

# A game-theoretic model for capacity-constrained fair bandwidth allocation

Yonghe Yan, Adel El-Atawy\*<sup>†</sup> and Ehab Al-Shaer

*School of Computing, DePaul University, Chicago, IL 60604, USA*

## SUMMARY

Data stream providers face a hard decision to satisfy the requirements of their subscribers. Each user has a minimum and a maximum required bandwidth. The server should be able to decide which requests can be satisfied and how much bandwidth will be allocated to each. We present a theoretical framework in a distributed mechanism for fair bandwidth allocation on a network with various bottleneck links. In our model, a user is guaranteed a minimum bandwidth and charged a price for the bandwidth allocated. A utility function is defined over the allocated bandwidth for a specific maximum requested bandwidth. We then present a non-cooperative game with social welfare function to resolve users' conflicting bandwidth capacity requests at bottleneck links. We also show that our proposed game-theoretic solution guarantees fair bandwidth allocation as defined in our residual capacity fairness. In order to guarantee the minimum bandwidth requirement, we integrate an admission control mechanism in our solution. However, global optimal admission conditions are not easy to implement for large networks. Therefore, we propose a distributed admission scheme. As a result, the paper presents fair and practical distributed algorithms for bandwidth allocation and admission control in enterprise networks. Our simulation and evaluation study shows that the distributed approach is sufficiently close to the global optimal solution. Copyright © 2008 John Wiley & Sons, Ltd.

## 1. INTRODUCTION

Quality of service (QoS) in the Internet has received significant attention in many research communities. The ability to provide differentiated services is essential for accommodating next-generation Internet objectives, including the support for various types of applications and services. Fair bandwidth allocation, as one of the most critical resources, is becoming increasingly important. Having a pricing model for such a crucial resource is essential in order to provide sufficient incentives for users to use the network resources efficiently, and for service providers to provide guaranteed QoS services in a healthy market environment.

In this paper, we consider a network model that is oblivious to the underlying topology and, as a consequence, to routing. In this model, there exist a set of service classes and a group of flows that belong to each. Each flow in a service class requests a certain bandwidth capacity to achieve the required QoS. A link in the network has a given finite bandwidth capacity that cannot be exceeded by the aggregate bandwidth of all flows using this link. Each flow should be assigned a bandwidth to share the link capacities in compliance with a fairness criterion. We propose a model that can devise a fairness criterion for links to assign a flow a fair share of the link capacities. The model also allows us to investigate insightful properties of prices among various service classes, and allocates bandwidths fairly such that user

---

\*Correspondence to: Adel El-Atawy, School of Computing, DePaul University, 243 S Wabash Avenue, Chicago, IL 60641, USA.

<sup>†</sup>E-mail: aelatawy@cs.depaul.edu

bandwidth requirements are guaranteed. These properties can also be used by service providers to price differentiated service classes properly. We assume that each user is charged a price for a certain bandwidth capacity at the edge of the network. The price considered in this paper is not a unit price. It is charged for a 2-tuple of bandwidth requirements; that is, the bandwidth capacity and the minimum bandwidth.

The bandwidth capacity, which is typically bought from a network service provider, represents the maximum bandwidth enforced at the network entry point. When congestions occur in the network, the user might be allocated a bandwidth that is lower than the user bandwidth capacity. However, the allocated bandwidth in this case should not be arbitrarily low in order to guarantee minimum bandwidth requirement for QoS demands as stated in the service-level agreement. Thus, users should be guaranteed a bandwidth between the bandwidth capacity and the minimum bandwidth.

To satisfy various bandwidth requirements of concurrent users, the network should use all available bandwidth to the fullest while maintaining certain fairness in bandwidth allocations. When a link's capacity is used to the fullest, congestion occurs at the link. Thus, increasing bandwidth of one flow implies decreasing the bandwidths of other flows. Therefore, we developed a non-cooperative game with social welfare function to solve users' conflicting bandwidth demands at the bottleneck links. The social welfare function is the total utility of a user's utility function, which reflects user bandwidth demand when a user requests a bandwidth capacity. The solution to the game is a fair bandwidth allocation that satisfies the residual capacity fairness criterion that takes into account a user's bandwidth capacity requests. In order to satisfy the guaranteed minimum bandwidth requirements, we integrate admission control into the bandwidth allocation such that each user is allocated a fair bandwidth that is between the requested bandwidth capacity and the minimum bandwidth. However, this centralized scheme is not easy to implement in a large network. We thus propose a distributed approach so that each bottleneck link independently allocates a fair share for each flow according to its capacity constraints. Therefore, all links can accommodate a flow when the flow takes a bandwidth that is not greater than any of the bandwidths allocated by these bottleneck links. With simulations, we evaluated the relative difference between the bandwidth allocations given by the centralized and the distributed schemes, and found that the relative difference is within a satisfactory range.

Although there are significant amounts of research in this area, considering users' capacity constraints and providing a fully distributed allocation scheme is a unique contribution of this paper.

The paper is organized as follows. Section 2 discusses the related work. In Section 3, we present the network model used in the paper and a social welfare game for the bandwidth allocation. A utility function is introduced in this section as well. Section 4 presents a notion of fairness for the bandwidth allocation: residual capacity fairness. In Section 5, we discuss the pricing framework that is integrated in the network model. Section 6 considers the global admission conditions and the bandwidth allocation for flows. Distributed admission conditions and bandwidth allocation are addressed in Section 7. In Section 8, we present our simulation results for evaluating a sample network and the accuracy of the distributed bandwidth allocation and admission control. Finally, Section 9 draws conclusions and future work.

## 2. RELATED WORK

Several models have been put forth for service providers to charge network services dynamically or statically. Paschalidis and co-workers [1,2] presented a revenue maximization problem to charge each bandwidth requirement a static price. In [3], optimal price and admission control policy are derived from a global revenue maximization problem. Savagaonkar *et al.* [4] also investigated a revenue maximization problem with dynamic pricing for bandwidth provisioning under the assumption that users' demands are known stochastic processes. These price schemes are designed primarily for a network where services are provided by one network owner or broker. But for large networks, such as the Internet, there are numerous network service providers. Global revenue maximization is not the primary concern of each individual service provider.

Wang and Schulzrinne [5] considered a price scheme for Differentiated Services architecture (DiffServ) [6]. The price for each service class depends on the average demand for that service class and is negotiated through a Resource Negotiation and Pricing protocol [7]. The utility function used in their work is based on a set of packet transmission parameters. Somewhat similar work can also be found elsewhere [8–10], and users are charged prices derived from measurements of packet transmissions. However, users who access the network wish to complete certain tasks (e.g., place an Internet phone call, request a web page, send an email) and packets are completely transparent to them. Thus, it is not clear whether packet-based pricing schemes are always appropriate [1].

Semret *et al.* [11] proposed a bandwidth capacity provisioning framework to sell bandwidth capacities in a dynamic market. A user is charged by a service provider for a certain bandwidth capacity, and the bandwidth capacity is actually the constraint enforced by the service provider at the network entry point. However, the service provider cannot guarantee that the user can always transmit the user's data at the speed of bandwidth capacity since congestions may occur in the network. Thus, this scheme lacks any control over bandwidth allocation, and users may be unfairly charged for a nominal bandwidth capacity.

Yaïche *et al.* [12] proposed a fair bandwidth allocation from the solution to the Nash bargaining problem [13,14]. Kelly *et al.* [15] introduced a notion of proportional fairness for an 'elastic' network (i.e., best-effort service network), where users' total utility is maximized when the bandwidth allocation fairness is achieved. This fair bandwidth sharing is further generalized to be weighted  $\alpha$ -bandwidth allocation in [16,17], which includes several fairness criteria, such as max–min fairness [18,19]. However, these fairness criteria only take into account links' capacities. They do not consider the users' bandwidth capacity requirements. The fairness criteria always favor increasing small bandwidth over large bandwidth [15,16]. Therefore, a user requesting a small bandwidth capacity may be allocated bandwidth that is much closer to the requested bandwidth capacity than that of a user requesting a large bandwidth capacity. It is not considered fair between users requesting different bandwidth capacities.

Various fluid models are used [20–24] for bandwidth allocation analysis of elastic networks. Alpcan and Başsar [25] proposed a broad class of utility functions that are non-decreasing and strictly concave on bandwidths over the interval  $(0, \infty)$ . All these utility functions are used to capture user demand for bandwidth in elastic networks. They are not applicable for inelastic networks, where a user is charged for a certain bandwidth capacity and guaranteed a minimum bandwidth.

### 3. NETWORK MODEL

We consider a fluid model of the network where the packets are infinitely divisible and small. The network is a set of links  $\mathbf{L}$  where each link  $j \in \mathbf{L}$  has a capacity  $C_j > 0$ , and let  $L = |\mathbf{L}|$  be the number of the links in the set. There is a set of users  $\mathbf{N}$  with the cardinality  $N = |\mathbf{N}|$ . The users compete for the use of the network. Each flow is associated with a route consisting of a subset of  $\mathbf{L}$ . Without loss of generality, we assume that each user  $i \in \mathbf{N}$  is associated with one flow (or connection) in the network. The user is charged a price  $p_i > 0$  for requesting bandwidth capacity  $R_i$  with a guaranteed minimum bandwidth  $r_i$ , so that the user can transmit data with a bandwidth of  $x_i$ , where  $R_i \geq x_i \geq r_i \geq 0$ . In this paper, we make the assumption that users' demands are stationary and therefore bandwidth capacity  $R_i$  and minimum bandwidth  $r_i$  are constants in the model. We define the matrix  $\mathbf{A} = (A_{ij}, i \in \mathbf{N}, j \in \mathbf{L})$  where  $A_{ij} = 1$  if flow  $i$  uses link  $j$  and  $A_{ij} = 0$ , otherwise. Let  $\mathbf{J}_i = \{j | A_{ij} = 1\}$  be the set of links that flow  $i$  uses and  $\mathbf{I}_j = \{i | A_{ij} = 1\}$  be the set of flows that use link  $j$ .

We assume that the total minimum bandwidths of flows using a link cannot exceed the link's capacity, that is, the minimum bandwidth  $\mathbf{r} = (r_1, \dots, r_N)^T$  satisfies the link capacity constraint  $\mathbf{A}^T \mathbf{r} \leq \mathbf{C}$ , where  $\mathbf{C} = (C_1, \dots, C_L)^T$ . Otherwise it is not feasible for the network to allocate each flow a bandwidth that is not less than its minimum bandwidth.

**Assumption 1** The bandwidth requirements of flows are feasible, that is,  $\mathbf{A}^T \mathbf{r} \leq \mathbf{C}$ .

In the rest of the paper, we implicitly consider feasible bandwidth requirements only. However, the total bandwidth capacities requested by flows using a link may or may not exceed the link's capacity.

When the bandwidth capacities of links are over-provisioned for flow  $i$ , that is,  $\sum_{k \in I_i} R_k < C_j$  for each  $j \in J_i$ , there is no congestion on the route of the flow. The links can always assign flow  $i$  a bandwidth that equals its bandwidth capacity  $R_i$ . However, we are interested in considering flows that do go through congestion links. Moreover, when congestion occurs at a link, the link should use its full capacity to serve the flows using it. Therefore, we make the following assumption.

**Assumption 2** There is at least one link  $j \in J_i$  on the route of flow  $i$ , for each  $i = 1, \dots, N$ , such that  $\sum_{k \in I_j} x_k = C_j$ .

Assumption 2 implies that there exists at least one bottleneck link  $j$  on the route of flow  $i$ . At the bottleneck link, the total bandwidth capacities requested by flows using the link exceeds the link's capacity (i.e.,  $\sum_{k \in I_j} R_k \geq C_j$ ).

A user is charged a price  $p_i$  for requesting bandwidth capacity  $R_i$ , which is the bandwidth constraint enforced by the service provider at the network entry point. Consumers in the real world generally try to obtain the best possible 'value' for the price they have been charged, subject to their budget and minimum quality requirements; in other words, a user is charged a price for a bandwidth capacity and the user obtains the best value from the network when the network allocates the user a bandwidth that equals the bandwidth capacity. The user has an aversion to any bandwidth that deviates away from the bandwidth capacity. Thus, the user's objective is to maximize the following utility function with respect to  $x_i$  over  $[r_i, R_i]$ :

$$U_i(x_i) = \frac{w_i}{(\alpha - 1)(R_i - x_i)^{\alpha - 1}} \quad i = 1, \dots, N \tag{1}$$

where  $w_i = m_i p_i (R_i - r_i)^\alpha$ . The parameter  $w_i$  is the weight of the user's utility. It represents the nature of the flow, and takes into account the price, the bandwidth requirements, and  $m_i$ , which is the number of links used by the flow. The parameter  $\alpha > 1$  represents the sensitivity to deviation of the allocated bandwidth  $x_i$  from the requested bandwidth capacity  $R_i$ . Therefore, the utility is more sensitive when  $\alpha$  is large and vice versa. We will discuss  $w_i$  and  $\alpha$  in more detail in Section 5. This utility function reflects the user's subjective preference of the network service. The best possible bandwidth the user expects is  $R_i$ . A user is fully satisfied when  $x_i \rightarrow R_i$ , thus  $U_i(x_i) \rightarrow \infty$  when  $x_i \rightarrow R_i$ . When  $x_i$  deviates away from  $R_i$ , the utility decreases because the user's perceived value of the network service decreases.

Under Assumption 2, there is at least one bottleneck link on the route of each flow. At a bottleneck link, the total bandwidth capacities requested by flows using the link exceed the link's capacity. Some flows cannot be assigned a bandwidth that equals its requested bandwidth capacity, because the total bandwidths of flows at a bottleneck link cannot exceed the link's capacity, and increasing the bandwidth of one flow implies decreasing bandwidths of other flows. To solve this conflicting problem with allocating bandwidth fairly among flows, we consider a non-cooperative game among the flows.

Each player of the game is a flow, which is associated with a Harsanyi-type social welfare function [26]:

$$U(x_i, \mathbf{x}_{-i}) = \sum_{k \in N} U_k(x_k) \tag{2}$$

where  $\mathbf{x}_{-i} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_N)^T$  and  $U_k(x_k)$  is the utility function of flow  $k$  defined in (1). The strategy set of the game is

$$\mathbf{S} = \{(x_i, \mathbf{x}_{-i}) \mid \mathbf{A}^T \mathbf{x} \leq \mathbf{C} \text{ and } \mathbf{R} \geq \mathbf{x} \geq \mathbf{r}\}$$

where

$$\begin{aligned} \mathbf{x} &= (x_1, \dots, x_N)^T; \\ \mathbf{C} &= (C_1, \dots, C_L)^T; \\ \mathbf{R} &= (R_1, \dots, R_N)^T; \\ \mathbf{r} &= (r_1, \dots, r_N)^T; \end{aligned}$$

and for notation convenience,  $(x_i, \mathbf{x}_{-i}) = \mathbf{x}$ . The strategy set consists of all the feasible bandwidth allocations.

To achieve a fair bandwidth allocation, each flow maximize guaranteed social welfare against other flows

$$\max_{x_i} \min_{\mathbf{x}_{-i}} U(x_i, \mathbf{x}_{-i}) \quad (x_i, \mathbf{x}_{-i}) \in \mathbf{S} \quad i = 1, \dots, N \quad (3)$$

Note that this objective function is identical for all users. Under Assumption 2, flow  $i$  always goes through at least one bottleneck link  $j$  such that  $x_i = C_j - \sum_{k \in I_j, k \neq i} x_k$ . Substituting it into (3), problem (3) becomes one minimization problem  $\min_{\mathbf{x}} U(x_i, \mathbf{x}_{-i})$ , that is:

$$\min U(\mathbf{x}) = \min \sum_{i \in N} U_i(x_i) \quad (P)$$

subject to

$$\mathbf{A}^T \mathbf{x} \leq \mathbf{C} \quad (4)$$

$$\mathbf{x} \leq \mathbf{R} \quad (5)$$

$$\mathbf{x} \geq \mathbf{r} \quad (6)$$

The inequality (4) expresses the link capacity constraints. Bandwidth requirements of each flow are represented by (5) and (6).

The constraint (6) expresses that the bandwidth allocation must satisfy the minimum bandwidth requirements. We relax the constraint to be  $\mathbf{x} \geq 0$ , while we attempt to find the fair bandwidth allocation as the solution to problem  $P$ . Constraint (6) can be enforced by an admission control mechanism. It is not necessary to take it into account while solving problem  $P$  for the fair bandwidth allocation. If the bandwidth allocation cannot satisfy (6), the admission control will not admit the flows into the network. The need for an admission control is consistent with Assumption 1, that is, flows with infeasible bandwidth requirements will not be admitted into the network. Hence problem  $P$  is defined on the set that is nonempty, convex, and compact. Also,  $U(\mathbf{x})$  is a convex function because its Hessian matrix  $\nabla^2 U(\mathbf{x})$  is positive definite. This implies that a unique solution to problem  $P$  exists while the relaxed constraints are satisfied. We will use the relaxed constraints implicitly until Section 6.

Let  $L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$  denote the Lagrangian where  $\mu_j \geq 0, j = 1, \dots, L$ , and  $\lambda_i \geq 0, i = 1, \dots, N$ , are the Lagrange multipliers associated with constraint (4) and (5), respectively. Then

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = U(\mathbf{x}) - \boldsymbol{\mu}^T (\mathbf{A}^T \mathbf{x} - \mathbf{C}) - \boldsymbol{\lambda}^T (\mathbf{x} - \mathbf{R})$$

The first-order Kuhn–Tucker conditions [27] are

$$\frac{w_i}{(R_i - x_i)^\alpha} - \sum_{j \in J_i} \mu_j - \lambda_i = 0, \quad i = 1, \dots, N$$

and

$$\lambda_i (x_i - R_i) = 0, \quad \lambda_i \geq 0, \quad i = 1, \dots, N$$

$$\mu_j \left( \sum_{k \in I_j} x_k - C_j \right) = 0, \quad \mu_j \geq 0, \quad j = 1, \dots, L$$

$$x_i \geq 0, \quad i = 1, \dots, N$$

Under Assumption 2, we see that the constraint  $\mathbf{x} < \mathbf{R}$  is inactive and hence  $\lambda_i = 0$  for all  $i = 1, \dots, N$ ; and there is at least one link  $j \in \mathbf{J}_i$  for flow  $i, i = 1, \dots, N$ , such that  $\sum_{k \in \mathbf{I}_j, k \neq i} x_k - C_j = 0$  and  $\mu_j > 0$ . Hence, we have the solution to problem  $P$ :

$$\dot{x}_i = R_i - \left( \frac{w_i}{\sum_{j \in \mathbf{J}_i} \mu_j} \right)^{1/\alpha}, i = 1, \dots, N \tag{7}$$

$$\mu_j \left( \sum_{k \in \mathbf{I}_j} \dot{x}_k - C_j \right) = 0, \mu_j \geq 0, j = 1, \dots, L \tag{8}$$

This solution is the max–min strategy of game (2). It is also one of the Nash equilibria [28] of game (2).

*Lemma:* A bandwidth allocation is a Nash equilibrium of game (2) if the bandwidth allocation satisfies assumption 2.

*Proof:* Suppose that  $\mathbf{x}^* = (x_i^*, \mathbf{x}_{-i}^*)$  is a bandwidth allocation that satisfies Assumption 2. Thus, there exists a link  $j$  such that  $x_i^* + \sum_{k \in \mathbf{I}_j, k \neq i} x_k^* = C_j$ . Assume that flow  $i$  is allocated bandwidth  $x_i$  such that  $x_i \neq x_i^*$ . The bandwidth allocation  $(x_i, \mathbf{x}_{-i}^*)$  must satisfy constraint (4) such that  $x_i + \sum_{k \in \mathbf{I}_j, k \neq i} x_k^* \leq C_j$ . Combining these relations, we have  $x_i \leq x_i^*$ . Because  $U(x_i, \mathbf{x}_{-i})$  is a strictly increasing function with respect to  $x_i$ , we obtain

$$U(x_i, \mathbf{x}_{-i}^*) \leq U(x_i^*, \mathbf{x}_{-i}^*), i = 1, \dots, N$$

Therefore, we have shown that  $\mathbf{x}^*$  is a Nash equilibrium of game (2).  $\square$

The solution given by (7) and (8) satisfies Assumption 2 and therefore it is in Nash equilibrium of game (2).

There is another strategy that may be a candidate solution to the game. Each flow just maximizes its social welfare function over set  $\mathbf{S}$ :

$$\max_{\mathbf{x}} U(x_i, \mathbf{x}_{-i}) = \max_{\mathbf{x}} U(\mathbf{x}) \mathbf{x} \in \mathbf{S}$$

Note that  $\partial U(\mathbf{x})/\partial x_i > 0$  and  $\partial^2 U(\mathbf{x})/\partial^2 x_i > 0$  on the interval  $[0, R_i]$ . Hence, while maximizing  $U(\mathbf{x})$ , the bandwidth of a flow may keep increasing until it reaches its bandwidth capacity. Therefore, the bandwidth for flow  $i$  in the solution to the problem is one of the values in the set

$$\left\{ 0, R_i, C_j - \sum_{k \in \mathbf{I}_j, k \neq i} \eta_k R_k \mid \exists j \in \mathbf{J}_i, \eta_k \in \{0, 1\} \right\} \tag{9}$$

The third value in (9) is the remaining capacity at a bottleneck link when some other flows reach their bandwidth capacities. Although there is congestion in the network, some flows are still assigned bandwidths that reach their requested bandwidth capacities. In this case, these flows are fully satisfied and they will not be affected by any congestion in the network. At the same time, some other flows are assigned zero or the remaining bandwidth. Therefore, this is not considered a fair bandwidth allocation when there is congestion in the network. However, in our solution to problem  $P$ , no flow is assigned a bandwidth that reaches its bandwidth capacity when there is congestion. Moreover, we will show in the next section that this solution satisfies residual capacity fairness.

#### 4. RESIDUAL CAPACITY FAIRNESS

The notion of fairness characterizes how competing flows should share the bottleneck capacities. In this section we define a fairness criterion called residual capacity fairness, and investigate the relationship between the solution to problem  $P$  and the fairness criterion.

Bandwidth  $\alpha$ -fairness is a generalized fairness criterion that has been well established [15–17]. A bandwidth vector  $\dot{\mathbf{x}} = (\dot{x}_1, \dots, \dot{x}_N)^T$  is bandwidth  $\alpha$ -fair if it is feasible, that is,  $\mathbf{A}^T \dot{\mathbf{x}} \leq \mathbf{C}$  and  $\dot{\mathbf{x}} \geq 0$ , and if for any other feasible vector  $\mathbf{x}$

$$\sum_{i \in \mathbf{N}} w_i \frac{x_i - \dot{x}_i}{\dot{x}_i^\alpha} \leq 0$$

where a bandwidth vector  $\mathbf{x}$  is feasible when  $\mathbf{A}^T \mathbf{x} \leq \mathbf{C}$  and  $\mathbf{x} \geq 0$ . This feasible bandwidth is not constrained by the user's bandwidth capacity  $R_i$  because bandwidth  $\alpha$ -fairness is the fairness criterion for the best-effort network, where the user's bandwidth is only constrained by the links' capacities. Therefore, a feasible bandwidth in the best-effort network may not be feasible in our network model, where a bandwidth vector  $\mathbf{x}$  is feasible when  $\mathbf{A}^T \mathbf{x} \leq \mathbf{C}$ ,  $\mathbf{x} < \mathbf{R}$ , and  $\mathbf{x} \geq 0$ .

Moreover, bandwidth  $\alpha$ -fairness favors increasing  $x_i$  and decreasing  $x_j$  when  $x_i$  is much less than  $x_j$  [15,16]. This may not be deemed fair in our model. For example, although  $x_i$  is much less than  $x_j$ , a fairness criterion for our model should not favor increasing  $x_i$  and decreasing  $x_j$  when  $R_i - x_i$  is also much less than  $R_j - x_j$ . We will refer to  $R_i - x_i$  as residual capacity of flow  $i$ , and define residual capacity fairness that concerns the residual capacities.

The residual capacity fairness is the fairness criterion for a network to fairly assign users bandwidths in bandwidth capacity trading markets, such as those described in [1,4,11], where users are charged for bandwidth capacities at the edge of networks. The fairness criterion captures the essence that the fair bandwidth allocation should be aggregately close to the bandwidth capacities that have been purchased by the users.

**Definition: residual capacity fair** Let  $\mathbf{w} = (w_1, \dots, w_N)^T$  be positive numbers and  $\alpha$  a number on the interval  $(1, \infty)$ . A bandwidth vector  $\dot{\mathbf{x}} = (\dot{x}_1, \dots, \dot{x}_N)^T$  is residual capacity fair if it is feasible, that is,  $\mathbf{A}^T \dot{\mathbf{x}} \leq \mathbf{C}$ ,  $\dot{\mathbf{x}} \leq \mathbf{R}$ , and  $\dot{\mathbf{x}} \geq 0$ , and if for any other feasible vector  $\mathbf{x}$

$$\sum_{i \in \mathbf{N}} w_i \frac{x_i - \dot{x}_i}{(R_i - \dot{x}_i)^\alpha} \leq 0 \quad (10)$$

where a bandwidth vector  $\mathbf{x}$  is feasible when  $\mathbf{A}^T \mathbf{x} \leq \mathbf{C}$ ,  $\mathbf{x} \leq \mathbf{R}$ , and  $\mathbf{x} \geq 0$ . To investigate the residual capacity fairness, we rewrite (10) as follows:

$$\sum_{i \in \mathbf{N}} w_i \frac{(R_i - \dot{x}_i) - (R_i - x_i)}{(R_i - \dot{x}_i)^\alpha} \leq 0 \quad (11)$$

When increasing  $x_i$  such that  $x_i > \dot{x}_i$ , bandwidths of some other flows have to be decreased when there is congestion. Therefore, residual capacity change  $(R_i - \dot{x}_i) - (R_i - x_i)$  of flow  $i$  is positive and the residual capacity changes of some other flows have to be negative. However, it is shown in (11) that this positive weighted change of residual capacity of flow  $i$  cannot compensate for the negative weighted changes of residual capacities of some other flows. Therefore, the bandwidth of flow  $i$  cannot move closer to its bandwidth capacity  $R_i$  without moving bandwidths of some other flows further away from their bandwidth capacities. Therefore, this fairness criterion takes into account the residual capacities other than just the bandwidths.

The following proposition clarifies the relation between the solution to problem  $P$  and residual capacity fairness.

**Proposition 1** If Assumption 2 holds, the solution to problem  $P$  is residual capacity fair.

*Proof:* Let  $\dot{\mathbf{x}}$  be the solution to problem  $P$  under Assumption 2. We show that  $\dot{\mathbf{x}}$  is residual capacity fair. We rewrite (7) as follows:

$$\frac{w_i}{(R_i - \dot{x}_i)^\alpha} = \sum_{j \in \mathbf{J}_i} \mu_j, \quad i = 1, \dots, N \quad (12)$$

Multiplying (12) by  $(x_i - \dot{x}_i)$  and summing over  $i$ , we obtain

$$\begin{aligned} \sum_{i \in N} w_i \frac{x_i - \dot{x}_i}{(R_i - \dot{x}_i)^\alpha} &= \sum_{i \in N} \left( (x_i - \dot{x}_i) \sum_{j \in J_i} \mu_j \right) \\ &= (\mathbf{x} - \dot{\mathbf{x}})^T \mathbf{A} \boldsymbol{\mu} \\ &= \boldsymbol{\mu}^T \mathbf{A}^T (\mathbf{x} - \dot{\mathbf{x}}) \end{aligned} \tag{13}$$

Summing (8) over  $j$  and rearranging the terms, we find  $\boldsymbol{\mu}^T \mathbf{A}^T \dot{\mathbf{x}} = \boldsymbol{\mu}^T \mathbf{C}$ . Multiplying (4) by  $\boldsymbol{\mu}^T$ , we get  $\boldsymbol{\mu}^T \mathbf{A}^T \mathbf{x} \leq \boldsymbol{\mu}^T \mathbf{C}$ . Combining these relations, we see that

$$\boldsymbol{\mu}^T \mathbf{A}^T \mathbf{x} \leq \boldsymbol{\mu}^T \mathbf{C} = \boldsymbol{\mu}^T \mathbf{A}^T \dot{\mathbf{x}}$$

Therefore,  $\boldsymbol{\mu}^T \mathbf{A}^T (\mathbf{x} - \dot{\mathbf{x}}) \leq 0$ , and combining this inequality with (13), we establish that

$$\sum_{i \in N} w_i \frac{x_i - \dot{x}_i}{(R_i - \dot{x}_i)^\alpha} \leq 0$$

We have shown that the solution to problem  $P$  is residual capacity fair.  $\square$

### 5. PRICING FRAMEWORK

When we define the utility function in (1) and residual capacity fairness in previous section, two parameters,  $w_i$  and  $\alpha$ , are used in the definitions. These two parameters are closely related to pricing mechanisms that are integrated into our model. We are going to investigate pricing framework in this section.

We have shown that when congestion occurs at link  $j$  the Lagrange multiplier  $\mu_j > 0$  is a positive number in the solution to problem  $P$ . From microeconomic theory, we know that the Lagrange multipliers in (7) and (8) can be interpreted as the marginal costs for link capacity expansion at each link [27,29]. Therefore flow  $i$  may be charged a price  $\bar{p}_i$  that equals the average total marginal costs of the flow:

$$\bar{p}_i = \frac{1}{m_i} \sum_{j \in J_i} \mu_j, \mu_j \geq 0, i = 1, \dots, N \tag{14}$$

where  $m_i = |J_i|$  is the number of links that flow  $i$  uses. Note that  $\mu_j > 0$  only when link  $j$  is a bottleneck link. If there is no congestion on the route of flow  $i$ , the price is zero. The price is positive only when there is a bottleneck link on the route of flows, and a bottleneck link can only exist when the total bandwidth capacities requested by flows using the link exceed the link's capacity. Therefore, the price is a penalty for users' excessive bandwidth capacity demands. Although a user is penalized by charging the price, the network should guarantee that the user is allocated a bandwidth that satisfies the minimum bandwidth requirement. Hence, combining (7) and (14), and letting  $\dot{x}_i \leq r_i$ , yields  $\dot{x}_i = R_i - (w_i / m_i \bar{p}_i)^{1/\alpha} \geq r_i$ . Then, we obtain  $w_i \leq m_i \bar{p}_i (R_i - r_i)^\alpha$ . Note that we set  $w_i = m_i p_i (R_i - r_i)^\alpha$  in the definition of the user's utility function (1) and therefore we have

$$p_i \leq \bar{p}_i, i = 1, \dots, N \tag{15}$$

The weight  $w_i$  used in the utility function helps a user to assess the price charged to her or him. As long as the price is not greater than the average total marginal costs, the flow is allocated a bandwidth that is not less than the minimum bandwidth.

However, equation (14) is not a closed form of  $\bar{p}_i$ . The average total marginal costs are a function of all prices, say,  $\bar{p}_i = f(p_i, \dots, p_N)$ . Therefore, there are competitions among the flows. A price charged to flow

$i$  is considered a rationally charged price when it satisfies (15). Thus, the bandwidth allocated to a flow is a function of the prices:

$$\dot{x}_i = R_i - (R_i - r_i) \left( \frac{p_i}{\bar{p}_i} \right)^{1/\alpha}$$

If the flow is among a large number of flows and the impact of price change of the flow on  $\bar{p}_i$  is negligible, we see that the bandwidth of flow  $i$  is a decreasing function with respect to  $p_i$ . This reflects the fact that the price charged to a flow is a penalty for the congestions which are caused by excessive bandwidth capacity requirements. And when there is no congestion on the route of a flow, the price for the flow is zero because the marginal costs are zero.

To show the price competitions and the impact of parameter  $\alpha$ , we consider a typical situation of the network. That is, each flow goes through only one bottleneck link, and the other links the flows goes through are not bottleneck links. Assume that link  $j$  is a bottleneck link and it is the only bottleneck link on the route of flow  $i$  for each  $i \in \mathbf{I}_j$ . Thus, from (7) and (8), we have

$$\dot{x}_i = R_i - (R_i - r_i) \left( \frac{m_i p_i}{\mu_j} \right)^{1/\alpha}, i \in \mathbf{I}_j$$

$$\sum_{i \in \mathbf{I}_j} \dot{x}_i - C_j = 0, \mu_j > 0$$

Solving these equations, we obtain

$$\dot{x}_i = (1 - \beta_i) R_i + \beta_i r_i, i \in \mathbf{I}_j \quad (16)$$

where

$$\beta_i = \frac{\sum_{k \in \mathbf{I}_j} R_k - C_j}{\sum_{k \in \mathbf{I}_j} (R_k - r_k) (m_k / m_i)^{1/\alpha} (p_k / p_i)^{1/\alpha}}$$

This shows that the price ratios have an impact on the bandwidth allocation and there are price competitions among flows, because it is the price ratio changes that can cause bandwidth allocation change. Moreover, the parameter  $\alpha$  controls the sensitivity of the changes.

When  $\alpha \rightarrow \infty$ , we have  $\beta_\infty = \lim_{\alpha \rightarrow \infty} \beta_i$ , and

$$\beta_\infty = \frac{\sum_{k \in \mathbf{I}_j} R_k - C_j}{\sum_{k \in \mathbf{I}_j} (R_k - r_k)} \quad (17)$$

Therefore, price ratios do not have any impact on the bandwidth allocation when  $\alpha \rightarrow \infty$ . Note that for any feasible bandwidth requirements at the bottleneck link we have  $\sum_{k \in \mathbf{I}_j} R_k \geq C_j \geq \sum_{k \in \mathbf{I}_j} r_k$ ; therefore we see that  $0 \leq \beta_\infty \leq 1$ . Consequently, the bandwidth allocation given by (16) always satisfies the bandwidth requirements of each flow (i.e.,  $R_i \geq x_i \geq r_i$ , for all  $i \in \mathbf{I}_j$ ), and the bandwidth allocation is only determined by the link capacity and bandwidth requirements of flows. This leads to a non-monetary bandwidth allocation algorithm at the bottleneck link.

When  $\alpha$  is a small number, the bandwidth allocation is sensitive to prices. When overcharged prices are imposed on flow  $i$ , the parameter  $\beta_i$  in (16) is going to be greater than one and the network cannot allocate the flow a bandwidth that satisfies its minimum bandwidth requirement. Therefore, the network should not admit such flows with this price and an admission control mechanism should be integrated into the solution to problem  $P$ .

## 6. GLOBAL ADMISSION CONDITIONS

To obtain the solution to problem  $P$  given by (7) and (8), we relaxed the minimum bandwidth constraint (6). Also, the discussion in section 5 shows that irrational price ratios may result in a bandwidth

allocation that does not satisfy the minimum bandwidth constraints. When we apply the minimum bandwidth constraint (6) to the solution given by (7) and (8), we obtain a global admission conditions. From (7), we have

$$\begin{aligned} \dot{x}_i &= R_i - (R_i - r_i) \left( \frac{m_i p_i}{\sum_{j \in J_i} \mu_j} \right)^{\frac{1}{\alpha}} \\ &= (1 - \beta_i) R_i + \beta_i r_i \end{aligned}$$

where  $\beta_i = \left( m_i p_i / \sum_{j \in J_i} \mu_j \right)^{1/\alpha}$ . Let  $\rho_i = m_i p_i / \sum_{j \in J_i} \mu_j$ . Thus,  $\dot{x}_i$  is guaranteed to be between bandwidth capacity  $R_i$  and minimum bandwidth  $r_i$  when  $0 \leq \rho_i \leq 1$ .

**Proposition 2: global admission conditions** Each flow is charged a price  $p_i$ , and requests bandwidth capacity  $R_i$  and minimum bandwidth  $r_i$  for all  $i = 1, \dots, N$ . The network is able to allocate these flows bandwidths that satisfy their bandwidth requirements if the global admission conditions are satisfied:

$$0 \leq \rho_i \leq 1, i = 1, \dots, N \tag{18}$$

where

$$\begin{aligned} \rho_i &= m_i p_i / \sum_{j \in J_i} \mu_j, i = 1, \dots, N \\ m_i &= |J_i|, i = 1, \dots, N \end{aligned}$$

and the bandwidth allocation  $\dot{\mathbf{x}}$ , which is residual capacity fair, is given by

$$\dot{x}_i = (1 - \beta_i) R_i + \beta_i r_i, i = 1, \dots, N \tag{19}$$

$$\mu_j \left( \sum_{k \in I_j} \dot{x}_k - C_j \right) = 0, \mu_j \geq 0, j = 1, \dots, L \tag{20}$$

$$\beta_i = \rho_i^{1/\alpha}, i = 1, \dots, N \tag{21}$$

*Proof:* Equations (19) and (20) give a bandwidth allocation that is the solution to problem  $P$  with the relaxed constraint. By Proposition 1, this bandwidth allocation is residual capacity fair. Combining the global admission conditions (18), (19), and (21), the bandwidth allocation satisfies the bandwidth requirement,  $\mathbf{r} \leq \dot{\mathbf{x}} \leq \mathbf{R}$ , for all flows. Therefore, the bandwidth allocation satisfies the constraints of problem  $P$  and the network can admit these flows.  $\square$

The global admission conditions actually give the solution to problem  $P$ , which satisfies all constraints of problem  $P$ .

For a given set of flows with feasible minimum bandwidth requirements (i.e.,  $\mathbf{A}^T \mathbf{r} \leq \mathbf{C}$ ), prices charged to flows are the decision variables of the global admission conditions (18). From the discussion in the previous section, we know that a price charged to a flow is overcharged if it is greater than the average total marginal costs of the flow. The global admission conditions (18) will not be satisfied when overcharged prices are imposed to some flows. Therefore, the prices need to be adjusted to fall below the average total marginal costs. On the other hand, the marginal total cost is a function of prices of all flows. A flow cannot know that the price is overcharged without comparing it with other prices through the global admission conditions. Therefore, there are global price competitions to satisfy the global admission condition.

When there is no congestion in the network, we see that  $\mu_j = 0$ , for all  $j = 1, \dots, L$ . Thus,  $\rho_i$  is not well defined when there is no congestion. However, only constraint (5) is active when there is no congestion. Therefore, the solution to problem  $P$  is a bandwidth allocation that each flow is assigned a bandwidth that equals its bandwidth capacity. Also, no admission controls are needed. Prices do not play any role for admission decision and bandwidth allocation. The admission control is to guarantee that admitted flows are allocated bandwidths that are not less than their minimum bandwidths if there are congestions in the network.

To explain the admission conditions, we consider a simple network of a single link and two flows. Suppose that the link's capacity is  $C$ , flow 1 and flow 2 are charged a price  $p_1$  and  $p_2$ , and request bandwidth capacity  $R_1$  and  $R_2$ , minimum bandwidth  $r_1$  and  $r_2$ , respectively. We should also assume that  $R_1 + R_2 \geq C \geq r_1 + r_2$  such that it is feasible for these two flows to compete for sharing the link. When the two flows attempt to share the link, congestion occurs at the link. Thus, the global admission conditions are

$$0 \leq \rho_i \leq 1, i = 1, 2 \quad (22)$$

$$\dot{x}_i = (1 - \rho_i^{1/\alpha})R_i + \rho_i^{1/\alpha}r_i, i = 1, 2 \quad (23)$$

$$\dot{x}_1 + \dot{x}_2 = C \quad (24)$$

$$\rho_i = \frac{p_i}{\mu}, \mu > 0, i = 1, 2 \quad (25)$$

Combining (23), (24) and (25), we obtain

$$\rho_1^{1/\alpha} = \frac{R_1 + R_2 - C}{(R_1 - r_1)p_1^{1/\alpha} + (R_2 - r_2)p_2^{1/\alpha}} p_1^{1/\alpha}$$

$$\rho_2^{1/\alpha} = \frac{R_1 + R_2 - C}{(R_1 - r_1)p_1^{1/\alpha} + (R_2 - r_2)p_2^{1/\alpha}} p_2^{1/\alpha}$$

Substituting  $\rho_1$  and  $\rho_2$  into (22), and rearranging the terms, we see that these two flows are able to share the link if the following conditions are satisfied:

$$\left(\frac{p_1}{p_2}\right)^{1/\alpha} \geq \frac{R_1 + r_2 - C}{R_1 - r_1} \quad (26)$$

$$\left(\frac{p_2}{p_1}\right)^{1/\alpha} \geq \frac{r_1 + R_2 - C}{R_2 - r_2} \quad (27)$$

Hence, when the price ratio satisfies these conditions, condition (22) is satisfied, and the network can achieve a bandwidth allocation that satisfies users' bandwidth requirements (i.e.,  $R_i \geq \dot{x}_i \geq r_i$ ). When flow 1 is overcharged, condition (27) is violated. On the other hand, condition (26) is violated when flow 2 is overcharged. Therefore, a price that is rationally charged to one flow is related to the price charged to the other flow. When prices are charged such that condition (26) reaches equality, we have  $\rho_2 = 1$ , and  $\dot{x}_2 = r_2$  and  $\dot{x}_1 = C - r_2$ . Also, flow 2 is considered more heavily charged than flow 1, but it is still within the range considered to be rational. Conversely, when condition (27) reaches equality, we see that  $\rho_1 = 1$ , and  $\dot{x}_1 = r_1$  and  $\dot{x}_2 = C - r_1$ . When neither condition (26) nor (27) takes the equality, both flows are allocated a bandwidth between its bandwidth capacity and minimum bandwidth.

Because this example is just a very simple small network, we can evaluate the global admission conditions for prices charged. When prices are overcharged, the network is not going to admit the flows. When the network is very large, this centralized approach is, if not impossible, difficult to implement. Thus, a distributed admission approach should be considered.

## 7. DISTRIBUTED ADMISSION CONDITIONS

To implement the global admission conditions, a centralized algorithm has to be used for admission control and bandwidth allocation. All the flows need to have perfect information of the network to make

decisions on their prices. Unfortunately, it is very difficult to implement for large networks. We have to consider a distributed algorithm for admission control, which is also related to distributed bandwidth allocation. In this section, we present distributed admission conditions and bandwidth allocation.

We rewrite the social welfare function (2) as follows:

$$\begin{aligned}
 U(\mathbf{x}) &= \sum_{i \in \mathbf{N}} U_i(x_i) \\
 &= \sum_{i \in \mathbf{N}} \frac{m_i p_i (R_i - r_i)^\alpha}{(\alpha - 1)(R_i - x_i)^{\alpha-1}} \\
 &= \sum_{j \in \mathbf{L}} \sum_{i \in \mathbf{I}_j} \frac{p_i (R_i - r_i)^\alpha}{(\alpha - 1)(R_i - x_i)^{\alpha-1}} \\
 &= \sum_{j \in \mathbf{L}} U^j(\mathbf{x}^j)
 \end{aligned}$$

where

$$U^j(\mathbf{x}^j) = \sum_{i \in \mathbf{I}_j} \frac{p_i (R_i - r_i)^\alpha}{(\alpha - 1)(R_i - x_i)^{\alpha-1}}, j = 1, \dots, L$$

and

$$\mathbf{x}^j = (x_{i_1}^j, x_{i_2}^j, \dots, x_{i_\ell}^j)^T, i_k \in \mathbf{I}_j, 1 \leq k \leq \ell = |\mathbf{I}_j|$$

The vector  $\mathbf{x}^j$  represents the flows using link  $j$ . Thus, the social welfare function  $U(\mathbf{x})$  is the sum of social welfare function  $U^j(\mathbf{x}^j)$  of each link. And we know that bandwidth allocations are determined by the bottleneck links when there is congestion. The optimization problem of global function  $U(\mathbf{x})$  may be solved approximately by individual optimization problem of  $U^j(\mathbf{x}^j)$  at each link. Hence, we consider the optimization problem  $P^j$

$$\min U^j(\mathbf{x}^j) \quad P^j$$

subject to

$$\begin{aligned}
 \sum_{k \in \mathbf{I}_j} x_k^j &\leq C_j \\
 r_i &\leq x_i^j \leq R_i, i \in \mathbf{I}_j
 \end{aligned}$$

where  $j = 1, \dots, L$ . In the same way that we solve problem  $P$ , we first relax the minimum bandwidth constraint. The Lagrangian is given by

$$\mathcal{L}^j(\mathbf{x}^j, \mu^j, \boldsymbol{\lambda}) = U^j(\mathbf{x}^j) - \mu^j \left( \sum_{i \in \mathbf{I}_j} x_i^j - C_j \right) - \boldsymbol{\lambda}^{jT} (\mathbf{x}^j - \mathbf{R}^j)$$

The first-order Kuhn–Tucker conditions [27] are

$$\begin{aligned}
 \frac{p_i (R_i - r_i)^\alpha}{(R_i - x_i)^\alpha} - \mu_j - \lambda_i &= 0, i \in \mathbf{I}_j \\
 \lambda_i (x_i^j - R_i) &= 0, \lambda_i \geq 0, i \in \mathbf{I}_j \\
 \mu_j \left( \sum_{k \in \mathbf{I}_j} x_k^j - C_j \right) &= 0, \mu_j \geq 0
 \end{aligned}$$

Thus, we find that the bandwidth allocated to flow  $i$  at link  $j$  is given by

$$\hat{x}_i^j = \begin{cases} R_i, & \sum_{k \in \mathbf{I}_j} \hat{x}_k^j < C_j \\ R_i - (R_i - r_i) \left( \frac{p_i}{\mu^j} \right)^{1/\alpha}, & \mu^j > 0, \sum_{k \in \mathbf{I}_j} \hat{x}_k^j = C_j \end{cases} \quad (28)$$

where  $i \in \mathbf{I}_j$  and  $j = 1, \dots, L$ .

However, a flow goes through several links, and each link may allocate the flow different bandwidths when each of these links allocates the flow a bandwidth independently. All the links can accommodate the flow when the flow takes the minimum value among the bandwidths allocated by these links. Hence, we obtain the bandwidth  $\bar{x}_i$  allocated to flow  $i$  as follows:

$$\bar{x}_i = \min \{ \hat{x}_i^j | j \in \mathbf{J}_i \}, i = 1, \dots, N \quad (29)$$

where  $\hat{x}_i^j$  is given by (28).

The bandwidth allocation  $\bar{x}$  may not be the same as the bandwidth allocation  $\hat{x}$ , which is given in global admission conditions. The reason to pursue the bandwidth allocation  $\bar{x}$  is not because it is the optimal solution to problem  $P$ , but because it satisfies the residual capacity fairness criterion. In the same way as we prove  $\hat{x}$  satisfies the residual capacity fairness criterion, we can also prove  $\bar{x}$  satisfies local residual capacity fairness at link  $j$ . That is, for any other feasible bandwidth allocation  $x^j$  at link  $j$ , we have

$$\sum_{k \in \mathbf{I}_j} \hat{w}_k \frac{x_k^j - \hat{x}_k^j}{(R_i - \hat{x}_k^j)^\alpha} \leq 0$$

where  $\hat{w}_k = p_k(R_k - r_k)^\alpha$  for each  $k \in \mathbf{I}_j$ . Therefore, the bandwidth  $\bar{x}_i$  given by (29) satisfies the local residual capacity fairness criterion at a bottleneck link. Bandwidth allocation  $\hat{x}$  and  $\bar{x}$  can be thought of as satisfying different fairness criteria. The question of which fairness criterion is more 'fair' is arguable. However, both fairness criteria can avoid the obvious 'unfair' bandwidth allocation given by (9), which allocates some flows bandwidths that reach their full bandwidth capacities and at the same time some other flows are allocated zero bandwidths.

Bandwidth allocation  $\bar{x}$  may not be as fair as bandwidth allocation  $\hat{x}$ . The advantage of bandwidth allocation  $\bar{x}$  is that distributed algorithms can be developed to implement the bandwidth allocation, which also leads to distributed admission controls. Note that the solution given in (28) needs only the local information of link  $j$ . Each link is able to allocate bandwidth to the flows with the local information of the link. The network allocates flow  $i$  the bandwidth  $\bar{x}_i$  given by (29), which can be implemented by a round trip probe on the route of the flow.

**Proposition 3: distributed admission conditions** Each flow at link  $j$  is charged price  $p_i$  and requests bandwidth capacity  $R_i$  and minimum bandwidth  $r_i$ , for all  $i \in \mathbf{I}_j$ :

- (a) if  $\sum_{k \in \mathbf{I}_j} R_k < C_j$ , link  $j$  is able to admit the flows and allocate flow  $i$  bandwidth  $\hat{x}_i^j = R_i$ , for all  $i \in \mathbf{I}_j$ ;
- (b) if  $\sum_{k \in \mathbf{I}_j} R_k \geq C_j$ , link  $j$  is able to admit the flows and the bandwidth allocation satisfies bandwidth requirements of the flows if the following conditions are satisfied:

$$\sum_{k \in \mathbf{I}_j} r_k \leq C_j \quad (30)$$

$$0 \leq \beta_i^j \leq 1, i \in \mathbf{I}_j \quad (31)$$

where

$$\beta_i^j = \frac{\sum_{k \in \mathbf{I}_j} R_k - C_j}{\sum_{k \in \mathbf{I}_j} (R_k - r_k) (p_k / p_i)^{1/\alpha}}$$

and link  $j$  is able to allocate flow  $i$  a bandwidth as follows:

$$\hat{x}_i^j = (1 - \beta_i^j) R_i + \beta_i^j r_i, i \in I_j \tag{32}$$

*Proof:* When  $\sum_{k \in I_j} R_k < C_j$  at link  $j$ , it is obvious as it stands in (a). We only need to show that when conditions (30) and (31) are satisfied, the bandwidth allocation satisfies the bandwidth requirements of the flows. The bandwidth allocation has a feasible solution to problem  $P^j$  when condition (30) is satisfied. Suppose condition (30) is satisfied, eliminating  $\mu^j$  from (28) gives a bandwidth allocation in (32). Applying the minimum bandwidth constraints to bandwidth allocation (32) and letting  $\hat{x}_i^j \geq r_i$ , for all  $i \in I_j$ , we obtain condition (31). Hence, when condition (31) is satisfied, the bandwidth allocation meets the bandwidth requirements of flows, i.e.,  $r_i \leq \hat{x}_i^j \leq R_i$ , for all  $i \in I_j$ . Therefore, the link is able to admit the flows and the bandwidth allocation is given in (32).  $\square$

From the distributed admission conditions, we can easily develop distributed admission controls and bandwidth allocation algorithms. First, we look into a simple non-monetary algorithm. Let  $\beta_\infty^j = \lim_{\alpha \rightarrow \infty} \beta_i^j$ , for all  $i \in I_j$  at link  $j$ , we have

$$\beta_\infty^j = \frac{\sum_{k \in I_j} R_k - C_j}{\sum_{k \in I_j} (R_k - r_k)} \tag{33}$$

Note that  $\beta_\infty^j$  is identical for all  $i \in I_j$  at link  $j$ . Also,  $\beta_\infty^j$  satisfies condition (31) when condition (30) holds. Therefore, a link can admit flows and allocate them bandwidths given by (32) when condition (30) is satisfied. Because prices do not play a role in (33), it is a non-monetary algorithm. After each link on the route of a flow admits the flow, the flow is allocated a bandwidth given by (29), and it satisfies the bandwidth requirements of the flow.

The distributed admission conditions can also help us to develop an algorithm for DiffServ. Before now, we only considered per-flow models, which can apply to Integrated Services architecture [30]. It is worth emphasizing that our model does not preclude DiffServ. Flows in DiffServ are marked into a small number of service classes. Flows in the same service class are treated equally. Now, we consider a network that accommodates two service classes: class 1 and class 2. A flow in class 1 and class 2 requests bandwidth capacity  $R_1$  and  $R_2$ , minimum bandwidth  $r_1$  and  $r_2$ , and is charged  $p_1$  and  $p_2$ , respectively. From the distributed admission conditions, we obtain the admission conditions at link  $j$  as follows:

$$\begin{aligned} n_1 r_1 + n_2 r_2 &\leq C_j \\ r_1 n_1 + \left( R_2 - (R_2 - r_2) \left( \frac{p_2}{p_1} \right)^{1/\alpha} \right) n_2 &\leq C_j \\ \left( R_1 - (R_1 - r_1) \left( \frac{p_1}{p_2} \right)^{1/\alpha} \right) n_1 + r_2 n_2 &\leq C_j \\ n_1 \geq 0, n_2 &\geq 0 \end{aligned}$$

where  $n_1$  and  $n_2$  are the numbers of flows in class 1 and class 2, respectively. These admission conditions can be represented by two systems of inequalities:

$$\begin{cases} p_1 \geq p_2 \\ r_1 n_1 + \left( R_2 - (R_2 - r_2) \left( \frac{p_2}{p_1} \right)^{1/\alpha} \right) n_2 \leq C_j \\ n_1 \geq 0, n_2 \geq 0 \end{cases} \tag{34}$$

and

$$\begin{cases} p_1 \leq p_2 \\ \left( R_1 - (R_1 - r_1) \left( \frac{p_1}{p_2} \right)^{1/\alpha} \right) n_1 + r_2 n_2 \leq C_j \\ n_1 \geq 0, n_2 \geq 0 \end{cases} \quad (35)$$

Each of the systems defines a set of  $(n_1, n_2)$  depicted in Figure 1. The shaded ranges are sets that satisfy (34) and (35), respectively. When  $(n_1, n_2)$  is in the shaded ranges, link  $j$  is able to admit the flows and allocate bandwidth given by

$$\hat{x}_i^j = (1 - \beta_i) R_i + \beta_i r_i$$

$$\beta_i = \frac{n_1 R_1 + n_2 R_2 - C_j}{n_1 (R_1 - r_1) p_1^{1/\alpha} + n_2 (R_2 - r_2) p_2^{1/\alpha}} p_i^{1/\alpha}$$

where  $i = 1, 2$ , and  $\hat{x}_i^j$  is guaranteed to satisfy the bandwidth requirements of flows.

Because each class involves multiple users, each user should only be a price taker. The network can implement a static price for each class. Suppose that class 1 is the high QoS class such that  $R_1 \geq R_2$  and  $R_1 - r_1 \leq R_2 - r_2$ . A flow in class 1 should be charged a higher price such that

$$\frac{R_2 - r_2}{R_1 - r_1} \geq \left( \frac{p_1}{p_2} \right)^{1/\alpha} \geq 1 \quad (36)$$

Then, we have

$$(R_2 - r_2) p_2^{1/\alpha} - (R_1 - r_1) p_1^{1/\alpha} \geq 0$$

$$R_2 - R_1 \leq 0$$

Hence

$$\left( (R_2 - r_2) p_2^{1/\alpha} - (R_1 - r_1) p_1^{1/\alpha} \right) \left( \frac{1}{\mu^j} \right)^{1/\alpha} \geq R_2 - R_1$$

Rearranging the terms and substituting (28) into it, we obtain  $\hat{x}_1^j \geq \hat{x}_2^j$ . Therefore, at a bottleneck link a flow in class 1 is always allocated a bandwidth that is not less than that of a flow in class 2. Moreover, class 1 is charged a higher price when relations (36) hold. Therefore, relations (36) can help the network to devise static pricing policies for these two service classes.

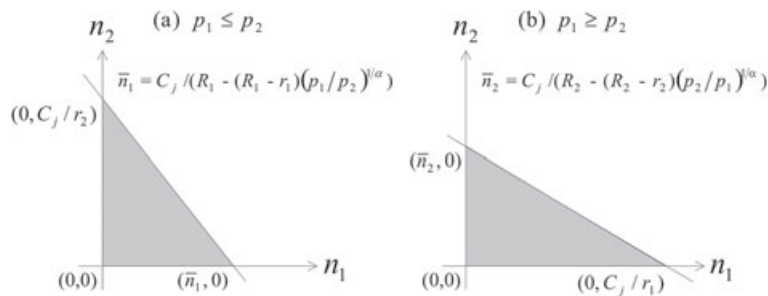


Figure 1. Feasible numbers of flows

We have shown that a simple distributed bandwidth allocation and admission control with a static pricing policy can be developed from the distributed admission conditions. Although we only considered two service classes, more service classes can be considered easily in the same way.

## 8. SIMULATION AND EVALUATION

In this section, we first show examples that illustrate the discussion in previous sections, and then present the simulation results on the relative difference between the centralized and distributed bandwidth allocations.

Consider a network that consists of one link and four flows where the link capacity is 1 and the minimum bandwidths of all flows are zero. The price charged to each flow in this case is  $p_i + 1 + R_i^\tau$ , where  $0 \leq \tau \leq 1$  is a constant. Hence, the price is composed of a fixed connection fee and an increasing concave function of requested bandwidth capacity of the flow. Tables 1 and 2 show bandwidth allocations with  $\tau = 0.5$  and  $\tau = 1$ , respectively. When  $\alpha = 1.01$  or  $\alpha = 2$ , the bandwidth allocations in Tables 1 and 2 are different because different pricing policies are used. When  $\alpha = 50$ , we see that prices have little impact on the bandwidth allocations. From (16) and (17), we obtain  $\hat{x}_i = R_i C / \sum_{k=1}^4 R_k = 0.4R_i$  for  $i = 1, \dots, 4$  when  $\alpha \rightarrow \infty$ . That is, each flow is allocated a bandwidth that is 40% of its bandwidth capacity and the bandwidth allocation is (0.1, 0.2, 0.3, 0.4) when  $\alpha \rightarrow \infty$ . Therefore, when  $\alpha$  is large enough, the bandwidth allocated to each flow is approximately 40% of its requested bandwidth capacity.

In Figure 2, prices charged to flows 1, 2, and 3 are those shown in column 3 in Table 1; the  $x$ -axis represents the price of flow 4 and the  $y$ -axis represents allocated bandwidth. In Figure 2(a) and (b),  $\alpha = 1.01$  and  $\alpha = 50$ , respectively. While price charged to flow 4 increases, the bandwidth allocated to flow 4 decreases in Figure 2(a). Note that the prices of flows 1, 2, and 3 in Figure 2 are fixed values. The bandwidths of these flows (dotted lines) increase in Figure 2(a) because the bandwidth of flow 4 decreases. Bandwidths in Figure 2(b) do not change much while the price of flow 4 changes. Therefore, we see that the price charged to a flow has a strong impact on the bandwidth allocated to the flow when  $\alpha$  is small. On the other hand, prices have little impact on bandwidth allocation when  $\alpha$  is large enough. With appropriately chosen  $\alpha$ , the network model can balance the impact of prices on bandwidth allocations.

Flow	$R$	Price	Bandwidth		
			$\alpha = 1.01$	$\alpha = 2$	$\alpha = 50$
1	0.25	1.50	0.128	0.115	0.101
2	0.50	1.71	0.223	0.212	0.200
3	0.75	1.87	0.296	0.298	0.300
4	1.00	2.00	0.352	0.376	0.399

Table 1. Bandwidth allocation with  $\tau = 0.5$

Flow	$R$	Price	Bandwidth		
			$\alpha = 1.01$	$\alpha = 2$	$\alpha = 50$
1	0.25	1.25	0.142	0.123	0.101
2	0.50	1.50	0.242	0.222	0.201
3	0.75	1.75	0.300	0.299	0.300
4	1.00	2.00	0.315	0.357	0.398

Table 2. Bandwidth allocation with  $\tau = 1$

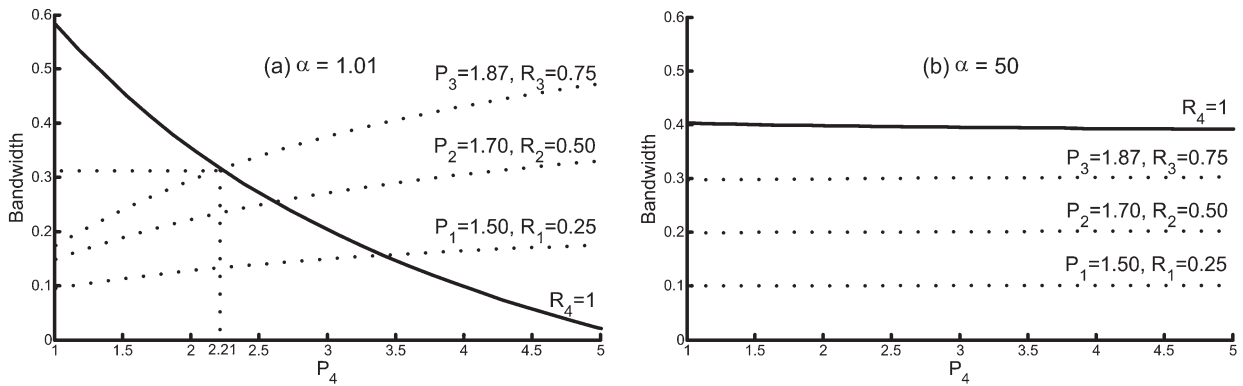


Figure 2. Bandwidth and price of flow 4

Although user adaptations are not in the scope of this paper, the pricing framework integrated in the network model gives some intuition for user adaptations and pricing policies. For example, when the network adopts a pricing policy  $\mathbf{p} = (1.50, 1.71, 1.87, p_4)$  such that  $p_4 > 2.21$ , we see, in Figure 2(a), that the bandwidth allocated to flow 4 is less than that of flow 3. Hence, it is rational to assume that user 4 will request a bandwidth capacity of 0.75 rather than 1. Thus, the price charged to flow 4 should not be greater than 2.21, so that user 4 is allocated a bandwidth that is greater than that of flow 3 and therefore user 4 may still request the bandwidth capacity of 1.

To show the relative difference between the centralized and distributed bandwidth allocations, we simulate a network of 10 links and 25 flows with random configurations (Note that simulations on a network with a large number of links are infeasible because computation complexity of global admission conditions is  $O(2^L)$ .) Each link capacity is randomly generated such that  $0.75 \leq C_j \leq 1$ . The bandwidth capacity and minimum bandwidth for each flow are random numbers such that  $0.05 \leq R_i \leq 0.25$  and  $0 \leq r_i \leq 0.5R_i$ . Matrix  $\mathbf{A}$  is also randomly generated such that each link is used by at least one flow, and each flow uses at least one link, and  $A_{ij} = B(p)$ , where  $B(p)$  is single Bernoulli trial with success probability  $p = L^{-1/2}$ . With this probability, the expected number of links that a flow goes through is  $L^{1/2}$  and therefore a flow goes through a reasonable number of links (i.e., the diameter of a 2D grid). For each random configuration, we simulate the global and distributed admission conditions to obtain bandwidth allocation  $\mathbf{x}_g$  and  $\mathbf{x}_d$  that satisfies global and distributed admission conditions, respectively. The relative difference is calculated as  $\sigma = |\mathbf{x}_g - \mathbf{x}_d| / |\mathbf{x}_g|$ . Figure 3 shows the complement cumulative density function (CCDF) of the relative difference between global/centralized and distributed results. In other words, CCDF (y-axis) gives the likelihood  $P(\sigma > c)$ , where  $c$  is a certain relative difference on the x-axis. In Figure 3, the four curves represent different congestion degrees based on the percentage of flows that goes through at least one bottleneck link. For example, the fourth curve represents that 40–50% of the flows uses at least one bottleneck link. Therefore, the distributed scheme performs very well when the ratio of congested flows to the total flows in any network is less than 50%, which is a very realistic bound in practice.

## 9. CONCLUSION AND FUTURE WORK

In this paper, we have presented a theoretical framework for bandwidth allocation and admission control at network links to meet network users' bandwidth requirements with a price charged by a network service provider. A utility function was defined to capture the bandwidth demands of network users when users are charged prices for certain bandwidth capacities. An optimization framework leads to a fair bandwidth allocation and global admission conditions, in which the prices are the decision variables

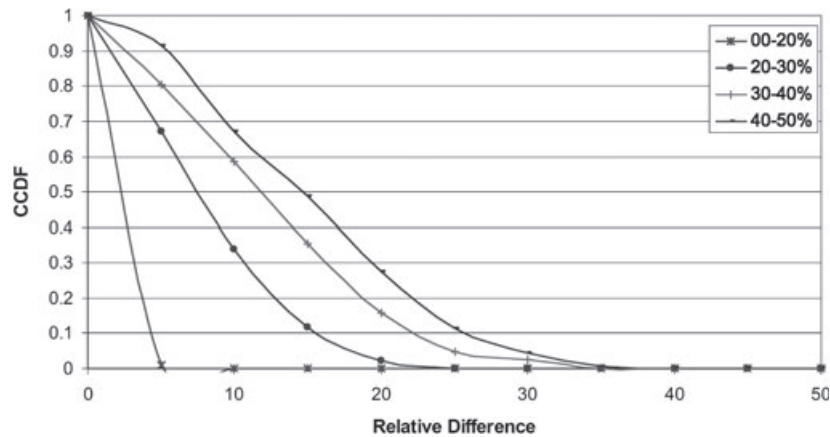


Figure 3. CCDF of relative difference

for given sets of feasible bandwidth requirements. Although the implementation of global admission conditions for a large network is still an open problem, we provided distributed admission conditions and bandwidth allocations, which are easy to implement. A non-monetary bandwidth allocation algorithm with a simple admission control is also given and is easy to implement as well. We have shown that when the network only needs to accommodate a small number of service classes, a bounded set of the numbers of flows that are able to be admitted into the network can be easily deduced from the distributed admission condition, and a static pricing policy is developed to charge flows in different services classes.

The simulations demonstrate that the model is well tunable using parameter  $\alpha$  and the prices charged to flows. Rational price ranges can also be obtained through price competitions. Moreover, the distributed approach was shown to perform very well under reasonable and normal conditions with a relative difference less than 10–20% from the global centralized solution.

In this paper, only stationary bandwidth requirements are considered. Dynamic bandwidth requirements and user adaptations should be considered in future work. Another aspect of our model is that the service provider is assumed to be in favor of optimizing the social welfare without having an agenda of its own. Such cases will also be addressed in future research.

## REFERENCES

1. Paschalidis C, Tsitsiklis JN. Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking* 2000; **8**(2): 171–184.
2. Paschalidis C, Liu Y. Pricing in multiservice loss networks: static pricing, asymptotic optimality, and demand substitution effects. *IEEE/ACM Transactions on Networking* 2002; **10**(3): 425–438.
3. Keon NJ, Anandalingam G. Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees. *IEEE/ACM Transactions on Networking* 2003; **11**(1): 66–80.
4. Savagaonkar U, Chong EKP, Givan RL. Online pricing for bandwidth provisioning in multi-class networks. *Computer Networks Journal* 2004; **44**: 835–853.
5. Wang X, Schulzrinne H. Pricing network resources for adaptive applications in a differentiated services network. In *Proceedings of IEEE INFOCOM 2001*, Anchorage, AK, 2001.
6. Blake S, Black D, Carlson M, Davies E, Wang Z, Weiss W. An architecture for differentiated services. *RFC 2475*, December 1998.
7. Wang X, Schulzrinne H. RNAP: a resource negotiation and pricing protocol. In *International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'99)*, Basking Ridge, NJ, 1999.

8. Kelly F, Weber R. Measurement-based charging in communication networks. *Operations Research* 2000; **48**: 535–548.
9. Heikkinen T, Multidimensional QoS and charging in a packet data network. In *2001 IEEE International Conference on Multimedia and Expo*, 2001; 282.
10. Ferrari D, Delgrossi L. Charging for QoS. In *Proceedings of IEEE/IFIP International Workshop on Quality of Service*, Napa Valley, CA, 18–20 May 1998.
11. Semret N, Liao RR-F, Campbell AT, Lazar AA. Pricing, provisioning and peering: dynamic markets for differentiated internet services and implications for network interconnections. *IEEE Journal on Selected Areas in Communications* 2000; **18**(12): 2499–2513.
12. Yaïche H, Mazumdar RR, Rosenberg C. A game theoretic framework for bandwidth allocation and pricing in broadband networks. *IEEE/ACM Transactions on Networking* 2000; **8**(5): 667–678.
13. Muthoo A. *Bargaining Theory with Applications*. Cambridge University Press: Cambridge, UK, 1999.
14. Nash J. The bargaining problem. *Econometrica* 1950; **18**: 155–162.
15. Kelly FP, Maulloo A, Tan D. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society* 1998; **49**: 237–252.
16. Mo J, Walrand J. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking* 2000; **8**(5): 556–567.
17. Bonald T, Massoulié L. Impact of fairness on Internet performance. In *Proceedings of the 2001 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, Vol. 29(1), 2001.
18. Massoulié L, Roberts J. Bandwidth sharing: objectives and algorithms. *IEEE/ACM Transactions on Networking* 2002; **10**(3): 320–328.
19. Bertsekas D, Gallager R. *Data Networks*. Prentice-Hall: Englewood Cliffs, NJ, 1987.
20. De Veciana G, Lee T-J, Konstantopoulos T. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking* 2001; **9**(1): 2–14.
21. Kelly FP, Williams RJ. Fluid model for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability* 2004; **14**: 1055–1083.
22. Roberts J, Massoulié L. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems* 2000; **15**: 185–201.
23. Bu T, Towsley D. Fixed point approximation for TCP behavior in an AQM network. *Proceedings of ACM Sigmetrics* 2001; **29**(1): 216–225.
24. Liu Y, Presti FL, Misra V, Towsley DF, Gu Y. Scalable fluid models and simulations for large-scale IP networks. *Transactions on Modeling and Computer Simulation* 2004; **14**(3): 305–324.
25. Alpcan T, Başsar T. A utility-based congestion control scheme for Internet-style networks with delay. In *Proceedings of IEEE INFOCOM 2003*, San Francisco, CA, 2003.
26. Harsanyi JC. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 1955; **63**(4): 309–321.
27. Hoy M, Livernois J, McKenna C, Rees R, Stengos T. *Mathematics for Economics* (2nd edn). MIT Press: Cambridge, MA, 2001.
28. Fudenberg D, Tirole J. *Game Theory*. MIT Press: Cambridge, MA, 1996.
29. Varian HR. *Microeconomic Analysis* (3rd edn). Norton: New York, 1992.
30. Braden R, Clark D, Shenker S. Integrated services in the Internet architecture: an overview. *RFC 1633*, June 1994.

#### AUTHOR'S BIOGRAPHIES

**Yonghe Yan** holds a PhD from the University of Hong Kong, a graduate degree from the University of Electronic Science and Technology of China and an undergraduate degree from the Chengdu Institute of Radio Engineering. Yonghe has delved into electronic commerce and information economics, artificial intelligence, expert systems and intelligent agents, distributed and collaborative computing, microeconomics and game theory. He chose to shift his career to the industry at 2007 when he left the School of Computing at DePaul University and joined Conoco-Philips.

**Adel El-Atawy** is a PhD candidate in the School of Computing at DePaul University. He received his MSc and BSc degrees in Computer Science from Alexandria University, Egypt, in 2003 and 2000 respectively. While in Egypt he

also co-founded eSpace, a leading software house in the Middle East. Currently, he is a research assistant at DePaul University. His research revolves around using information theoretic analysis and statistical techniques in network security systems.

**Ehab Al-Shaer** is an associate professor and the director of the Security and Multimedia Networking Research Lab (SMNLAB) in the School of Computing at DePaul University. Prof. Al-Shaer received his PhD in Computer Science, MS in Computer Science University, and BS in Computer Engineering from Old Dominion University, Northeastern University and KFUPM in 1998, 1994, and 1990 respectively. His primary research areas are firewall optimization, configuration management, and fault diagnosis. Prof. Al-Shaer has co-edited six books and published more than 80 refereed articles. He was also the General Chair for the *16th ACM Conference on Computer and Communication Security (CCS)* in 2009. Prof. Al-Shaer has served as a Program Co-chair for number of conferences, including Automated Network Management (ANM-INFOCOM 2008), POLICY 2008, Integrated Management (IM 2007), MMNS 2001, and E2EMON 2004–2005. He has received a number of Best Paper and Fellowship awards. His research is supported by NSF, Intel, Cisco and Sun Microsystems.