43rd IEEE Conference on Decision and Control
December 14-17, 2004
Atlantis, Paradise Island, Bahamas

Minimal Feedback Optimal Algorithms for Traffic Engineering in Computer Networks

Bernardo A. Movsichoff, Constantino Lagoa
Department of Electrical Engineering
The Pennsylvania State University, University Park, PA 16802. USA.

Abstract— **This paper addresses the problem of Traffic Engineering in Computer Networks. More precisely, optimal data rate adaptation laws are provided, which are applicable to networks where both multiple paths are available between any pair of source/destination nodes and multiple Classes of Service are to be provided. In particular, it is shown that the algorithms presented only need a minimal amount of information to achieve the optimal operating point. More precisely, they only require knowledge of whether a path is congested or not. Hence, the control laws provided in this paper require much less feedback than currently available ones. The proposed approach is applicable to utility functions of a very general form and endows the network with the very important property of robustness with respect to node/link failures; i.e., upon the occurrence of such failure, the presented control laws reroute traffic away from the inoperative node/link and converge to the optimal allocation for the "reduced" network.**

I. Introduction

The work presented in this paper is integrated in the line of research of [5] where Sliding Modes ([1], [11]) are used for the development of algorithms for optimal TE. In particular, the approach taken in [5] led to adaptation laws that maximize a given utility function (of a very general form) while providing several Classes of Service (CoSs) and allowing for multiple paths for the calls that share the network. This was accomplished by using binary congestion information fed back from each congested link. However, obtaining an accurate estimate of the number of congested links can result in significant overhead traffic; i.e., it leads to less available resources to the users. Hence, it is desirable to reduce this overhead. However, achieving optimality using reduced feedback laws remained an open problem until now. The objective of this paper is then to provide optimal adaptation laws that, using only *binary information per path*, achieve an optimal operating point as measured by a given utility function.

The results by Massoulié et al. [7] show that the control laws presented in [5] may converge to a point different than the optimal if the true number of congested links is not known. This fact prompted the need to analyze the behavior of our algorithms under the assumption that only a reduced amount of feedback information is available. These laws were found to be quasi-optimal; e.g., see [9] and [8]. In contrast to these previous results, this paper presents *new* control laws that are indeed optimal in such a scenario.

Several other methodologies have been proposed to address the problem of optimal Traffic Engineering (TE); e.g., see [2], [3], [4], [6]. However, to the best of these authors knowledge, this is the first work to propose a comprehensive optimal solution in the multipath multi-CoS case environment, using only *per path* binary congestion information.

The remainder of the paper is organized as follows: Section II presents notation and assumptions used throughout this paper, Section III provides a precise statement of the problem to be solved, while Section IV introduces the proposed optimal solution. Section V on the other hand, discuss some implementation issues while Section VI provides some simulation results. Finally, Section VII provides some conclusions and the Appendix presents the proof of the results in this paper.

II. Preliminaries

Throughout this paper, it is assumed that traffic flows can be described by a fluid flow model, where the only resource taken into account is link bandwidth. Consider a computer network where calls of different *types* are present. In this paper *types* denote aggregate of calls with the same ingress and egress node as well as service requirements; i.e., calls that share a given set of paths connecting the same ingress/egress node pair and whose service requirements are to be satisfied by the aggregate, not by individual calls.

More precisely, consider a computer network whose set of links is denoted by $\mathscr{L}$ and let c_l be the capacity of link $l \in \mathscr{L}$. Let n be the number of types of calls, n_i be the number of paths available for calls of type i and $\mathscr{L}_{i,j}$ be the set of links used by calls of type i taking path j; i.e., if $B_{i,j} = \mathrm{card}(\mathscr{L}_{i,j})$, the cardinality of the set $\mathscr{L}_{i,j}$, then $B_{i,j}$ is the number of links in this path. Given calls of type i, let $x_{i,j}$ be the total data rate of calls of type i using path j. Also, let

$$\mathbf{x}_i \doteq [x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}] \in \mathbf{R}^{n_i}$$

denote the vector containing the data rates allocated to the different paths taken by calls of type i and

$$\mathbf{x} \doteq \left[\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T\right]^T \in \mathbf{R}^N$$

the vector containing all the data rates allocated to different call types and respective paths, where $N = n_1 + n_2 + \cdots + n_n$.

Now, a link $l \in \mathscr{L}$ is said to be congested if the aggregated data rate of the calls using the link reaches its capacity c_l. The congestion information $cg_{i,j}$ for calls $x_{i,j}$; i.e., calls of type i taking path j, is defined as

$$cg_{i,j} \doteq \begin{cases} 1 & \text{if any link } l \in \mathscr{L}_{i,j} \text{ is congested} \\ 0 & \text{otherwise} \end{cases},$$

and $\overline{cg_{i,j}}$ denotes the logical not operation on $cg_{i,j}$.

A. Classes of Service

The service requirements mentioned above, are classified as a Class of Service (CoS). In this paper two distinct CoSs are assumed to be provided: Calls of type $i = 1,2,\dots,s$ are assumed, without loss of generality, to be of Assured Service (AS) category. By AS it is meant that a target rate for the call should be guaranteed in an average sense. More precisely, assuming that the target rate for $\mathbf{x}_i$ is Λ_i, the objective is to allocate the data rates in such a way that

$$\sum_{j=1}^{n_i} x_{i,j} = \Lambda_i,$$

for all $i = 1,2,\dots,s$. Calls of types $i = s+1, s+2, \dots, n$ are assumed to be of BE category; i.e., these calls utilize whatever resources are available after AS is satisfied and do not have any requirements.

Other classes of service can be addressed, such as the ones defined in [5], but due to space constraints only the case of AS and BE is addressed in this paper. The results can be easily extended to these additional service requirements.

III. Problem Statement

Let $U(\mathbf{x})$ be a given utility function representing the desired policy for assigning resources in the network. The results in this paper aim at maximizing utility functions the form

$$U(\mathbf{x}) \doteq \alpha \sum_{i=1}^{n} U_i(\mathbf{x}_i) \doteq \alpha \sum_{i=1}^{n} U_i\big(x_{i,1}, x_{i,2}, \dots, x_{i,n_i}\big),$$

where $U_i(\cdot)$, $i = 1,2,\dots,n$, are differentiable concave functions, strictly increasing in each of their arguments, and α is a positive scaling constant.

Given this, the problem of optimal Traffic Engineering can be precisely stated as the following optimization problem:

$$\max_{\mathbf{x}} U(\mathbf{x})$$

subject to the network capacity constraints

$$\sum_{i,j:\, l \in \mathscr{L}_{i,j}} x_{i,j} - c_l \le 0 \qquad l \in \mathscr{L},$$

AS requirements

$$\sum_{j=1}^{n_i} x_{i,j} - \Lambda_i = 0; \qquad i = 1,2,\dots,s,$$

and non-negativity of all the data rates; i.e., $x_{i,j} \ge 0$, for all i and all j.

Clearly this is a convex optimization problem. However, the solution is not trivial if one imposes decentralized algorithms and restricts knowledge on the satisfaction of the capacity constraints.

IV. Family of Decentralized Control Laws

Before presenting the main results in this paper, this section introduces the proposed solution to the optimization problem above, a family of control laws that achieve optimal rate allocation.

Let $f_{i,j}$ be defined as

$$f_{i,j}(\mathbf{x}) \doteq \big(1 - e^{-\partial U/\partial x_{i,j}}\big),$$

and let $z_{i,j}(t,\mathbf{x})$ be positive scalar functions for all i and all j. Now, define the following family of control laws: For $i = 1,2,\dots,s$; i.e., AS calls, let

$$\dot{x}_{i,j} = z_{i,j}(t,\mathbf{x})\Big[f_{i,j}(\mathbf{x}) - \big(1 - \overline{cg_{i,j}}\, r_i\big)\Big],$$

where

$$r_i(\mathbf{x}_i) = \begin{cases} r_{min} < 1 & \text{if } \sum_{j=1}^{n_i} x_{i,j} > \Lambda_i \\ r_{max} > 1 & \text{if } \sum_{j=1}^{n_i} x_{i,j} < \Lambda_i, \end{cases}$$

with r_{min} and r_{max} predetermined positive constants. Finally, for $i = s+1, s+2, \dots, n$; i.e., BE calls, let

$$\dot{x}_{i,j} = z_{i,j}(t,\mathbf{x})\Big[f_{i,j}(\mathbf{x}) - \big(1 - \overline{cg_{i,j}}\big)\Big].$$

These expressions for the derivative $\dot{x}_{i,j}$ do not include any term that will prevent the data rates to become negative. Therefore, if at any time, $x_{i,j} = 0$ and $\dot{x}_{i,j} < 0$, the derivative is set to zero; i.e., $\dot{x}_{i,j} = 0$. Otherwise, the expressions above are used as they stand. For notational convenience they will be left in this way.

Note: The proof presented in the Appendix is applied to the control laws as written above. However, by forcing the derivative to zero as explained, one obtains the same behavior as the one would obtain by adding extra parameters in the control laws to deal with non-negativity constrains. The reasoning is identical to the one in [9] and [10].

A. Main Result

The main result of this paper establishes that the control laws presented above, converge to the solution of the optimization problem posed. This is formally stated in the following theorem.

Theorem 1: Assume that all data rates are bounded; i.e., there exists $\rho \in \mathbf{R}$ such that the data rate vector $\mathbf{x}$ always belongs to the set

$$\mathscr{X} \doteq \{\mathbf{x} \in \mathbf{R}^{n_1+n_2+\dots+n_n} : \; x_{i,j} \le \rho,\; l \in \mathscr{L}_{i,j},\; j = 1,2,\dots,n_i,\; i = 1,2,\dots,n\}.$$

Also, assume that at the optimal traffic allocation, each congested link has at least one BE call with non-zero data rate and that the elements of the gradient of the utility function

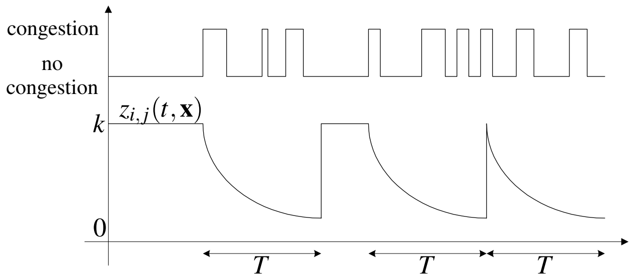


Fig. 1. Example of a scaling function $z_{i,j}(\cdot)$

are bounded in $\mathscr{X}$. Let $\zeta > 0$ be a given (arbitrarily small) constant and let $z_{i,j}(t,\mathbf{x})$ be scalar continuous functions satisfying

$$z_{i,j}(t,\mathbf{x}) > \zeta$$

for all $t > 0$ and all $\mathbf{x} \in \mathscr{X}$. Furthermore, let

$$0 < r_{min} < r_{lower} < 1 < r_{upper} < r_{max},$$

where

$$r_{lower} = e^{-v_{k,\max}}, \quad r_{upper} = e^{v_{k,\max}}$$

and

$$v_{k,\max} = \max_{i,j} B_{i,j} \max_{i,j,\mathbf{x} \in \mathscr{X}} \frac{\partial U}{\partial x_{i,j}}.$$

The quantity $B_{i,j}$, as defined in Section II, is the number of links in path j taken by calls of type i.

Then, the control laws presented above converge to a traffic allocation that maximizes the utility function $U(\mathbf{x})$ subject to the network's capacity constraints, AS requirements and non-negativity of all the data rates.

V. Implementation Considerations

The results presented in Section IV provide a large set of control laws. In this section some of the issues related to a practical implementation are discussed.

A. Discrete-time Control Laws

In a real network, the control laws have to be implemented in discrete-time. This is accomplished in this paper by means of a backward rule approximation: Let

$$\dot{x}_{i,j} = g_{i,j}(\mathbf{x},t)$$

denote the continuous time laws derived in Section IV. Then the discrete-time counterpart is given by

$$x^d_{i,j}\left[(k+1)t_d\right] = x^d[kt_d] + t_d g_{i,j}\big(\mathbf{x}(kt_d), kt_d\big);\ k = 0,1,\ldots,$$

where t_d is the integration interval. Now, Sliding Mode theory does not apply to this approximation. However, some of the results in [1] can be used to show that by a suitable choice of the parameters of the control laws one can be arbitrarily close to the continuous time trajectory. In particular, given $\delta > 0$ an arbitrarily small constant, t_d and $z_{i,j}$ should be chosen to satisfy

$$t_d z_{i,j}(t,\mathbf{x}) < \delta.$$

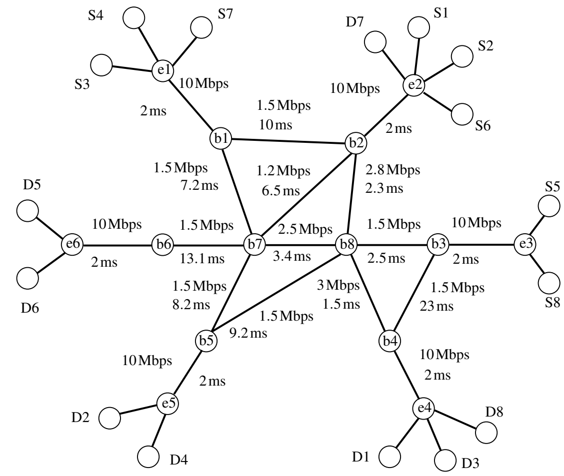


Fig. 2. Topology of the network

B. Adaptive Oscillation Reduction

Several non-ideal conditions that exist in a real network lead to the undesirable effect of oscillation. In particular, the existence of delays in the propagation of information packets as well as congestion information, together with discretization and quantization can lead to quite large oscillation in the data rates.

To mitigate this undesired phenomenon, the scheme originally presented in [5] is proposed. Let $T > 0$ and $k > 0$ be given. Set $z_{i,j}(0,\mathbf{x}) = k$. If congestion is detected at time $t = t_0$ for calls of type i taking path j let

$$z_{i,j}(t,\mathbf{x}) = \omega(t - t_0); \quad \text{for } t_0 \le t < t_0 + T,$$

where $\omega : [0,T] \to [\delta, k]$ is a decreasing function of time and δ is a small positive constant. Now, simply repeat this procedure ad infinitum. The desired behavior is depicted in Fig. 1.

Proceeding in this way, ensures that the network can adapt to changes in the demand while diminishing the speed of change when congestion is detected. Note that oscillation occurs due to non-ideal switching around the sliding surfaces and this occurs only if there is congestion.

VI. Simulation Examples

In this section simulation examples are presented, that help in the understanding of the behavior of the proposed control laws. In particular, is is shown that the control laws converge to the optimal traffic allocation while satisfying service requirements and that they provide an optimal way of reacting to link failures. These examples implement a discrete-time version of the control laws and use a flow approximation for the calls.

A. Simulation Setup

The model of the network used for these examples is the same in [5] which was originally used by La, et al.([4]). The topology is shown in Fig. 2 along with all link capacities

TABLE I

PATHS AVAILABLE FOR EACH TYPE OF CALLS

type 1 $n_1 = 4$	$x_{1,1} : e_2b_2b_8b_4e_4$ $x_{1,2} : e_2b_2b_8b_3b_4e_4$ $x_{1,3} : e_2b_2b_7b_8b_3b_4e_4$ $x_{1,4} : e_2b_2b_7b_8b_4e_4$	type 5 $n_5 = 2$	$x_{5,1} : e_3b_3b_8b_7b_6e_6$ $x_{5,2} : e_3b_3b_4b_8b_5b_7b_6e_6$
type 2 $n_2 = 3$	$x_{2,1} : e_2b_2b_8b_5e_5$ $x_{2,2} : e_2b_2b_7b_5e_5$ $x_{2,3} : e_2b_2b_1b_7b_5e_5$	type 6 $n_6 = 3$	$x_{6,1} : e_2b_2b_1b_7b_6e_6$ $x_{6,2} : e_2b_2b_8b_7b_6e_6$ $x_{6,3} : e_2b_2b_7b_6e_6$
type 3 $n_3 = 2$	$x_{3,1} : e_1b_1b_7b_8b_4e_4$ $x_{3,2} : e_1b_1b_2b_8b_4e_4$	type 7 $n_7 = 3$	$x_{7,1} : e_1b_1b_2e_2$ $x_{7,2} : e_1b_1b_7b_2e_2$ $x_{7,3} : e_1b_1b_7b_8b_2e_2$
type 4 $n_4 = 4$	$x_{4,1} : e_1b_1b_7b_5e_5$ $x_{4,2} : e_1b_1b_7b_8b_5e_5$ $x_{4,3} : e_1b_1b_2b_7b_5e_5$ $x_{4,4} : e_1b_1b_2b_8b_5e_5$	type 8 $n_8 = 2$	$x_{8,1} : e_3b_3b_4e_4$ $x_{8,2} : e_3b_3b_8b_4e_4$

and assumed delays. There are overall $n = 8$ types of calls as given by the source/destination pairs indicated in the figure. The paths available for each one of these calls are indicated in Table I, where n_i is the number of paths available for each type of calls.

Utilization will be measured by the utility function

$$U(\mathbf{x}) = \sum_{i=1}^{8} 0.1 \log\left(0.5 + \sum_{j=1}^{n_i} x_{i,j}\right),$$

where n_i is again indicated in the table. The term 0.5 is included to avoid an infinite derivative at 0 data rate. For the AS service requirements, calls of types $i = 3$ and $i = 5$ are assumed to have target rates $\Lambda_3 = \Lambda_5 = 1\,\text{Mbps}$.

Given this, the control laws presented in Section IV become: For $i = 1,3$ and $j = 1,2$; i.e., AS calls

$$\dot{x}_{i,j} = z_{i,j}(t,\mathbf{x})\left[\left(1 - e^{-0.1\left(\sum_{j=1}^{n_i} x_{i,j}+0.5\right)^{-1}}\right) - \left(1 - \overline{cg_{i,j}} r_i\right)\right]$$

and for $i = 1,2,4,6,7,8$ and $j = 1,\ldots,n_i$; i.e., BE calls

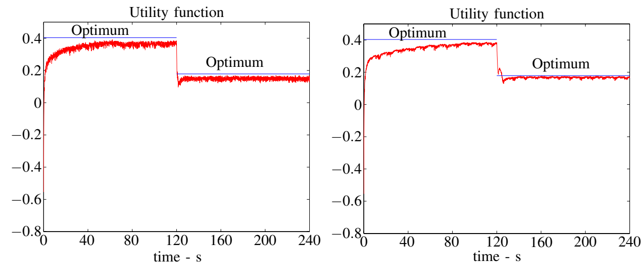
$$\dot{x}_{i,j} = z_{i,j}(t,\mathbf{x})\left[\left(1 - e^{-0.1\left(\sum_{j=1}^{n_i} x_{i,j}+0.5\right)^{-1}}\right) - \left(1 - \overline{cg_{i,j}}\right)\right],$$

where r_i was chosen with a margin of ± 0.001 on the bounds set forth in Theorem 1. On the other hand, the oscillation reducing functions $z_{i,j}$ were taken as $z_{i,j}(t) = \omega(t - T_0)$, where $T_0 = 10\,\text{s}$ and

$$\omega(t) = 1.8(0.25 + 0.65^t).$$

However, the first simulations were conducted with a constant $z_{i,j}$. These simulations showed the need for the oscillation reduction scheme as can be easily be seen by comparing Fig. 3(a) (constant $z_{i,j} = \omega(0) = 2.25$) and Fig. 3(b) (time varying $z_{i,j}$). These plots show that the control laws converge to the optimum as expected. Moreover, the oscillation reduction scheme provides better behavior in terms of oscillation and proximity to the optimal utilization. Finally, the integration step was chosen as $t_d = 5\,\text{ms}$.

Note: Non-ideal implementation such as delays and discretization (that were not considered in Section IV) prevent the algorithm from reaching the exact optimum. Instead, convergence to a small neighborhood is achieved. This neighborhood is seen to be smaller for the laws with a time-varying $z_{i,j}$.



(a) Constant $z_{i,j}$ (b) Time-varying $z_{i,j}$

Fig. 3. Utility function

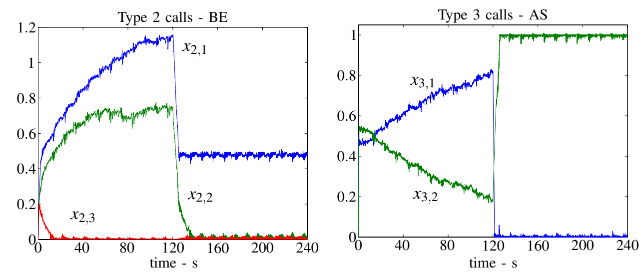


Fig. 4. Example of data rates

Other factor that has a large impact on the magnitude of the oscillation is the value of r_i. Since this value is directly related to the gradient of the utility function, it becomes apparent that the factor of 0.1 in the utility function contributes to reducing oscillation.

Figure 4 show some representative examples of data rates. Calls of type $i = 3$ of the AS type are seen to satisfy the imposed target rate. Calls of type $i = 5$ have a similar response and is not included due to space constraints. Also shown in the figure are BE calls of type $i = 2$.

B. Robustness Against Link Failures

The control laws presented in this paper excel at re-routing traffic upon a failure in a node or link. In order to show this feature, the link connecting nodes b_7 and b_8 was opened at time $T = 120\,\text{s}$. The behavior of the control laws can be seen in Figs. 3 and 4 from time 120 s on. Note from Table I that both AS calls lose one of the two paths they have available so it can be considered to be an extreme situation. In this scenario, for example calls of type $i = 2$ have to "kill" all traffic on one of the available paths and greatly reduce another. Also, note that the control laws implemented at the edge nodes are oblivious to the failure. They simply react to what they perceive as congestion.

VII. CONCLUSION

In this paper a novel solution to the problem of optimal Traffic Engineering was presented. This solution encompasses networks were both multiple paths between each pair of source/destination nodes are available and provides multiple CoS. Moreover, they require minimal information on the network status and, hence, only a very small amount of overhead is needed to achieve optimality. These laws also provide a way of optimally tackling failures in the network.

Effort is now being put in the implementation of the control laws presented in this paper. In particular, these laws have several parameters for which only bounds are provided. Hence, criteria is now being developed for the determination of these "free parameters".

ACKNOWLEDGMENTS

This research was made possible thanks to funding provided by the National Science Foundation under grants ECS-9984260 and ANI-0125653.

REFERENCES

[1] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, ser. Mathematics and Its Applications. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1988.

[2] S. J. Golestani and S. Bhattacharyya, "A class of end-to-end congestion control algorithms for the Internet," in *Proc. Int. Conf. Network Protocols, ICNP*, Austin, TX, USA, Oct. 1998, pp. 137–150.

[3] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar. 1998.

[4] R. J. La and V. Anantharam, "Charge-sensitive TCP and rate control in the Internet," in *Proc. IEEE INFOCOM 2000*, vol. 3, Tel Aviv, Israel, Mar. 2000, pp. 1166–1175.

[5] C. M. Lagoa, H. Che, and B. A. Movsichoff, "Adaptive control algorithms for decentralized optimal traffic engineering in the Internet," *IEEE/ACM Trans. Networking*, vol. 12, no. 3, pp. 415–428, June 2004.

[6] S. H. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," *IEEE/ACM Trans. Networking*, vol. 7, no. 6, pp. 861–874, Dec. 1999.

[7] L. Massoulié and J. Roberts, "Bandwidth sharing: Objectives and algorithms," in *Proc. IEEE INFOCOM '99*, vol. 3, New York, NY, USA, Mar. 1999, pp. 1395–1403.

[8] B. A. Movsichoff, C. M. Lagoa, and H. Che, "A sliding mode approach to traffic engineering in computer networks," in *Advances in Communication Control Networks*, ser. Lecture Notes in Control and Information Sciences (LCNCIS), C. T. Abdallah, J. N. Chiasson, and S. Tarbouriech, Eds. Spriger-Verlag, 2004, to be published.

[9] B. A. Movsichoff and C. M. Lagoa, "Near-optimal decentralized algorithms for traffic engineering in the Internet," 2003, submitted to Automatica.

[10] ——, "A novel approach to traffic engineering for ad-hoc networks," in *Proc. 42nd IEEE CDC*, vol. 1, Maui, HI, USA, Dec. 2003, pp. 576–581.

[11] V. I. Utkin, *Sliding Modes in Control and Optimization*, ser. Communications and Control Engineering Series. Berlin, Heidelberg: Spriger-Verlag, 1992, vol. 66, no. 1.

APPENDIX
PROOF OF MAIN RESULTS

In this appendix, the proof of the Theorem 1 is presented. We set the stage by introducing some additional notation. Due to space constraints, only the main steps of the proof are presented.

To simplify the exposition to follow, let the problem at hand be recast in the following form

$$\max_{\mathbf{x}} U(\mathbf{x})$$

subject to inequality (capacity and non-negativity) constraints

$$h_k(\mathbf{x}) \leq 0; \quad k = 1,2,\ldots,m$$

and the equality (AS) constraints

$$h_k(\mathbf{x}) = 0 \quad k = m+1, m+2, \ldots, L.$$

Now, let the admissible domain be defined as the set

$$\mathscr{C} = \left\{ \mathbf{x} \in \mathbf{R}^N : h_k(\mathbf{x}) \leq 0 \text{ for } k = 1,2,\ldots,m \right\};$$

i.e., the set of data rates that can be admitted by the network without any further constraints. Also, let the feasible set be defined as

$$\mathscr{D} = \left\{ \mathbf{x} \in \mathscr{C} : h_k(\mathbf{x}) = 0 \text{ for } k = m+1, m+2, \ldots, L \right\};$$

i.e., the set of data rates satisfying all the constraints of the optimization problem. The proof follows by first observing that convergence to the admissible set is obtained in finite time. Then, once inside $\mathscr{C}$ the adaptation laws can be shown to evolve towards the optimum.

Lemma 1: Let r_i satisfy the conditions set forth in Theorem 1. Then vector $\mathbf{x}$ converges to the admissible domain $\mathscr{C}$ in finite time.

Proof: Let $x_{i,j} \geq 0$, for any given i and j, such that $\mathbf{x} \notin \mathscr{C}$ and let $\varepsilon > 0$ be an arbitrarily small constant. By construction of the control laws it holds that $\dot{x}_{i,j} \leq -\varepsilon < 0$. Therefore, since the derivative is strictly negative outside $\mathscr{C}$, $x_{i,j}$ reaches the admissible region $\mathscr{C}$ in finite time. ■

The following Lemma, central to the proof of the results in this paper, provides an alternative representation of the proposed control laws.

Lemma 2: For all $\mathbf{x} \in \mathscr{C}$, the control laws above can be expressed as

$$\dot{\mathbf{x}} = \mathbf{Z}(\mathbf{x},t)\left[\nabla U(\mathbf{x}) - \mathbf{H}(\mathbf{x})\mathbf{v}(\mathbf{x})\right],$$

where $\mathbf{Z}(\mathbf{x},t)$ is a positive definite matrix and

$$\mathbf{H}(\mathbf{x}) = \left[\nabla h_1(\mathbf{x}), \nabla h_2(\mathbf{x}), \ldots, \nabla h_L(\mathbf{x})\right].$$

Proof: The control laws presented in Section IV can be formulated as follows: Let $\mathscr{I}_{i,j}$ be the set of indices $k \in \{1,2,\ldots,m\}$ such that the capacity constraints $h_k(\mathbf{x})$ involve the data rate $x_{i,j}$. Also, let $\mathscr{I}_i^{\text{as}}$ be the set of indices $k \in \{m+1, m+2, \ldots, L\}$ such that the constraints $h_k(\mathbf{x})$, $k \in \mathscr{I}_i^{\text{as}}$ impose AS requirements on the data rate $x_{i,j}$. Note that this set is empty if calls of type i are of the BE class.

Then,

$$\dot{x}_{i,j} = z_{i,j}\left[f_{i,j}(\mathbf{x}) - \left(1 - \prod_{k \in \mathscr{I}_{i,j} \cup \mathscr{I}_i^{\text{as}}} u_k\right)\right],$$

where the quantities u_k are defined as follows: For $k \in \mathscr{I}_i^{\text{as}}$, $i = 1,2,\ldots,s$; i.e., for AS constraints, let $u_k \doteq r_i$. For $k \in \mathscr{I}_{i,j}$, $i = s+1, s+2, \ldots, n$; i.e., for capacity constraints, let $u_k \doteq \overline{cg_{i,j}}$.

Now, since $f_{i,j}(\mathbf{x}_i) > 0$, when $\mathbf{x} \in \mathscr{C}$ either $\mathbf{x}$ is an inner point of $\mathscr{C}$ where by definition $u_k = 1$ or a sliding mode occurs on some surface $s(\mathbf{x}) = 0$, where $\mathbf{x} \in \partial\mathscr{C}$ (the boundary of $\mathscr{C}$). In the latter case, using the equivalent control method (for a definition when the differential inclusion(equation) is not an affine function of the equivalent control, see [1], [11]) there exists $u_{k,\mathrm{eq}}$, such that

$$\dot{x}_{i,j}(t) = z_{i,j}(\mathbf{x},t)\left[-\big(1-f_{i,j}(\mathbf{x})\big) + \prod_{k\in\mathscr{I}_{i,j}\cup\mathscr{I}_{i,j}^{m}} u_{k,\mathrm{eq}}\right].$$

Moreover, since $\max_{\mathbf{x}\in\mathscr{C}} f_{i,j}(\mathbf{x}_i) = \mu < 1$, then there exists a constant $\chi > 0$ such that

$$\chi < u_{k,\mathrm{eq}} < 1, \quad \forall \mathbf{x}\in\mathscr{C}.$$

Hence, given that the log function has a bounded derivative in the interval $\left[\min\big(1-\mu,\chi^{B_{i,j}}\big),1\right]$, the evolution of $x_{i,j}$ can be represented as

$$\begin{aligned}\dot{x}_{i,j} &= \hat{z}_{i,j}(\mathbf{x},t)\left[-\log\big(1-f_{i,j}(\mathbf{y})\big)+\log\prod_{k\in\mathscr{I}_{i,j}\cup\mathscr{I}_{i}^{\mathrm{as}}} u_{k,\mathrm{eq}}\right]\\ &= \hat{z}_{i,j}(\mathbf{x},t)\left[\log\frac{1}{1-f_{i,j}(\mathbf{y})} - \sum_{k\in\mathscr{I}_{i,j}\cup\mathscr{I}_{i}^{\mathrm{as}}}\log\frac{1}{u_{k,\mathrm{eq}}}\right]\\ &= \hat{z}_{i,j}(\mathbf{x},t)\left[\frac{\partial U}{\partial x_{i,j}} - \sum_{k\in\mathscr{I}_{i,j}\cup\mathscr{I}_{i,j}^{\mathrm{as}}}\log\frac{1}{u_{k,\mathrm{eq}}}\right],\end{aligned}$$

where $\hat{z}_{i,j}(\mathbf{x},t) \geq \hat{\mu} > 0$.

Hence,

$$\dot{\mathbf{x}} = \mathbf{Z}(\mathbf{x})\big[\nabla U(\mathbf{x}) - \mathbf{H}(\mathbf{x})\mathbf{v}(\mathbf{x})\big],$$

where

$$\mathbf{v}(\mathbf{x}) = [\log(1/u_{1,\mathrm{eq}}), \log(1/u_{2,\mathrm{eq}}), \ldots, \log(1/u_{L,\mathrm{eq}})]^T$$

and $\mathbf{Z}(\mathbf{x},t)$ is a positive definite diagonal matrix with elements $\hat{z}_{i,j}$; i.e,

$$\mathbf{Z} \doteq \mathrm{diag}\big[\hat{z}_{1,1},\ldots,\hat{z}_{1,n_1},\ldots,\hat{z}_{n,1},\ldots,\hat{z}_{n,n_n}\big].$$

■

Now, define the auxiliary function

$$\widehat{U}(\mathbf{x}) = U(\mathbf{x}) - \Xi(\mathbf{x}),$$

where

$$\Xi(\mathbf{x}) = \big[h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_L(\mathbf{x})\big]\mathbf{v}(\mathbf{x}).$$

The proof relies on the following three results whose proof is very similar to the ones in [11, Chapter 15, Sections 3 and 4]. Hence, they are presented here without proof.

Lemma 3: If the set of all maximum points is bounded (which is our case) then $\mathbf{x}$ will converge to this set from any initial condition.

Theorem 2: Let $\mathbf{v}^0$ be a vector whose entries are of the form

$$\begin{aligned} 0 \leq v_k \leq \gamma_k\,; &\quad k = 1,2,\ldots,m\\ -\xi_k \leq v_k \leq \xi_k\,; &\quad k = m+1, m+2,\ldots,L,\end{aligned}$$

where $v_k = 0$ for non-binding constraints. Then, the maximum of $\widehat{U}(\mathbf{x})$ coincides with the optimal $U(\mathbf{x}^*)$ if and only if there exists $\mathbf{x}^*$ such that

$$\nabla U(\mathbf{x}^*) = \mathbf{H}(\mathbf{x}^*)\mathbf{v}^0.$$

Theorem 3: The control laws presented above converge to the set of maximum points of the utility function $U(\mathbf{x})$ if this set is bounded, the condition of Theorem 2 is satisfied and vector $\mathbf{v}^0$ is an inner point of the set defined in Theorem 2, except for the non-binding constraints.

In order for these results to hold, it is necessary to show several properties of the proposed control laws. These properties can be formally stated with the following lemmas.

Lemma 4: The function $\widehat{U}(\mathbf{x})$, for $\mathbf{x}\in\mathscr{C}$ does not decrease along the trajectories.

Proof: If a sliding mode does not occur then

$$\begin{aligned}\frac{d\widehat{U}}{dt} &= \big[\nabla U - \mathbf{H}(\mathbf{x})\mathbf{v}(\mathbf{x})\big]^T\dot{x}\\ &= \big[\nabla U - \mathbf{H}(\mathbf{x})\mathbf{v}(\mathbf{x})\big]^T\mathbf{Z}(t,\mathbf{x})\big[\nabla U - \mathbf{H}(\mathbf{x})\mathbf{v}(\mathbf{x})\big] \geq 0\end{aligned}$$

since matrix $\mathbf{Z}(t,\mathbf{x})$ is positive definite.

Now, assume that a sliding mode occurs in the intersection of the surfaces $h_k(\mathbf{x}) = 0$, $k\in\mathscr{I}$. Let $\mathbf{H}_1(\mathbf{x})$ be the matrix whose columns are $\nabla h_k(\mathbf{x})$ for $k\in\mathscr{I}$ (and in the same order as in $\mathbf{H}(\mathbf{x})$). Also, let $\mathbf{H}_2(\mathbf{x})$ be the matrix with columns $\nabla h_k(\mathbf{x})$ for $k\notin\mathscr{I}$ (again in the same order as in $\mathbf{H}(\mathbf{x})$). Then, given that a sliding mode occurs in the intersection of the surfaces $h_k(\mathbf{x}) = 0$, $k\in\mathscr{I}$, we have

$$\mathbf{H}_1(\mathbf{x})^T\mathbf{Z}(t,\mathbf{x})\big[\nabla U - \mathbf{H}_1(\mathbf{x})\mathbf{v}_1(\mathbf{x}) - \mathbf{H}_2(\mathbf{x})\mathbf{v}_2(\mathbf{x})\big] = 0$$

where $\mathbf{v}_1(\mathbf{x})$ is the vector containing $v_k(\mathbf{x})$, for $k\in\mathscr{I}$ and $\mathbf{v}_2(\mathbf{x})$ is the vector containing $v_k(\mathbf{x})$, for $k\notin\mathscr{I}$. Now, assume that $\det\big[\mathbf{H}_1(\mathbf{x})^T\mathbf{Z}(t,\mathbf{x})\mathbf{H}_1(\mathbf{x})\big] \neq 0$ (a reasoning similar to the one in [11] can be done to address the case where this does not happen). From now on, to simplify the exposition, we drop the dependency on $\mathbf{x}$. Then, the equivalent control is

$$\mathbf{v}_{1,\mathrm{eq}} = \big(\mathbf{H}_1^T\mathbf{Z}\mathbf{H}_1\big)^{-1}\big(\mathbf{H}_1^T\mathbf{Z}\nabla U - \mathbf{H}_1^T\mathbf{Z}\mathbf{H}_2\mathbf{v}_2\big)$$

and the sliding motion that results from using the equivalent control in $\mathbf{x}$ is

$$\dot{\mathbf{x}} = \sqrt{\mathbf{Z}}\mathbf{P}\sqrt{\mathbf{Z}}\big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big),$$

where $\sqrt{\mathbf{Z}}$ is well defined since $\mathbf{Z}$ is a diagonal and positive definite matrix and

$$\mathbf{P} \doteq I - \sqrt{\mathbf{Z}}\mathbf{H}_1\big(\mathbf{H}_1^T\sqrt{\mathbf{Z}}\sqrt{\mathbf{Z}}\mathbf{H}_1\big)^{-1}\mathbf{H}_1^T\sqrt{\mathbf{Z}}.$$

Now, let Ξ_1 be the elements h_k of Ξ with $k\in\mathscr{I}$ (in the same order as in Ξ). Also, let Ξ_2 be the elements h_k of Ξ with $k\notin\mathscr{I}$ (again in the same order as in Ξ). Since a sliding mode occurs, during this motion we have

$$\widehat{U} = U - \Xi_2\mathbf{v}_2.$$

Now, since U and Ξ_2 are continuously differentiable and along this sliding motion $\mathbf{v}_2$ is constant, we have

$$\frac{d\widehat{U}}{dt} = \big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big)^T\dot{\mathbf{x}}.$$

Now, notice that $\mathbf{P} = \mathbf{P}^T = \mathbf{P}^2$. Hence,

$$\frac{d\widehat{U}}{dt} = \big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big)^T \sqrt{\mathbf{Z}}\mathbf{P}\sqrt{\mathbf{Z}}\big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big) = \left\|\mathbf{P}\sqrt{\mathbf{Z}}\big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big)\right\|^2 \geq 0.$$

■

Lemma 5: The time derivative of $\widehat{U}(\mathbf{x})$, for $\mathbf{x} \in \mathscr{C}$, is zero only when $\dot{\mathbf{x}} = 0$.

Proof: If a sliding mode does not occur we have

$$\frac{d\widehat{U}}{dt} = \big[\nabla U - \mathbf{H}\mathbf{v}\big]^T \mathbf{Z}\big[\nabla U - \mathbf{H}\mathbf{v}\big]$$

and since $\mathbf{Z}$ is positive definite

$$\frac{d\widehat{U}}{dt} = 0 \Rightarrow \nabla U - \mathbf{H}\mathbf{v} = 0 \Rightarrow \mathbf{Z}\big[\nabla U - \mathbf{H}\mathbf{v}\big] = 0 \Rightarrow \dot{\mathbf{x}} = 0.$$

Now assume that a sliding mode occurs in the intersection of the surfaces $h_k(\mathbf{x}) = 0$, $k \in \mathscr{I}$. In this case,

$$\frac{d\widehat{U}}{dt} = \left\|\mathbf{P}\sqrt{\mathbf{Z}}\big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big)\right\|^2.$$

Hence,

$$\frac{d\widehat{U}}{dt} = 0 \Rightarrow \mathbf{P}\sqrt{\mathbf{Z}}\big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big) = 0 \Rightarrow \sqrt{\mathbf{Z}}\mathbf{P}\sqrt{\mathbf{Z}}\big(\nabla U - \mathbf{H}_2\mathbf{v}_2\big) = 0 \Rightarrow \dot{\mathbf{x}} = 0.$$

■

Lemma 6: The stationary points of $\widehat{U}$ are the maximum points of $\widehat{U}$.

Proof: Let $\mathbf{x}_0$ be a stationary point on the intersection of surfaces given by $\mathbf{H}_1 = 0$. In this case, we have

$$\nabla U(\mathbf{x}_0) - \mathbf{H}_1(\mathbf{x}_0)\mathbf{v}_{1,\text{eq}}(\mathbf{x}_0) - \mathbf{H}_2(\mathbf{x}_0)\mathbf{v}_2(\mathbf{x}_0) = 0.$$

Now, consider the function

$$\widehat{U}^*(\mathbf{x}) = U(\mathbf{x}) - \mathbf{H}_1(\mathbf{x})\mathbf{v}_{1,\text{eq}}(\mathbf{x}_0) - \mathbf{H}_2(\mathbf{x})\mathbf{v}_2(\mathbf{x}).$$

Given that the components of the equivalent control belong to the convex hull of the set of possible vectors $\mathbf{v}$,

$$\mathbf{H}_1(\mathbf{x})\mathbf{v}_{1,\text{eq}}(\mathbf{x}_0) \leq \mathbf{H}_1(\mathbf{x})\mathbf{v}_1(\mathbf{x})$$

and, as a consequence

$$\widehat{U}^*(\mathbf{x}) \geq \widehat{U}(\mathbf{x}).$$

Now, since $\widehat{U}$ is a concave function, h_k are convex functions for $1 \leq k \leq m$ and h_k are linear functions for $m+1 \leq k \leq L$ then $\widehat{U}^*$ is a concave function and hence it has a unique maximum. Moreover, $\widehat{U}^*$ is continuously differentiable and

$$\nabla\widehat{U}^*(\mathbf{x}_0) = 0.$$

Therefore

$$\max_{\mathbf{x}} \widehat{U}^*(\mathbf{x}) = \widehat{U}^*(\mathbf{x}_0).$$

Now, since $\widehat{U}^*(\mathbf{x}) \geq \widehat{U}(\mathbf{x})$ and $\widehat{U}^*(\mathbf{x}_0) = \widehat{U}(\mathbf{x}_0)$, we conclude that $\widehat{U}(\mathbf{x})$ reaches its maximum at $\mathbf{x}_0$. Therefore, any stationary point of the optimization procedure is a maximum of $\widehat{U}(\mathbf{x})$. Now, assume that a maximum point $\mathbf{x}^*$ of $\widehat{U}(\mathbf{x})$ is not a stationary point. Then, we have

$$\frac{d\widehat{U}(\mathbf{x}^*)}{dt} > 0$$

and so $\widehat{U}$ will increase along the trajectory which contradicts the fact that $\mathbf{x}^*$ is a maximum point of $\widehat{U}(\mathbf{x})$. ■

We are now finally ready to address the proof of the main result in this paper.

A. Proof of Theorem 1

The definition of $f_{i,j}$ together with the conditions on r_{lower} and r_{upper} imply that the Lagrange multipliers at the KKT point of the optimization problem at hand lie in the convex hull generated by the set of all possible $\mathbf{v}$. Indeed, if each congested link is traversed by a BE call, then in the KKT conditions at the optimum $\mathbf{x}^*$

$$\nabla U(\mathbf{x}^*) = \mathbf{H}(\mathbf{x}^*)\mathbf{v}^0$$

the components of $\mathbf{v}^0$ associated with capacity constraints; i.e., v_k^0 for $k = 1,2,\ldots,\text{card}\,(\mathscr{L})$, appear in a set of equations decoupled from the remaining components of $\mathbf{v}^0$. Then, the worst case (larger) value of v_k^0, $k = 1,2,\ldots,\text{card}\,(\mathscr{L})$ is

$$v_{k,max}^0 = \max_{i,j,\mathbf{x}\in\mathscr{X}} \frac{\partial U(\mathbf{x})}{\partial x_{i,j}}$$

Now, using this information in the remaining equations, it is possible to solve for $v_{k,max}^0$, $k = m+1, m+2, \ldots, L$. Since $U(\mathbf{x})$ is an increasing function in all its arguments $x_{i,j}$, the largest absolute value of v_k^0 associated with CoS constraints is given by

$$v_{k,max}^0 = \sum_{\kappa\in\mathscr{K}} v_{\kappa,max}^0 = \max_{i,j} B_{i,j} \max_{i,j,\mathbf{x}\in\mathscr{X}} \frac{\partial U(\mathbf{x})}{\partial x_{i,j}},$$

where

$$\mathscr{K} \doteq \Big\{\kappa\colon\ \kappa \in \mathscr{I}_{i,j^*};\ j^* = \arg\max_{j=1,2,\ldots,n_i} \text{card}(\mathscr{I}_{i,j});\ i\colon\ k \in \mathscr{I}_i^{\text{as}}\Big\}.$$

Hence, it should hold that

$$v_k \leq v_{k,max}^0 < \gamma_k; \qquad k = 1,2,\ldots,m$$
$$|v_k| \leq \big|v_{k,max}^0\big| < \xi_k; \qquad k = m+1,\ldots,L.$$

That is: For capacity constraints, $k = 1,2,\ldots,m$,

$$u_k < e^{-v_{k,max}};$$

and for AS constraints, $k = m+1, m+2, \ldots, L$,

$$u_k < e^{-v_{k,max}} \quad \text{and} \quad u_k > e^{v_{k,max}}.$$

For the capacity constraints the condition is trivially satisfied with $u_k = 0$, while for AS constraints, these are the conditions imposed on r_{lower} and r_{upper}.

Therefore the Lemma 3 and Theorems 2 and 3 hold. Hence, the family of adaptation laws proposed in this paper converge to the maximum of the utility function $U(\mathbf{x})$ subject to $\mathbf{x} \in \mathscr{D}$. In other words, they converge to the optimum of the optimization problem at hand. ■