Utility Fair Congestion Control For Real-Time Traffic

Tobias Harks* and Tobias Poschwatta**

Abstract. This paper deals with a new approach to integrate congestion control for real-time applications and elastic traffic into a unified framework. In our previous work, we proposed a new fairness criterion, *utility proportional fairness*, that takes characteristics of real-time applications into account. We complement this framework by deriving a general method to generate utility functions for layered multimedia applications. Finally, we demonstrate our approach through ns-simulations.

1 Introduction

In the last years, there have been several papers [1,2,3,4] that interpreted congestion control of communication networks as a distributed algorithm at sources and links in order to solve a global optimization problem. Each user is associated with an increasing strictly concave bandwidth utility function representing elastic traffic. The congestion control algorithms are aimed to maximize aggregate utility subject to capacity constraints on the links. Even though considerable progress has been made in this direction, the existing work focuses only on elastic traffic, such as file transfer (FTP, HTTP) or electronic mail (SMTP). As shown in [5], some applications, especially real-time applications have non-concave 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 described by a step-function.

Instead of relying on traditional fairness criteria (max-min or proportional fairness), we turn the focus on fairness of user-received utility. 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.

In [6], Cao and Zegura present a link algorithm that achieves a utility max-min fair bandwidth allocation, where for each link the utility functions of all flows sharing that link is maintained. In [7], 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. Liao et. al. present in [8] a utility-based approach for wireless access networks, where the links maintain a per-flow aggregate to allocate resources utility fair.

In [9], we have proposed a new fairness criterion, *utility proportional fairness*, that includes utility max-min fairness as a special case. The utility proportional fair bandwidth allocation is derived from the solution of an associated convex program, which is obtained via transforming non-concave bandwidth utilities into appropriate new concave *second order utility functions*. This problem is of the same type as those considered in [1,2,3,4], but it generalizes the fairness characterizations given in [4] and extends previous work by including real-time applications into a unified optimization framework. The contribution of this paper is to translate the theoretical framework into practical application. The missing link in [9] is how to generate utility functions for real-time applications respecting a minimum slope requirement. Here, we present a fairly general method to construct smooth real-time utility functions satisfying that requirement. Armed with this method, we present simulation results, demonstrating stable network performance for real-time applications coexisting with elastic traffic. We emphasize that our implementation does not rely on per-flow states. Instead, we use AQM techniques such as Random Early Marking (REM)[10].

2 Fluid Flow 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

* Tobias Harks works at Zuse Institut Berlin (http://www.zib.de) and has been supported by the German research funding agency 'Deutsche Forschungsgemeinschaft' under the graduate program 'Graduiertenkolleg 621 (MAGSI/Berlin)'

** Tobias Poschwatta works at the Telecommunication Networks Group (TKN) of Technische Universität Berlin (http://www.tkn.tu-berlin.de) and has been supported by the WWI Ambient Networks project, part of the Sixth Framework Programme (FP6) of the European Commission (Project 507134)

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. A transmission rate $x_s \in X_s = [0, x_s^{max}]$ in *packets per second* is associated with each source s. A rate vector $x = (x_s, s \in S)$ is said to be *feasible* if it satisfies the conditions: $x_s \in X_s \ \forall s \in S$ and $\sum_{s \in S(l)} x_s \leq c_l \ \forall l \in L$. With each link l, a scalar positive congestion-measure p_l, called *price*, is associated. Let $y_l = \sum_{s \in S(l)} x_s$ be the aggregate transmission rate of link l, i.e. the sum over all rates using that link, and let $q_s = \sum_{l \in L(s)} p_l$ be the end-to-end congestion measure of source s. 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

$$U_s(x_s) \in C^1 : X_s \to Y_s,\ Y_s = [U_s(0), U_s(x_s^{max})] = [u_s^{min}, u_s^{max}],\ U_s'(x_s) > 0\ \forall x_s \in X_s,\ s \in S. \tag{1}$$

The positivity assumption on the first derivative ensures the existence of the inverse function $U_s^{-1}(\cdot)$ over the range $[u_s^{min}, u_s^{max}]$. Since we focus on utility fairness, an equilibrium point 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. To follow this paradigm, we 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, the *available utility*, as the transformation of the congestion measure q_s by a *transformation function* $f_s(q_s)$. The transformation function is assumed to be continuous and strictly decreasing, i.e. $f_s'(q_s) < 0\ \forall q_s > 0$.

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: $U_s(x_s) = [f_s(q_s)]_{u_s^{min}}^{u_s^{max}}, s \in S,$ where $[w]_a^b := \min\{\max\{w, a\}, b\}$. 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 $q_s < q_s^{min} := f_s^{-1}(u_s^{max})$, then the source transmits data at maximum rate $x_s^{max} := U_s^{-1}(u_s^{max})$; if q_s is above a threshold $q_s > q_s^{max} := f_s^{-1}(u_s^{min})$, 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 $x_s = U_s^{-1}(f_s(q_s))$. Using $f_s'(q_s) < 0\ \forall q_s > 0$ and the chain rule we get:

Lemma 1. *The demand curve $G_s(q_s) = U_s^{-1}([f_s(q_s)]_{u_s^{min}}^{u_s^{max}})$ 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.*

2.1 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^*. The sending rates $x_s^* \in X_s,\ s \in S$ satisfy:

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

With the demand curve $G_s(q_s)$, we construct the strictly concave second order utility functions as

$$F_s(x_s) = \int G_s^{-1}(x_s) dx_s \quad \text{with} \quad F_s'(x_s) = G_s^{-1}(x_s). \tag{4}$$

Using the second order utilities, we interpret the resource allocation problem as the optimization problem

$$\max_{x \in X_s} \sum_{s \in S} F_s(x_s) \quad \text{s.t.:} \quad \sum_{s \in S(l)} x_s \leq c_l\ \forall l \in L. \tag{5}$$

At fixed prices $q_s = \sum_{l \in L(s)} p_l$, i.e. Lagrange multipliers, the problem can be decomposed into S subproblems that can be solved by the users using (2). To get optimal prices $p_l, l \in L$, we can apply a gradient projection method to the dual:

$$\frac{d}{dt} p_l(t) = \begin{cases} \gamma_l (y_l(t) - c_l + \alpha_l b_l(t)) & \text{if } p_l(t) > 0 \\ \gamma_l [y_l(t) - c_l + \alpha_l b_l(t)]^+ & \text{if } p_l(t) = 0, \end{cases} \tag{6}$$

where $[z] = max\{0, z\}$ and $\gamma_l, \alpha_l > 0$. (6) is implemented in REM [10]. Due to the transformation of non-concave utilities in strictly concave ones, the resulting system optimization problem is of the same type as those considered in [1,2,3,11]. Using Lyapunov techniques along the lines of [11], it can be shown that (6) combined with the static source law (2) converges asymptotically to the solution of (5) from any initial condition.

2.2 Utility Proportional Fairness

Before we come to the definition of utility proportional fairness, 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. A utility max-min fair rate allocation is characterized that a user cannot increase its utility, without decreasing the utility of another user, which receives already a smaller utility [6]. Now we come to our proposed fairness criterion, based on the second order utility optimization framework.

Definition 1. *Assume, all second order utilities $F_s(\cdot)$ are of the form (4) and all users use the same transformation function $f_s(x_s) = f(x_s),\ s \in S$. A rate vector $x = (x_s, s \in S)$ is called* utility proportional fair *if for any other feasible rate vector $(y_s, s \in S)$ the following optimality condition is satisfied:*

$$\sum_{s \in S} \frac{\partial F_s}{\partial x_s}(x_s)(y_s - x_s) = \sum_{s \in S} G_s^{-1}(x_s)(y_s - x_s) = \sum_{s \in S} f^{-1}(U_s(x_s))(y_s - x_s) \leq 0. \tag{7}$$

The above definition ensures, that any proportional utility fair rate vector will solve the utility optimization problem (5). We have derived in [9] the following properties of a utility proportional fair rate allocation.

Theorem 1. *Let the rate vector $x = (x_s \in X_s,\ s \in S)$, and $x(\kappa),\ \kappa > 0$ (for (iii)) be utility proportional fair, i.e. the unique solution of (5). Then the following properties hold:*

(i) The rate allocation for $s \in S_p$ sharing the same path, i.e. $L(s) = L_p,\ s \in S_p$, is utility max-min fair.
(ii) If s_1 uses a subset of links of s_2, i.e. $L(s_1) \subseteq L(s_2)$, and $U_{s_1}(x_{s_1}) < u_{s_1}^{max}$, then $U_{s_1}(x_{s_1}) \geq U_{s_2}(x_{s_2})$.
(iii) If all users use the transformation function $f_s(q_s, \kappa) = q_s^{-\frac{1}{\kappa}},\ s \in S,\ \kappa > 0$ then $x(\kappa)$ approaches the utility max-min fair rate allocation as $\kappa \to \infty$.

The above theorem allows to approximate utility max-min fairness to any degree: the higher the parameter κ, the less will flows be penalized for using more resources (see Fig. 3(b), 3(c), and 3(d)).

3 Utility Function Generation for Adaptive Real-Time Traffic

In this paper, we consider a multi-layer video streaming application as an example for real-time traffic [12]. The user-perceived satisfaction of video applications with multi-layer streams, i.e. a base layer and multiple enhancement-layer streams, can be described by multiple-step functions. Each step corresponds to an encoding layer and represents the achieved utility at that layer.[1]
We will briefly outline a constructive method to approximate single- and multiple-step functions. To accommodate this, we propose the following sigmoidal logistic function with strictly positive derivative (i.e. respecting (1)) to approximate a single-step function:

$$U_s(x_s) = \frac{a_s}{1 + \exp(-b_s(x_s - c_s))^{d_s}} + e_s, \tag{8}$$

where $a_s > 0, b_s > 0, c_s \geq 0, d_s > 0, e_s \geq 0$ are normalization factors.

Definition 2. *A rising range $R(\epsilon) = [r_1(\epsilon), r_2(\epsilon)]$ is a real interval such that for a fixed $\epsilon > 0$ the following holds: $U_s'(x_s) > \epsilon$, for all $x_s \in R(\epsilon)$.*

Lemma 2. *The rising range $R(\epsilon)$ of function (8) goes to $\{c_s\}$ as $d_s \to \infty$, (see Fig. 1).*

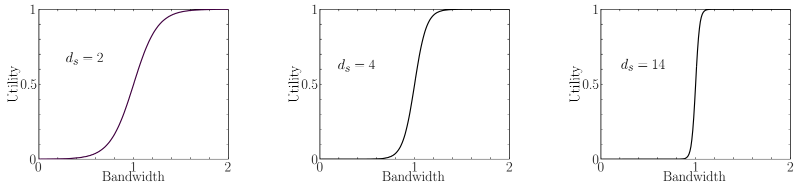


Fig. 1. Fixed values $a_s = 1, b_s = 4, c_s = 1, e_s = 0$ and variable d_s

Proof. Let $\epsilon > 0$ be fixed. We prove the Lemma by contradiction. Suppose there exist real numbers $y_s \in R(\epsilon)$ with $y_s \neq c_s$. Then,

$$U_s^{'}(y_s) = \frac{a_s b_s d_s \exp(-b_s(y_s - c_s))^{d_s}}{(1 + \exp(-b_s(y_s - c_s))^{d_s})^2} \leq \frac{a_s b_s d_s \exp(-b_s(y_s - c_s))^{d_s}}{(\exp(-b_s(y_s - c_s))^{d_s})^2} = \frac{a_s b_s d_s}{(\exp(-b_s(y_s - c_s))^{d_s})} < \epsilon,$$

for d_s large enough. □

Thus, we have found a function, which approximates the single-step function to any degree and respects requirement (1). To obtain a smooth multiple step function, we combine several smooth single-step functions. Let us consider a real-time application (i.e. multi-layered video) that can adapt its sending rate according to n different coding layers $\{c_{s,1}, c_{s,2}, ..., c_{s,n}\}$ with $c_{s,i} < c_{s,i+1},\ i = 1, ..., n-1$. We will construct the corresponding approximation of the multiple step function in a way that the coding levels $c_{s,i}$ are the inflection points of the corresponding i-th single-step function. More generally, the coding levels $c_{s,i}$ lie in the rising range $R_i(\epsilon)$ for all $\epsilon \leq \max_{x_s \in X_s} U_s^{'}(x_s)$. Using similar ideas as in Lemma 3, we can prove the following:

Lemma 3. *Let us consider the following approximation of a multiple step function:*

$$U_s(x_s) = \frac{a_{s,i}}{1 + \exp(-b_{s,i}(x_s - c_{s,i}))^{d_{s,i}}} + \sum_{j=1}^{i-1} a_{s,j},\ for\ c_{s,i} - \alpha_{s,i} \leq x_s < c_{s,i+1} - \alpha_{s,i+1},\ i = 1, ...n \qquad (9)$$

with $\alpha_{s,1} := c_{s,1},\ \alpha_{s,i} := \frac{c_{s,i} - c_{s,i-1}}{2} > 0,\ i = 2, ...n$. *Then we have for* $i = 1, ...n$*:*

(i) $c_{s,i} \in R_i(\epsilon)\ \forall \epsilon \leq \max_{x_s \in X_s} U_s^{'}(x_s)$

(ii) $U_s(x_s) \in C^1[0, x_s^{max}]$, *i.e.* $U_s(x_s)$ *is continuously differentiable over the range* $[0, x_s^{max}]$ *for* $d_s \to \infty$.

4 Simulation Results

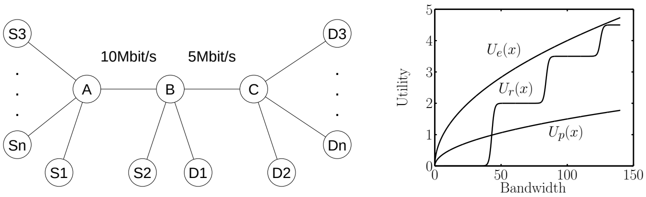
To demonstrate the effectiveness of our utility-based congestion control approach we performed simulations using the *NS-2* network simulator[13] emphasizing on adaptation of real-time flows, prioritization and utility proportional fair bandwidth allocation. Our senders are rate-based, i.e. packets are sent at intervals of $1/x_s$ seconds. We limit the changes of x_s to not increase by more than 1 and not decrease by more than half in one round-trip time to smooth sender behaviour on startup and when adapting to changing network load. Link prices are accumulated in a double precision floating point field in the packet headers that receivers return to the senders in acknowledgement packets. Data packets have a fixed size of 1500 bytes and receivers acknowledge every data packet immediately.

Figure 2 shows the network topology and bandwidth utility functions used in the simulations. There are two elastic utility functions, U_e and prioritized U_p[2], and one real-time utility function U_r that models a video streaming application with three supported coding layers: 512kbit/s ($c_{r,1} = 43$pps), 1024kbit/s ($c_{r,2} = 84$pps), and 1.5Mbit/s ($c_{r,3} = 125$pps). All senders use the same transformation function $f_s(q_s) = q_s^{-\frac{1}{\kappa}}$.

The results of four simulation setups are shown in Figure 3. Our first simulation focuses on adaptation of a real-time flow when network load changes and on prioritization of flows. The four graphs in Figure 3(a)

[1] More important than the absolute value of utility values is the distance to utility functions of competing traffic.

[2] For a given utility value, a prioritized flow receives a larger bandwidth share than competing non-prioritized flows.



$$U_r(x) = \begin{cases} \frac{2}{1+\exp(-(x-43))}, & 0 \le x < 63.5 \\ \frac{1.5}{1+\exp(-(x-84))} + 2, & 63.5 \le x < 104.5 \\ \frac{1}{1+\exp(-(x-125))} + 2 + 1.5, & 104.5 \le x \le 130 \end{cases}$$

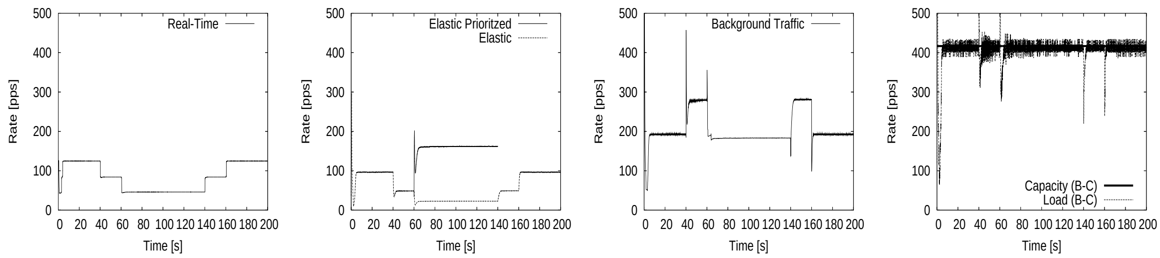
$$U_e(x) = 0.4\sqrt{x}$$

$$U_p(x) = 0.15\sqrt{x}$$

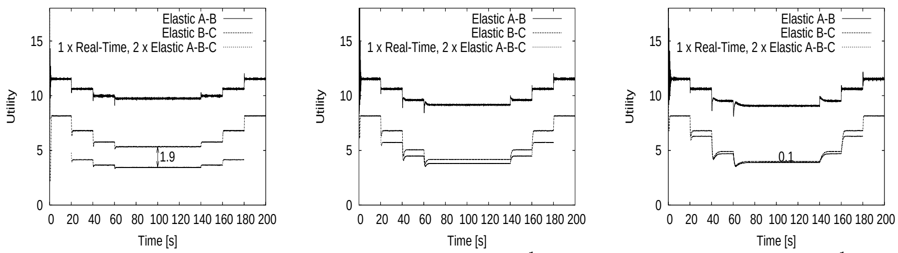
Fig. 2. Network topology and utility functions using (9) with $a_{r,1} = 2, a_{r,2} = 1.5, a_{r,3} = 1, b_{r,i} = 1, d_{r,i} = 1$

show the sending rate of one real-time flow (using U_r), the rates of two elastic flows (prioritized U_p and non-prioritized U_e) and the aggregated rates of additional background traffic (up to four elastic and two real-time flows), which together yield the total load shown in the fourth graph. Starting with the highest encoding rate (125pps), the real-time flow switches to lower rates as more flows start after 40 and 60 seconds. Flows terminate after 140 and 160 seconds which allows the real-time flow to switch back to higher encoding rates. The prioritized elastic flow starts after 60 and ends after 140 seconds and receives a signifcant higher bandwidth share, although it receives the same utility as the other flows. In this setup, there are no short flows using only the A-B or B-C links and all senders have $\kappa = 1$.

The remaining three Figures 3(b), 3(c) and 3(d) show the effect of an increasing κ for $\kappa = 1, 2$ and 3. Here, two short flows using the A-B and B-C links respectively, compete with up to three long flows using the whole A-B-C path. As can be seen, the difference in received utility between short and long flows competing for the same bottleneck link (B-C) decreases for increasing values of κ.



(a) Adaptation and prioritization of flows



(b) $f_s(q_s) = \frac{1}{q}$ (c) $f_s(q_s) = q^{-\frac{1}{2}}$ (d) $f_s(q_s) = q^{-\frac{1}{3}}$

Fig. 3. Simulation Results

5 Conclusion

The concept of Utility Proportional Fairness is a promising approach to integrate congestion control for real-time and elastic traffic into a unified framework. Here, a crucial problem is the construction of real-time utility functions. In this paper, we have presented a mathematical method that enabled us to efficiently construct bandwidth utility functions for real-time applications. Moreover, we have applied this method to layered video applications and validated our approach via simulations. One of the next steps to improve our framework is to address the issue of choosing parameters γ_l, α_l. The tradeoff between stability (w.r.t. delay) and performance needs further investigation.

References

1. Low, S.H., Lapsley, D.E.: Optimization Flow Control I. IEEE/ACM Trans. on Networking **7** (1999) 861–874
2. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate Control in Communication Networks: Shadow Prices, Proportional Fairness, and Stability. Journal of the Operational Research Society **49** (1998) 237–52
3. Low, S., Paganini, F., Doyle, J.C.: Internet Congestion Control. IEEE Control Systems Magazine **22** (2002)
4. Mo, J., Walrand, J.: Fair end-to-end window-based congestion control. IEEE/ACM Trans. on Networking **8** (2000) 556–567
5. Shenker, S.: Fundamental Design Issues for the Future Internet. IEEE JSAC **13** (1995) 1176–88
6. Cao, Z., Zegura, E.: Utility max-min: An application-oriented bandwidth allocation scheme. In: Proceedings of IEEE INFOCOM'99. (1999) 793–801
7. Cho, J., Chong, S.: Utility Max-Min Flow Control Using Slope-Restricted Utility Functions. Available at `http://netsys.kaist.ac.kr/Publications` (2004)
8. Liao, R.F., Campbell, T.: A utility-based approach for quantitative adaption in wireless packet networks. Wireless Networks **7** (2001) 541–557
9. Harks, T.: Utility Proportional Fair Resource Allocation: An Optimization Oriented Approach. In: Proceedings of QoS in Multiservice IP Networks, Catania, Italy, Springer (2005) 61–74 to appear.
10. Athuraliya, S., Li, V.H., Low, S.H., Yin, Q.: REM: Active Queue Management. IEEE Network **15** (2001) 48–53
11. Paganini, F.: A global stability result in network flow control. Systems and Control Letters **46** (2002) 165–172
12. Gharavi, H., Partovi, M.H.: Multilevel video coding and distribution architectures for emerging broadband digital networks. IEEE Trans. on Circuits and Systems for Video Technology **6** (1996) 459–469
13. McCanne, S., Floyd, S.: The network simulator - ns-2. http://www.isi.edu/nsnam/ns/ (2000)