Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

TOBIAS HARKS[1]

Utility Proportional Fair Bandwidth Allocation - An Optimization Oriented Approach

[1]This work has been supported by the German research funding agency 'Deutsche Forschungsgemeinschaft' under the graduate program 'Graduiertenkolleg 621 (MAGSI/Berlin)'

Utility Proportional Fair Bandwidth Allocation: An Optimization Oriented Approach

Tobias Harks[†]
Konrad Zuse Zentrum für Informationstechnik Berlin (ZIB)
harks@zib.de

Berlin, August 2004

Abstract

In this paper, we present a novel approach to the congestion control and resource allocation problem of elastic and real-time traffic in telecommunication networks. With the concept of utility functions, where each source uses a utility function to evaluate the benefit from achieving a transmission rate, we interpret the resource allocation problem as a global optimization problem. The solution to this problem is characterized by a new fairness criterion, *utility proportional fairness*. We argue that it is an application level performance measure, i.e. the utility that should be shared fairly among users. As a result of our analysis, we obtain congestion control laws at links and sources that are globally stable and provide a utility proportional fair resource allocation in equilibrium. We show that a utility proportional fair resource allocation also ensures utility max-min fairness for all users sharing a single path in the network. As a special case of our framework, we incorporate utility max-min fairness for the entire network. To implement our approach, neither per-flow state at the routers nor explicit feedback beside ECN (Explicit Congestion Notification) from the routers to the end-systems is required.

Keywords: Utility proportional fairness, resource allocation, congestion control, optimization, real-time applications

1 Introduction

In this paper, we present a network architecture that considers an application-layer performance measure, called *utility*, in the context of bandwidth allocation schemes. In the last years, there have been several papers [1,5-10] that interpreted congestion control of communication networks as a distributed algorithm at sources and links in order to solve a global optimization problem. Even though considerable progress has been made in this direction, the existing work focusses on elastic traffic, such as file transfer (FTP, http) or electronic mail (SMTP). In [2], elastic applications are characterized by their ability to adapt the sending rates in presence of congestion and to tolerate packet delays and losses rather gracefully. From a user perspective, common to all elastic applications is the request to transfer data in a short time. To model these

[†]This work has been supported by the German research funding agency 'Deutsche Forschungsgemeinschaft' under the graduate program 'Graduiertenkolleg 621 (MAGSI/Berlin)'

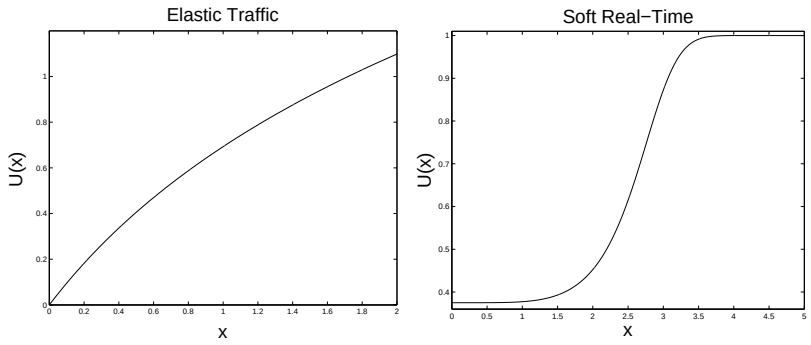

Figure 1: Utilities for elastic traffic and adaptive real-time traffic

characteristics, we resort to the concept of utility functions. Following [2] and [5], traffic that leads to an increasing, strictly concave (decreasing marginal improvement) utility function is called *elastic* traffic. We call such a utility function *bandwidth utility* since the utility function evaluates the benefit from achieving a certain transmission rate. The proposed source and link algorithms are designed to maximize the aggregate bandwidth utility (sum over all bandwidth utilities) subject to capacity constraints at the links. Kelly introduced in [5] the so called *bandwidth proportional fair* allocation, where bandwidth utilities are logarithmic. The algorithms at the links are based on Lagrange multiplier methods coming from optimization theory, so the concavity assumption seems to be essential. As shown in [2], some applications, especially real-time applications have nonconcave bandwidth utility functions. A voice-over-IP flow, for instance, receives no bandwidth utility, if the rate is below the minimum encoding rate. Its bandwidth utility is at maximum, if the rate is above its maximum encoding rate. Hence, its bandwidth utility can be approximated by a step function. According to Shenker [2], the bandwidth utility of adaptive real-time applications can be modeled as an S-shaped utility function (a convex part at low rates followed by a concave part at higher rates) as shown in Figure 1. The paradigm of the work dealing with bandwidth utility functions of elastic applications in the context of congestion control is to maximize the bandwidth utilization of the network (bandwidth system optimum) under specific bandwidth fairness aspects (bandwidth max-min, bandwidth proportional fair).
The central part of this work is to turn the focus on fairness of user-received utility of different applications including nonelastic applications with nonconcave bandwidth utility functions. A user running an application does not care about any fair bandwidth shares, as long as his application performs satisfactory. Hence, we argue that it is an application performance measure, i.e. the utility that should be shared fairly among users. To motivate this new paradigm, we refer to the concept of *utility max-min fairness* introduced by Cao and Zegura in [11]. Let us consider a network consisting of a single link of capacity one shared by two users. One user transfers data according to an elastic application with strictly increasing and concave bandwidth utility $U_1(\cdot)$. The other user transfers real-time video data with a nonconcave bandwidth utility function $U_2(\cdot)$. Figure 2 shows, how different bandwidth allocations affect the received utility.

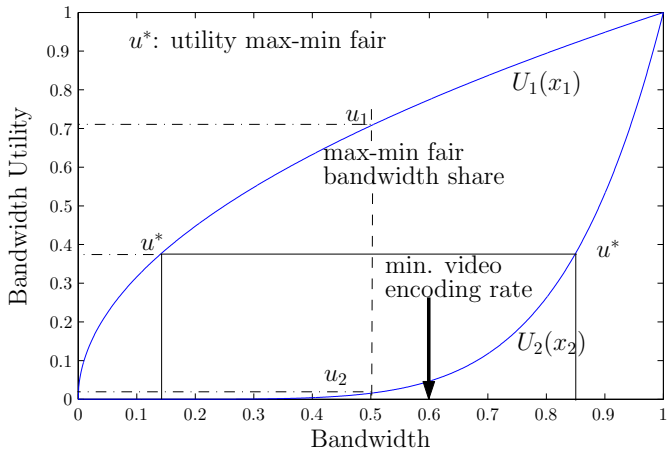

Figure 2: Utility max-min and bandwidth max-min fairness

If the bandwidth is shared equally, what is referred to as *max-min bandwidth* allocation in this example, user 1 receives a much larger utility than user 2. Conversely, user 2 would not be satisfied since he does not receive the minimum video encoding bandwidth. If we want to share utility equally, instead of bandwidth, we would like to have a resource allocation, where the received utilities are equal or *utility max-min fair*, i.e. $U_1(x_1) = U_2(x_2) = u^*$.
In [11], Cao and Zegura present a link algorithm that achieves a utility max-min fair bandwidth allocation, where each for each link the utility functions of all flows sharing that link is maintained. In [12], Cho and Song present a utility max-min architecture, where each link communicates a supported utility value to sources using that link. Then sources adapt their sending rates according to the minimum of these utility values.
In this paper, we extend the utility max-min architecture and propose a new fairness criterion, *utility proportional fairness*, which includes the utility max-min fair resource allocation as a special case. A utility proportional fair bandwidth allocation is characterized by the solution of an associated optimization problem. The benefit a user s gains when achieving a bandwidth utility value $U_s(x_s)$ is evaluated by a new *second order utility* $F_s(x_s)$ and the objective is to maximize aggregate second order utility subject to capacity constraints. The second order utilities are assumed to be strictly concave, whereas the bandwidth utilities can be chosen arbitrarily. We only assume that the bandwidth utilities are monotonic increasing in a given interval. This is a natural assumption since any application will profit from receiving more bandwidth in a certain bandwidth interval. We emphasize, that our distributed algorithm does not need any per-flow information at the links. The feedback from links to sources does not include overhead, such as explicit utility values as done in [12]. It merely relies on the communication of Lagrange multipliers, called shadow prices, from the links to the sources. This can be achieved by an Active Queue Management (AQM) scheme, such as Random Early Marking (REM) [9] using Explicit Congestion Notification (ECN) [4].
The rest of the paper is organized as follows. In the next section, we describe

our model, the second order utility optimization problem and its dual based on ideas of [1, 5, 8]. Given a specific bandwidth utility, we describe a constructive method to find the second order utility function $F_s(\cdot)$. In Section 3, we present a static primal algorithm at the sources and a dynamic dual algorithm at the links solving the global optimization problem and its dual. We further present a global stability result for the dual algorithm based on Lyapunov functions along the lines of [13]. In Section 4, we define a new fairness criterion, *utility proportional fairness*, and show that our algorithms achieve utility max-min fairness in equilibrium for users sharing a single path in the network. We further incorporate utility max-min fairness for the entire network as a special case of our framework. Finally, we conclude in Section 5 with remarks on open issues.

2 Analytical Model

Considerable progress has recently been made in bringing analytical models into congestion control and resource allocation problems [1, 5, 6, 7, 8]. Key to these works has been to explicitly model the *congestion measure* that is communicated implicitly or explicitly back to the sources by the routers. It is assumed that each link maintains a variable, called *price*, and the sources have information about the aggregate price of links in their path. These assumptions are implicitly present in current TCP protocols. TCP Reno uses loss as an indication of congestion, whereas Vegas uses queuing delay to measure the state of congestion in the network. The equilibrium structure of TCP protocols can be interpreted as the solution of a global optimization problem, where the objective is to maximize the aggregate bandwidth utility over transmission rates. Since TCP is designed for elastic traffic, the underlying bandwidth utility functions are strictly concave [2].
In this section, we describe a fluid-flow model, similar to that in [1, 5, 8]. We interpret an equilibrium point as the unique solution of an associated optimization problem. The resulting resource allocation is aimed to provide a fair share of an application layer performance measure, i.e. the utility to users. In contrast to [1,5-8], we do not pose any restrictions on the bandwidth utility functions, except for monotonicity.

2.1 Model

We model a packet switched network by a set of nodes (router) connected by a set L of unidirectional links (output ports) with finite capacities $c = (c_l, l \in L)$. The set of links are shared by a set S of sources indexed by s. A source s represents an end-to-end connection and its route involves a subset $L(s) \subset L$ of links. Equivalently, each link is used by a subset $S(l) \subset S$ of sources. The sets $L(s)$ or $S(l)$ define a routing matrix

$$R_{ls} = \begin{cases} 1 & \text{if } l \in L(s), \\ 0 & \text{else.} \end{cases}$$

A transmission rate x_s in *packets per second* is associated with each source s. We assume, that the rates $x_s, s \in S$ lie in the interval $X_s = [0, x_s^{max}]$, where

x_s^{max} is the maximum sending rate of source s. This upper bound may differ substantially for different applications. A subset of sources $S_r \subset S$ transferring real-time data, for instance, may have a maximum encoding rate $x_s^{max}, s \in S_r$, which can be much lower than the upper bound $x_s^{max}, s \in S \setminus S_r$ of elastic applications, which are greedy for any available bandwidth in the network. Thus, sending rates of elastic applications are constrained by bottleneck links in the network.

Definition 2.1 *A rate vector $x = (x_s, s \in S)$ is said to be* feasible *if it satisfies the conditions:*

$$x_s \in X_s \ \forall s \in S \text{ and } Rx \leq c.$$

With each link l, a scalar positive congestion-measure p_l, called *price*, is associated. In current TCP implementations, the congestion measure or price is based upon information about either loss (Drop Tail), queue length (RED), or marks (REM), which requires an ECN bit [7]. Let

$$y_l = \sum_{s \in S} R_{ls} x_s$$

be the aggregate transmission rate of link l, i.e. the sum over all rates using that link, and let

$$q_s = \sum_{l \in L} R_{ls} p_l$$

be the end-to-end congestion measure of source s. Note that taking the sum of congestion measures of a used path is essential to maintain the interpretation of p_l as dual variables [1]. Source s can observe its own rate x_s and the end-to-end congestion measure q_s of its path. Link l can observe its local congestion measure p_l and the aggregate transmission rate y_l. When the transmission rate of user s is x_s, user s receives a benefit measured by the bandwidth utility $U_s(x_s)$, which is a scalar function and has the following form:

$$\begin{aligned} U_s &: X_s \to Y_s \\ x_s &\mapsto U_s(x_s), \end{aligned}$$

where $Y_s = [U_s(0), U_s(x_s^{max})] = [u_s^{min}, u_s^{max}]$, $U_s(0) = u_s^{min}, U_s(x_s^{max}) = u_s^{max}$.

Assumption 2.2 *The bandwidth utility functions $U_s(\cdot)$ are continuous, differentiable, and strictly increasing, i.e. $U_s'(x_s) > 0$ for all $x_s \in X_s$, $s \in S$.*

This assumption ensures the existence of the inverse function $U_s^{-1}(\cdot)$ over the range $[u_s^{min}, u_s^{max}]$. It should be noted, that we do not assume the bandwidth utilities to be strictly concave as done in [1, 7, 8, 9]. Thus the shape of feasible bandwidth utilities can be arbitrarily chosen, except for the nonnegativity assumption on the first derivative. Before we present a constructive method to generate second order utility functions, we briefly restate the overall paradigm. An optimal operation point or equilibrium should result in almost equal utility values for different applications. The exact definition of the proposed resource allocation, i.e. *utility proportional fair* resource allocation, will be given below. If we want to follow this philosophy, we must translate a given congestion level of a path, represented by q_s, into an appropriate utility value the network can offer to source s. We model this utility value, called *available utility*, as the transformation of the congestion measure q_s by a function $f_s(q_s)$, called *transformation function*. This function is assumed to be strictly decreasing.

Assumption 2.3 *The transformation function $f_s(\cdot)$ describing the available utility of a path used by sender s is assumed to be a continuous, differentiable, and strictly decreasing function of the aggregate congestion measure q_s, i.e. $f_s'(q_s) < 0$ for all $q_s \geq 0$ and $s \in S$.*

This assumption is reasonable, since the more congested a path is, the smaller will be the available utility of an application. The main idea is, that each user s should send at data rates x_s in order to match its own bandwidth utility with the available utility of its path. This leads to the following equation:

$$U_s(x_s) = [f_s(q_s)]_{u_s^{min}}^{u_s^{max}}, s \in S, \tag{1}$$

where $[w]_a^b := min\{max\{w, a\}, b\} = \begin{cases} w, & \text{if } a \leq w \leq b \\ a, & \text{if } w < a \\ b, & \text{if } w > b. \end{cases}$

Note that the utility a source can receive is bounded by the minimum and maximum utility values u_s^{min} and u_s^{max}. Hence, the source rates x_s are adjusted according to the available utility $f_s(q_s)$ of their used path as follows:

$$x_s = U_s^{-1}([f_s(q_s)]_{u_s^{min}}^{u_s^{max}}), \quad s \in S. \tag{2}$$

A source $s \in S$ reacts to the congestion measure q_s in the following manner: if the congestion measure q_s is below a threshold $q_s < q_s^{min} := f^{-1}(u_s^{max})$, then the source transmits data at maximum rate $x_s^{max} := U_s^{-1}(u_s^{max})$; if q_s is above a threshold $q_s > q_s^{max} := f^{-1}(u_s^{min})$, the source sends at minimum rate $x_s^{min} := U_s^{-1}(u_s^{min})$; if q_s is in between these two thresholds $q_s \in Q_s := [q_s^{min}, q_s^{max}]$, the sending rate is adapted according to $x_s = U_s^{-1}(f_s(q_s))$.

Lemma 2.4 *The function $G_s(q_s) = U_s^{-1}([f_s(q_s)]_{u_s^{min}}^{u_s^{max}})$ is positive, differentiable, and strictly monotone decreasing, i.e. $G_s^{'}(q_s) < 0$ on the range $q_s \in Q_s$, and its inverse $G_s^{-1}(\cdot)$ is well defined on X_s.*

Proof: Since $U_s(\cdot)$ is defined on X_s, $U_s^{-1}(\cdot)$ is always nonnegative. Since $f_s(\cdot)$ is differentiable over Q_s, and $U_s^{-1}(\cdot)$ is differentiable over Y_s, the composition $G_s(q_s) = U_s^{-1}(f_s(q_s))$ is differentiable over Q_s. We compute the derivative using the chain rule: $G_s^{'}(q_s) = U_s^{-1'}(f_s(q_s))f_s^{'}(q_s)$. The derivative of the inverse $U_s^{-1}(f_s(q_s))$ can be computed as

$$U_s^{-1'}(f_s(q_s)) = \frac{1}{U_s'(U_s^{-1}(f_s(q_s)))} > 0.$$

With the inequality $f_s^{'}(\cdot) < 0$, we get $G_s^{'}(q_s) < 0$, $q_s \in Q_s$. Hence, $G_s(q_s)$ is strictly monotone decreasing in Q_s, so its inverse $G_s^{-1}(x_s)$ exists on X_s. □

2.2 Equilibrium Structure and Second Order Utility Optimization

In this section we study the above model at equilibrium, i.e. we assume, that rates and prices are at fixed equilibrium values x^*, y^*, p^*, q^*. From the above

model, we immediately have the relationships:

$$y^* = Rx^*$$
$$q^* = Rp^*.$$

In equilibrium, the sending rates $x_s \in X_s,\ s \in S$ satisfy:

$$x_s^* = U_s^{-1}([f_s(q_s^*)]_{u_s^{min}}^{u_s^{max}}) = G_s(q_s^*). \tag{3}$$

Since q_s represents the congestion in the path $L(s)$, the sending rate will be decreasing at higher q_s, and increasing at lower q_s. Now we consider the inverse $G_s^{-1}(x_s)$ of the above function on the interval X_s, and construct the second order utility $F_s(x_s)$ as the integral of $G_s^{-1}(x_s)$. Hence $F_s(\cdot)$ has the following form and property:

$$\begin{aligned} F_s(x_s) &= \int G_s^{-1}(x_s)dx_s \\ F_s^{'}(x_s) &= G_s^{-1}(x_s). \end{aligned} \tag{4}$$

Lemma 2.5 *The second order utility $F_s(\cdot)$ is a positive, continuous, strictly increasing, and strictly concave function of $x_s \in X_s$.*

Proof: This follows directly from Lemma 2.4 and the relation

$$F_s^{''}(x_s) = G_s^{-1'}(x_s) = \frac{1}{G_s'(q_s)} < 0. \qquad \square$$

The construction of $F_s(\cdot)$ leads to the following property:

Lemma 2.6 *The equilibrium rate* (3) *is the unique solution of the optimization problem:*

$$\max_{x_s \geq 0} F_s(x_s) - q_s x_s. \tag{5}$$

Proof: The first order necessary optimality condition to problem (5) is:

$$\begin{aligned} F_s^{'}(x_s) &= q_s \\ \Leftrightarrow G_s^{-1}(x_s) &= q_s \\ \Leftrightarrow \qquad x_s &= U_s^{-1}([f_s(q_s)]_{u_s^{min}}^{u_s^{max}}) \end{aligned}$$

Due to the strict concavity of $F(\cdot)$ on X_s, the second order sufficient condition is also satisfied completing the proof. $\square$

The above optimization problem can be interpreted as follows. $F_s(x_s)$ is the second order utility a source receives, when sending at rate x_s, and $q_s x_s$ is the price per unit flow the network would charge. The solution to (5) is the maximization of individual utility profit at fixed cost q_s.
Now we turn to the overall system utility optimization problem. The aggregate prices q_s ensure that individual optimality does not collide with social optimality. An appropriate choice of prices $p_l, l \in L$ must guarantee that the solutions of (5) also solve the system utility optimization problem:

$$\max_{x \geq 0} \sum_{s \in S} F_s(x_s) \tag{6}$$

$$\text{subject to} \quad Rx \leq c. \tag{7}$$

This problem is a convex program, similar to the convex programs in [1, 8, 10], for which a unique optimal rate vector exist. For solving this problem directly global knowledge about actions of all sources is required, since the rates are coupled through the shared links. This problem can be solved by considering its dual [10].

3 Dual Problem and Global Stability

In accordance with the approach in [1], we introduce the Lagrangian and consider prices $p_l, l \in L$ as Lagrange multipliers for (6),(7). Let

$$L(x,p) = \sum_{s\in S} F_s(x_s) - \sum_{l\in L} p_l(y_l - c_l) = \sum_{s\in S} F_s(x_s) - q_s x_s + \sum_{l\in L} p_l c_l$$

be the Lagrangian of (6) and (7). The dual problem can be formulated as:

$$\min_{p_l\geq 0} \sum_{s\in S} V_s(q_s) + \sum_{l\in L} p_l c_l, \tag{8}$$

where

$$V_s(x_s) = \max_{x_s\geq 0} F_s(x_s) - q_s x_s, \quad x_s \in X_s. \tag{9}$$

Due to the strict concavity of the objective and the linear constraints, at optimal prices p^*, the corresponding optimal x^* solving (9) is exactly the unique solution of the primal problem (6),(7). Note that (5) has the same structure as (9), so we only need to assure that the prices q_s given in (5) correspond to Lagrange multipliers q_s given in (9).
As shown in [10], a straightforward method to guarantee that equilibrium prices are Lagrange multipliers is the gradient projection method applied to the dual problem (8):

$$\frac{d}{dt}p_l(t) = \begin{cases} \gamma_l(p_l(t))(y_l(t) - c_l) & \text{if } p_l(t) > 0 \\ \gamma_l(p_l(t))[y_l(t) - c_l]^+ & \text{if } p_l(t) = 0, \end{cases} \tag{10}$$

where $[z] = max\{0, z\}$ and $\gamma_l(p_l) > 0$ is a nondecreasing continuous function. A discrete time version of (10) is:

$$p_l(t+1) = \begin{cases} p_l(t) + \gamma_l(p_l(t))(y_l(t) - c_l) & \text{if } p_l(t) > 0 \\ p_l(t) + \gamma_l(p_l(t))[y_l(t) - c_l]^+ & \text{if } p_l(t) = 0. \end{cases}$$

This algorithm can be implemented in a distributed environment. The information needed at the links is the link bandwidth c_l and the aggregate transmission rate $y_l(t)$, both of which are available. In equilibrium, the prices satisfy the complementary slackness condition, i.e. $p_l(t)$ are zero for non-saturated links and non-zero for bottleneck links. In this section, we state the global convergence of the dual algorithm (8) combined with the static source law (5) using Lyapunov techniques along the lines of [13]. We only assume that the routing matrix R is nonsingular. This guarantees that for any given $q_s \in S$ there exists a unique vector $(p_l, l \in L_s)$ such that $q_s = \sum_{l\in L_s} p_l$.

Theorem 3.1 *Assume the routing matrix R is nonsingular. Then the dual algorithm* (10) *starting from any initial state converges asymptotically to the unique solution of* (6) *and* (7)*.*

The proof of this theorem can be found in Appendix A. For further analysis of the speed of convergence, we refer to [1].

4 Utility Proportional Fairness

Kelly et al. [8] introduced the concept of *proportional fairness*. They consider elastic flows with corresponding strictly concave logarithmic bandwidth utility functions. A proportional fair rate vector $(x_s, s \in S)$ is defined such that for any other feasible rate vector $(y_s, s \in S)$ the aggregate of proportional change is nonpositive:

$$\sum_{s \in S} \frac{y_s - x_s}{x_s} \leq 0.$$

This definition is motivated by the assumption that all users have the same logarithmic bandwidth utility function $U_s(x_s) = \log(x_s)$. By this assumption, a first order necessary and sufficient optimality condition for the system bandwidth optimization problem

$$\begin{aligned} \max_{x_s \geq 0} & \sum_{x_s \geq 0} U_s(x_s) \\ \text{subject to} \quad & Rx \leq 0 \end{aligned}$$

is

$$\sum_{s \in S} \frac{\partial U_s}{\partial x_s}(x_s)(y_s - x_s) = \sum_{s \in S} \frac{y_s - x_s}{x_s} \leq 0.$$

This condition is known as the *variational inequality* and it corresponds to the definition of proportional fairness.
Before we come to our new fairness definition, we restate the concept of utility max-min fairness. It is simply the translation of the well known bandwidth max-min fairness applied to utility values.

Definition 4.1 *A set of rates $(x_s, s \in S)$ is said to be* utility max-min fair*, if it is feasible, and for any other feasible set of rates $(y_s, s \in S)$, the following condition hold: if $U_s(y_s) > U_s(x_s)$ for some $s \in S$, then there exists $k \in S$ such that $U_k(y_k) < U_k(x_k)$ and $U_k(x_k) \leq U_s(x_s)$.*

Suppose we have a utility max-min fair rate allocation. Then, a user cannot increase its utility, without decreasing the utility of another user, which receives already a smaller utility. We further apply the above definition to a utility allocation of a single path.

Definition 4.2 *Consider a single path in the network denoted by a set of adjacent links $(l \in L_p)$. Assume a set of users $S_{L_p} \subset S$ share this path, i.e. $L(s) = L_p$ for $s \in S_{L_p}$. Then, the set of rates $x_s, s \in S$ is said to be* path utility max-min fair *if the rate allocation on such a path is utility max-min fair.*

Now we come to our proposed new fairness criterion, based on the second order utility optimization framework.

Definition 4.3 *Assume, all second order utilities $F_s(\cdot)$ are of the form (4). A rate vector $(x_s, s \in S)$ is called* utility proportional fair *if for any other feasible rate vector $(y_s, s \in S)$ the following optimality condition is satisfied:*

$$\begin{aligned} \sum_{s\in S} \frac{\partial F_s}{\partial x_s}(x_s)(y_s - x_s) &= \sum_{s\in S} G_s^{-1}(x_s)(y_s - x_s) \\ &= \sum_{s\in S} f_s^{-1}(U_s(x_s))(y_s - x_s) \\ &\leq 0 \end{aligned} \tag{11}$$

The above definition ensures, that any proportional utility fair rate vector will solve the utility optimization problem (6), (7). If we further assume, all users have the same transformation function $f(\cdot) = f_s(\cdot), s \in S$, then we have the following properties of a utility proportional fair rate allocation.

Theorem 4.1 *Suppose all users have a common transformation function $f(\cdot)$ and all second order utility functions are defined by (4). Let the rate vector $(x_s \in X_s,\ s \in S)$ be proportional utility fair, i.e. the unique solution of (6). Then the following properties hold:*

(i) The rate vector $(x_s \in X_s,\ s \in S)$ is path utility max-min fair.

(ii) If $q_{s_1} \in Q_{s_1}$, $q_{s_2} \in Q_{s_2}$ and $q_{s_1} \leq q_{s_2}$ for sources s_1, s_2, then $U_{s_1}(x_{s_1}) \geq U_{s_2}(x_{s_2})$.

(iii) If source s_1 uses a subset of links that s_2 uses, i.e. $L(s_1) \subseteq L(s_2)$, and $U_{s_1}(x_{s_1}) < u_{s_1}^{max}$, then $U_{s_1}(x_{s_1}) \geq U_{s_2}(x_{s_2})$.

Proof: To (i): if sources $s \in S_{L_p}$ share the same path, they receive the same aggregate congestion feedback in equilibrium $q_p = q_s,\ s \in S_{L_p}$. Two cases are of interest.
(a) Suppose for all sources the following inequality holds: $f(q_p) < u_s^{max},\ \ s \in S_{L_p}$. Hence, all sources adapt their sending rates according to the available utility $f(q_p) = U_s(x_s)$. This corresponds to the trivial case of path utility max-min fairness, since all sources receive equal utility.
(b) Suppose a set $s \in Q \subset S_{L_p}$ receives utility $U_s(x_s) = u_s^{max} < f(q_p),\ \ s \in Q$ in equilibrium. We prove the theorem by contradiction. Assume the utility proportional fair rate vector $(x_s, s \in S)$ is not path utility max-min fair with respect to the path L_p. By definition, there exists a feasible rate vector $y_s, s \in S$ with

$$U_j(y_j) > U_j(x_j) \ \text{ for } j \in S_{L_p} \setminus Q \tag{12}$$

such that for all $k \in S_{L_p} \setminus (Q \cup \{j\})$ with $U_k(x_k) \leq U_j(x_j)$ the inequality

$$U_k(y_k) \geq U_k(x_k) \tag{13}$$

holds. In other words, we can increase the utility of a single source rate $U_j(x_j)$ to $U_j(y_j)$ by increasing the rate x_j to y_j without decreasing utilities $U_k(y_k), k \in S_{L_p} \setminus (Q \cup \{j\})$ which are already smaller. We represent the rate increase of source j by $y_j = x_j + \xi_j$, where $\xi_j > 0$ will be chosen later on. Here again, we have to consider two cases:

(b1) Suppose, there exists a sufficiently small $\xi_j > 0$ that we do not have to decrease any source rate of the set
$\{y_k, k \in S_{L_p} \setminus (Q \cup \{j\})\}$ to maintain feasibility. Hence, the rate vector
$y := (x_1, x_2, ..., x_{j-1}, y_j, x_{j+1}, ..., x_{|S|})$ is feasible and with (12) the difference of aggregate second order utility $\sum_{s\in S} F_s(y_s) - \sum_{s\in S} F_s(x_s) = F_j(y_j) - F_j(x_j) > 0$ is positive. Therefore x is not optimal to problem (6) contradicting the utility proportional fairness property of x.
(b2) Suppose, we have to decrease a set of utilities $(U_k(y_k), k \in K)$, which are higher then $U_j(x_j)$, i.e. $U_k(y_k) < U_k(x_k)$ with $U_k(y_k) > U_j(x_j), k \in K \subset S_{L_p} \setminus (Q \cup \{j\})$. This correspond to decreasing the set of rates $y_k = x_k - \xi_k$, $k \in K$ with $\sum_{k\in K} \xi_k \leq \xi_j$. Due to the strict concavity of the objective functions of (6), we get the following inequalities:

$$F_j^{'}(x_j) = f^{-1}(U_j(x_j)) > f^{-1}(U_k(y_k)) = F_k^{'}(y_k),\ k \in K \subset S_{L_p} \setminus (Q \cup \{j\}).$$

Due to the continuity of $F_s^{'}(\cdot), s \in S$, we can choose ξ_j with $y_j = x_j + \xi_j$ such that

$$F_j^{'}(x_j + \upsilon_j \xi_j) > F_k^{'}(y_k) \ \text{ for all } \ k \in K \subset S_{L_p} \setminus (Q \cup \{j\}) \text{ and } \ \upsilon_j \in (0,1).$$

Comparing the aggregate second order utilities of the rate vectors x and y using the mean value theorem, we get:

$$\begin{aligned}
\sum_{s\in S} F_s(x_s) - \sum_{s\in S} F_s(y_s) &= \sum_{k\in K} (F_k(x_k) - F_k(y_k)) + F_j(x_j) - F_j(y_j) \\
&= \sum_{k\in K} (F_k(y_k + \xi_k) - F_k(y_k)) + F_j(x_j) - F_j(x_j + \xi_j) \\
&= \sum_{k\in K} (F_k(y_k) + F_k^{'}(y_k + \upsilon_k \xi_k)\xi_k - F_k(y_k)) \\
&\quad + F_j(x_j) - (F_j(x_j) + F_j^{'}(x_j + \upsilon_j \xi_j)\xi_j \\
&= \sum_{k\in K} F_k^{'}(y_k + \upsilon_k \xi_k)\xi_k - F_j^{'}(x_j + \upsilon_j \xi_j)\xi_j \\
&\leq \sum_{k\in K} \xi_k \max_{k\in K}(F_k^{'}(y_k + \upsilon_k \xi_k)) - F_j^{'}(x_j + \upsilon_j \xi_j)\xi_j \\
&\leq \xi_j (\max_{k\in K}(F_k^{'}(y_k + \upsilon_k \xi_k)) - F_j^{'}(x_j + \upsilon_j \xi_j)) \\
&< 0, \quad \upsilon_j \in (0,1),\ \upsilon_k \in (0,1), k \in K.
\end{aligned}$$

The last inequality shows that x is not the optimal solution to (6). Thus, x cannot be utility proportional fair. This contradicts the assumption and proves that x is path utility max-min fair.

To (ii):Assume $q_{s_1} \in Q_{s_1}$, $q_{s_2} \in Q_{s_2}$ and $q_{s_1} \leq q_{s_2}$ for sources s_1, s_2. Applying (1) to given q_{s_1}, q_{s_2}, we have $f(q_{s_1}) = U_{s_1}(x_{s_1}) \geq f(q_{s_2}) = U_{s_2}(x_{s_2})$ because of the monotonicity of $f(\cdot)$.

To (iii): From $L(s_1) \subseteq L(s_2)$ it follows, that $q_{s_1} \leq q_{s_2}$. Since the available utility $f(\cdot)$ is monotone decreasing in q_s and the bandwidth utility $U_{s_1}(x_{s_1}) < u_{s_1}^{max}$ of user s_1 is not bounded by its maximum value, it follows, that $f(q_{s_1}) = U_{s_1}(x_{s_1}) \geq [f(q_{s_2})]_{u_{s_2}^{min}}^{u_{s_2}^{max}} = U_{s_2}(x_{s_2})$. $\square$

It is a well-known property of the concept of proportional fairness that flows traversing several links on a route receive a lower share of available resources than flows traversing a part of this route provided all utilities are equal. The rationale behind this is that these flows use more resources, hence short connections should be favored to increase system utility. Transferring this idea to utility proportional fairness, we get a similar result. Flows traversing several links receive less utility compared to shorter flows, provided a common transformation function is used. If this feature is undesirable, since the path a flow takes is chosen by the routing protocol and beyond the reach of the single user, the second order utilities can be modified to compensate this effect. We show that an appropriate choice of the transformation functions $f_s(\cdot)$ will assure a utility max-min bandwidth allocation in equilibrium.

Theorem 4.2 *Suppose all users have the same parameter dependent transformation function* $f_s(q_s, \kappa) = q_s^{-\frac{1}{\kappa}}$, $s \in S$, $\kappa > 0$. *The second order utilities* $F_s(x_s, \kappa)$, $s \in S$ *are defined by* (4). *Let the sequence of rate vectors* $x(\kappa) = (x_s(\kappa) \in X_s,\ s \in S)$ *be utility proportional fair. Then* $x(\kappa)$ *approaches the utility max-min fair rate allocation as* $\kappa \to \infty$.

Proof: Since all elements of the sequence $x(\kappa)$ solve (6) subject to (7), the sequence is bounded. Hence, we find a subsequence $x(\kappa_p), p \in \mathbb{N}^+$, such that $\lim_{\kappa_p \to \infty} = x$. We show, that this limit point x is utility max-min fair. The uniqueness of the utility max-min fair rate vector x will ensure that every limit point of $x(\kappa)$ is equal x. This proves the convergence of $x(\kappa)$ to x.

Since all users $s \in S$ use the same transformation function $f_s(q_s) = q_s^{-\frac{1}{\kappa}}$, $s \in S$, the second order utility and its derivative applied to the rate vector $x_s(\kappa)$ have the following form:

$$\begin{aligned} F_s(x_s(\kappa)) &= \int U_s(x_s(\kappa))^{-\kappa} dx_s(\kappa) \\ \frac{\partial F_s}{\partial x_s(\kappa)} &= U_s(x_s(\kappa))^{-\kappa},\ s \in S. \end{aligned}$$

We assume that the limit point $x = (x_s \in X_s,\ s \in S)$ is not utility max-min fair. Then we can increase the bandwidth utility of a user j while decreasing the utilities of other users $k \in K \subset S \setminus \{j\}$ which are larger than $U_j(x_j)$. More formal, it exists a rate vector $y = (y_s \in X_s,\ s \in S)$ and an index $j \in S$ with $U_j(y_j) > U_j(x_j),\ j \in S$ and $U_k(y_k) < U_k(x_k)$ with $U_k(y_k) > U_j(x_j)$ for a subset $k \in K \subset S \setminus \{j\}$. We choose κ_0 so large that for all elements of the subsequence $x(\kappa_p)$ with $\kappa_p > \kappa_0$ the inequalities $U_j(y_j) > U_j(x_j(\kappa_p),\ j \in S$, and $U_k(y_k) < U_k(x_k(\kappa_p))$ with $U_k(y_k) > U_j(x_j(\kappa_p))$ for a subset $k \in K \subset S \setminus \{j\}$ hold. With the inequality $U_j(x_j(\kappa_p)) < U_k(x_k(\kappa_p)),\ k \in K$, we can choose $\kappa_1 > \kappa_0$ large enough such that

$$U_j(x_j(\kappa_p))^{-\kappa_p} > C \cdot U_k(x_k(\kappa_p))^{-\kappa_p}, \tag{14}$$

for all $k \in K$, $\kappa_p > \kappa_1$, and $C > 0$ an arbitrary constant. Hence, there exists a κ_1 large enough that the following inequality holds:

$$U_j(x_j(\kappa_p))^{-\kappa_p} > \sum_{k \in K} \underbrace{(x_k(\kappa_p) - y_k)}_{>0} \max_{k \in K} U_k(x_k(\kappa_p))^{-\kappa_p},\ \kappa_p > \kappa_1. \tag{15}$$

We evaluate the variational inequality (11) given in the definition of utility proportion fairness for the candidate rate vector $(y_s \in X_s,\ s \in S)$ and $\kappa_p > \kappa_1$.

$$\begin{aligned}
& \sum_{s\in S} \frac{\partial F_s}{\partial x_s(\kappa_p)}(x_s(\kappa_p))(y_s - x_s(\kappa_p)) = \sum_{s\in S} U_s(x_s(\kappa_p))^{-\kappa_p}(y_s - x_s(\kappa_p)) \\
= \;& U_j(x_j(\kappa_p))^{-\kappa_p}(y_j - x_j(\kappa_p)) + \sum_{k\in K} U_k(x_k(\kappa_p))^{-\kappa_p}(y_k - x_k(\kappa_p)) \\
= \;& U_j(x_j(\kappa_p))^{-\kappa_p}(y_j - x_j(\kappa_p)) - \sum_{k\in K} U_k(x_k(\kappa_p))^{-\kappa_p}(x_k(\kappa_p) - y_k) \\
> \;& U_j(x_j(\kappa_p))^{-\kappa_p}(y_j - x_j(\kappa_p)) - \max_{k\in K} U_k(x_k(\kappa_p))^{-\kappa_p} \sum_{k\in K}(x_k(\kappa_p) - y_k) \\
> \;& 0, \text{ using (15).}
\end{aligned}$$

Hence, the variational inequality is not valid contradicting the utility proportional fairness property of $x(\kappa_p)$. □

5 Conclusion

We have obtained decentralized flow control laws at links and sources, which are globally stable and provide a utility proportional fair resource allocation in equilibrium. This new fairness criterion ensures that bandwidth utility values of users (applications), rather than rates, are proportional fair in equilibrium. We further showed that a utility proportional fair resource allocation also ensures utility max-min fairness for all users sharing a single path in the network. As a special case of our model, we incorporate utility max-min fairness for all users sharing the network. To the best of our knowledge, this is the first paper dealing with resource allocation problems in the context of global optimization, that includes non-concave bandwidth utility functions. We believe that this framework has a great potential in providing real-time services for a growing number of multimedia applications in future networks.
An open issue and challenge is to design bandwidth utility functions that accurately map the bandwidth allocated for any application into user-perceived satisfaction. Furthermore, pricing issues must be considered when designing such a network architecture. For example, a selfish user can choose a too slowly increasing bandwidth utility function for his application, and will get a higher unfair bandwidth share from the network.

Appendix A

Proof of Theorem 3.1:
Let x^* be the unique set of optimal rates solving (6). With the equation $q_s^* = G_s^{-1}(x_s^*) = f^{-1}(U_s(x_s^*))$, we also have the uniqueness of q_s^*. Further the nonsingularity assumption of R ensures that the vector p^* is also unique. Let $x = (x_s, s \in S)$ be the rate vector with $x = G(q)$, where $q = (q_s, s \in S)$ and $G = (G_s, s \in S)$ are in vector form. If we apply the Karush-Kuhn-Tucker optimality conditions to problem (6), we get the following complementary slackness

condition at the links $l \in L$:

$$\begin{aligned} \sum_{l \in L} p_l^*(y_l^* - c_l) &= 0 \\ p_l^* &\geq 0. \end{aligned}$$

Hence, we have the following condition at each link:

$$(y_l^* = c_l) \text{ or } (y_l^* < c_l \text{ and } p_l^* = 0).$$

Next we consider the following Lyapunov function

$$V(p(t)) = \sum_{l \in L}(c_l - y_l^*)p_l(t) + \sum_{s \in S}\int_{q_s^*}^{q_s(t)}(x_s^* - G_s(w))dw.$$

Differentiating with respect to time t yields:

$$\begin{aligned} \frac{dV}{dt} &= \sum_{l \in L}(c_l - y_l^*)\dot{p}_l + \sum_{s \in S}(x_s^* - G_s(q_s))\dot{q}_s \\ &= (c - y^*)^T\dot{p} + (x^* - x)^T\dot{q} \\ &= (c - y^*)^T\dot{p} + (x^* - x)^T(\sum_{l \in L_s}\dot{p}_l)_{s \in S} \\ &= (c - y^*)^T\dot{p} + (x^* - x)^T R^T\dot{p} \\ &= (c - y^*)^T\dot{p} + (y^* - y)^T\dot{p} \\ &= (c - y)^T\dot{p} \\ &= ((c - y)^T\Gamma(p)(y - c)_p^+ \\ &\leq 0, \end{aligned}$$

where $\Gamma(p) = diag(\gamma_l(p_l))$ is a diagonal matrix and $z_p^+ := \begin{cases} z, & \text{if } p > 0 \\ z^+, & \text{if } p = 0. \end{cases}$

The dynamics of $V(\cdot)$ becomes zero, i.e. $\dot{V} = 0$ only when each link satisfies the conditions $(y_l = c_l)$ or $(y_l < c_l$ and $p_l = 0)$. Thus, the complementary slackness condition is satisfied and the system converges to the unique optimal solution of (6), (7). $\square$

References

[1] S.H. Low and D. E. Lapsley, Optimization Flow Control, I: Basic Algorithm and Convergence, IEEE/ACM Trans. Net., vol. 7, no. 6, pp. 861-874, Dec. 1999.

[2] S. Shenker, Fundamental Design Issues for the Future Internet, IEEE JSAC, vol. 13, 1995, pp. 1176-88.

[3] Schulzrinne, A., Casner, S., RTP: A Transport Protocol for Real-Time Applications, Internet Engineering Task Force, Internet Draft, RFC 1889.

[4] S. Floyd, TCP and Explicit Congestion Notification, ACM Comp. Commun. Review, vol. 24, no. 5, Oct. 1994, pp. 10-23.

[5] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, Rate Control in Communication Networks: Shadow Prices, Proportional Fairness, and Stability, J. Operational Research Society, vol. 49, 1998, pp. 237-52, `http://www.statslab.cam.ac.uk/~frank/rate.html`.

[6] R.J. Gibbens and F.P. Kelly, Resource pricing and the evolution of congestion control, Automatica, no. 35, pp. 1969-1985

[7] S.H. Low, A duality model of TCP flow controls, In Proceedings of ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management.

[8] S.H. Low, F. Paganini, J. Doyle, Internet congestion control, IEEE Control Systems Magazine

[9] S. Athuraliya, V. H. Li, S. H. Low and Q. Yin, REM: Active queue management, IEEE Network 15, 2001, pp. 48-53

[10] S.H. Low, F. Paganini, J. C. Doyle, Scalable Laws for Stable Network Congestion Control, Proceedings of Conference of Decision and Control, 2001

[11] Z. Cao, E.W. Zegura, Utility max-min: An application-oriented bandwidth allocation scheme, Proceedings of IEEE INFOCOM'99, 1999, pp. 793-801

[12] J. Cho, S. Chong, Utility Max-Min Flow Control Using Slope-Restricted Utility Functions, `http://netsys.kaist.ac.kr/Publications`, 2004

[13] F. Paganini, A global stability result in network flow control, Systems and Control Letters 46, no.3, 2002, pp. 153-163 Proceedings of Conference of Decision and Control, 2001