Convergence of trajectories and optimal buffer sizing for AIMD congestion control

Konstantin Avrachenkov[1], Urtzi Ayesta[2,3], Alexei Piunovskiy[4]

[1] INRIA, 2004 route des Lucioles, 06902 Sophia Antipolis, France
[2] BCAM, Basque Center for Applied Mathematics, 48160 Derio, Spain
[3] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
[4] University of Liverpool, Department of Mathematics, L69 7ZL Liverpool, UK

Abstract

We study the interaction between the AIMD (Additive Increase Multiplicative Decrease) multi-socket congestion control and a bottleneck router with Drop Tail buffer. We consider the problem in the framework of deterministic hybrid models. First, we show that trajectories always converge to limiting cycles. We characterize the cycles. Necessary and sufficient conditions for the absence of multiple jumps in the same cycle are obtained. Then, we propose an analytical framework for the optimal choice of the router buffer size. We formulate this problem as a multi-criteria optimization problem, in which the Lagrange function corresponds to a linear combination of the average goodput and the average delay in the queue. Our analytical results are confirmed by simulations performed with MATLAB Simulink.

1 Introduction

Most traffic in the Internet is governed by TCP/IP (Transmission Control Protocol and Internet Protocol) [1, 15]. TCP protocol tries to adjust the sending rate of a source to match the available bandwidth along the path. During the principal Congestion Avoidance phase the current TCP New Reno version uses AIMD (Additive Increase Multiplicative Decrease) binary feedback congestion control scheme. In the absence of congestion signals from the network TCP increases congestion window linearly in time, and upon the reception of a congestion signal TCP reduces the congestion window by a multiplicative factor. Congestion signals can be either packet losses or ECN (Explicit Congestion Notifications) [24]. At the present state of the Internet, nearly all congestion signals are generated by packet losses. Packets can be dropped either when the router buffer is full or when AQM (Active Queue Management) scheme is employed [11]. Given an ambiguity in the choice of the AQM parameters [8, 18], so far AQM is rarely used in practice. On the other hand, in the basic Drop Tail routers, the buffer size is the only one parameter to tune apart from the router capacity. In fact, the buffer size is one of few parameters of the TCP/IP network that can be managed by network operators. This makes the choice of the router buffer size a very important consideration in the TCP/IP network design. A significant increase of link capacities has posed a challenge to the current TCP implementation. The current TCP New Reno version is not able to utilise efficiently high speed links. To mitigate this problem several new TCP versions have been proposed (an extensive overview and comparison of different TCP version for high capacity links is given in [17]). One possible solution, which is also simple, is to use multiple TCP sockets in parallel [2, 9]. This approach is implemented in the GridFTP protocol [13] and is used in Grid computing projects such as Atlas (atlas.ch), EU DataGrid (eu-datagrid.web.cern.ch) and Globus (www.globus.org).

The paper is composed of two main parts. In the first part (Sections 2-5) we analyze the interaction between the multi-socket AIMD congestion control and the bottleneck router with Drop Tail buffer. This interaction can adequately be described by a hybrid modelling approach. There are several hybrid models of the interaction between TCP and the bottleneck router [5, 7, 14]. Here we analyze the model of [14]. In our opinion, this model takes into account all essential details of TCP and at the same time leads to a tractable analysis. The hybrid model correctly represents the dynamics of the system at the scale of round trip time. We note that the other widely used TCP models [6, 16, 19] do not account for queueing dynamics. Since we deal with multiple TCP sockets in parallel, we need to model how congestion signals are distributed among TCP sockets. We assume that during the congestion event, congestion signals are sent to a random subset of sockets of a fixed size. By choosing the subset size we can model various degrees of synchronization. Our model includes two important particular cases: if only one socket reduces its congestion window during the congestion event (the subset is a singleton), we have the complete desynchronization case; and if all the sockets reduce their congestion windows during the congestion event (the subset coincides with the original set), we have the complete synchronization case. We would like to mention that the hybrid modelling approach is adequate for the time scale on the round trip time order. The contributions in the first part are as follows: we show that the system always converges to a limiting behavior. In particular, we demonstrate that two different limiting regimes can coexist and the convergence to one or to the other depends on the initial conditions. Then, we provide necessary and sufficient conditions for the absence of subsequent packet losses. The absence of subsequent packet losses benefits the TCP performance as well as the quality of service for end users. We note that in [14] there is no characterization of limiting regimes. Furthermore, in [14] only a sufficient condition for the absence of multiple jumps was obtained and the sufficient condition of [14] is loose for some values of the decrease factor.

In the second part of the paper (Sections 6-7) we study the optimal choice of the buffer size in the bottleneck routers. There are some empirical rules for the choice of the router buffer size. The first proposed rule of thumb for the choice of the router buffer size was to choose the buffer size equal to the BDP (Bandwidth-Delay Product) of the outgoing link [26]. This recommendation is based on very approximative considerations and it can be justified only when a router is saturated with a single long-lived TCP connection. The case of multiple competing TCP connections was studied in [3, 4]. The authors of [3] suggest that the minimal buffer size for the full system utilization should be chosen inversely proportional to the quare root of the number of competing connections. The authors of [5, 10, 12, 22] advocate that even smaller buffers are needed. There are also proposals to tune the router buffer size adaptively [21, 25]. We refer the interested reader to [28, 27] and references therein for more information on the problem of optimal choice of buffer size.

All the above mentioned works on the router buffer sizing are based on quite rough approximations and strictly speaking do not take into account the feedback nature of TCP protocol. Here we propose a mathematically solid framework to analyze the interaction of TCP with the finite buffer of an IP router. In particular, we state a criterion for the choice of the optimal buffer size in a mathematical form.

All proofs are provided in the Appendix.

2 Mathematical model

Consider n long-lived AIMD TCP connections that share a bottleneck router. Denote by $w_i(t)$ the instantaneous congestion window of connection $i = 1, 2, \dots, n$ at time $t \in [0, \infty)$. Let $x(t)$ be the amount of data in the bottleneck queue at time t, B be the size of the Drop Tail buffer, and μ be the capacity of the bottleneck router.

If $x(t) < B$, the evolution of $w_i(t)$ is given by the differential equation

$$\frac{dw_i}{dt} = \frac{m_i}{T_i + x(t)/\mu}.$$

Here T_i is the two way propagation delay and m_i is a constant. Note that $T_i + x(t)/\mu$ corresponds to the Round Trip Time (RTT) at time moment t. Consequently, the sending rate of the i-th connection is given by $\lambda_i(t) = \frac{w_i(t)}{T_i + x(t)/\mu}$.

Since we consider the multi-socket connection (or just a single information flow), we concentrate on the symmetric case $T_i \equiv T = const$, $m_i \equiv m_0 = const$. Now the total congestion window $w(t) = \sum_{i=1}^{n} w_i(t)$ satsifies equation

$$\frac{dw}{dt} = \frac{m}{T + x(t)/\mu}, \tag{1}$$

where $m = n \cdot m_0$. The total sending rate of the window based congestion control is given by

$$\lambda(t) = \sum_{i=1}^{n} \lambda_i(t) = \frac{w(t)}{T + x(t)/\mu}. \tag{2}$$

We emphasize that the time parameter t corresponds to the local time observed at the router.

When x reaches B at time t^*, i.e. $x(t^*) = B$, the buffer starts to overflow. The overflow of the buffer will be noticed by the sender only after the time delay $\delta = T + B/\mu$. Upon the reception of the congestion signal at time $t^* + \delta$, the congestion window is reduced according to $w(t^* + \delta + 0) = \beta^k w(t^* + \delta - 0)$. Usually, $k = 1$, but sometimes it is necessary to send several congestion signals in order to reduce the sending rate below the transmission capacity of the bottleneck router.

Let us assume that when $x(t^*) = B$ congestion signals are sent to $\tilde{n} \in \{1, ..., n\}$ random sockets. The parameter $\tilde{n}$ represents a degree of synchronization and it is assumed to be fixed. For example, if $\tilde{n} = 1$, only one socket reduces its congestion window during the congestion event and the sockets are completely desynchronizaed. If $\tilde{n} = n$, all the sockets are synchronized and reduce their congestion windows during the congestion event.

Under the assumption of fixed $\tilde{n}$, the total sending rate is reduced during the congestion event by the factor

$$\beta = 1 - (1 - \beta_0)\frac{\tilde{n}}{n}, \tag{3}$$

where β_0 is the window reduction factor for a single connection. In the particular case of total synchronization we have $\beta = \beta_0$, and in the particular case of no-synchronization, we have $\beta = 1 - (1 - \beta_0)/n$. We note that in practice various degrees of synchronization can take place (see e.g. [23, 29] and Example 1). In TCP New Reno version the window reduction factor β_0 is equal to one half. If n is big, in the complete synchronization case β can be close to 1.

To justify formula (3), we argue by induction with respect to $\tilde{n}$. If $\tilde{n} = 1$, then only one connection k_1 out of n, chosen uniformly, will receive the congestion signal; consequently

$\mathbb{E}[\lambda_{k_1}(t^*+\delta-0)] = \frac{\lambda(t^*+\delta-0)}{n}$. Suppose

$$\mathbb{E}\left[\mathbb{E}[\lambda_{k_{\tilde{n}}}(t^*+\delta-0)|k_1,\ldots,k_{\tilde{n}-1}]\right] = \frac{\lambda(t^*+\delta-0)}{n},$$

where $k_1,\ldots,k_{\tilde{n}-1}$ are the numbers of connections which receive the congestion signals. Then, we have

$$\mathbb{E}\left[\mathbb{E}[\lambda_{k_{\tilde{n}+1}}(t^*+\delta-0)|k_1,\ldots,k_{\tilde{n}}]\right] = \mathbb{E}\left[\frac{\lambda(t^*+\delta-0)-\sum_{i=1}^{\tilde{n}}\lambda_{k_i}}{n-\tilde{n}}\right] = \frac{\lambda(t^*+\delta-0)}{n}.$$

Consequently, we obtain

$$\frac{\mathbb{E}[\lambda(t^*+\delta+0)]}{\lambda(t^*+\delta-0)} = \frac{\mathbb{E}[\beta_0\sum_{i=1}^{\tilde{n}}\lambda_{k_i}(t^*+\delta-0)+\lambda(t^*+\delta-0)-\sum_{i=1}^{\tilde{n}}\lambda_{k_i}(t^*+\delta-0)]}{\lambda(t^*+\delta-0)}$$

$$= \frac{\lambda(t^*+\delta-0)[\beta_0\frac{\tilde{n}}{n}+1-\frac{\tilde{n}}{n}]}{\lambda(t^*+\delta-0)} = 1-(1-\beta_0)\frac{\tilde{n}}{n}.$$

Since we consider a fluid model for the data, between the instantaneous jumps of variable w, we have the following differential equation description for the evolution of the buffer content

$$\dot{x} = \begin{cases} \lambda(t)-\mu, & \text{if } 0 < x(t) < B, \text{ or} \\ & x(t)=0 \text{ and } \lambda(t)\geq\mu, \text{ or} \\ & x(t)=B \text{ and } \lambda(t)\leq\mu; \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\lambda(t)$ and $w(t)$ are given by (2) and (1).

Let us make the change of time scale according to $ds \overset{\triangle}{=} dt/(T+x(t)/\mu)$, and the change of variables: $v(s) \overset{\triangle}{=} w(t(s))/m, \quad y(s) \overset{\triangle}{=} x(t(s))/m$. The new time s can be viewed as a counter for Round Trip Times. Now the dynamics of the system between the jumps is described by equations

$$\frac{dv}{ds} = 1, \tag{5}$$

$$\frac{dy}{ds} = \begin{cases} v(s)-y(s)-q, & \text{if } 0 < y(s) < b, \text{ or} \\ & y(s)=0 \text{ and } v(s)\geq q, \text{ or} \\ & y(s)=b \text{ and } v(s)\leq q+b; \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $q=\mu T/m$ is the maximal number of packets that can be fit in the pipe per connection, in other words Bandwidth-Delay Product (BDP) in packets per connection, and $b=B/m$ is the maximal number of packets that can be fit in the router buffer per connection. Let s^* be a moment in the new time scale when component y reaches value b. Then, we have $v(s^*+1+0)=\beta^k v(s^*+1-0)$, where $k=\min\{i : \beta^i v(s^*+1-0) < b+q\}$.

Remark 1 *Because of the delay in the information propagation, the congestion window is reduced after the delay $\delta = T+B/\mu$ in the original time scale, or, equivalently, after* 1 *time unit in the new time scale s. The value of k is such that, after sending k congestion signals, the amount of data x (and y) starts to decrease.*

Example 1 (Independence assumption validation)

A critical assumption in the derivation of formula (3) is the independence of the congestion signal target socket from the values of congestion windows. We have carried out NS simulations to investigate when this assumption is reasonable.

We consider a bottleneck link of 100Mbps capacity and 20ms propagation delay. Two TCP sockets are connected to the bottleneck link with an access link of 1000Mbps capacity and 10ms propagation delay each. The packet size is 500bytes. Thus, the bandwidth delay product in the network is $(100000000 \times 0.008)/4000 = 200$ *packets. We vary the buffer length between 100 and 340 packets.*

In every congestion epoch we verify which connection reduces its congestion window. In the particular case when the buffer length is 200 packets, we obtain the next contingency table, where the first column indicates the interval of the congestion window of each socket during the congestion epoch.

	TCP 1	*TCP 2*	*TCP 1 & 2*
$[25,162][25,162]$	*750*	*765*	*169*
$[25,162][162,300]$	*542*	*576*	*219*
$[162,300][25,162]$	*707*	*765*	*300*
$[162,300][162,300]$	*231*	*253*	*123*

In this case, it can be easily checked that in 85% of the congestion epochs only one connection reduces the congestion window, thus we can assume that in this example $\tilde{n} = 1$. *We ignore the third column then, and we carry out Pearson's Chi-Square test to check the null hypothesis that which connection reduces its congestion window is independent of the particular value of the congestion window. In order to do so we first calculate the "theoretical frequency"* E_{ij} *for every cell in the table with*

$$E_{ij} = \frac{\sum_{k=1}^{2} O_{ik} \sum_{k=1}^{4} O_{kj}}{\sum_{k=1}^{4} \sum_{j=1}^{2} O_{kj}},$$

where O_{ij} *denotes the* ij *entry of the table. The value of the test-statistic is*

$$\chi^2_{sim} = \sum_{i=1}^{4} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

The number of degrees of freedom is equal to $(4-1) \times (2-1)$ *(4 and 2 being the numbers of rows and columns in the table, respectively). We then determine* $\chi^2_{0.05}$*, the value such that the probability that the* χ^2 *distribution exceed this value is precisely* 0.05. $\chi^2_{0.05} > \chi^2_{sim}$ *is interpreted as a justification for not rejecting the null hypothesis that the row variable is unrelated to the column variable.*

In our particular example of a buffer length of 200 packets, we get $\chi^2_{0.05} = 7.8147$ *and* $\chi^2_{sim} = 0.8422$*, and we can thus conclude that at the significance level of 0.05 we cannot reject the hypothesis that the congestion signal (to the first or to the second connection) is independent of the congestion window.*

In Figure 1 we plot $\chi^2_{0.05}$ *and* χ^2_{sim} *for several values of the buffer length. We note that for buffer lengths larger than 150 packets* $\chi^2_{0.05} > \chi^2_{sim}$*, which indicates that for large buffer lengths, the two variables become less related.*

Our conclusion from this example is that the target of congestion signals is weakly related to the size of congestion window. The relation becomes especially weak for large values of the buffer size. Of course, we acknowledge that the independence of the target of congestion signals from the size of congestion window is just an assumption and should be taken with care.

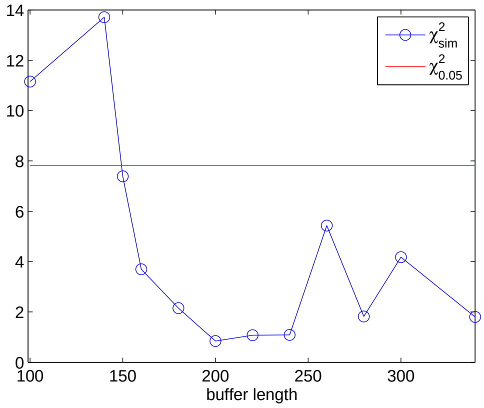

Figure 1: $\chi^2_{0.05}$ and χ^2_{sim} with respect to the buffer length

3 Convergence of the system trajectories

The dynamics are defined by three parameters β, q, and b, and a system trajectory remains in the region $\Omega = \{0 \le y \le b,\ v > 0\}$, provided the initial condition is there.

Suppose a trajectory starts at $s = 0$ from initial condition $y_0 = b$, $\beta(q+b) \le v_0 < q+b$,[1] and s^* is the first positive moment when $y(s^*) = b$. Let $v_1 = v(s^*+1+0)$. We introduce mapping φ such that $v_1 \overset{\triangle}{=} \varphi(v_0)$. Consider the iterations $v_{i+1} \overset{\triangle}{=} \varphi(v_i)$, $i = 0, 1, \ldots$.

Theorem 1 *There exists* $\lim_{i\to\infty} v_i = V(v_0)$ *with*

$$V(v) = \begin{cases} V_1, & \text{if } v \in [\beta(q+b), d]; \\ V_2, & \text{if } v \in (d, q+b), \end{cases} \tag{7}$$

for some constant d. In particular, one of the above intervals can be empty.

The detailed proofs and formulae for calculating d can be found in the Appendix.

Definition 1 *Suppose a trajectory starting at $s = 0$ from initial condition $y_0 = b$, $v_0 < b + q$ reaches the same point, for the first time, at some time moment $S \ge 1$. Then this finite trajectory is called a* cycle. *A cycle with component y remaining zero for a positive time interval is called* clipped (*see Figure 2*). *If a cycle touches the axis $y = 0$ only at a single point, we call such cycle* critical (*see Figure 3*).

Corollary 1 (from Theorem 1) *Any cycle has a single time moment, when a (multiple) jump occurs.*

The number k of instant jumps of component v is called a *cycle order*. We call such cycles k-cycles for brevity. If one of the intervals in (7) is empty then only a single cycle exists (Figure 2). Otherwise, two cycles exist simultaneously (Figures 3-5); their orders are two subsequent positive integers. According to Theorem 1, which cycle is realized depends on the initial conditions.

[1] Initial conditions outside the region $[\beta(q+b), q+b)$ are of no interest because, after the very first (multiple) jump we have $v(s^*+1+0) \in [\beta(q+b), q+b)$.

4 Properties of cycles

In this section we characterize the shape of the cycles. In other words, for given parameters β, q and b, we would like to know if the limit cycles of the system trajectories are clipped or unclipped and what orders the cycles have. For fixed values of β and q, we define the following quantities:

$$N \triangleq \min\left\{ i \geq 1 : \ \frac{\beta^i}{1-\beta^i} < q \right\}; \tag{8}$$

$$D \triangleq \ln(1-\beta^N) + \frac{2\beta^N}{1-\beta^N}; \tag{9}$$

$$C \triangleq -\ln(1-\beta^N) - \beta^N; \tag{10}$$

θ_k is the single positive solution to equation

$$\ln\frac{\theta}{1-e^{-\theta}} + \frac{\beta^k\theta}{1-\beta^k} = q - \frac{\beta^k}{1-\beta^k}, \qquad k = N, N+1; \tag{11}$$

$$b_{0,k} \triangleq \frac{\theta_k}{1-e^{-\theta_k}} - \ln\frac{\theta_k}{1-e^{-\theta_k}} - 1. \tag{12}$$

Then, we define the set of quantities which do not depend on q:
τ_k is the single positive solution to equation

$$\frac{\tau}{1+\frac{\beta^{k-1}-\beta^k}{1-\beta^k}(\tau+1)} = 1-e^{-\tau}, \quad k = 2, 3, \ldots \tag{13}$$

$$A_k^* \triangleq \frac{\beta^{k-1}(\tau_k+1)}{1-\beta^k}; \tag{14}$$

$$q_k^* \triangleq \frac{\beta^k}{1-\beta^k}(\tau_k+1) + \ln\frac{\tau_k}{1-e^{-\tau_k}}; \tag{15}$$

It is convenient to put τ_1, A_1^* and q_1^* equal to $+\infty$. Finally, in case $q \leq D$, one has to solve equation $e^{-r}+r-1 = \beta^N(q+r+1)-q$. It has no more than two positive solutions $\underline{r} \leq \bar{r}$ which define

$$\underline{b} \triangleq e^{-\underline{r}} + \underline{r} - 1; \quad \bar{b} \triangleq e^{-\bar{r}} + \bar{r} - 1. \tag{16}$$

Note that $\underline{b} \leq \bar{b}$. If $q \leq q_{N+1}^*$ then $q \leq D$ and $\bar{b} \geq A_{N+1}^* - q$.

We note that all the above defined quantities do not depend on b. Thus, from now on we assume that β and q are fixed and we are going to describe what kind of cycles exist for different values of b. Thus, we study what effect the router buffer size has on the limiting behavior of TCP/IP. There are three cases:

Case $A_{N+1}^* < q$: If $b \in \left[0, \frac{\beta^{N-1}}{1-\beta^{N-1}} - q\right]$ then only the cycle of order N exists. In case $N = 1$, we put $\frac{\beta^0}{1-\beta^0} = +\infty$ for generality. Suppose $N > 1$. Then for $b \in \left(\frac{\beta^{N-1}}{1-\beta^{N-1}} - q, A_N^* - q\right]$ two cycles, of orders N and $N-1$ exist simultaneously. For $b \in \left(A_N^* - q, \frac{\beta^{N-2}}{1-\beta^{N-2}} - q\right]$, there exists only a single cycle of order $N-1$. And so on; for $b > A_2^* - q$, only 1-cycle exists. The N-cycle is clipped for $b \in [0, b_{0,N})$. Cycles of lower orders are unclipped for all values of b, if they exist. The N-cycle touches the v-axis at a single point iff $b = b_{0,N}$. Thus, if $b = b_{0,N}$ there exists a critical N-cycle. No critical cycles of lower orders exist.

Example 2 *Let us illustrate this with a numerical example. If we take $q = 0.9$ and $\beta = 1/2$. Then $N = 2$, $A_2^* = 1.4965$, $A_3^* = 0.3910$. If $b \in [0, 0.1]$ we have only 2-cycles; if $b \in (0.1, 0.5965]$ we have 1-cycles and 2-cycles (see Figures 3-5); and if $b > 0.5965$ we have only 1-cycles. For each $b < b_{0,2} = 0.0617$, there exists only a clipped 2-cycle (see Figure 2). As one can see on Figure 3, when $b = b_{0,2} = 0.0617$, the 2-cycle becomes critical. All figures have been generated with MATLAB Simulink.*

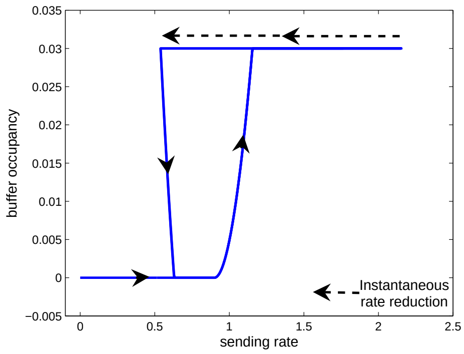

Figure 2: Phase portrait of a clipped 2-cycle. Case $A_2^* < q$.

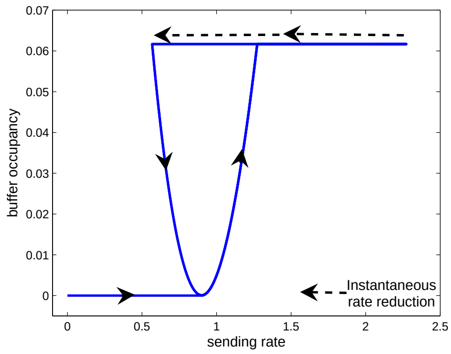

Figure 3: Phase portrait of a critical 2-cycle. Case $A_2^* < q$.

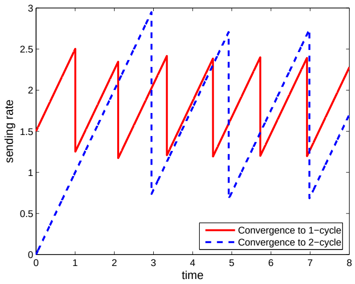

Figure 4: Sending rate versus time for the case $A_2^* < q$.

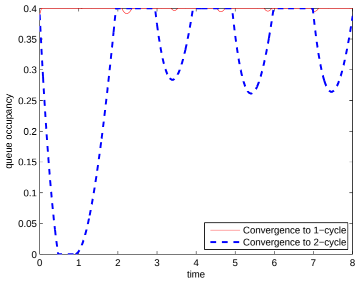

Figure 5: Buffer occupancy over time for the case $A_2^* < q$.

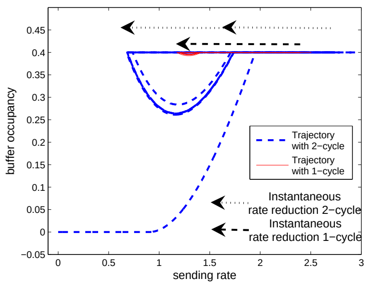

Figure 6: Phase portrait for the case $A_2^* < q$.

Case $q \leq q^*_{N+1}$: If $b \in [0, \underline{b})$, then only the N-cycle exists. If $b \in [\underline{b}, A^*_{N+1} - q]$, then two cycles of orders N and $N+1$ exist simultaneously. For $b \in (A^*_{N+1} - q, \frac{\beta^{N-1}}{1-\beta^{N-1}} - q]$, again, only the N-cycle exists. The N-cycle is clipped for $b \in [0, b_{0,N})$; the $(N+1)$-cycle is clipped for $b \in [\underline{b}, b_{0,N+1})$. These cycles become critical at $b = b_{0,N}$ and $b = b_{0,N+1}$, respectively. Cycles of lower orders are unclipped for all values of b, if they exist. If $N > 1$ then, similarly to the case $A^*_{N+1} < q$, the order of the cycle decreases as b increases above $\frac{\beta^{N-1}}{1-\beta^{N-1}} - q$.

Case $q^*_{N+1} < q \leq A^*_{N+1}$: If $C \leq A^*_{N+1} - q$ [2] and $q \leq D$, then everything is similar to the case $q \leq q^*_{N+1}$. The difference is that the $(N+1)$-cycle is clipped and cannot be critical; it exists simultaneously with the N-cycle for $b \in [\underline{b}, \bar{b}]$. If $b \in \left(\bar{b}, \frac{\beta^{N-1}}{1-\beta^{N-1}} - q\right]$, only the N-cycle exists. The latter interval is non-empty. If $C > A^*_{N+1} - q$ or $D < q$, then everything is exactly as in case $A^*_{N+1} < q$.

[2] Actually C cannot be equal to $A^*_{N+1} - q$.

5 Conditions for the absence of multiple jumps

The regime with multiple jumps is not desirable. The multiple jump corresponds to the loss of more than one packet in a single congestion window. Subsequent packet losses can force TCP to switch from the Congestion Avoidance TCP phase to the Slow Start phase and lead to lengthy timeouts. Furthermore, the absence of subsequent packet losses is beneficial not only to the TCP performance but also to the quality of service provided to the end users. In the next theorem we provide necessary and sufficient conditions for the absence of multiple jumps.

Theorem 2 *The following mutually exclusive conditions fully characterise all possible cases when only a single cycle of order* 1 *exists:*

(a) $\frac{\beta}{1-\beta} \geq q$ *and* $b + q > A_2^*$;

(b) $A_2^* < q$ *(b can be arbitrary);*

(c) $\frac{\beta}{1-\beta} < q \leq q_2^*$ *and* $b \notin [\underline{b}, A_2^* - q]$;

(d) $\max\left\{\frac{\beta}{1-\beta},\ q_2^*\right\} < q \leq A_2^* - C$, $q \leq D$ *and* $b \notin [\underline{b}, \bar{b}]$;

(e) $\max\left\{\frac{\beta}{1-\beta},\ q_2^*\right\} < q \leq A_2^* - C$, $q > D$ *(b can be arbitrary);*

(f) $\max\left\{\frac{\beta}{1-\beta},\ q_2^*,\ A_2^* - C\right\} < q \leq A_2^*$ *(b can be arbitrary).*

In the following corollary we provide a simple sufficient condition for absence of multiple jumps.

Corollary 2 *Condition* $b + q > A_2^*$ *is sufficient for the absence of cycles of orders* $k > 1$.

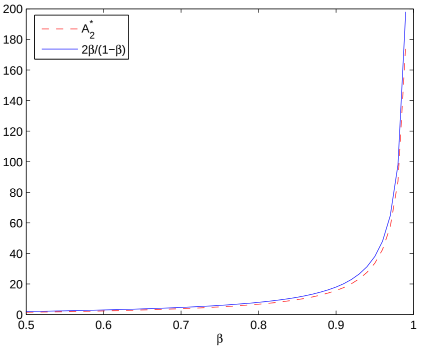

Figure 7: Comparison of sufficient conditions for absence of multiple jumps.

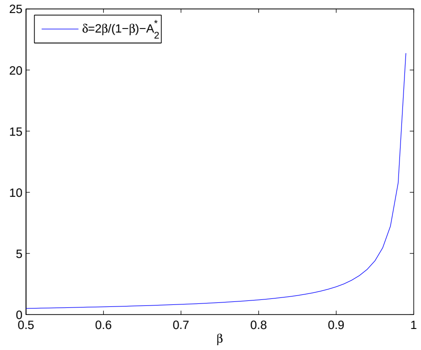

Figure 8: The value of $2\beta/(1-\beta) - A_2^*(\beta)$.

Recall that A_2^* depends only on β. In particular, if $\beta = 1/2$, $A_2^* = 1.4965$. We would like to note that the above sufficient condition is tighter than the sufficient condition for the absence of multiple jumps provided in [14]: $b+q > 2\beta/(1-\beta)$. To compare the two sufficient conditions we plot $A_2^*(\beta)$ and $2\beta/(1-\beta)$ in Figure 7 and the difference $2\beta/(1-\beta) - A_2^*(\beta)$ in Figure 8. Strictly speaking we have

Proposition 1 *The difference* $\delta \overset{\triangle}{=} \frac{2\beta}{1-\beta} - A_2^*$ *is always positive and* $\lim_{\beta\to 1}\delta = +\infty$.

Nevertheless, the simple sufficient condition of [14] appears to be quite good except for values of β that are close to one. Note that formula (3) implies that β can be close to one in the case of many asynchronous connections.

6 Pareto set for optimal buffer sizing and full system utilization

Let us study what effect has the choice of the buffer size on the performance of TCP. In particular, we are interested in the optimal buffer sizing. On one hand, we want to obtain as large goodput as possible: $\bar{g} = \lim_{t\to\infty} \frac{1}{t}\int_0^t g(s)ds \to \max$, where the instantaneous goodput $g(t)$ is defined by

$$g(t) = \begin{cases} \lambda(t), & \text{if} \quad x(t) < B, \\ \mu, & \text{if} \quad x(t) = B. \end{cases}$$

Here we concentrate on the incoming data that will go through the router (e.g. based on the first-in-first-out discipline): if $x(t) = B$ and $\lambda(t) > \mu$ then $(\lambda(t) - \mu)$ units will be lost per unit of time. One can introduce the goodput based on the outcoming data:

$$\tilde{g}(t) = \begin{cases} \lambda(t), & \text{if} \quad x(t) = 0, \\ \mu, & \text{if} \quad x(t) > 0. \end{cases}$$

If $x(t) = 0$ and $\lambda(t) < \mu$ then only $\lambda(t)$ units will be served per time unit. The average values

$$\bar{g} = \lim_{t\to\infty} \frac{1}{t}\int_0^t g(s)ds = \lim_{t\to\infty} \frac{1}{t}\int_0^t \tilde{g}(s)ds$$

coincide.

At the same time, we are interested in making the delay of data in the buffer (or, equivalently, the average amount of data in the buffer) as small as possible: $\bar{x} = \lim_{t\to\infty} \frac{1}{t}\int_0^t x(s)ds \to \min$. Clearly, these two goals are contradictory. A standard approach is to consider the optimization of one criterion under constraints for the other criteria (see e.g., [20]). Namely, one can consider problem

$$\max\{\bar{g} : \bar{x} \le \bar{x}_*\}. \tag{17}$$

or problem

$$\min\{\bar{x} : \bar{g} \ge \bar{g}_*\}. \tag{18}$$

The solution to the above constrained optimization problems can be obtained from the Pareto set. As is known, see e.g. [20], the Pareto set can be constructed by solving the optimization problem

$$\max\left\{\lim_{t\to\infty} \frac{1}{t}\int_0^t [c_1 g(s) - c_2 x(s)]ds\right\}. \tag{19}$$

To be more precise, the Pareto Set is formed by the pairs of objectives $(\bar{g}, \bar{x})$ that solve (19) for different $(c_1, c_2) \in R_+^2$. An example of Pareto set is given in Figure 11. Each point of the Pareto set corresponds to a solution of optimization problem(19) for some choice of c_1 and c_2. Once we obtain the Pareto set, it is very easy to deduce solution of the constrained problems (17) and (18). For instance, if one requires that the utilization of the bottleneck router is not less than, say, 95%, one has to be ready to accept the delays that are equal or greater than x_*.

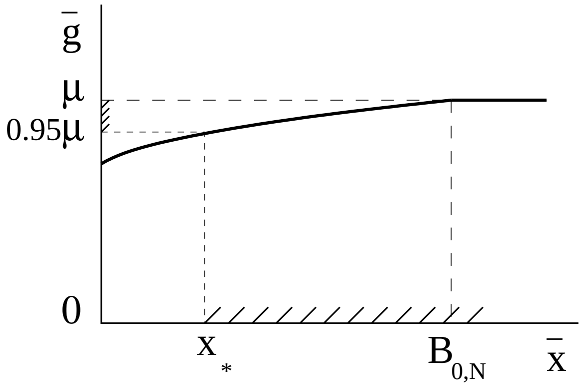

Figure 9: Pareto set.

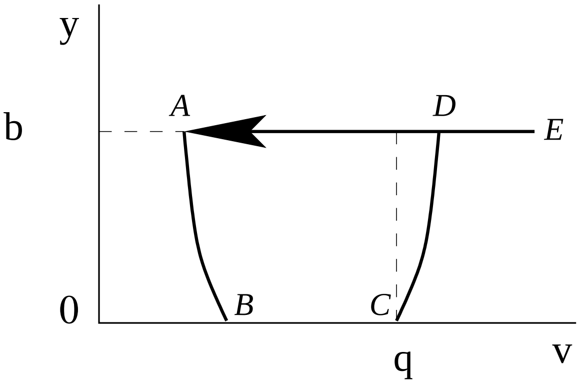

Figure 10: Phases of the clipped cycle.

All three optimization problems (17), (18) and (19) can be regarded as mathematical formulation of the lingual criterion “find the link buffer size that accommodates both TCP and UDP traffic” given in [12]. Since UDP traffic does not contribute much in terms of the load, for the design of IP routers one can use for instance optimization problem (17) where the delay constraint is imposed by the UDP traffic.

We note that here we deal with the optimal impulse control problem of a deterministic system with long-run average optimality criterion. To the best of our knowledge there are no available results on such type of problems in the literature. In principle, the control policy in our model can depend on x and λ. In practice, however, all currently implemented buffer management schemes (e.g., AQM, DropTail) send congestion signals based only on the state of the buffer. Thus, we also limit ourselves to the case when the control depends only on the amount of data in the buffer. Furthermore, we restrict the control action only to the choice of the buffer size. Thus, the control signal is only sent at the moment when the buffer gets full.

Define

$$y_{CD}(u) = e^{-u} + (u-1),$$

and

$$y_{AB}(s) = [\frac{B}{m} + (1-\beta)(1+\frac{\mu T}{m}) - \beta S_{CD}]e^{-s} + (s-1) + \beta(S_{CD}+1) - (1-\beta)\frac{\mu T}{m},$$

where S_{CD} and S_{AB} are the solutions of the equations

$$e^{-S_{CD}} + S_{CD} - 1 = \frac{B}{m},$$

$$\left[\frac{B}{m} - \beta S_{CD} + (1-\beta)(1+\frac{\mu T}{m})\right] e^{-S_{AB}} + S_{AB} + \beta S_{CD} - (1-\beta)(1+\frac{\mu T}{m}) = 0,$$

and

$$y(s) = \left[1 + \frac{\mu T + B}{m} - v_0\right] e^{-s} + (s-1) + v_0 - \frac{\mu T}{m},$$

where v_0 is given by (24) with $k = 1$.

The following theorem provides expressions for the average sending rate, goodput and queue size under condition $q > A_2^*$, which guarantees the absence of multiple jumps for any value of the buffer size. Remember that A_2^* depends only on β (see (13),(14)). In particular, the expressions allow us to plot the Pareto set parametrized by the buffer size.

Theorem 3 *Suppose $\frac{\beta}{1-\beta} < \frac{\mu T}{m}$ and let the condition $\mu T/m > A_2^*$ be satisfied. Then, for $B \in [0, mb_{0,1}]$ the average sending rate, goodput and queue size are given by*

$$\bar{\lambda} = \frac{m(1-\beta^2)}{2T_{cycle}}\left(1 + \frac{\mu T}{m} + S_{CD}\right)^2,$$

$$\bar{g} = \frac{m}{T_{cycle}}\left[\frac{1}{2}\left(\frac{\mu T}{m} + S_{CD}\right)^2 - \frac{\beta^2}{2}\left(1 + \frac{\mu T}{m} + S_{CD}\right)^2 + \frac{\mu T + B}{m}\right],$$

$$\bar{x} = \frac{1}{T_{cycle}}\left[mT\left(\int_0^{S_{AB}} y_{AB}(s)ds + \int_0^{S_{CD}} y_{CD}(u)du\right)\right.$$
$$\left.+\frac{m^2}{\mu}\left(\int_0^{S_{AB}} y_{AB}^2(s)ds + \int_0^{S_{CD}} y_{CD}^2(u)du + \frac{B(\mu T + B)}{m^2}\right)\right],$$

respectively, where T_{cycle} is the cycle duration given by

$$T_{cycle} = (1-\beta)(1 + \frac{\mu T}{m} + S_{CD})T + \frac{B}{\mu} + \frac{m}{\mu}\left(\int_0^{S_{AB}} y_{AB}(s)ds + \int_0^{S_{CD}} y_{CD}(u)du\right).$$

For $B \in (mb_{0,1}, \infty)$, we have

$$\bar{\lambda} = \frac{m}{2T_{cycle}}\frac{1+\beta}{1-\beta}(s_1+1)^2,$$

$$\bar{g} = \mu,$$

$$\bar{x} = \frac{1}{T_{cycle}}\left[mT\int_0^{s_1} y(s)ds + \frac{m^2}{\mu}\left(\int_0^{s_1} y^2(s)ds + \frac{B(\mu T + B)}{m^2}\right)\right],$$

where

$$T_{cycle} = T(s_1+1) + \frac{m}{\mu}\left(\int_0^{s_1} y(s)ds + \frac{B}{m}\right),$$

and s_1 is defined by (25) *with $k = 1$.*

Example 3 *Let us illustrate the Pareto set for a benchmark example of the TCP/IP network created with the help of NS-2 simulator. The network consists of a single bottleneck link of capacity $\mu = 10Mbps$ which is shared by n long-lived TCP connections. The propagation delay for each connection is $T = 0.24s$ and $\beta = 1/2$. The packet size is $4000bits$. Thus, we have that $m_0 = 4000bits$ as well. In Figure 11 we plot the Pareto set for $n = 10$ (and $m = m_0 n = 40,000$) using the derived analytic formulae of Theorem 3 and measurements obtained from NS simulations. As one can see, the two curves match well.*

In Figure 9, again using the formulae of Theorem 3, we plot the average goodput and the average sending rate as functions of the buer size for m = 60.

We note that $\bar{g} \leq \mu$, but the average sending rate $\bar{\lambda}$ can exceed the router capacity μ (see Figure 12). Nevertheless, as the next Proposition 2 states, the difference between the average sending rate and the router capacity goes to zero as B increases. In particular, this means that when the Drop Tail router is used, the rate of lost (and then retransmitted) information eventually diminishes to zero as the buffer size increases.

Proposition 2 *When $B \to \infty$, the difference $\Delta = \bar{\lambda} - \mu$ approaches zero from above.*

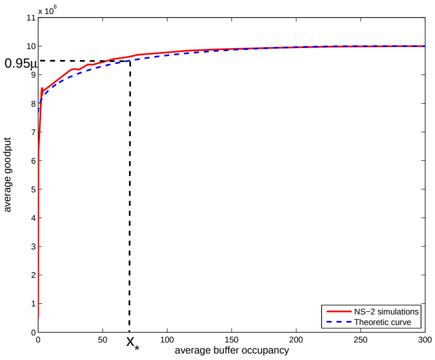

Figure 11: Pareto set: Numeric calculations and NS-2 simulations.

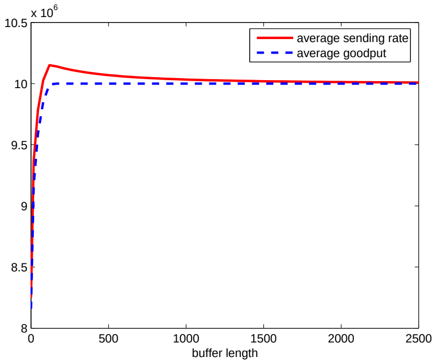

Figure 12: Non-monotonicity of the average sending rate.

7 Minimal buffer size for the full system utilization

In the case of multiple TCP connections competing for resource of the bottleneck router we have $m = nm_0$. Here n is the number of competing TCP connections. Let us study how the minimal buffer size for the full system utilization depends on n or, equivalently, on m. The minimal buffer size for the full system utilization is the buffer size when the Pareto set touches the level μ (see Figure 9). It corresponds to the critical cycle of minimal order. The next two statements describe the dependence of the minimal buffer size for the full system utilization on m.

Proposition 3 (a) *For a fixed N, the value of $B_{0,N} = mb_{0,N}$ decreases as m increases.*
(b) *The value of $B_{0,N}$ increases as N increases.*

Corollary 3 *The buffer size $B_{0,N}$ of the minimal order critical cycle is a piece-wise differentiable function of m, decreasing on the intervals $[m_i, m_{i+1})$;*

$$\lim_{m \to m_{i+1}-0} B_{0,N} < B_{0,N}(m_{i+1}), \quad i = 0, 1, 2, \dots$$

Here $m_i \overset{\triangle}{=} \mu T(1-\beta^i)/\beta^i$; the value of N equals $i+1$ on the interval $[m_i, m_{i+1})$ (see (8)).

Moreover, $\lim_{m \to m_N - 0} B_{0,N} = 0$, $\lim_{m \to m_N - 0} \frac{dB_{0,N}}{dm} = 0$, $\lim_{m \to 0+} B_{0,1} = \mu T(1-\beta)/\beta$, $\lim_{N \to \infty} m_{N-1} B_{0,N}(m_{N-1}) = 0.5(\mu T(1-\beta))^2$ *and hence* $\lim_{m \to \infty} B_{0,N} = 0$.

Example 3(cntd.) *In Figure 13 we plot the buffer size $B_{0,N}$ of the minimal order critical cycle and the curve $f(m) = (1-\beta)^2(\mu T)^2/(2m)$ for $\mu T = 600 packets$ and $\beta = 1/2$. The curve $f(m)$ indeed approaches fast the local maxima of $B_{0,N}$ as m increases. In Figure 14 we make a zoom on the interval with smaller values of m. As one can see, when m goes to zero, the value of $B_{0,N}$ approaches 600packets, which is the BDP in this network example.*

We note that by Corollary 3 for small values of m in case $\beta = 1/2$ the minimal buffer size for the full system utilization is approximately equal to μT, BDP of the bottleneck link. This is in agreement with the empirical conclusion of [26]. In [3] the authors suggested that the minimal buffer size for the full system utilization should decrease as $(\mu T)/\sqrt{m}$ as the number of connections m increases. We note that the authors of [3] have assumed that the competing TCP connections are not synchronized. That is, only a single connection reduces its congestion window when the buffer becomes full. In our model we assume full synchronization of competing

TCP connections. Namely, when the buffer is full, all connections simultaneously reduce their congestion windows. We expect that the situation in real networks is in between these two extremes. And thus, the model of [3] provides a lower bound and our model provides an upper bound. Furthermore, it was believed previously that if the competing TCP connections are synchronized, one has to provide BDP of buffering to guarantee the full system utilization. From Figure 14 one can see that the minimal buffer requirement decreases with increasing m even in the case of complete synchronization. Finally, we would like to mention that the value of $B_{0,N}$ is non-monotonous with respect to m, even though it eventually decreases to zero (see Figure 13). Curiously enough, the experiments of [28] with the router, running FreeBSD dummynet software, have also shown the non-monotonous behavior of the minimal buffer requirement in the case of synchronized connections (see Figure 1 in [28]).

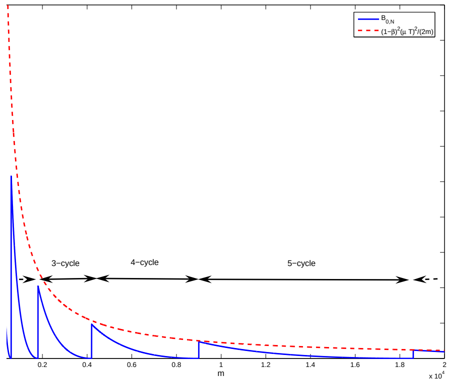

Figure 13: The minimal buffer size (in packets) for the full system utilization, as function of $m = m_0 n$.

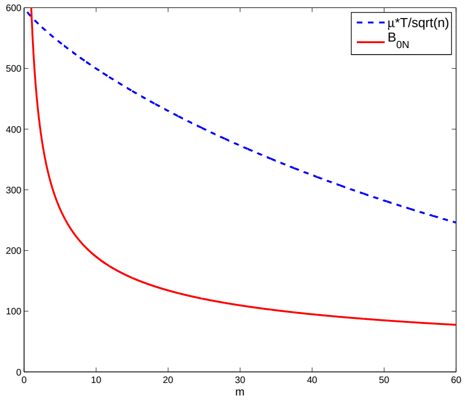

Figure 14: The minimal buffer size for the full system utilization (zoom).

8 Conclusion

We have analyzed a hybrid model which describes an interaction between Multi-Socket AIMD TCP and a bottleneck Drop Tail IP router. The model accurately takes into account feedback delay as well as non-linear sending rate evolution. It allows one to study the effect of different degrees of synchronization. We have completely characterized the dynamics of the system. In particular, we have classified the limiting cycles and found necessary and sufficient conditions for the absence of subsequent packet losses. It was believed previously that if the competing TCP connections are synchronized, one has to provide BDP of buffering to guarantee the full system utilization. From Figure 13 one can see that the minimal buffer requirement decreases with increasing m even in the case of complete synchronization. We have also demonstrated a non-monotonous behavior of $B_{0,N}$ with respect to $m = m_0 n$, even though it eventually decreases to zero (see Figure 13). Curiously enough, the experiments of [28] with the router, running FreeBSD dummynet software, have also shown the non-monotonous behavior of $B_{0,N}$ in the case of synchronized connections (see Figure 1 in [28]).

References

[1] M. Allman, V. Paxon, and W. Stevens. “TCP congestion control”. *RFC2581*, April 2002.

[2] E. Altman, D. Barman, B. Tuffin, and M. Vojnovic. “Parallel TCP Sockets: Simple Model, Throughput and Validation”. In *Proceedings of IEEE INFOCOM*, 2006.

[3] G. Appenzeller, I. Keslassy, and N. McKeown. “Sizing router buffers”. In *ACM SIGCOMM*, 2004.

[4] K.E. Avrachenkov, U. Ayesta, E. Altman, P. Nain, and C. Barakat. “The effect of router buffer size on the TCP performance”. In *Proceedings of LONIIS*, 2002.

[5] K.E. Avrachenkov, U. Ayesta, and A. Piunovskiy. “Optimal choice of the buffer size in the internet routers”. In *Proceedings of IEEE CDC-ECC*, 2005.

[6] F. Baccelli, and D. Hong. “Interaction of TCP flows as billiards”. In *IEEE INFOCOM 2003*.

[7] S. Bohacek, J.P. Hespanha, J. Lee, and K. Obraczka. “A hybrid systems modeling framework for fast and accurate simulation of data communication networks”. In *Proceedings of ACM SIGMETRICS*, 2003.

[8] M. Christiansen, K. Jeffay, D. Ott, and F.D. Smith. “Tuning RED for Web traffic”. *IEEE/ACM Transactions on Networking*, v.9(3), pp.249–264, 2001.

[9] J. Crowcroft and P. Oechslin. “Differentiated end-to-end Internet services using a weighted proportional fair sharing TCP”. *ACM SIGCOMM Computer Communication Review*, v.28(3), pp.53–69, 1998.

[10] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden. “Routers with very small buffers”. In *Proceedings of IEEE INFOCOM*, 2006.

[11] S. Floyd and V. Jacobson. “Random early detection gateways for congestion avoidance” *IEEE/ACM Transactions on Networking*, v.1(4), pp.397–413, 1993.

[12] S. Gorinsky, A. Kantawala, and J. Turner. “Link buffer sizing: a new look at the old problem”. In *Proceedings of IEEE ISCC*, 2005.

[13] GridFTP. : Protocol description and implementation. http://www.globus.org/grid-software/data/gridftp.php.

[14] J.P. Hespanha, S. Bohacek, K. Obraczka, , and J. Lee. “Hybrid modeling of tcp congestion control”. In *HSCC ’01: Proceedings of the 4th International Workshop on Hybrid Systems*, pp.291–304. Springer-Verlag, 2004.

[15] V. Jacobson. “Congestion avoidance and control”. In *Proceedings of ACM SIGCOMM*, pp.314–329, August 1988.

[16] F. Kelly. “Fairness and stability of end-to-end congestion control”. *European journal of control*, v.9, pp.159–176, 2003.

[17] Y. Li, D.J. Leith, and R.N. Shorten. “Experimental Evaluation of TCP Protocols for High-Speed Networks”. *IEEE/ACM Transaction on Networking*, v.15(5), pp.1109–1122, 2007.

[18] M. May, J. Bolot, C. Diot, and B. Lyles. "Reasons Not to Deploy RED". In *Proceedings of 7th International Workshop on Quality of Service (IWQoS'99)*, 1999.

[19] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. "Modeling TCP throughput: A simple model and its empirical validation". In *ACM SIGCOMM 1998*.

[20] A.B. Piunovskiy. *Optimal Control of Random Sequences in Problems with Constraints.* Kluwer Academic Publishers: Dordrecht, 1997.

[21] R. Prasad, C. Dovrolis, and M. Thottan. "Router buffer sizing revisited: The role of the output/input capacity ratio". In *Proceedings of ACM CoNext*, 2007.

[22] G. Raina and D.J. Wischik. "Buffer sizes for large multiplexers: TCP queueing theory and instability analysis". In *Proceedings of Euro NGI*, 2005.

[23] G. Raina, D. Towsley and D. Wischik. "Part II: Control theory for buffer sizing", *ACM SIGCOMM Computer Communication Review*, v.35(3), pp.79-82, 2005.

[24] K.K. Ramakrishnan, S. Floyd, and D. Black. "The Addition of Explicit Congestion Notification (ECN) to IP". *RFC3168*, September 2001.

[25] R. Stanojevic, R. Shorten, and C. Kellet. "Adaptive tuning of drop-tail buffers for reducing queueing delays". *IEEE Communications Letters*, v.10(7), pp.570–572, 2006.

[26] C. Villamizar and C. Song. "High Performance TCP in the ANSNET". *ACM SIGCOMM Computer Communication Review*, v.24(5), pp.45–60, 1994.

[27] A. Vishwanath, V. Sivaraman, and M. Thottan. "Perspectives on router buffer sizing: recent results and open problems". *ACM SIGCOMM Computer Communication Review*, v.39(2), pp.34–39, 2009.

[28] G. Vu-Brugier, R.S. Stanojevic, J. Leith, and R.N. Shorten. "A critique of recently proposed buffer-sizing strategies". *ACM SIGCOMM Computer Communication Review*, v.37(1), pp.43–48, 2007.

[29] L. Zhang, S. Shenker, and D.D. Clark, "Observations on the dynamics of a congestion control algorithm: the effects of two-way traffic". *ACM SIGCOMM Computer Communication Review*, v.21(4), pp.133-147, 1991.

Appendix

Unclipped cycles.

In this and the next subsection, we ignore the requirement that $y \geq 0$. Thus dynamics is described by equations

$$\left\{ \begin{array}{l} \frac{dv}{ds} = 1; \\ \frac{dy}{ds} = \left\{ \begin{array}{ll} v - y - q, & \text{if } y < b, \text{ or} \\ & y = b \text{ and } v \leq A; \\ 0 & \text{otherwise,} \end{array} \right. \end{array} \right. \tag{20}$$

where

$$A \stackrel{\triangle}{=} b + q.$$

The jumps occur according to (7) as before.

Definition 2 *Let $y_0 = b$ and $v_0 < A$ be the initial conditions. A piece of trajectory on the time interval $[0, s^* + 1 + 0]$ is called a pseudo-cycle of order k (see (7)). If $v(s^* + 1 + 0) = v_0$ then the pseudo-cycle is called a k-cycle.*

Later, it will be shown that if a clipped k-cycle exists then the unclipped k-cycle exists, too (Corollary 6). Clearly, (20) has a single solution

$$\left\{ \begin{array}{l} v(s) = v_0 + s; \\ y(s) = (1 + q + y_0 - v_0)e^{-s} + s - 1 + v_0 - q. \end{array} \right. \tag{21}$$

Theorem 4 *An (unclipped) k-cycle exists iff*

$$A \in \left(\frac{\beta^k}{1 - \beta^k},\ A_k^* \right], \tag{22}$$

where

$$A_k^* \triangleq \left\{ \begin{array}{ll} \frac{\beta^{k-1}(\tau_k + 1)}{1 - \beta^k}, & \textit{if } k > 1, \\ \infty, & \textit{if } k = 1 \end{array} \right. \tag{23}$$

and, for $k > 1$, τ_k is the single positive solution to (13).

<u>Proof.</u> Obviously, parameters of a k-cycle, v_0 and time interval s_1 can be found from equations

$$y(s_1) = b; \qquad \beta^k v(s_1 + 1) = v_0,$$

which are equivalent to

$$v_0 = \frac{\beta^k (s_1 + 1)}{1 - \beta^k}. \tag{24}$$

$$1 - e^{-s_1} = \frac{s_1}{1 + A - \frac{\beta^k (s_1 + 1)}{1 - \beta^k}}. \tag{25}$$

A k-cycle exists iff (25) has a positive solution and v_0 given by (24) satisfies inequality $v_0 \geq \beta A$. (Otherwise, if $v_0 < \beta A$, there is no need to reduce v so many times.) Equation (25) has a positive solution iff

$$1 + A - \frac{\beta^k}{1 - \beta^k} > 0 \quad \text{and} \quad \frac{d}{ds} \left[\frac{s}{1 + A - \frac{\beta^k (s+1)}{1 - \beta^k}} \right] \Bigg|_{s=0} \ < \ 1$$

(see Fig.15), or, equivalently, iff

$$\beta^k (1 + A) < A. \tag{26}$$

Put

$$K \triangleq \min\{i \geq 1 : \ \ \beta^i < \frac{A}{1 + A}\}. \tag{27}$$

Before proceeding further, we need the following statements.

Lemma 1 *If $v_0 \in [\beta A, A)$ then, starting from v_0, $y_0 = b$, the next instant series of $K + 1$ jumps results in the value $v < A$. Hence the order of any cycle cannot exceed $K + 1$ (and clearly cannot be smaller than K).*

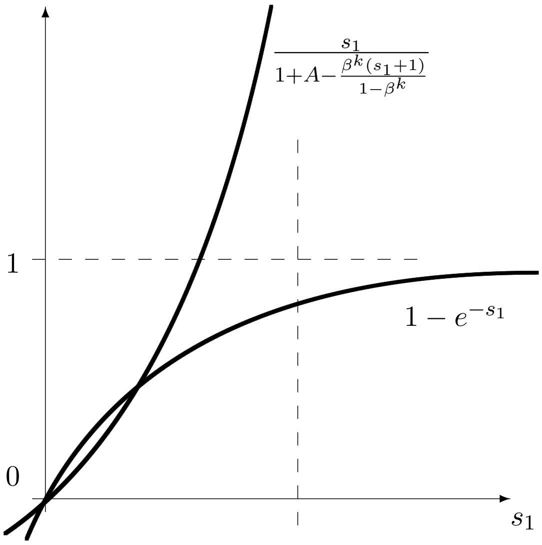

Figure 15: Graphical solution to equation (25).

Proof. Suppose $\hat{v}_0 = \beta A$. Then, after the next instant series of $K+1$ jumps, the value $\hat{v}$ is not smaller than v. To put it differently,

$$v \le \beta^{K+1}[\beta A + \hat{s} + 1], \tag{28}$$

where $\hat{s}$ solves equation

$$\begin{aligned} &(1 + A - \beta A)e^{-s} + s - 1 + \beta A = A \\ \iff &\frac{s}{1 + (1-\beta)A} - 1 + e^{-s} = 0. \end{aligned} \tag{29}$$

If we substitute

$$\tilde{s} \overset{\triangle}{=} \frac{A+1}{\beta} - \beta A - 1 < \frac{A}{\beta^{K+1}} - \beta A - 1$$

into (29) we obtain, using equality $A = \frac{\beta\tilde{s}+\beta-1}{1-\beta^2}$:

$$\frac{\tilde{s}}{1+(1-\beta)A} - 1 + e^{-\tilde{s}} = \frac{\tilde{s}(1+\beta)}{\beta(2+\tilde{s})} - 1 + e^{-\tilde{s}} > \frac{2\tilde{s}}{2+\tilde{s}} - 1 + e^{-\tilde{s}} > 0.$$

Used inequalities: $\frac{1+\beta}{\beta} > 2$ for $\beta < 1$ and $\tilde{s} - 2 + 2e^{-\tilde{s}} + \tilde{s}e^{-\tilde{s}} \ge 0$. (This function increases from zero for positive $\tilde{s}$.) When s increases from zero, the lefthand side of (29) initially decreases from zero and increases thereafter. Hence $\tilde{s} > \hat{s}$ and (28) implies

$$v < \beta^{K+1}[\beta A + \tilde{s} + 1] < \beta^{K+1}[\beta A + \frac{A}{\beta^{K+1}} - \beta A - 1 + 1] = A.$$

■

Lemma 2 *Suppose $\beta \in (0,1)$ is fixed and consider function*

$$f_k(A) \overset{\triangle}{=} (1-\beta^k)A - \beta^{k-1}(s_1+1), \tag{30}$$

where s_1 solves (25). The domain of f is given by (26). Then

(a) $\forall k \geq 2$ $\frac{df_k(A)}{dA} > 0$;

(b) $f_1(A) < 0$ for all $A > \frac{\beta}{1-\beta}$;

(c) $\forall k > 1$ equation $f_k(A) = 0$ has a single finite solution A_k^* given by (23); A_k^* decreases as k increases.

(d) $\forall k > 1$ $A_k^* > \frac{\beta^{k-1}}{1-\beta^{k-1}}$; $\forall k > 2$, $A_k^* \leq \frac{\beta^{k-2}}{1-\beta^{k-2}}$.

Proof. (a) According to the rule of implicit differentiation, applied to equation

$$\left(1 + A - \frac{\beta^k(s_1+1)}{1-\beta^k}\right)\left(1 - e^{-s_1}\right) - s_1 = 0,$$

we have

$$\frac{ds_1}{dA} = -\frac{(1-e^{-s_1})^2(1-\beta^k)}{e^{-s_1}s_1(1-\beta^k) - (1-e^{-s_1})^2\beta^k - (1-\beta^k)(1-e^{-s_1})}.$$

The denominator equals

$$-(1 - e^{-s_1} - s_1e^{-s_1}) - \beta^k(s_1e^{-s_1} - e^{-s_1} + e^{-2s_1})$$

$$< \beta^k(e^{-s_1} - e^{-2s_1} - s_1e^{-2s_1}) - (1 - e^{-s_1} - s_1e^{-s_1})$$

$$= (1 - e^{-s_1} - s_1e^{-s_1})(\beta e^{-s_1} - 1) < 0;$$

hence $\frac{ds_1}{dA} > 0$ for $s_1 > 0$.

Now

$$\frac{df_k(A)}{dA} = (1-\beta^k) - \beta^{k-1}\frac{ds_1}{dA}$$

$$= \frac{(1-\beta^k)[(1-\beta^k)(s_1e^{-s_1} - 1 + e^{-s_1}) + \beta^{k-1}(1-e^{-s_1})^2(1-\beta)]}{s_1e^{-s_1}(1-\beta^k) + (1-e^{-s_1})(\beta^k e^{-s_1} - 1)}.$$

The denominator is negative (see above). Consider the nominator at $k \geq 2$:

$$g_1(s_1) = (1-\beta^k)(s_1e^{-s_1} - 1 + e^{-s_1}) + \beta^{k-1}(1-e^{-s_1})^2(1-\beta).$$

Clearly, $g_1(0) = 0$ and

$$\frac{dg_1}{ds_1} = e^{-s_1}\{2\beta^{k-1}(1-\beta)(1-e^{-s_1}) - s_1(1-\beta^k)\}.$$

Function in the parentheses $g_2(s_1) = 2\beta^{k-1}(1-\beta)(1-e^{-s_1}) - s_1(1-\beta^k)$ is negative if $s_1 > 0$ since $g_2(0) = 0$ and

$$\frac{dg_2}{ds_1} = 2\beta^{k-1}(1-\beta)e^{-s_1} - 1 + \beta^k < 2\beta^{k-1}(1-\beta) - 1 + \beta^k = 2\beta^{k-1} - \beta^k - 1 < 0$$

because $1 + \beta^k > 2\beta^{k/2} \geq 2\beta^{k-1}$. Hence $\frac{df_k(A)}{dA} > 0$.

(b) It is sufficient to prove that

$$s_1 > S \overset{\triangle}{=} A(1-\beta) - 1,$$

where s_1 solves (25) at $k = 1$.

Case $S < 0$ is trivial, thus assume that $S > 0$. Let us substitute S into the both sides of (25) and estimate the difference:

$$\frac{S}{1 + A - \frac{\beta(S+1)}{1-\beta}} - 1 + e^{-S} = \frac{S}{1 + \frac{S+1}{1-\beta} - \frac{\beta(S+1)}{1-\beta}} - 1 + e^{-S} = e^{-S} - \frac{2}{S+2} < 0,$$

because function $(S+2)e^{-S}$ decreases from 2 at $S = 0$. To complete this part of the proof, it is sufficient to notice that, on the interval

$$0 < S < \frac{A(1-\beta)}{\beta} + \frac{1-2\beta}{\beta},$$

the righthand side of (25) is smaller than the lefthandside iff $S < s_1$.

(c) The first part is obvious: A_k^* is given by (23), provided equation (13) has a single positive solution. The latter statement follows from the fact that function

$$g(\tau) = (1 - e^{-\tau})(1 + \alpha(\tau + 1))/\tau$$

decreases to $\lim_{\tau\to\infty} g(\tau) = \alpha$, starting from $\lim_{\tau\to 0} g(\tau) = 1 + \alpha$. Here

$$\alpha \stackrel{\triangle}{=} \frac{\beta^{k-1} - \beta^k}{1 - \beta^k}. \tag{31}$$

Indeed,

$$\frac{dg}{d\tau} = \frac{e^{-\tau}[1 + \alpha + \tau(1 + \alpha + \alpha\tau)] - (1 + \alpha)}{\tau^2} < 0$$

in case $\alpha < 1$, and

$$\alpha = \frac{\beta^{k-1}}{1 + \beta + \ldots + \beta^{k-1}} \leq \frac{1}{1 + 1/\beta} < 1/2. \tag{32}$$

Now, look what happens as k increases. Obviously, functions $\frac{\beta^{k-1}}{1-\beta^k} = \frac{\beta^k}{1-\beta^k} \cdot \frac{1}{\beta}$ and $\alpha = \frac{\beta^k}{1-\beta^k}(\frac{1}{\beta} - 1)$ (see (31)) decrease. According to (23) it remains to prove that τ_k given by (13) increases with α. We rewrite (13) as $(1 + \alpha(\tau + 1))(1 - e^{-\tau}) - \tau = 0$. Hence

$$\frac{d\tau_k}{d\alpha} = -\frac{(\tau_k + 1)(1 - e^{-\tau_k})}{\alpha(1 - e^{-\tau_k}) + (1 + \alpha(\tau_k + 1))e^{-\tau_k} - 1} = -\frac{(\tau_k + 1)^2(1 - e^{-\tau_k})^2}{h(\tau_k)},$$

where $h(\tau) = -2 + 3e^{-\tau} - e^{-2\tau} + \tau e^{-\tau} + \tau^2 e^{-\tau}$. (We have substituted $\alpha = \frac{\tau_k - 1 + e^{-\tau_k}}{(1-e^{-\tau_k})(\tau_k+1)}$.) We intend to prove that

$$\frac{dh}{d\tau} = -2e^{-\tau} + 2e^{-2\tau} + \tau e^{-\tau} - \tau^2 e^{-\tau} < 0 \tag{33}$$

when $\tau > 0$. Clearly (33) holds for $\tau \geq 1$.

Suppose $\tau \in (0, 1)$. Then

$$\frac{d^2h}{d\tau^2} = 3e^{-\tau} - 4e^{-2\tau} - 3\tau e^{-\tau} + \tau^2 e^{-\tau} < 3e^{-\tau} - 4e^{-2\tau} - 2\tau e^{-\tau} = e^{-\tau}(3 - 4e^{-\tau} - 2\tau).$$

Expression in the brackets has a negative maximum at $\tau = \ln 2$. Therefore, $\frac{d^2h}{d\tau^2} < 0$ and $\frac{dh}{d\tau} < 0$. Finally, $h(\tau) < 0$ for all $\tau > 0$, because $h(0) = 0$.

(d) To estimate A_k^* from below, we use statement (a): it is sufficient to establish that $f_k\left(\frac{\beta^{k-1}}{1-\beta^{k-1}}\right) < 0$, ie $s_1 + 1 > \frac{1-\beta^k}{1-\beta^{k-1}} \iff \frac{S}{1+\frac{\beta^{k-1}}{1-\beta^{k-1}}-\frac{\beta^k(S+1)}{1-\beta^k}} < 1 - e^{-S}$ for $S = \frac{1-\beta^k}{1-\beta^{k-1}} - 1$. (The argument is similar to (b).) But

$$\frac{S}{\frac{1}{1-\beta^{k-1}} - \frac{\beta^k}{1-\beta^k}\cdot\frac{1-\beta^k}{1-\beta^{k-1}}} - 1 + e^{-S} = \frac{S}{S+1} - 1 + e^{-S} = e^{-S} - \frac{1}{1+S} < 0$$

because function $e^{-S}(1+S)$ decreases from 1 at $S = 0$.

Finally, in case $k > 2$, suppose $A_k^* > \frac{\beta^{k-2}}{1-\beta^{k-2}}$. Then for parameters values β and $A \in \left(\frac{\beta^{k-2}}{1-\beta^{k-2}}, A_k^*\right)$ we have that (26) holds for $k-2$, $k-1$, and k and simultaneously $f_k(A) < 0$, $f_{k-1}(A) < 0$, $f_{k-2}(A) < 0$: see (a), and (c), and (b) in case $k = 3$. According to the beginning of the proof of Theorem 1, cycles of orders k, $k-1$, and $k-2$ exist which contradicts Lemma 1. ■

Now we can easily finish the proof of Theorem 4. Suppose a k-cycle exists. Then, according to (26), $A > \frac{\beta^k}{1-\beta^k}$. Lemma 2 guarantees that

$$A_k^* > \frac{\beta^{k-1}}{1-\beta^{k-1}} > \frac{\beta^k}{1-\beta^k},$$

and, as was mentioned earlier, inequality $v_0 \geq \beta A$ must be valid (see (24)), which is equivalent to $A \leq A_k^*$. Finally, if (22) holds then (25) has a positive solution (see (26)) and $v_0 \geq \beta A$; hence a k-cycle exists. ■

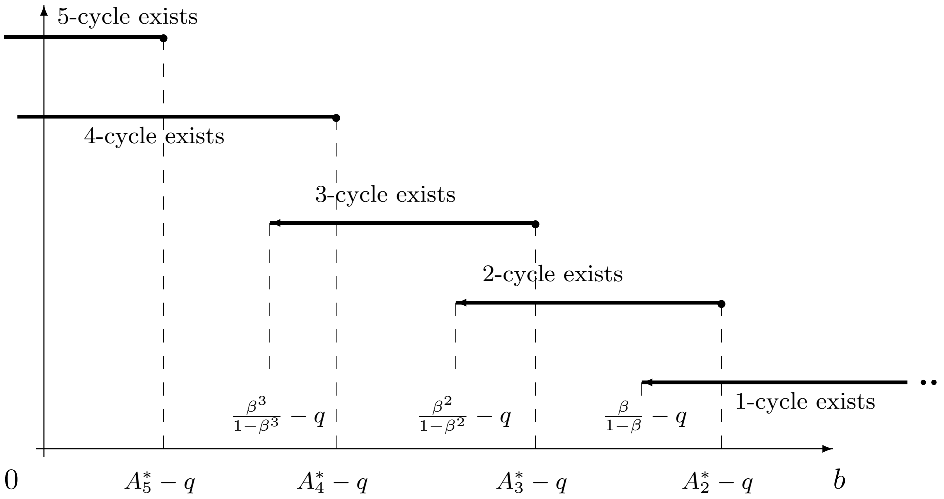

Figure 16: Existence of unclipped cycles; $N = 4$, case $A_{N+1}^* > q$.

Remember that $A = b + q$. Thus, if q is fixed and b increases from 0, unclipped cycles have orders N (see (8)) and, possibly, $N + 1$, if $A_{N+1}^* - q > 0$. Later, as b increases, the order of cycles decreases according to Fig. 16.

Stability of unclipped cycles.

We intend to study the mapping φ introduced just before Theorem 1. Since we study only unclipped cycles, this map is a little different and will be denoted $\tilde{\varphi}$. But firstly we concentrate on a different mapping:

$$\Phi^k(v_0) = \beta^k(v_0 + s^* + 1)$$

defined for $v_0 \in [\beta A, A]$ under a fixed $k \geq 1$. Here $s^* \overset{\triangle}{=} 0$ if $v_0 = A$; in case $v_0 < A$, $s^* > 0$ is the first moment when $y(s^*) = b$ starting from $y(0) = b$, $v(0) = v_0$.

Lemma 3 $\left|\frac{d\Phi^k(v_0)}{dv_0}\right| < \beta^k$ *and hence* Φ^k *is a contraction. Function* Φ^k *is decreasing.*

Proof. Assuming that $v_0 < A$, s^* is a single positive solution to equation

$$(1 + A - v_0)(1 - e^{-s}) - s = 0, \tag{34}$$

hence

$$\frac{ds^*}{dv_0} = \frac{1 - e^{-s^*}}{(1 + A - v_0)e^{-s^*} - 1} = \frac{1 - e^{-s^*}}{\frac{s^*}{1-e^{-s^*}} \cdot e^{-s^*} - 1}$$

and

$$\frac{d\Phi^k}{dv_0} = \beta^k(1 + \frac{ds^*}{dv_0}) = \beta^k e^{-s^*} \frac{s^* - 1 + e^{-s^*}}{s^* e^{-s^*} + e^{-s^*} - 1} < 0.$$

Finally,

$$e^{-s^*} \frac{s^* - 1 + e^{-s^*}}{s^* e^{-s^*} + e^{-s^*} - 1} + 1 = e^{-s^*} \frac{e^{s^*} - e^{-s^*} - 2s^*}{1 - e^{-s^*} - s^* e^{-s^*}} > 0,$$

because the nominator increases, starting from 0 at $s^* = 0$. Therefore $\frac{d\Phi^k}{dv_0} > (-\beta^k)$. ■

Lemma 4 *(a)* $A \in \left(A^*_{K+1}, \frac{\beta^{K-1}}{1-\beta^{K-1}}\right]$ *iff* $d < \beta A$, *where* d *is a solution to* $\Phi^K(d) = A$. *Here and below,* $\frac{\beta^{K-1}}{1-\beta^{K-1}} \overset{\triangle}{=} \infty$ *if* $K = 1$; K *is defined by (27).*

In this case, $\forall v_0 \in [\beta A, A)$, *the mapping* $\tilde{\varphi}(v_0)$ *coincides with* $\Phi^K(v_0)$.

(b) If $A \in \left(\frac{\beta^K}{1-\beta^K}, A^*_{K+1}\right]$ *the following statements hold:*

(α) $\forall v_0 \in [\beta A, d]$, $\tilde{\varphi}(v_0) = \Phi^{K+1}(v_0) \in [\beta A, d]$;

(β) $\forall v_0 \in (d, A)$, $\tilde{\varphi}(v_0) = \Phi^K(v_0) \in (d, A)$.

See Fig.17. (Note that, according to the definition of K, $A \in \left(\frac{\beta^K}{1-\beta^K}, \frac{\beta^{K-1}}{1-\beta^{K-1}}\right]$; *according to Lemma 2,* $A^*_{K+1} \in \left(\frac{\beta^K}{1-\beta^K}, \frac{\beta^{K-1}}{1-\beta^{K-1}}\right]$.*)*

Proof. (a) According to the definition, $d = \frac{A}{\beta^K} - s^* - 1$, where s^* solves (34) under $v_0 = d$. If $d = \beta A$ then

$$\begin{cases} (1 + A - \beta A)(1 - e^{-s^*}) &= s^*; \\ \frac{A}{\beta^K} - s^* - 1 &= \beta A, \end{cases}$$

or equivalently

$$\begin{cases} A &= \frac{\beta^K(s^*+1)}{1-\beta^{K+1}}; \\ (1 + A - \frac{\beta^{K+1}(s^*+1)}{1-\beta^{K+1}})(1 - e^{-s^*}) &= s^*. \end{cases}$$

To put it differently, we have $A = A^*_{K+1}$ if $d = \beta A$.

It remains to prove that $d - \beta A = \frac{A}{\beta^K} - s^* - 1 - \beta A$ is a decreasing function of A. Since s^* satisfies equation

$$\left(1 + A - \frac{A}{\beta^K} + s^* + 1\right)(1 - e^{-s^*}) - s^* = 0,$$

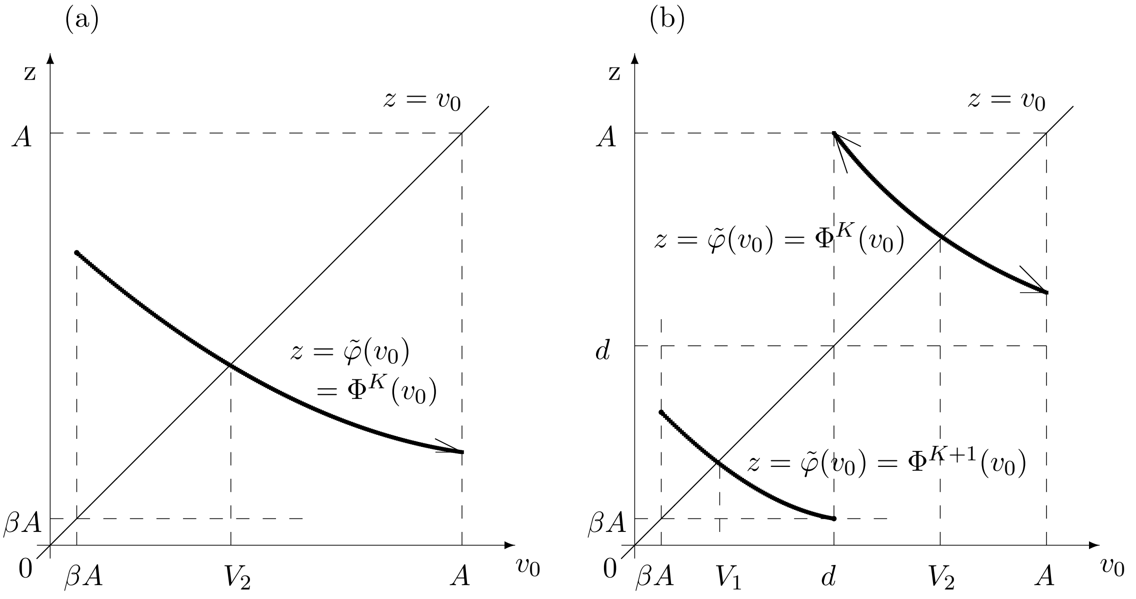

Figure 17: Graphs of $\tilde\varphi(v_0)$.

$$\frac{ds^*}{dA} = \frac{(1-\beta^K)(1-e^{-s^*})^2}{\beta^K e^{-s^*}(e^{-s^*}+s^*-1)}$$

and

$$\frac{d(d-\beta A)}{dA} = \frac{1}{\beta^K} - \frac{ds^*}{dA} - \beta = \frac{s^*e^{-s^*}+e^{-s^*}-1+\beta^K(1-e^{-s^*})^2-\beta^{K+1}e^{-s^*}(e^{-s^*}+s^*-1)}{\beta^K e^{-s^*}(e^{-s^*}+s^*-1)}.$$

The denominator is obviously positive for $s^* > 0$. The nominator equals zero when $s^* = 0$, its derivative equals

$$e^{-s^*}[-s^* + 2\beta^K(1-e^{-s^*}) - \beta^{K+1}(2-2e^{-s^*}-s^*)].$$

Expression in the square brackets equals zero when $s^* = 0$ and has derivative

$$-1 + 2\beta^K e^{-s^*} - 2\beta^{K+1}e^{-s^*} + \beta^{K+1} \overset{\triangle}{=} g(s^*, \beta).$$

Clearly,

$$\frac{\partial g(s^*,\beta)}{\partial s^*} = 2\beta^K e^{-s^*}(\beta-1) < 0,$$

and finally $g(0,\beta) = -1 + 2\beta^K - \beta^{K+1} < 0$ for all $\beta \in (0,1)$ because $g(0,1) = 0$ and $\frac{dg(0,\beta)}{d\beta} = \beta^{K-1}[K(1-\beta)+K-\beta] > 0$. Therefore $\frac{d(d-\beta A)}{dA} < 0$.

According to Lemma 1, $\tilde\varphi$ can coincide with Φ^K or Φ^{K+1} only. In case (a), $\Phi^K(A) < A$ because $\lim_{v_0\to A} s^* = 0$ (see (27)). Function Φ^K increases as v_0 decreases (Lemma 3), but $\Phi^K(v_0) = A$ when $v_0 = d < \beta A$. Thus, $\forall v_0 \in [\beta A, A)$ $\Phi^K(v_0) < A$, $(K+1)$ instant jumps are never needed and $\tilde\varphi = \Phi^K$.

(b) In this case, $d \geq \beta A$ according to (a). Since $\Phi^K(d) = A$ and Φ^K is a decreasing function (Lemma 3), $\Phi^K(v_0) \geq A$ if $v_0 \in [\beta A, d]$ and $\tilde\varphi(v_0) = \Phi^{K+1}(v_0)$, as K jumps are not sufficient. Obviously, $\tilde\varphi(d) = \Phi^{K+1}(d) = \beta A$. Now

$$\tilde\varphi(\beta A) = \Phi^{K+1}(\beta A) = \Phi^{K+1}(d) - \int_{\beta A}^{d} \frac{d\Phi^{K+1}(v_0)}{dv_0}dv_0 < \Phi^{K+1}(d) + (d-\beta A) = d$$

according to Lemma 3, and statement (α) is proved.

In case (β), $\Phi^K(v_0) < A$, hence $\tilde{\varphi}(v_0) = \Phi^K(v_0)$. We know that $\Phi^K(d) = A$. Using Lemma 3, we conclude that

$$\Phi^K(A) = \Phi^K(d) + \int_d^A \frac{d\Phi^K(v_0)}{dv_0} dv_0 > A - (A - d) = d.$$

■

Corollary 4 *Theorem 1 and Corollary 1 hold for unclipped cycles.*

Proof. (See Fig.17.) Under conditions (a) of Lemma 4, $\tilde{\varphi}$ has a stable stationary point V_2 coincident with that of Φ^K. (Note that $\Phi^K(A) < A$, so that $V_2 \in [\beta A, A)$.)

Consider case (b) of Lemma 4.

If $v_0 \in [\beta A, d]$ then $\tilde{\varphi} = \Phi^{K+1}$ is a contraction defined on this interval; so that the statement follows.

If $v_0 \in (d, A)$, $\tilde{\varphi}$ has a stable stationary point V_2 coincident with that of Φ^K. (Note that $\Phi^K(A) < A$, hence $d < A = \Phi^K(d)$, so that $V_2 \in (d, A)$.)

Corollary 1 is obvious. ■

Critical cycles.

Remind that a cycle is called critical if $\min_s y(s) = 0$ From (21,24,25) it is clear that the minimum is attained at

$$s_0 = \ln \frac{s_1}{1 - e^{-s_1}}, \tag{35}$$

where s_1 solves (25).

Lemma 5 *Suppose, an unclipped k-cycle exists.*

(a) $y(s_0)$ increases with A.

(b) For cycles of order $k = 1$, $\exists \varepsilon > 0\ \exists \delta > 0$: $\frac{dy(s_0)}{dA} > \varepsilon$ as soon as $A > \frac{\beta}{1-\beta} + \delta$. Consequently $y(s_0) \to \infty$ as $A \to \infty$.

Proof. (a) After rewriting (25) in the form

$$\left(1 + A - \frac{\beta^k(s_1 + 1)}{1 - \beta^k}\right)(1 - e^{-s_1}) - s_1 = 0,$$

we obtain:

$$\frac{ds_1}{dA} = \frac{1 - e^{-s_1}}{\frac{\beta^k}{1-\beta^k}(1 - e^{-s_1}) - e^{-s_1}\frac{s_1}{1-e^{-s_1}} + 1} = \frac{(1 - e^{-s_1})^2(1 - \beta^k)}{1 - e^{-s_1} - s_1 e^{-s_1}(1 - \beta^k) - \beta^k e^{-s_1} + \beta^k e^{-2s_1}}. \tag{36}$$

The denominator has derivative (wrt $s_1 > 0$)

$$s_1 e^{-s_1}(1 - \beta^k) + 2\beta^k(e^{-s_1} - e^{-2s_1}) > 0$$

and hence increases starting from 0 when $s_1 = 0$. Therefore $\frac{ds_1}{dA} > 0$.

Since

$$y(s_0) = (1 + A - v_0)e^{-s_0} + s_0 - 1 + v_0 - q = v_0 + s_0 - q \tag{37}$$

we conclude that

$$\frac{dy(s_0)}{dA} = \left(\frac{dv_0}{ds_1} + \frac{ds_0}{ds_1}\right)\frac{ds_1}{dA} = \left(\frac{\beta^k}{1 - \beta^k} + \frac{1 - e^{-s_1} - s_1 e^{-s_1}}{s_1(1 - e^{-s_1})}\right)\frac{ds_1}{dA} > 0.$$

(b) Note that the denominator in (36) is a bounded function of s_1. Thus $\exists\varepsilon > 0\ \exists\delta_1 > 0$: $\frac{ds_1}{dA} > \varepsilon$ as soon as $s_1 > \delta_1$, or, equivalently, as soon as $A > \frac{\beta}{1-\beta} + \delta$, where $\delta > 0$ exists because s_1 monotonically increases with A. Remember that $\lim_{A\to\frac{\beta}{1-\beta}} s_1 = 0$. ■

Lemma 6 *Suppose, all parameters, apart from b, are fixed.*

(a) A critical cycle of order k exists (for some positive value of b) if and only if

$$\frac{\beta^k}{1-\beta^k} < q \le q_k^*, \tag{38}$$

where q_k^* is given by (15). The corresponding value of b equals $b_{0,k}$, see (12).

(b) The boundary q_k^* satisfies inequalities

$$\frac{\beta^{k-1}}{1-\beta^{k-1}} \le q_k^* < A_k^*. \tag{39}$$

(In case $k = 1$, $q_1^* = +\infty$.)

Proof. (a) Necessity. Let $k > 1$ and suppose a critical cycle of order k exists. Then, if we increase b up to $b^* = A_k^* - q$, this k-cycle (equipped with an asterisk) must remain unclipped (Lemma 5):

$$y^*(s_0^*) = v_0^* + s_0^* - q \ge 0 \tag{40}$$

(see (37)), ie $q \le v_0^* + s_0^*$. Here $s_0^* = \ln\frac{s_1^*}{1-e^{-s_1^*}}$ (see 35)), s_1^* solves (25) under A_k^* and hence coincides with τ_k defined by (13); v_0^* is defied by (24). Therefore, $v_0^* + s_0^* = q_k^*$.

Obviously, system of equations (24,25,35) and

$$v_0 + s_0 - q = 0$$

(see (37)) must be compatible, ie equation

$$h(s_1) = \frac{\beta^k(s_1+1)}{1-\beta^k} + \ln\frac{s_1}{1-e^{-s_1}} - q = 0 \tag{41}$$

must have a positive solution. One can easily check that h increases to infinity with s_1, starting from $\lim_{s_1\to 0} h(s_1) = \frac{\beta^k}{1-\beta^k} - q$. Hence $q > \frac{\beta^k}{1-\beta^k}$.

In case $k = 1$ we put $q_1^* = +\infty$, so that (38) transforms to $q > \frac{\beta}{1-\beta}$, and the proof of the latter inequality remains unchanged.

Before proving sufficiency, we firstly prove part (b).

(b) Let $k > 1$;

$$h \stackrel{\triangle}{=} q_k^* - \frac{\beta^{k-1}}{1-\beta^{k-1}} = \frac{\beta^k}{1-\beta^k}(\tau_k+1) + \ln\frac{\tau_k}{1-e^{-\tau_k}} - \frac{\beta^{k-1}}{1-\beta^{k-1}}$$

$$= \ln\frac{\tau_k}{1-e^{-\tau_k}} - \alpha(\tau_k+1) - \alpha(\tau_k+1)\gamma + \tau_k\gamma,$$

where $\alpha \stackrel{\triangle}{=} \frac{\beta^{k-1}-\beta^k}{1-\beta^k}$, $\gamma \stackrel{\triangle}{=} \frac{\beta^{k-1}}{1-\beta^{k-1}}$. Using (13), the last expression can be rewritten as

$$h = 1 - \frac{\tau_k}{1-e^{-\tau_k}} + \ln\frac{\tau_k}{1-e^{-\tau_k}} + \left(1 - \frac{\tau_k}{1-e^{-\tau_k}}\right)\gamma + \tau_k\gamma.$$

For $k > 1$ one can easily check that $\gamma \geq \frac{\alpha}{1-2\alpha}$; therefore, since $1 - e^{-\tau_k} - \tau_k e^{-\tau_k} \geq 0$,

$$\begin{aligned} h &\geq 1 - \frac{\tau_k}{1-e^{-\tau_k}} + \ln\frac{\tau_k}{1-e^{-\tau_k}} + \frac{1-e^{-\tau_k}-\tau_k e^{-\tau_k}}{1-e^{-\tau_k}} \cdot \frac{\alpha}{1-2\alpha} \\ &= 1 - \frac{\tau_k}{1-e^{-\tau_k}} + \ln\frac{\tau_k}{1-e^{-\tau_k}} + \frac{1-e^{-\tau_k}-\tau_k e^{-\tau_k}}{1-e^{-\tau_k}} \cdot \frac{\tau_k - 1 + e^{-\tau_k}}{3-3e^{-\tau_k}-\tau_k-\tau_k e^{-\tau_k}}. \end{aligned} \tag{42}$$

(We have used (13) to express α in terms of τ_k.)

During the proof of Lemma 2(c), we established that τ_k increases with $\alpha \in (0, 1/2)$, starting from 0 when $\alpha = 0$. Hence $\tau_k \in (0, \tau)$, where τ is the single positive solution to equation

$$(1 - e^{-\tau})(1 + \frac{1}{2}(\tau + 1)) = \tau.$$

(The solvability was established in the Proof of Lemma 2(c).)

Now the righthand side of (42) is non-negative if $\tau_k \in (0, \tau)$. This statement was accuratly checked numerically; the analytical proof is problematic.

The second inequality, to be verified, is obvious:

$$q_k^* - A_k^* = \frac{\beta^k}{1-\beta^k}(\tau_k + 1) + \ln\frac{\tau_k}{1-e^{-\tau_k}} - \frac{\beta^{k-1}(\tau_k+1)}{1-\beta^k} = \ln[1 + \alpha(\tau_k + 1)] - \alpha(\tau_k + 1) < 0.$$

(a) Sufficiency. Suppose inequalities (38) hold. Then for $b \in [0, A_k^* - q]$ (unclipped) k-cycles exist according to Theorem 4, see Fig.16. (Remember that $A_1^* = q_1^* = +\infty$.) Note that, in case $k > 1$, $q < A_k^*$ due to (b). In this case, for $b = b^* = A_k^* - q$,

$$y^*(s_0^*) = v_0^* + s_0^* - q = q_k^* - q \geq 0$$

(see (40)) and this particular cycle is really unclipped. In case $k = 1$, according to Lemma 5(b), $y(s_0) > 0$ for sufficiently large b. Now, if b decreases then the minimal value of y over a cycle decreases (Lemma 5(a)) and, being continuous, becomes zero, since $y(s_0) < 0$ for the unclipped k-cycle corrresponding to $b = 0$.

To calculate the critical value of b, note that equation (41) has a single positive solution s_1. Now, if we take

$$b = \frac{s_1}{1-e^{-s_1}} + \frac{\beta^k(s_1+1)}{1-\beta^k} - 1 - q = \frac{s_1}{1-e^{-s_1}} - \ln\frac{s_1}{1-e^{-s_1}} - 1$$

then, according to (24,25), the corrresponding cycle will be critical. (One can easily see that $b > 0$.) It remains to notice that equation (41) is identical with (11). ■

Corollary 5 *Let N be defined by (8). Then critical cycles of orders $k < N$ cannot exist.*

Proof. According to (8), $q \leq \frac{\beta^k}{1-\beta^k}$, if $n < N$. The statement follows from Lemma 6(a). ■

Clipped cycles.

Proof of Theorem 1. Let S be the single positive solution to equation

$$(1 + b + S)e^{-S} = 1.$$

Then a continuous trajectory (21) starting from ($y_0 = b,\ v_0 = q - S$) touches the axis $y = 0$ at a single point, at time moment S.

(a) In case $S > q - \beta A$ it is obvious that starting from any point $(y_0 = b, v_0 \in [\beta A, A))$, the trajectory never touches the axis $y = 0$. The statements follow now from Corollary 4: the mappings φ and $\tilde{\varphi}$ coincide.

(b) Suppose that $S \leq q - \beta A$ and $q - S < V$, where V ($= V_1$ or V_2) is the minimal stationary point of the mapping $\tilde{\varphi}$ (see Lemma 4 and Fig.17). Then, starting from any point $(y_0 = b, v_0 \in [\beta A, A))$, at most $\varphi(\varphi(v_0))$ is such that the further trajectory never touches the axis $y = 0$: see Lemmas 3 and 4. To put it differently, $\varphi^n(v_0) > q - S$ for $n \geq 2$. The required statements again follow from Corollary 4. The mappings φ and $\tilde{\varphi}$ coincide on the domain $[q - S, A)$.

(c) Suppose that $S \leq q - \beta A$, $q - S \geq V_2$, where V_2 is the maximal stationary point of the mapping $\tilde{\varphi}$ (see Lemma 4 and Fig.17). Then, starting from any point $(y_0 = b, v_0 \in [\beta A, A))$ $\forall n \geq 2$, $\varphi^n(v_0) = \varphi(\varphi(v_0))$ because $\varphi(v_0)$, $\varphi(\varphi(v_0)) \leq V_2 \leq q - S$. Note that in terms of Theorem 1, $d < \beta(q + b)$, $V_2 = \varphi(\varphi(v_0))$ is different from (smaller than) V_2 shown on Fig.17.

(d) Suppose that $S \leq q - \beta A$, case (b) (Lemma 4) takes place and $V_1 \leq q - S \leq d$ (see Fig.17). Then, if $v_0 \in [\beta A, d]$, the situation is similar to (c): $\forall n \geq 2$ $\varphi^n(v_0) = \varphi(\varphi(v_0))$, because $\varphi(v_0)$, $\varphi(\varphi(v_0)) \leq V_1 \leq q - S \leq d$.

If $v_0 \in (d, A)$ then the trajectory never touches axis $y = 0$ because $\forall n$ $\varphi^n(v_0) = \tilde{\varphi}^n(v_0) > d \geq q - S$. The statements follow from Corollary 4.

(e) Suppose that $S \leq q - \beta A$, case (b) (Lemma 4) takes place and $d < q - S < V_2$ (see Fig.17). Then situation is similar to (b). Starting from any point $(y_0 = b,\ v_0 \in [\beta A, A))$, at most $\varphi(\varphi(v_0))$ is such that the further trajectory never touches the axis $y = 0$, the mappings φ and $\tilde{\varphi}$ coincide on the domain $[q - S, A)$ and the required statements follow from Corollary 4. ■

Corollary 1 is now obvious.

Corollary 6 *If a clipped k-cycle exists then an unclipped k-cycle exists, too. (See (20).)*

Proof. As is clear from the proof of Theorem 1, $0 < S \leq q - \beta A$ and $\varphi(q - S) = \Phi^k(q - S)$. To put it differently, the domain of Φ^k is non-empty, so that the corresponding stationary point V_1 or V_2 (Fig.17) does exist and defines the unclipped k-cycle. ■

Corollary 7 *The order of a clipped cycle can be N or $N + 1$ only (see (8)).*

Proof. Suppose all parameters are fixed, apart from b. For very small values of b, obviously, only a clipped N-cycle is realised. Conditions when a clipped $(N + 1)$-cycle exists, are left till the next subsection.

Suppose $N > 1$. When we increase b, k-cycles with $k < N$ appear: see Fig.16. If b is close (but bigger) to $\frac{\beta^k}{1-\beta^k} - q$ then the k-cycle has a very short continuous part. From the proof of Lemma 5, we have

$$\lim_{b \to \frac{\beta^k}{1-\beta^k} - q} s_0 = 0 \text{ and } \lim_{b \to \frac{\beta^k}{1-\beta^k} - q} y(s_0) = \frac{\beta^k}{1 - \beta^k} - q > 0.$$

(See (35,37)). Therefore, using Lemma 5(a) we conclude that all k-cycles remain unclipped indeed. See also Corollary 5. ■

Effects of the router buffer b.

The goal of this subsection is to justify all the statements of Section 4.

Case $A^*_{N+1} < q$ is trivial: see Fig.16, Lemmas 5,6, Corollary 7 and its proof.

Case $q \le q^*_{N+1}$. According to Lemma 6, here the $(N+1)$-cycle appears and becomes critical before it extincts at $b = A^*_{N+1} - q$.

Consider the continuous trajectory (21) staring from $(y_0 = 0, v_0 = q)$:

$$\begin{cases} y(r) = e^{-r} + r - 1; \\ v(r) = q + r. \end{cases}$$

Clearly, there is $1-1$ correspondance between parameters r and b given by equation

$$e^{-r} + r - 1 = b. \tag{43}$$

The $(N+1)$-cycle cannot be realised if

$$\beta^N(q + r + 1) < y(r) + q = e^{-r} + r - 1 + q = b + q.$$

Let us study the difference

$$\Delta(r) \overset{\triangle}{=} e^{-r} + r - 1 + q - \beta^N(q + r + 1). \tag{44}$$

Since $\frac{d\Delta(r)}{dr} = 1 - e^{-r} - \beta^N$, this difference has a minimum at $r = -\ln(1-\beta^N)$ (corresponding to $b = C$, see (10)) which equals

$$q(1-\beta^N) - 2\beta^N - (1-\beta^N)\ln(1-\beta^N) = (1-\beta^N)(q - D),$$

see (9). Since the critical $(N+1)$-cycle exists, we are sure that $q \le D$ and the values $\underline{b}$ and $\bar{b}$ (16) are well defined. These equal the minimal and the maximal values providing $\Delta(r(b)) = 0$. Here and below, $r(b)$ is the positive solution to (43). Note that the clipped $(N+1)$-cycle appears when $b = \underline{b}$ and becomes critical at $b = b_{0,N+1}$. The value $\bar{b}$ does not play any role because $\bar{b} \ge b_{0,N+1}$.

Case $q^*_{N+1} < q \le A^*_{N+1}$. Here the $(N+1)$-cycle cannot be critical (Lemma 6). According to Lemma 5, it also cannot be unclipped because unclipped cycle becomes critical when b decreases. Sometimes $(N+1)$-cycles are not realised at all. Firstly, the latter happens if $D < q$. But even if $D \ge q$, it can happen that $\underline{b} > A^*_{N+1} - q$, so that the $(N+1)$-cycle does not exist in view of Corollary 6.

Lemma 7 *Suppose* $q^*_{N+1} < q \le A^*_{N+1}$.

(a) For a given value of b*, the clipped* $(N+1)$*-cycle exists iff* $\Delta(r(b)) \le 0$ *and* $b \le A^*_{N+1} - q$.

(b) $\Delta(r(A^*_{N+1} - q)) > 0$.

(c) Suppose that $D \ge q$*. Then* $\underline{b} > A^*_{N+1} - q$ *iff* $C > A^*_{N+1} - q$*;* $\bar{b} < A^*_{N+1} - q$ *iff* $C < A^*_{N+1} - q$.

Proof. (a) The necessity is obvious: see Corollary 6 and Fig.16.

Suppose $\Delta(r(b)) \le 0$ and $b \le A^*_{N+1} - q$. For the unclipped $(N+1)$-cycle, the minimal value of y is negative; let us denote the corresponding minimal value of v by $\hat{v}$. Then, starting from $(y_0 = 0, v_0 = q)$, the trajectory (21) reaches the level $y = b$, and, after $(N+1)$ instant reductions of v, reaches point $(y = b, v < \hat{v})$. After that, the trajectory goes down up to the axis $y = 0$, and the clipped $(N+1)$-cycle is well defined.

(b) Value $b = A^*_{N+1} - q$ is the largest buffer size when the unclipped $(N+1)$-cycle exists: see Fig.16. The corresponding minimal value y_{min} is negative and, starting from $(y_0 = y_{min}, v_0 = y_{min} + q)$ trajectory (21) reaches level $y = b$ at such value of v that $\beta^N(v+1) = b + q$. Therefore, starting from $(y_0 = 0, v_0 = q)$, trajectory (21) reaches level $y = b$ at a smaller value of v, and smaller than $(N+1)$ reductions of v are needed, meaning that $\Delta(r(b)) > 0$.

(c) Obviously, $\underline{b} \le C \le \bar{b}$. Thus the necessity is trivial. The sufficiency follows from (b) because $A^*_{N+1} - q \notin [\underline{b}, \bar{b}]$. ■

Corollary 8 *In case $q^*_{N+1} < q \le A^*_{N+1}$, $q \le D$, the value of C cannot equal $A^*_{N+1} - q$.*

The proof follows directly from statement (b), Lemma 7.

Corollary 9 *Suppose N is fixed.*
(a) For all $q \in (q^*_{N+1}, A^*_{N+1}]$ the value of N remains unchanged.
(b) If $D \ge A^*_{N+1}$ then $\forall q \in (q^*_{N+1}, A^*_{N+1}]$ $C > A^*_{N+1} - q$.
(c) If $q^*_{N+1} < D < A^*_{N+1}$ then either $C > A^*_{N+1} - D$ and $\forall q \in (q^*_{N+1}, D]$ $C > A^*_{N+1} - q$, or $C < A^*_{N+1} - D$ and $\forall q \in (q^*_{N+1}, D]$ $C < A^*_{N+1} - q$.
(d) If $\beta \in (0, 1)$ varies, equality $C = A^*_{N+1} - D$ can hold only in the area where $D \le q^*_{N+1}$.

Proof. (a) The assertion follows from inequalities

$$\frac{\beta^N}{1-\beta^N} \le q^*_{N+1} < A^*_{N+1} \le \frac{\beta^{N-1}}{1-\beta^{N-1}},$$

see Lemma 2(d) and Lemma 6(b). As usual, $\frac{\beta^{N-1}}{1-\beta^{N-1}} = +\infty$ if $N = 1$.

(b) Clearly, if $q = A^*_{N+1} = \min\{A^*_{N+1}, D\}$ then $C > 0 = A^*_{N+1} - q$. If we decrease q up to q^*_{N+1}, the values of C and A^*_{N+1} remain unchanged and situation $C = A^*_{N+1} - q$ is excluded due to Corollary 8.

(c) The proof is similar to (b): take $q = D = \min\{A^*_{N+1}, D\}$ and reduce its value.

(d) In case $C = A^*_{N+1} - D$ and $D > q^*_{N+1}$ we have a contradiction to (c). ■

Theorem 2 follows directly from Section 4.

Proof of Proposition 1. According to definition (14), $\delta = \frac{\beta(1+2\beta-\tau)}{1-\beta^2}$, where τ solves equation

$$\frac{\tau(1+\beta)}{1+2\beta+\beta\tau} = 1 - e^{-\tau}.$$

The both functions on the left and on the right increase from zero, and τ is smaller than θ which solves equation $\frac{\theta(1+\beta)}{1+2\beta+\beta\theta} = 1$, ie $\tau < \theta = 2\beta + 1$. Now

$$\frac{\tau(1+\beta)}{1+2\beta+\beta\tau} < 1 - e^{-(2\beta+1)} \Longrightarrow \tau < \frac{(1+2\beta)(1-e^{-(2\beta+1)})}{1+\beta e^{-(2\beta+1)}}$$

and

$$\delta > \frac{\beta}{1-\beta^2} \cdot \frac{(1+2\beta)(\beta+1)e^{-(2\beta+1)}}{1+\beta e^{-(2\beta+1)}} \to \infty \text{ as } \beta \to 1.$$

■

Proof of Theorem 3.

First we consider the case $b \in [0, b_{0,1}]$. In this case, the cycle is clipped or critical (see Figure 10). According to Condition (b) of Theorem 2, if $q > A^*_2$ the cycle does not have multiple jumps for any size of the buffer. Without loss of generality, we assume that the zero time moment corresponds to the time moment just after the jump (Point A). Recall that we denote the transformed time by s and the original time by t. We denote by S_A the transformed time when the system reaches point A, by S_B the transformed time when the system reaches point B, and so on. Without loss of generality, we assume that $S_A = 0$. We also use the notation: $S_{AB} = S_B - S_A = S_B$, $S_{BC} = S_C - S_B$, and so on.

From (21) we have

$$y(S_C + u) = y_{CD}(u) = e^{-u} + (u-1), \quad \text{for} \quad u \in [0, S_{CD}],$$

so that

$$y(S_D) = e^{-S_{CD}} + S_{CD} - 1 = b.$$

We note that $v(S_C) = q$. Consequently, $v(S_D) = q + S_{CD}$, $v(S_E) = q + S_{CD} + 1$ and $v(S_A) = \beta(q + S_{CD} + 1)$. Again, from (21) we have

$$y(s) = (1 + q + y(S_A) - v(S_A))e^{-s} + s - 1 + v(S_A) - q,$$

and

$$y(S_B) = [y(S_A) + 1 + q - v(S_A)]e^{-S_{AB}} + [S_{AB} - 1] + v(S_A) - q.$$

Thus, we have the following equation for S_{AB}

$$[b - \beta S_{CD} + (1-\beta)(1+q)]e^{-S_{AB}} + S_{AB}$$

$$+\beta S_{CD} - (1-\beta)(1+q) = 0.$$

Now, we can calculate the cycle duration in the original and transformed times. Denote these quantities by T_{cycle} and S_{cycle}, respectively. Note that $S_{cycle} = s_1 + 1$ (see (25) with $k = 1$). From the equation $v(S_E) = v(S_A) + S_{cycle}$ we obtain

$$S_{cycle} = (1-\beta)(q + S_{CD} + 1),$$

and, consequently,

$$T_{cycle} = \int_0^{T_{cycle}} dt = \int_0^{S_{cycle}} \left(T + \frac{x(s)}{\mu}\right) ds = TS_{cycle} + \frac{m}{\mu}\left(b + \int_{S_A}^{S_B} y(s)ds + \int_{S_C}^{S_D} y(s)ds\right).$$

Next, we calculate the average queue size

$$\bar{x} = \frac{1}{T_{cycle}} \int_0^{T_{cycle}} x(t)dt = \frac{1}{T_{cycle}} \int_0^{S_{cycle}} x(s)\left(T + \frac{x(s)}{\mu}\right) ds$$

$$= \frac{1}{T_{cycle}}\left[mT\left(\int_{S_A}^{S_B} y(s)ds + \int_{S_C}^{S_D} y(s)ds\right) + \frac{m^2}{\mu}\left(\int_{S_A}^{S_B} y^2(s)ds + \int_{S_C}^{S_D} y^2(s)ds\right) + B\left(T + \frac{B}{\mu}\right)\right].$$

Now we calculate the average sending rate

$$\bar{\lambda} = \frac{1}{T_{cycle}} \int_0^{T_{cycle}} \lambda(t)dt$$

Using (2), we have

$$\bar{\lambda} = \frac{1}{T_{cycle}} \int_0^{T_{cycle}} \frac{w(t)}{T + x(t)/\mu} dt = \frac{m}{T_{cycle}} \int_0^{S_{cycle}} v(s)ds$$

$$= \frac{m}{T_{cycle}} \int_0^{S_{cycle}} (\beta(q + 1 + S_{CD}) + s)\, ds = \frac{m}{T_{cycle}} \frac{1}{2}(1-\beta^2)(q + 1 + S_{CD})^2.$$

For the calculation of the average goodput we use the following formula:

$$\bar{g} = \frac{1}{T_{cycle}}\left[\int_{T_A}^{T_D} \lambda(t)dt + \mu\left(T + \frac{B}{\mu}\right)\right] = \frac{m}{T_{cycle}}\left[\int_{S_A}^{S_D} v(s)ds + q + b\right].$$

In the case $b \in (b_{0,1}, \infty)$ the cycle is unclipped. Consequently, the calculation of the average quantities are more straightforward than in the previous case and are based on the knowledge of only one parameter S_{cycle}. ■

Proof of Proposition 2.

If $B \to \infty$ (equivalently, $b \to \infty$), then $s_1 \to \infty$ (see equation (25)). According to Theorem 3, we have

$$T_{cycle} = \frac{m}{\mu}\left[q(s_1+1) + \int_0^{s_1} y(s)ds + b\right]$$

$$= \frac{m}{\mu}\left[1 + 2b + 2q - s_1 - (1 + b + q - \frac{\beta(s_1+1)}{1-\beta})e^{-s_1} + \frac{s_1^2}{2} + \frac{\beta(s_1^2-1)}{1-\beta}\right],$$

and, consequently,

$$\Delta = \mu\frac{\frac{1+\beta}{2(1-\beta)}(2s_1+1) - 1 - 2b - 2q + s_1 + (1 + b + q - \frac{\beta(s_1+1)}{1-\beta})e^{-s_1} + \frac{\beta}{1-\beta}}{1 + 2b + 2q - s_1 - (1 + b + q - \frac{\beta(s_1+1)}{1-\beta})e^{-s_1} - \frac{\beta}{1-\beta} + \frac{1+\beta}{2(1-\beta)}s_1^2}$$

$$\sim \frac{(2 + \frac{2(1-\beta)}{1+\beta})s_1}{s_1^2} = \frac{4}{1+\beta}\frac{1}{s_1} \to 0+, \quad \text{as} \quad s_1 \to \infty.$$

■

Proof of Proposition 3. (a) Suppose N is fixed and $q = \frac{\mu T}{m}$ changes, i.e., increases starting from $\frac{\beta^N}{1-\beta^N}$. Using (11),(12) and omitting for brevity N as the power and the index, we obtain:

$$\frac{dB_0}{dm} = m\frac{db_0}{d\theta}\cdot\frac{d\theta}{dq}\cdot\frac{dq}{dm} + b_0$$

$$= m\left[\frac{1 - e^{-\theta} - \theta e^{-\theta}}{(1 - e^{-\theta})^2} \times \frac{\theta - 1 + e^{-\theta}}{\theta}\right]\left[\frac{1}{\frac{1-e^{-\theta}-\theta e^{-\theta}}{\theta(1-e^{-\theta})} + \frac{\beta}{1-\beta}}\right]\left[-\frac{\mu T}{m^2}\right]$$

$$+\frac{\theta}{1 - e^{-\theta}} - \ln\frac{\theta}{1 - e^{-\theta}} - 1.$$

We used the implicit differentiation theorem for $\frac{d\theta}{dq}$. Note that $\frac{\mu T}{m} = q$ and express q using (11):

$$\frac{dB_0}{dm} = \left[\frac{\theta}{1 - e^{-\theta}} - \ln\frac{\theta}{1 - e^{-\theta}} - 1\right]$$

$$-\left[\frac{(1 - e^{-\theta} - \theta e^{-\theta})(\theta - 1 + e^{-\theta})}{\theta(1 - e^{-\theta})^2} \times \frac{\ln\frac{\theta}{1-e^{-\theta}} + \frac{\beta(\theta+1)}{1-\beta}}{\frac{1-e^{-\theta}-\theta e^{-\theta}}{\theta(1-e^{-\theta})} + \frac{\beta}{1-\beta}}\right].$$

The second square bracket, $f(\frac{\beta}{1-\beta})$ is a monotonous function of $\frac{\beta}{1-\beta}$.

(α) If $(\theta+1)\frac{1-e^{-\theta}-\theta e^{-\theta}}{\theta(1-e^{-\theta})} - \ln\frac{\theta}{1-e^{-\theta}} \geq 0$ then $f(\cdot)$ does not decrease and hence

$$\frac{dB_0}{dm} \leq \frac{\theta}{1 - e^{-\theta}} - \ln\frac{\theta}{1 - e^{-\theta}} - 1 - f(0)$$

$$= \frac{\theta}{1 - e^{-\theta}} - \ln\frac{\theta}{1 - e^{-\theta}} - 1 - \frac{\ln\frac{\theta}{1-e^{-\theta}}\cdot(\theta - 1 + e^{-\theta})}{1 - e^{-\theta})}$$

$$= \frac{\theta}{1-e^{-\theta}} - 1 - \frac{\theta}{1-e^{-\theta}} \ln \frac{\theta}{1-e^{-\theta}} = \gamma - 1 - \gamma \ln \gamma < 0,$$

because $\gamma \overset{\triangle}{=} \frac{\theta}{1-e^{-\theta}} \in (1, \infty)$ for $\theta > 0$ and function $\gamma - 1 - \gamma \ln \gamma$ has the maximum which is equal to zero at $\gamma = 1$.

(β) If

$$(\theta+1)\frac{1-e^{-\theta}-\theta e^{-\theta}}{\theta(1-e^{-\theta})} - \ln \frac{\theta}{1-e^{-\theta}} < 0 \tag{45}$$

then $f(\cdot)$ decreases and hence

$$\frac{dB_0}{dm} < \frac{\theta}{1-e^{-\theta}} - \ln \frac{\theta}{1-e^{-\theta}} - 1 - \lim_{y\to\infty} f(y)$$

$$= \frac{\theta}{1-e^{-\theta}} - \ln \frac{\theta}{1-e^{-\theta}} - 1 - \frac{(1-e^{-\theta}-\theta e^{-\theta})(\theta-1+e^{-\theta})(\theta+1)}{\theta(1-e^{-\theta})^2}.$$

Using (45), we have

$$\frac{dB_0}{dm} < \frac{\theta}{1-e^{-\theta}} - \frac{(\theta+1)(1-e^{-\theta}-\theta e^{-\theta})}{\theta(1-e^{-\theta})} - 1$$

$$- \frac{(1-e^{-\theta}-\theta e^{-\theta})(\theta-1+e^{-\theta})(\theta+1)}{\theta(1-e^{-\theta})^2}$$

$$= \frac{3e^{-\theta}+\theta^2 e^{-\theta}+\theta e^{-\theta}-2-e^{-2\theta}}{(1-e^{-\theta})^2} < 0, \text{ if } \theta > 0.$$

Indeed, consider function $g(\theta) = 3e^{-\theta}+\theta^2 e^{-\theta}+\theta e^{-\theta}-2-e^{-2\theta}$. Clearly $g(0) = 0$;

$$\frac{dg}{d\theta} = e^{-\theta}[\theta + 2e^{-\theta} - 2 - \theta^2];$$

$$\left.\frac{dg}{d\theta}\right|_{\theta=0} = 0; \qquad \frac{d[\theta + 2e^{-\theta} - 2 - \theta^2]}{d\theta} = 1 - 2e^{-\theta} - 2\theta < 0$$

because the latter function decreases starting from -1 at $\theta = 0$.

Note that

$$\frac{dB_0}{dm} \to 0- \text{ as } \theta \to 0+ \tag{46}$$

(b) Obviously, without loss of generality we can put $N = 1$ and prove that $B_{0,N}$ increases as $\beta \in (0,1)$ decreases. Like previously, we omit N as the power and the index. Now again using the implicit differentiation theorem we obtain

$$\frac{dB_0}{d\beta} = m\frac{db_0}{d\theta} \cdot \frac{d\theta}{d\beta}$$

$$= m\left[\frac{(1-e^{-\theta}-\theta e^{-\theta})(\theta-1+e^{-\theta})}{(1-e^{-\theta})^2\theta}\right]\left[\frac{-\frac{1+\theta}{(1-\beta)^2}}{\frac{1-e^{-\theta}-\theta e^{-\theta}}{\theta(1-e^{-\theta})} + \frac{\beta}{1-\beta}}\right] < 0.$$

■

<u>Proof of Corollary 3.</u> The first part follows directly from Proposition 3, if we notice that N remains unchanged on intervals $m \in \left[\frac{\mu T(1-\beta^{N-1})}{\beta^{N-1}}, \frac{\mu T(1-\beta^N)}{\beta^N}\right)$ and increases by 1 at points m_{i+1}.

When $m \to m_N - 0$, q approaches $\frac{\beta^N}{1-\beta^N}$ and θ_N goes to zero (see (15)). According to (12), $b_{0,N} \to 0+$, hence $B_{0,N} = mb_{0,N} \to 0+$. Equality (46) implies that $\frac{dB_{0,N}}{dm} \to 0-$.

Suppose $m \to 0+$, $q \to \infty$, $N = 1$, $\theta_1 \to \infty$. Then $\frac{\ln \frac{\theta_1}{1-e^{-\theta_1}}}{\theta_1} \to 0$ and $\frac{q}{\theta_1} \to \frac{\beta}{1-\beta}$ according to (11). Therefore

$$B_{0,1} = mb_{0,N} = \frac{\mu T}{q}\theta_1 \left(\frac{b_{0,1}}{\theta_1}\right) \to \frac{\mu T(1-\beta)}{\beta}.$$

Consider $B_{0,N}(m_{N-1})$, ie put $q = \frac{\beta^{N-1}}{1-\beta^{N-1}}$ and study equations (11),(12). Let θ_N be the positive solution to

$$\ln \frac{\theta}{1-e^{-\theta}} + \frac{\beta^N}{1-\beta^N} \cdot \theta = \frac{\beta^{N-1}}{1-\beta^{N-1}} - \frac{\beta^N}{1-\beta^N}; \tag{47}$$

then

$$B_{0,N}(m_{N-1}) = \frac{\mu T(1-\beta^{N-1})}{\beta^{N-1}} \left[\frac{\theta_N}{1-e^{-\theta_N}} - \ln \frac{\theta_N}{1-e^{-\theta_N}} - 1\right].$$

Obviously, $\lim_{N\to\infty} \theta_N = 0$, hence, directly from (47) we obtain:

$$\lim_{N\to\infty} \left[\frac{\ln \frac{\theta_N}{1-e^{-\theta_N}}}{\theta_N} \cdot \frac{\theta_N}{\beta^N} + \frac{\theta_N}{1-\beta^N}\right] = \frac{1}{2} \lim_{N\to\infty} \frac{\theta_N}{\beta^N} = \lim_{N\to\infty} \left[\frac{\beta^{-1}}{1-\beta^{N-1}} - \frac{1}{1-\beta^N}\right] = \frac{1-\beta}{\beta}$$

and finally

$$\lim_{N\to\infty} m_{N-1}B_{0,N}(m_{N-1}) = \lim_{N\to\infty} \left[\left(\frac{\theta_N}{1-e^{-\theta_N}} - \ln \frac{\theta_N}{1-e^{-\theta_N}} - 1\right) \Big/ \theta_N^2\right]$$

$$\times \lim_{N\to\infty} \left(\frac{\theta_N}{\beta^N}\right)^2 \beta^2(\mu T)^2 \lim_{N\to\infty} (1-\beta^{N-1})^2$$

$$= \frac{1}{8}\left[\frac{2(1-\beta)}{\beta}\right]^2 \beta^2(\mu T)^2 = \frac{1}{2}(1-\beta)^2(\mu T)^2.$$

■