How (Un)Fair are the ABR Binary Schemes, Actually?

MILAN VOJNOVIĆ*AND JEAN-YVES LE BOUDEC
Institute for Computer Communications and Applications (ICA)
Swiss Federal Institute of Technology at Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{Milan.Vojnovic,Jean-Yves.Leboudec}@epfl.ch

Abstract – It is well known that a simple binary feedback rate–based congestion avoidance scheme cannot ensure a fairness goal of the Available Bit Rate (ABR) service, namely, max–min fairness. In this paper we show how the rates are distributed for the network consisting of the binary switches, and end–systems employing an additive–increase/multiplicative decrease rate control. The modeling assumptions fairly resembles the ABR congestion avoidance, and applies to an arbitrary network topology. The results are obtained on the basis of a stochastic modeling, upon which we obtain certain analytical results, and conduct a numerical simulation. We validate the stochastic modeling through a discrete–event simulation. We believe that modeling presented in this paper enlight the performance issues of the binary ABR schemes.

Keywords – ABR, ATM, congestion control, binary scheme, EFCI, fairness, max–min, proportional fairness, stochastic approximation, ODE, Lyapunov, Runge-Kutta.

1 Introduction

Six different Asynchronous Transfer Mode (ATM) service classes are defined by the ATM Forum: CBR (constant bit rate), rt-VBR (real time Variable Bit Rate), nrt-VBR (non–real time Variable Bit Rate), UBR (unspecified bit rate), ABR (available bit rate), and GFR (guaranteed frame rate). All of them are specified in the traffic management specification [1], except the GFR that gained attention only very recently. CBR, rt–VBR, and nrt-VBR assume that some traffic characteristics are known before data transmission, upon which connection admission control and resource reservations are performed in order to provide certain quality of service guarantees. Data applications have an inherit property that the generated traffic is bursty and it is very hard (or even infeasible) to specify traffic characteristics prior the data transmission. In addition, data applications can tolerate delay and delay variation, but are sensitive to data loss. UBR is the simplest solution where the higher–level packets are fragmented into the ATM cells, which are transmitted without any specific congestion control; thus it relies on the higher–level congestion control like TCP congestion avoidance [2]. On the other hand, ABR is a far more complex with defined end–to–end rate–based control. UBR does not provide any guarantees at all, while ABR provides limited guarantees in the form of guaranteed minimum cell rate. Due to complexity of the ABR service and some discovered shortcomings to support prevailing data applications based on TCP, recently, GFR service class is proposed. Essentially, GFR is an enhanced UBR to provide packet level guarantees by combining open–loop UBR with packet level discarding mechanisms. For a brief up-to-date overview of ATM traffic management issues reader is referred to an article by Ghani, Nananukul, and Dixit [3].

In our previous work [4] we showed how the rates are distributed for an additive–increase/multiplicative–decrease congestion control with constant rate updating intervals associated with each source. The

*The simulation experiments presented in this paper were conducted within Department of Electronics, University of Split, Split, Croatia.

result holds for an arbitrary network topology with multiple bottlenecks and heterogeneous round–trip times. Underlying assumptions are the regime of rare negative feedback and rate proportional feedback. The modeling is particularly suited to analysis of TCP congestion avoidance. In addition, in [5] and [6], we addressed transient behavior of the explicit–rate schemes, and impact of the variable available bandwidth on the rate allocation performance, respectively.

It is well known that a simple binary congestion avoidance scheme (Explicit Forward Congestion Indication; EFCI, and Relative Rate Marking; RRM [1]) cannot ensure a fairness goal of the TM4.0 specification, namely, max–min fairness. Throughout the existing literature it is claimed that the ABR connections tranversing a larger number of links (hops) are discriminated; it is referred to this phenomena as a "beat–down problem". This is commonly argumented with a rather intuitive fact that the connections tranversing a larger number of links have a higher probability of marking their cells. In this paper we examine this phenomena through stochastic modeling, upon which we obtain some analytical results and conduct a numerical solving. The modeling is based on a theory of stochastic approximation algorithms [7]. The model applies to an arbitrary network topology of multiple bottlenecks. Also, we compare the model results with corresponding results of the discrete–event simulation. It is an objective of this paper to clarify the fairness issues, particularly, to evaluate divergence of the fairness achieved with the binary feedback and simple additive–increase/multiplicative decrease rate control, from the one determined by the max–min allocation. Although, recently, there has been a growing interest for explicit–rate (ER) schemes, due to superiority in terms of performance in respect to the binary schemes, we believe that performance issues of the binary schemes are still relevant, particularly due to existing first generation binary switches, and issues of interoperability of the binary and ER schemes.

The remainder of the paper is structured as follows. In Section 2 we give an overview of the ABR congestion control. Section 3 describes the stochastic model and applies obtained results to the parking–lot network. In Section 4 we validate the stochastic modeling results through both numerical simulation of the stochastic model, and a discrete–event simulation. In Section 5 we discuss obtained results and related work. The paper is concluded with some remarks in Section 6.

2 An Outline of the ABR Congestion Control

In this section we introduce notation and definitions that are used throughout subsequent presentation. Let $\mathcal{S}$ and $\mathcal{L}$ be sets of the connections and links, respectively. With each connection $i \in \mathcal{S}$ associate source rate x_i; similarly, to the link $l \in \mathcal{L}$ associate the link capacity c_l. Routing setting of the connections (topology) we define by matrix $A_{l,i}$; $A_{l,i} = 0$, if connection i does not tranverse link l, and $A_{l,i} = 1$ otherwise. Subsequently, let r_i and η_i be an additive–increase and multiplicative–decrease parameters, respectively. Let $I_{i,n}$ is a binary indication at the nth rate updating of connection i, with values in $\{0, 1\}$.

In the current form [1] ABR protocol operates as follows. Sending rate of the ABR source is controlled on the basis of a close–loop control that consists of a binary and/or explicit–rate feedback; the end–system algorithm is defined to operate with a mix of both feedbacks. The control loop is established by inserting resource management (RM) cells each N_{rm} data cells into the cell stream generated by the source. Both data and RM cells contain an indicator bit that is set to 0 by the source; referred to as an EFCI and CI bit for data and RM cells, respectively. In addition, RM cells contain a no increase (NI) bit and explicit rate (ER) field. Specification [1] distinguishes three types of switches: (1) Explicit Forward Congestion Indication (EFCI), (2) Relative Rate Marking (RRM), and (3) Explicit Rate (ER) switches. Along the forward connection path, the EFCI switches set EFCI bit of the data cells, the RRM switches set CI and NI bits of the RM cells, and the ER switches set the ER field of the RM cells. We refer to the EFCI and RRM switches as *binary switches* (resp. to the rate control based on the EFCI and RRM feedback as *binary schemes*). The forward RM cells are turned around by the destination, send back to the source, and potentially further modified in the backward path. Basically, transmission rate of the source end–system, referred to as an Allowed Cell Rate (ACR), is adjusted upon receipt of the backward RM cells as follows.

Let us consider an ABR connection $i \in \mathcal{S}$ with initial, minimum, and peak cell rate parameters ICR_i, MCR_i, and PCR_i, respectively. Then define $r_i = RIF_i PCR_i$ and $\eta_i = RDF_i$, where RIF_i, RDF_i are respective rate increase and rate decrease parameters, defined on $(0, 1]$[1]. Parameters MCR_i, PCR_i, RIF_i, and RDF_i are negotiated in the connection setup phase. The ACR of the connection i is adjusted upon receipt of the nth backward RM cell

$$\begin{aligned} &x_{i,0} = ICR_i, \\ &x_{i,n+1} = \Pi_{[MCR_i, PCR_i]}\{\min(x_{i,n} + r_i(1 - I_{i,n}) - \eta_i x_{i,n} I_{i,n}, ER_{i,n})\},\ n \geq 1, \end{aligned} \tag{1}$$

where $ER_{i,n}$ is the explicit rate contained in the latest received backward RM cell, and $\Pi_{[MCR_i, PCR_i]}(\cdot)$ denotes projection on the interval $[MCR_i, PCR_i]$, i.e. $\Pi_{MCR_i, PCR_i}(\cdot) = \min(\max(\cdot, MCR_i), PCR_i)$.

Note that (1) is a simple additive–increase and multiplicative–decrease algorithm upper bounded by $\{ER_{i,n}\}_{n>0}$ and constrained on $[MCR_i, PCR_i]$. In principle, ABR source end–system sets ER field of the forward RM cells to the PCR value. Therefore, for a connection tranversing no ER switches, one can ommit $\min(\cdot)$ and $ER_{i,n}$ in (1) and obtain a simple additive–increase and multiplicative–decrease algorithm constrained on $[MCR_i, PCR_i]$.

It should be noted that the rate adjustment (1) is not complete comparing to [1] that defines several mechanisms to alleviate certain abnormal conditions, for instance

- exponential backoff due to absence of backward RM cells (source rule 6 [1]),
- upper bound of the rate adjustment interval (source rule 3.a.i [1]).

In the modeling given later in this paper we do not consider aforementioned special conditions of the rate adjustment.

Behavior of the ABR source and destination systems is specified in detail by TM4.0 [1], however, switch behavior is not standardized and is leftover to the equipment vendors to choose one particular algorithm. Given a textual description of ABR source and destination algorithms [1], Lee, Ramakrishnan, and Moh developed an extended finite state machine (EFSM) in [8]. On the basis of the EFSM model, correctness of the TM4.0 specification, and interoperability of EFCI and ER schemes are examined in [9]. For a good surveys of ABR protocol the reader is referred to [10, 11, 12]. An extensive coverage of setting of the ABR end–system parameters is given by Fahmy, Jain, Goyal, and Vandalore in [13].

2.1 Binary Schemes

Considerable amount of work has been conducted on analysis of the binary ABR congestion control. Fundamental principles of additive–increase/multiplicative–decrease are addressed in Chiu and Jain [14]. Bonomi, Mitra, and Seery [15] modeled a single bottleneck with heterogeneous round–trip times as a system of the first–order delay differential equations. Ohsaki at al. [16] and Ritter [17] analyzed a steady–state of multiple ABR connections sharing a single bottleneck, with equal end–system parameters and network propagation delays; particularly, they identified phases and modeled corresponding rate evolutions with differential equations, upon which they computed maximum queue length.

Recently, Ait–Hellal and Altman [18] presented analytical modeling of the ABR congestion control for the tandem of multiple switches. They showed that it is not necessary that the bottleneck link is the one with the slowest transmission rate. Also, it is shown that the maximum queue lengths may be underestimated by approximating the network with a single bottleneck.

In order to alleviate unfairness of the simple binary schemes two approaches appeared: (1) Per-VC queueing, and (2) selective feedback. The former impose significant implementation and storage complexity due to per–VC queueing. The selective feedback was proposed as a part of the seminal series of reports on DECbit scheme [19].

[1] [1] defines RIF and RDF on a discrete set $\{\frac{1}{2^n},\ 1 \leq n \leq 15\}$.

2.2 Explicit–Rate Schemes

It was recognized that explicit–rate schemes provide superior performance than the binary schemes in respect to the steady–state rate dynamics, speed of convergence, stability, and fairness. One of the early proposals of distributed explicit–rates computation was done by Charny, Clark, and Jain [20]. Later on many other algorithms were proposed, e.g. [21, 22, 23], with objectives to improve performance, and decrease time/memory complexity of the ER computation.

In the recent sequel of papers [24, 25], Abraham and Kumar considered the max–min allocation with non-zero MCRs; they proved that unique max–min allocation can be obtanied as a solution of a certain vector equation [24], and showed how the max–min allocation can be computed in a distributed manner by applying a theory of stochastic approximation algorithms [25]. Also, a sliding–window estimation of the available bandwidth, based on a theory of effective bandwidth, is proposed in [26] by the same authors.

2.3 Fairness Issues

The target fair rate allocation, selected for the ABR service [1], is centered around a max–min fairness. There are several definitions of the max–min fair rate allocation; maybe the most common one is the following. First, let $x = (x_i, i \in \mathcal{S})$ be a vector of rates, then, x is said to belong to the set of feasible rate vectors if $\sum_{i \in \mathcal{S}} x_i \leq c_l$, for all $l \in \mathcal{L}$. A vector x is said to be max–min fair if it is feasible and for each $i \in \mathcal{S}$, x_i cannot be increased while maintaining feasibility without decreasing x_j for some flow j for which $x_j \leq x_i$ [27]. Notion of the max–min fairness was introduced by Jaffe [28], who considered the rate distribution to achieve an ideal trade–off between high throughput and low delay. For a textbook treatment of the max–min fairness refer to [27]. There are also some other notions of the fairness; perhaps the most prevailing one is the proportional fairness [29]. A vector x is said to be proportionaly fair if x maximizes $\sum_{i \in \mathcal{S}} \log x_i$ within the set of feasible vectors. Recently, Mo and Walrand [30] defined (p, α)–proportionally fair allocation, where $p = (p_i,\ i \in \mathcal{S})$, $p_i > 0$ for all $i \in \mathcal{S}$, and $\alpha > 0$. Among other results, the authors showed that max–min and proportional fairness are the special cases of (p, α)–proportional fairness for $\alpha \to \infty$ and $\alpha = 1$, respectively. In this section we briefly discussed the max–min and proportional fairness that are further considered in the next sections.

3 Modeling

The main modeling assumptions are: (1) sources are persistent (greedy, infinite), meaning that they always have a cell awaiting for transmission, (2) routing setting is static determined by matrix A; lifetime of the connections span the time interval of concern.

Subsequently, we build further notation and definitions on the one introduced in the begining of Section 2. Let subscripts i, n of $x_{i,n}$ denote the nth temporal sample of the ith component of x. Let $\{\tau_{i,n}\}_{n \geq 0}$ be a sequence of updating times of rate x_i. Then, define an auxiliary variable $\delta\tau_{i,n}$ equal to the nth updating interval $\delta\tau_{i,n} = \tau_{i,n+1} - \tau_{i,n}$. Also, we introduce temporal quantities $\Delta_{i,l}$ and $\Delta_{l,i}$ corresponding to delay from the source i to the link l, and vice–versa, respectively. Then, we define $\Delta_{i,j}$ as $\Delta_{i,j} = \Delta_{j,l} + \Delta_{l,i}$; clearly $\Delta_{i,i}$ is a round–trip time of connection i. Let $x_i(\cdot)$ be a piece–wise constant interpolation of x_i on the real time (or scaled real time), such that $x_i(t) = x_i$, for $\tau_i \leq t < \tau_{i+1}$. Then, we define overall load at the link l present at time t as

$$f_l(t) = \sum_{j \in \mathcal{S}} A_{l,j} x_j (t - \Delta_{j,l}).$$

Let $x = (x_i, i \in \mathcal{S})$ be an asymptotic limit, $x_i = \lim_{n \to \infty} x_{i,n}$, for all $i \in \mathcal{S}$, then, a set of feasible rate allocations contains all x such that

$$\sum_{j \in \mathcal{S}} A_{l,j} x_j \leq c_l, \text{ for all } l \in \mathcal{L}. \tag{2}$$

Kushner and Yin [7] obtained a weak convergence result (convergence in probability) for the class of asynchronous distributed stochastic approximation algorithms with a small constant gain γ. Specifically, they considered an algorithm of the form

$$\begin{aligned} x_{i,n+1} &= \Pi_{[a_i,b_i]}\left(x_{i,n} + \gamma H_{i,n}(x_j(\tau_{i,n+1} - \Delta_{i,j,n}), j \in \mathcal{S})\right) = \\ &= x_{i,n} + \gamma H_{i,n}(x_j(\tau_{i,n+1} - \Delta_{i,j,n}), j \in \mathcal{S}) + \gamma Z_{i,n},\ i \in \mathcal{S}, \end{aligned} \tag{3}$$

where $\Pi_{[a_i,b_i]}(\cdot)$ denotes projection of the argument on $[a_i, b_i]$; for x constrained on $C = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_S, b_S]$, and $Z^{\gamma}_{i,n}$ is a reflection term. Note that in our context $a_i \equiv MCR_i$ and $b_i \equiv PCR_i$.

$\Delta_{i,j,n}$ denotes already defined $\Delta_{i,j}$ at the nth update of the ith component of x. It is assumed that the communication delays are bounded in the sense

$$\gamma \Delta_{i,j,n} \to 0, \text{ as } \gamma \to 0. \tag{4}$$

The algorithm is asynchronous since the updating of the ith component is not aligned in respect to updating of other components of x, and distributed since the components are delayed due to communication delays.

From the Theorem 3.1 [7] Ch. 12.3, p. 364–365, it follows that the weak convergence subsequence is the limit set of an ordinary differential equation (ODE)

$$\frac{dx_i}{dt} = \frac{\bar{h}_i(x)}{\bar{u}_i(x)} + z_i,\ i \in \mathcal{S}, \tag{5}$$

where

$$E_{i,n}[H_{i,n}] = h_{i,n}(x_j(\tau_{j,n} - \Delta_{i,j,n}), j \in \mathcal{S}),$$

$$E_{i,n}[\delta\tau_{i,n+1}] = u_{i,n+1}(x_j(\tau_{i,n+1} - \Delta_{i,j,n+1}), j \in \mathcal{S})$$

and $E_{i,n}[\cdot]$ is a conditional expectation defined as $E_{i,n}[\cdot] = E[\cdot | x(s), s < \tau_{i,n}, \tau_{j,n} < \tau_{i,n+1}, j \in \mathcal{S}]$. Thus, $E_{i,n}[\cdot]$ is a conditional expectation given all past data before time $\tau_{i,n+1}$, including $x_{j,k+1}$ and $\tau_{j,k+1}$, for all $j \in \mathcal{S}$. Then, let $\bar{h}_i(\cdot)$ and $\bar{u}_i(\cdot)$ be respective asymptotic averages of $h_{i,n}(\cdot)$ and $u_{i,n}(\cdot)$, as $\gamma \to 0$ and $n \to \infty$. For a complete treatment of the underlying theory refer to [7], and its application to TCP congestion avoidance refer to [4].

Basic principle behind the proof of the convergence is an observation of the interpolation of the discrete process x_i on the real time scaled by the step size γ. Then, in the asymptotic case, whereas $\gamma \to 0$ and number of iterations $n \to \infty$, one neglect delays that satisfy condition (4).

Note that for an unconstrained $x(\cdot)$, element z_i in (5) is set to 0. It turns out that the limit mean ODE is the same as in the synchronous case; only exception is a factor $1/\bar{u}_i$ (5) that captures diversity of the rate updating frequency, i.e. inverse of the rate updating interval.

In the steady–state, the frequency of the rate updating $1/\bar{u}_i(\cdot)$ is proportional to rate $x_i(\cdot)$ delayed by a value greater than one round–trip propagation delay; strictly equal to propagation delay if the queueing delay, and other delays involving processing time at the switches, source, and destination end–systems can be neglected. Thus, in the asymptotic case the delays are omitted and we may write

$$\bar{u}_i(x) = \frac{N_i}{x_i}. \tag{6}$$

where N_i stands for end–system parameter N_{rm} associated with connection $i \in \mathcal{S}$. Here we should note following. In several studies [16, 17] of the single bottleneck link of capacity c, with N identical end-systems, the rate of the backward RM cells is taken proportional to the rate delayed by the round–trip time, if the buffer is empty, otherwise, the rate is set to $\frac{1}{N_{rm}}\frac{c}{N}$, if the buffer is not empty. Therefore, it is assumed that each connection contains an equal share of the queue content. In the general case of an arbitrary setting of end–system parameters, it is reasonable to assume that the rate of the backward RM cells is proportional

to $x_i(t-\Delta_{i,i})$ shifted by the queueing delay. Therefore, in both cases the rate of the backward RM cells is proportional to $\frac{1}{N_{rm}}x_i(t-\Delta_{i,i})$. This assumption is particularly plausible if the RM cells are queued separately and given a higher priority over the data cells; in this case, obviously, the rate of the backward RM cells is independent of the queue backlog of the data cells.

We assume that parameters of additive–increase and multiplicative–decrease are small[2], and define r_i^γ and η_i^γ such that $r_i = \gamma r_i^\gamma$ and $\eta_i = \gamma \eta_i^\gamma$. Then, on the basis of (1) we identify function $\bar{h}_i$ as

$$\bar{h}_i(x) = r_i^\gamma - (r_i^\gamma + \eta_i^\gamma x_i) P_{i,n}(I_{i,n}=1), \tag{7}$$

thus, with a given $P_{i,n}(I_{i,n}=1)$, (6) and (7) determine the limit mean ODE (5).

Queue backlog at the link l is goverened by the following differential equation

$$\dot{q}_l(t) = \begin{cases} f_l(t) - c_l, & q(t) > 0, \\ \max(0, f_l(t) - c_l), & q(t) = 0, \end{cases} \tag{8}$$

with an initial value $q_l(0)$. Similar modeling of the queueing was used for instance in [15] to model a single–bottleneck case. However, note that for an arbitrary topology $f_l(\cdot)$ must account all upstream queueing delays, and have to satisfy feasibility constraints (2). With a congestion detection on the basis of queue threshold (single or double threshold) congestion is indicated, $I_{i,n}=1$, if $q_l(\tau_{i,n+1}-\Delta_{l,i})$ is exceeding the predetermined queue threshold. Therefore, the system evolution is completely determined by the coupled systems of equations (5) and (8). However, analytical analysis becomes cumbersome, hence, we abstract the queuing through a notion of the link cost function $g_l : [0,\infty) \to [0,1]$ that is a function of load $f_l(\cdot)$. Intuitively, one can think about the link cost function as to be related to the tail distribution of the queue length $P(q_l > q_t | f_l)$ given a load f_l, $q_t \geq 0$, which corresponds to probability of setting the congestion indication $I_{i,n}$. Indeed, the queue tail distribution is expressed in terms of an average load, but for the small variations of the rates in the steady state it seems reasonable to replace the average total load with a value of the current total load. Otherwise, one can interprete g_l as a relation between the current load and probability of marking the cells. Actually, in this case, congestion detection is based on the load, i.e. first derivative of the queue length. One such congestion detection mechanism, named EFCI-ECD (Early Congestion Detection), was proposed by Zhao, Li, and Sigarto in [31].

It should be noted that the system evolution determined by (5) and (8), with a given inital values, is fully deterministic. However, with a modeling based on (5) and a concept of the link cost functions we shift to a stochastic modeling. Finally, replacing (6) and (7) into (5), we obtain an initial value problem

$$\frac{dx_i}{dt} = \frac{1}{N_i} x_i [r_i^\gamma - (r_i^\gamma + \eta_i^\gamma x_i) P_{i,n}(I_{i,n}=1)],\ i \in \mathcal{S}, \tag{9}$$

on $[0, t_f]$, with $x_i(0) = ICR_i$.

In the sequel, we assume the probability of negative feedback of the form

$$P(I_i = 1) = 1 - \prod_{l: A_{l,i} > 0} (1 - g_l(f_l)), \tag{10}$$

that corresponds to probability of marking a single data cell along the connection path.

The limit set of the ODE (5) can be computed by identifying a Lyapunov function, which maximization yields an attractor towards solutions of (5) converge [4]. Let us write (9) in the form

$$\frac{dx_i}{dt} = \frac{x_i}{r_i + \eta_i x_i} \frac{\partial}{\partial x_i} \{F_{ABR}(x) - G(x)\}, \tag{11}$$

[2]This assumption is less restrictive for the binary schemes for which it is observed that conservative smaller values of RIF and RDF are desirable to avoid high rate oscillations and congestion loss [13].

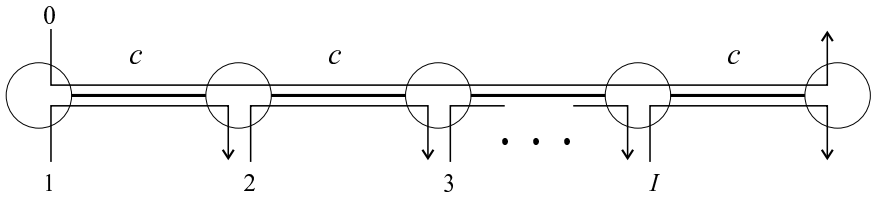


Figure 1: A Parking–lot network.

where

$$F_{ABR}(x) = \sum_{i \in \mathcal{S}} \frac{r_i}{N_i \eta_i} \log(r_i + \eta_i x_i), \tag{12}$$

and $G(x)$ is a primitive of (10). Note that the maximization of (12) is subject to the constraints (2).

Alternatively, one can solve system (5) by numerical methods; this approach is particularly amenable to incorporate constraints $[MCR_i, PCR_i]$, $i \in \mathcal{S}$. Note that we are interested in an asymptotic solution $x(t_f)$, as $t_f \to \infty$, however, in reality, we compute numerically $x(t_f)$ for large enough $t_f < \infty$.

In the sequel of the paper we refer to the rate distribution derived from (12) as *an analytical result*; computed through numerical solving of (5) as *a numerical result*; measured through discrete–event simulation as *a simulation result.*

3.1 A Parking–lot

We validate our analytical results through simulation of the parking–lot network topology depicted in Fig. 1. The network consists of a tandem of I links of capacity c and arbitrary propagation delay δ. Let access links be of an infinite capacity and arbitrary propagation delay. Particularly, we restrict ourselves to the two network scales; all access delays equal to zero (HETRTT), and access delays set such that all the connections tranversing a given link have equal source–to–link and link–to–destination delay (HOMRTT). We distinguish a multi–hop connections (class 0) and a single–hop connections (class i, $i = 1, 2, \ldots, I$). In order to shrink set of parameters values we assume the following. There is v connections of class 0, and w connections of class i. Class 0 sources are associated with r_0, η_0, and N_0 parameters, likewise, class i connections are associated with r_I, η_I, and N_I parameters.

We express (12) in terms of x_0 utilizing $x_I = (c - vx_0)/w$ from (2). Then it is straightforward to compute x_0 for which maximum of $F_{ABR}(x_0)$ is attained

$$\frac{x_0}{c} = \begin{cases} \frac{1 - w\frac{r_I}{\eta_I}(\frac{N_0}{N_I} I - 1)}{\frac{N_0}{N_I} v + wI\frac{r_I \eta_0}{r_0 \eta_I}}, & \frac{r_I}{\eta_I} < \frac{1}{w} \frac{1}{I - \frac{N_I}{N_0}} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

where we abuse previous notation and write r_0 and r_I to denote additive–increase parameters normalized by c. Assuming identical setting of the end–system parameters, on the basis of (13), it is obvious that class 0 connections get not more throughput then the one determined by proportional fairness ($\frac{1}{v+wI}$) [29]. Obviously, disparity between (13) and the proportional fairness, for identical setting of the configuration parameters for all sources, is particularly emphasized for larger $\frac{r}{\eta}$ ratio (where $r_0 = r_I = r$ and $\eta_0 = \eta_I = \eta$), larger number of single–hop connections w, and number of links I.

Henceforth, it is evident that there is a strong bias against connections tranversing a large number of hops, and the rate distribution substantially diverse from the one determined by the max–min fair rate distribution ($\frac{1}{v+w}$).

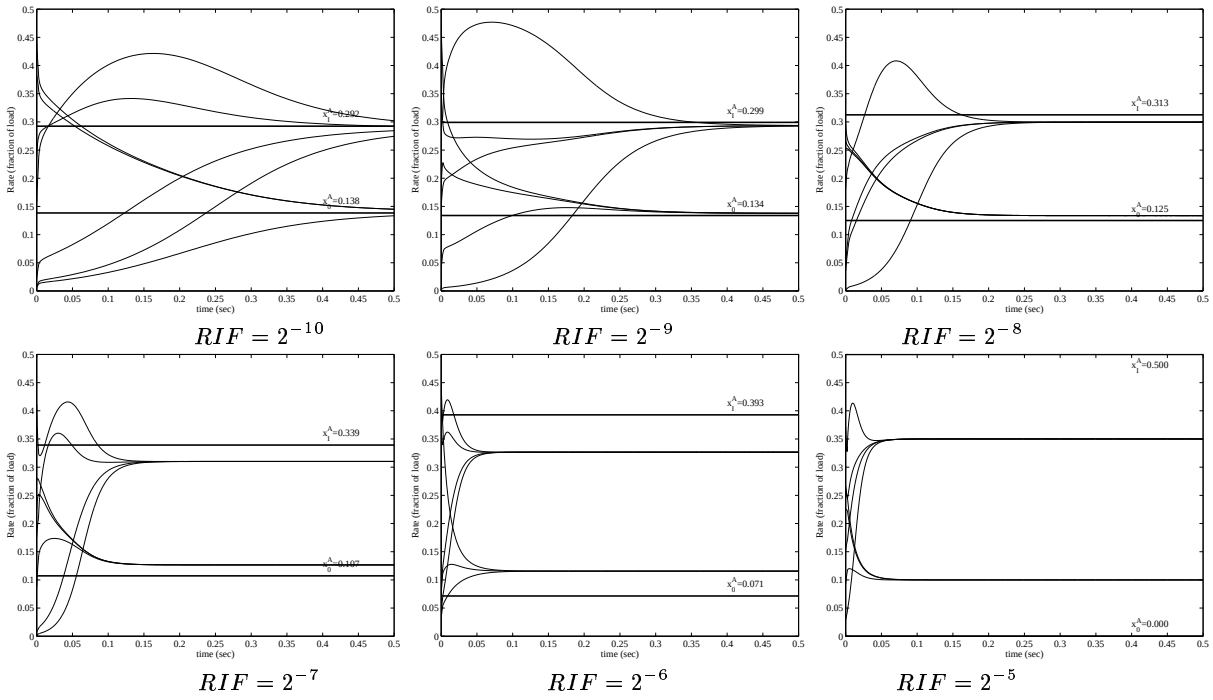


Figure 2: Numerical simulation – fraction of load allocated to the sources versus RIF; $RDF = 2^{-4}$, $c = 150$ Mbps, $PCR = c$, $MCR = 1$ Mbps, $N_{rm} = 32$, $I = 2$, $v = 3$, $w = 3$, $d = 0.5$, $p = 5$ (thick lines indicate the analytical result).

4 Numerical and Simulation Results

We conduct the numerical simulation of the limit mean ODE (5) by using the Runge–Kutta method contained in the MATLAB ODE suite (`ode45`) [32]. All the numerical results are obtained for the link cost function of the form $g_l(f_l) = 0$, for $f_l < 0$, $g_l(f_l) = 1$, for $f_l > c_l$, $g_l(f_l) = (\frac{f_l/c_l - d}{1-d})^p$, otherwise, where $d \in [0, 1]$, and $p > 0$. This form of $g_l(\cdot)$ is also used in [4] and elsewhere.

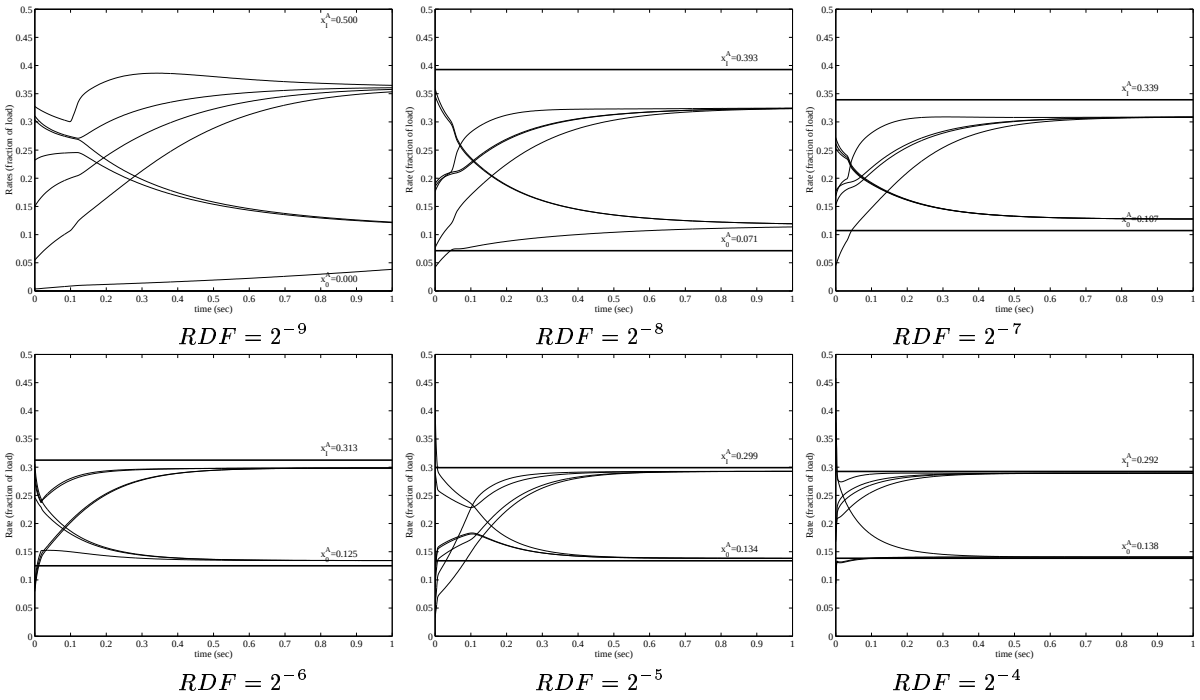
Besides numerical simulation of the model we performed a discrete–event simulation of the system so that the impact of the queueing is evaluated. The simulation model is built on the basis of a discrete–event class library CNCL [33] that we extended with ATM classes.

The model of ATM switch is a simple output–queueing switch, where we distinguish two service disciplines. First, switch treats data and RM cells transparently, where cells share a common per–Port FIFO buffer. Later, data and RM cells are queued separately with respective per–Port FIFO buffers that are served giving higher priority to the RM cells, i.e. whenever the RM buffer is non–empty that buffer is served. We refer to the later service discipline as an express RM queueing.

The congestion is detected on the basis of the instanteneous queue length and a double queue thresholds; low q_l, and high q_h. Whenever the instanteneous queue length is larger than q_h the port is in the congestion state. The port moves from the congestion state to the non–congestion state when the queue length becomes less than q_l. Objective of using two thresholds instead of a single one is to avoid oscillations between the two states. However, in order to shrink the parameter space, in majority of the simulations, we assume $q_l = q_h (= q_t)$, which corresponds to the single threshold congestion detection mechanism.

We assume an infinite buffer size, hence, the cell loss is not occuring at the switches. In reality, this corresponds to a well–engineered network, where in the steady state the queue length is not exceeding the buffer size at any time.

The network parameters are set as $c = 150$ Mbps, and $\delta = \{1, 50\}$ km. We refer to the respective δ



$RDF = 2^{-9}$ $RDF = 2^{-8}$ $RDF = 2^{-7}$

$RDF = 2^{-6}$ $RDF = 2^{-5}$ $RDF = 2^{-4}$

Figure 3: Numerical simulation – fraction of load allocated to the sources versus RDF; $RIF = 2^{-10}$, $c = 150$ Mbps, $PCR = c$, $MCR = 1$ Mbps, $N_{rm} = 32$, $I = 2$, $v = 3$, $w = 3$, $d = 0.5$, $p = 5$ (thick lines indicate the analytical result).

values as a LAN and MAN cases. For $I = 2$ scenario these two cases correspond to the fixed round–trip time, of the class 0 sources, of about 22μs, and 1 ms, respectively. Larger round–trip time case (tens of ms), corresponding to a WAN scale, is not considered in this paper due to time and space limitations. Relevant ABR configuration parameters are set as follows: $N_{rm} = 32$, $PCR = c$, $MCR = 0.5$ Mbps, $ICR =$ random uniformly distributed on $[0, \frac{c}{v+w}]$, $RIF = \{2^{-n}, n = 4, 5, \ldots, 11\}$, $RDF = 2^{-4}$.

The asymptotic rate allocated to the class 0 sources is estimated based on the average of the last half of the trace; captured over time interval taken sufficiently large for the rates to converge. For the majority of the simulations total simulation time is set to 2 seconds. Simulation experiments are run 5 times, upon which an average and 95% confidence interval is computed. It should be noted that in order to take into account diversity of the link utilization, the rates are normalized in respect to the average link load. First, we consider the numerical simulation results shown in Fig. 2 and Fig. 3 for varying RIF and RDF parameters, respectively. It can be observed that discrepancy between the analytical and numerical results is higher as the additive–increase RIF is higher (fixed RDF), and as the multiplicative–decrease RDF is lower (fixed RIF). This observation can be explained with the fact that larger RIF to RDF ratio moves the network operating point towards higher link utilization, consequently, g_l becomes larger, and impact of the $G(\cdot)$ element in (11) becomes more significant. Relation of the RIF (RDF) and the link utilization/cost function is shown in Fig. 4a (resp. Fig. 4b). Additionally, it is evident that higher RIF and RDF yield faster convergence to the steady–state rates.

Second, we study the simulation results shown in Fig. 5 and 6; for a reference we plotted also the numerical results for $d = 0.8$ and $p = 5$. For the LAN setting, Fig. 5a, and $RIF \leq 1/128$, there is a fair matching of the analytical result and a simulation one for both non express RM and express RM cases. For higher RIF values, the express RM case exhibits expected behavior, while the non express RM case deviates from the non–increasing trend. This is not suprising since the interarrival times of the backward RM cells, for the non express RM case, are influenced by high periodic oscillations of the queue

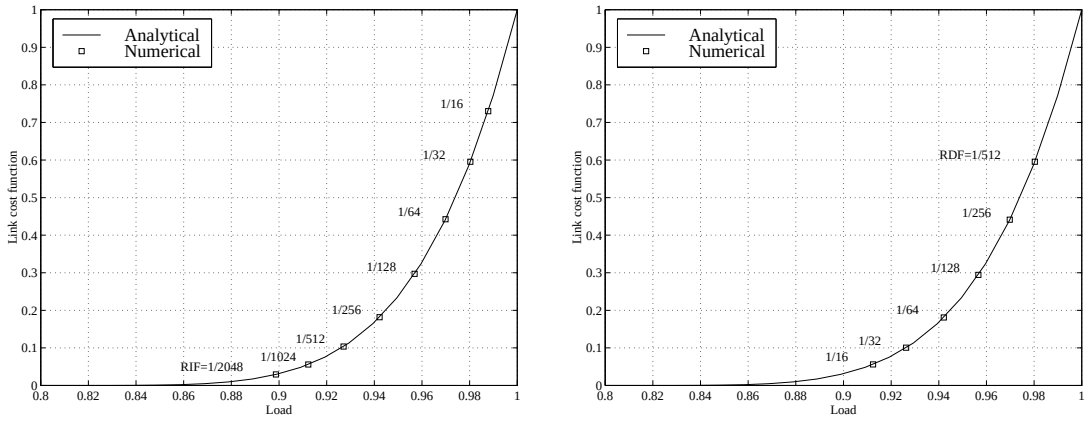


Figure 4: Numerical simulation – average load and link cost function versus RIF ($RDF = 1/16$), (b) RDF ($RIF = 1/1024$); $c = 150$ Mbps, $PCR = c$, $MCR = 0.5$ Mbps, $I = 2$, $v = 3$, $w = 2$.

backlog, and consequently, differ from the assumed relation (6). Similar observations also hold for the MAN settting Fig. 5b and Fig. 5c. Note that the bandwidth delay product for the MAN setting is about 350 cells (counting only fixed propagation delay); value in between the queue thresholds 250 and 500 (due to space limitations we ommit the results for the later case). Note also an abnormal confidence interval for $RIF = 1/16$ and MAN setting (Fig. 5b) that appeared due to rate starvation of the class 0 sources for certain settings of the initial cell rates. Subsequently, results obtained for varying RDF parameter and fixed $RIF = 1/1024$, shown in Fig. 6, exhibit expected behavior. Higher values of RDF, for identical setting of the end–system parameters, yield an allocation closer to the proportional fairness (as observed on the basis of (13)). Also, again, it is evident that for smaller values of RDF there is a higher discrepancy between the numerical/simulation result and the corresponding analytical one, due to aforementioned reasons.

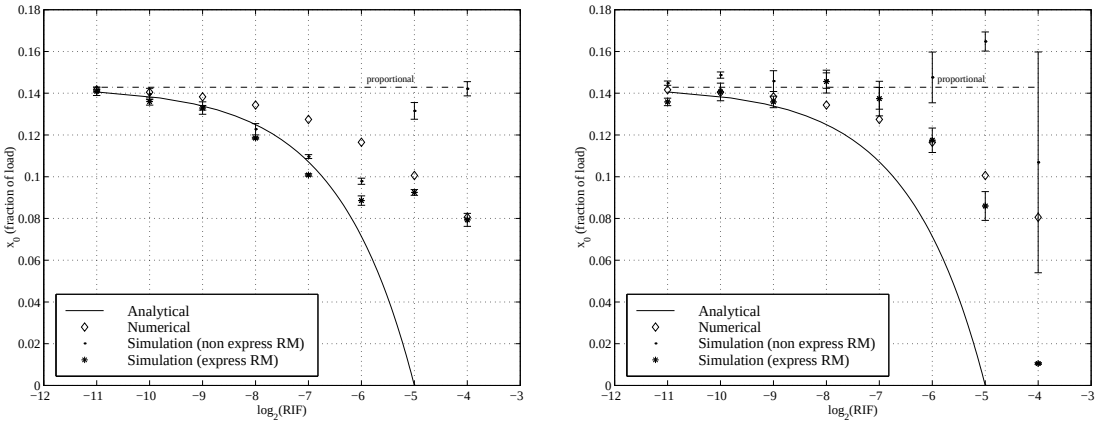
5 Discussion of the Results and Related Work

On the basis of the obtained results we observe following. The analytical result (12), and a special case (13) indicate that ratio $\frac{r_i}{\eta_i}$ plays an important role in the rate distribution. First, it is a weight (divided by N_i) of the rate utility function of the source i (12), thus, higher $\frac{r_i}{\eta_i}$ implies higher rate utility. Second, it determines significance of the elements in the sumation of the $\log(\cdot)$ argument. For a moment, let r_i, η_i, and N_i be set to equal values for all sources, then, for $r_i << \eta_i x_i$ rate distribution is close to the proportional fairness.

Next, we observe that in respect to the asymptotic limit ($t \to \infty$ and small r_i and η_i parameters), rate distribution does not depend explicitly on the round–trip times; this is not the case with TCP congestion avoidance (see [4]).

The result obtained for the parking–lot network (13) suggests following selection of the parameters in order to achieve max–min fair allocation. Ratio $\frac{r_I}{\eta_I}$ should be kept small (relatively in respect to the number of connections w and number of hops I[3]), while setting of $\frac{r_0}{\eta_I}$ should be equal to the value I times larger than $\frac{r_I}{\eta_I}$. This observation is validated through simulation for the range of varying RIF_0 parameter with fixed RIF_I, see Fig. 7. The analytical, numerical, and simulation results conform to each other, and it is shown that different fairness objectives can be achieved by proper selection of the end–system parameters. Validation of the above reasoning for a more extensive set of parameters, and an arbitrary network topology remains to be done.

[3]This certanly implies scalability problems.



(a) LAN, HETRTT; $q_t = 250$ cells (b) MAN, HETRTT; $q_t = 250$ cells

Figure 5: Fraction of load allocated to class 0 sources versus additive–increment RIF; $RDF = 1/16$, $c = 150$ Mbps, $PCR = c$, $MCR = 0.5$ Mbps, $I = 2$, $v = 3$, $w = 2$.

It is commonly claimed that for the binary ABR schemes RIF and RDF parameters have to be set to small values [13]. However, on the basis of our results we induce that RIF and RDF should comply to the following setting rules (relative within interval $[0, 1]$)

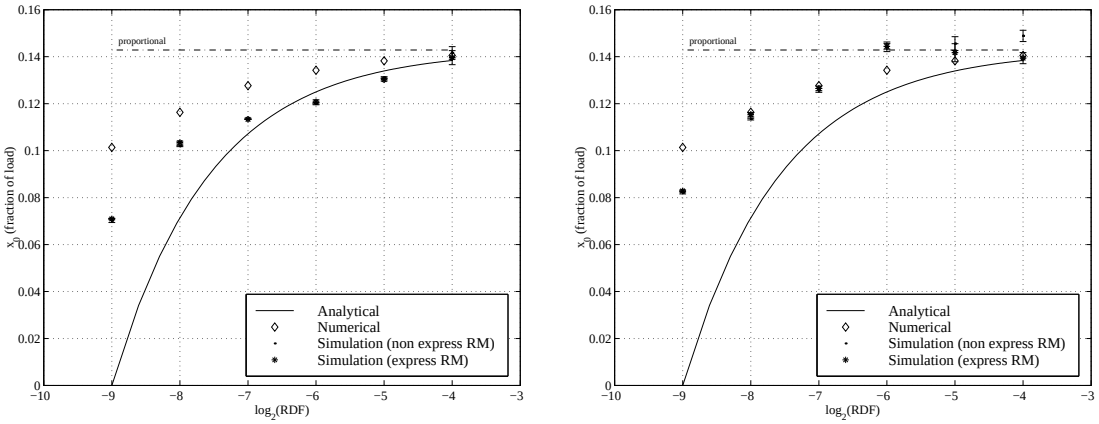
- *speed of convergence* – large RIF and large RDF,
- *utilization* – large RIF and small RDF,
- *fairness* – small RIF and large RDF.

Obviously, there is a trade–off in selection of both RIF and RDF parameters.

Interoperability of EFCI and ER Schemes – In practice it is likely that EFCI and ER switches coexist, hence, interoperability of these two schemes in the heterogeneous network environment has drawn attention and it was covered in [34], and more recently in [35]. Lai and Lin [35] addressed impact on the use and setting of a certain configuration parameters in the mixed environment. Results regarding optimal locality of ER switch, in all EFCI network, given in [34] and [35] are inconsistent. Furthermore, Plotkin and Sydir [36] identified a rate mismatch problem, occuring in a heterogeneous network of oscillatory (binary) and non-oscillatory (ER) schemes. This effect is explained to occur due to difference in rapidity of the rate increase that is caused by the non-bottleneck switches that control the rate during the rate–increase phase; the resulting effect is unfairness.

TCP over ABR – Some work on performance of TCP over ABR has already been done. The existing work mainly concentrates on comparative analysis of TCP performance over ABR, versus TCP over UBR [37, 38]. In practice, ABR segement is rarely employed end–to–end; rather it is commonly a segment within a communication path between source and destination. The former corresponds to an ATM interconnection of two ATM workstations, while the later corresponds to an interconnection of two legacy LAN networks through ATM backbone. It is known that ABR congestion mechanism pushes congestion points towards the source end–system, i.e. edge of the network [37]. Henceforth, keeping in mind complexity of the ABR and inherit overhead due to RM cells, it became questonable wheather ABR provides any benefits in respect to UBR accompanied with a packet level buffer management. Nevertheless, ABR could remedy existing bias [4] against TCP connections with long round–trip times.

Fairness Issues – In majority of the work it is implicitly assumed that the utility, relating the rate allocation to user utility, is the same for all sources, or more restrictive that the relationship between utility and the



(a) LAN, HETRTT; $q_t = 250$ cells (b) MAN, HETRTT; $q_t = 250$ cells

Figure 6: Fraction of load allocated to class 0 sources versus multiplicative–decrease RDF; $RIF = 1/1024$, $c = 150$ Mbps, $PCR = c$, $MCR = 0.5$ Mbps, $I = 2$, $v = 3$, $w = 2$.

rate is linear. Accordingly, most of the rate allocation schemes are based on bandwidth exclusively. However, there has been an incentive to allocate bandwidth so that the user utility is taken into account, see for example [39, 40]. Those proposals are often in conjuction with the integrated pricing, with an objective to maximize a social welfare, defined as a subtraction of the user utility and willingness–to–pay, which result in an economically efficient utilization of network resources [40]. However, maximizing a given overall utility function is often leading to an unfair allocation of the bandwidth in respect to TM4.0 notion of fairness. Recently, Cao and Zegura [41] showed how a bandwidth max–min can be generalized to utility max–min, in order to take into account diversity of application utilities.

6 Concluding Remarks

We evaluated fairness of the rate distribution provided by the binary ABR scheme through a stochastic modeling. The model is derived under certain reasonable set of assumptions, and validated through numerical simulation of the associated limit mean ODE, and through discrete–event simulation. The results clarify intuitive arguments explaining the observed bias against the ABR connections tranversing a large number of binary ATM switches. Further work might be directed towards a more extensive validation of the result through simulation over larger set of parameters; particularly, for larger values of the propagation delays. Nevertheless, the setting of the ABR parameters to improve fairness of the binary schemes should be validated for other multiple–bottleneck topologies. Finally, a numerical solving of the coupled system of equations (rate (5) and queueing (8) dynamics) deserves further study due to speed of numerical simulation comparing to the discrete–event one.

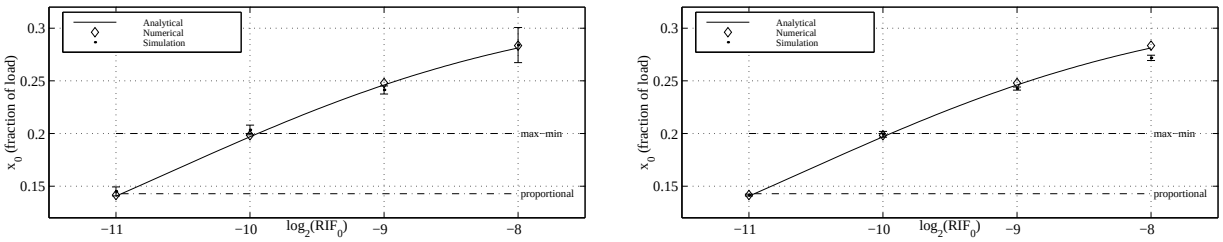
References

[1] The ATM Forum Technical Committee, "Traffic managment specification version 4.0," Tech. Rep. af-tm-0056.000, ATM Forum, April 1996.

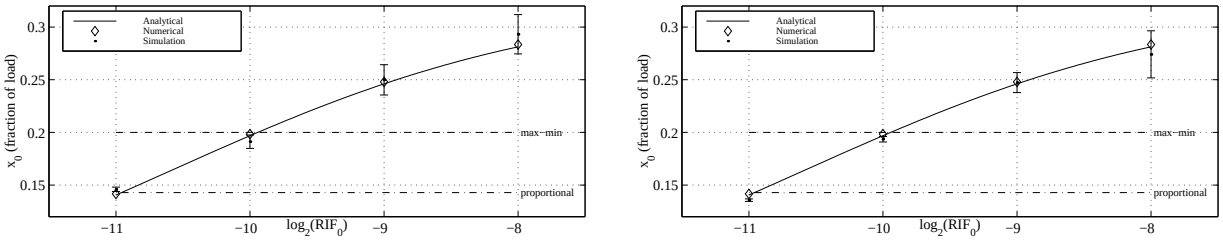
[2] V. Jacobson and M. J. Karels, "Congestion avoidance and control," in *Proc. of the ACM SIGCOMM'88*, (Stanford), pp. 314–329, August 1988.

[3] N. Ghani, S. Nananukul, and S. Dixit, "ATM traffic management considerations for facilitating broadband access," *IEEE Communications Magazine*, pp. 98–105, November 1998.

[4] M. Vojnović, J.-Y. Le Boudec, and C. Boutremans, "Global fairness of additive–increase and multiplicative–decrease with heterogeneous round–trip times," in *submitted to IEEE INFOCOM'2000*, (Tel Aviv, Israel), March 2000.



(a) LAN; non express RM (left), express RM (right)



(b) MAN; non express RM (left), express RM (right)

Figure 7: Fraction of load allocated to class 0 sources versus additive–increment RIF_0 with fixed $RIF_I = 1/2048$; $RDF_0 = RDF_I = 1/16$, $c = 150$ Mbps, $PCR = c$, $MCR = 0.5$ Mbps, $I = 2$, $v = 3$, $w = 2$.

[5] M. Vojnović and N. Rožić, "Analytical and simulation analysis of the explicit-rate ABR flow control algorithms: Transient behavior," in *Proc. of the Third IEEE Symposium on Computers and Communications*, (Athens, Greece), June/July 1998.

[6] M. Vojnović and N. Rožić, "An evaulation of the ABR explicit-rate allocation interfering with the guaranteed services traffic," *submitted to Int'l Journal of Computer and Telecommunications Networking*, September 1999. (an abbreviated version appeared in Proc. of of the IFIP ATM'98, Ilkley, UK, July 1998).

[7] H. J. Kushner and G. G. Yin, *Stochastic Approximations Algorithms and Applications.* Springer-Verlag, 1997.

[8] D. Lee, K. K. Ramakrishnan, and W. M. Moh, "A formal specification of the ATM ABR rate control scheme," *Computer Networks and ISDN Systems*, vol. 30, pp. 1735–1748, 1998.

[9] D. Lee, K. K. Ramakrishnan, M. Moh, and A. U. Shankar, "Performance and correctness of the ATM ABR rate control scheme," in *IEEE INFOCOM'97*, (Kobe, Japan), pp. 785–794, April 1997.

[10] T. M. Chen, S. S. Liu, and V. K. Samalam, "The available bit rate service for data in ATM networks," *IEEE Communications Magazine*, pp. 56–71, May 1996.

[11] R. Jain, "Congestion control and traffic management in ATM networks: Recent advances and a survey," *Computer Networks and ISDN Systems*, vol. 28, February 1995.

[12] A. Arulambalam, X. Chen, and N. Ansari, "Allocating fair rates for available rate service in ATM networks," *IEEE Communications Magazine*, pp. 92–100, November 1996.

[13] S. Fahmy, R. Jain, R. Goyal, and B. Vandalore, "ABR engineering: Roles and guideliness for setting ABR parameters," *submitted to Journal of Computer Networks, (also available through http://www.cis.ohio-state.edu/ jain/)*, 1998.

[14] D. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, pp. 1–14, June 1989.

[15] F. Bonomi, D. Mitra, and J. B. Seery, "Adaptive algorithms for feedback–based flow control in high–speed, wide–area ATM networks," *IEEE Journal on Selected Areas in Communications*, pp. 1267–1282, September 1995.

[16] H. Ohsaki, M. Murata, H. Suzuki, and C. I. H. Miyahara, "Analysis of rate-based congestion control algorithms for ATM networks – part 1: Steady state analysis," in *Proc. of the IEEE GLOBECOM'95*, pp. 296–303, IEEE, 1995.

[17] M. Ritter, "Network buffer requirements of the rate-based control mechanism for ABR services," in *Proc. of the IEEE INFOCOM'96*, no. 3, (San Francisco, California), pp. 1190–1197, IEEE, March 1996.

[18] O. Ait-Hellal and E. Altman, "Performance evaluation of the rate–based flow control mechanism for ABR service: Generalization," in *Proc. of the IEEE INFOCOM'99*, (New York, USA), pp. 819–826, March 1999.

[19] K. K. Ramakrishnan, R. Jain, and D.-M. Chiu, "Congestion avoidance in computer networks with a connectionless network layer, part IV: A selective binary feedback scheme for general topologies," Tech. Rep. DEC–TR–510, Digital Equipment Corporation, 550 King St., Littleton, MA 01460, August 1987.

[20] A. Charny, D. D. Clark, and R. Jain, "Congestion control with explicit rate indication," in *Proc. of the IEEE ICC'95*, (Seattle, WA), pp. 1954–63, IEEE, June 1995.

[21] L. Roberts, "Enhanced PRCA (proportional rate-control algorithm)," Tech. Rep. 94-0735R1, ATM Forum, ATM Systems, Foster City, CA, August 1994.

[22] R. Jain, S. Fahmy, S. Kalynaraman, and R. Goyal, "The ERICA switch algorithm for ABR traffic management in ATM networks, part I: Description," *submitted to the IEEE/ACM Trans. on Networking*, January 1997.

[23] N. Ghani and J. W. Mark, "An enhanced distributed explicit rate allocation algorithm for ABR services," in *In 15th Int. Teletraffic Congress (ITC15)*, (Washington D.C.), IEEE, June 1997.

[24] S. P. Abraham and A. Kumar, "Max–min fair rate control of ABR connections with nonzero MCRs," in *Proc. of the IEEE GLOBECOM'97*, (Phoenix, AZ, USA), pp. 498–502, November 1997.

[25] S. P. Abraham and A. Kumar, "A stochastic approximation approach for max–min fair adaptive rate control of ABR sessions with MCRs," in *Proc. of the IEEE INFOCOM'98*, (San Francisco, California, USA), 1998.

[26] S. P. Abraham and A. Kumar, "A simulation study of an adaptive distributed algorithm for max–min fair rate control of ABR sessions," in *Proc. of the CCBR'98*, (Ottawa, Canada), 1998.

[27] D. Bertsekas and R. Gallager, *Data Networks.* Prentice Hall, 2 ed., 1992.

[28] J. M. Jaffe, "Bottleneck flow control," *IEEE Trans. on Communications*, vol. COM-29, July 1981.

[29] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, 1998.

[30] J. Mo and J. Walrand, "Fair end–to–end window–based congestion control," Tech. Rep. UCB/ERL Report M98/46, University of California at Berkeley, 1998. (submitted to the IEEE/ACM Trans. on Networking; an abbreviated version appeared in Int'l Symp. on Voice, Video and Data Communications (SPIE'98), October 1998.).

[31] Y. Zhao, S. qi Li, and S. Sigarto, "An improved EFCI scheme with early congestion detection," Tech. Rep. af-94-0682, ATM Forum, Department of ECE, University of Texas at Austin, March 1996.

[32] L. F. Shampine and M. W. Reichelt, "The matlab ODE suite," *SIAM Journal on Scientific Computing*, vol. 18, no. 1, 1997.

[33] M. Junius, M. Steppler, N. Bauter, and D. Pesch, "CNCL - Communication Networks Class Library 1.10 edt.." Aachen University of Technology, D-52056 Aachen, Germany, 1996. `e-mail: cncl-adm@comnets.rwth-aachen.de`.

[34] Y. Chang, N. Golmie, and D. Su, "Study of interoperability between EFCI and ER switch mechanisms for ABR traffic in an ATM network," *Computer Communications*, vol. 19, pp. 653–658, 1996.

[35] Y.-C. Lai and Y.-D. Lin, "Interoperability of EFCI and ER switches for ABR services in ATM networks," *IEEE Network*, pp. 34–42, January/February 1998.

[36] N. T. Plotkin and J. J. Sydir, "The rate mismatch problem in heterogeneous ABR flow control," in *Proc. of the INFOCOM'97*, (Kobe, Japan), 1997.

[37] S. Manthorpe and J.-Y. Le Boudec, "A comparison of ABR and UBR to support TCP traffic," Tech. Rep. DI 97/224, Ecole Polytechnique Federal de Lausanne, Deparment d'Informatique, Laboratoire de Reseaux de Communication, Lausanne, Switzerland, April 1997.

[38] J. J. Sydir, N. Taft-Plotkin, and N. Akar, "Using ATM services for (in)efficient support of TCP," Tech. Rep. ITAD-1617-TR-98-113, SRI International, Menlo Park, CA 94025, September 1998.

[39] Z. Cao and E. W. Zegura, "ABR service for applications with non-linear bandwidth utility functions," in *Proc. of the IEEE ICNP'97*, October 1997.

[40] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis, "Integration of pricing and flow control for available bit rate service in ATM networks," in *Proc. of the IEEE GLOBECOM'96*, (London, UK), 1996.

[41] Z. Cao and E. W. Zegura, "Utility max–min: An application–oriented bandwidth allocation scheme," in *Proc. of the IEEE INFOCOM'99*, vol. 2, (New York, USA), pp. 793–801, March 1999.