Utility based fair bandwidth allocation

Corinne Touati, Eitan Altman, Jérôme Galtier
{ctouati,altman,galtier}@sophia.inria.fr

Abstract For over a decade, the Nash Bargaining Solution (NBS) concept from cooperative game theory has been used in networks as a concept that allows sharing resources fairly. Due to its many appealing properties, it has recently been used for assigning bandwidth in a general topology network between applications that have linear utilities. In this paper, we use this concept for the bandwidth allocation between applications with general concave utilities and focus on the case of quadratic utility.

Keywords Nash bargaining, bandwidth allocation, fairness

1 Introduction

Fair bandwidth assignment has been one of the important challenging areas of research and development in networks supporting elastic traffic. Indeed, Max-min fairness has been adopted by the ATM forum for the Available Bit Rate (ABR) service of ATM [1]. Although the max-min fairness has some optimality properties in the sense of Pareto, it has been argued that it favors too much long connections and does not make efficient use of available bandwidth. In contrast, the concept of proportional fairness (of the throughput assignment) has been proposed by Kelly [10, 6], and gives rise to a more efficient solution in terms of network resources by providing more resources to shorter connections. An assignment is proportionally fair if any change in the distribution of the assigned rates would result in the sum of the proportional changes to be non-positive.

Although the object that is shared fairly seems to be a very specific one, the throughput, it is shown in [10, 6] that in fact, the starting point for obtaining (weighted) proportional fairness of the throughput can be a general (concave) utility function per connection; it is then shown that a global minimization of the (weighted) sum of these utilities leads to a weighted proportional fair assignment of the throughput. As opposed to this approach, we wish to use a fairness concept which is defined directly in terms of the utilities of users rather than in terms of the throughputs they are assigned. Yet, as in weighted proportional fairness, it would be desirable to obtain this concept as the solution of a utility maximization problem, since it makes it possible to use recent algorithms for utility maximization in networks, along with decentralized implementations [9, 8, 11].

The Nash Bargaining Solution (NBS) is a natural framework that allows us to define and design fair assignment of bandwidth between applications with different concave utilities and has already been used in networking problems [13, 7]. It is characterized by a set of axioms that are appealing in defining fairness. As already recognized in [6] and later in [7], proportional fairness agrees with NBS when the object that is shared fairly is the throughput (and the minimum required rate is zero). We use NBS to study the fairness of an assignment where connection i has a concave utility over an interval $[MR_i, PR_i]$. It thus has a minimum rate requirement MR_i and does not need more than PR_i. Utility functions with similar features have been identified in [15] for representing some real time applications such as voice and video, and for elastic traffic in the case $MR_i = 0$.

We study in this paper the way the concavity of the utilities affects the bandwidth assignment according to NBS, as well as according to a generalized version of the proportional fairness (in which the utilities that correspond to different assignments, instead of the throughputs, are fairly allocated). Both notions are introduced in Sec. 2. We then propose in Sec. 3 a quadratic approximation for the utility of each connection, which allows us to parameterize the degree of concavity of the utility function using a single parameter. In Section 4 we provide an example of allocation of bandwidth in a large network, and we conclude in Section 5 with some final comments.

2 General problem

2.1 Utility function

The fairness problem which we consider is how to allocate bandwidth to connections beyond their minimum required bandwidth (MR). (We assume that if the minimum required bandwidth is not available then the connection is not accepted by the network.) The fairness issue is only interesting when the utility of an application strictly increases when allocating more bandwidth than its MR. Connections with on/off utility functions (which characterize some applications with hard real-time requirements, [15]) are thus ignored in allocating extra bandwidth once they receive their MR.

Two kinds of applications are considered in [15] for which the fair allocation is relevant:

Elastic applications: Examples of such applications are file transfer or email. The typical utility function is concave increasing without a required minimum rate, see Fig. 1.

"Delay adaptive" or "rate adaptive" applications: These are typically real time applications such as voice or video over IP. The utility functions that we use for these applications (Fig. 1) are slightly different than those in [15]. In [15], the utility is always strictly positive for non null bandwidth and tends to zero when the bandwidth does. We consider in contrast the utility being equal to zero below a certain value, as in [7]. Indeed, in many voice applications, one can select the transmission rate by choosing an appropriate compression mechanism and existing compression software have an upper bound on the compression, which implies a lower bound on the transmission rate for which a communication can be initiated. If there is no sufficient bandwidth, the connection is not initiated. This kind of behavior generates utility functions that are zero for bandwidth below MR and which are not differentiable at the point $(MR, 0)$.

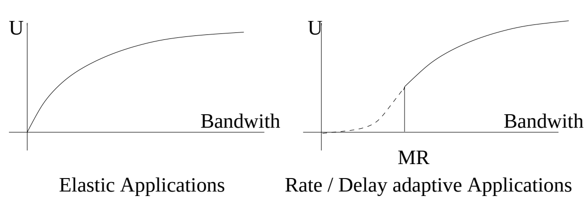

Figure 1: Utility function of elastic (left) and of rate-adaptive or delay-adaptive (right) applications

2.2 Fair allocations

Several concepts of fairness are known in the literature: the max-min fairness [3], (as well as the more general concept of weighted max-min fairness) which has been adopted by the ATM-forum [1] for ABR traffic, the proportional fairness [6], the harmonic mean fairness [12], the general fairness criterion that bridges all the above concepts [14] and the Nash Bargaining Solution (NBS).

Nash Bargaining Solution (NBS) Our starting point is the NBP (Nash Bargaining Point) concept [7] for fair allocation, which is frequently used in cooperative game theory. It deals directly with fair allocation of achievable utilities of players (and does not require to relate them to the original objects, throughputs in our case, that generate these utilities). Let there be n users (or connections). Let $\mathbb{R}^n$ represent the vectors of utilities of the form $(f_1, ..., f_n)$. Let $u^0 = (u_i^0)_{i \in 1 \cdots n}$ be a minimum required performance of user i.[1] If X is the set of all achievable vectors of bandwidths, then $U = \{f(x) | x \in X\}$. Let $\mathcal{G} = \{(U, u^0) | U \subset \mathbb{R}^n\}$: it denotes the class of sets of performance measures that satisfy the minimum performance bound u^0 (it contains achievable performances obtained for different utility functions f; in fact, in order to define NBP one has to introduce its performance w.r.t. other utilities, as is seen from property 3 and 5 in the definition below).

Definition 2.1. *A mapping $S : \mathcal{G} \to \mathbb{R}^n$ is said to be a NBS (Nash Bargaining Solution) if*

1. *$S(U, u^0) \in U^0 := \{u \in U | u \geq u^0\}$, i.e. it guarantees the minimum required performances.*
2. *It is Pareto optimal.* [2]
3. *It is linearly invariant, i.e. the bargaining point is unchanged if the performance objectives are affinely scaled. More precisely, if $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map such that its i^{th} component is given by $\phi_i(u) = a_i u_i + b_i$, then $S(\phi(U), \phi(u^0)) = \phi(S(U, u^0))$.*
4. *S is symmetric i.e. does not depend on the specific labels, i.e. users with the same minimum performance measures and the same utilities will have the same performances.*
5. *S is not affected by enlarging the domain if a solution to the problem with the larger domain can be found on the restricted one. More precisely, if $V \subset U$, $(V, u^0) \in \mathcal{G}$, and $S(U, u^0) \in V$ then $S(U, u^0) = S(V, u^0)$.*

The definition of NBP is thus given through axioms that game theorists find natural to require in seeking for fair assignment. Having defined this concept through the achievable utilities, we define the NBS (Nash Bargaining Solution) in terms of the corresponding strategies (i.e. the allocation of bandwidth

[1] In our context, $u_i^0 = f_i(MR_i)$ where f_i is concave increasing.

[2] An allocation f is said to be Pareto optimal if it is impossible to strictly increase the allocation of a connection without strictly decreasing another one. The Pareto axiom assures that no bandwidth is "wasted".

that results in the NBP), and then present its characterization through a utility optimization approach.

Definition 2.2. *The point $u^* := S(U, u^0)$ is called the Nash Bargaining Point and $f^{-1}(u^*)$ is called the set of Nash Bargaining Solutions.*

Define $X_0 := \{x \in X | f(x) \geq u^0\}$.

Theorem 2.1. *[7, Thm. 2.1, Thm 2.2]. Let the utility functions f_i be concave, upper-bounded, defined on X which is a convex and compact subset of $\mathbb{R}^n$. Let $\mathcal{J}$ be the sets of achievable indices of users able to achieve a performance strictly superior to their initial performance, i.e. $\mathcal{J} = \{J \subseteq \{1, \ldots, n\}, \exists x \in X_0, \forall j \in J, f_j(x) > u_j^0\}$. Assume that for all $J \in \mathcal{J}$, $\{f_j\}_{j \in J}$ are injective. Then there exists a unique NBP as well as a unique NBS x that verifies $f_j(x) > u_j^0, j \in J$, and is the unique solution of the problem P_J:*

$$(P_J) \quad \max \prod_{j \in J} (f_j(x) - u_j^0), \quad x \in X_0. \tag{1}$$

Equivalently, it is the unique solution of:

$$(P'_J) \quad \max \sum_{j \in J} \ln(f_j(x) - u_j^0), \quad x \in X_0.$$

Before examining some qualitative implications of the definition, we introduce the very related notion of generalized proportional fairness.

Generalized proportional fairness (GPF). An assignment $x \in X$ is said to be (generalized) proportionally fair with respect to a utility f, if for any other assignment $x^* \in X$, the aggregate of proportional changes in the utilities is zero or negative

$$\sum_{i=0}^{n} \frac{f_i(x_i^*) - f_i(x_i)}{f_i(x_i)} \leq 0. \tag{2}$$

Thus, an allocation is GPF if any change in the distribution of the rates would result in the sum of the proportional changes of the utilities to be non-positive. This concept has been defined and applied without considering any utility, i.e. by restricting the object that is assigned fairly to the rates [10, 6] (see also [4, 12, 14]). This amounts in taking in (2) $f_i(x_i) = x_i$. Yet, there is no conceptual difference in defining it as we do, i.e. with respect to utilities. In particular, by simply replacing x_i by $f_i(x_i)$, we have the following property (established for the special case $f_i(x_i) = x_i$) of the solution x^{GPF}:

$$\begin{aligned} &x^{GPF} \text{ maximizes } \sum_{i=1}^{n} \ln f_i(x_i) \text{ over } X \quad \text{or} \\ &x^{GPF} \text{ maximizes } \prod_{i=1}^{n} f_i(x_i) \text{ over } X. \end{aligned} \tag{3}$$

The Internet is an example where proportional fairness is used. Indeed, congestion control mechanisms based on linear increase and multiplicative decrease (such as TCP) achieve proportional fairness upper appropriate conditions [6]. The (weighted) proportional fairness is also advocated for future developments of TCP [5].

Comparing with Thm. 2.1, we conclude that GPF coincides with the NBS of [7] when the MR_i's equal zero, and to the original proportional fairness when further restricting to the identity utilities.

We finally note that due to (3) it follows that GPF is invariant under a scale change, i.e. if we multiply the utility f_i of a connection i by a positive constant c_i, the GPF assignment will not change. Yet in general, it will not remain the same under translation by a constant as in NBS.

General fairness criterion. We present another general fairness criterion [14] but apply it to fair allocation of utilities rather than of the rate. Given a positive constant $\alpha \neq 1$, consider the problem

$$\max_x \frac{1}{1-\alpha} \sum_{i=1}^{n} f_i(x_i)^{1-\alpha}, \ \alpha \geq 0, \alpha \neq 1. \tag{4}$$

subject to the problem's constraints. This defines a unique allocation which is called the α-bandwidth allocation. This allocation corresponds to the globally optimal allocation as $\alpha \to 0$, to the (generalized) *proportional fairness* when $\alpha \to 1$, to the generalized *harmonic mean fairness* when $\alpha \to 2$, and to the generalized *max-min* allocation when $\alpha \to \infty$ [14].

2.3 Statement of the general problem

We focus in the paper on the computation of the NBS and briefly compare it to the GPF allocation. Using Thm.2.1, the NBS is the unique solution $x = x_1, x_2, ..., x_n$ (with n the number of connections) of:

$$\begin{gathered} \max_{x \in X_0} \prod_{i=1}^{n} (f_i(x_i) - f_i(MR_i)) \text{ where } X_0 := \\ \{x | \forall l = 1, ..., L, (Ax)_l \leq (C)_l, MR_i \leq x_i \leq PR_i\}, \end{gathered} \tag{5}$$

with L the number of links, A the routing matrix (the element $A_{l,i}$ being equal to 1 if connection i goes through link l, 0 otherwise), and C the capacity vector (C_l is the capacity of link l). $(Ax)_l \leq (C)_l$ are the standard capacity constraints. We assume that the network has sufficient bandwidth to satisfy all the users' minimum requirements i.e. $\forall l \in 1..L$ we have $\sum_{i=1}^{N} A_{l,i} MR_i < C_l$.

3 Quadratic utility functions

3.1 Definition of the utility function

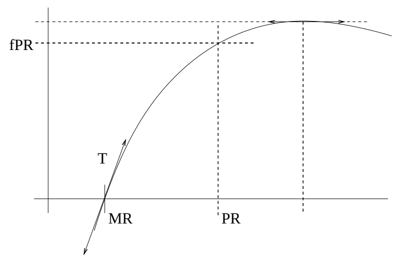

Figure 2: Quadratic utility function.

The utility function of both "elastic traffic" and "delay adaptive" applications have a minimum value MR_i below which it equals zero (in the former case, $MR_i = 0$). As the NBS is shift invariant, we can assume without loss of generality that $f_i(MR_i) = 0$. Beyond MR_i the function is concave and increasing with the bandwidth. We can approximate such a utility function with a parabola with several parameters that may depend on the applications (see Fig.2):

- PR_i: maximum throughput needed by the application
- T_i: tangent of the utility function at the point $(MR_i, 0)$
- fPR_i: utility value at point PR_i

Note that the utility function is defined only until the point PR_i, so we may ignore the whole right part of the parabola (and in particular, the part in which the function decreases).

As the utility function is a parabola, its general equation has the form: $f_i(x_i) = c_i - a_i(x_i - b_i)^2$. Obviously, f_i can equally be defined by a_i, b_i, c_i or through the equations $f_i(MR_i) = 0$, $f_i(PR_i) = fPR_i$ and $f'_i(MR_i) = T_i$. We should note that, since PR_i is in the increasing part of the function,

$$\frac{1}{2}T_i(PR_i - MR_i) \leq fPR_i \leq T_i(PR_i - MR_i).$$

We thus define the *concavity of the utility*, β_i through

$$fPR_i = T_i \cdot \beta_i \cdot (PR_i - MR_i)$$

We have: $1/2 \leq \beta_i \leq 1$ and the smaller β_i is, the more concave is the utility. The limit $\beta_i = 1$ is the linear case (studied in [7]).
Finally:

$$a_i = T_i \frac{1-\beta_i}{PR_i - MR_i}, \qquad b_i = \frac{PR_i - (2\beta_i - 1)MR_i}{2(1-\beta_i)}$$
$$\text{and} \quad c_i = \frac{T_i}{4}\frac{PR_i - MR_i}{1-\beta_i}.$$

We next present several examples where we use our parabolic utility functions.

3.2 The linear network example

We consider the problem in Fig. 3 in which the squares represent the links and the lines represent the routes. We have $N = L + 1$ connections sharing L links. Connection 0 uses all the links, whereas each of the other L connections only goes through a single link (connection i uses link i). All proofs of the following theorems and propositions can be found in [16].

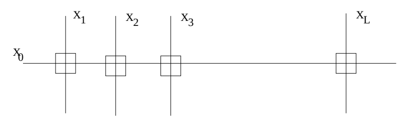

Figure 3: A linear network.

To obtain the NBS, we need to maximize

$$\prod_{i \in \{0,\ldots,L\}} f_i(x_i). \tag{6}$$

But, as NBS is Pareto optimal, we have the following constraints for $i = 1, ..., L$: $x_0 + x_i = C_i$ as well as $MR_i \leq x_i \leq PR_i$. This implies $b_i - \sqrt{\frac{c_i}{a_i}} \leq x_i \leq b_i$.

We make two significant assumptions. First, that each link has the same capacity *cap*. Therefore, it is straightforward to notice that each connection i with $1 \leq i \leq L$ will get the same bandwidth at the equilibrium point. Secondly, we suppose that each of these connections has the same utility function: $\forall i \in \{2, \ldots, L\}, a_i = a_1, b_i = b_1, c_i = c_1$.

Therefore the term to maximize in equation (6) becomes: $f_0(x)(f_1(cap - x))^L$ if we denote by x the throughput of the connection x_0.

Solution of the linear problem. By differentiating (6) we obtain:

$$\begin{aligned} a_0(x - b_0)(c_1 - a_1(cap - x - b_1)^2) = \\ La_1(cap - x - b_1)(c_0 - a_0(x - b_0)^2) \end{aligned} \tag{7}$$

which is a polynomial of the third degree. This can be explicitly solved.

Possible limits. We are interested in the possible limits x_{lim} of the bandwidth assigned to connection x_0 as L grows to infinity.

Lemma 3.1. *Assume $MR_0 + PR_1 \geq cap$. As L grows to infinity, the only possible limit x_{lim} of the bandwidth assigned to connection x_0 is $x_{lim}^{(3)} = b_0 - \sqrt{c_0/a_0} = MR_0$.*

It is interesting to note that the limit x_{lim} does not depend on any parameter of the $i^{th}(i \geq 1)$ connections, or any parameter related to the concavity of the utility function of the connection 0. We show in Fig. 4 how the system converges to the solution as L grows.

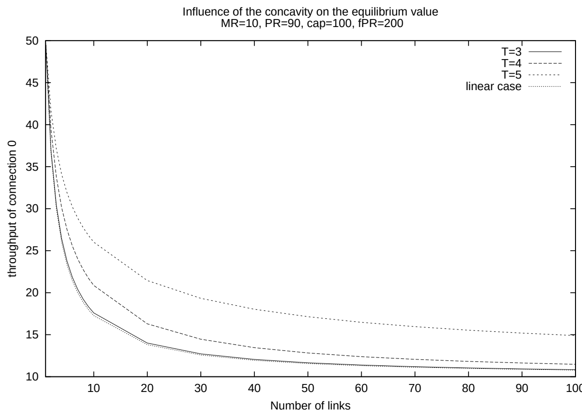

Figure 4: NBS for the linear network.

Remark 3.1. *The condition $MR_0 + PR_1 \geq cap$ in Lemma 3.1 (and in the next propositions) is not restrictive. If it does not hold then we can replace (for any L) MR_0 by $MR'_0 := cap - PR_1$ without affecting the NBS, and then apply Lemma 3.1 for MR'_0. Indeed, let x^* be NBS for the original problem. Then $x^* \geq MR'_0$ due to the Pareto optimality of the NBS (2nd element in Definition 2.1). Then it is the NBS for the new problem due to the 5th element in Definition 2.1.*

Asymptotic Analysis. We further refine the analysis of the limit as L becomes large, show that it exists and obtain the rate at which x converges to x_{lim}.

Proposition 3.1. Suppose that $MR_0 + PR_1 \geq cap$, then x verifies:

$$x = MR_0 + Z + o(1/L) \tag{8}$$

with: $Z = \dfrac{cap - MR_0 - MR_1}{2L}\left[1 - \dfrac{PR_1 - MR_1}{denom}\right]$,

$$denom = 2(1-\beta_1)(cap - MR_0) - PR_1 + MR_1(2\beta_1 - 1)$$

where $o(1/L)$ is a function that, when divided by L, *tends to zero as L grows to infinity.*

We can notice that:

- the convergence of x is in $1/L$,
- the result does not depend on T_0 nor T_1 (scale invariant),
- in the asymptote, none of the parameters of the 0^{th} connection but MR_0 appears, so that the results are independent of the shape of the utility function of x_0,
- the larger β_1 is, the smaller x gets.

In (8), we can easily check the asymptotes for special cases:

- when $\beta_1 \to 1$ we obtain: $Z = \dfrac{cap - MR_0 - MR_1}{L}$ (linear case).
- when $\beta_1 \to 1/2$ we get: $Z = \dfrac{cap - MR_0 - MR_1}{2L}\left[1 - \dfrac{PR_1 - MR_1}{cap - MR_0 - PR_1}\right]$.

4 Numerical results

In this section we present a network example for which we obtain the fair allocation. More details on these examples and on the program used to obtain them can be found in [16]. All links are assumed to have equal capacity (although the program allows to handle different capacities without increasing the complexity). We present two figures. The first with the set of links and nodes and the second with the set of connections and amount of assigned bandwidth. All connections had the same quadratic utility with the parameters $MR = 10$ and $PR = 80$, $T = 3$, $fPR = 200$. We took $cap = 100$ for all links. Bandwidth parameters and assignments are given in percent of full link capacity.

We considered the COST experimental network [2], depicted in Fig. 5. It contains 11 nodes, representing the main European capitals. We have considered the 30 connections with the highest forecast demand (We did not include the connections whose forecast demand, based on experiments dating from 1993, were inferior to 2.5 Gb/s).

Connection	Bandwith	Connection	Bandwith
London−Paris	33.93	Zurich−Milano	71.58
London−Bruxelles	80.00	Copenhaguen−Berlin	80.00
London−Amsterdam	76.27	Copenhaguen−Prague	80.00
Amsterdam−Berlin	27.11	Berlin−Amsterdam−Bruxelles	22.04
Amsterdam−Bruxelles	49.54	Paris−Bruxelles−Amsterdam	28.42
Bruxelles−Paris	43.66	Milano−Viena	63.00
Paris−Berlin	80.00	Berlin−Amsterdam−Luxembourg	27.11
Paris−Zurich	33.19	Zurich−Prague−Berlin	50.00
Paris−Milano	47.34	Zurich−Luxembourg−Amsterdam	35.79
Zurich−Viena	55.06	Zurich−Luxembourg−Bruxelles	35.79
Paris−Zurich−Viena	25.48	Viena−Zurich−Paris−London	19.46
London−Paris−Milano	24.74	Milano−Zurich−Luxembourg−Amsterdam	28.42
Milano−Viena−Berlin	37.00	London−Paris−Zurich	21.87
Milano−Paris−Bruxelles	27.93	London−Amsterdam−Berlin	23.73
Berlin−Prague	50.00	Berlin−Viena	63.00

Figure 7: Bandwidth allocation for COST network.

The experiments allow to identify several independent systems, such as isolated connections (London-Brussels, Paris-Berlin, Berlin-Copenhaguen and Copenhagen-Prague) that are all served at their maximum rate of 80%, connections sharing only one link (Zurich-Prague), served both at 50%, two connections sharing two links (Milano- Viena-Berlin) served at 63%, 63% and 37% respectively.

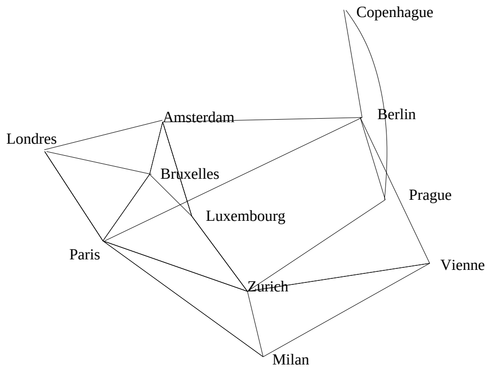

Figure 5: COST network: links.

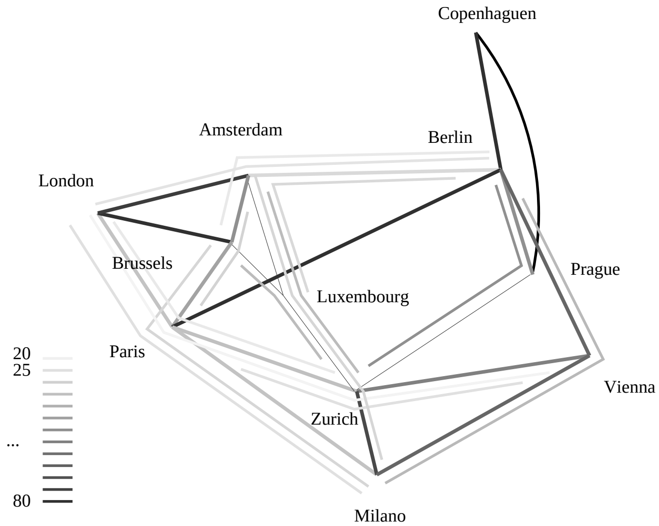

Figure 6: COST network: solution.

For the other connections, we notice several unsaturated links (Luxembourg-Brussels, Luxembourg-Amsterdam). One link receives at most four connections, and the proportional fairness never assigns less than 19.46% of the bandwidth to an individual connection, which is remarquably high, enlightening the interest of our approach.

5 Conclusion

We have applied in this paper the NBS approach for bandwidth allocation, as well as the GPF concept that is sensitive to the utilities of connections. We have studied some of the characteristics of these concepts, and showed that they are indeed more suitable for applications that have concave utility. We proposed a simple parameterization of the concavity of the utility function using quadratic functions. Computational approaches are proposed in [16]: a Lagrangian approach that allows us to handle large networks and an alternative centralized computation approach based on Semi-Definite-programming.

References

[1] Traffic management specification, version 4.0. Technical Report AF-TM-0056.000, ATM Forum Traffic Management Working Group, April 1996.

[2] Cost 239. Ultra high capacity optical transmission networks, final report of Cost project 239. *ISBN 953-184-013-X*, 1999.

[3] D. Bertsekas and R. Gallager. *Data Networks.* Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

[4] T. Bonald and L.Massoulié. Impact of fairness on internet performance. Sigmetrics, 2001.

[5] Jon Crowcroft and P. Oechslin. Differentiated end-to-end internet services using a weighted proportional fair sharing TCP. *Computer Comm. Rev.*, 28(3):53–69, 1998.

[6] A. Maulloo F.P. Kelly and D. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *J. Oper. Res. Society*, 49:237–252, 1998.

[7] R. R. Mazumdar H. Yaiche and C. Rosenberg. A game theoretic framework for bandwidth allocation and pricing in broadband networks. *IEEE/ACM Trans. on Networking*, 8(5):667–677, 2000.

[8] K. Kar, S. Sarkar, and L. Tassiulas. Optimization based rate control for multirate multicast sessions. In *IEEE INFOCOM'01*, Anchorage, Alaska, April 2001.

[9] K. Kar, S. Sarkar, and L. Tassiulas. A simple rate control algorithm for maximizing total user utility. In *IEEE INFOCOM'01*, Anchorage, Alaska, April 2001.

[10] F. P. Kelly. Charging and rate control for elastic traffic. *European Trans. on Telecom.*, 8:33–37, 1998.

[11] S. Kunniyur and R. Srikant. A time scale decompositon approach to adaptive ECN marking. In *IEEE INFOCOM'01*, Anchorage, Alaska, April 2001.

[12] L.Massoulié and J. W. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecom. Systems*, 2000.

[13] R. Mazumdar, L. G. Mason, and C. Douligeris. Fairness in network optimal flow control: optimality of product forms. *IEEE Trans. on Comm.*, 39:775–782, 1991.

[14] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. In *SPIE '98, International Symposium on Voice, Video and Data Communications*, 1998.

[15] Scott Shenker. Fundammental design issues for the future internet. *IEEE J. on Selected Areas in Comm.*, 13(7):1176–1188, 1995.

[16] C. Touati, E. Altman, and J. Galtier. On fairness in bandwidth allocation. *INRIA Research Report*, 2001.