Centrum Wiskunde & Informatica

REPORTRAPPORT

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Design issues of a back-pressure-based congestion control mechanism

R. Malhotra, M.R.H. Mandjes, W.R.W. Scheinhardt,
J.L. van den Berg

REPORT PNA-E0804 JULY 2008

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Design issues of a back-pressure-based congestion control mechanism

ABSTRACT
Congestion control in packet-based networks is often realized by feedback protocols -- in this paper we assess the performance under a back-pressure mechanism that has been proposed and standardized for Ethernet metropolitan networks. Relying on our earlier results for feedback fluid queues, we derive explicit expressions for the key perfomance metrics, in terms of the model parameters, as well as the parameters agreed upon in the service level agreement. Numerical experiments are performed to evaluate the main trade-offs of this model (for instance the trade-off between the signaling frequency and the throughput). These can be used to generate design guidelines. The paper is concluded by an elementary, yet powerful, Markovian model that can be used as an approximative model in situations of large traffic aggregates feeding into the system; the trade-offs and guidelines identified for the feedback fluid model turn out to carry over to this more stylized model.

2000 Mathematics Subject Classification: 60K25, 90B18
Keywords and Phrases: Fluid models, feedback, Ethernet

Design issues of a back-pressure-based congestion control mechanism

Richa Malhotra, Michel Mandjes, Werner Scheinhardt, and Hans van den Berg [*]

July 2, 2008

Abstract

Congestion control in packet-based networks is often realized by feedback protocols – in this paper we assess the performance under a back-pressure mechanism that has been proposed and standardized for Ethernet metropolitan networks. Relying on our earlier results for feedback fluid queues, we derive explicit expressions for the key perfomance metrics, in terms of the model parameters, as well as the parameters agreed upon in the service level agreement. Numerical experiments are performed to evaluate the main trade-offs of this model (for instance the trade-off between the signaling frequency and the throughput). These can be used to generate design guidelines. The paper is concluded by an elementary, yet powerful, Markovian model that can be used as an approximative model in situations of large traffic aggregates feeding into the system; the trade-offs and guidelines identified for the feedback fluid model turn out to carry over to this more stylized model.

[*] RM is with Alcatel-Lucent, Capitool 5, 7521 PL Enschede, the Netherlands, `rimalhotra@alcatel-lucent.com`, and also with University of Twente, the Netherlands. MM is with Korteweg-de Vries Institute for Mathematics, Plantage Muidergracht 24, 1018 TV Amsterdam, the Netherlands, `mmandjes@science.uva.nl`, and also with CWI, Amsterdam, the Netherlands, and EURANDOM, Eindhoven, the Netherlands; part of this work was done while he was at Stanford University, Stanford, CA 94305, US. WS is with University of Twente, Faculty of Electrical Engineering, Mathematics, and Computer Science, P.O. Box 217, 7500 AE Enschede, the Netherlands, `w.r.w.scheinhardt@utwente.nl`, and also with CWI, Amsterdam, the Netherlands. HvdB is with TNO ICT, P.O. Box 5050, 2600 GB Delft, the Netherlands, `j.l.vandenberg@tno.nl`, and also with University of Twente, the Netherlands.

1 Introduction

Over the past decades a broad variety of mechanisms has been proposed to control congestion in packet networks. A well-known example is *random early detection*, as proposed in e.g. [4], where incipient congestion is notified to the users by dropping packets (or by setting a bit in packet headers); when the queue size exceeds a preset threshold, each arriving packet is dropped (or marked) with a certain probability that depends on the buffer content and its evolution in the recent past; for more insights into this type of schemes, see e.g. [5] and the early reference [13]. Similar feedback-based mechanisms have been proposed and standardized for congestion control in *Ethernet metropolitan networks.* The back-pressure scheme defined in IEEE 802.3x [6], is intended to provide flow control on a hop-by-hop basis by allowing ports to *turn off* their upstream link neighbors for a period of time. For a full-duplex connection, this mechanism is based on a special frame called *pause frame* in which the pause period is specified. The end-station (or router) receiving the pause frame looks at the pause period, and does not transmit or attempt transmission for that amount of time. Alternatively, an ON/OFF pause message can be sent signaling the beginning and end of the transmission pause phase. Importantly, this congestion control method is usually implemented by using *two* thresholds, viz. a high threshold to detect the onset of a congestion period, and a low threshold to detect its end. When the queue occupancy exceeds the high threshold the *PauseOn* message is sent and transmission is temporarily stopped; when the queue occupancy drops below the low threshold the *PauseOff* message is sent and consequently transmission is resumed.

There are hardly any performance evaluation studies available on the above-described back-pressure mechanisms for Ethernet congestion control. Previous works [8, 10, 12] predominantly concentrated on the throughput gain which can be achieved. Recently, however, we have been able to develop a rather detailed, and analytically tractable, model of the mechanism [9]. This model belongs to the class of fluid models [1, 7], in which the steady-state distribution of the buffer content is expressed in terms of the solution of a system of linear differential equations, which, after imposing the proper boundary conditions, can be solved by standard techniques from linear algebra. Models with a single threshold to signal both the onset and end of congestion had been analyzed before, see e.g. [11], but it turned out that the analysis complicated substantially due to the fact that we have *two* thresholds in our back-pressure model. The main (mathematical) difficulties related to (i) the behavior of the storage level close to the thresholds, and (ii) the generation of a sufficient number of boundary conditions to solve for the remaining unknown constants; [9] provides us with a solution to these problems, resulting in a procedure to numerically determine the steady-state buffer content distribution.

In [9] it was mentioned that the model presented (and solved) there could be relied on when configuring the high and low thresholds, thus addressing a pivotal design criterion for the Ethernet congestion avoidance scheme. We also remarked in [9] that the back-pressure scheme has the attractive property that the signaling overhead (in terms of the number of pause messages sent per unit time) is lower than when using just one threshold (that detects both start and end of congestion periods), but we did not systematically quantify this effect. Also, the reduction of signaling overhead may be at the expense of a loss in throughput, or degraded performance in terms of delay. The primary goal of the present paper is to demonstrate the effect of the thresholds, and to obtain insight into the trade-offs mentioned above. In order to do so, we also derive analytic formulas for the performance metrics of interest. In a substantial part of the paper we focus on the single-source model, that may be viewed as a benchmark model that provides useful insights. Later in the paper we also introduce a model for the multiple-source case that indicates that most of the effects observed in the single-source model carry over to considerably more general settings.

The organization of this paper is as follows. Section 2 describes our fluid model, specializing to the situation of just one source feeding into the queue. It also recapitulates the main results from [9]. Then Section 3 presents derivations of the main performance metrics considered in this paper: the throughput, the the mean packet delay, signaling frequency, and the mean transmission time of a burst of packets. Here we note that packet delays are of crucial interest for *streaming* applications; these generate traffic with an 'intrinsic duration and rate (which is generally variable) whose time integrity must be preserved by the network' [14] — think of telephony, streaming video and audio. On the other hand, the transmission time, to be thought of as the time it takes for bursts of packets ('jobs') to go through a node, is a main performance metric for *elastic* applications, such as email, file transfer, but also pictures or video sequences transferred for local storage before viewing. Section 4 presents the numerical experiments that demonstrate how to evaluate the trade-offs mentioned above, and presents a number of general guidelines. We also include in Section 5 a model and corresponding numerical experiments that indicate that the main findings carry over to the situation in which there is a substantial number of concurrent users. Section 6 concludes.

2 Model and preliminaries

In this section we describe the model of which we analyze a number of key performance metrics in Section 3, and which we numerically assess in Section 4. To this end, we first define generator matrices Q^+ and Q^- on the state space $\{1,2\}$:

$$Q^+ = \begin{pmatrix} -p_1 & p_1 \\ p_2 & -p_2 \end{pmatrix}; \quad Q^- = \begin{pmatrix} -m_1 & m_1 \\ m_2 & -m_2 \end{pmatrix}.$$

Also, we introduce traffic rate vectors $\boldsymbol{r}^+ = (r_p, 0)^{\mathrm{T}}$ and $\boldsymbol{r}^- = (r_m, 0)^{\mathrm{T}}$, with $r_p > c$ and $r_m > c$; these should be thought of as rates at which traffic is generated, in that traffic flows into the system at rate r_i^ℓ if a background process $X^\ell(\cdot)$, governed by generator matrix Q^ℓ, is in state i (with $\ell \in \{+,-\}$, and $i \in \{1,2\}$). In other words, we identify the on-state with state 1 ('burst'), and the off-state with state 2 ('silence'). The capacity of the buffer is assumed to be infinite (a similar analysis can be done for the finite-buffer case, though).

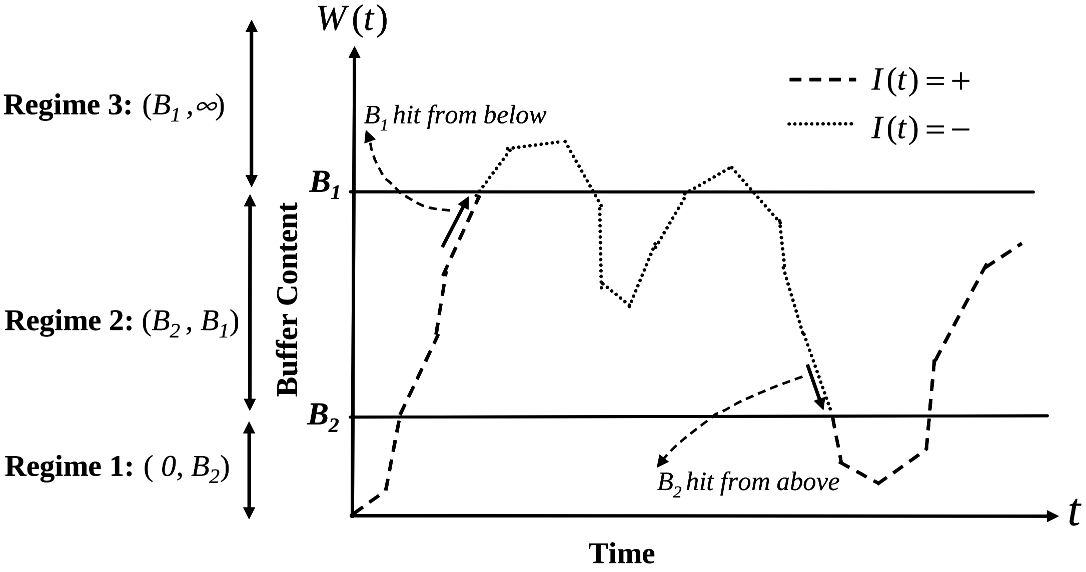

Figure 1: Schematic illustration of different regimes for the buffer content $W(t)$.

In this paper we consider the model of [9], featuring the special case that the dimension of the underlying sources is 2. In this feedback fluid model, the input stream alternates between two 'modes' (also referred to as 'phases'). In one mode the input process behaves like a Markov fluid source with generator Q^{+} and traffic rate vector $\boldsymbol{r}^{+}$: when the background process is in state $i \in \{1, 2\}$ at time t, traffic is generated at a constant rate r_i^{+}, whereas the queue is drained at a constant rate c. Similarly, in the other mode it behaves like a Markov fluid source with generator Q^{-} and traffic rate vector $\boldsymbol{r}^{-}$.

The queueing process alternates between the two above-mentioned modes as follows. We first introduce the indicator variable process $I(\cdot)$, taking values in $\{+, -\}$, which gives the current mode of operation of the input source. It is important to note that whenever $I(t)$ switches from one mode to another, the background process $X(t)$ stays in the same state; only its dynamics will from that time onwards behave according to the other generator matrix. However, the rate at which the fluid buffer receives fluid *does* change instantaneously from r_i^{+} to r_i^{-} (or vice versa), when the background process $X(t)$ is in state i at the switching instant. Which of the two modes is currently valid at some time t depends on the behavior of the content process $W(t)$ relative to two thresholds, an upper threshold B_1 and a lower threshold B_2. The first mode ('+') applies as long as $W(t)$ has not reached the upper threshold B_1 from below. As soon as that happens, $I(t)$ switches to the other mode ('−'), until $W(t)$ hits the lower threshold B_2 from above, etc. The queueing dynamics are illustrated by Fig. 1.

It is not hard to verify that the equilibrium condition of this model is

$$\frac{m_2}{m_1 + m_2} \cdot r_m < c,$$

i.e., in the '$-$'-phase there should be a negative drift. We let

$$F_i^\ell(x) := \mathbb{P}(I = \ell, X = i, W \le x),$$

with $x \ge 0$, $i \in \{1,2\}$, and $\ell \in \{+,-\}$, be the steady-state distribution of the workload W, jointly with the state of the background process $X \in \{1,2\}$, and the phase $I \in \{+,-\}$. In [9] we presented an algorithm to compute $F_i^\ell(\cdot)$, as follows. Let z^ℓ ($\ell \in \{+,-\}$) be the non-zero eigenvalue of the matrix $Q^\ell(R^\ell - cI)^{-1}$. It is easily verified that

$$z^+ = \frac{p_2}{c} - \frac{p_1}{r_p - c}, \text{ and } z^- = \frac{m_2}{c} - \frac{m_1}{r_m - c}.$$

Notice that $z^- < 0$ because of the stability condition. Then the analysis in [9] entails that 3 regimes should be distinguished, cf. Fig. 1. More precisely, there are constants $\gamma_{j,i}^\ell$, $\delta_{j,i}^\ell$, $\varepsilon_{j,i}^\ell$ (with regime $j \in \{1,2,3\}$, state $i \in \{1,2\}$, and mode $\ell \in \{-,+\}$), such that

$$\begin{aligned} F_i^-(x) &= 0, \quad x \le B_2; \\ F_i^-(x) &= \gamma_{2,i}^- + \delta_{2,i}^- e^{z^- x} + \varepsilon_{2,i}^- x, \quad B_2 < x \le B_1; \\ F_i^-(x) &= \gamma_{3,i}^- + \delta_{3,i}^- e^{z^- x}, \quad x > B_1; \end{aligned}$$

also

$$\begin{aligned} F_i^+(x) &= \gamma_{1,i}^+ + \delta_{1,i}^+ e^{z^+ x}, \quad x \le B_2; \\ F_i^+(x) &= \gamma_{2,i}^+ + \delta_{2,i}^+ e^{z^+ x} + \varepsilon_{2,i}^+ x, \quad B_2 < x \le B_1; \\ F_i^+(x) &= F_i^+(B_1), \quad x > B_1. \end{aligned}$$

In [9] a procedure is detailed that enables us to compute these 10 constants, by introducing 10 linear constraints to be imposed on the parameters.

3 Performance metrics

In this section we derive (or recall) formulas for a number of performance metrics.

Throughput. We have the evident formula, already given in [9],

$$\vartheta = r_p \cdot F_1^+(\infty) + r_m \cdot F_1^-(\infty).$$

Alternatively, it is clear that the througput can be written as (realize that $F_1^+(0) = 0$)

$$\vartheta = c \cdot \mathbb{P}(W > 0) = c\left(1 - F_2^+(0)\right). \tag{1}$$

Packet delays. The delay D is defined as the delay experienced by an arbitrary packet (in our model an infinitesimally small 'fluid particle'), and is hence a 'traffic-average'. This performance metric is particularly relevant for streaming traffic, as argued in the introduction, due to its inherent time-integrity requirements. The distribution of D was given in [9]:

$$\mathbb{P}(D \leq t) = \frac{r_p F_1^+(tc) + r_m F_1^-(tc)}{r_p F_1^+(\infty) + r_m F_1^-(\infty)};$$

note that the denominator can be interpreted as the average amount of fluid that arrives per unit of time, whereas the numerator is the fraction thereof that corresponds to a delay smaller than t. The mean delay can be computed as

$$\mathbb{E}D = \int_0^\infty \mathbb{P}(D > t)\mathrm{d}t = \int_0^\infty \left(1 - \frac{r_p F_1^+(tc) + r_m F_1^-(tc)}{r_p F_1^+(\infty) + r_m F_1^-(\infty)}\right) \mathrm{d}t.$$

Signaling frequency. The signaling frequency is defined as the expected number of phase transitions per unit time, and is a measure for the signaling overhead. With $f_i^\ell(x) := \mathrm{d}F_i^\ell(x)/\mathrm{d}x$, we first observe that the expected number of upcrossings per unit time through level x is, reasoning as in, e.g., [2, 15],

$$f_1^+(x) \cdot (r_p - c) + f_1^-(x) \cdot (r_m - c); \tag{2}$$

here the first (second) term reflects the number of upcrossings while in the '$+$'-phase ('$-$'-phase). Likewise the expected number of downcrossings per unit time is given by

$$f_2^+(x+) \cdot c + f_2^-(x+) \cdot c. \tag{3}$$

As an aside we mention that, as argued in [2, 15], expressions (2) and (3) should match, since for any level the mean number of upcrossings per unit time equals the mean number of downcrossings per unit time.

Relying on the above reasoning it is now directly seen that the expected number of phase-transitions per unit time equals

$$\varphi := f_1^+(B_1) \cdot (r_p - c) + f_2^-(B_2) \cdot c = 2f_1^+(B_1) \cdot (r_p - c) = 2f_2^-(B_2) \cdot c;$$

here the $f_1^+(B_1) \cdot (r_p - c)$ term corresponds to the number of upcrossings per unit of time through B_1 while in (to be understood as 'coming from') the '$+$'-phase, and the $f_2^-(B_2) \cdot c$ term to the number of downcrossings per unit of time through B_2 while in (i.e., coming from) the '$-$'-phase. It is further noted the last two equalities are due to the fact that the number of upcrossings per unit time through B_1 while in the '$+$'-mode should match the number of downcrossings per unit time through B_2 while in the '$-$'-mode.

Transmission and sojourn time. The next performance metric, T, is the transmission time of a burst, i.e., the time it takes to put the entire burst into the buffer. Let $f_T(\cdot)$ be the density of T. Consider the event $\{T = x\}$. We list three useful properties:

- A first observation is that if $x > B_1/(r_p - c)$, the system must have been in the '$-$'-phase during at least part of the transmission time (as the buffer content grows at rate $r_p - c$ while in the '$+$'-phase).

- A second observation is the following. Suppose the elastic job enters the system when there is y in the buffer. If x is larger than $(B_1 - y)/(r_p - c)$ and the phase is '+', then the phase shifts from '+' to '$-$' during the transmission time.

- A third observation is that if the phase is '$-$' upon arrival, the phase remains '$-$' during the entire transmission time.

It leads to the following expression, with $f^\ell(\cdot)$ the density of the buffer content *seen by an arriving job*, intersected with being in the ℓ-phase, $\ell \in \{+, -\}$:

$$
\begin{aligned}
f_T(x) &= \int_0^{\max\{B_1-(r_p-c)x,0\}} p_1 e^{-p_1 x} f^+(y)\mathrm{d}y \\
&+ \int_{\max\{B_1-(r_p-c)x,0\}}^{B_1} \exp\left(-p_1 \cdot \frac{B_1 - y}{r_p - c}\right) \cdot m_1 \exp\left(-m_1\left(x - \frac{B_1 - y}{r_p - c}\right)\right) f^+(y)\mathrm{d}y \\
&+ \int_{B_2}^{\infty} m_1 e^{-m_1 x} f^-(y)\mathrm{d}y; \qquad (4)
\end{aligned}
$$

the first term corresponds to the situation in which the queue was in the '+'-phase at the arrival epoch of the burst, and remains in the '+'-phase during the transmission time, whereas in the second term the queue makes a transition to the '$-$'-phase during the transmission time; in the third term the queue was in the '$-$'-phase at the arrival epoch of the burst, and remains (automatically) in the '$-$'-phase during the transmission time. From the density, the mean transmission time $\mathbb{E}T$ can be computed.

It now remains to identify $f^+(y)$ and $f^-(y)$. As a burst enters while the source is in the off-state, i.e., $X = 2$, and taking into account the different rates at which the source can transmit when switching on,

$$
f^+(y) = \frac{f_2^+(y)p_2}{\int_0^\infty (f_2^+(x)p_2 + f_2^-(x)m_2)\mathrm{d}x}; \quad f^-(y) = \frac{f_2^-(y)m_2}{\int_0^\infty (f_2^+(x)p_2 + f_2^-(x)m_2)\mathrm{d}x}.
$$

The numerator of $f^+(y)$ is to be interpreted as the rate at which the source turns on while the phase is '+' and the buffer is y, whereas the denominator is the rate at which the source turns on, irrespective of the phase and buffer content; the expression for $f^-(y)$ can be interpreted likewise.

We now see how the formulas change when we do not consider the time it takes before the burst is stored in the buffer, but instead the time before the entire burst has left the queue, which we will refer to as the sojourn time S. This random variable is most easily expressed in terms of its Laplace transform. We have to distinguish between the same three cases as in (4). Regarding the first term, observe that if the initial buffer level is y and the on-time is x, the entire burst has left the queue after

$$
x + \frac{y + (r_p - c)x}{c} = \frac{y}{c} + \frac{r_p x}{c}
$$

units of time. Regarding the second term, the amount of traffic in the buffer at the end of the transmission time is

$$
B_1 + (r_m - c)\left(x - \frac{B_1 - y}{r_p - c}\right),
$$

and hence the sojourn time is

$$x+\frac{1}{c}\left(B_1+(r_m-c)\left(x-\frac{B_1-y}{r_p-c}\right)\right)=\frac{r_mx}{c}+\frac{r_p-r_m}{r_p-c}\frac{B_1}{c}+\frac{r_m-c}{r_p-c}\frac{y}{c}.$$

Regarding the third term, then the sojourn time is

$$x+\frac{y+(r_m-c)x}{c}=\frac{y}{c}+\frac{r_mx}{c}.$$

We thus obtain

$$\begin{aligned}\mathbb{E}e^{-\alpha S}&=\int_0^\infty\int_0^{\max\{B_1-(r_p-c)x,0\}}p_1e^{-p_1x}f^+(y)\exp\left(-\alpha\left(\frac{y}{c}+\frac{r_px}{c}\right)\right)\mathrm{d}y\mathrm{d}x\\&+\int_0^\infty\int_{\max\{B_1-(r_p-c)x,0\}}^{B_1}\exp\left(-p_1\cdot\frac{B_1-y}{r_p-c}\right)\cdot m_1\exp\left(-m_1\left(x-\frac{B_1-y}{r_p-c}\right)\right)f^+(y)\\&\qquad\exp\left(-\alpha\left(\frac{r_mx}{c}+\frac{r_p-r_m}{r_p-c}\frac{B_1}{c}+\frac{r_m-c}{r_p-c}\frac{y}{c}\right)\right)\mathrm{d}y\mathrm{d}x\\&+\int_0^\infty\int_{B_2}^\infty m_1e^{-m_1x}f^-(y)\exp\left(-\alpha\left(\frac{y}{c}+\frac{r_mx}{c}\right)\right)\mathrm{d}y\mathrm{d}x.\end{aligned}$$

By differentiating, inserting $\alpha := 0$, and multiplying with -1, we obtain $\mathbb{E}S$. The formulas do not provide much additional insight, and we have decided to omit them here.
The transmission time and sojourn time are specifically meaningful in the case of elastic traffic. Then we let the size of the elastic job (in, say, bits) be exponentially distributed with mean μ^{-1}, and choose $p_1=\mu r_p$ and $m_1=\mu r_m$. In this situation, the amount of traffic to be sent has a fixed distribution (viz. exponentially with mean μ^{-1}). The mean sojourn time reads

$$\mathbb{E}S=\frac{1}{c}\int_0^\infty y(f^+(y)+f^-(y))\mathrm{d}y+\frac{1}{\mu c},$$

where the first term represents the mean amount of time needed to serve all traffic the tagged job sees in the queue upon arrival, and the second term the time needed to serve the tagged job itself.

Multi-dimensional sources. The above results can be extended to sources with dimension higher than 2 (and hence also to the situation of multiple sources), as the model of [9] presents the steady-state distribution for any dimension of the underlying Markov fluid source; in fact, the formulas for the throughput and the (packet-)delay distribution were already given in [9]. The formula for the signaling frequency follows along the same lines as sketched above, by an upcrossings/downcrossings argument, where all states should be taken into account in which B_1 can be reached from below while being in the '+'-phase, as well as all states in which B_2 can be reached from above while being in the '$-$'-phase.

4 Numerical experiments

In this section we describe a number of experiments, that assess the impact of the model parameters on the performance. Four key metrics are considered, viz. (i) throughput, (ii) signaling frequency, (iii) expected (packet) delay (streaming traffic), (iv) expected transmission time (elastic traffic). We then indicate how our model can be used in the design of the back-pressure system, or, more specifically, when selecting suitable values for the thresholds. The last part of the section addresses an alternative model that can be used in case of larger aggregates feeding into the queue.

4.1 Experiments

Experiment I: Effect of the thresholds – streaming traffic. In this first experiment we study the effect of the thresholds on the performance in case of streaming traffic. In [10] we found (for a considerably more stylized model) that, for a given value of the upper threshold B_1, the throughput was maximized by choosing the lower threshold B_2 as closely as possible to B_1. What we did not address in [10] is to what extent this affects the signaling frequency, packet delays, and transmission times.

In this example we chose the following parameters, with $c = 10$:

$$Q^+ = Q^- = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}; \quad \boldsymbol{r}^+ = \begin{pmatrix} 25 \\ 0 \end{pmatrix}; \quad \boldsymbol{r}^- = \begin{pmatrix} 15 \\ 0 \end{pmatrix}.$$

Remark that this situation is typical for a streaming user: when there is low (high, respectively) congestion, it is allowed to transmit at a high (low) rate, but the generator matrices, i.e., Q^+ and Q^-, are *not* affected by the level of congestion. In other words: a sample-path of the process consists of a sequence of on- and off-times. The results are presented in Fig. 2. It is noted that the mean buffer content and the mean packet delay can be easily translated in one another, noticing that (due to Little's formula) the mean buffer content equals the product of the throughput and the mean packet delay. This motivates why we have chosen to show just the throughput and the mean packet delay, and to leave out the mean buffer content; the reader can compute the mean buffer content easily. We mention that in all our experiments the mean buffer content showed the same qualitative behavior as the mean packet delay.

Consider the situation of a fixed value of B_1, and compare the situations of (A) $B_2 < B_1$ and (B) $B_2 = B_1$. From the graphs we will see that, compared to situation (B), under (A) the throughput, signaling frequency, and mean packet delay are lower. In other words: there is a trade-off between throughput on one hand, and signaling frequency and mean packet delay on the other hand.

These trends can be explained as follows. First observe that epochs at which the buffer content is B_1 and the phase jumps from '+' to '−' are regeneration epochs, in that the process *probabilistically* starts all over. Let time 0 be such a regeneration epoch, and let $W_{\rm A}(t)$ be the workload process in situation (A), and $W_{\rm B}(t)$ the workload in situation (B). Then it is seen that $W_{\rm A}(t) \leq W_{\rm B}(t)$ sample-path-wise, and hence $\mathbb{P}(W_{\rm A} = 0) \geq \mathbb{P}(W_{\rm B} = 0)$, and hence, according to (1), the throughput is indeed lower under (A) than under (B). Likewise, it can be argued that regeneration cycles last shorter under (B), and as there are two signals per regeneration period, the signaling frequency

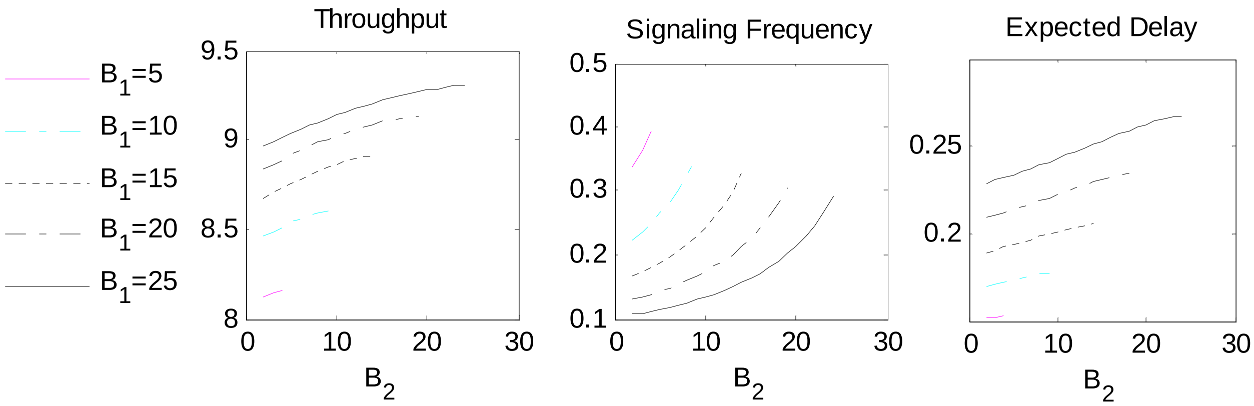

Figure 2: Effect of thresholds on streaming traffic.

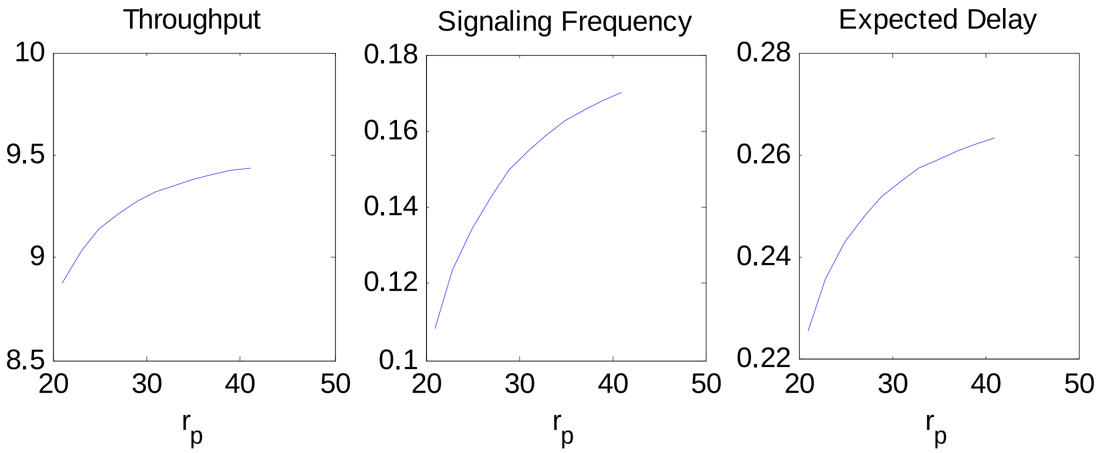

Figure 3: Effect of transmission rate r_p on streaming traffic.

under (A) is lower than under (B). With a similar argumentation, it also follows that the mean packet delay is lower under (A) than under (B).

Experiment II: Effect of the transmission rate – streaming traffic. In this experiment we study the effect of the peak rate r_p on the performance. In the service level agreement, typically the r_p will be specified. The effect of having a higher r_p is the following. Observe that regeneration periods become shorter when r_p increases, and hence the signaling rate increases. Also (on a sample-path basis) the workload process increases in r_p, leading to a higher throughput and mean packet delay. Hence, we see a similar effect as in Experiment I. Doubling the peak rate r_p, though, does clearly not lead to doubling the throughput. Remark that it may, at first glance, be slightly counterintuitive that the performance in term of packet delay *degrades* when increasing r_p, but this effect is due to the fact that the buffer content increases. In the numerical experiment, we use the parameters of Experiment I (except that we vary the value of r_p). We chose $B_1 = 25$ and

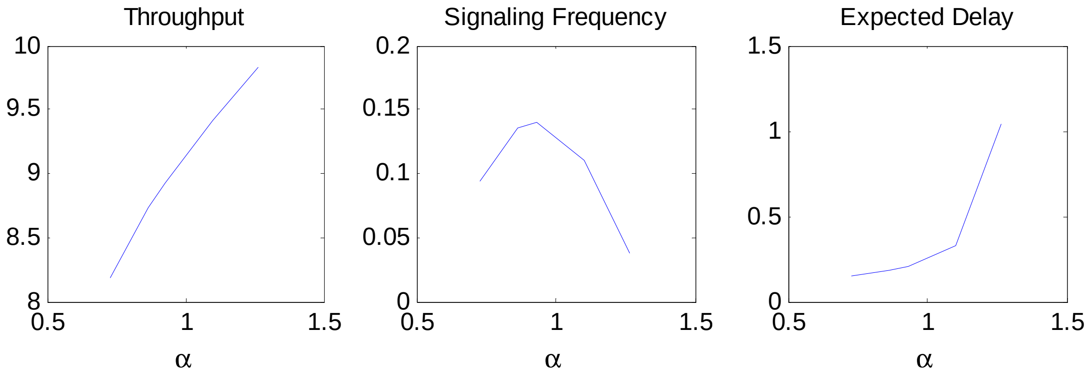

Figure 4: Effect of multiplying r_p and r_m by factor α (α is such that the stability condition is satisfied).

$B_2 = 10$; the graphs are shown in Fig. 3.

We also include a related experiment here, where both r_p and r_m are multiplied by α (but α is such that the stability condition remains fulfilled). Now the '$+$'-phase lasts shorter, while the '$-$'-phase lasts longer. Hence it can be argued that both the delay and throughput increase when r_p and r_m grow, but it is not *a priori* clear what happens with the signaling frequency. The results are presented in Fig. 4; it is seen that the signaling frequency shows non-monotone behavior in α.

Notice that the above insights are of interest for the user. The r_p is the fastest rate he can transmit at, whereas the r_m can be regarded as some minimally guaranteed transmission rate. These are rates that are agreed upon in the service level agreement. Clearly, the higher the transmission rates, the more the customer will be charged. The figures may guide the user in choosing his r_p and r_m, taking into account this trade-off.

Experiment III: Effect of the thresholds – elastic traffic. In this third experiment we study the effect of the thresholds on the performance for the case of elastic traffic. We wonder if, in order to maximize the throughput, just as in the case of streaming traffic, it is again optimal to choose $B_2 = B_1$; we are also interested in the impact of the choice of the thresholds on the other performance metrics.

In this example we chose the following parameters, with $c = 10$:

$$Q^+ = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}; \quad Q^- = \begin{pmatrix} -\frac{3}{5} & \frac{3}{5} \\ 1 & -1 \end{pmatrix}; \quad \boldsymbol{r}^+ = \begin{pmatrix} 25 \\ 0 \end{pmatrix}; \quad \boldsymbol{r}^- = \begin{pmatrix} 15 \\ 0 \end{pmatrix}.$$

Remark that this situation is typical for an elastic user: when there is low (high, respectively) congestion, it is allowed to transmit at a high (low) rate, but the generator matrices, i.e., Q^+ and Q^-, are now adapted too in order to reflect the fact that the burst lasts longer when the transmission rate is reduced. A sample-path of the process is now a sequence of job sizes (i.e., measured in *volume*, in, say, bits — hence *not* time) and off-times (measured in *time*, to be interpreted as 'read-times'). In this example the job sizes have an exponential distribution with mean

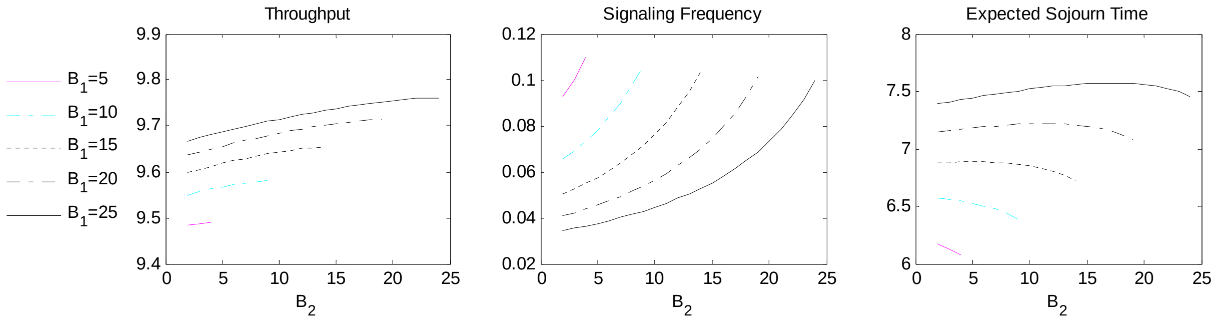

Figure 5: Effect of thresholds on elastic traffic.

$1/\mu = r_m/m_1 = r_p/p_1 = 25$; and the read-times have an exponential distribution with mean 1. The numerical outcome is presented in Fig. 5.

Consider again the situation of a fixed value of B_1, and compare the situations of (A) $B_2 < B_1$ and (B) $B_2 = B_1$. Under (A) regeneration cycles are longer than under (B), and hence the signaling frequency is lower. We have not found, however, a sound argumentation that reveals in which situation the throughput and packet delay are higher. Intuitively one would think that under (B) throughput is higher, which is confirmed by the graphs. The expected sojourn time $\mathbb{E}S$ turns out to have non-monotone behavior in this parameter setting; varying B_2 has clearly impact on the buffer content seen by an arriving job, but in a rather unpredictable way.

We mention that in this experiment the parameters are chosen such that the 'mean drift' while being in the '+'-phase is positive, which implies that the upper threshold B_1 will be reached in a relatively short time (roughly equal to $B_1 - B_2$ divided by this mean drift). The case of a negative 'mean drift' while being in the '+'-phase is less interesting, as it can then be argued that the process will be in the '+'-phase most of the time, and the queue roughly behaves as a non-feedback queue with generator Q^+ and traffic rates $\boldsymbol{r}^+$. In other words: in this case the value of B_1 has hardly any impact on the throughput.

Experiment IV: Effect of the transmission rate – elastic traffic. When r_p increases (with μ held constant, i.e., p_1 increases as well), regeneration cycles become shorter, and hence the signaling frequency increases. As could be intuitively expected also the throughput and expected sojourn time increase, but again we lack a solid argumentation; see Fig. 6.

4.2 Design issues

Above we saw that there is a trade off between the signaling frequency and the throughput, and it is the network provider's task to balance these, according to his (subjective) preference. We here sketch how such a decision is facilitated by our model. Figs. 7 and 8 depict the trade-off between the throughput ϑ and the time between two subsequent signals $\psi := 1/\varphi$, for a given B_1 by varying $B_2 \in [0, B_1]$; it provides us with a (decreasing) function $\vartheta = g(\psi)$ (see the left panels in Figs. 7 and 8). The provider having objective function $f(\vartheta, \psi)$, increasing in both ϑ and ψ, is

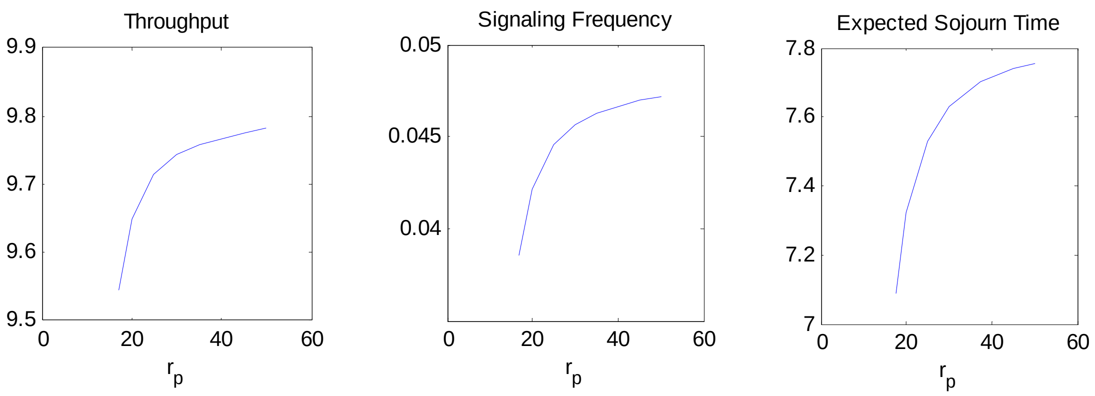

Figure 6: Effect of varying r_p and p_1 while keeping their ratio fixed at $r_p/\ p_1 = 25$.

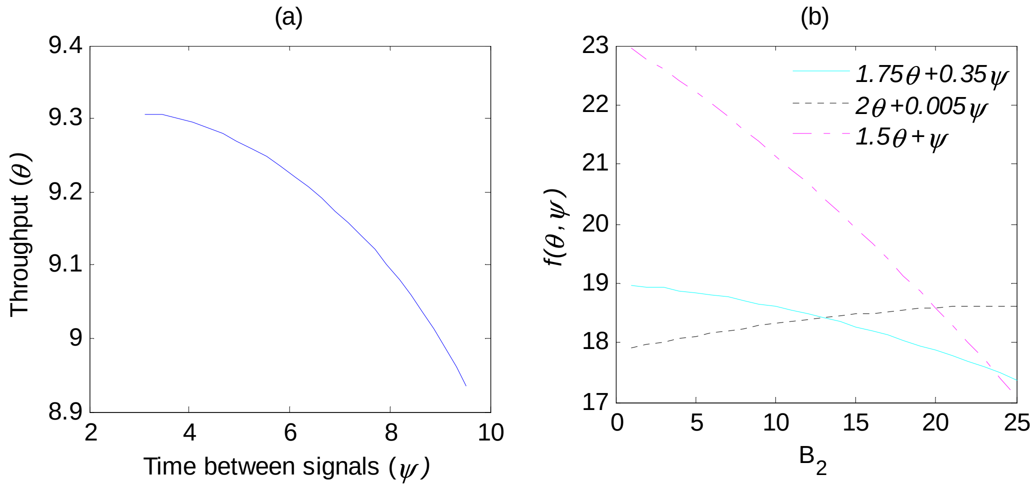

Figure 7: Trade-off between throughput ϑ and time between signals ψ (streaming traffic).

faced with the following optimization problem:

$$\max_{\vartheta,\psi} f(\vartheta, \psi) \quad \text{under} \quad \vartheta = g(\psi).$$

Having identified the optimally achievable pair $(\vartheta^\star, \psi^\star)$, we can now reconstruct what the corresponding value $B_2^\star$ was. Clearly, a similar procedure can be set up with both B_1 and B_2 being decision variables.

In Figs. 7-8 we graphically illustrate how to identify the optimum for the objective function $f(\vartheta, \psi) = \xi_1\vartheta + \xi_2\psi$. Fig. 7 uses the parameters of Experiment I, whereas Fig. 8 uses the parameters of Experiment III; B_1 is chosen equal to 25. The left panels show the trade-off between ϑ and ψ, whereas the right panels show the value of the objective function (for several choices of ξ_1 and ξ_2) as a function of B_2. The right panel of Figs. 7-8 shows that in some of these examples it turns out that the objective function is maximized by choosing B_2 as small as possible (which

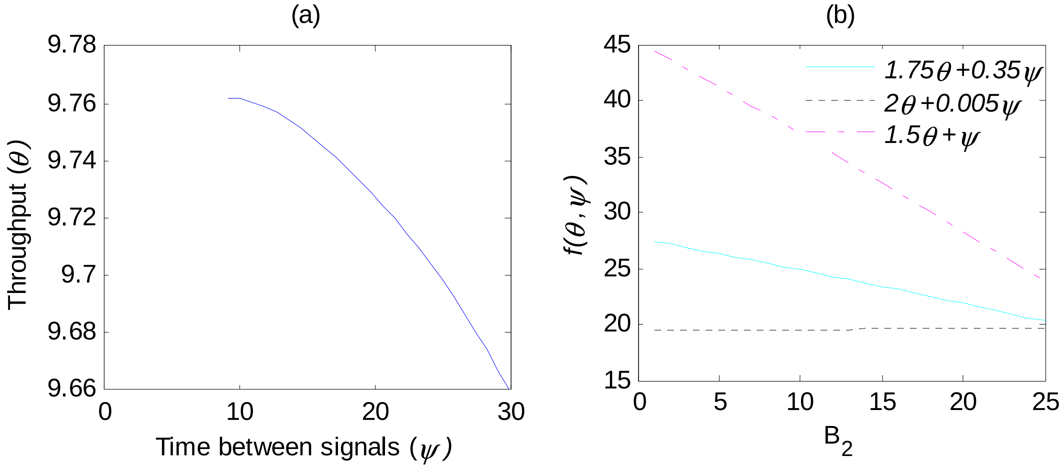

Figure 8: Trade-off between throughput ϑ and time between signals ψ (elastic traffic).

underscores the use of having two different thresholds rather than one). Evidently, this result is specific for the performance measures (ϑ and ψ) and the objective function chosen; other choices may lead to a structurally different outcome (see the curves in Figs. 7-8 in which B_2 should be chosen close to 25).

5 A model for higher aggregation levels

The experiments in the previous section involved a single source, but the main findings carry over to the situation with multiple sources. This can be validated in detail by redoing the numerical computations, but we here take an alternative approach. This approach is simpler, and somewhat less precise, but still capable of capturing the main trends.

Instead of having both a fluid content process (recorded by $W(t)$) and one or multiple sources (recorded by $X(t)$), we model the buffer content (resulting from the *ensemble* of all sources) by a birth-death-like process: during the '$+$'-phase the buffer content behaves as an M/M/1 queue with arrival rate λ^+ and departure rate μ^+, whereas during the '$-$'-phase it behaves as an M/M/1 queue with arrival rate λ^- (of traffic quanta of size, say, 1) and departure rate μ^-. Thus, the rate of change of the buffer is no longer determined by vectors $\boldsymbol{r}^-$ and $\boldsymbol{r}^+$ and generator matrices Q^- and Q^+ as before, but simply by birth-and-death parameters λ^+, μ^+, λ^-, μ^-. What remains the same as before, is that these depend on the current mode (that is, '+' or '$-$'), just as $\boldsymbol{r}$ and Q did before. To analyze a given situation, we can tune the λ^+, μ^+, λ^-, μ^- (satisfying the equilibrium condition $\lambda^- < \mu^-$), so that they roughly match the first and second order characteristics of the buffer dynamics. For this model we verify whether the trends, as observed in Section 4.1, still apply.

Let τ^+ be the duration of the '$+$'-phase, and τ^- the duration of the '$-$'-phase. It is immediate (for instance from Wald's theorem) that

$$\mathbb{E}\tau^- = \frac{B_1 - B_2}{\mu^- - \lambda^-}.$$

The computation of $\mathbb{E}\tau^+$ is standard, but a bit more tedious. With a_i denote the mean time until B_1 is reached, starting in $i \in \{0, \dots, B_1 - 1\}$, it is evident that

$$(\lambda^+ + \mu^+)a_i = \lambda^+ a_{i+1} + \mu^+ a_{i-1} + 1, \tag{5}$$

for $i = 1, \dots, B_1 - 1$; also $\lambda^+ a_0 = \lambda^+ a_1 + 1$ and $a_{B_1} = 0$. With $b_i = a_{i+1} - a_i$, Eqn. (5) can be rewritten as $\lambda^+ b_i = \mu^+ b_{i-1} - 1$, where $b_0 = -1/\lambda^+$. It is then easy to verify that

$$b_i = -\frac{(\mu^+)^i}{(\lambda^+)^{i+1}} - \left(1 - \left(\frac{\mu^+}{\lambda^+}\right)^i\right) \Bigg/ \left(1 - \frac{\mu^+}{\lambda^+}\right) = \frac{1}{\lambda^+ - \mu^+}\left(\left(\frac{1}{\varrho^+}\right)^{i+1} - 1\right),$$

with $\varrho^+ := \lambda^+/\mu^+$, and realizing that $-a_i = b_i + \cdots + b_{B_1-1}$ (use $a_{B_1} = 0$),

$$\mathbb{E}\tau^+ = a_{B_2} = -\sum_{j=B_2}^{B_1-1} b_j = \frac{B_1 - B_2}{\lambda^+ - \mu^+} - \frac{1}{\lambda^+}\frac{1}{1-\varrho^+}\frac{(\varrho^+)^{B_2 - B_1} - 1}{(\varrho^+)^{B_2} - (\varrho^+)^{B_2 - 1}};$$

if $\lambda^+ > \mu^+$, then this may be (roughly) approximated by $(B_1 - B_2)/(\lambda^+ - \mu^+)$ (as could be expected), whereas if $\lambda^+ < \mu^+$, then it roughly equals

$$\frac{1}{\mu^+}\frac{1}{(1-\varrho^+)^2}\left(\frac{1}{\varrho^+}\right)^{B_1}.$$

The signaling frequency equals $\varphi = 2/(\mathbb{E}\tau^+ + \mathbb{E}\tau^-)$, by virtue of 'renewal reward'. As is easily verified, the mean time per cycle spent in state 0 is

$$\mathbb{E}\tau_0^+ = \frac{(\varrho^+)^{B_1 - B_2} - 1}{(\varrho^+)^{B_1} - (\varrho^+)^{B_1 - 1}},$$

so that the throughput is given by

$$\vartheta = \frac{\mathbb{E}\tau^+ - \mathbb{E}\tau_0^+}{\mathbb{E}\tau^+ + \mathbb{E}\tau^-}\cdot\frac{1}{\mu^+} + \frac{\mathbb{E}\tau^-}{\mathbb{E}\tau^+ + \mathbb{E}\tau^-}\cdot\frac{1}{\mu^-}.$$

The thresholds B_1 and B_2 can be optimally selected by following a scheme similar to the one sketched in Section 4.2. In Fig. 9 we consider an example that again focuses on the trade-off between the throughput ϑ and the time between two consecutive signals ψ. We see the same type of behavior as in the single-source case. The input parameters are $\lambda^+ = 7$, $\mu^+ = 5$, $\lambda^- = 4$, $\mu^- = 5$, so that there is a positive drift during the '+'-phase. The thresholds are $B_1 = 25$ and $B_2 = 10$.

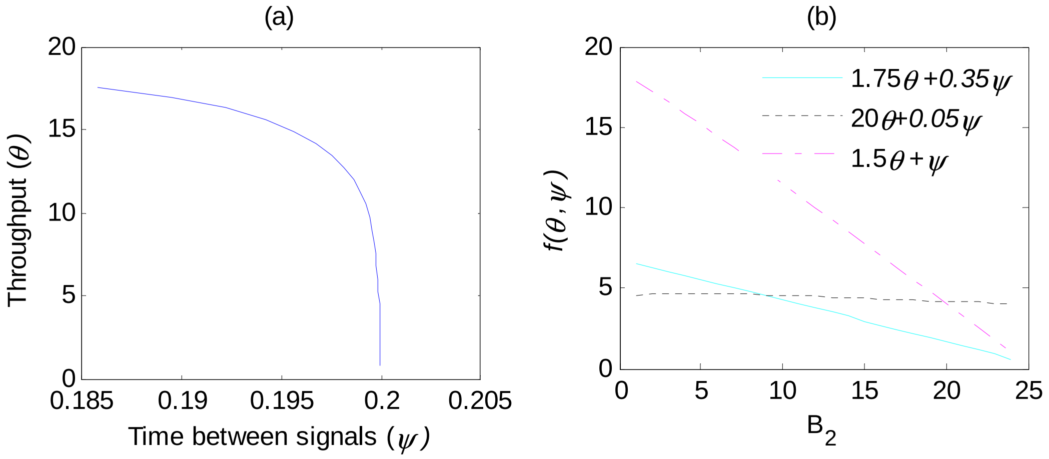

Figure 9: Trade-off between throughput ϑ and time between signals ψ for higher traffic aggregation model.

6 Concluding remarks

This paper addressed a methodology for resolving design issues in back-pressure-based control mechanisms. Relying on a feedback fluid model [9], we derived closed form expressions (in terms of the solution of certain eigensystems, and additionally a system of linear equations) for a number of key performance metrics. It enables us to investigate in detail the trade-offs involved – for instance the trade-off between throughput and the signaling overhead – and thus facilitates a proper selection of the protocol's design parameters (such as the values of the thresholds). It also sheds light on the effect of changing the transmission rates. We also presented a more stylized model, that is particularly useful when the input consists of a substantially larger aggregate of users.

References

[1] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61: 1871-1894, 1982.

[2] O. Boxma, H. Kaspi, O. Kella, and D. Perry. On/off storage systems with state dependent input, output and switching rates. *Probability in the Engineering and Informational Sciences*, 19: 1-12, 2005.

[3] S. Floyd. TCP and Explicit Congestion Notification. *ACM Computer Communication Review*, 24: 10-23, 1994.

[4] S. Floyd and V. Jacobson. Congestion gateways for packet neworks. *IEEE/ACM Transactions on Networking*, 1: 397-413, 1993.

[5] R. Gibbens and F. Kelly. Resource pricing and the evolution of congestion control. *Automatica*, 35: 1969-1985, 1999.

[6] IEEE STANDARD 802.3. Carrier Sense Multiple Access with Collision Detection (CSMA/CD) access method and physical layer specification, Annex 31B, 1998 Edition.

[7] L. KOSTEN. Stochastic theory of a data handling systems with groups of multiple sources. In: *Performance of Computer Communication Systems*, eds. H. Rudin and W. Bux, Elsevier, Amsterdam, the Netherlands, 321-331, 1984.

[8] R. MALHOTRA, R. VAN HAALEN, R. DE MAN, and M. VAN EVERDINGEN. Managing SLAs for metropolitan Ethernet networks. *Bell Labs Technical Journal,* 8: 83-95, 2002.

[9] R. MALHOTRA, M.R.H. MANDJES, W.R.W. SCHEINHARDT, and J.L. VAN DEN BERG. A feedback fluid queue with two congestion control thresholds. To appear in *Mathematical Methods in Operations Research.* `http://ftp.cwi.nl/CWIreports/PNA/PNA-E0803.pdf`

[10] R. MALHOTRA, R. VAN HAALEN, M. MANDJES, and R. NÚÑEZ-QUEIJA. Modeling the interaction of IEEE 802.3x hop-by-hop flow control and TCP End-to-end Flow Control, *Proc. Next Generation Internet Networks,* 260-267, 2005.

[11] M. MANDJES, D. MITRA, and W. SCHEINHARDT. Models of network access using feedback fluid queues. *Queueing Systems,* 44: 365-398, 2003.

[12] W. NOUREDDINE and F. TOBAGI. Selective back-pressure in switched Ethernet LANs. *Global Telecommunications Conference* 2, 1256-1263, 1999.

[13] K. RAMAKRISHNAN and R. JAIN. A binary feedback scheme for congestion avoidance in computer networks. *ACM Transactions on Computer Systems,* 8: 158-181, 1990.

[14] J. ROBERTS. Engineering for Quality of Service. In: *Self-Similar Network Traffic and Performance Evaluation,* Chapter 16, Wiley-Interscience, Chichester, UK, 401-420, 2000.

[15] W. SCHEINHARDT, N. VAN FOREEST, and M. MANDJES. Continuous feedback fluid queues. *Operations Research Letters* 33: 551-559, 2005.