Capacity Management and Equilibrium for Proportional QoS

Ishai Menache and Nahum Shimkin, *Senior Member, IEEE*

***Abstract*— Differentiated services architectures are scalable solutions for providing class-based Quality of Service (QoS) over packet switched networks. While qualitative attributes of the offered service classes are often well defined, the actual differentiation between classes is left as an open issue. We address here the *proportional* QoS model, which aims at maintaining pre-defined ratios between the service class delays (or related congestion measures). In particular, we consider capacity assignment among service classes as the means for attaining this design objective.**

Starting with a detailed analysis for the single hop model, we first obtain the required capacity assignment for fixed flow rates. We then analyze the scheme under a reactive scenario, in which self-optimizing users may choose their service class in response to capacity modifications. We demonstrate the existence and uniqueness of the equilibrium in which the required ratios are maintained, and address the efficient computation of the optimal capacities. We further provide dynamic schemes for capacity adjustment, and consider the incorporation of pricing and congestion control to enforce absolute performance bounds on top of the proportional ones. Finally, we extend our basic results to networks with general topology.

I. Introduction

A. Background and Motivation

The need for providing service differentiation over the Internet is an ongoing concern in the networking community. The Differentiated Services architecture [2] has been proposed as a scalable solution for QoS provisioning. Instead of reserving resources per session (e.g., as in the Integrated Services (IntServ) model [3]), packets are marked to create a smaller number of packet classes, which offer different service qualities. The premise of differentiated services is to combine simple priority mechanisms at the network core with admission control mechanisms at the network edges only, in order to create diverse end-to-end services.

Several service classes in specific architectures such as Diffserv [2] have been formally defined. For instance, the purpose of the Expedited Forwarding (EF) class [4] is to provide no-loss and delay reduction to its subscribers. The Assured Forwarding (AF) [5] services are intended for users who need reliable forwarding even in times of network congestion. A Service Level Agreement (SLA) is formed between the user and the network provider, in which the user commits to interact with the network in a given way, usually reflected by the allowed bandwidth for each service class. Yet, current technical specifications (e.g., Diffserv standards) deliberately do not quantify the provider part of the agreement, i.e., the actual *service characteristics*, which users will obtain by using the above mentioned classes. Hence, the two elements that jointly determine the QoS in the different service classes, namely, the resource allocation policy (which may be carried out by internal packet scheduling rules) and the regularization of user traffic (e.g., by admission control or pricing) are left as open design issues. Apparently, service characteristics would have to be defined and publicly declared in order to make the distinction between the service classes meaningful to the user and possibly worth paying for.

Differentiated services networks cannot offer strict quality guarantees, as resources are allocated to the service classes based on some average network conditions [6]. Hence, these networks are considered as a "soft QoS" model [7]. The provider may thus declare loose upper bounds on QoS measures, or alternatively provide probabilistic or time-dependent guarantees. Another option, which we consider here, is to announce *relative* quality guarantees, which means that some traffic is simply intended to be treated better than other traffic (faster handling and lower average loss rate). The proportional QoS model [8] has been introduced in order to add concreteness to the notion of relative guarantees. In this model, pre-defined QoS ratios between the classes are to be maintained. These announced ratios are independent of the congestion level of the network. Thus, when a user signs an SLA for a class based on relative performance guarantees, it always gets a concrete performance enhancement over lower service classes. Consequently, the QoS can be easily quantified and advertised as, for example: "service class K provides half the delay of service class $K+1$, at any given time". We note that this proportional QoS can be offered alongside absolute bounds on the relevant performance measure, as we elaborate below (Section V).

In this paper we concentrate on delay-like performance measures, which are formulated through general congestion-dependent cost functions. Delay quality is essential for several modern applications, such as carrying voice over the Internet. Delay ratios are easier to maintain in comparison with absolute end-to-end delay guarantees, primarily because they may hold for different levels of congestion, and secondly because keeping the ratios locally (on a link basis) leads to fulfilling this objective on the network level. Although our focus is on delay, other QoS measures may potentially be included within the proportional QoS framework. We briefly consider in Section II-D the extension of the proportional QoS model to relative packet-loss differentiation.

Dovrolis et al. [8] proposed a class of schedulers, based on the Proportional Delay Differentiation (PDD) model, which aims at providing predetermined delay ratios. The schedulers are implemented by observing the history of the encountered

Both authors are with the Faculty of Electrical Engineering, Technion, Israel. A preliminary version of this work was presented at the WINE'05 workshop, Hong Kong [1].

delays (or alternatively, by measuring the delay of the packet at the head of each service class), and serving the class which most exceeds its nominal delay ratio relative to other classes. In [9], proportional delays are maintained by modifying the weights of a Weighted Fair Queuing (WFQ) scheduler [10] based on predictions of the average delays. Several other schedulers were suggested for obtaining proportional QoS over other congestion measures, separately or simultaneously (see [11] for a survey). These schedulers have been incorporated in various applications such as class provisioning [12] and optical burst switching networks [13].

In the present work we do not impose the delay ratios on a per packet basis (which usually requires time monitoring of queued packets), but rather propose using capacity allocation for this objective. While the capacity allocation model abstracts away the details of packet scheduling, it may be considered an approximated model for existing schedulers such as WFQ. This is further discussed in Section II. We model the differentiation mechanism as a general capacitated link, in which the capacity (e.g., in bits per seconds) assigned to each service class determines its performance. Since users are free to modify their flows subsequent to each capacity allocation, we examine the viability of our framework under a *reactive* user model. The network provider assigns a *capacity manager* in order to maintain the delay-ratios design objective, in face of changing network conditions. We pose the overall model as a *non-cooperative game* between the manager and the network users, and explore the associated capacity management policies and equilibrium conditions.

An important consideration for our model is the time scale at which capacity updates take place. We view capacity management as the means to regulate proportional QoS where service quality is averaged over relatively long time scales, ranging perhaps from minutes to hours. Even longer time scales should be considered for fixed capacity allocation. Accordingly, capacity allocation and updates should be based on average network conditions over time intervals of corresponding duration.

Several previous papers have addressed differentiated services models with reactive users. Perhaps the simplest approach is Odlyzko's Paris Metro Pricing (PMP) proposal [14], where differentiation is induced by assigning a different price to separated service classes. Other papers explicitly consider the connection between the network (social or economic) objective, scheduling mechanism, and the underlying user model. For example, [15] and [16] determine the prices that maximize the provider's profits using a priority queue and a WFQ scheduler, respectively. In [17], the authors focus on incentive prices in priority queues, leading to a socially optimal equilibrium. Our approach differs from the above references, by considering the maintenance of the relative service characteristics as a primary management priority.

Our model allows users to choose between the service classes according to their needs. This choice can be viewed as a selection of a route, hence leading to a *selfish routing* problem. Other papers have considered selfish routing with infinitesimal users, originated in the transportation literature [18] and has been extensively studied since (see [19] for a recent survey). The case of finitely many users, each carrying substantial flow has been introduced to the networking literature more recently (see [20]–[22]). We shall use a similar routing model to represent the user's choice of service class in response to given capacity allocation.

B. Contribution and Organization

Our starting point is a single hop (or single link) network which offers a fixed set of service classes. This model will later serve as a building block for the general network case. It can also be considered as an approximation of a single path in a network, neglecting variations in traffic over intersecting network paths. Under mild assumptions on the delay functions, we first show that there exists a unique capacity assignment which induces the ratio objective for every set of *fixed* flows, and show how it may be efficiently computed. Having established these properties, we extend our analysis of the capacity allocation problem within a reactive user environment. Using a standard flow model (see [23], Sec. 5.4) we represent the user population as a set of self-optimizing decision makers, who may autonomously decide on their flow assignment to each service class. Users differ by their flow utility, their sensitivity to delay and by the way in which they sign up to the network. We show that for every required delay ratios, there exists an equilibrium point, in which the declared ratios are fulfilled; this equilibrium is unique under some further conditions. We provide an efficient computation scheme of the optimal capacity allocation, which can be used when full information of user-specific characteristics is available. However, since in most cases a user model can only be estimated, two alternative reactive schemes are suggested, in which the network manager adapts its capacity based on current per-class conditions only. We provide partial convergence results for the reactive schemes.

The basic model described above is extended in several directions. We address the incorporation of pricing and congestion control mechanisms alongside capacity management. We show that keeping the *total* incoming flow (over all service classes) below a certain level, suffices to ensure upper bounds (which correspond to the same pre-specified ratios) on the delays of *each* service class. This leaves a degree of freedom regarding the regulation of the individual service class flows, which can be exploited to satisfy supplementary network objectives that include profit maximization and fairness. We finally show how proportional QoS can be carried over to general network topologies, by maintaining the specified proportions over each link separately.

The organization of the paper is as follows: The basic network model is described in Section II. We then consider the calculation of the manager's capacity allocation for fixed network flows (Section III). The equilibrium analysis for a reactive user model is given in Section IV. Section V concentrates on pricing and congestion control issues. General network topologies are considered in Section VI. Conclusions and further research directions are outlined in Section VII.

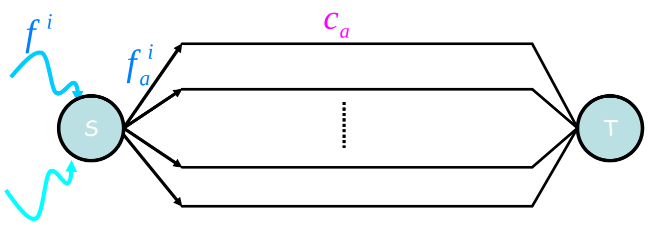

Fig. 1. The single hop network.

II. The Single-Hop Model

A. Network Description

In our single hop model all users employ the same link for shipping their flow. Let $\mathcal{I} = \{1, 2, \ldots, I\}$ be a finite set of users, which share a link that offers a set of service classes $\mathcal{A} = \{1, 2, \ldots, A\}$. We consider the link with its respective service classes as a two terminal (source-destination) network, which is connected by a set of parallel arcs (see Figure 1). Each arc represents a different service class. Thus, the set of arcs is also denoted by $\mathcal{A}$, and the terms service class and arc are used interchangeably. Denote by f_a^i the flow which user i ships on arc a, and by $f_a = \sum_{i \in \mathcal{I}} f_a^i$ the total flow on that arc. The network manager has available a constant link capacity C, to be divided between the service classes. This capacity is to be dynamically assigned in order to address different network conditions. We denote by c_a the allocated capacity at arc a. The capacity allocation of the manager is then the vector $\mathbf{c} = (c_1, \ldots, c_A)$. An allocation $\mathbf{c}$ is feasible if its components obey the nonnegativity and total capacity constraint, namely (i) $c_a \geq 0$, $a \in \mathcal{A}$ and (ii) $\sum_{a \in \mathcal{A}} c_a = C$. The set of all feasible capacity allocations $\mathbf{c}$ is denoted by $\mathbf{\Gamma}$.

B. Latency Functions

Let D_a be the latency (delay) function at service class a. An important example of a latency function is the well known *M/M/1 delay function*, namely

$$D_a(c_a, f_a) = \begin{cases} \frac{1}{c_a - f_a} & f_a < c_a \\ \infty & \text{otherwise} \end{cases} \tag{1}$$

(with the possible addition of a fixed propagation delay, see [23]). More generally, D_a may stand for a general measure of link congestion. We shall consider latency models that comply with the following assumptions.

B1 The delay in each service class a is a function of c_a and f_a only. Namely, $D_a(\mathbf{c}, \mathbf{f}) = D_a(c_a, f_a)$.

B2 D_a is positive, finite and continuous in each of its two arguments for $c_a > f_a$.

B3 $D_a = \infty$ for $c_a < f_a$, and $D_a \to \infty$ for $f_a \to c_a$.

B4 D_a is strictly increasing, convex and continuously differentiable in f_a for $c_a > f_a$.

B5 D_a is strictly decreasing in c_a for $c_a > f_a$.

Assumption B1 implies that the performance in each service class is unaffected by the capacity and traffic intensity in other service classes. This would be the case when separate network resources are allocated at any given time to each service class (e.g., separate queues and wavelengths in a WDM fiber system). In shared multiplexing systems such as WFQ and Weighted Round Robin (WRR) [24], the same assumption may be still applied to approximate the delays under heavy traffic conditions (in which differentiation is mostly crucial) [16]. Assumption B3 induces a strict meaning to the notion of capacity as an upper bound on sustainable flow, which is central to this paper. The monotonicity and continuity properties in Assumptions B2, B4, B5 are natural, while the convexity assumption in B4 is necessary for the analysis and is consistent with common latency models like (1). An additional assumption which will be required for some of our results is:

B6 D_a is a function of $(c_a - f_a)$ only, and is strictly decreasing in $(c_a - f_a)$ over $c_a > f_a$.

Note that the last assumption, together with Assumption B4, implies that if $D_a(c_a, f_a) > D_a(\hat{c}_a, \hat{f}_a)$ then $\frac{\partial D_a(c_a, f_a)}{\partial f_a} > \frac{\partial D_a(\hat{c}_a, \hat{f}_a)}{\partial f_a}$. Consequently, $\frac{\partial D_a}{\partial f_a}$ is uniquely determined by the value of D_a.

C. The manager's objective

We can now precisely formulate the proportional QoS objective. Taking the delay of class 1 as a reference, the ratios are described by a vector $\rho = (\rho_1, \ldots, \rho_A)$, $0 < \rho_a < \infty$, where $\rho_1 \triangleq 1$. The manager's objective is to have the delays $D_1, \ldots, D_A$ satisfy

$$D_a(c_a, f_a) = \rho_a D_1(c_1, f_1) \quad \forall\ a \in \mathcal{A}. \tag{2}$$

We refer to that relation as the *fixed ratio objective*. For concreteness, we shall assume that $\rho_1 \leq \rho_2 \leq \cdots \leq \rho_A$, so that service classes are ordered from best to worst. It will be convenient to define the following cost function for the manager which complies with the fixed ratio objective:

$$J^M(\mathbf{c}, \mathbf{f}) = \begin{cases} 0 & \text{if (2) holds,} \\ \infty & \text{otherwise.} \end{cases} \tag{3}$$

D. QoS Criteria

While the treatment in this paper is geared towards delay as a main QoS measure, in some applications other QoS measures, such as packet-loss or jitter, may be as important. Observe that our model definitions and assumptions throughout this section are not specific to delay. Any congestion measure (or combination of several congestion measures) which can be quantified through an appropriate flow-based model may be considered, as long as (i) the set of assumptions B1-B5 is obeyed for that measure, and (ii) the measure is *additive* over the path links, a property which is required for distributed capacity allocation in general network topologies (see Section VI).

As an example, we address the possible incorporation of packet-loss performance within the proportional QoS framework. Packet losses naturally occur as a consequence of finite buffer space. A *proportional-loss* objective may be considered, where the manager is interested in maintaining predetermined loss-probability ratios between the service classes (perhaps in conjunction with certain delay ratios). An interesting issue here is the meaning of "capacity" in relation to the loss metric. Capacity may stand for service-rate (as in (1)), buffer space, or a combination of them both.

Note that although the loss metric is multiplicative, loss can be approximately treated as additive, assuming low loss rates, which hold under moderate congestion levels. Verifying compliance with Assumptions B1-B5 requires explicit models for loss over the Internet, which are usually unavailable. However, examination of a simplified queueing model, the M/M/1/K queue [25], indicates that when capacity corresponds to service-rate, the average loss complies with most of these assumptions (excluding B3, which is not needed when the loss is not excessive). The detailed study of loss and other performance measures, as well as multiple QoS criteria, is left for future research.

III. Capacity Assignment with Fixed Flows

In this section we consider the network manager's optimal capacity assignment, i.e., an assignment which induces the ratio objective for given network flows. We show the existence and uniqueness of this assignment, and moreover provide the means for its calculation. We exemplify our results by the M/M/1 delay model. The proofs for the results of this section appear in Appendix I.

The analysis of the fixed flow case is significant on its own, as it suggest that the manager is always able to induce the ratio objective, at least for short terms. Additionally, the optimal capacity assignment for fixed flows naturally serves as the manager's best response in the game formulation with reactive users, to be considered in Section IV.

A. Basic Properties

We consider the manager's optimal capacity assignment to a *given* set of service-class flows $f_1, \ldots, f_A$. Henceforth, we will refer to this assignment as the *best response* capacity allocation. The best response here is a capacity allocation which minimizes (3); note that if the minimum is finite, then (2) is satisfied. We show next that the manager has a unique best response, which can be computed by a monotone search over a scalar variable.

Proposition 1: Consider the single-hop model with latency functions obeying Assumptions B1-B5 and a desired ratio vector ρ. Let $(f_1, \ldots, f_A)$ be a fixed flow configuration with $\sum_{a \in \mathcal{A}} f_a < C$. Then (i) there exists a *unique* capacity allocation $\mathbf{c} \in \mathbf{\Gamma}$ such that the ratio objective (2) is met. (ii) This capacity allocation can be obtained by a monotone search procedure over the scalar c_1, which is specified in the proof.

Instead of monotone search, the best response capacity allocation can also be obtained as the solution of a convex optimization problem.

Proposition 2: Under the conditions of Proposition 1, the best response capacity allocation can be obtained by solving the following convex optimization problem:

$$\min_{\mathbf{c}=(c_1,\ldots,c_A)} \hat{J}^M(\mathbf{c},\mathbf{f}), \text{ s.t. } c_a \geq f_a \; \forall a \in \mathcal{A}, \quad \sum_{a=1}^{A} c_a = C,$$

where

$$\hat{J}^M(\mathbf{c},\mathbf{f}) = -\sum_{a=1}^{A} \int_{f_a+\epsilon_a}^{c_a} g\big(\rho_a^{-1} D_a(x_a, f_a)\big) dx_a \tag{4}$$

and $g(\cdot)$ is any continuous and strictly increasing function with $g(0) = 0$, $g(\infty) = \infty$. Further, $\epsilon_a \geq 0$ is a small non-negative constant, which is taken to be strictly positive if $\int_{f_a}^{f_a+\delta} g\big(\rho_a^{-1} D_a(x_a, f_a)\big) dx_a = \infty$ for every $\delta > 0$.

For example, if we take $g(x) = x$ then $\epsilon_a > 0$ is required for the M/M/1 delay function (1), while $\epsilon_a = 0$ can be used for $D_a(c_a, f_a) = \frac{1}{\sqrt{c_a - f_a}}$.

B. Iterative Capacity Assignment

A rather different approach for obtaining the best response allocation during network operation would be to simply update the capacities at each service class based on the observed deviations from the ratio objective. This approach eliminates the need for an explicit calculation of the best response capacity allocation. Define the average normalized delay $\hat{D}(\mathbf{c},\mathbf{f}) \triangleq \frac{1}{A} \sum \rho_a^{-1} D_a(c_a, f_a)$. The required capacities can be obtained by the following update rule:

$$c_a := c_a + \epsilon \alpha_a \big(D_a(c_a, f_a) - \rho_a \hat{D}(\mathbf{c},\mathbf{f})\big). \tag{5}$$

Here $\epsilon > 0$ is a small step-size, and $\alpha_a = \rho_a^{-1}$. Note that the choice of $\{\alpha_a\}$ guarantees that $\sum_a c_a$ is kept fixed, which is required by the fixed capacity constraint.

For ϵ small, the update rule (5) may be approximated by the differential equation

$$\frac{d}{dt} c_a = \alpha_a \big(D_a(c_a, f_a) - \rho_a \hat{D}(\mathbf{c},\mathbf{f})\big). \tag{6}$$

We then have the following convergence result:

Proposition 3: Under the conditions of Proposition 1, the update rule (6) converges asymptotically to the (unique) best response capacity allocation.

Observe that in either approach (namely, a direct or an iterative calculation) the information required by the manager is the *total* flow at each service class. Alternatively, an estimate of the current delay at each service class could be used directly, as these are the required parameters in (4)–(5).

C. M/M/1 Latency functions

The M/M/1 delay model has special significance, as it is frequently used for estimating queuing delays [23]. Clearly, the results obtained so far in this section hold for the M/M/1 delay model case, since it obeys Assumptions B1-B5. Yet, there are some additional distinctive features of this specific model, which we highlight next. Our first result considers the manager's best-response capacity allocation for a given set of per-class flows.

Proposition 4: Consider the single-hop model with M/M/1 latency functions (1) and a desired ratio vector ρ. Then under the conditions of Proposition 1,

$$c_a = f_a + (C - \sum_{\alpha \in \mathcal{A}} f_\alpha) \frac{\rho_a^{-1}}{\sum_{\alpha \in \mathcal{A}} \rho_\alpha^{-1}} \ . \qquad (7)$$

This formula provides an explicit solution for the best response capacity allocation. Note that the excess capacity $(C - \sum_\alpha f_\alpha)$ is divided between the service classes, where each class $a \in \mathcal{A}$ obtains a share which is inversely proportional to ρ_a. Interestingly, a similar expression is obtained for the classical capacity assignment problem of minimizing the network average delay (see [25] Vol. II, p.331).

We next provide a concrete expression for the cost function (4) which is minimized by satisfying the fixed ratio objective.

Proposition 5: Consider the single-hop model with M/M/1 latency functions (1) and a desired ratio vector ρ. Let

$$\bar{J}^M(\mathbf{c}, \mathbf{f}) \triangleq \sum_{a \in \mathcal{A}} w_a D_a(c_a, f_a), \qquad (8)$$

where $w_a = \frac{1}{\rho_a^2}$. Then the (unique) feasible capacity allocation which minimizes (8) is also the best response capacity allocation.

Proof: The function $\bar{J}^M$ is obtained by setting $g(x) = x^2$ in (4). Thus, it follows from Proposition 2 that the manager's best responses for J^M and $\bar{J}^M$ are identical. □

We conclude from the last proposition that by achieving the ratio objective, the manager in fact minimizes a reasonable social cost function, which is just a weighted sum of the delays over the different service classes. Indeed, the weights w_a are inversely proportional in ρ_a^2, which gives higher weight to better, and naturally more expensive, service classes.

IV. Reactive Users

In this section we turn our attention to the case of reactive users, namely, users who modify their flow allocation according to network conditions, and particularly in response to capacity changes. We begin by formulating the user model which leads to a noncooperative game description of the users-manager interaction. We then analyze the properties of the (Nash) equilibrium point of this game. We further consider the convergence of adaptive algorithms for capacity assignment. The proofs of the results in this section are provided in Appendix II.

A. User Model

Recall that we consider a finite set $\mathcal{I}$ of users. Each user i is free to choose its flow f_a^i for each service class $a \in \mathcal{A}$. We allow for pre-specified upper bounds s_a^i on the users flow, so that $0 \le f_a^i \le s_a^i$. The total flow of user i is denoted by $f^i \triangleq \sum_{a \in \mathcal{A}} f_a^i$, and its flow configuration is the vector $\mathbf{f}^i = (f_1^i, \dots, f_A^i)$. The flow configuration $\mathbf{f}$ is the vector of all user flow configurations, $\mathbf{f} = (\mathbf{f}^1, \dots, \mathbf{f}^I)$. A user flow configuration $\mathbf{f}^i$ is *feasible* if its components obey the flow constraints as described above. We denote by $\mathbf{F}^i$ the set of all feasible user flow configurations $\mathbf{f}^i$, and by $\mathbf{F}$ the set of all feasible flow configurations $\mathbf{f}$. Finally, a system configuration $(\mathbf{c}, \mathbf{f})$ is feasible if it consists of a feasible flow configuration and a feasible capacity allocation (defined in Section II-A).

Users are distinguished first by their flow utility function $U^i(f^i)$, which quantifies their subjective utility for shipping a total flow f^i. Thus, we accommodate users with *elastic* flow demand. We make the following assumptions regarding U^i: *For every user $i \in \mathcal{I}$, the utility function $U^i : \Re \to \Re$ is bounded above, concave and continuously differentiable.* We note that utility functions with the above characteristics are commonly used within the networking pricing literature [6], [26]. The total cost J^i for user i is given by

$$J^i(\mathbf{c}, \mathbf{f}) = \beta^i \sum_{a=1}^{A} f_a^i D_a(c_a, f_a) + \sum_{a=1}^{A} f_a^i p_a^i - U^i(f^i). \qquad (9)$$

The left term of J^i represents the delay cost, which is the total delay of the user, multiplied by its delay sensitivity $\beta^i > 0$. The middle term stands for the network usage price, where we assume linear tariffs [6], i.e., p_a^i is the price per unit flow of user i in class a.

The above cost function allows to treat different types of network users in a unified mathematical framework. This includes:

1) Elastic or Plastic users. An elastic user's total flow is generally not constant, and varies according to the network conditions. A plastic user is interested in shipping a fixed amount of total flow into the network. Such a user can be modeled by a flow utility function which has a sharp maximum at the required total rate.

2) Static SLA users. Static SLAs are typically negotiated on a regular (e.g., monthly or yearly) basis. The agreement means that the users can start data transmission (subject to the rates they buy) whenever they wish without signaling their Internet service providers [3]. Thus, from user i's point of view, static SLAs are manifested by the maximal flow rates s_a^i in each service class. In this paper we do not consider the establishment phase of static SLAs, and therefore their associated prices are irrelevant to our analysis. Accordingly, we have $p_a^i = 0$ for every static SLA user, since payment was already transferred for acquiring each s_a^i.

3) Dynamic SLA users. Dynamic SLA users buy differentiated services on-demand, meaning that they pay a price per unit traffic p_a^i over each service class. In a reasonable pricing model, these prices could be identical for all users, that is $p_a^i = p_a$ for every $i \in \mathcal{I}$. As these users are not limited by static SLAs, we can set $s_a^i = M$, where $M > C$ is an arbitrary large constant.

The prices of both static or dynamic SLAs may be viewed as an indirect means for congestion control [6], and (among other things) prevent flooding of the premium service classes. In this paper, however, we concentrate on capacity assignment as the management tool, assuming that prices are static (or change on a slower time scale). The issue of price setting in our context is addressed later in Section V[1].

Remark 1: Strictly plastic users, i.e., users $i \in \mathcal{I}$ with a constant rate f^i, require the additional constraint $\sum_a f_a^i =$

[1] One would expect that better service classes (with a lower delay) would be more expensive, although this is not required for our derivations.

f^i. For simplicity, we will not explicitly consider the case of strictly plastic users, yet all the results in this section still hold if strictly plastic users are incorporated in our model (provided their total flow requirement $\sum f^i$ is less than the total capacity C).

B. The Game Formulation

Having defined cost functions for all parties involved, the interaction between the manager and the users may now be considered as a non-cooperative game, and will be referred to as the *users-manager game*. Note that the manager in our case is not adversarial to the users, but simply wishes to impose its ratio objective. A Nash Equilibrium Point (NEP) of our game is a feasible system configuration $(\tilde{\mathbf{c}}, \tilde{\mathbf{f}})$ such that all costs (J^M, J^i, $i \in \mathcal{I}$) are finite, and the following conditions hold:

$$\begin{aligned} J^M(\tilde{\mathbf{c}}, \tilde{\mathbf{f}}) &= \min_{\mathbf{c} \in \Gamma} J^M(\mathbf{c}, \tilde{\mathbf{f}}), \qquad (10)\\ J^i(\tilde{\mathbf{c}}, \tilde{\mathbf{f}}^i, \tilde{\mathbf{f}}^{-i}) &= \min_{\mathbf{f}^i \in \mathbf{F}^i} J^i(\tilde{\mathbf{c}}, \mathbf{f}^i, \tilde{\mathbf{f}}^{-i}) \quad \text{for every } i \in \mathcal{I} \end{aligned}$$

where $\tilde{\mathbf{f}}^{-i}$ stands for the flow configurations of all users except for the ith one. Namely, the NEP is a network operating point which is stable in the sense that neither any user, nor the manager, finds it beneficial to unilaterally change its flow or capacity allocation, respectively.

Before analyzing the overall users-manager game, let us consider the users reaction to a *fixed* capacity allocation $\mathbf{c} = (c_1, \ldots, c_A)$. In this case we still have a non-cooperative game between the I users. The definition of the Nash equilibrium of this game remains as in (10), but excluding the network manager's minimization of J^M. The next result establishes the uniqueness of the user-equilibrium flow of that game. We note that this result is particular to the single hop (parallel arc) case treated here, and does not fully extend to the general network case treated in Section VI.

Proposition 6: Consider the single-hop model with latency functions obeying Assumptions B1-B5. Then for every capacity allocation $\mathbf{c}$, the resulting user-equilibrium flow configuration is unique.

C. Analysis of the Users-Manager Equilibrium

We now return to the complete game model, where the users react to the network congestion conditions, while the capacity manager is concerned with keeping the delay ratios and modifies the capacity allocation accordingly. Our next result establishes the existence of an equilibrium point, in which the desired ratios are met.

Theorem 7: Consider the single-hop model with latency functions obeying Assumptions B1-B5 and a desired ratio vector ρ. Then there exists a Nash equilibrium point for the users-manager game. Any such NEP exhibits finite costs for both the manager and the users. In particular, the ratio objective (2) is satisfied.

The additional Assumption B6 on the latency function is required to establish the *uniqueness* of the NEP.

Theorem 8: Consider a single-hop network with latency functions obeying Assumptions B1-B6. The Nash equilibrium point for the users-manager game in this network is unique.

Besides existence and uniqueness, an additional appealing feature of our framework is the computational complexity of calculating the equilibrium point. Generally, equilibrium computation schemes for selfish routing even for a network with fixed capacities are quite involved (see [27] for a survey). Moreover, most schemes apply under certain conditions, which need not hold in our case. However, in our framework which includes capacity adjustment according to a pre-specified delay ratio, the calculation becomes tractable. In the sequel we provide the means for the explicit calculation of an equilibrium point in the users-manager game. Let us start by assuming that one of the equilibrium delays, say D_1, is given. Our next lemma shows that in this case, the equilibrium point can be efficiently calculated via a set of quadratic problems.

Lemma 1: Consider the single-hop model with latency functions obeying Assumptions B1-B6. Assume that the delay D_1 at the NEP is given. Let $\{D_a = \rho_a D_1\}$ and $\{D'_a \triangleq \frac{\partial D_a}{\partial f_a}\}$ be the equilibrium delays and their derivatives, as determined by D_1. Then the flows at the NEP can be calculated by solving the following I quadratic optimization problems (with A variables each), one for each user $i \in \mathcal{I}$:

$$\min_{\mathbf{f}^i} \Big\{ \sum_a \frac{1}{2} \beta^i D'_a {f^i_a}^2 + f^i_a \left(\beta^i D_a + p^i_a\right) - U^i\big(\sum_a f^i_a\big) \Big\}$$
$$\text{subject to}: \quad 0 \le f^i_a \le s^i_a. \qquad (11)$$

Once the equilibrium flows at each service class are determined, it is a simple matter to calculate the equilibrium capacities according to the delay formulas. Since the number of classes in a Diffserv-like network is small, the calculation procedure (11) is computationally manageable. The only issue that needs to be resolved for a complete calculation scheme is the determination of D_1 at equilibrium. We next show that D_1 can be obtained as the fixed point of a monotone map. Let D_1 be an *estimate* of the equilibrium delay. Solving (11) for each user yields (aggregate) flows $\{f_a\}$ which in turn can be used in (4) for obtaining the manager's best response. This best response, together with the flows $\{f_a\}$ yield service class delays which meet the ratio objective. We denote these delays by $\tilde{D}_a(D_1)$, emphasizing that they are a function of the original estimation of D_1. Our next result shows that the equilibrium delay of class 1 can be obtained via an efficient iterative procedure.

Proposition 9: Consider a network with latency functions obeying Assumptions B1-B6. Then,

1) $\tilde{D}_1(D_1)$ is monotonously decreasing in D_1.
2) The equilibrium delay D_1 is the unique solution of the equation $D_1 - \tilde{D}_1(D_1) = 0$.
3) The equilibrium delay of class 1 may thus be obtained by a monotone search over the scalar D_1, where each stage involves the calculation of $\tilde{D}_1(D_1)$ through the solution of (4) and (11).

D. Adaptive Algorithms for Capacity Assignment

Adaptive algorithms are required to account for the reactive and non-stationary nature of the network users. In this section we consider two plausible options for adaptive capacity management, which are based on the fixed-flow analysis of Section III. Propositions 1 and 2 allow the manager to directly calculate its *best response* assignment, namely the capacity assignment that will satisfy the fixed ratio objective given the *current* network flows. This forms the basis for our first algorithm.

Algorithm 1: The network manager periodically observes the current network flows over each service class, and modifies its capacity allocation to its best response capacity allocation (described in Propositions 1–2).

An alternative approach is based on the iterative update rule (5).

Algorithm 2: The network manager periodically observes the current network flows, and adapts its capacity allocation according to (5).

Algorithm 1 is essentially designed to reach the ratio objective within a small number of capacity re-allocations. Indeed, if the users are not reactive so that network flows are fixed, the network manager would satisfy its objective within a single capacity modification. This algorithm inherently incorporates substantial capacity changes in each step, as the current best response capacity allocation may correspond to a significantly different flow distribution in comparison with the last capacity allocation. Algorithm 2 represents a different approach, in which capacity modifications are simple and carried out gradually, in small steps. Sudden changes in capacity can thus be avoided, possibly at the cost of slower convergence toward the required equilibrium. An attractive hybrid scheme may be considered, where Algorithm 1 makes the initial capacity updates, followed by Algorithm 2 which smoothly adapts the capacities towards the required equilibrium point.

We next address some convergence properties of both algorithms under a reactive user environment. To start with, we assume that the users flow configuration reaches the user-equilibrium point (unique by Proposition 6), and moreover that the network manager adapts its previous capacity allocation only *after* the users flow configuration is at equilibrium. It can then be shown that both algorithms converge to the Nash equilibrium point in the case of *two service classes*. A precise statement of these results and the (somewhat lengthy) proofs are omitted here for lack of space and can be found in [28]. We point out that the analysis in [28] relies heavily on monotonicity properties and is not readily extendible to more than two service classes.

Convergence to the user-equilibrium point under best-response (or similar) dynamics which is required for our analysis is essentially an open problem which is resolved in simplified scenarios only (see [20], [29], [30]). However, assuming that users reach a stationary working point approximates a reasonable scenario, where the network manager operates on a slower time scale than that of the users. In this section we have obviously ignored transient conditions, such as changes in the user population and their flow requirements. Hence, the above results should not be considered as ensuring the convergence of the suggested iterative algorithms under general conditions, but rather as an indication for their viability within a more stable environment.

E. Discussion

We pause here to discuss some consequences of the previous results, as well as some aspects of our modeling assumptions. Our central result is Theorem 7 (existence of a Nash equilibrium point), which implies that the ratio objective is feasible for any congestion level. This suggests that the network has a stable operating point which satisfies the ratio objective even when users are reactive and modify the flow and service class selection according to perceived congestion.

Theorem 8 (uniqueness of a Nash equilibrium point) implies that there exists only one such operating point which is stable under unilateral deviations of self-optimizing users. We note that the uniqueness property requires the additional Assumption B6, which is not needed for existence of the equilibrium. Generally, when the equilibrium is not unique, the network behavior becomes less predictable. Simulation results or computation of the equilibrium cannot be relied on to give a complete picture of the network operation. However, the possible existence of multiple equilibria in our case can be tolerated, at least with respect to the network manager's objective, since the required ratios are met in every equilibrium point.

From an analytical perspective, the computation of the equilibrium point (Lemma 1 and Proposition 9) scales well (linearly) with number of users, unlike general non-cooperative games, in which computation is hard from three players and above (see [31] for a recent survey). From the operational perspective, the ability to compute the equilibrium capacities suggests that the manager can set the equilibrium capacities just once, and wait for the users to reach the equilibrium flows. In terms of game theory, this approach is related to a Stackelberg game [32], where the leading player (in our case the network manager) announces its strategy first, and the other players react to this strategy. Proposition 6 ensures that the resulting equilibrium flows are unique, thus the ones expected by the manager. Note that in our specific game the Stackelberg strategy leads to an equilibrium point which coincides with the (unique) NEP desired by the network manager. This property essentially follows from the structure of the manager's cost (3) which assumes only two values (zero or infinity): indeed, in every NEP the manager's cost must be zero, and this cannot be improved upon even when the manager acts as a leader.

We emphasize that the explicit computation of the equilibrium capacities requires the manager to possess considerable per-user information, including user preferences, which can only be estimated. Accordingly, the use of adaptive algorithms seems more practical in our case. Still, rough estimates of the user preferences can be used for estimating the equilibrium capacities, which in turn may serve as a good starting point for an adaptive algorithm.

V. Pricing and Congestion Control

In the previous sections we have established that, under our assumptions on the latency functions, proportional QoS can be maintained at any congestion level. Still, excessive congestion should obviously be avoided. We briefly consider in this section two alternatives for regulating the flow level of the network, namely pricing and congestion control. The underlying objective in either case is to guarantee an upper-bound $D_a^{\max}$ on the average delay at each service class a, while still keeping the ratio objective, thus

$$D_a^{\max} = \rho_a D_1^{\max}, \quad a \in \mathcal{A}. \tag{12}$$

Accordingly, both pricing and congestion control are assumed to take place alongside capacity management, which is still responsible for keeping the delay ratios. Furthermore, in the differentiated services context, pricing and congestion control can be considered as complementary, in the sense that pricing operates on a slower time scale than congestion control. Prices are expected to vary slowly (see, e.g., [6] p. 256), to give the users enough time to evaluate their price-quality tradeoff and decide which service class to join. Thus, price regulation is carried out by taking into account *average network conditions*. Congestion control mechanisms, on the other hand, should discard excess flow (or block user access) when momentary performance is not adequate. We will demonstrate here that the ratio objective, which is enforced by capacity management, fits well with both pricing and congestion control. In a sense, their implementation in a shared-resources multi-class network can become easier, when performed in concert with the ratio maintenance.

The user cost model (9) of the previous sections already includes pricing at linear tariffs. We will not fully address here the price setting issue, which is a broad research area in communication networks (e.g., [6], [26], [33] and references therein). We do provide a qualitative monotonicity result regarding the effect of prices on equilibrium delays, which should be significant to the implementation of any pricing scheme. Our focus is on the dynamic SLAs framework (see Section II), where price is identical for all users, i.e., $p_a^i = p_a$.

Theorem 10: Consider the single-hop model, where the delay functions obey Assumptions B1-B6, and two pricing vectors $\mathbf{p} = (p_1, \ldots, p_A)$ and $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_A)$. If $\tilde{\mathbf{p}} \geq \mathbf{p}$ (i.e., $\tilde{p}_a \geq p_a$ for all $a \in \mathcal{A}$) then (i) $\tilde{D}_a \leq D_a$ at equilibrium for *every* service class $a \in \mathcal{A}$. (ii) If, moreover, $\tilde{p}_a > p_a$ and $0 < f_a^i < s_a^i$ for some $i \in \mathcal{I}$ and $a \in \mathcal{A}$, then $\tilde{D}_a < D_a$ at equilibrium for every $a \in \mathcal{A}$.

Proof: (outline) It can be shown that all users submit less or equal total flow when prices are higher, thus delays are lower. The full details can be found in [28]. □

The result above indicates that increasing the prices for any subset of service classes results in reduced delays at *all* service classes. This is of course a consequence of capacity re-allocation that takes place to maintain the delay ratios. The theorem above leaves a degree of freedom as to which of the prices should be modified in order to satisfy the required upper bounds (12) on $\{D_a\}$. This should be determined by other (economic) pricing objectives along with the required delay ratios.

We now turn our attention to active congestion control. In a shared-resource multi-class environment, the question of congestion control becomes multidimensional, as removing traffic from one service class also affects the others. Our goal here is to provide guidelines for determining the target flow levels in each service class at congestion periods. We emphasize that the short time scale on which congestion control must operate and react, necessarily leads us to consider its effect relative to the current flow demands. This stands in contrast to the above analysis of pricing, which considers its effect on *equilibrium* flows.

The next central result fully characterizes the admissible region of service-class flows, which allow to maintain the required delay bounds.

Theorem 11: Consider the single-hop model, where the delay functions obey Assumptions B1-B6. Assuming that capacities are set according to (4), there exists a critical flow level $f^{\max}$ so that $D_a \leq D_a^{\max}$ holds for every class if and only if $\sum_{a\in\mathcal{A}} f_a \leq f^{\max}$.

Proof: The claim immediately follows from the following fact: Let $\mathbf{f}$ and $\hat{\mathbf{f}}$ be two fixed flow vectors. If $\sum_a f_a \leq \sum_a \hat{f}_a$, then the respective best response capacity allocations yield class delays which satisfy $D_a \leq \hat{D}_a$ for every $a \in \mathcal{A}$. To prove this, Assume by contradiction that $D_a > \hat{D}_a$ for some a (hence for every a). Then by Assumption B6, $c_a - f_a < \hat{c}_a - \hat{f}_a$. Summing this inequality over all service classes, and noting that the total capacity is fixed, we obtain that $\sum_a f_a > \sum_a \hat{f}_a$, which is a contradiction. □

The significance of this result is threefold. First, as a consequence of the underlying capacity management, the admissible set of flows $\sum_{a\in\mathcal{A}} f_a \leq f^{\max}$ is simple and requires to regulate the total flow only. Second, the set of feasible flows is naturally expanded (as compared with maintaining the required delay bound at each service class without capacity sharing). Third, the network faces a degree of freedom in setting the target flow levels in each class, which could be exploited to promote diverse objectives. We outline here two such options:

1. *Profit maximization.* Assume that users pay for their goodput only (i.e., they do not pay for their discarded flow). Keeping that in mind, the network could be interested in maximizing its profits by discarding the excess flow from cheaper service classes (recall that class prices are *fixed* during short time scales which are considered here). However, the network should usually keep an adequate flow rate at each service class (e.g., due to static SLA commitments). To formalize this tradeoff, denote by f_a^d the current total user demand for class a (without congestion control) and by f_a^0 the rate for class a which the network must allow. Whenever congestion control is called upon, the allowed input rates can be obtained from the following optimization problem.

$$\max_{f_1,\ldots,f_A} \Big\{ \sum_a f_a p_a \Big\} \tag{13}$$
$$\text{s.t. } \sum_a f_a = f^{\max}, \quad \min\{f_a^d, f_a^0\} \leq f_a \leq f_a^d \ \forall a \in \mathcal{A}.$$

The solution to this optimization problem follows easily by ordering the service classes in increasing price order, and discarding flow according to this order (while obeying the f_a^0 constraint) until the total flow reaches $f^{\max}$.

2. *Fairness.* Consider the following network-wide performance criterion $\prod_a \frac{f_a}{D_a}$, which is known as the product form of the user's powers (rate over delay) [34]. Maximizing this criterion captures a natural tradeoff between class utilization and delay.

Noting that the target delays are known under congestion, the target flow levels could be chosen as the solution to the following optimization problem

$$\max_{f_1,\ldots,f_A} \Big\{ \prod_a \frac{f_a}{D_a^{\max}} \Big\}, \tag{14}$$

subject to the same constraints as in (13). This is in fact a geometric program ([35], p.160), which can be easily converted into a convex optimization problem, and consequently solved efficiently. The maximizer of (14) is known to obey certain fairness properties among the service classes, which coincide with the Nash bargaining solution. The properties of the bargaining solution and their association with fairness are summarized in [34], [36].

We emphasize that as long as the admitted flow does not exceed $f^{\max}$, any other alternative for excess flow removal would ensure compliance with the upper bound delays. We conclude this section with a brief summary of implementation issues related to the removal of excess flow from overloaded service classes. Perhaps the simplest way to attain the required flow levels in each service class is through call admission control (CAC), i.e., denying service from some users which access an overloaded service class. In this context, denying service from dynamic SLA users (who contract with the network through short term agreements) seems more natural; static SLA users, on the other hand, usually cannot be denied, unless their contracts explicitly allow that. The important issue of determining which specific users to reject is beyond the scope of this discussion.

Note that the total user demand f_a^d in each class, which is required in (13) and (14), is comprised of the currently admitted flow f_a, and the sum of user requests for obtaining class a's service (e.g., through dynamic SLAs). In practice, the first quantity can be assessed through edge router measurements of the actual flow-rates at each service class. The latter quantity is available, for example, at the Bandwidth Broker (BB) entity, part of the Diffserv architecture [3].

VI. General Network Topologies

In this section we extend our results to the general network case, in which every user has its own source and destination, and a given unique route which leads from that source to the destination[2]. We assume that the route is predetermined by some routing protocol, and is not a part of the user decisions. Note that since users choose and maintain a service class on an end-to-end basis, we cannot reduce the network case to an independent game over each link. Thus, those results obtained for the single hop model which involve user choice of service class need not carry over to general network topologies.

A. Model Definition

We consider a network of general topology which consists of a set of links $\mathcal{L} = \{1, \ldots, L\}$. As before, let $\mathcal{I} = \{1, 2, \ldots, I\}$ be the set of users, which share the network.

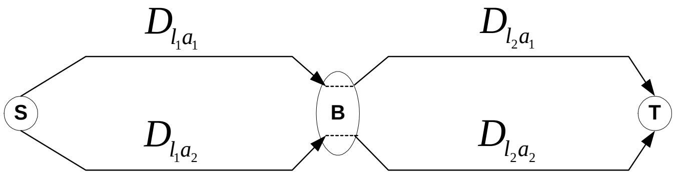

Fig. 2. An example of a general topology network. This network has a source S, an intermediate node B and a destination T. There are two links, $l_1 = SB$ and $l_2 = BT$. In this network $A = 2$, thus each link contains two parallel arcs. We emphasize in the drawing that flows do not switch from one service class to the other within a path.

We associate with each user i a route $R^i = (l^{i,1}, \ldots, l^{i,N^i})$, where N^i is the length of the route and $l^{i,k}$ is the kth link traversed by user i. Let $\mathcal{I}_l \subset \{1, 2, \ldots, I\}$ be the set of users for which $l \in R^i$. Each link l carries the set of service classes $\mathcal{A} = \{1, 2, \ldots, A\}$. As before, each link is thus represented by a set of A parallel arcs (see Figure 2). Additionally, each link l has a total capacity C_l, which is to be divided between its service classes. Denoting by c_{la} the capacity assigned to class a in link l, a capacity assignment is feasible as long as (i) $c_{la} \geq 0$ $\forall l, a$ and (ii) $\sum_{a \in \mathcal{A}} c_{la} = C_l$. Denote by $\mathbf{c}_l = (c_{l1}, \ldots, c_{lA})$ the capacity allocation vector for link l. Let Γ_l denote the set of all feasible capacity allocations for link l, and let $\Gamma = \Gamma_1 \times \ldots \times \Gamma_L$ be the set of all feasible network capacity allocations $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_L)$ (where $\mathbf{c}$ can be viewed as matrix of dimension $L \times A$).

As user routes are fixed, the user's only decision is how to set its flow rate in each of the service classes. Denoting by $f_a^{i,k}$ the flow assigned to the kth link of user i in service class a, we have $f_a^i = f_a^{i,1} = \ldots = f_a^{i,N^i}$; namely, once the user has determined the inter-class flow distribution, it remains fixed along the entire path. We emphasize that each user i can adjust only the flow rates f_a^i on its entire path, but *cannot* adjust the flow rates separately on each individual link thereof. Using the same notations as in the single hop case, a feasible flow configuration $\mathbf{f}^i$ further obeys $0 \leq f_a^i \leq s_a^i$ for every $a \in \mathcal{A}$. We adopt some additional notations from the single hop case, namely $\mathbf{F}^i$, $\mathbf{f}$ and $\mathbf{F}$, the definition of which is given in Section IV-A. Finally, a system configuration is feasible if it is composed of feasible flow configurations and feasible capacity allocations.

Turning our attention to some link $l \in \mathcal{L}$, let f_{la} be the total flow in link l which is assigned to service class a, i.e., $f_{la} = \sum_{i \in \mathcal{I}_l} f_a^i$. The delay of service class a at link l is denoted by D_{la}. We adopt the same assumptions as in Section II regarding the delay functions; thus D_{la} is a function of c_{la} and f_{la} only. Let D_a^i be the end-to-end delay of user i in service class a, namely $D_a^i = \sum_{l \in R^i} D_{la}(c_{la}, f_{la})$. The cost function of each user $i \in \mathcal{I}$ is then given by

$$J^i(\mathbf{c}, \mathbf{f}) = \beta^i \sum_{a=1}^{A} f_a^i D_a^i + \sum_{a=1}^{A} f_a^i p_a^i - U^i(f^i). \tag{15}$$

The objective of the network remains to impose predetermined ratios between the delays of the service classes. Formally, given a ratio vector ρ, the network's goal is to have the delays

[2] An extension to this model, where each user has multiple destinations, will not affect the results of this section (excluding Theorem 16). For simplicity of exposition, we focus here on the single user-path case.

D_a^i, $a \in \mathcal{A}$, $i \in \mathcal{I}$ obey

$$D_a^i = \rho_a D_1^i. \tag{16}$$

As in the single-hop case, we could now proceed to define the Nash equilibrium of the users-manager game, based on the above ratio objective. However, in the network context the relation (16) can in general be realized in many different ways, as it only specifies a requirement on the end-to-end delay, but not on specific link delays. Indeed, the following example demonstrates that there could be more than a single capacity allocation which meets the required delay ratios for a fixed flow configuration, using a simple scenario.

Example 1: Consider the network in Figure 2, where the delay in each arc is given by the M/M/1 formula (1). Assume that the network's objective is that the end-to-end delay in service class a_2 would be twice larger than that of service class a_1. The link capacities are $C_{l_1} = C_{l_2} = 11$. Further assume a fixed flow configuration $f_{la_1} = f_{la_2} = 4$, $l = \{l_1, l_2\}$. The capacity allocation $c_{l_1a_1} = c_{l_2a_1} = 6$, $c_{l_1a_2} = c_{l_2a_2} = 5$ maintains the required ratios in every link, and thus end-to-end. A different capacity allocation which will comply with the same delay ratio is $c_{l_1a_1} = c_{l_1a_2} = 5.5$, $c_{l_2a_1} = 6.342$, $c_{l_2a_2} = 4.658$. Essentially, infinitely many capacity allocation which meet the required ratio of $1:2$ may be suggested, by inducing an arbitrary finite ratio at the first link, and then compensating for it on the second link.

We advocate here a natural link-level scheme, in which capacity adaptation is performed independently at each link with the objective of locally preserving the delay ratios. Formally, this objective is given by

$$D_{la}(c_{la}, f_{la}) = \rho_a D_{l1}(c_{l1}, f_{l1}), \tag{17}$$

for every l and a. Obviously, (17) is sufficient for maintaining the network's objective (16). This link-level approach is attractive from a practical point of view, as capacity assignment can be implemented in a distributed manner, and it does not require links to communicate their current status. For a game-theoretic formulation we assign to each link l its own capacity manager M_l, equipped with capacity C_l. Accordingly, a feasible system configuration $(\tilde{\mathbf{c}}, \tilde{\mathbf{f}})$ is a NEP if the following conditions hold:

$$\begin{aligned} J^{M_l}(\tilde{\mathbf{c}}_l, \tilde{\mathbf{c}}_{-l}, \tilde{\mathbf{f}}) &= \min_{\mathbf{c}_l \in \Gamma_l} J^{M_l}(\mathbf{c}_l, \tilde{\mathbf{c}}_{-l}, \tilde{\mathbf{f}}), \ l \in \mathcal{L}, \\ J^i(\tilde{\mathbf{c}}, \tilde{\mathbf{f}}^i, \tilde{\mathbf{f}}^{-i}) &= \min_{\mathbf{f}^i \in \mathbf{F}^i} J^i(\tilde{\mathbf{c}}, \mathbf{f}^i, \tilde{\mathbf{f}}^{-i}), \ i \in \mathcal{I}, \end{aligned} \tag{18}$$

where

$$J^{M_l}(\mathbf{c}, \mathbf{f}) = \begin{cases} 0 & \text{if } D_{la} = \rho_a D_{l1}, \\ \infty & \text{otherwise.} \end{cases} \tag{19}$$

Note that if each manager's cost is finite, then the ratio objective (16) is maintained. We will refer to this model as the (link-based) *users-managers* game.

B. Main Results

We next present our main results for general topology networks. The longer proofs are deferred to Appendix III. Our focus in this section is on the link-level approach which maintains the delay ratios on a link basis. A specific benefit of this approach is that each link manager can apply its local best response map, based on the same methods that were applied for the single hop case. This is formalized in the following proposition.

Proposition 12: Consider the general network model with latency functions obeying Assumptions B1-B5. Let $\mathbf{f}$ be a fixed flow configuration with $\sum_{a \in \mathcal{A}} f_{la} < C_l$, $l \in \mathcal{L}$. Then there exists a *unique* capacity allocation $\mathbf{c} \in \mathbf{\Gamma}$ such that the delay ratios are met locally in every link. This capacity allocation can be obtained for each link independently using the results of Propositions 1–2.

Proof: The proof follows directly from the proofs of Propositions 1–2. Since the best response in each link is unique, it follows that there is a unique best response at the network level, where the capacity assignment is separately calculated in every link. □

Similarly, the iterative capacity assignment (5) can be applied separately in each link for obtaining the required network capacity allocation. The convergence of this scheme (for small ϵ) follows directly from Proposition 3. Focusing on the M/M/1 delay model, we may still interpret the fixed ratio objective as a social objective as in Proposition 5.

Proposition 13: Consider a general topology network with M/M/1 latency functions (1) and a desired ratio vector ρ. Let

$$\bar{J}^{R^i}(\mathbf{c}, \mathbf{f}) \triangleq \sum_{a \in \mathcal{A}} w_a D_a^i, \tag{20}$$

where $w_a = \frac{1}{\rho_a^2}$. Then the (unique) feasible capacity allocation which minimizes (20) for every $i \in \mathcal{I}$ is also the best response capacity allocation.

Proof: Immediate from Proposition 5 by noting that $\min \sum_{a \in \mathcal{A}} w_a D_a^i = \sum_{l \in R^i} \min w_a D_{la}(c_{la}, f_{la})$. □

The significance of the last result is that under the best response capacity allocation, the weighted sum of the end-to-end delays is minimized for *each* user path.

Returning to the reactive user model, we show next that an equilibrium which maintains the delay ratios on a per-link basis (as defined in (18)) always exists.

Theorem 14: Consider a general network with latency functions obeying Assumptions B1-B5. Then there exists an equilibrium point for the link-based users-managers game at which the delay ratios are met.

The existence of this Nash equilibrium is a basic indication for the viability of the proportional QoS approach in general network topologies. We next consider a special case, where all users are strictly plastic (i.e., users whose total demand f^i is fixed, see Remark 1). As we show in the next theorem, the equilibrium point of the users-managers game in this case is unique, and further computable via quadratic optimization problems, whose complexity remains the same as in the single-hop case.

Theorem 15: Consider a general network with latency functions obeying Assumptions B1-B6. Assume that all users are strictly plastic, i.e., the total demand f^i is constant for every $i \in \mathcal{I}$. In addition, assume that the users' flow can be accommodated by the network, i.e., $\sum_{i \in \mathcal{I}_l} f^i < C_l$ for every $l \in \mathcal{L}$. Then (i) there exists a *unique* equilibrium point of the link-based users-managers game. (ii) the equilibrium can then be efficiently calculated by solving I quadratic problems with A variables each.

Similar results to Theorem 15 regarding uniqueness of the equilibrium are not currently available for general elastic users. That is, we cannot rule out the existence of two equilibrium points, each with a different set of capacity allocations and user flow demands. Nonetheless, in the following theorem we provide sufficient conditions, under which uniqueness of the equilibrium does hold for general users.

Theorem 16: Consider a general network with latency functions obeying Assumptions B1-B6. Uniqueness of the equilibrium point for the users-managers game holds when the user paths satisfy either one of the following: (a) Every link is shared by either all users, or by one user at most. (b) Every link is shared by at most two users.

The incorporation of pricing and congestion control mechanisms for general network topologies is naturally more involved than in the single hop case. For instance, maintaining end-to-end performance guarantees for each user (which are fair in some sense) becomes a considerable issue. In this context, possible solution concepts may combine complementary routing mechanisms (e.g., constraint based routing [3]) alongside pricing and congestion control mechanisms, which could reduce congestion at overloaded links.

VII. Conclusion

This paper considered an approach for capacity allocation in differentiated services networks which focuses on maintaining a fixed proportion of certain congestion measures across the different service classes. The congestion model we consider incorporates fairly general delay functions at each service class, and furthermore takes into account a reactive and heterogenous user environment. An attractive feature of the suggested capacity allocation schemes is the ability to implement per-link distributed algorithms, alongside an efficient computation procedure for the required capacity assignment. We have also shown how the proposed ratio objective fits seamlessly with congestion control and pricing mechanisms, which may be invoked to ensure delay bounds at each service class.

We have presented a comprehensive analysis of the single-hop case, and partially extended these results to a general network topology. Some specific issues that remain for further study within the general network model include: (1) more general conditions for the uniqueness of the equilibrium, (2) convergence properties of distributed capacity management schemes with reactive users, and (3) the incorporation of pricing and congestion control mechanisms on an end-to-end basis.

The scope of our model may be enhanced in several respects. We purpose to examine coupled latency functions, which allow some dependence in performance of each service class on the congestion level at other classes. These latency functions may provide more accurate models for common scheduling schemes such as WFQ and WRR. The simultaneous consideration of several QoS measures is of obvious interest. Finally, an important future direction would be to consider routing alongside capacity assignment as a complementary mechanism which balances the traffic in the network.

Appendix I
Proofs for Section III

Proof of Proposition 1: The proof of existence and uniqueness of the manager's best response rests on monotonicity properties of the best-response capacities.

Define the mapping $c_1 \mapsto T(c_1) \in \Re^+$ as follows: for each $c_1 > f_1$ (which induces a unique delay $D_1(c_1, f_1)$), set the remaining service class capacities $c_2, \ldots, c_A$ so that the required delay ratios $D_a(c_a, f_a) = \rho_a D_1(c_1, f_1)$ are met for every $a = 2, \ldots, A$. Note that c_a is uniquely determined due to the monotonicity of D_a in c_a (Assumption B5). We define $T(c_1)$ as the sum of these capacities (including c_1), i.e., $T(c_1) = \sum_{a \in A} c_a$. Note that $T(c_1)$ need not be equal to C, as the total capacity constraint is not enforced here. It follows that $(c_1, \ldots, c_A)$ is a best response to $(f_1, \ldots, f_A)$ if and only if $f_1 \leq c_1 \leq C$ and $T(c_1) = C$. To establish uniqueness, it remains to show that there is a unique c_1 with these properties. For that purpose, the following observation is required.

Lemma 2: The mapping $c_1 \mapsto T(c_1)$ is strictly increasing and continuous in c_1.

Proof: Immediate by the continuity and strict monotonicity of each delay function D_a in c_a. □

The existence and uniqueness of the manager's best response now follows easily. Note first that if we set $c_1 = f_1$ then $T(c_1) = \sum f_a < C$ (as assumed in the proposition's conditions). Setting $c_1 = C$ obviously yields $T(c_1) \geq C$. Then by Lemma 2, it follows that there exists a unique value of $c_1 \in [f_1, C]$ such that $T(c_1) = C$. This value of c_1 induces a unique feasible capacity allocation over the remaining service classes, such that the ratios are satisfied. This establishes part (i) of the proposition.

We next consider the proof of part (ii). A straightforward conclusion from Lemma 2 is that the required capacity allocation can be obtained by a simple search over the scalar c_1, that will induce the required delay ratios. Based on the mapping $T(c_1)$ defined above, we search for c_1 so that $T(c_1) = C$. Since $T(c_1)$ is monotonous in c_1, several well-known techniques could be applied for an efficient search, such as the bisection method [35]. □

Proof of Proposition 2: (outline) The key idea in the proof is to use Lagrangian techniques to establish that optimality conditions for (4) are equivalent to the ratio objective equations (2). Thus, by solving (4), the capacity allocation which meets the required delay ratios is obtained. The full details can be found in [28].

Proof of Proposition 3: (outline) Denote the (unique) capacity allocation which induces the required ratios by $\mathbf{c}^* = (c_1^*, \ldots c_A^*)$. Define the following potential function:

$$V(\mathbf{c}) = \frac{1}{2} \sum_b (c_b - c_b^*)^2. \quad (21)$$

It can be shown that (21) is a Lyapunov function for the system (6) implying its global stability. Details can be found in [28].

Proof of Proposition 4: Noting (1) and (2), the best response capacity allocation is derived by solving the following set of *linear* equations:

$$\sum_{a=1}^{A} c_a = C; \quad \rho_a(c_a - f_a) = (c_1 - f_1), \ a = 2, \ldots, A. \tag{22}$$

The unique solution of these equations is easily seen to be given by (7). □

Appendix II
Proofs for Section IV

Proof of Proposition 6: This proposition extends Theorem 2.1 in [20] regarding the uniqueness of the Nash equilibrium for the parallel arcs network. The extensions are: 1) User demands are elastic [37]. 2) Users have an upper bound s_a^i on their allowed flow in each service class. Since the proof technique resembles the one used in [20], a full proof is omitted here, and can be found in [28].

Proof of Theorem 7: We established in Proposition 2 that the manager's best-response capacity allocation is also a solution to a convex optimization problem (4). Since each user i's cost function is convex in its decision vector $\mathbf{f}^i$, then the existence of a NEP in our model essentially follows from a well known result regarding the existence of a NEP in convex games [38], [39]. Nonetheless, since some technical points related to infinite costs in our model impede a direct application of this result, we provide a proof which is a direct application of the Kakutani fixed point theorem (see, e.g., [40]).

Let us first precisely state the Kakutani's fixed point theorem, along with the necessary mathematical definitions. These are taken from [40].

Theorem 17: (Kakutani) Let S be a compact and convex subset of $\mathbb{R}^n$, and let Λ be an upper semicontinuous function which assigns to each $x \in S$ a closed and convex subset of S. Then there exists some $x \in S$ such that $x \in \Lambda(x)$.

Recall that Λ is said to be upper semicontinuous (usc) at a point $x_0 \in S$, if for any sequence (x_i) converging to x_0 and any sequence $(y_i \in \Lambda(x_i))$ converging to y_0, we have $y_0 \in \Lambda(x_0)$. The function Λ is upper semicontinuous if it is usc at each point of S.

Let $S \triangleq \mathbf{F} \times \mathbf{\Gamma}$. Note that any NEP belongs to S, as a point outside S is not a feasible system configuration. We further define the point-to-set mapping $(\mathbf{c}, \mathbf{f}) \in S \mapsto \Lambda(\mathbf{c}, \mathbf{f})$, as follows.

$$\begin{aligned}\Lambda(\mathbf{c}, \mathbf{f}) = \big\{(\hat{\mathbf{c}}, \hat{\mathbf{f}}) \in S : \hat{\mathbf{c}} &\in \operatorname*{argmin}_{\tilde{\mathbf{c}} \in \mathbf{\Gamma}} J^M(\tilde{\mathbf{c}}, \mathbf{f}) \\ \hat{\mathbf{f}}^i &\in \operatorname*{argmin}_{\tilde{\mathbf{f}}^i \in \mathbf{F}^i} J^i(\mathbf{c}, \tilde{\mathbf{f}}^i, \mathbf{f}^{-i}) \ \forall i \in \mathcal{I}\big\}.\end{aligned} \tag{23}$$

Note that $\Lambda(\mathbf{c}, \mathbf{f})$ is comprised of the best response correspondences of each player (user or manager). It is readily seen that Λ is usc for the points $(\mathbf{c}, \mathbf{f}) \in S$ such that $\sum_a f_a < C$. Indeed, J^i is continuous in $(\mathbf{c}, \mathbf{f})$ and convex in $\mathbf{f}^i$, guaranteeing the usc of the user's best response (see, e.g., [39]). Further, we may replace J^M in the definition of Λ above by $\hat{J}^M$ (the objective function in (4)), since the best responses of the two coincide (see the proof of Proposition 2). Note that $\hat{J}^M$ is continuous in $(\mathbf{c}, \mathbf{f})$ and convex in $\mathbf{c}$ in the neighborhood of the best response allocation (due to Proposition 2). Thus overall Λ is usc (note that by the strict convexity property of the cost functions in their respective decision variables, Λ is in fact a point-to-point mapping). For the case where $\sum_a f_a \geq C$, the user's best response still maintains the continuity, finiteness and convexity properties, as users can always "ignore" infinite delay arcs by shipping a zero flow into them. The manager is indifferent as to its "best response" to $\mathbf{f}$ (since the manager will obtain an infinite cost regardless of the chosen capacity allocation), thus $\{\hat{\mathbf{c}} | (\hat{\mathbf{c}}, \hat{\mathbf{f}}) \in \Lambda(\mathbf{c}, \mathbf{f})\} = \mathbf{\Gamma}$. This implies that for any sequence $\{\Lambda(\mathbf{c}^k, \mathbf{f}^k)\}$ converging to some $(\mathbf{c}_0, \mathbf{f}_0)$, we have that $(\mathbf{c}_0, \mathbf{f}_0) \in \Lambda(\mathbf{c}, \mathbf{f})$. For both the cases above, it is readily seen that $\Lambda(\mathbf{c}, \mathbf{f})$ is a closed and convex set.

Note that the finiteness of the NEP is guaranteed, since if not all costs are finite, then at least one player with infinite cost can change its own flow configuration to make its cost finite. This argument is valid, since for the case where $\sum_a f_a \geq C$ there exists a user who ships flow to at least a single arc with an infinite delay. This user can make its cost finite by unilaterally reducing (possibly nullifying) its flow in the infinite delay arcs. For the case where $\sum_a f_a < C$ all players can employ their best response to obtain a finite cost. Applying the Kakutani fixed point theorem with the above definitions of S and Λ, we conclude that there exists a NEP. This NEP is finite as shown above, and it is also a NEP where the delay ratios are met.□

Proof of Theorem 8: The idea of this proof is to establish uniqueness of the user best response for the case where the class delays are given, and then argue that the delay values are identical in every equilibrium point. The proof proceeds through the next three lemmas.

Lemma 3: Let $D_1, \ldots, D_A$ be the class delays at some NEP, and let $D'_a \triangleq \frac{\partial D_a}{\partial f_a}$. Then the following equations are met at the equilibrium for every $i \in \mathcal{I}$ and every $a \in \mathcal{A}$

$$\begin{aligned}\beta^i \left(D_a + f_a^i D'_a\right) + p_a^i &\leq U^i(f^i)' \quad \text{if } f_a^i = s_a^i, \\ \beta^i \left(D_a + f_a^i D'_a\right) + p_a^i &= U^i(f^i)' \quad \text{if } 0 < f_a^i < s_a^i, \\ \beta^i \left(D_a + f_a^i D'_a\right) + p_a^i &\geq U^i(f^i)' \quad \text{if } f_a^i = 0,\end{aligned} \tag{24}$$

where $U^i(f^i)' \triangleq \frac{dU^i(f^i)}{df^i}$.

Proof: Let $i \in \mathcal{I}$. Observe that $\frac{\partial J^i(\mathbf{c}, \mathbf{f})}{\partial f_a^i} = \beta^i \left(D_a(c_a, f_a) + f_a^i D'_a(c_a, f_a)\right) + p_a^i - U^i(f^i)'$. Then (24) is readily seen to be the KKT optimality conditions [35] for minimizing the cost function (9) of user i subject to the flow constraint $0 \leq f_a^i \leq s_a^i$. These conditions are necessary and sufficient by the convexity of J^i in (9) in $\mathbf{f}^i$. □

Lemma 4: Consider a NEP with given class delays $D_1, \ldots, D_A$. Then the respective equilibrium flows f_a^i are uniquely determined.

Proof: Assume fixed delays D_a, $a \in \mathcal{A}$. As mentioned, D'_a is uniquely determined by D_a by Assumptions B4 and B6. For every $i \in \mathcal{I}$, consider the optimization problem given in (11). Note that (11) is a strictly convex optimization problem, since the objective function is the sum of a diagonal quadratic term (with $\beta^i D'_a > 0$ for every a) and the negation of U^i, where U^i is concave. Thus, this problem has a unique minimum, which is characterized by the KKT optimality conditions. It is now readily seen that the KKT conditions for (11) coincide with the conditions in (24). Thus, by Lemma 3, any set of equilibrium flows $(f_a^i)_{a\in\mathcal{A}}$ is a solution of (11). But since this solution is unique, the equilibrium flows for every $i\in\mathcal{I}$ are uniquely determined. □

Lemma 5: Consider two Nash equilibrium points $(\mathbf{c},\mathbf{f})$ and $(\tilde{\mathbf{c}},\tilde{\mathbf{f}})$. Then $D_a(c_a,f_a)=D_a(\tilde{c}_a,\tilde{f}_a)$ for every $a\in\mathcal{A}$.

Proof: Denote $D_a \triangleq D_a(c_a,f_a)$ and $\tilde{D}_a \triangleq D_a(\tilde{c}_a,\tilde{f}_a)$. Assume that $\tilde{D}_a > D_a$ for some $a\in\mathcal{A}$. Then $\tilde{D}_a > D_a$ for every $a\in\mathcal{A}$ since the ratios are met in both equilibria. It follows by Assumption B6 that $\tilde{c}_a-\tilde{f}_a < c_a-f_a$ for every a. Since the total capacity C is fixed in both equilibria, then summing the last inequality over all service classes yields that $\sum_{a\in\mathcal{A}}\tilde{f}_a > \sum_{a\in\mathcal{A}} f_a$. This implies that there exists some user $j\in\mathcal{I}$ for which

$$\tilde{f}^j=\sum_{a\in\mathcal{A}}\tilde{f}_a^j>\sum_{a\in\mathcal{A}}f_a^j=f^j. \qquad (25)$$

We next contradict (25) by invoking the next two implications:

$$(a)\ f_a^j=0\Rightarrow\tilde{f}_a^j=0;\quad (b)\ f_a^j>0\Rightarrow f_a^j>\tilde{f}_a^j. \qquad (26)$$

Their proof is based on the KKT conditions (24). Since the utility U^j is concave, then by (25) we have $\lambda^j\triangleq U^j(f^j)'\geq U^j(\tilde{f}^j)'\triangleq\tilde{\lambda}^j$. If $f_a^j=0$, then $\beta^jD_a+p_a^j\geq\lambda^j\geq\tilde{\lambda}^j$. Since $\tilde{D}_a>D_a$, then $\beta^j\tilde{D}_a+p_a^j>\tilde{\lambda}^j$, hence $\tilde{f}_a^j=0$. To prove (26)(b) note first that it holds trivially if $\tilde{f}_a^j=0$ or $f_a^j=s_a^j$. Next assume $\tilde{f}_a^j>0$ and $f_a^j<s_a^j$. Then by (24)

$$\beta^j(D_a+D'_af_a^j)+p_a^j\geq\lambda^j\geq\tilde{\lambda}^j\geq\beta^j(\tilde{D}_a+\tilde{D}'_a\tilde{f}_a^j)+p_a^j. \quad (27)$$

Since $\tilde{D}_a>D_a$ (hence $\tilde{D}'_a>D'_a$ by assumptions B4 and B6), and each β^j is positive, we must have $f_a^j>\tilde{f}_a^j$ in order for (27) to hold, which establishes (26)(b). Summing user j's flows according to (26) yields $\sum_{a\in\mathcal{A}}\tilde{f}_a^j\leq\sum_{a\in\mathcal{A}}f_a^j$, which contradicts (25). Thus $\tilde{D}_a\leq D_a$. Symmetrical arguments will lead to $\tilde{D}_a\geq D_a$, hence $\tilde{D}_a=D_a$ for every $a\in\mathcal{A}$. □

The last two lemmas imply that the user flows and the class delays in equilibrium are unique. The capacities in the equilibrium must also be unique by uniqueness of the manager's best response (Proposition 1). This establishes the uniqueness of the NEP, and completes the proof of Theorem 8. □

Proof of Lemma 1: Given the (assumed) equilibrium delay D_1, the remaining equilibrium delays are uniquely determined. Hence, the conditions of Lemma 4 are established, and the result directly follows. □

Proof of Proposition 9:

1) Formally, we have to prove the following: *Let D_1 and $\hat{D}_1$ be two estimates of the equilibrium delay in class* 1. *Then if $\hat{D}_1>D_1$, it follows that $\tilde{D}_1(\hat{D}_1)<\tilde{D}_1(D_1)$.* For the proof, we assume that the estimate D_1 is such that the resulting total flow which is obtained from (11) is positive. This assumption is practically met for any search scheme, since if the network is indeed utilized in equilibrium (i.e., $\sum f_a>0$), then any plausible search method would tune its estimates of D_1 to a range in which the resulting user-equilibrium total flow is positive.

Observe first that the quantity $\sum f_a$ is non-increasing with D_1 as the (aggregate) solution to the user optimization problems (11). This fact was established in the proof of Lemma 5. It may be easily verified that the same quantity strictly decreases in D_1 if the respective solutions to the user optimization problems are such that $\sum_a f_a>0$ (by showing that (26)(b) holds for at least a single (user, service class) pair). Hence, since $\hat{D}_1>D_1$ it follows that

$$\sum\hat{f}_a<\sum f_a. \qquad (28)$$

Assume by contradiction that $\tilde{D}_1(\hat{D}_1)\geq\tilde{D}_1(D_1)$. This means that $\tilde{D}_a(\hat{D}_1)\geq\tilde{D}_a(D_1)$ for every a. Hence, By Assumption B6 we have that $\hat{c}_a-\hat{f}_a\leq c_a-f_a$. Summing up on all service classes we obtain that $\sum\hat{f}_a\geq\sum f_a$ contradicting (28). Thus, $\hat{D}_1>D_1$ implies that $\tilde{D}_1(\hat{D}_1)<\tilde{D}_1(D_1)$.

2) Observe that $\tilde{D}_1(D_1)$ is obtained through best response map from D_1, and therefore $D_1=\tilde{D}_1(D_1)$ must hold for the equilibrium value of D_1 (recall that the equilibrium is unique by Theorem 8). It therefore follows by part 1 of the proposition and by the continuity of the associated cost functions, that the equilibrium delay is the unique solution to $D_1-\tilde{D}_1(D_1)=0$. This is a direct consequence of the mean value theorem.

3) The results above indicate that search methods with polynomial complexity may be applied for efficiently calculating the equilibrium delays. The key idea of any iterative search in our context is to decrease in each step the distance between D_1 and $\tilde{D}_1(D_1)$. For example, if the bisection method [35] is applied, then the new guess in each step is the average of the previous D_1 and $\tilde{D}_1(D_1)$. The number of steps which are required by the method for a precision of ϵ is given by $\log_2|D_1^0-\tilde{D}_1(D_1^0)|-\log_2\epsilon$, where D_1^0 is the initial estimate of D_1. □

Appendix III
Proofs for Section VI

Theorem 14 follow similarly to the single-hop case, while Theorems 15 and 16 require specific proofs.

Proof of Theorem 14: (outline) As in the single link case, the existence of a NEP essentially follows from the existence of a NEP in convex games [38], [39]. Note first that the best response capacity allocation of each capacity manager M_l, $l\in\mathcal{L}$ is also a solution to a convex optimization problem (4) (Proposition 2). As to the network users, observe that the delay cost (15) of each user i can be written as $\beta^i\sum_{l\in R^i}\sum_{a=1}^{A}f_a^iD_{la}(c_{la},f_{la})$. Since $\sum_{a=1}^{A}f_a^iD_{la}(c_{la},f_{la})$ is convex in $\mathbf{f}^i$ for every $l\in R^i$, the delay cost is convex in $\mathbf{f}^i$ as the sum of convex functions. Noting that the other cost terms in (15) remain the same as the single link case, we conclude that each user's cost function is convex in its decision vector. Thus, the basic conditions for the existence of a NEP in our game (as a convex game) are established. Some technical points related to infinite costs are resolved in a similar manner as in the proof of existence of a NEP for the single hop case (Theorem 7). □

Proof of Theorem 15(i) : (outline) The key idea is to show that the delays at equilibrium are unique, and then establish the uniqueness of the equilibrium flows similarly to Lemma

4. A detailed proof is omitted due to lack of space, and can be found in [28].

Proof of Theorem 15(ii) : (outline) The calculation of the equilibrium flows and capacities is even easier than in the single-hop case with elastic users. Briefly, the delays can be obtained just from the set of total flows over each link (which is fixed here). Consequently, (11) is solved for each user to obtain the exact flow distribution, leading to the capacity allocation to each link. A detailed description of this calculation can be found in [28].

Proof of Theorem 16: The proofs here resemble the proof of Lemma 5, see [28] for details.

References

[1] I. Menache and N. Shimkin, "Proportional QoS in differentiated services networks: Capacity management, equilibrium analysis and elastic demands," in *WINE*, 2005, pp. 728 – 737.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," RFC 2475, 1998.

[3] X. Xiao and L. M. Ni, "Internet QoS: A big picture," *IEEE Network*, vol. 36, no. 11, pp. 8–18, 1999.

[4] B. Davie, A. Charny, J. Bennett, K. Benson, J. Boudec, W. Courtney, S. Davari, V. Firoiu, and D. Stiliadis, "An expedited forwarding PHB (per-hop behavior)," RFC 3246, 2001.

[5] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured forwarding PHB group," RFC 2597, 1999.

[6] C. Courcoubetis and R. Weber, *Pricing Communication Networks: Economics, Technology and Modelling*. Wiley, 2003.

[7] *Internetworking Technologies Handbook*. Cisco documentation, 2005.

[8] C. Dovrolis, D. Stiliadis, and P. Ramanathan, "Proportional differentiated services: Delay differentiation and packet scheduling," *IEEE/ACM Transactions on Networking*, pp. 12–26, 2002.

[9] C. Li, S. Tsao, M. Chen, Y. Sun, and Y. Huang, "Proportional delay differentiation service based on weighted fair queuing," in *Proceedings of the Ninth International Conference on Computer Communications and Networks*, 2000, pp. 418–423.

[10] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queuing algorithm," *Journal of Internetworking: Research and Experience*, vol. 1, no. 6, pp. 3–26, 1990.

[11] Y. C. C. Qiao, M. Hamdi, and D. Tsang, "Proportional differentiation: a scalable QoS approach," *IEEE Communications Magazine*, vol. 41, no. 6, pp. 52–58, 2003.

[12] C. Dovrolis and P. Ramanathan, "Class provisioning using proportional delay differentiation," in *the Scalability and Traffic Control in IP Networks conference*, 2001.

[13] Y. Chen, M. Hamdi, D. H. K. Tsang, and C. Qiao, "Providing proportionally differentiated services over optical burst switching networks," in *Proceedings of IEEE Global Telecommunications Conference*, 2001, pp. 1510–1514.

[14] A. M. Odlyzko, "Paris metro pricing for the internet," in *Proceedings of the ACM Conference on Electronic Commerce*, 1999, pp. 140–147.

[15] M. Mandjes, "Pricing strategies under heterogeneous service requirements," in *Proceedings of IEEE INFOCOM*, 2003, pp. 1210–1220.

[16] Y. Hayel, D. Ros, and B. Tuffin, "Less-than-best-effort services: Pricing and scheduling," in *Proceedings of IEEE INFOCOM*, 2004, pp. 66–75.

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

[18] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Proc. Inst. Civ. Eng*, vol. 2, pp. 325–378, 1952.

[19] E. Altman and L. Wynter, "Equilibrium, games, and pricing in transportation and telecommunications networks," *Networks and Spacial Economics*, vol. 4, no. 1, pp. 7–21, 2004.

[20] A. Orda, R. Rom, and N. Shimkin, "Competitive routing in multi-user environments," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 510–521, 1993.

[21] Y. A. Korilis, A. A. Lazar, and A. Orda, "Architecting noncooperative networks," *IEEE Journal on Selected Areas in Communication*, vol. 13, no. 7, pp. 1241–1251, 1995.

[22] R. Azouzi and E. Altman, "Constrained traffic equilibrium in routing," *IEEE Transactions on Automatic Control*, vol. 48, no. 9, pp. 1656–1660, 2003.

[23] D. Bertsekas and R. Gallager, *Data Networks*. Prentice-Hall, 1992.

[24] M. Katevenis, S. Sidiropoulos, and C. Courcoubetis, "Weighted round-robin cell multiplexing in a general-purpose ATM switch chip," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 8, pp. 1265–1279, 1991.

[25] L. Kleinrock, *Queueing systems*. John Wiley and Sons, 1975.

[26] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[27] M. Patriksson, "Algorithms for computing traffic equilibria," *Networks and Spacial Economics*, vol. 4, no. 1, pp. 23–38, 2004.

[28] I. Menache and N. Shimkin, "Capacity management and equilibrium for proportional QoS," Department of Electrical Engineering, Technion, Tech. Rep. CCIT No. 575, March 2006.

[29] T. Jimenez, E. Altman, T. Basar, and N. Shimkin, "Competitive routing in networks with polynomial cost," *IEEE Trans. on Automatic Control*, vol. 47, pp. 92–96, 2002.

[30] ——, "Routing into two parallel links: game-theoretic distributed algorithms," *Journal of Parallel and Distributed Computing*, vol. 61, no. 9, pp. 1367–1381, 2001.

[31] C. H. Papadimitriou, "Recent developments in equilibria algorithms," in *Proceedings of WINE*, 2005, pp. 1–2.

[32] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991.

[33] P. Marbach, "Analysis of a static pricing scheme for priority services," *IEEE/ACM Transactions on Networking*, vol. 12, pp. 312–325, 2004.

[34] R. Mazumdar, L. Mason, and C. Douligeris, "Fairness in network optimal flow control: Optimality of product forms," *IEEE Trans. on Communications*, vol. 39, no. 5, pp. 775–782, 1991.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2003.

[36] R. D. Luce and H. Raiffa, *Games and Decisions*. John Wiley, 1957.

[37] E. Altman, R. E. Azouzi, T. Basar, and R. Srikant, "Combined competitive flow control and routing games," in *Workshop on Networking Games and Resource Allocation, Petrozavodsk*, 2002, pp. 12–15.

[38] G. Debreu, "A social equilibrium existence theorem," *Proceedings of the. National Academy of Science*, vol. 38, pp. 886–893, 1952.

[39] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica*, vol. 33, no. 3, pp. 520–534, 1965.

[40] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. Academic Press, 1995.

Ishai Menache received the B.Sc. and M.Sc. degrees in Electrical Engineering from the Technion, Haifa, Israel, in 1998 and 2003, respectively. Between 1997 and 2000 he worked at Intel as an engineer in the network communications group. He is currently working toward a Ph.D. degree in the Department of Electrical Engineering at the Technion. His research interests include QoS in communication networks and reinforcement learning.

Nahum Shimkin received the Ph.D. degree in Electrical Engineering from the Technion, Israel, in 1991. Subsequently he spent a year as a Postdoctoral fellow at the Institute of Mathematics and its Applications, University of Minnesota, and two years as a Senior Research Engineer in Rafael. He is currently an associate professor of Electrical Engineering at the Technion. His research interests include stochastic systems and control, reinforcement learning, dynamic games, queueing systems, and competitive behavior in multiuser systems.