Fair Admission Control to Achieve Guaranteed Bandwidth Allocation

Yonghe Yan
School of Computer Science, Telecommunication, and Information Systems
DePaul University
246 South Wabash, Chicago, IL 60604

Abstract

In this paper, we present a theoretic framework for admission controls and bandwidth allocations at network links to achieve guaranteed bandwidth allocations, which guarantee to allocate admitted flows data rates that are above their required minimum bandwidths. The admission control and bandwidth allocation are devised from the optimal solution of maximizing aggregate social welfare of the network. We define a utility function to capture network user's demand for guaranteed bandwidth requirements with a price charged by a network service provider at the edge of the network. A fairness criterion is introduced for network links to allocate bandwidths. We first consider global admission conditions, which are deduced from the social welfare maximization problem, and then present distributed admission conditions, which can be used by each network link to make admission decisions locally. The bandwidth allocation resulted from the distributed admission conditions is asymptotically optimal with respect to the bandwidth allocation resulted from the global admission conditions. We show that the admission control framework can provide guidance for network service providers to charge users that require guaranteed bandwidths for data transmissions.

1. Introduction

During last several years, Quality of Service (QoS) issues in the Internet has attracted significant research interests. QoS provisioning requires the next-generation Internet to have the capability to provide differentiated services and accommodate differentiated classes of service to support various types of applications and business requirements. Fair bandwidth allocation and pricing of QoS services are becoming increasingly important. It provides sufficient incentives: (i) for users to use the network resources efficiently and (ii) for service providers to provide guaranteed QoS services in a healthy market environment.

In this paper, we consider a communication network with fixed routing that accommodates multiple service classes. Each service class has different bandwidth requirements for guaranteed QoS. A link in the network has a given finite bandwidth capacity and the total data rates of all flows using the link cannot exceed the link's bandwidth capacity. Each link should assign a flow using the link a share of its bandwidth capacity in compliance with a fairness criterion, and guarantee data rate of the flow is above a minimum bandwidth. We propose a model that can devise a fairness criterion for links to assign a flow a fair share of their bandwidth capacities with a guaranteed minimum bandwidth. The model also allow us to investigate insight properties of prices among various service classes, and allocate bandwidth fairly such that QoS requirements are guaranteed while total social welfare is maximized, or asymptotically maximized. These properties give guidance for service providers to price differentiated service classes, which have different bandwidth requirements.

Several models have been put forth for service providers to sell bandwidth capacities. Paschalidis et al [5, 6] presented a revenue maximization problem for a service provider to charge each bandwidth requirement a static price. In [2], optimal prices and admission control policy are given from a global revenue maximization problem. Savagaonkar et al [7] investigated revenue maximization problem of dynamic pricing for bandwidth provisioning with an assumption that user demands are known stochastic processes. Semret et al [10] considered a bandwidth capacity provisioning framework to allocate bandwidths in a dynamic market. In these researches, bandwidth capacities are offered at the edge of a network. The problem about how to deliver the sold bandwidth capacities in the core of the network is still open. We consider a solution to the problem of delivering the services that have already been sold to users with service level agreements. Note that the network users can be end-users or next-level service providers that may sell their bandwidth capacities bought to other users.

A network user is charged for a finite bandwidth capacity offered from a service provider under a service level agreement. The bandwidth capacity bought is actually the maximum data rate that the network user can transmit data through the network. The service level agreement between network user and service provider, either explicitly or implicitly, requires the network to allow the user to transmit data above a minimum data rate, so that the bandwidth perceived by the user will not be too much lower than the bandwidth capacity bought by the user. This also implies that expected bandwidths for data transmissions are always in an interval between the minimum bandwidth and the maximum bandwidth (i.e., bandwidth capacity bought by the user). Clark [16] has pointed out that users should be charged by the expected bandwidths. The higher minimum bandwidth a user is able to transmit his/her data, the higher expected bandwidth the user perceives. Therefore users' QoS requirements will be guaranteed to be satisfied with respect to a guaranteed minimum bandwidth.

To satisfy various maximum and minimum bandwidth requirements of concurrent flows, networks should use all available bandwidth to the fullest while maintaining certain fairness in allocations to these flows. Yaïche et al [19] proposed a fair bandwidth allocation from the solution of Nash bargaining problem [20, 21]. Kelly et al [8] introduced a notion of proportional fairness for "elastic" network (i.e., best-effort service network), where user total utility is maximized when the bandwidth allocation fairness is achieved. The fair bandwidth sharing is further generalized to be weighted α-bandwidth allocation in [3, 13]. The allocation corresponds to the maximum throughput fairness [13] when $\alpha \to 0$, the max-min fairness [12] when $\alpha \to \infty$, the proportional fairness [8] when $\alpha \to 1$, and the minimum potential delay fairness [11] when $\alpha \to 2$. A class of utility functions of data rate on the interval $(0,\infty)$ is used in fluid model in their analysis of α-bandwidth allocation. Similar fluid models are also used in [4, 9, 15, 17, 22] for bandwidth allocation analysis in elastic networks. Alpcan and Başşar [14] proposed a broad class of utility functions that are non-decreasing and strictly concave on data rates 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, in which users have been charged for finite bandwidth capacities with guaranteed minimum bandwidth.

We propose a utility function that captures user demand for bandwidth in inelastic networks and deduce key characteristics for fair bandwidth allocation from fluid model. These characteristics give guidance for devising admission control policies and pricing different service classes. Although a fluid model is used in our analysis, it is worth emphasizing that our per-flow model does not preclude Differentiated Services architecture (DiffServ) [1]. On the contrary, it can provide guidance on how to price different service classes in DiffServ. Flows in DiffServ are marked into a small number of service classes. A link is expected to treat flows in a same service class equally and flows belonging to different classes differently. Each flow in Integrated Services (IntServ) architecture [18] is treated as one service class by a link and therefore each flow is processed differently in InterServ.

The paper is organized as follows: In Section 2, we present the network model used in the paper and the bandwidth allocation optimization problem for flows which are constrained with both their bandwidth capacities and minimum bandwidths, and a utility function is introduced in this section as well. Section 3 presents a notion of fairness for the optimal bandwidth allocation. In section 4, we discuss the utility function that is used in the network model of this paper. Section 5 considers the global admission conditions and the optimal bandwidth allocation for admitted flows. Section 6 presents distributed admission conditions and distributed bandwidth allocation. Finally, Section 7 draws conclusions.

2. Network Model

We consider a network as 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 (connection) in the network. The user has been charged a price $p_i > 0$ by a network service provider for using the network. The user is provided a bandwidth capacity R_i with a guaranteed minimum bandwidth r_i, so that the user can transmit data with a data rate of x_i, where $R_i \ge x_i \ge r_i \ge 0$. 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 also $\mathbf{J}_i = \{j \mid A_{ij} = 1\}$ be the set of links that flow i uses and $\mathbf{I}_j = \{i \mid A_{ij} = 1\}$ be the set of flows that use link j.

We consider a fluid model of the network where the packets are infinitely divisible and small. After a user has bought a bandwidth capacity from a service provider, the user's objective is to maximize the following utility function with respect to x_i over $[r_i, R_i]$:

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

where $\alpha > 1$ corresponds to a class of utility functions. We call the parameter α intensity factor. The parameter $w_i(\cdot)$ is a positive number, which is the weight of user's utility. The parameter $w_i(\cdot)$ is a function of the price charged by a service provider and the QoS purchased by the user. We will explain the parameter α and introduce $w_i(\cdot)$ formally later in the next section. Apparently, this utility function reflects user's prospective. After the user has bought bandwidth capacity R_i, the maximum bandwidth the user expected is R_i because $U_i(x_i) \to \infty$ when $x_i \to R_i$.

To best satisfy all the users of the network, the network's objective is to maximize the social welfare of the network. Therefore, the optimal bandwidth allocation $\dot{\mathbf{x}}$ that maximize the social welfare are the solution of the optimization problem *P*:

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

subject to

$$\mathbf{A}^T \mathbf{x} \le \mathbf{C} \qquad (2)$$
$$\mathbf{x} \le \mathbf{R} \qquad (3)$$
$$\mathbf{x} \ge \mathbf{r} \qquad (4)$$

where

$$\mathbf{x} = (x_1, \cdots, x_N)^T$$
$$\mathbf{C} = (C_1, \cdots, C_L)^T$$
$$\mathbf{R} = (R_1, \cdots, R_N)^T$$
$$\mathbf{r} = (r_1, \cdots, r_N)^T.$$

The inequality (2) expresses the link capacity constraints, that is, the aggregate of data rates of flows that use a link cannot exceed the capacity of the link. Bandwidth requirements of each flow are represented by inequality (3) and (4).

We assume that the sum of minimum bandwidths of flows that go through a link cannot exceed the capacity of the link (i.e. $\mathbf{A}^T \mathbf{r} \le \mathbf{C}$), otherwise the solution set defined by inequalities (2), (3), and (4) is empty and thus there is no solution for Problem *P*.

Assumption 1: The bandwidth requirements of flows are feasible for the network, that is, $\mathbf{A}^T \mathbf{r} \le \mathbf{C}$.

We further relax the constraint (4) to be $\mathbf{x} \ge 0$, and will apply it later in global admission conditions. Hence the optimization problem *P* is defined on the set that is nonempty, convex, and compact. It implies that the solution to optimization problem *P* exists on the set defined on the relaxed constraints.

Apparently, when the bandwidth capacities of the links of the network are over-provisioned, meaning that $\mathbf{A}^T \mathbf{R} \le \mathbf{C}$; the optimal solution to the Problem *P* is $\dot{\mathbf{x}} = (R_1, \cdots, R_N)$. It implies that there are no congestions in the network. We believe that bandwidth capacities of network links are scarce resources, and over-provisioning of bandwidth capacities is an unlikely situation. We are interested to solve the problem that congestions do occur in the network. Therefore, we make the following assumption.

Assumption 2: There exists an $j \in \mathbf{J}_i$ for flow *i*, $i = 1, \cdots, N$, such that $\sum_{k \in \mathbf{I}_j} R_k \ge C_j$.

Assumption 2 implies that there is at least one bottleneck link j on the route of flow *i*, for all $i = 1, \cdots, N$.

Let $\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ denote the Lagrangian where $\mu_j \ge 0$, $j = 1, \cdots, L$ and $\lambda_i \ge 0$, $i = 1, \cdots, N$ are the Lagrange multipliers associated with the constraint (2) and (3), 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 [23] are:

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

and

$$\lambda_i(x_i - R_i) = 0, \qquad \lambda_i \geq 0, \quad i = 1, \cdots, N,$$
$$\mu_j(\sum_{k \in \mathbf{I}_j} x_k - C_j) = 0, \qquad \mu_j \geq 0, \quad j = 1, \cdots, L,$$
$$x_i \geq 0, \qquad i = 1, \cdots, N.$$

Under the assumption 1 and assumption 2, we see that the constraint $\mathbf{x} \leq \mathbf{R}$ is inactive and hence $\lambda_i = 0$ for all $i = 1, \cdots, N$; and there exists an $j \in \mathbf{J}_i$ for flow *i*, $i = 1, \cdots, N$ such that $(\sum_{k \in \mathbf{I}_j} x_k - C_j) = 0$ and $\mu_j > 0$. Hence, we have a solution of Problem *P*:

$$\dot{x}_i = R_i - \left(\frac{w_i}{\sum_{j \in \mathbf{J}_i} \mu_j} \right)^{\frac{1}{\alpha}}, \qquad i = 1, \cdots, N, \tag{5}$$

and

$$\mu_j(\sum_{k \in \mathbf{I}_j} \dot{x}_k - C_j) = 0, \quad \mu_j \geq 0, \quad j = 1, \cdots, L, \tag{6}$$

From the network model, we have shown that when there is no congestion in the network, each flow is allocated the maximum bandwidth that is the bandwidth capacity purchased by the user. When there is at least one bottleneck link on the route of a flow, the bandwidth allocation is given from the relation (5) and (6). We will show in the next section that this bandwidth allocation is fair with respect to the bandwidth capacities purchased by users.

3. Bandwidth Capacity Fairness

In this section we present the bandwidth capacity fairness for the bandwidth allocation resulting from the solution of Problem *P*. Bandwidth α-fairness [3, 13] have been proposed to be the criterion for elastic network. In the elastic networks, there are no constraints of bandwidth capacity for each user. The bandwidth allocation is only constrained by the capacities of the network links. The bandwidth capacity fairness is used for a market where bandwidth capacities are sold to users at the edge of networks, for example, revenue optimization problems for service providers described in [4, 6, 11]. The fairness captures the essence that fair bandwidth allocation should be aggregately close to the bandwidth capacities that have been purchased by the users.

Definition *bandwidth capacity α-fair*: Let $\mathbf{w} = (w_1, \cdots, w_N)^T$ be positive numbers, α is the intensity factor and a number on the interval $(1, \infty)$. A vector of data rates $\dot{\mathbf{x}} = (\dot{x}_1, \cdots, \dot{x}_N)^T$ is bandwidth capacity α-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 $\mathbf{x}$

$$\sum_{i \in \mathbf{N}} w_i \frac{x_i - \dot{x}_i}{(R_i - \dot{x}_i)^\alpha} \leq 0. \tag{7}$$

We give the following proposition to show the relationship between the solution of Problem *P* and the definition.

Proposition 1: Under the assumption 1 and 2, the solution to Problem *P* is bandwidth capacity α-fair.

Proof: Let $\dot{\mathbf{x}}$ be the solution of Problem *P* under the assumption 1 and 2. We rewrite identity (5) to be

$$\frac{w_i}{(R_i - \dot{x}_i)^\alpha} = \sum_{j \in \mathbf{J}_i} \mu_j, \qquad i = 1, \cdots, N \tag{8}$$

Multiplying identity (8) by $(x_i - \dot{x}_i)$ and summing over *i*, we obtain

$$\begin{aligned} \sum_{i \in \mathbf{N}} w_i \frac{x_i - \dot{x}_i}{(R_i - \dot{x}_i)^\alpha} &= \sum_{i \in \mathbf{N}} \left((x_i - \dot{x}_i) \sum_{j \in \mathbf{J}_i} \mu_j \right) \\ &= (\mathbf{x} - \dot{\mathbf{x}})^T \mathbf{A} \boldsymbol{\mu} \\ &= \boldsymbol{\mu}^T \mathbf{A}^T (\mathbf{x} - \dot{\mathbf{x}}) \end{aligned} \tag{9}$$

Summing identity (6) over *j* and rearrange the terms, we have $\boldsymbol{\mu}^T \mathbf{A}^T \dot{\mathbf{x}} = \boldsymbol{\mu}^T \mathbf{C}$. Multiplying inequality (2) by $\boldsymbol{\mu}^T$, we get $\boldsymbol{\mu}^T \mathbf{A}^T \mathbf{x} \leq \boldsymbol{\mu}^T \mathbf{C}$. Combining these relations, we obtain

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

Hence, $\boldsymbol{\mu}^T \mathbf{A}^T (\mathbf{x} - \dot{\mathbf{x}}) \leq 0$, and combining this inequality with identity (9). We establish that

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

We have shown that the solution of Problem *P* is bandwidth capacity α-fair. ■

When we define the user's utility function, the weight of the utility function have not been given yet. We will determine the weight and formally define the utility function in the next section.

4. The utility function

We have shown that if there is no congestion on the route of flow *i*, the bandwidth allocated to the flow is R_i, which is the maximum bandwidth for flow *i*. When congestion does occur at link *j*, the Lagrange multiplier $\mu_j > 0$ is a positive number in the solution of Problem *P*. From microeconomic theory, we know that the Lagrange multipliers in identity (5) and (6) can be interpreted to be the marginal costs for capacity expansion at each link [23, 24]. Therefore we interpret the price p_i paid by user *i* to be the average aggregate of marginal costs for flow *i*, that is,

$$p_i = \frac{1}{m_i} \sum_{j \in \mathbf{J}_i} \mu_j , \quad \mu_j > 0 , \quad i = 1, \cdots, N , \qquad (10)$$

where $m_i = |\mathbf{J}_i|$ is the number of links that flow *i* goes through. The price paid by a user requires the network to guarantee that the user is allocated the minimum bandwidth r_i even when congestion occurs at every link that flow *i* goes through (i.e., $\mu_j > 0$, for all $j \in \mathbf{J}_i$). Hence, combining identity (5) and (10), and let $\dot{x}_i = r_i$, we have

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

Then, we obtain

$$w_i = m_i p_i (R_i - r_i)^{\alpha} , \quad i = 1, \cdots, N .$$

Hence, we formally define the utility function as

$$U_i(x_i) = \frac{m_i p_i (R_i - r_i)^{\alpha}}{(\alpha - 1)(R_i - x_i)^{\alpha - 1}} , \quad i = 1, \cdots, N \qquad (11)$$

In the next section, we will clarify that parameter α is the intensity factor, which represents the intensity of the effect that network congestion status and prices are on the bandwidth allocations.

We set the price for an individual flow to be the average aggregate of the marginal costs, instead of aggregate of marginal costs, for the flow to be allocated the required minimum bandwidth, since we believe this price is rational. To explain this, we consider a linear network depicted in Figure **1**. In the figure, the circles represent links and lines represent the routes. The linear network consists of *m* links with the same capacity *C*. The flow x_0 crosses every link, and flow x_k uses link *k* alone, for all $k = 1, \cdots, m$. All the flows request the same bandwidth capacity *R* and minimum bandwidth *r*.

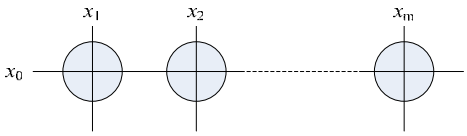

Figure 1. Linear network

The price we considered is charged by a network service provider at the edge of the network. Each flow concerns only the data rate allocated to it under the conditions of its bandwidth requirements and the price charged. Since all the flows require the same bandwidth capacity and minimum bandwidth, each flow should be charge the same price *p* for all the flows. It is also rational to expect that the bandwidth allocated to each flow should be the same.

Suppose that $2r \le C \le 2R$, so that congestion occurs at every link and a feasible bandwidth allocation exists. When we let the price be the aggregate of marginal costs for a flow to be allocated minimum bandwidth, the weight of the utility function should be set to $w = (R - r)^{\alpha} p$, for all the flows. The bandwidth allocation from the solution of Problem *P* is:

$$\begin{aligned}
&\dot{x}_0 = R - (R - r) \left(\frac{p}{m\mu} \right)^{\frac{1}{\alpha}} \\
&\dot{x}_k = R - (R - r) \left(\frac{p}{\mu} \right)^{\frac{1}{\alpha}} , \quad && k = 1, \cdots, m , \\
&\dot{x}_0 + \dot{x}_k = C , && k = 1, \cdots, m , \\
&\mu > 0 .
\end{aligned}$$

Thus, eliminating the Langrage multiplier, the bandwidth for flow 0 is

$$\dot{x}_0 = R - (2R - C) \frac{1}{1 + m^{1/\alpha}} ,$$

and bandwidth for all other flows

$$\dot{x}_k = R - (2R - C)\frac{m^{1/\alpha}}{1 + m^{1/\alpha}}, \quad k = 1, \cdots, m .$$

Apparently, the flow 0 is allocated more bandwidth than that of each other flow. It contradicts our expectation that the bandwidths should be the same, since all the flows require the same bandwidth capacity and minimum bandwidth, and the prices charged are also the same.

When we let the price be the average aggregate of marginal costs for a flow to be allocated minimum bandwidth, the weight of the utility function should be set to $w_0 = m(R - r)^\alpha p$ for flow 0, and $w_k = (R - r)^\alpha p$ for flows $k = 1, \cdots, m$. Thus, all the flows are allocated the same bandwidth, that is, the half of the link capacity, $\dot{x}_k = C/2$, for all the flows, $k = 0, 1, \cdots, m$. It is the rational bandwidth allocation we expect. Therefore, a flow's price charged at the edge of the network should be the average aggregate of marginal costs when congestions occur at every link on the route of the flow.

5. Global admission conditions

To obtain the solution of Problem *P*, in section 1 we relaxed the minimum bandwidth constraints for all the flows. When we apply the minimum bandwidth constraint (4) to the solution of Problem *P* and if the solution still satisfies the minimum bandwidth constraint (4), apparently the network is able to admit all the flows and each flow is allocated a data rate that is above its minimum bandwidth. Thus, we can obtain a global admission conditions from the solution of Problem *P*.

Combining identity (5) and (11), we have

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

where $\beta_i = \left(m_i p_i \Big/ \sum_{j \in \mathbf{J}_i} \mu_j\right)^{\frac{1}{\alpha}}$, $i = 1, \cdots, N$. Let $\rho_i = m_i p_i \Big/ \sum_{j \in \mathbf{J}_i} \mu_j$ denote the price congestion ratio. The bandwidth $\dot{x}_i$ allocated to flow i is guaranteed to be between bandwidth capacity R_i and minimum bandwidth r_i when the price congestion ratio satisfies the relation: $0 \le \rho_i \le 1$, $i = 1, \cdots, N$. Apparently, the price congestion ratios are determined by the prices that users paid and the current congestion status of the network.

Proposition 2. Global admission conditions: Flows, which are willing to pay price p_i and require bandwidth capacity R_i and minimum bandwidth r_i for all $i = 1, \cdots, N$, are able to be admitted into the network, if the global admission conditions are satisfied:

$$0 \le \rho_i \le 1, \quad i = 1, \cdots, N \tag{12}$$

where

$$\rho_i = \frac{m_i p_i}{\sum_{j \in \mathbf{J}_i} \mu_j}, \quad i = 1, \cdots, N$$
$$m_i = |\mathbf{J}_i|, \quad i = 1, \cdots, N$$

and the bandwidth allocation $\dot{\mathbf{x}}$ for all flows is bandwidth capacity α-fair, and are given by:

$$\dot{x}_i = (1 - \beta_i)R_i + \beta_i r_i, \quad i = 1, \cdots, N, \tag{13}$$

$$\mu_j\left(\sum_{k \in \mathbf{I}_j} \dot{x}_k - C_j\right) = 0, \ \mu_j \ge 0, \ j = 1, \cdots, L, \tag{14}$$

$$\beta_i = \rho_i^{\frac{1}{\alpha}}, \quad i = 1, \cdots, N . \tag{15}$$

Proof: Apparently, identity (13) and (14) are the solution of problem *P*. Hence, the bandwidth allocation is bandwidth capacity α-fair. Combining the global admission conditions (12), identity (13), (14), and (15), it shows that the bandwidth allocation meets the bandwidth requirement, $\mathbf{r} \le \dot{\mathbf{x}} \le \mathbf{R}$, and link capacity constraint, $\mathbf{A}^T \dot{\mathbf{x}} \le \mathbf{C}$, for all flows. Therefore, admitting the flows does not violate any bandwidth requirements and the flows are able to be admitted into the network. ■

Price congestion ratios in the global admission conditions reflect the congestion status of the network. After the flows are admitted, the allocated data rates are effected by β_i for all $i = 1, \cdots, N$. From identity (15), we know that $\beta_i = 0$ when the intensity factor is infinite (i.e., $\alpha \to \infty$). This implies that the congestion

status of the network has no impact on the bandwidth allocation for all flows. And each flow is allocated the maximum required bandwidth, the bandwidth capacity. The global admission conditions degenerate to be the link capacity constraints only, that is, $\sum_{k\in \mathbf{I}_j} R_k \le C_j$, $j = 1, \cdots, L$. Note that this bandwidth allocation is the same as that when there is no congestion in the network. It implies that when bandwidth requirements of flows will cause congestion in the network, the network cannot admit the flows. We have $\beta_i = \rho_i$, $i = 1, \cdots, N$ when $\alpha \to 1$, such that the congestion status has the most strong impact on the bandwidth allocation after the flows are admitted. Therefore, we show that the parameter α is the intensity factor of user's utility function.

When the congestion status has no impact on the bandwidth allocation, each link is able to make admission decision based on its local information of flows, and consequently this leads to a distributed admission algorithm. Hence, the admission conditions are distributed admission conditions when $\alpha \to \infty$. Each admitted flow is strictly allocated a data rate that is its required bandwidth capacity. The network loses the flexibility to service more flows with data rates that are lower than their bandwidth capacities. Therefore setting the intensity factor be infinite is a unlike situation for the network.

For a given a set of flows with feasible bandwidth requirements, prices of the flows are the decision variables of the global admission condition (12). It reflects the global price competitions among feasible bandwidth requirements when congestions exist in the network. When there is no congestion in the network, all flows can be admitted into network, and each flow is assigned a data rate of its bandwidth capacity. Prices do not play any role for admission decision and bandwidth allocation. The prices are charged for the network to guarantee admitted flows to be allocated data rates that are greater than or equal to their minimum bandwidths.

To explain the admission conditions and show the price competitions among feasible flows, we consider a simple network of a single link and two flows. Suppose that the link capacity is *C*, the first flow pays a price p_1 and requests bandwidth capacity R_1 and minimum bandwidth r_1, and the second flow pays a price p_2, and requests bandwidth capacity R_2 and minimum bandwidth r_2. Of course, we should also assume that $R_i \le C$, $i = 1, 2$, and $r_1 + r_2 \le C$ such that it is feasible for the two flows to compete for sharing the link. When the link is only used by one of the flows, the flows is assigned the bandwidth $\dot{x}_i = R_i$, $i = 1, 2$. When the two flows attempt to share the link, congestion occurs at the link, assuming $R_1 + R_2 \ge C$. Thus, the global admission conditions are:

$$0 \le \rho_i \le 1, \quad i = 1, 2, \tag{16}$$

$$\dot{x}_i = (1 - \rho_i^{\frac{1}{\alpha}}) R_i + \rho_i^{\frac{1}{\alpha}} r_i, \quad i = 1, 2, \tag{17}$$

$$\dot{x}_1 + \dot{x}_2 = C, \tag{18}$$

$$\rho_i = \frac{p_i}{\mu}, \quad \mu > 0, \quad i = 1, 2. \tag{19}$$

Combining equation (17), (18) and (19), we obtain

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

and

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

Bringing ρ_1 and ρ_2 into the relation (16), and rearranging the terms, we obtain that the two flows is able to share the link if the following relations are satisfied:

$$\left(\frac{p_1}{p_2}\right)^{\frac{1}{\alpha}} \ge \frac{R_1 + r_2 - C}{R_1 - r_1} \tag{20}$$

and

$$\left(\frac{p_2}{p_1}\right)^{\frac{1}{\alpha}} \ge \frac{r_1 + R_2 - C}{R_2 - r_2}. \tag{21}$$

These inequalities show that the prices are constrained by bandwidth requirements, and the price ratio determines whether the two flows are able to share the link, In other words, the network is able to admit the two flows if the price ratio satisfies the relation (20) and (21).

To use the link efficiently, it is rational for us to suppose that $R_1 = R_2 = C$. Hence, the price ratio is constrained with the relation:

$$\frac{r_2}{C - r_1} \le \left(\frac{p_1}{p_2}\right)^{\frac{1}{\alpha}} \le \frac{C - r_2}{r_1}. \qquad (22)$$

When $r_1 + r_2 < C$, the relation (22) reflects the price competition between the two flows. When the price ratio $(p_1/p_2)^{1/\alpha} = r_2/(C - r_1)$, the bandwidth allocation favors flow 1, $\dot{x}_1 = C - r_2$ and $\dot{x}_2 = r_2$. When the price ratio $(p_1/p_2)^{1/\alpha} = (C - r_2)/r_1$, flow 2 is allocated better data rate $\dot{x}_2 = C - r_1$, and $\dot{x}_1 = r_1$.

When $r_1 + r_2 = C$, the relation (22) requires that the two flows must be charged the same price for them to share the link, and each flow is allocated the minimum bandwidth, $\dot{x}_1 = r_1$ and $\dot{x}_2 = r_2$. It shows that the prices are for the network to allocate flows the guaranteed minimum bandwidths.

6. Distributed admission conditions

When networks admit flows and allocate bandwidths to admitted flows as given in global admission conditions, a global algorithm has to be used to implement the admission control and bandwidth allocation. Unfortunately, global algorithm is, if not impossible, very difficult to be implemented for large networks. All the flows need to have perfect information of the network to make decision on their prices and bandwidth requirements. 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, which is asymptotically optimal.

The network's objective is to maximize the social welfare for the network. We rewrite the total utility of the network with respect to the utility function (11):

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

where

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

and

$$\mathbf{x}^j = \{(x_{i_1}^j, x_{i_2}^j, \cdots, x_{i_l}^j) \mid i_k \in \mathbf{I}_j, 1 \le \mathrm{k} \le l = |\mathbf{I}_j|\}$$

The vector $\mathbf{x}^j$ is the flows that go through link *j*. Thus, the total utility of the network is represented as the aggregate utilities of all the links. When a link consider maximizing its social welfare independently, we form the optimization Problem P^j:

$$\max U^j(\mathbf{x}^j)$$

subject to

$$\sum_{k \in \mathbf{I}_j} x_k^j \le C_j$$

$$r_i \le x_i^j \le R_i, \quad i \in \mathbf{I}_j$$

where $j = 1, \cdots, L$. In the same way we resolve Problem *P*, we first relax the lower bound constraint of flows to be just positive. 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 [23] are:

$$\frac{p_i (R_i - r_i)^{\alpha}}{(R_i - x_i)^{\alpha}} - \mu_j - \lambda_i = 0, \quad i \in \mathbf{I}_j,$$

and

$$\lambda_i^j (x_i^j - R_i) = 0\,, \quad \lambda_i^j \ge 0\,, \quad i \in \mathbf{I}_j\,,$$
$$\mu^j (\sum_{k \in \mathbf{I}_j} x_k^j - C_j) = 0\,, \qquad \mu^j \ge 0\,.$$

The bandwidth allocated to flow *i* at link *j* is given by

$$\hat{x}_i^j = \begin{cases} R_i\,, & \sum_{k \in \mathbf{I}_j} \hat{x}_k^j < C_j \\ R_i - (R_i - r_i)\left(\dfrac{p_i}{\mu^j}\right)^{\frac{1}{\alpha}}\,, \ \mu^j > 0, & \sum_{k \in \mathbf{I}_j} \hat{x}_k^j = C_j \end{cases} \tag{23}$$

where $i \in \mathbf{I}_j$ and $j = 1, \cdots, L$. Suppose that the bandwidth $\bar{x}_i$ allocated by the network to flow *i* is the minimum bandwidth allocated among the links that the flow goes through. We obtain the bandwidth $\bar{x}_i$ of flow *i*:

$$\bar{x}_i = \min\{\hat{x}_i^j \mid j = \mathbf{J}_i\}\,, \quad i = 1, \cdots, N\,, \tag{24}$$

and $\hat{x}_i^j$ is the solution from identity (23). The network utility for the bandwidth allocation $\bar{\mathbf{x}}$ is less than or equal to that of the optimal bandwidth allocation $\dot{\mathbf{x}}$:

$$\overline{U}(\bar{\mathbf{x}}) = \sum\nolimits_{j \in \mathbf{L}} U^j(\mathbf{x}^j) \le U(\dot{\mathbf{x}}) \tag{25}$$

where $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_N)^T$, and $\dot{\mathbf{x}}$ is the optimal bandwidth allocation of Problem *P*. Obviously, when a flow or more flows in the network are allocated bandwidths that are the flows' bandwidth capacities (i.e., $x_i = R_i$, for some $i \in \mathbf{N}$), the relation (25) reaches the equality, $\overline{U}(\bar{\mathbf{x}}) = U(\dot{\mathbf{x}}) = \infty$. When the network is fully loaded with all the flows being allocated the minimum bandwidths, $\dot{\mathbf{x}} = \bar{\mathbf{x}} = \mathbf{r}$, the relation (25) reaches the equality again. Thus, the bandwidth allocation $\bar{\mathbf{x}}$ is asymptotical optimal with respect to the optimal solution of Problem *P*.

The advantage of the bandwidth allocation $\bar{\mathbf{x}}$ is that a distributed algorithm can be developed for the bandwidth allocation, which also leads to a distributed admission control. Note that the algorithm given in (23) needs only the local information of link *j*. Thus, each link is able to allocate bandwidth to the flows that go through the link with the local information of the link. The network allocates flow *i* the bandwidth $\bar{x}_i$ that is given in (24), which is able to be determined by a round trip probe along the route of the flow.

Proposition 3. Distributed admission conditions: The flows at link *j* are willing to pay price p_i , and require 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 not a congestion link. 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 $\sum_{k \in \mathbf{I}_j} R_k \ge C_j$, link *j* is a congestion link. Link *j* is able to admit the flows and the bandwidth allocation satisfies bandwidth requirements of the flows if the following conditions are satisfied:

$$\sum_{k \in \mathbf{I}_j} r_k \le C_j\,, \tag{26}$$

and

$$\sum_{k \in \mathbf{I}_j} (R_k - r_k) p_k^{\frac{1}{\alpha}} - (\sum_{k \in \mathbf{I}_j} R_k - C_j) p_i^{\frac{1}{\alpha}} \ge 0\,,\ i \in \mathbf{I}_j\,. \tag{27}$$

And link *j* is able to allocate flow *i* bandwidth

$$\hat{x}_i^j = R_i - \frac{\sum_{k \in \mathbf{I}_j} R_k - C_j}{\sum_{k \in \mathbf{I}_j} (R_k - r_k) p_k^{\frac{1}{\alpha}}} (R_i - r_i) p_i^{\frac{1}{\alpha}}\,,\ i \in \mathbf{I}_j\,. \tag{28}$$

Proof: When link *j* is not a congestion link, it is obvious as it stands in (a). We only need to show that when condition (26) and (27) are satisfied, the bandwidth allocation meets the bandwidth requirements of the flows. The bandwidth allocation has a feasible solution to Problem P^j when condition (26) is satisfied. Suppose condition (26) is satisfied, eliminating μ^j from identity (23), we obtain a bandwidth allocation given in (28). Applying the minimum bandwidth constraints to bandwidth allocation (28) to let $\hat{x}_i^j \ge r_i$, for all $i \in \mathbf{I}_j$, we obtain the condition (27). Hence, when condition (27) is satisfied, the bandwidth allocation meets the bandwidth requirements of flows, i.e., $r_i \le \hat{x}_i^j \le R_i$, for all $i \in \mathbf{I}_j$. Therefore, the link is able to admit the flows and the bandwidth allocation is given in (28). ■

In the admission condition (27), the prices are the decision variables for given bandwidth requirements of flows. This reflects the price competition among flows at each link. While each link makes admission decision independently with its local information, a flow is admitted into the network when the flow is admitted by all the links that the flow goes through, and the bandwidth allocated to the flow by the network is given in (24).

The distributed admission control is particularly useful for Diffserv since each link make bandwidth allocation decisions locally. Flows in DiffServ are marked into a small number of service classes. Each service class has the same bandwidth requirements. We consider a network that accommodates two service classes with different bandwidth requirements. Each class requires bandwidth capacity R_1 , R_2 and minimum bandwidth r_1 , r_2 , respectively. Flows are charged p_1 for flows in class 1, and p_2 for flows in class 2. Suppose that link j is a congestion link and admission condition (26) is satisfied at the link. Thus, admission condition (27) becomes:

$$r_1 n_1 + (R_2 - (R_2 - r_2)\left(\frac{p_2}{p_1}\right)^{\frac{1}{\alpha}})n_2 \le C_j , \qquad (29)$$

$$(R_1 - (R_1 - r_1)\left(\frac{p_1}{p_2}\right)^{\frac{1}{\alpha}})n_1 + r_2 n_2 \le C_j , \qquad (30)$$

$$n_1 \ge 0 , \quad n_2 \ge 0 . \qquad (31)$$

where n_1 and n_2 are the numbers of flows in class 1 and class 2, respectively. Admission condition (29), (30), and (31) defines a feasible set of (n_1, n_2) depicted in **Figure 2**. The shaded areas are sets of (n_1, n_2) that are numbers of flows in class 1 and class 2, respectively. When the flow numbers (n_1, n_2) is in the shaded areas, link j is able to admit the flows and allocate bandwidth given by

$$\hat{x}_i^j = R_i - \frac{n_1 R_1 + n_2 R_2 - C_j}{n_1 (R_1 - r_1) p_1^{\frac{1}{\alpha}} + n_2 (R_2 - r_2) p_2^{\frac{1}{\alpha}}} (R_i - r_i) p_i^{\frac{1}{\alpha}} ,$$

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

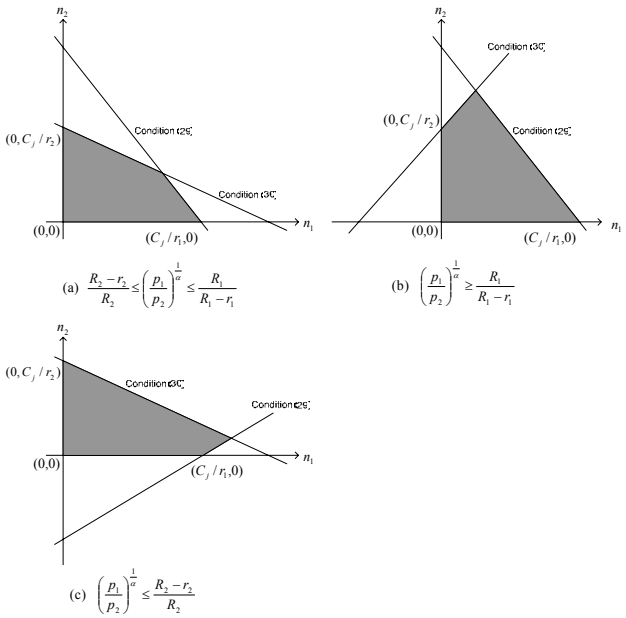

Figure 2. The feasible set of flow numbers with different price ratios.

In **Figure 2**, it also shows that price ratio can affect the feasible set of flow numbers in each class. The shaded area in picture (a) is the feasible set of flow numbers when the price ratio satisfies the relation,

$$\frac{R_2 - r_2}{R_2} \le \left(\frac{p_1}{p_2}\right)^{\frac{1}{\alpha}} \le \frac{R_1}{R_1 - r_1} ;$$

picture (b) gives the feasible set of flow numbers when the price ratio satisfies the relation,

$$\left(\frac{p_1}{p_2}\right)^{\frac{1}{\alpha}} \ge \frac{R_1}{R_1 - r_1} ;$$

picture (c) gives the feasible set of flow numbers when the price ratio satisfies the relation,

$$\left(\frac{p_1}{p_2}\right)^{\frac{1}{\alpha}} \le \frac{R_2 - r_2}{R_2} .$$

Therefore, given the prices and bandwidth requirements of each class, network links can determine the numbers of flows that the link can admit. The admitted flows are guaranteed to be allocated data rates that are greater than or equal to their minimum bandwidth requirements.

7. Conclusion

In this paper we have presented a theoretic framework for admission control and bandwidth allocation at network links to meet network users' bandwidth requirements with a price charged by a network service provider at the edge of the network. A utility function was defined to capture the bandwidth demands of network users when users are charged prices for finite network capacities at the edge of the network. An optimization framework leads to fair bandwidth allocation and global admission condition, in which the prices are the decision variables for given sets of feasible bandwidth requirements. We have also provided a distributed admission condition and bandwidth allocation, which is asymptotically optimal with respect to the global optimal bandwidth allocation. 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 the bounds of the set is determined by bandwidth requirements and the price ratios of the service classes.

Reference

1. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. RFC 2475: An Architecture for Differentiated Services, December 1998.

2. Neil J. Keon and G. Anandalingam, Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees, IEEE/ACM Transactions on Networking, Vol. 11, No. 1, February 2003.

3. Jeonghoon Mo and Jean Walrand, Fair End-to-End Window-Based Congestion Control, IEEE/ACM Transactions On Networking, Vol. 8, No. 5, October 2000.

4. De Veciana, G., Lee, T.-J. And Konstantopoulos, T., Stability and performance analysis of networks supporting elastic services. IEEE/ACM Transactions on Networking, Volume 9, Issue 1, pp2–14, 2001.

5. C. Paschalidis and J. N. Tsitsiklis, "Congestion-Dependent Pricing of Network Services," IEEE/ACM Transactions on Networking, vol. 8, no. 2, pp. 171–184, 2000.

6. C. Paschalidis and Y. Liu, "Pricing in Multiservice Loss Networks: Static Pricing, Asymptotic Optimality, and Demand Substitution Effects," IEEE/ACM Transactions on Networking, vol. 10, no. 3, pp. 425–438, 2002.

7. Uday Savagaonkar, Edwin K.P. Chong, Robert L. Givan, Online Pricing For Bandwidth Provisioning In Multi-Class Networks, Computer Networks Journal, Vol. 44, pp. 835-853, 2004.

8. 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.

9. F.P. Kelly and R. J. Williams, Fluid model for a network operating under a fair bandwidth-sharing policy, Annals of Applied Probability, Vol.14, pp1055-1083, 2004.

10. Nemo Semret, Raymond R.-F. Liao, Andrew T. Campbell, and Aurel A. Lazar, Pricing, Provisioning and Peering: Dynamic Markets for Differentiated Internet Services and Implications for Network Interconnections, IEEE Journal on Selected Areas in Communications, vol. 18, no. 12, December 2000.

11. Laurent Massoulié and James Roberts, Bandwidth Sharing: Objectives and Algorithms, IEEE/ACM Transactions on Networking, Vol. 10, No. 3, June 2002.

12. D. Bertsekas and R. Gallager, Data Networks. Englewood Cliffs, NJ: Prentice Hall, 1987.

13. Thomas Bonald and Laurent Massoulié, Impact of Fairness on Internet Performance, Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, Vol. 29, Iss. 1, 2001.

14. Tansu Alpcan and Tamer Başşar, A Utility-Based Congestion Control Scheme for Internet-Style Networks with Delay, Proceedings of IEEE INFOCOM 2003, San Franciso, CA, 2003.

15. Roberts, J. And Massoulié, L., Bandwidth sharing and admission control for elastic traffic. Telecommunication Systems, Vol. 15, pp185–201, 2000.

16. David D. Clark, A Model for Cost Allocation and Pricing in the Internet, Journal of Electronic Publishing, Volume 2, Issue 1, May, 1996

17. Tian Bu, Don Towsley, Fixed point approximation for TCP behavior in an AQM network, Proc. ACM Sigmetrics, Volume 29, Issue 1, 2001.

18. R. Braden, D. Clark, and S. Shenker. RFC 1633: Integrated Services in the Internet Architecture: an Overview, June 1994.

19. Haïkel Yaïche, Ravi R. Mazumdar, and Catherine Rosenberg, A Game Theoretic Framework for Bandwidth Allocation and Pricing in Broadband Networks, IEEE/ACM Transactions on Networking, Vol. 8, No. 5, October 2000.

20. Muthoo, Bargaining Theory with Applications. Cambridge, U.K.: Cambridge Univ. Press, 1999.

21. J. Nash, The bargaining problem, Econometrica, vol. 18, pp. 155–162, 1950.

22. Yong Liu, Francesco L. Presti, Vishal Misra, Donald F. Towsley, Yu Gu, Scalable fluid models and simulations for large-scale IP networks, Volume 14 Issue 3, 2004.

23. Michael Hoy, John Livernois, Chris McKenna, Ray Rees, Thanasis Stengos, Mathematics for Economics, 2nd Edition, MIT Press, 2001.

24. H.R. Varian, Microeconomic Analysis, Third edition, Norton: New York, 1992.