price. Thus, for the rest of this proof we may disregard the empty service classes, and assume for convenience that all service classes are non-empty.

From Lemma 4, we know that existence of an optimal price vector is equivalent to the existence of a vector $\mathbf{p}$ which satisfies (8). We proceed to show that the latter is equivalent to condition (7) in the theorem claim.

In one direction, (7) follows trivially from (8) since $\sum_{(k,l)\in C} p(k,l) = 0$ for any cycle C.

To establish the opposite implication, assume that (7) holds. We shall construct a price vector $\mathbf{p}$ which satisfies (8). In fact, we shall construct a set of numbers $P = \{p(k,l) : k,l \in \mathcal{K}\}$ which satisfy (8) and $\sum_C p(k,l) = 0$ for any cycle C. An appropriate price vector $\mathbf{p}$ may then be trivially extracted (e.g., fix some l and set $p_k = p(k,l)$).

Construction of P will proceed iteratively, using graph-theoretic principles. For simplicity we refer to the pairs (k,l) as links. We start by noting that since (7) holds for any simple cycle, it also holds for any cycle. We now set $p(k,l) = \alpha(k,l) - \epsilon$, with $\epsilon > 0$ small enough so that

$$\sum_C p(k,l) \geq 0 \quad \text{for any cycle } C. \tag{17}$$

The key idea is to reduce the values $p(k,l)$ involved in cycles with strictly positive values, so as to nullify the values of these cycles, while not violating the non-negativity requirement for other cycles.

A link (k',l') is said to be *all positive* if $\sum_C p(k,l) > 0$ for every cycle C which includes (k',l'). We shall rely on the following claim: if $\sum_C p(k,l) \geq 0$ for every cycle C, with at least one of these inequalities being strict, then there exists an all-positive link (k',l'). To establish this claim, let C_0 be the cycle with $\sum_{C_0} p(k,l) > 0$. We shall show that this cycle contains an all-positive link. Assume, by way of contradiction, that for every link $(k,l) \in C_0$ there exists a closed cycle $C = C(k,l)$ which includes (k,l) such that $\sum_C p(k',l') = 0$. Let us write the cycle $C(k,l)$ as $(k,l) \cup \tilde{C}(l,k)$, where $\cup$ denotes here the concatenation of links, and $\tilde{C}(l,k)$ is a path which connects l to k. Note that $C_1 := \cup_{(k,l)\in C_0}\tilde{C}(l,k)$ is also a cycle, hence we obtain the contradiction

$$0 = \sum_{(k,l)\in C_0} 0 = \sum_{(k,l)\in C_0}\sum_{(k',l')\in C(k,l)} p(k',l') = \sum_{(k,l)\in C_0} p(k,l) + \sum_{(k',l')\in C_1} p(k',l') > 0\,.$$

The last inequality follows since the first term is strictly positive by choice of C_0, while the second is non-negative by our standing assumption (17). This establishes the required claim.

We may now continue with our construction. If $\sum_C p(k,l) = 0$ for every simple cycle C, then we are done. Otherwise, by the claim above there exists at least one all-positive link. We may now reduce the value $p(k,l)$ of this link just until the value of one of the (simple) cycles which include this link is reduced to zero. As a result, this link ceases to be all-positive, and the number of all-positive links is reduced at least by one. Note, furthermore, that conditions (17) and $\{p(k,l) < \alpha(k,l)\}$ are preserved. Thus, proceeding this way for every remaining all-positive link, the process terminates with P as required. □

Proof of Lemma 10. Let $\frac{\partial\lambda^*(\mathbf{p})}{\partial\mathbf{p}}$ denote the Jacobian matrix of λ^*, with $\frac{\partial\lambda_k^*(\mathbf{p})}{\partial\mathbf{p}_l}$ as the kl element. Note that, by Assumption (I3), the function λ^* is Lipschitz continuous, hence absolutely continuous, therefore that Jacobian exists almost everywhere, and its integral equals λ^*. We start by showing that $-\frac{\partial\lambda^*(\mathbf{p})}{\partial\mathbf{p}}$ satisfies the properties (P1–P4) above. By definition of $\lambda^*(\mathbf{p})$, it is evident that $\lambda_k^*(\mathbf{p})$ is monotone non-increasing in p_k (increasing the price of service class k decreases its total flow). Together with the continuity assumption (I3), it follows that $\frac{\partial\lambda_k^*}{\partial p_k}$ takes values in $[-\alpha_k, 0]$. Similarly, for $l \neq k$ it is obvious that $\lambda_l^*(\mathbf{p})$ is monotone non-decreasing in p_k, hence $\frac{\partial\lambda_k^*}{\partial p_l}$ is non-negative. Furthermore, denoting by $\Delta\lambda^*$ the change in $\lambda^*(\mathbf{p})$ corresponding to a change $\Delta p_l > 0$ in p_l, we obtain by flow conservation that $\sum_k \Delta\lambda_k^* = 0$. It follows that $\sum_k \frac{\partial\lambda_k^*}{\partial p_l} = 0$, namely $\frac{\partial\lambda^*}{\partial\mathbf{p}}$ has zero column sums. Finally, note that increasing any component p_l of $\mathbf{p}$ produces the same differential pricing (hence the same user-optimal allocation) as *reducing* all other components of $\mathbf{p}$ by the same amount; this implies that $\frac{\partial\lambda^*}{\partial p_l} = -\sum_{k\neq l}\frac{\partial\lambda^*}{\partial p_k}$, hence $\frac{\partial\lambda^*}{\partial\mathbf{p}}$ has zero row sums. Combining the above, it follows that $-\frac{\partial\lambda^*}{\partial\mathbf{p}}$ satisfies properties (P1–P4).

We may now write

$$\lambda^*(\hat{\mathbf{p}}) - \lambda^*(\mathbf{p}) = \int_{\mathbf{p}}^{\hat{\mathbf{p}}} \frac{\partial\lambda^*}{\partial\mathbf{p}}(\mathbf{p}')\,d\mathbf{p}'$$

$$= \left\{ \int_0^1 \frac{\partial \lambda^*}{\partial p}(p + \alpha(\hat{p} - p))d\alpha \right\} (\hat{p} - p)$$
$$:= -M(p, \hat{p}) \cdot (\hat{p} - p) \; . \tag{18}$$

For the second equality the change of variables $p' = p + \alpha(\hat{p} - p)$ was used. Note that the defining expression of the matrix $M(p, \hat{p})$ may be interpreted as a convex combination of matrices $-\frac{\partial \lambda^*}{\partial p}$. Recalling that the latter satisfy properties (P1–P4), and noting that properties (P1–P4) are invariant under convex combinations, it follows that the matrix $M(p, \hat{p})$ satisfies these properties as well. □

Proof of Lemma 12. The basic idea of the proof is to define K partial sums of the components of $y(t)$, with the k-th sum taken over the k largest elements of $y(t)$ (with appropriate weighting), and showing that these sums are non-increasing. (To make a connection with stability theory of dynamic systems, we note that these partial sums may be regarded as vector Lyapunov functions.)

Assume for simplicity that the coefficients $\{g_k\}$ are arranged in decreasing order, i.e., $g_1 \geq g_2 \geq \ldots \geq g_K$ (this may always be arranged by re-ordering the classes). Given a K-vector y, let $L_k(y)$ denote the k-th largest element of y, and define the partial sums:

$$S_k(y) = \sum_{k'=1}^{k} g_{k'} L_{k'}(y) \, , \quad k = 1, \ldots, K \, .$$

We first establish that, if A is a matrix which satisfies properties (Q2–Q4) and $z = Ay$, then $S_k(z) \leq S_k(y)$ for every k. Let k be fixed. Denoting by $i(k')$ the index of the k'-th largest element in z, we obtain

$$S_k(z) = \sum_{k'=1}^{k} g_{k'} z_{i(k')} = \sum_{k'=1}^{k} g_{k'} (\sum_{l=1}^{K} a_{i(k'),l} y_l) = \sum_{l=1}^{K} h_l y_l \, , \tag{19}$$

where

$$h_l := \sum_{k'=1}^{k} g_{k'} a_{i(k'),l} \, .$$

By properties (Q2) and (Q3) of A it follows that

$$h_l \leq g_l \, , \quad l = 1 \ldots, K \, , \tag{20}$$

and from property (Q4) we obtain, by changing the order of summation,

$$\sum_{l=1}^{K} h_l = \sum_{k'=1}^{k} g_{k'} \sum_{l=1}^{K} a_{i(k'),l} = \sum_{k'=1}^{k} g_{k'} \, . \tag{21}$$

The required upper bound on $S_k(z)$ may now be obtained by bounding the last expression in (19), subject to the derived constraints on the magnitude and sum of $\{h_l\}$. Note first that

$$S_k(z) = \sum_{l=1}^{K} h_l y_l \leq \sum_{l=1}^{K} L_l(h) L_l(y)$$

since an inner product of two vectors is maximized if their elements are re-ordered so that the largest element of one multiplies the largest element of the other, etc. Expanding the right-hand side we obtain

$$\begin{aligned}
S_k(z) &\leq \sum_{l=1}^{k} g_l L_l(y) + \sum_{l=1}^{k} (L_l(h) - g_l) L_l(y) + \sum_{l=k+1}^{K} L_l(h) L_l(y) \\
&\leq \sum_{l=1}^{k} g_l L_l(y) + L_k(y) \sum_{l=1}^{k} (L_l(h) - g_l) + L_k(y) \sum_{l=k+1}^{K} L_l(h) \\
&= S_k(y) + L_k(y) (\sum_{l=1}^{K} L_l(h) - \sum_{l=1}^{k} g_l) \\
&= S_k(y)
\end{aligned}$$

The second inequality follows since (20) implies that $(L_l(h) - g_l) \leq 0$ (recall that the $\{g_l\}$ coefficients are already in decreasing order), $h_l \geq 0$, and $\{L_l(y)\}$ is decreasing in l by definition. The last equality follows from (21). The required bound has thus been established.

Returning to the system (13), we conclude that, for each k, the partial sum sequence $S_k(y(t))$ in monotone non-increasing. On the other hand, using property (Q4) of $A(t)$ it is easily verified that the elements of $y(t)$ are lower bounded, hence the partial sums are lower bounded. It follows that all the partial sum sequences converge to a finite limit. This implies that, for each k, the sequence $L_k(y(t))$ converges to a finite limit. Thus, we have shown that the largest element of the vectors $y(t)$ converges, as does the second-largest element, and so on. Let us denote the limiting vector by z_∞ (with elements arranged, say, in descending order).

We proceed to show that convergence of $y(t)$ may be inferred from the established convergence of $\{L_k(y(t))\}$. This is where property (Q1), which has not been used thus far, will come in. First, observe that $y(t)$ converges to z_∞ up to reordering of elements at each t; this means that all possible limit points of $\{y(t)\}$ are permutations of z_∞, hence that $\{y(t)\}$ may be divided into a finite number of subsequences, each converging to a different permutation of z_∞. To show that $y(t)$ itself converges, it remains to show that only one such limit point (or subsequence) can exist. For that, it is sufficient to show that no "jumps" between different subsequences are possible; that is, if $z = Ay$, with A satisfying (Q1–Q4), then z cannot be "very close" to a permutation of y (other than y itself). Since the set of matrices satisfying (Q1–Q4) is compact, it is actually sufficient to show that z cannot be *equal* to a permutation of y.

To establish the latter, we note that (Q1) enables to decompose A as follows:

$$A = \beta_0 I + (1 - \beta_0)\tilde{A}\,,$$

where I is the unit matrix, $0 < \beta_0 < 1$, and $\tilde{A}$ is a matrix which satisfies properties (Q2–Q4). Consequently, $z = Ay$ may be written as the following convex combination:

$$z = \beta_0 y + (1 - \beta_0)\tilde{y} \quad \text{where} \;\; \tilde{y} := \tilde{A}y\ .$$

Now, as was established above, since $\tilde{A}$ satisfies properties (Q2–Q4), the partial sums of $\tilde{y}$ are not larger than those of y (i.e., $S_k(\tilde{y}) \le S_k(y)$ for every k). It is now a simple arithmetic matter to deduce that, unless $\tilde{y} = y$,, the partial sums of z cannot equal those of y. It follows that z cannot be any permutation of y other than y itself. With this established, the proof of convergence of $y(t)$ is complete. □

Remark. Consider the modified algorithm based on (11), except that negative prices are truncated to zero. We claim that the convergence result in Theorem 9 is not affected. We will only give a sketch of a possible proof. First, is is easily verified that the iteration (11) does not modify the sum $\sum_k d_k^{-1} p_k(t)$. Thus, each time that a negative price p_k is truncated up to 0 this sum is increased by $d_k^{-1}|p_k|$, and is therefore lower bounded by the sum of all truncations times a constant. On the other hand, it may be independently verified that the prices are bounded. It then follows that the total amount of truncations is bounded, which implies that amount of truncation at each step must decrease to zero asymptotically. Hence, the effect of truncations on the the algorithm becomes negligible, and converge can be established as before.

Acknowledgements

This research was supported by a grant from the Ministry of Science of Israel, in cooperation with the French Ministry of Research and Technology. The authors would like to thank Eitan Altman and Adam Shwartz for valuable discussions and suggestions. We would also like to thank the referees for valuable observations, in particular for pointing out the relations with general equilibrium theory, and for useful technical comments.

References

[1] C. Partridge, *Gigabit Networking*, Addison-Wesley, 1994.

[2] S. Shenker, C. Partridge, and R. Guérin, Specification of guaranteed quality of service. INTERNET-RFC 2212, Internet Engineering Task Force, September 1997.

[3] J. Wroclawski, Specification of the controlled-load network element Service. INTERNET-RFC 2211, Internet Engineering Task Force, September 1997.

[4] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, Pricing in computer networks: motivation, formulation and examples, IEEE/ACM Transactions on Networking 1, 1993, pp. 614–627.

[5] S. H. Low and P. P. Varaiya, A new approach to service provisioning in ATM networks, IEEE/ACM Transactions on Networking 1, pp. 547–553, 1993.

[6] F. P. Kelly, Tarrifs and effecitve bandwidths in multiservice networks, in: *Proc. ITC'94*, pp. 401–410, Elsevier Science, 1994.

[7] R. P. McLean and W. W. Sharkey, An approach to the pricing of broadband telecommunication services, Telecommunication Systems 2, pp. 159–184, 1994.

[8] S. Jordan and H. Jiang, Connection establishment in high speed networks, IEEE Journal on Selected Areas in Communications 13, pp. 1150–1161, September 1995.

[9] J. K. MacKie-Mason and H. R. Varian, Pricing congestable network resources, IEEE Journal on Selected Areas in Communications 13, pp. 1141–1149, September 1995.

[10] R. J. Edell, N. McKeown, and P. P. Varaiya, Billing users and pricing for TCP, IEEE Journal on Selected Areas in Communications, vol. 13, pp. 1162–1175, September 1995.

[11] J. Sairamesh, D. F. Ferguson, and Y. Yemini, An approach to pricing, optimal allocation and quality of service provisioning in high-speed packet networks, in: *Proc. IEEE INFOCOM'95*, Boston, MA, pp. 1111–1119, April 1995.

[12] S. Shenker, D. Clark, D. Estrin and S. Herzog, Pricing in Computer Networks: reshaping the research agenda, Telecommunications Policy 20, pp. 183–201, 1996.

[13] F. Toutain and O. Huber, A General preemption-based admission policy using a smart market approach, in *Proc. IEEE INFOCOM'96*, San Francisco, CA, pp. 794–801, March 1996.

[14] L. McKnight and J. Bailey (Eds.), *Internet Economics*, MIT Press, Cambridge, MA, 1997.

[15] F. P. Kelly, A. K. Maulloo and D. K. H. Tan, Rate control in communication networks:

shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49, pp. 237–252, 1998.

[16] Y. A. Korilis, T. A. Varvarigou and S. R. Ahuja, Incentive-compatible pricing strategies in noncooperative networks, in: *Proc.IEEE INFOCOM'98*, San Francisco, CA, April 1998.

[17] R. J. Gibbens and F. P. Kelly, Resource pricing and the evolution of congestion control, Automatica 35(12), 1999.

[18] Y. A. Korilis and A. Orda, Incentive compatible pricing strategies for QoS routing, in: *Proc.IEEE INFOCOM'99*, New York, NY, March 1999.

[19] U. Yechiali, Customers' optimal joining rules for the GI/M/s queue, Management Science 18, pp. 434–443, 1972.

[20] H. Mendelson and U. Yechiali, Controlling the GI/M/1 queue by conditional acceptance of customers, European Journal of Operations Research 7, pp. 77–85, 1981.

[21] H. Mendelson and S. Whang, Optimal incentive-compatible priority pricing for the M/M/1 queue, Operations Research 38, pp. 870–883, 1990.

[22] S. Rao and E. R. Petersen, Optimal pricing of priority services, Operations Research 46, pp. 46–56, 1998.

[23] H. R. Varian, *Microeconomic Analysis*, W. W. Norton & Company, 3rd ed., 1992.

[24] A. Orda and N. Shimkin, Incentive pricing in multi-class communication networks, EE Pub. No. 1031, Department of Electrical Engineering, Technion, Haifa, Israel, June 1996. May be obtained from
"ftp://ftp.technion.ac.il/pub/supported/ee/Network/orda.shimkin.price96.ps".

[25] S. Bikhchandani and J. W. Mamer, Competitive equilibrium in an exchange economy with indivisibilities, Journal of Economic Theory 74, pp. 385–413, 1997.

[26] F. Hahn, Stability, in *Handbook of Mathematical Economics, Vol. II* (K. J. Arrow and M. D. Intrilgator, eds.), North-Holland Publishing Company, 1982.

[27] R. B. Myerson, *Game Theory: Analysis of Conflict*, Harvard University Press, 1991.

[28] S. C. Dafermos, Traffic equilibrium and variational inequalities, Transportation Science 14, pp. 42–54, 1980.