Performance Evaluation of the Rate-Based flow control Mechanism for ABR service:

Generalization

Omar Ait-Hellal Eitan Altman
INRIA B.P. 93
Projet Mistral
2004 Route des Lucioles
06902 Sophia Antipolis Cedex
FRANCE
E-mail: {oaithel, altman}@sophia.inria.fr

Abstract

In this paper we investigate the performances of the EFCI-based (Explicit Forward Congestion Indication) and ER-based (Explicit Rate) algorithms for the rate-based flow control of the ABR (Available Bit Rate) traffic in an ATM network. We consider the case of multiple switches in tandem. We present several definitions of a bottleneck, and provide conditions that determine which queue is the bottleneck. We show that it is not necessarily the queue with the slowest transmission rate that is "responsible" for a bottleneck. We derive analytic formulas for the maximum queue length. We compare our results to those obtained by approximating a network by a simpler one, containing only the bottleneck switch. We show that the maximum queue lengths under the approximating approach may largely underestimate the ones obtained in the real network.

Keywords

ATM, Available Bit Rate, bottleneck, closed-loop congestion control, differential equations, fluid approximation.

1 Introduction

Adaptive mechanisms for congestion control have a central role in the efficient sharing of the network resources between many users. These mechanisms have also the role of preventing congestion in the network [2, 6]. The fact, however, that the control is performed by the sources (that are not policed by the network), and not by the network, make it hard to protect the network from applications that might not use such mechanisms (e.g. from video conferences that use UDP).

In ATM networks, the ABR (Available Bit Rate) service [1] has been defined for supporting best effort applications, in which the control decisions are taken by the network in the switches, ensuring to achieve fairness among the active connections and controlling the loss cell ratios. This service manages the bandwidth leftover by applications that have guaranteed performance (VBR and CBR), and shares it between the ABR sources by signaling to them their allowable transmission rate. The behavior of the source and destination is specified in [1] as well as the manner in which feedback information should be conveyed back to the source [1, 10]. The behavior of the switches, however, is left to the designer of the switch.

Several controllers have been proposed for the switches. They are either based on (i) the EFCI (Explicit Forward Congestion Indication) bit which originates from the approach of the DEC bit [14]; it indicates, whether the congestion is detected or not, or (ii) on ER (Explicit Rate) which informs the source on the bandwidth that is available (see next section and [4, 5, 8, 9, 13, 16] and references therein). The different control mechanisms include many parameters that influence the performances and the Quality of Services of the ABR connections. Some of these are fixed, such as the buffer size, the round trip times and the available bandwidth. Others may be negotiated at the session establishment. It is therefore crucial to have methods that allow the network to choose those parameters as a function of the fixed parameters and the quality of service which the network wishes to provide (in particular, the Cell Loss Ratio).

In order to propose such methods, simple queueing models have been developed and analyzed in the past years. In particular, models have been developed based on the simplifying assumption that the network can be modeled by a single bottleneck queue [3, 12, 15, 17]. Bounds for the buffer occupancy are then computed for EFCI and ER-based switches as functions of the parameters. This approach enables then to come up with the values of the negotiated parameters so as to avoid queueing overflow.

The main purpose of our paper is to study more accurately the case when the connection actually goes across more than one switch; we examine then validity of the simplified approach of modeling the system by a single bottleneck queue.

It is easy to see that if the first switch on the path of a virtual circuit has a lower bandwidth than the others, then queueing will not occur in downstream switches, and therefore the approach of a single

bottleneck can be used. We therefore restrict our study to the opposite case; we consider multiple (N) switches along the path of a virtual circuit, where the bandwidth available for the ABR traffic in the ith switch is larger than in the $(i+1)$th one.

We consider both EFCI and ER-based switches. The first objective of the paper is to obtain better understanding of the bottleneck phenomenon. We show that it is not necessarily the queue with the slowest transmission rate that is "responsible" for a bottleneck. We present and examine, in particular, several alternative definitions of a bottleneck queue. We obtain conditions under which a given queue is the bottleneck. The second objective of the paper is a quantitative one: obtain bounds for the buffer requirements in the all the queues as a function of the fixed and the negotiated parameters. This permits, in particular, to compute the control parameters that are required in order to have no losses in the case that the buffer size is given a priori.

We restrict our analytical study to the case where the Nth buffer never empties (after some initial time). This regime corresponds to a full utilization of the network, which is a desirable operation condition.

In comparing the lower bounds on the maximum queue size, to the maximum queue length obtained by simulating the simplified single-queue model [12, 15], we see there are cases in which the simplified model is not applicable, and large differences appear.

We present both a transient analysis, in order to obtain bounds on the queues which are uniform for all time, as well as a steady state analysis, which provides smaller bounds. The steady state bounds allow us to obtain cell loss rates that are negligible in the long run using smaller buffers, at the price of having losses in some initial period.

Another reason for which the transient behavior might be of interest, is that the queue responsible for the bottleneck might change in time, as shown by simulation.

The paper is structured as follows: in section 2, we describe the rate-based congestion control mechanism, and the model. The main analytic results are presented in Section 3. The exact calculations are delayed to the Appendix, in Section A. In Section 4, numerical results obtained by using our analytic formulas are compared to simulations.

2 Description and the model

In the following section we briefly describe the behavior of the *Source End System* (SES) [10], the *Destination End System* (DES) and the way control information are conveyed back to the source. The source sends data cells at rate no more than its Allowed Cell Rate (ACR) which varies according to the congestion state of the network. At a connection setup, an Initial Cell Rate (ICR), a Minimum Cell Rate (MCR, which we assume 0 in this paper) and a Peak Cell Rate (PCR) are negotiated. The source begins

to send with a rate ICR, and its ACR may vary between MCR and PCR.

The ATM forum *Traffic Management Specification,* Version 4.0 [1], has specified the structure of the RM (Resource Management) cells which are sent by the source and make the round trip between the source, destination (forward RM) and back to the source (backward RM). An RM cell is sent every N_{rm} data cells. According to the degree of the network congestion, the switches may alter the content of RM cells in the two directions (forward or backward). At the arrival of a backward RM cell, the source (SES) adjusts its rate according to the *congestion indication* (CI) bit and the *Explicit Rate* field, of the RM cell.

If $CI = 1$ then the ACR shall be reduced, multiplicatively, by $N_{rm} * ACR/RDF$ down to MCR, where RDF is called the *Reduction Decrease Factor*. However if $CI = 0$, the ACR shall be increased by no more than $N_{rm} * AIR$ but not beyond the PCR, where AIR is called *Additive Increase Rate* (we assume that the NI bit is not used, i.e. its value is zero, as is the case in many switches). If the value of the ER field in the received RM cell is lower than the current ACR and higher than MCR, then ACR is set to this value. The CI bit and the ER field are updated by the switches in the following manner depending on the switch architecture:

- *EFCI-based switch*:
  If the queue length exceeds a certain threshold Q_H (in our case each switch has a different threshold Q_i $i = 1, \cdots, N$, see Figure 1), the switch sets the EFCI bit in the header of the data cells, until the queue length goes bellow the threshold (the switch is then considered to be not congested)[13, 15]. When the DES receives a forward RM cell, it sets the CI bit to the EFCI of the last received data cell. The switch may also set the bit CI of backward RM cells if congestion is detected, which we consider in this paper.

- *ER-based switch*:
  ER-based control has an intelligent marking and a capability to estimate the available bandwidth which permits to reduce, selectively, the rates of ABR sources by setting the CI bit or by updating the ER field in forward and/or backward RM cells. In this paper, we consider that backward and forward RM cells are updated by the ER-based switch (the case where only forward RM cells are updated for EFCI-based switch as well as for ER-based switch, can be seen to perform worst, since the information about the congestion arrives later. The queue lengths are then larger. The qualitative results, however, remain the same).

  During a congestion, there are different ways to signal ERs (Fair share, load factor, load adjustment factor... [4, 7, 8, 13]). As in [15], the method called *fair share* (FS) is considered in this paper. When no congestion is detected, the ACR is increased in the same manner as for EFCI-switch. If congestion is detected (queue length exceeds Q_H), the switch computes a FS depending on the

available bandwidth (bottleneck rate) and on the MCRs of active connections. In our case, each switch computes its FS ($FS_i,\ i = 1, \cdots, N$) if congestion is detected. If all switches are congested, the value of the ER field is set to the $min(FS_1, \cdots, FS_N)$, which we consider equals to FS_N. The FS_i is given by $FSi = \mu_i - MCR$ where μ_i is the available bandwidth at the switch i. This FS_i multiplied by an *Explicit Reduction Factor* ($ERF_1 = ERF_2 = \cdots = ERF_N$) smaller than 1 is added to the MCR, and the resulting rate is, then, signaled to the source.

$$ER_i = MCR + ERF * FS_i \tag{1}$$

Hence, we define after considering $MCR_i = 0\ \forall i$

$$ER_i \triangleq ERF * \mu_i$$

There are timers for recovering from *starvation* situations where ACR is zero for a long time (generally 100 ms). The source, in this case, is allowed to send RM cells called *out-of-rate*, at a rate no more than $TCR = 10\ cells/s$ with lower priority (see [10] for more details). Such mechanisms are not taken into account in the present paper.

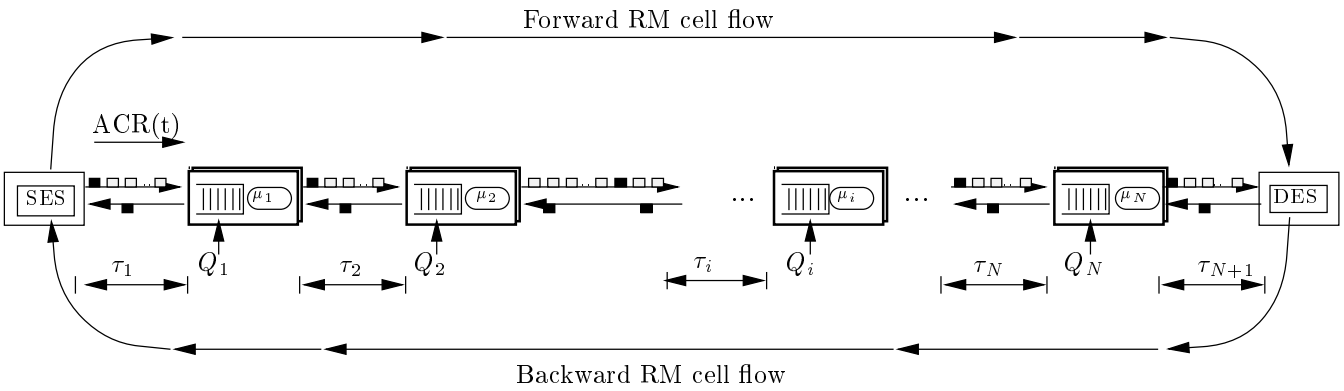


Figure 1: The Model with multiple bottlenecks

We consider a saturated ABR source (SES) (i.e. it has always cells to transmit) sending to a sink (DES). The source and the destination are separated by multiple (N) switches, with available bandwidths of $\mu_1, \cdots, \mu_N$, respectively. We assume that $PCR \triangleq \mu_0 > \mu_1 > \mu_2 > \cdots > \mu_N > MCR$. We assume that the buffer sizes are sufficiently large so that no losses occur (we compute below sizes of the buffers that guarantee no losses). Data cells and RM cells go through the N switches before they reach the destination (Figure 1). The round trip time is τ (i.e. the time that is necessary for an RM cell to reach the destination and to return back to the source in an empty system). Cells spend τ_1 from the source to reach the first switch and, generally, τ_i between the $(i-1)$th switch and the ith switch as depicted in

figure 1. The time that is necessary for an RM cell to go from the last switch to the destination and back to the source, in an empty system, is $\tau' \triangleq \tau - \sum_{\ell=1}^{N} \tau_\ell$.

The variation of the allowed cell rate at the source ($ACR(t)$) is cyclic. We define a cycle as the time separating the two moments when $ACR(t)$ is increasing and equals to μ_N, these moments are refered in what follows by $T_N^n - \sum_{\ell=1}^{N} \tau_\ell$ and $T_N^{n+1} - \sum_{\ell=1}^{N} \tau_\ell$

Define :

Q_i is the queue threshold for the ith switch.

$T_i^n \triangleq$ Time at which queue i starts to build, in the nth cycle (i.e. $ACR(T_i^n - \sum_{\ell=1}^{i} \tau_\ell) = \mu_i$).

$Q_i^n \triangleq$ The queue length in the ith switch at time T_i^n, $Q_i^0 = 0 \ \forall i$.

$Q_{max}^{i,n} \triangleq$ The maximum queue length at the ith switch in the nth cycle.

$Q_i(t) \triangleq$ The queue length at the ith switch at time t.

$T_{Q_i}^n \triangleq$ The time at which queue i reaches Q_i in the nth cycle, while increasing, if no congestion is detected.

$T_0^{n,i} \triangleq$ Time at which queue i empties in the n th cycle. By convention we consider that $T_0^{n,0} = T_{Q_N}^n + \sum_{\ell=1}^{N} \tau_\ell$.

$T_{Q_i'}^n \triangleq$ The time at which queue i reaches Q_i in the nth cycle, while decreasing.

By convention, we consider that the available bandwith at the source is PCR (that is the source can't send more than PCR), hence we define $ACR(T_0^n) = \mu_0 = PCR$.

We assume that after the first cycle, queue N never empties (i.e. full utilization. We establish later, the conditions for which this assumption holds: see Theorem 3.3 equation (4) and equation (10) in Theorem 3.4) and the congestion is known by the source $\sum_{\ell=1}^{i} \tau_\ell$ after it was detected by the ith switch (i.e. time between backward RM cells is neglected). In practice, it may take longer from the time that the congestion occurs, till the source is informed about it; the extra delay is bounded by the time between two consecutive RM cells, i.e. $(N_{rm} - 1)/\mu_N$. For the transient analysis, we assume, also, that $ICR = \mu_N$. (It is not reasonable to negotiate an ICR larger than the available bandwidth; $ICR = \mu_N$ thus corresponds to the largest initial burst.)

The following properties hold:

- **P1)** If queue i is not empty at time t then queue $i+1$ is not empty at time $t + \tau_{i+1}$.
- **P2)** If the queue length at the ith switch exceeds the threshold Q_i, in the nth cycle, then queue $i-1$ will be empty at the end of that cycle (in fact all queues $l \leq i-1$ will be empty at the end of

that cycle).

The first property is obvious, since if queue i is not empty at time t then the input rate at switch $i+1$ at time $t+\tau_{i+1}$ equals to $\mu_i > \mu_{i+1}$. For the second property, if queue i exceeds the threshold Q_i and queue $i-1$ is nonempty, then queue i continues to increase. Thus it decreases (at time t) only if queue $i-1$ empties (at time $t-\tau_i$), which should happen (after some delay) since $ACR(t)$ decreases as long as queue i exceeds Q_i. From property (**P1**) it follows that all queues $l < i-1$ will be empty. We shall focus, mainly, on the case when queue N is the bottleneck (see definition 3.1).

3 Main results

In this section we present the results obtained for the transient analysis and the steady state analysis. The transient analysis will provide for bounds on the maximum queue lengths Q_i^t in switches i, that will be uniform for all time. Thus, if the actual buffer size is larger than the computed maximum queue length, there will be no losses at any time. The steady state analysis will provide bounds for the case that the queue is initially in its steady state. We thus obtain negligible losses at the long run for buffer sizes that are larger than the computed maximum queue sizes Q_i^s (but the buffer sizes may be smaller than the ones computed for the transient case).

Supported by simulations, we assume that a steady state periodic regime indeed exists, and is reached in finite time. The maximum length of queue i at steady state is then $Q_i^s \stackrel{\Delta}{=} \overline{\lim}_{n\to\infty}(Q_{max}^{i,n})$, $i = 1, \cdots, N$.

Definition 3.1. i) *Queue i is the strict bottleneck if and only if queue j ($j \neq i$) remains always empty (from the properties above, only queue N can be the strict bottleneck).*
ii) *Queue i is the essential bottleneck if and only if queue j ($j \neq i$) never reaches the threshold Q_j.*
iii) *We say that congestion is always due to the ith switch if and only if queue i reaches its threshold Q_i, not later than $\sum_{\ell=1}^{i}\tau_\ell - \sum_{\ell=1}^{j}\tau_\ell$ before queue j ($\forall j \neq i$) reaches its Q_j in all cycles (the first RM cell indicating congestion is updated by switch i: $T_{Q_i}^n + \sum_{\ell=1}^{i}\tau_\ell \leq T_{Q_j}^n + \sum_{\ell=1}^{j}\tau_\ell$; $\forall n\ \forall j \neq i$).*

Theorem 3.1. *When the congestion is always due to the Nth switch, the maximum queue sizes are obtained in the first cycle:*

$$Q_i^t \stackrel{\Delta}{=} \max_{n\geq 0}(Q_{max}^{i,n}) = Q_{max}^{i,0}$$

Proof:

We first note that $Q_i^n = 0\ \forall n$ for $i < N$. This is due to property **P2**. For all n we have $Q_i^n \geq Q_i^0$, since we assume that the queue i is initially empty. We now make use of the following monotone property:

Assume that $Q_i^n \geq Q_i^m$ for some n and m. Then the time that queue i takes to reach Q_i in the nth cycle is shorter than the time it takes to reach Q_i in the mth cycle. It now follows that the ACR in cycle n at time $T_{Q_i}^n - \sum_{\ell=1}^{i} \tau_\ell + t$ is smaller than the ACR at cycle m at time $T_{Q_i}^m - \sum_{\ell=1}^{i} \tau_\ell + t$ for $t \geq 0$. It then follows that the size of the ith queue in cycle n after time $T_{Q_i}^n + t$ is smaller than in cycle m at time $T_{Q_i}^m + t$, $t \geq 0$. The proof follows by taking $m = 0$ and arbitrary n. ∎

Remark 3.1. *In the proof of Theorem 3.1 we established the following property: If $Q_i^n \geq Q_i^m$ for some n and m then $Q_{max}^{i,n} \leq Q_{max}^{i,m}, i = 1, \cdots, N$.*

Define $A_{i,n}$, $B_{i,n}$, $C_{i,n}$ as

$$A_{i,n} \triangleq \begin{cases} \sqrt{\dfrac{2(Q_i - Q_i^n)}{AIR * \mu_N}} + \dfrac{\mu_i - \mu_N}{AIR * \mu_N} & \text{if } \dfrac{\mu_{i-1} - \mu_i}{AIR * \mu_N} \geq \sqrt{\dfrac{2(Q_i - Q_i^n)}{AIR * \mu_N}} \\ \dfrac{Q_i - Q_i^n}{\mu_{i-1} - \mu_i} + \dfrac{\mu_{i-1} + \mu_i - 2\mu_N}{2 * AIR * \mu_N} & \text{otherwise} \end{cases}$$

$$B_{i,n} \triangleq \left(2 \sum_{\ell=1}^{N} \tau_\ell + A_{N,n} - \frac{\mu_{i-1} - \mu_N}{AIR * \mu_N} \right)^+ \quad \text{and} \quad C_{i,n} \triangleq min \left(2 \sum_{\ell=1}^{N} \tau_\ell + A_{N,n} - \frac{\mu_{i-1} - \mu_N}{AIR * \mu_N}, \frac{\mu_{i-1} - \mu_i}{AIR * \mu_N} \right)$$

where $(x)^+ \triangleq max(0, x)$. Q_i^n is given in Theorem 3.3 and Theorem 3.4.

3.1 Results for the Transient case

Theorem 3.2. *For both EFCI-based switch and ER-based switch, we have*

i) Queue i starts to build first, if and only if

$$\frac{\mu_i - \mu_j}{AIR * \mu_N} \leq \sum_{\ell=i+1}^{N} \tau_\ell - \sum_{\ell=j+1}^{N} \tau_\ell \ , \ \forall j \neq i.$$

Particular case: $\tau_i = \tau_j = \tau_1 \ \forall i, j$, we have queue i starts to build first if and only if

$$\frac{\mu_i - \mu_j}{AIR * \mu_N} \leq (j - i)\, \tau_1, \quad \forall j \neq i$$

ii) The following are equivalent:

- *(ii.a) The congestion is always due to the ith switch.*
- *(ii.b)* $\sum_{\ell=j+1}^{N} \tau_\ell + \sum_{\ell=1}^{i} \tau_\ell - \sum_{\ell=1}^{j} \tau_\ell - \sum_{\ell=i+1}^{N} \tau_\ell \leq A_{j,n} - A_{i,n}, \ \forall n, \quad j < N.$

Particular case: for switch N, the following are equivalent:

- *(ii.a.1) The congestion is always due to the switch N.*
- *(ii.b.1)* $2\sum_{\ell=j+1}^{N}\tau_\ell \le A_{j,n} - A_{N,n},\ \forall n, \quad j < N,.$
- *(ii.c.1)* $2\sum_{\ell=j+1}^{N}\tau_\ell \le A_{j,0} - A_{N,0}, \quad \forall j < N$

Theorem 3.3. *If the congestion is always due to the Nth switch, then we have*
i) For both EFCI and ER-based switch, the following are equivalent:

- *(i.a) Queue N is the strict bottleneck*
- *(i.b)* $2\sum_{i=1}^{N}\tau_\ell \le \frac{\mu_i - \mu_N}{AIR * \mu_N} - A_{N,n}\ ;\ \forall n,\ i < N.$
- *(i.c)* $2\sum_{i=1}^{N}\tau_\ell \le \frac{\mu_{N-1} - \mu_N}{AIR * \mu_N} - A_{N,0}$

i.1) and the following also are equivalent

- *(i.1.a) All the queues build up*
- *(i.1.b)* $2\sum_{i=1}^{N}\tau_\ell > \frac{\mu_1 - \mu_N}{AIR * \mu_N} - A_{N,0}$

ii) For the EFCI-based switch, if queue N is not the strict bottleneck, then the maximum queue length at switch i (Q_i^t which is obtained in the first cycle, $Q_i^0 = 0\ \forall i$) is given by

$$\begin{aligned} Q_1^t &= \frac{AIR * \mu_N}{2}\ {C_{1,0}}^2 + (\mu_0 - \mu_1)\, B_{1,0} + RDF * AIR * C_{1,0} \\ &\quad - \frac{RDF * \mu_1}{\mu_N}\log\left(1 + \frac{AIR * \mu_N}{\mu_1} C_{1,0}\right) \end{aligned} \tag{2}$$

and for $i > 1$, we have

$$\begin{aligned} Q_i^t &\cong \frac{AIR * \mu_N}{2} {C_{i,0}}^2 + (\mu_{i-1} - \mu_i)\left[B_{i,0} + \frac{RDF}{\mu_N}\log\left(1 + \frac{AIR * \mu_N}{\mu_1} C_{1,0}\right)\right. \\ &\quad \left. + \sum_{j=1}^{i-1}\left(\frac{Q_j^t}{\mu_j} + \frac{RDF(\mu_j - \mu_{j+1})}{\mu_N \mu_j}\right)\right] \end{aligned} \tag{3}$$

The following bounds hold (for the remaining cells in queue N at the end of a given cycle):

$$\left(Q_N - \mu_N\left(\tau + \frac{1}{2*AIR}\right)\right)^+ \leq Q_N^n \leq \left(Q_N + RDF\left(1 - e^{-\frac{\mu_N}{RDF}\tau}\right)\right.$$
$$\left. - \frac{\mu_N}{2*AIR}\left(2*AIR*\tau + \left(1 - e^{-\frac{\mu_N}{RDF}\tau}\right)^2\right)\right)^+ \quad (4)$$

iii) For the ER-based switch, the maximum queue length at the switch i (Q_i^t) is given by

$$Q_i^t = \frac{AIR*\mu_N}{2}\,{C_{i,0}}^2 + (\mu_{i-1} - \mu_i)\left(\sum_{\ell=1}^{i-1}\frac{Q_\ell^t}{\mu_\ell - ER_N} + B_{i,0}\right) \quad (5)$$

iv) Queue N is the essential bottleneck (ER-based switch) if

$$\sum_{\ell=1}^{i-1}\frac{Q_\ell^t}{\mu_\ell - ER_N} < \sqrt{\frac{2Q_i}{AIR*\mu_N}} - B_{i+1,0} \quad \forall i < N$$

3.2 Results for the steady state case

In what follows, we shall restrict to the case (in steady state) where congestion is always due to the Nth switch. Since we assume steady state, this is less restrictive than the corresponding assumption for the transient behavior. Indeed, the steady state behavior is a periodic regime obtained after some transient period, so congestion is always due to the Nth switch if and only if $T_{Q_N}^n + \sum_{\ell=1}^{N}\tau_\ell \leq T_{Q_i}^n + \sum_{\ell=1}^{i}\tau_\ell$ for some n and $\forall i < N$. The above restriction applies in almost all simulations that we performed, in the case where the Nth queue never empties. We almost always obtained the following behavior: after a finite transient period, the steady state is reached, and the congestion *in steady state* is always due to the Ntn switch.

Theorem 3.4. *If queue N is not the strict bottleneck and the congestion is always due to the Nth switch, then*

i) For the EFCI-based switch, the maximum queue length at switch i (Q_i^s) is given by

$$Q_1^s = Q_1^\infty + \frac{AIR*\mu_N}{2}\,{C_{1,\infty}}^2 + (\mu_0 - \mu_1)\,B_{1,\infty} + RDF*AIR*C_{1,\infty}$$
$$- \frac{RDF*\mu_1}{\mu_N}\log\left(1 + \frac{AIR*\mu_N}{\mu_1}C_{1,\infty}\right) \quad (6)$$

and for $i > 1$ we have

$$Q_i^s \cong Q_i^\infty + \frac{AIR*\mu_N}{2}{C_{i,\infty}}^2 + (\mu_{i-1} - \mu_i)\left[B_{i,\infty} + \frac{RDF}{\mu_N}\log\left(1 + \frac{AIR*\mu_N}{\mu_1}C_{1,\infty}\right)\right.$$
$$\left. + \sum_{j=1}^{i-1}\left(\frac{Q_j^s}{\mu_j} + \frac{RDF(\mu_j - \mu_{j+1})}{\mu_N\mu_j}\right)\right] \quad (7)$$

$$\textit{Where} \quad Q_i^\infty = 0 \ \forall i < N \quad \textit{and} \quad Q_N^\infty = \left(Q_N - \mu_N \left(\tau + \frac{1}{2 * AIR} \right) \right)^+ \tag{8}$$

ii) For the ER-based switch, the maximum queue length at the ith switch (Q_i^s) is given by

$$Q_i^s \quad = \quad Q_i^\infty + \frac{AIR * \mu_N}{2} \ {C_{i,\infty}}^2 + (\mu_{i-1} - \mu_i) \left(\sum_{\ell=1}^{i-1} \frac{Q_\ell^s}{\mu_\ell - ER_N} + B_{i,\infty} \right) \tag{9}$$

$$\textit{Where} \quad Q_i^\infty = 0 \ \forall i < N \quad \textit{and} \quad Q_N^\infty = \left(Q_N - (\mu_N - ER_N) \left(\tau + \frac{\mu_N - ER_N}{2 * AIR * \mu_N} \right) \right)^+ \tag{10}$$

4 Numerical results and discussions

In this section we present some numerical examples, in order to validate our analytical results with simulation ones, and compare them to the approximating model in which the network is replaced by a single bottleneck node (exactly the same parameters as for our model without the $N-1$ first switches (switches 1 to $N-1$ are deleted)). Only the transient state (first cycle) is taken into account, since the maximum queue lengths are obtained in this case (see Theorem 3.1). We compute the queue length (Q_i^t) in each switch as a function of $2\sum_{\ell=1}^{N} \tau_\ell$.

As we have seen (Theorem 3.3), the maximum queue length in the switches during the first cycle, is not a function of the round trip time τ. This is because switches have the capability to alter the content of backward RM cells (i.e. when congestion is detected, also RM cells in backward direction are updated). However, in the steady state, the maximum queue lengths are function of both τ and $2\sum_{\ell=1}^{N} \tau_\ell$ since, queue N can be nonempty at the end of the cycle ($Q_N^n > 0$ which is function of τ, see equations (8) and (10)). The sufficient conditions that allow a full utilization, which is a desirable operation, can be obtained, for both ER-based and EFCI-based switches, from euqations (8) and (10).

Remark 4.1. *As discussed in section 2, the time between RM cells is neglected in the paper. However, this time will have some effects on the actual results. To take this time into account, we added the quantity $(N_{rm}-1)/\mu_N$ to $A_{i,0}$, when computing analytically the queue lengths (ana-q_i in the figures), since when the congestion occurs at switch i, it is detected by the source at most $\sum_{\ell=1}^{i} \tau_\ell$ after.*

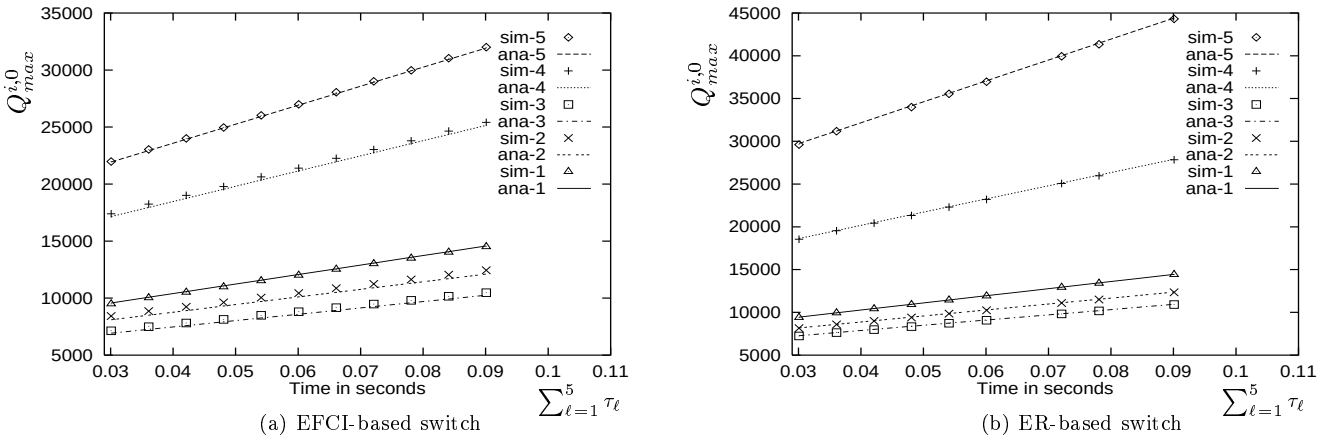
In all the following simulations, we considered $PCR = 141.33\ Mbps = 333333.25\ cells/sec$ and hence the transmission time of a cell is about 3 μsec, $MCR = 0.0\ Mbps$ $AIR = 0.1\ Mbps \cong 250\ cells/sec$, $RDF = 512$ and $ERF = 0.8$. We will take different theresholds for each queue, in order to show the effect of homogeneous queues on the maximum queue length at the Nth switch.

In the first example we consider five switches with the following thresholds: $Q_1 = 20000$, $Q_2 = 15000$, $Q_3 = 10000$, $Q_4 = 10000$ and $Q_5 = 5000$. The available bandwith in each switch are respectively

$\mu_1 = 106\ Mbps = 250000\ cells/sec$, $\mu_2 = 84.8\ Mbps = 200000\ cells/sec$, $\mu_3 = 70.66\ Mbps = 166666.67\ cells\,\text{sec}$, $\mu_4 = 42.4\ Mbps = 100000\ cells/sec$ and $\mu_5 = 21.2\ Mbps = 50000\ cells/sec$, then we have $A_{1,0} = 0.2593$, $A_{2,0} = 0.314$, $A_{3,0} = 0.3106$, $A_{4,0} = 0.1566$ and $A_{5,0} = 0.102$. From Theorem 3.2 we have, the congestion is always due to queue N if and only if $2\sum_{\ell=j+1}^{5}\tau_\ell \leq A_{j,0} - A_{N,0}$, $\forall j < 5$, which is equivalent to

$$\tau_5 \leq 0.0273sec\ , \quad \tau_4 \leq 0.1043 - \tau_5 sec\ , \quad \tau_3 \leq 0.106 - \tau_4 - \tau_5 sec\ , \quad \tau_2 \leq 0.078667 - \tau_3 - \tau_4 - \tau_5 sec$$

On the other hand, from Theorem 3.3 *(i.1)*, we must have $\sum_{\ell=1}^{5}\tau_\ell > -0.0794$ which means that all the queues build up. We choose the following initial values for τ_i $i = 2 \cdots 5$: $\tau_5 = 2\ ms$, $\tau_4 = 2\ ms$, $\tau_3 = 4\ ms$ and $\tau_2 = 4ms$. When we compare the analytical results to the simulation ones as function of the round trip time, we increase only τ_1 (from 3 ms to 24 ms by a step of 3 ms, see figure 2) since the queue lenghts are function of $\sum_{\ell=1}^{N}\tau_\ell$.



(a) EFCI-based switch (b) ER-based switch

Figure 2: Queue length as a function of the round trip time

References

[1] The ATM forum Technical Committee, *Traffic Management Specification,* Version 4.0, April 1996.

[2] L.S. Brakmo, L.L. Peterson, TCP Vegas : End to End Congestion Avoidance on a Global Internet," *IEEE Journal on selected Areas in communications,* vol. 13, pp. 1465-1480, 1995.

[3] M. O. Brouillet and U. Madhow, "Rate control for adaptive bit rate sources on ATM networks using one bit congestion notification", Research report UILU-ENG-96-2214, University of Illinois at Urbana-Champaign, May, 1996.

[4] Y. Chang, N. Golmie, D. Siu, "A Rate-Based Flow Control Switch Design for ABR Service in an ATM Network," *Twelfth International Conference on Computer Communication ICCC'95,* August 1995.

[5] A. Charny, D.D. Clark, R.Jain, "Congestion Control With Explicit Rate Indication,"*Proc. ICC'95,* June 1995, 10 pp.

[6] Van Jacobson, "Congestion avoidance and control," *ACM SIGCOMM 88,* pages 273-288, 1988. ftp.ee.lbl.gov/papers/congavoid.ps.Z

[7] R. Jain, S. Kalyanaraman, R. Goyal, S. Fahmy, R. Viswanathan, "ERICA switch algorithm: A Complete Description," *ATM Forum/96-1172.*

[8] R. Jain, S. Kalyanaraman, R. Viswanathan, "The OSU Scheme for Congestion Avoidance in ATM Networks using Explicit Rate Indication," *Proceedings WATM'95 First Workshop on ATM Traffic Management,* Paris, December 1995.

[9] R. Jain, "Congestion Control and Traffic Management in ATM Networks: Recent Advances and A Survey," *invited submission to Computer Networks and ISDN Systems.*

[10] R. Jain, S. Kalyanaraman, R. Goyal, S. Fahmy, "Source Behavior for ATM ABR Traffic Management: An Explanation," *IEEE Communications Magazine,* November, 1996.

[11] S. Keshav, "REAL: A network simulator," *Department of Computer Science, UC Berkeley,* Technical Report 88/472, 1988.

[12] H. Ohsaki, G. Hasegawa, M. Murata and A.Miyahara, "Parameter Tuning of Rate-Based Congestion Control Algorithms and its Application to TCP over ABR," *Proceedings WATM'95 First Workshop on ATM Traffic Management,* Paris, December 1995.

[13] H. Ohsaki, M. Murata, H. Suzuki, C. Ikeda and H. Miyahara, "Rate-based control for ATM Networks," *Computer Communication Review, acm-sigcomm,* special issue in ATM, ed. R. Jain and K.Y. Siu, Vol 25, No. 2, pp. 60-71, 1995.

[14] K. Ramakrishnan and R. Jain, "A Binary Feedback Scheme for Congestion Avoidance in Computer Networks with Connectionless Network Layer," *Proc. SIGCOMM'88,* August 1988.

[15] Michael Ritter, "Network Buffer Requirements of the Rate-Based Control Mechanism for ABR services," *IEEE INFOCOM'96,* san francisco, USA, March 1996.

[16] K.Y. Siu and H.Y. Tzeng, "Intelligent Congestion Control for ABR Service in ATM Networks," *Computer Communication Review, acm-sigcomm,* Vol 24, No. 5, pp. 81-106, 1995.

[17] N. Yin and M. G. Hluchyj, "On closed-loop rate control for ATM cell relay networks", *Proc. IEEE Infocom'94*, pp. 99-108, June 1994.

A Appendix: Analysis and proof of Theorems

A.1 Evolution of the ACR

The evolution of the $ACR(t)$ follows two phases in our case ($ICR = \mu_N$, queue N never empties and RM cells have no priority), depending on whether the congestion is detected or not.

Phase 1: No congestion and queue N nonempty

This phase concerns (i) the first cycle ($n = 0$) which begins after the first RM cell returns back; we consider, then, $ACR(T_N^0 - \sum_{\ell=1}^{N} \tau_\ell) = ACR(\tau) = \mu_N$, and (ii) all other cycles while no congestion is detected. The evolution of the $ACR(t)$ in this phase is given by

$$\frac{dACR(t)}{dt} = AIR * \mu_N$$

$$\text{from which we get} \qquad ACR(t) - ACR(t_0) = AIR * \mu_N (t - t_0) \tag{11}$$

t_0 is the beginning instant of this phase which ends once an RM cell with bit CI set is received.

Phase 2: Congestion is detected

When congestion is detected, ACR(t) is reduced multiplicatively down to MCR and its evolution is given by

$$\frac{dACR(t)}{dt} = -\frac{\mu_N}{RDF} ACR(t)$$

$$\text{from which we get} \qquad ACR(t) = ACR(t_1) e^{-\frac{\mu_N}{RDF}(t - t_1)} \tag{12}$$

t_1 is the beginning instant of this phase, which corresponds to the ending time of phase 1.

A.2 Transient Analysis

First we compute T_i^n, the time when queue i starts to build. We have by definition

$$ACR\left(T_i^n - \sum_{\ell=1}^{i} \tau_\ell\right) = \mu_i$$

Since queue i is assumed always nonempty, the evolution of the $ACR(t)$ follows equation (11). Hence

$$ACR\left(T_i^n - \sum_{\ell=1}^{i} \tau_\ell\right) = \mu_N + AIR * \mu_N \left(T_i^n - T_N^n + \sum_{\ell=i+1}^{N} \tau_\ell\right) = \mu_i,$$

$$\text{from which we get} \qquad T_i^n - T_N^n = \frac{\mu_i - \mu_N}{AIR * \mu_N} - \sum_{\ell=i+1}^{N} \tau_\ell \tag{13}$$

Queue i starts to build up first iff $T_i^n - T_j^n < 0, \forall j \neq i$, hence from (13) we get Theorem 3.2 (i).

RM cells with a CI bit cleared are sent by the switches till one of the queues reaches the threshold Q_i ($T_{Q_i}^n$) which we compute next as if there is no congestion (only equation (11) is used). For $T_{Q_i}^n$, we distinguish two cases:

1) $T_{Q_i}^n - \tau_i \leq T_{i-1}^n$

$$Q_i - Q_i^n = \int_0^{T_{Q_i}^n - T_i^n} AIR * \mu_N * x \; dx = \frac{1}{2} AIR * \mu_N (T_{Q_i}^n - T_i^n)^2$$

$$\text{from which we get} \qquad T_{Q_i}^n = \sqrt{\frac{2(Q_i - Q_i^n)}{AIR * \mu_N}} + T_i^n \tag{14}$$

Case (1) is equivalent to $\frac{\mu_{i-1} - \mu_i}{AIR * \mu_N} \geq \sqrt{\frac{2(Q_i - Q_i^n)}{AIR * \mu_N}}$ as can be seen by combining (13) and (14).

2) In other cases, we have

$$Q_i - Q_i^n = \int_0^{T_{i-1}^n - T_i^n + \tau_i} AIR * \mu_N * x \; dx + (\mu_{i-1} - \mu_i)(T_{Q_i}^n + \tau_i - T_{i-1}^n)$$

from which we get, after substituting (13)

$$T_{Q_i}^n = \frac{Q_i - Q_i^n}{\mu_{i-1} - \mu_i} + \frac{\mu_{i-1} - \mu_i}{2 * AIR * \mu_N} + T_i^n \tag{15}$$

The congestion is always due to the ith switch if and only if $T_{Q_i}^n + \sum_{\ell=1}^{i} \tau_\ell \leq T_{Q_j}^n + \sum_{\ell=1}^{j} \tau_\ell, \; \forall j \neq i$, from which the equivalence between (ii.a) and (ii.b) in part (ii) of Theorem 3.2 follows. This is obtained after substituting (13) and remarking that $T_{Q_i}^n = A_{i,n} - \frac{\mu_i - \mu_N}{AIR * \mu_N} + T_i^n$.

The particular case is obtained by replacing i by N. Now we prove the equivalence between (ii.b.1) and (ii.c.1).

(ii.b.1) implies (ii.c.1) follows trivially, since the condition should be satisfied for all cycles.

Now, we prove that $2 \sum_{\ell=j+1}^{N} \tau_\ell \leq A_{j,0} - A_{N,0} \; \forall j \leq N \Longrightarrow 2 \sum_{\ell=j+1}^{N} \tau_\ell \leq A_{j,n} - A_{N,n} \; \forall n, j$.

Since $Q_N^n \geq 0 = Q_N^0$ for all integers n, we have $A_{N,0} \geq A_{N,n} \; \forall n$. $2 \sum_{\ell=j+1}^{N} \tau_\ell \leq A_{j,0} - A_{N,0} \quad \forall j < N$, implies that the congestion is due to the Nth switch in the first cycle, i.e. queue N reaches its threshold Q_N, not later than $\sum_{\ell=j+1}^{N} \tau_\ell$ before queue j reaches its Q_j for all $j < N$. Hence $Q_j^1 = 0, \; \forall j < N$ (property **P2**). Then it follows that $A_{j,1} = A_{j,0}, \; \forall j < N$, so that $2 \sum_{\ell=j+1}^{N} \tau_\ell \leq A_{j,1} - A_{N,0} \; \forall j < N$. Combining this with the fact that $A_{N,0} \geq A_{N,n} \; \forall n$, we have $2 \sum_{\ell=j+1}^{N} \tau_\ell \leq A_{j,1} - A_{N,1}$ which means that the congestion is due to switch N in the second cycle. Repeating this argument we obtain (ii.b.1).

We shall assume

- (**H1**) The congestion is always due to the switch N ($T^n_{Q_N} + \sum_{\ell=1}^{N} \tau_\ell \leq T^n_{Q_j} + \sum_{\ell=1}^{j} \tau_\ell \ \forall j < N$)

Queue N is the strict bottleneck if and only if

$$T^n_{Q_N} + \sum_{\ell=1}^{N} \tau_\ell \leq T^n_i - \sum_{\ell=1}^{i} \tau_\ell, \ \forall i < N \tag{16}$$

from which we get after substituting (13)

$$2\sum_{\ell=1}^{N} \tau_\ell \leq \frac{\mu_i - \mu_N}{AIR * \mu_N} - A_{N,n} \quad \forall i < N.$$

Since $A_{N,n} < A_{N,0}, \ \forall n$ and the condition should be satisfied in the first cycle, then the relation above is equivalent to

$$2\sum_{\ell=1}^{N} \tau_\ell \leq \frac{\mu_i - \mu_N}{AIR * \mu_N} - A_{N,0} \quad \forall i < N.$$

By remarking that $\mu_0 < \mu_1 < \cdots < \mu_{N-1} < \mu_N$ and the relation above, again, should be satisfied for the first queue $i = 1$, we get the equivalence between *(i.b)* and *(i.c)* in part *(i)* of theorem 3.3.

The part *(i.1)* is obtained by using the same arguments and inverting the inequalities in (16).

For the maximum queue length in each switch, we consider both ER-based and EFCI-based switches. Let $T_P^{n,i} \triangleq min(T^n_{i-1} - \sum_{\ell=1}^{i-1} \tau_\ell, \ T^n_{Q_N} + \sum_{\ell=1}^{N} \tau_\ell)$, and $ACR^n_{max} \triangleq min\left(PCR, \ ACR(T^n_{Q_N} + \sum_{\ell=1}^{N} \tau_\ell)\right)$

We assume, also, that all the queues build up (*(i.1)* in theorem 3.3 is true).

A.2.1 EFCI-based switch

The first RM cell with CI bit set to one returns back to the source at time $T^n_{Q_N} + \sum_{\ell=1}^{N} \tau_\ell$ (backward RM cells are updated when congestion is detected), from this time the evolution of the $ACR(t)$ corresponds to phase 2, and then, the instant when $ACR(t) = \mu_i$, in the nth cycle, while decreasing, is given by

$$T^n_{\mu_i} = \frac{RDF}{\mu_N} \log\left(\frac{ACR^n_{max}}{\mu_i}\right) + T^n_{Q_N} + \sum_{\ell=1}^{i} \tau_\ell$$

Define, also, $T_M^{n,i} \triangleq max(T^n_{\mu_i}, \ T_0^{n,i-1} - \sum_{\ell=1}^{i-1} \tau_\ell)$

Since the congestion is always due to queue N (**H1**), then

$$Q^{i,n}_{max} = Q^n_i + \int_{T^n_i - \sum_{\ell=1}^{i} \tau_\ell}^{T_P^{n,i}} (ACR(x) - \mu_i)\, dx + \int_{T_P^{n,i}}^{T_0^{n,i-1} - \sum_{\ell=1}^{i-1} \tau_\ell} (\mu_{i-1} - \mu_i)\, dx$$

$$
\begin{aligned}
&+\int_{T_0^{n,i-1}-\sum_{\ell=1}^{i-1}\tau_\ell}^{T_M^{n,i}} (ACR(x)-\mu_i)\ dx \\
= \quad & Q_i^n + \int_0^{T_P^{n,i}-T_i^n+\sum_{\ell=1}^{i}\tau_\ell} AIR*\mu_N*x\ dx + (\mu_{i-1}-\mu_i)\left(T_0^{n,i-1}-\sum_{\ell=1}^{i-1}\tau_\ell - T_P^{n,i}\right) \\
&+\int_0^{T_M^{n,i}-T_0^{n,i-1}+\sum_{\ell=1}^{i-1}\tau_\ell} \left(ACR(T_0^{n,i-1}-\sum_{\ell=1}^{i-1}\tau_\ell)e^{-\frac{\mu_N}{RDF}x}-\mu_i\right)\ dx \\
= \quad & Q_i^n + \frac{AIR*\mu_N}{2}\left(T_P^{n,i}-T_i^n+\sum_{\ell=1}^{i}\tau_\ell\right)^2 + (\mu_{i-1}-\mu_i)\left(T_0^{n,i-1}-\sum_{\ell=1}^{i-1}\tau_\ell - T_P^{n,i}\right) \\
&+\int_{T_0^{n,i-1}-\sum_{\ell=1}^{i-1}\tau_\ell}^{T_M^{n,i}} (ACR(x)-\mu_i)\ dx
\end{aligned} \tag{17}
$$

Now we need to compute $T_0^{n,i}$. We have

$$
Q_{max}^{i,n} - Q_i^{n+1} = \int_{T_M^{n,i}}^{T_0^{n,i}-\sum_{\ell=1}^{i}\tau_\ell} (\mu_i - ACR(x))\ dx = \int_0^{T_0^{n,i}-\sum_{\ell=1}^{i}-T_M^{n,i}} \left(\mu_i - ACR(T_M^{n,i})\ e^{-\frac{\mu_N}{RDF}x}\right)\ dx
$$

Let $y_i \triangleq T_0^{n,i} - \sum_{\ell=1}^{i} - T_M^{n,i}$, then we have

$$
Q_{max}^{i,n} - Q_i^{n+1} = \mu_i y_i + \frac{RDF}{\mu_N} ACR(T_M^{n,i})\left(e^{-\frac{\mu_N}{RDF}y_i}-1\right) \tag{18}
$$

Combining equation (18) with equation (17), and after some calculations we get ($Q_i^N = 0\ \forall i < N$, and $C_{i,n} \triangleq T_P^{n,i} - T_i^n + \sum_{\ell=1}^{i}\tau_\ell$)

$$
\begin{aligned}
Q_{max}^{i,n} \quad = \quad & Q_i^n + \frac{AIR*\mu_N}{2} {C_{i,n}}^2 + \int_{T_0^{n,i-1}-\sum_{\ell=1}^{i-1}\tau_\ell}^{T_M^{n,i}} (ACR(x)-\mu_i)\ dx \\
& + (\mu_{i-1}-\mu_i)\left[(T_0^{n,0}-T_P^{n.i}) + \sum_{j=1}^{i-1}\left(\frac{Q_{max}^{j,n}}{\mu_j} + (T_M^{n,j} - T_0^{n,j-1} + \sum_{\ell=1}^{j-1}\tau_\ell)\right.\right. \\
& \left.\left. + \frac{RDF*ACR(T_M^{n,j})}{\mu_N*\mu_j}\left(1-e^{-\frac{\mu_N}{RDF}\left(T_0^{n,j}-\sum_{\ell=1}^{j}\tau_\ell - T_M^{n,j}\right)}\right)\right)\right]
\end{aligned} \tag{19}
$$

Recal that $T_0^{n,0} = T_{Q_N}^n + \sum_{\ell=1}^{N}\tau_\ell$. Now we distinguish between the case $i = 1$ and $i > 1$. Since $Q_i^n = 0\ \forall n,\ i < N$, then we have we have, after remarking that $B_{i,n} \triangleq T_{Q_N}^n + \sum_{\ell=1}^{N}\tau_\ell - T_P^{n,i}$

$$
Q_{max}^{1,n} \quad = \quad Q_1^n + \frac{AIR*\mu_N}{2} {C_{1,n}}^2 + \int_{T_{Q_N}^n+\sum_{\ell=1}^{N}\tau_\ell}^{T_{\mu_1}^n} (ACR(x)-\mu_i)\ dx + (\mu_0-\mu_1)\left(T_{Q_N}^n + \sum_{\ell=1}^{N}\tau_\ell - T_P^{n.1}\right)
$$

$$
\begin{aligned}
&= Q_1^n + \frac{AIR * \mu_N}{2} {C_{1,n}}^2 + (\mu_0 - \mu_1)\, B_{1,n} + RDF * AIR * C_{1,n} \\
&\quad - \frac{RDF * \mu_1}{\mu_N} \log\left(1 + \frac{AIR * \mu_N}{\mu_1} C_{1,n}\right) \qquad (20)
\end{aligned}
$$

from which we get equation (2) in theorem 3.3. For $i > 1$, the integral in the first line of equation (19) can be neglected since $T_M^{n,i} \cong T_0^{n,i-1} - \sum_{\ell=1}^{i-1} \tau_\ell \quad \forall i > 1$. Hence from (2), we have

$$
\begin{aligned}
Q_{max}^{i,n} &\cong Q_i^n + \frac{AIR * \mu_N}{2} {C_{i,n}}^2 + (\mu_{i-1} - \mu_i) \left[(T_M^{n,1} - T_P^{n.i}) \right. \\
&\quad \left. + \sum_{j=1}^{i-1} \left(\frac{Q_{max}^{j,n}}{\mu_j} + \frac{RDF * ACR(T_M^{n,j})}{\mu_N * \mu_j} \left(1 - e^{-\frac{\mu_N}{RDF}\left(T_0^{n,j} - \sum_{\ell=1}^{j} \tau_\ell - T_M^{n,j}\right)}\right)\right)\right] \\
&\cong Q_i^n + \frac{AIR * \mu_N}{2} {C_{i,n}}^2 + (\mu_{i-1} - \mu_i) \left[(T_{\mu_1}^n - T_P^{n.i}) \right. \\
&\quad \left. + \sum_{j=1}^{i-1} \left(\frac{Q_{max}^{j,n}}{\mu_j} + \frac{RDF}{\mu_N} \left(e^{-\frac{\mu_N}{RDF}\left(T_M^{n,j} - T_{\mu_j}^n\right)} - e^{-\frac{\mu_N}{RDF}\left(T_0^{n,j} - \sum_{\ell=1}^{j} \tau_\ell - T_{\mu_j}^n\right)} \right)\right)\right] \\
&\cong Q_i^n + \frac{AIR * \mu_N}{2} {C_{i,n}}^2 + (\mu_{i-1} - \mu_i) \left[\left(\frac{RDF}{\mu_N} \log\left(\frac{ACR_{max}^n}{\mu_1}\right) + T_{Q_N}^n + \sum_{\ell=1}^{N} \tau_\ell - T_P^{n.i} \right) \right. \\
&\quad \left. + \sum_{j=1}^{i-1} \left(\frac{Q_{max}^{j,n}}{\mu_j} + \frac{RDF}{\mu_N} \left(1 - e^{-\frac{\mu_N}{RDF}\left(T_{\mu_{j+1}}^n - T_{\mu_j}^n\right)}\right)\right)\right] \qquad (21)
\end{aligned}
$$

From which we get equation (3) in theorem 3.3. The number of remaining cells in queue N at the end of the cycle,is given by

$$
\begin{aligned}
Q_N^{n+1} &= Q_N + \int_{T_{Q'_N}^n - \sum_{\ell=1}^{N} \tau_\ell}^{T_{Q'_N}^n - \sum_{\ell=1}^{N} \tau_\ell + \tau} (ACR(x) - \mu_N)\, dx + \int_{T_{Q'_N}^n - \sum_{\ell=1}^{N} \tau_\ell + \tau}^{T_N^{n+1} - \sum_{\ell=1}^{N} \tau_\ell} (ACR(x) - \mu_N)\, dx \\
&= Q_N + ACR(T_{Q'_N}^n - \sum_{\ell=1}^{N} \tau_\ell) \frac{RDF}{\mu_N} \left(1 - e^{-\frac{\mu_N}{RDF}\tau}\right) \\
&\quad - \mu_N * \tau - \frac{\left(\mu_N - ACR(T_{Q'_N}^n - \sum_{\ell=1}^{N} \tau_\ell)\; e^{-\frac{\mu_N}{RDF}\tau}\right)^2}{2 * AIR * \mu_N} \qquad (22)
\end{aligned}
$$

Inequalities (4) in Theorem 3.3 are derived since $0 \leq ACR(T_{Q'_N}^n - \sum_{\ell=1}^{N} \tau_\ell) \leq \mu_N$.

A.2.2 ER-based switch

For ER-based switch once the congestion is detected, equation (1) is used. $ACR(t)$ is then set to $ER_N < \mu_N$. The maximum queue length at switch i is reached either $\sum_{\ell=1}^{i} \tau_\ell$ after an RM cell with a CI bit set

to one has arrived to the source or τ_i after queue $i-1$ empties. Since we assumed that all queues build up then it is the second case which happens. Hence we have

$$\begin{aligned} Q_{max}^{i,n} &= Q_i^n + \int_{T_i^n - \sum_{\ell=1}^{i}\tau_\ell}^{T_P^{n,i}} (ACR(x) - \mu_i)\, dx + \int_{T_P^{n,i}}^{T_0^{n,i-1} - \sum_{\ell=1}^{i-1}\tau_\ell} (\mu_{i-1} - \mu_i)\, dx \\ &= Q_i^n + \int_0^{T_P^{n,i} - T_i^n + \sum_{\ell=1}^{i}\tau_\ell} AIR * \mu_N * x\, dx + (\mu_{i-1} - \mu_i)(T_0^{n,i-1} - \sum_{\ell=1}^{i-1}\tau_\ell - T_P^{n,i}) \\ &= Q_i^n + \frac{AIR * \mu_N}{2}\left(T_P^{n,i} - T_i^n + \sum_{\ell=1}^{i}\tau_\ell\right)^2 + (\mu_{i-1} - \mu_i)\,(T_0^{n,i-1} - \sum_{\ell=1}^{i-1}\tau_\ell - T_P^{n,i}) \end{aligned} \tag{23}$$

$T_0^{n,i}$ satisfies the following:

$$\begin{aligned} Q_{max}^{i,n} - Q_i^{n+1} &= \int_{T_0^{n,i-1} - \sum_{\ell=1}^{i-1}\tau_\ell}^{T_0^{n,i} - \sum_{\ell=1}^{i}\tau_\ell} (\mu_i - ER_N) dx \\ &= (\mu_i - ER_N)(T_0^{n,i} - T_0^{n,i-1} - \tau_i) \end{aligned} \tag{24}$$

$$T_0^{n,i} = \frac{Q_{max}^{i,n} - Q_i^{n+1}}{\mu_i - ER_N} + T_0^{n,i-1} + \tau_i \tag{25}$$

Since $T_0^{n,0} = T_{Q_N}^n + \sum_{\ell=1}^{N}$, then from (25), we get

$$T_0^{n,i} = \sum_{\ell=1}^{i}\left(\frac{Q_{max}^{\ell,n} - Q_\ell^{n+1}}{\mu_\ell - ER_N} + \tau_\ell\right) + T_{Q_N}^n + \sum_{\ell=1}^{N}\tau_\ell \tag{26}$$

Since $Q_i^n = 0,\ \forall i < N$ and Q_i^t is obtained in the first cycle, then after substituting $C_{i,n}$ and $B_{i,n}$, already defined, we get equation (5) in Theorem 3.3.

Queue N is the essential bottleneck iff $Q_{max}^{i,n} < Q_i\ \forall i < N$ equivalent to $Q_{max}^{i,0} < Q_i\ \forall i < N$. From (23), since $T_P^{i,n} \le T_0^{n,i-1} - \sum_{\ell=1}^{i-1}$, we have

$$Q_{max}^{i,n} \le Q_i^n + \frac{AIR * \mu_N}{2}\left(\sum_{\ell=1}^{i-1}\frac{Q_{max}^{\ell,n}}{\mu_\ell - ER_N} + B_{i+1,n}\right)^2$$

then the sufficient condition for that $Q_{max}^{i,n} < Q_i\ \forall n,\ i < N$ is

$$Q_{max}^{i,n} \le \frac{AIR * \mu_N}{2}\left(\sum_{\ell=1}^{i-1}\frac{Q_{max}^{\ell,n}}{\mu_\ell - ER_N} + B_{i+1,n}\right)^2 < Q_i\ \forall n,\ i < N$$

From which we get (iv) in Theorem 3.3, after remarking that the maximum queue size is obtained during the first cycle ($Q_i^0 = 0,\ \forall i$).

In case of ER-based switch, $Q_N^{n+1} = Q_N^n \ \forall n \geq 1 \ (Q_i^n = 0, \ \forall n, i < N)$; as we can see bellow, this is due to the fact that $ACR(t) = ER_N$ is constant, once the congestion is detected and it doesn't depend on its maximum value during a cycle. When queue N falls bellow Q_N, the source is informed $\sum_{\ell=1}^{N} \tau_\ell + 2\tau_{N+1}$ (the switches can't update the CI bit in case of non congestion if that bit was set to one by the destination) after, by receiving an RM cell with a bit CI cleared, and increases its rate according to phase one. Hence the time ellapsed since the queue size drops bellow Q_N till queue N starts to build again, is given by $T_N^{n+1} - T_{Q'_N}^n = \tau + \frac{\mu_N - ER_N}{AIR * \mu_N}$ (see eq. (11). Hence

$$Q_N^{n+1} = Q_N + (ER_N - \mu_N) \ \tau + \int_0^{\frac{\mu_N - ER_N}{AIR * \mu_N}} ((ER_N - \mu_N) + AIR * \mu_N * x) \ dx, \tag{27}$$

A.3 Steady state analysis

The conclusions that we drew for the transient analysis held for $Q_i^n = 0$, but the formulas that we developed hold for any initial queue lengths, and in particular, to the ones obtained in steady state, in the case of a periodic behavior. (A periodic behavior is indeed obtained for the ER switch, as the queue lengths become constant in steady state. For EFCI, we have observed a periodic behavior in most of the simulations).

In what follows, we do not assume that the steady state has a periodic behavior. For EFCI switches, we use the formulas developed for the transient behavior in order to compute bounds on the maximum of the queue lengths. We obtain below an asymptotic upper bound Q_i^s on the maximum length of queue i by using equations (20) and(21), in which we substitute Q_N^n by an asymptotic *lower bound* of Q_N^n:

$$Q_N^\infty \stackrel{\Delta}{=} \underline{\lim}_{n \to \infty} (Q_N^n)$$

and taking $Q_i^n = 0 \ \forall i < N$. The fact that a lower bounds on Q_N^n yields an upper bound on the maximum queue length follows from the arguments in the proof of Theorem 3.1 and from Remark 3.1. Similarly, one may obtain an asymptotic lower bound on the maximum length of the queues by substituting in (20) and (21) an asymptotic upper bound for Q_N^n.

A.3.1 EFCI-based switch

From (22), we have

$$Q_N^\infty \quad = \quad \underline{\lim}_{n \to \infty} \left(Q_N + ACR \left(T_{Q'_N}^n - \sum_{\ell=1}^{N} \tau_\ell \right) \frac{RDF}{\mu_N} \left(1 - e^{-\frac{\mu_N}{RDF} \tau} \right) \right.$$

$$-\mu_N * \tau - \frac{\left(\mu_N - ACR\left(T^n_{Q'_N} - \sum_{\ell=1}^{N} \tau_\ell\right)\ e^{-\frac{\mu_N}{RDF}\tau}\right)^2}{2 * AIR * \mu_N}\Bigg)$$

Since $\underline{\lim}_{n\to\infty} ACR\left(T^n_{Q'_N} - \sum_{\ell=1}^{N} \tau_\ell\right) = \underline{\lim}_{n\to\infty} \mu_N\ e^{-\frac{\mu_N}{RDF}\left(T^n_{\mu_N} - T^n_{Q'_N} + \sum_{\ell=1}^{N} \tau_\ell\right)} = 0$

and $Q_N^\infty \geq 0$, we obtain equation (6) from equation (20) and equation (7) from (21). The proof for Theorem 3.4 *(i)* is then established.

A.3.2 ER-based switch

For an ER-based switch we have $Q_N^n = Q_N^{n+1}\ \forall n \geq 1$. The steady state regime is, then, reached just after the first cycle. Hence, since $Q_N^\infty \geq 0$, we have from (27)

$$Q_N^\infty = \underline{\lim}_{n\to\infty}\left(Q_N^n\right) = \left(Q_N - \left(\mu_N - ER_N\right)\left(\tau + \frac{\mu_N - ER_N}{2 * AIR * \mu_N}\right)\right)^+.$$

The proof of Theorem 3.4 *(ii)* follows after substituting Q_N^∞ and (26) in (23).