Analysis of an $M/G/1$ queue with customer impatience and an adaptive arrival process

O.J. Boxma* and B.J. Prabhu†

October 15, 2009

Abstract

We study an $M/G/1$ queue with impatience and an adaptive arrival process. The rate of the arrival process changes according to whether an incoming customer is accepted or rejected. We analyse two different models for impatience : *(i)* based on workload, and *(ii)* based on queue length. For the workload-based model, we obtain the Laplace-Stieltjes Transform of the joint stationary workload and arrival rate process, and that of the waiting time. For the queue-length based model we obtain the analogous z-transform. These queueing models also capture the interaction between congestion control algorithms and queue management schemes in the Internet.

1 Introduction

Data traffic in the Internet is regulated by means of distributed algorithms in which each flow adapts its sending rate in order to match the bandwidth offered to it. The bandwidth offered to a flow on a link varies in time according to the number of concurrent flows traversing that link, and is signalled by the link to each flow by means of a binary feedback instructing the flow to either increase or decrease its sending rate. The link generates these feedback signals as a function of the occupancy of its input buffer : the higher the occupancy, the higher is the level of congestion, leading to a larger number of decrease signals. A frequently employed binary feedback signal is packet admission/rejection : admission signals an increase and rejection signals a decrease. An easy to implement example of a packet admission control policy is to reject an incoming packet if the buffer is full and to accept it otherwise. However, this policy leads

*EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands. Email : boxma@win.tue.nl

†LAAS-CNRS, Université de Toulouse, Avenue du Colonel Roche, 31077 Toulouse, France. Email : bjprabhu@laas.fr

to synchronization among concurrent flows and hence to inefficient utilization of the link bandwidth, *cf.* [7]. To overcome such effects, various alternative policies, including probabilistic admission control, have been proposed, *cf.* [7] and [14].

The aim of this paper is to model and analyze the above described interaction between a data source and a link that has a probabilistic packet admission control policy. The packet arrival process is modelled as a time inhomogeneous Poisson process whose intensity varies depending on whether an incoming packet is accepted or rejected. The packet sizes are assumed to be independent and identically distributed, and the link is modelled as an infinite buffer served at a fixed rate. An incoming packet is accepted to the queue with a probability that is a function of the current workload in the queue. We obtain the Laplace-Stieltjes Transform (LST) of the joint stationary workload and arrival intensity process, which then leads to the LST of the waiting time of accepted arrivals. In addition, we also study the model in which incoming packets are accepted depending upon the number of packets in the system. For this model we give the z-transform of the joint stationary distribution of the queue length and the arrival intensity process, which then leads to the z-transform of the number in the system seen by accepted arrivals.

1.1 Related work

Performance of congestion control algorithms has been widely studied analytically using stochastic models in a single-source setting, *cf.* [12], and deterministic models for a more general network setting, *cf.* [11]. Congestion control algorithms can be classified according to the increase function (i.e., how much to increase the sending rate for each positive feedback) and the decrease function (i.e., how much to decrease it for each negative feedback). In most of the studies, the focus is on obtaining the stationary distribution of the sending rate process for various increase/decrease functions and packet drop probability functions. However, the analytical results for the distribution are obtained by decoupling the interaction of the data source and the link buffer, i.e., by assuming that the probability of negative feedback depends only on the current sending rate. Hence, these studies focus on congestion control algorithms without explicitly incorporating queue-length based admission control.

On the other hand, there is a body of literature which models the effect of admission control policies on the queue length but with the restriction of a constant arrival rate, i.e., there is admission control but no congestion control. In [4], an $M/M/1/K$ queue with probabilistic admission control was studied assuming a constant rate for the Poisson input process. In [8], the authors studied an $M/M/1/K$ with an admission probability that depends on the exponentially averaged queue length instead of the current queue length. Such an admission control policy, called *Random Early Detection (RED)*, was proposed in [7]. They obtained the joint distribution of the instantaneous queue length and the

average queue length process as a solution of a system of differential equations, for which they gave an analytical expression for $K = 1$ and $K = 2$, and numerical results for larger buffer sizes. In [2], a singular perturbation technique was applied to obtain the joint distribution of the instantaneous and averaged queue lengths for general values of buffer sizes when the averaging parameter is close to zero.

Due to its complexity, the analytical study of the interaction between congestion control algorithms and admission control policies has been more or less restricted to studying the dynamics of the expected values using deterministic differential equations (with or without feedback delay), *cf.* [13] and [9]. The model considered in the present paper attempts to capture this interaction within a stochastic framework for a certain class of probabilistic admission control policies.

A related work in which this interaction has been studied is that of [16] in which the authors model the TCP source as a fluid source whose rate varies depending upon whether the finite buffer is full or not. The trajectory of the buffer content process is a continuous function of time whose dynamics are governed by a set of differential equations. Although there are similarities in the idea of modelling the interaction between the source and the buffer, there are several differences in the modelling approach that we take. As opposed to their fluid model, we model the TCP source as one which emits packets at distinct epochs thereby causing jumps in the buffer content process which no longer has a continuous sample path. Another important difference is in the feedback model itself. In [16] there is a positive feedback if and only if the buffer is not full. It is thus a model for a Drop-Tail policy, whereas we study a probabilistic feedback policy - the feedback is positive with a probability that decreases as the buffer level increases.

The present model is also strongly related to some of the existing models on impatience investigated in queueing theory, and in the following we describe this connection.

1.2 Connection with queueing theory

The connection between queueing theory and performance analysis of congestion control algorithms has been known since long, *cf.* [17]. The modelling of a link as a server with a finite buffer is a natural one which immediately brings out this connection, and this has been used to study admission control policies as mentioned above. From the queueing theory side, in [15] the authors considered various rejection rules, i.e., admission control policies in our context, for the $M/G/1$ queue. However, the arrival rate to the queue in their model does not change with the decision to accept or reject a packet. Another somewhat non-intuitive connection between the two occurs when analysing congestion control without admission control. It has been shown, for example, that the sending rate

process of a fairly general class of congestion control algorithms is equivalent to the workload process in a queue with state-dependent input and service rates, *cf.* [3] and [1].

The model under investigation in this paper can be seen as a generalization of the $MAP/G/1$ queue with impatient customers which was studied by Combé in [6]. In that model, the arrival process changes states at each packet arrival instant. However, the dynamics of the arrival process do not depend on whether a packet is accepted or rejected, which makes our model a generalization of the one studied in [6]. Our method of analysis is similar to the one in [6] in that we obtain a system of Volterra integral equations of the second kind for the joint stationary distribution of the workload and arrival rate process, from which we obtain a system of recursive equations for the LST of the joint stationary process.

A part of the present work containing the LST of the joint stationary workload and arrival rate process appeared in [5].

1.3 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we describe the system model and state the assumptions. In Section 3, we present the analysis of the model with workload-based impatience leading to the computation of the LST of the joint stationary workload and arrival rate process. Based upon this LST, we give the LST of the waiting time of the accepted arrivals. In Section 4, we analyse the queue-length-based model and obtain the z-transform of joint stationary queue length and arrival rate process. Based upon this z-transform, we give the z-transform of the number of customers seen in the system by accepted arrivals. Finally, we summarize the results and state possible extensions in Section 5.

2 Model description

Consider a variable data rate source which generates packets at Poisson intensity $\lambda(t) \in \mathcal{L}$, where $\mathcal{L}$ is a finite set of cardinality N. The packet sizes are assumed to be i.i.d. with distribution function $B_i(\cdot)$, mean μ_i^{-1}, and Laplace-Stieltjes transform $\mathsf{B_i}(\cdot)$, when the data source is in state i. These packets arrive at a queue, say a router in the Internet, which admits the packets based on the following admission control policy. An incoming packet which sees a workload level of x is admitted to the queue with probability $f(x)$, and rejected otherwise. We do not model the possibility of a rejected packet re-entering the queue at a later instant. We shall assume that f has the form

$$f(x) = \exp(-\nu x),$$

independently of the state of the input process. The function $f(x)$ can also be thought of as an impatience function associated with customers arriving to a server. If an incoming customer sees a higher waiting time, then it is less likely to join the queue.

For a router in the Internet the buffer occupancy in bits (the workload) is generally known to the router, and hence a workload-based impatience model is better suited for the analysis of such systems. However, in some cases the impatience probability may depend on the number of customers in the system rather than the total workload which may be unknown. For such a queue-length-based model, we shall assume that if an incoming packet sees n packets in the system, then it is admitted with probability p^n, $0 < p < 1$.

We shall assume that the variable data rate source is informed immediately whether a packet was admitted or rejected. In practice, there is a delay after which the source receives this information, and this delay could depend on the queue length itself. The source reacts to the admission control policy by adapting its data rate in the following way. With state i of the source we associate a Poisson intensity λ_i. The state of the source jumps from i to j with probability p_{ij} if a packet is rejected, and with probability p^*_{ij} if a packet is accepted. Thus, the intensity of the arrival process potentially changes with each arrival to the queue. In a protocol like TCP, the state of the source will jump to a state $j \leq i$ if a packet is rejected and to a state $j \geq i$ if a packet is accepted. However, we shall not assume any particular structure for the matrices $\mathbf{P} = [p_{ij}]$ and $\mathbf{P}^* = [p^*_{ij}]$.

3 Impatience based on workload

Let $V_i(t, x)$ denote the joint probability that at time t the workload is less than or equal to x and the input process is in state i. The server is assumed to work at unit rate. There are three possible events that can happen in a small interval $[t, t + \delta t)$: (i) there are no arrivals, in which case the workload is drained by an amount δt; (ii) an arrival occurs and is rejected, in which case the input process changes state; and (iii) an arrival occurs and is accepted, in which case the input process changes state and there is a jump in the workload process. The three terms on the RHS in the following equation correspond to the above three possible events.

$$
\begin{aligned}
V_i(t+\delta t, x) &= (1-\lambda_i \delta t) V_i(t, x+\delta t) \\
&\quad + \sum_j p_{ji} \lambda_j \delta t \cdot \int_{0^-}^{x} (1-\exp(-\nu y)) dV_j(t,y) \\
&\quad + \sum_j p^*_{ji} \lambda_j \delta t \int_{0^-}^{x} B_j(x-y) \exp(-\nu y) dV_j(t,y),\ x>0,\ 1 \le i \le N.
\end{aligned}
$$

From the above dynamics, we can derive the following integral equation for $i = 1, \ldots, N$:

$$
\begin{aligned}
\frac{\partial V_i(t,x)}{\partial t} &= \frac{\partial V_i(t,x)}{\partial x} - \lambda_i\ V_i(t,x) + \sum_j p_{ji} \lambda_j \int_{0^-}^{x} (1-\exp(-\nu y)) dV_j(t,y) \quad (1) \\
&\quad + \sum_j p^*_{ji} \lambda_j \int_{0^-}^{x} B_j(x-y) \exp(-\nu y) dV_j(t,y). \quad (2)
\end{aligned}
$$

Let us now discuss the issue whether the joint steady-state distribution of workload and arrival rate process exists.

Proposition 1 (Stability)**.** *If*

1. $\rho_{max} := \sup_i \lambda_i \mu_i^{-1}$ *is finite, and*
2. $\lim_{x\to\infty} f(x) = 0$,

then the joint workload and arrival rate process is stable.

Proof. If $\lim_{x\to\infty} f(x) = 0$, then $\exists x^* < \infty$ such that $\rho_{max} f(x) < 1,\ \forall x > x^*$. That is, if the workload in the queue is greater than x^* then the traffic intensity is less than unity, which implies that the workload process will cross the level x^* infinitely often. □

In the sequel we assume that the two conditions of the above proposition hold.

Let

$$
\Phi_i(s) = \int_{0^-}^{\infty} \exp(-sx) dV_i(x),
$$

denote the Laplace-Stieltjes Transform of the joint distribution function, and let $\overline{\Phi}(s) := [\Phi_i(s)]$ denote the row vector of the LST of the joint stationary distribution. Also, let $V_i(0)$ be the stationary joint probability that the workload is zero and the arrival intensity is λ_i, and $\overline{V}(0) := [V_i(0)]$ be the row vector of these joint probabilities. The following result relates $\overline{\Phi}(s)$ to $\overline{V}(0)$.

Lemma 1. *The LST of the joint distribution function* $\overline{\Phi}(s)$ *is given by the following infinite sum:*

$$\overline{\Phi}(s) = \overline{V}(0)\left[\sum_{i=0}^{\infty}\mathbf{D}(s+i\nu)\mathbf{A}^{-1}(s+i\nu)\left[\prod_{j=0}^{i-1}\mathbf{C}(s+j\nu)\mathbf{A}^{-1}(s+j\nu)\right]\right], \quad (3)$$

where the empty product is assumed to be unity, and

$$\mathbf{A}(s) = s\mathbf{I} - \mathbf{\Lambda}(\mathbf{I}-\mathbf{P}), \quad (4)$$
$$\mathbf{D}(s) = s\mathbf{I}, \quad (5)$$
$$\mathbf{C}(s) = \mathbf{\Lambda}(\mathbf{P} - \mathbf{B}(s)\mathbf{P}^*), \quad (6)$$

$\mathbf{\Lambda}$ *is a diagonal matrix with* λ_i *as its* i*th diagonal entry,* $\mathbf{B}(s)$ *is a diagonal matrix with* $\mathsf{B}_i(s)$ *as its* i*th diagonal entry.*

Proof. By taking the limit $t \to \infty$ in (2), we obtain the following integral equation for the joint steady-state distribution of the workload and the input process.

$$\frac{dV_i(x)}{dx} = \lambda_i V_i(x) - \sum_j p_{ji}\lambda_j \int_{0^-}^{x}(1-\exp(-\nu y))dV_j(y) - \sum_j p^*_{ji}\lambda_j \int_{0^-}^{x} B_j(x-y)\exp(-\nu y)dV_j(y),\ x>0,\ 1\le i\le N. \quad (7)$$

The above integral equations can also be derived by writing the balance equations.

Taking the LST of (7) gives

$$\Phi_i(s) - V_i(0) = \lambda_i\frac{\Phi_i(s)}{s} - \sum_j p_{ji}\lambda_j\left(\frac{\Phi_j(s)-\Phi_j(s+\nu)}{s}\right) - \sum_j p^*_{ji}\lambda_j\frac{\mathsf{B}_j(s)}{s}\Phi_j(s+\nu). \quad (8)$$

On rearranging (8), we obtain the following system of recursive equations:

$$\overline{\Phi}(s)\mathbf{A}(s) = \overline{V}(0)\mathbf{D}(s) + \overline{\Phi}(s+\nu)\mathbf{C}(s), \quad (9)$$

which upon iterating leads to (3). □

We next proceed to determine the constants $V_i(0), i = 1, 2, ..., N$, which will then completely characterize $\overline{\Phi}(s)$.

Let $\gamma_i,\ i = 1, 2, ..., N$, denote the ith eigenvalue of $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$, such that $\gamma_i \le \gamma_j$ for $i < j$, and $\underline{\boldsymbol{\alpha}}_i$ denote the corresponding right eigenvector. Since $\mathbf{P}$ is a stochastic matrix, we can explicitly obtain the first eigenvector, $\underline{\boldsymbol{\alpha}}_1$, to be equal to $[1\ 1\ldots 1]^T$ with eigenvalue $\gamma_1 = 0$. For the location of the other $N-1$ eigenvalues of $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$, we have the following result.

Lemma 2. *The eigenvalues of the matrix* $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$ *have positive real parts.*

Proof. Applying Geršgorin's circle theorem, *cf.* [10], every eigenvalue of $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$ lies in at least one of the disks

$$\{s : |s-\lambda_i(1-p_{ii})| \leq \sum_j |\lambda_i p_{ij}| = \lambda_i(1-p_{ii})\}.$$

Thus, for every i, the real part of γ_i is positive. □

Since $\mathbf{A}(s) = s\mathbf{I}-\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$, $\mathbf{A}(s)$ is singular at the eigenvalues of $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$, i.e., $\det(\mathbf{A}(s)) = 0$ at $s = \gamma_i, i = 1, 2, ..., N$. However, $\overline{\Phi}(s)$ is analytic in the half-plane $\mathrm{Re}(s) \geq 0$, and hence the constants $V_i(0), i = 1, 2, ..., N$, are such that the RHS of (3) is finite at $s = \gamma_i, i = 1, 2, ..., N$.

In order to compute $\overline{V}(0)$ we shall make use of the above fact and the following representation

$$\overline{\Phi}(s)\mathbf{A}(s) = \overline{V}(0)\left[\sum_{i=0}^{\infty}\mathbf{D}(s+i\nu)\left[\prod_{j=0}^{i-1}\mathbf{A}^{-1}(s+(j+1)\nu)\mathbf{C}(s+j\nu)\right]\right], \quad (10)$$

which, for simplicity, we rewrite as

$$\overline{\Phi}(s)\mathbf{A}(s) = \overline{V}(0)\mathbf{M}(s). \quad (11)$$

Before stating the main result, we first make an assumption under which the result holds.

Assumption 1. *For* $j \geq 1$, $\mathbf{A}(s+j\nu)$ *is invertible at* $s = \gamma_i$, *which is equivalent to the condition that* $\gamma_i \neq \gamma_k + j\nu$ *for* $i \neq k$ *and for every* j, *i.e., no two eigenvalues differ by an integer multiple of* ν.

The above assumption ensures that $\mathbf{A}(s+j\nu)$ is invertible in the right-half plane for $j \geq 1$. We shall later observe using numerical computations that when two eigenvalues differ by an integer multiple of ν, we can obtain the constants $V_i(0)$ by perturbing the entries of the matrix $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$.

Theorem 1. *The joint probability vector* $\overline{V}(0)$ *is the unique solution of the following set of* N *linear equations:*

$$\overline{V}(0) = \begin{bmatrix} (\mathbf{I}+\mathbf{M}(\nu)\mathbf{A}^{-1}(\nu)\mathbf{\Lambda}\boldsymbol{\mu}^{-1})\underline{\boldsymbol{\alpha}}_1 \\ \mathbf{M}(\gamma_2)\underline{\boldsymbol{\alpha}}_2 \\ \vdots \\ \mathbf{M}(\gamma_N)\underline{\boldsymbol{\alpha}}_N \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (12)$$

Proof. Due to the form of $\mathbf{A}(s)$, every right eigenvector of $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$, $\underline{\boldsymbol{\alpha}}_i$, is also a right eigenvector of $\mathbf{A}(s)$ with eigenvalue $(s-\gamma_i)$. For $i = 2, 3, ..., N$, we right multiply (11) by $\underline{\boldsymbol{\alpha}}_i$ and set $s = \gamma_i$ to get the following $N-1$ equations

$$0 = \overline{V}(0)\mathbf{M}(\gamma_i)\underline{\boldsymbol{\alpha}}_i,\ i = 2, 3, ..., N. \tag{13}$$

For the final equation, we first note that $\mathbf{M}(s)$ is singular at $s = 0$. To see this, rewrite $\mathbf{M}(s)$ as

$$\mathbf{M}(s) = \mathbf{D}(s) + \mathbf{M}(s+\nu)\mathbf{A}^{-1}(s+\nu)\mathbf{C}(s), \tag{14}$$

and right multiply by $\underline{\boldsymbol{\alpha}}_1$. Using (5) and (6), we see that

$$\mathbf{M}(s)\underline{\boldsymbol{\alpha}}_1 = s\underline{\boldsymbol{\alpha}}_1 + \mathbf{M}(s+\nu)\mathbf{A}^{-1}(s+\nu)\mathbf{\Lambda}(\mathbf{I}-\mathbf{B}(s))\underline{\boldsymbol{\alpha}}_1 \tag{15}$$

is equal to zero at $s = 0$, and that

$$\lim_{s\to 0}\frac{\mathbf{M}(s)\underline{\boldsymbol{\alpha}}_1}{s} = (\mathbf{I} + \mathbf{M}(\nu)\mathbf{A}^{-1}(\nu)\mathbf{\Lambda}\boldsymbol{\mu}^{-1})\underline{\boldsymbol{\alpha}}_1, \tag{16}$$

where $\boldsymbol{\mu}$ is a diagonal matrix with μ_i as its ith diagonal entry. We right multiply (11) by $\underline{\boldsymbol{\alpha}}_1$ and use the normalization equation $\overline{\mathrm{V}}(0)\underline{\boldsymbol{\alpha}}_1 = 1$ to obtain

$$1 = \overline{V}(0)(\mathbf{I} + \mathbf{M}(\nu)\mathbf{A}^{-1}(\nu)\mathbf{\Lambda}\boldsymbol{\mu}^{-1})\overline{\boldsymbol{\alpha}}_1. \tag{17}$$

Combining (17) and (13), we obtain the system of equations (12). □

3.1 An example with $N = 2$

To illustrate the computation of the probability vector, $\overline{V}(0)$, we consider the following example with $N = 2$. Let the packet sizes be exponentially distributed with mean μ^{-1}. The transition probability matrices are

$$\mathbf{P} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},\ \text{and}\ \mathbf{P}^* = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

That is, the source transmits at rate λ_1 as long as packets are rejected, and switches to λ_2 and continues to transmit at that rate as long as packets are accepted. For this example, the matrices $\mathbf{A}, \mathbf{C}$ and $\mathbf{D}$ are

$$\mathbf{A}(s) = \begin{bmatrix} s & 0 \\ \lambda_2 & (s-\lambda_2) \end{bmatrix}, \mathbf{C}(s) = \begin{bmatrix} \lambda_1 & -\lambda_1\frac{\mu}{s+\mu} \\ \lambda_2 & -\lambda_2\frac{\mu}{s+\mu} \end{bmatrix},\ \text{and}\ \mathbf{D}(s) = s\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The eigenvalues and the corresponding right eigenvectors of $\mathbf{\Lambda}(\mathbf{I}-\mathbf{P})$ are $\gamma_1 = 0$ with $\underline{\boldsymbol{\alpha}}_\mathbf{1} = [1\ 1]^T$ and $\gamma_2 = \lambda_2$ with $\underline{\boldsymbol{\alpha}}_2 = [0\ 1]^T$.

Let $\lambda_1 = 0.5$, $\mu = 1$ and $\nu = 1$. In Fig. 1, we plot $V_1(0)$ and $V_2(0)$ for various values of λ_2 which is also equal to γ_2. For our analysis, we had assumed that $\gamma_2 \neq k\nu$ (see Assumption 1). In the numerical computations as well, we cannot use (12) to compute $V_1(0)$ and $V_2(0)$ when $\lambda_2 = k$, and hence the discontinuities in the plot at integral values of λ_2. However, this numerical example shows that the values of $V_1(0)$ and $V_2(0)$ for $\lambda_2 = k\nu$ could be approximated closely by assuming $\lambda_2 = k\nu + \epsilon$.

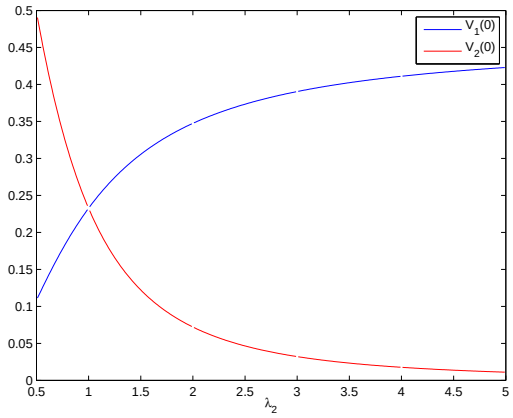

Figure 1: $V_1(0)$ and $V_2(0)$ as a function of λ_2. $\lambda_1 = 0.5$, $\nu = 1$ and $\mu = 1$.

3.2 Waiting times

Let W be the waiting time of the accepted customers in steady state. The following result relates the LST of W to Φ_i, which is the LST of the the joint steady-state probability that the workload is less than or equal to x and the input process is in state i.

Proposition 2.

$$E[\exp(-sW)] = \frac{\sum_i \lambda_i \Phi_i(s+\nu)}{\sum_i \lambda_i \Phi_i(\nu)}.$$

Proof. In steady state, $dV_i(x)$ represents the fraction of time the virtual workload is in the interval $[x, x+dx]$ and the background state is i. Since the arrival rate is λ_i when the background state is i, the fraction of accepted arrivals that see a workload in the interval $[x, x+dx]$ and the background state i is given by

$$\frac{\lambda_i \exp(-\nu x) dV_i(x)}{\sum_i \int_{0-}^{\infty} \lambda_i \exp(-\nu x) dV_i(x)},$$

and the fraction of accepted arrivals that see a workload of $[x, x+dx]$ is

$$\frac{\sum_i \lambda_i \exp(-\nu x) dV_i(x)}{\sum_i \int_{0-}^{\infty} \lambda_i \exp(-\nu x) dV_i(x)},$$

which is thus the probability that an accepted arrival sees a workload of $[x, x+$

dx]. Hence we can conclude that

$$\begin{aligned} E[\exp(-sW)] &= \int_{x=0-}^{\infty} \exp(-sx) \frac{\sum_i \lambda_i \exp(-\nu x) dV_i(x)}{\sum_i \int_{0-}^{\infty} \lambda_i \exp(-\nu x) dV_i(x)} \\ &= \frac{\sum_i \lambda_i \Phi_i(s+\nu)}{\sum_i \lambda_i \Phi_i(\nu)}. \end{aligned}$$

□

Remark 1. *Combé [6] derives a similar result for the special case that the dynamics of the input process do not depend on whether a packet is accepted or rejected.*

4 Impatience based on number of customers

In this section we shall consider a discrete state-space model to study the joint behaviour of a variable data rate source and the queue length at the buffer. As in the previous section, we assume that the variable rate source generates packets according to a Poisson process of rate λ_i when it is in state i, $i = 1, \ldots, N$. The packets arrive at a single server queue which admits the packets based on the following admission control policy. If an incoming packet sees n packets in the system, then it is admitted to the queue with probability p^n, with $0 < p < 1$. The background state now changes to j w.p. $p^*_{i,j}$. It is rejected with probability $1 - p^n$, and then the background state changes to j w.p. $p_{i,j}$. Unlike in the previous section however, we shall restrict ourselves to the case when packet sizes are i.i.d. and exponentially distributed with rate μ_i when the background state is i (i.e., the service speed may depend on the background state).

As in Proposition 1, we may conclude that the system is stable when $p < 1$. Let $\{q_{n,i}\}$, $n = 0, 1, \ldots,$ and $i = 1, \ldots, N$, be the steady-state probability that the system contains n customers while the background state is i. In the following, we shall obtain the z-transform of $\{q_{n,i}\}$ defined as

$$Q_i(z) := \sum_{n=0}^{\infty} z^n q_{n,i},\ i = 1, \ldots, N, \tag{18}$$

$$\text{and } \overline{Q}(z) := [Q_1(z)\ Q_2(z)\ \ldots Q_N(z)]. \tag{19}$$

The steps to obtain $\overline{Q}(z)$ follow closely the steps for obtaining $\overline{\Phi}(s)$ in Section 3. Let $\boldsymbol{\mu}$ (resp. $\boldsymbol{\Lambda}$) be a $N \times N$ diagonal matrix with μ_i (resp. λ_i) as its ith diagonal element, and let $\overline{q}_0 := \overline{Q}(0)$. Then,

Lemma 3. *The joint transform vector*

$$\overline{Q}(z) = \overline{q}_0 \left[\sum_{i=0}^{\infty} \mathbf{D}(p^i z) \mathbf{A}^{-1}(p^i z) \left[\prod_{j=0}^{i-1} \mathbf{C}(p^j z) \mathbf{A}^{-1}(p^j z) \right] \right], \tag{20}$$

where the empty product is assumed to be unity, and

$$\begin{aligned}
\mathbf{A}(z) &:= (z-1)\boldsymbol{\mu} + z\boldsymbol{\Lambda}(\mathbf{I}-\mathbf{P}), && (21)\\
\mathbf{C}(z) &:= -z\boldsymbol{\Lambda}\mathbf{P} + z^2\boldsymbol{\Lambda}\mathbf{P}^*, && (22)\\
\mathbf{D}(z) &:= (z-1)\boldsymbol{\mu}. && (23)
\end{aligned}$$

Proof. The joint process of the number of customers and background state forms a two-dimensional Markov process. Its balance equations are: For $n = 1, 2, \ldots$, $i = 1, \ldots, N$,

$$\begin{aligned}
(\lambda_i - \lambda_i(1-p^n)p_{i,i} + \mu_i)q_{n,i} = &\sum_j \lambda_j p^{n-1} p^*_{j,i} q_{n-1,j}\\
&+ \sum_{j\neq i} \lambda_j (1-p^n) p_{j,i} q_{n,j} + \mu_i q_{n+1,i}, \qquad (24)
\end{aligned}$$

and for $n = 0$ and $i = 1, \ldots, N$,

$$\lambda_i q_{0,i} = \mu_i q_{1,i}. \qquad (25)$$

This yields, for $i = 1, \ldots, N$:

$$\begin{aligned}
&\mu_i[Q_i(z) - q_{0,i}] + \lambda_i(1-p_{i,i})Q_i(z) + \lambda_i p_{i,i} Q_i(pz)\\
= \;& \frac{\mu_i}{z}[Q_i(z) - q_{0,i}] + z\sum_j \lambda_j p^*_{j,i} Q_j(pz) + \sum_{j\neq i} \lambda_j p_{j,i}[Q_j(z) - Q_j(pz)], \qquad (26)
\end{aligned}$$

and hence

$$\begin{aligned}
&[\mu_i(z-1) + \lambda_i z(1-p_{i,i})]Q_i(z) - z\sum_{j\neq i} \lambda_j p_{j,i} Q_j(z)\\
= \;& \mu_i(z-1)q_{0,i} + z^2 \sum_j \lambda_j p^*_{j,i} Q_j(pz) - z\sum_j \lambda_j p_{j,i} Q_j(pz). \qquad (27)
\end{aligned}$$

This formula can be written in the following matrix-form:

$$\overline{Q}(z)\mathbf{A}(z) = \overline{Q}(pz)\mathbf{C}(z) + \overline{q}_0\mathbf{D}(z). \qquad (28)$$

The solution to the above system of equations can be expressed in terms of the infinite sum (20). □

In order to determine $\overline{q}_0$ we shall couple the fact that $\overline{Q}(z)$ is analytic in the unit disk $\{z : |z| \leq 1\}$, i.e. it has no poles in the unit disk, with the fact that $\det(\mathbf{A}^{-1}(z))$ has N poles in the unit disk, and deduce that $\overline{q}_0$ is such that the RHS of (20) should remain analytic at these N singularities.

We now show that $\det(\mathbf{A}^{-1}(z))$ has N poles in the unit disk. Let ζ_i, $i = 1, 2, \ldots, N$, be the N zeros of $\det(\mathbf{A}(z))$.

Lemma 4. *The N zeros of the polynomial det$(\mathbf{A}(z))$ lie in the disk $\{z : |z| \leq 1\}$.*

Proof. Denote $\mathbf{U}(z) := (\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}\mathbf{A}(z)$. We shall deduce the location of the ζ_i-s from the location of the zeros of $\det(\mathbf{U}(z))$. From the definition of $\mathbf{U}(z)$ and (21),

$$\begin{aligned}
\mathbf{U}(z) &= (\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}((z-1)\boldsymbol{\mu} + z\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P})) \\
&= (\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}((z(\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P})) - \mu) \\
&= z\mathbf{I} - (\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}\boldsymbol{\mu} \\
&= z\mathbf{I} - (\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}(\boldsymbol{\mu}^{-1})^{-1} \\
&= z\mathbf{I} - (\boldsymbol{\mu}^{-1}(\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1} \\
&= z\mathbf{I} - (\mathbf{I} + \boldsymbol{\mu}^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}.
\end{aligned}$$

From the above equation we can infer that the zeros of $\det(\mathbf{U}(z))$ are the same as the eigenvalues of $(\mathbf{I} + \boldsymbol{\mu}^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}$.

Also from the definition of $\mathbf{U}(z)$, we have the relation

$$\det(\mathbf{U}(z)) = \det((\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1})\det(\mathbf{A}(z))$$

which, assuming $\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P})$ is invertible, says that the zeros of $\det(\mathbf{A}(z))$ are the same as the zeros of $\det(\mathbf{U}(z))$.

From the two preceding arguments we can conclude that the set of eigenvalues of $(\mathbf{I} + \boldsymbol{\mu}^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))^{-1}$ is the same as the set of zeros of $\det(\mathbf{A}(z))$.

We now show that the eigenvalues of $\mathbf{I} + \boldsymbol{\mu}^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P})$ lie on or outside the unit circle, which would then prove the lemma. For this we shall apply Geršgorin's circle theorem to $\mathbf{I} + \boldsymbol{\mu}^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P})$ and conclude that its eigenvalues lie in the set

$$\begin{aligned}
&\cup_{i=1}^{N}\{z : |z - (1 + \mu_i^{-1}\lambda_i(1 - p_{ii}))| \leq \sum_{j=1}^{N} |\mu_i^{-1}\lambda_i p_{ij}|\} \\
&= \cup_{i=1}^{N}\{z : |z - (1 + \mu_i^{-1}\lambda_i(1 - p_{ii}))| \leq \mu_i^{-1}\lambda_i(1 - p_{ii})\},
\end{aligned}$$

and hence lie on or outside the unit circle $|z| = 1$. □

Let $\underline{\alpha}_i$, $i = 1, 2, ..., N$ denote the right eigenvectors of $\mathbf{I} + \boldsymbol{\mu}^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P})$ corresponding to the eigenvalue γ_i, and let

$$\mathbf{M}(z) = \sum_{i=0}^{\infty}\left[\prod_{j=0}^{i-1}\mathbf{C}(p^j z)\mathbf{A}^{-1}(p^{j+1}z)\right]\mathbf{D}(p^i z). \tag{29}$$

In order to determine $\overline{q}_0$, we make the following assumption on the eigenvalues γ_i.

Assumption 2. *For any pair i_1 and i_2 such that $i_1 \neq i_2$,*

$$\gamma_{i_1} \neq p^j \gamma_{i_2}.$$

We now have the following result.

Theorem 2. *The probabilities $\overline{q}_0$ are the unique solution to*

$$\overline{q}_0 = \begin{bmatrix} \mathbf{M}(1)\mathbf{A}^{-1}(1)\underline{\boldsymbol{\alpha}}_1 \\ \mathbf{M}(\gamma_2)\underline{\boldsymbol{\alpha}}_2 \\ \vdots \\ \mathbf{M}(\gamma_N)\underline{\boldsymbol{\alpha}}_N \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{30}$$

Proof. Let us order the eigenvalues such that $|\gamma_i| \leq |\gamma_j|$ for $i \leq j$. The first eigenvalue $\gamma_1 = 1$ with eigenvector $\underline{\alpha}_1 = [1\,1 \ldots 1]^T$.

We can rewrite (20) as

$$\overline{Q}(z)\mathbf{A}(z) = \overline{q}_0\mathbf{M}(z) \tag{31}$$

Right multiplying the LHS of (31) by $\overline{\alpha}_i$, we get

$$\begin{aligned} \overline{Q}(z)\mathbf{A}(z)\underline{\alpha}_i &= \overline{Q}(z)(\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))\mathbf{U}(z)\underline{\alpha}_i \\ &= \overline{Q}(z)(\boldsymbol{\mu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}))(z - \gamma_i^{-1})\underline{\alpha}_i. \end{aligned} \tag{32}$$

On right multiplying (31) and substituting $z = \gamma_i^{-1}$, we get

$$0 = \overline{q}_0\mathbf{M}(\gamma_i^{-1})\underline{\alpha}_i, \tag{33}$$

which would give us $N-1$ equations corresponding to $i = 2, \ldots, N$, whereas for $i = 1$, the RHS of the above equation is 0 as well. To see this is true, we first rewrite $\mathbf{M}(z)$ as

$$\mathbf{M}(z) = \mathbf{D}(z) + \mathbf{M}(pz)\mathbf{A}^{-1}(pz)\mathbf{C}(z). \tag{34}$$

On right multiplying the above equation by $\underline{\alpha}_1$, we get

$$\begin{aligned} \mathbf{M}(z)\overline{\alpha}_1 &= (z-1)\boldsymbol{\mu}\underline{\alpha}_1 + (-z + z^2)\mathbf{M}(pz)\mathbf{A}^{-1}(pz)\boldsymbol{\Lambda}\overline{\alpha}_1 \\ &= (z-1)(\boldsymbol{\mu} + z\mathbf{M}(pz)\mathbf{A}^{-1}(pz)\boldsymbol{\Lambda})\underline{\alpha}_1. \end{aligned}$$

The final equation is obtained from the normalization equation, $\overline{Q}(1)\underline{\alpha}_1 = 1$, which concludes the proof. □

4.1 Number in the system as seen by accepted arrivals

Let L be the number of customers in the system as seen by an accepted arrival in steady state. We can derive the z-transform of L by using arguments similar to those used for deriving the LST of the waiting time in the previous section.

Proposition 3.

$$E[z^L] = \frac{\sum_i \lambda_i Q_i(zp)}{\sum_i \lambda_i Q_i(p)}.$$

Proof. Following arguments similar to those used in Proposition 2, we can deduce that the fraction of accepted arrivals that see n customers in the system is

$$\frac{\sum_i \lambda_i p^n q(n,i)}{\sum_i \sum_n \lambda_i p^n q(n,i)},$$

which is thus the probability that an accepted arrival sees n customers in the system. Hence we can conclude that

$$\begin{aligned} E[z^L] &= \sum_n z^n \frac{\sum_i \lambda_i p^n q_{n,i}}{\sum_i \sum_n \lambda_i p^n q_{n,i}} \\ &= \frac{\sum_i \lambda_i Q_i(zp)}{\sum_i \lambda_i Q_i(p)}. \end{aligned}$$

□

5 Conclusions and future work

In this paper, we have obtained the Laplace-Stieltjes Transform of the joint workload and arrival rate process of an $M/G/1$ queue with customer impatience and an adaptive arrival process. This queueing model was motivated by applications in the Internet, and models the interaction between congestion control and admission control algorithms. We have also analysed a variant in which customer impatience is based on the queue-length, and have obtained the z-transform of the joint queue-length and arrival rate process. We have enhanced the model by allowing service times to depend on the state of the input process.

Possible extensions could consider (i) infinite support for the arrival rate process, and (ii) packets that re-enter the system upon rejection. These extensions could be studied with either workload based rejection or packet based rejection rules.

6 Acknowledgments

The research of Onno Boxma was performed in the framework of the BRICKS program and the European Network of Excellence EURO-NF. The research of Balakrishna Prabhu was partly carried out while he was a post-doctoral researcher with the Centrum voor Wiskunde and Informatica (CWI), Amsterdam; the Eindhoven University of Technology (TU/e), Eindhoven; and EURANDOM,

Eindhoven. The authors would like to thank Offer Kella (Hebrew University of Jerusalem), Rudesindo Núñez-Queija (CWI and University of Amsterdam), and David Perry (University of Haifa) for fruitful discussions on this subject.

References

[1] E. Altman, K.E. Avrachenkov, A.A. Kherani, and B.J. Prabhu. Performance analysis and stochastic stability of congestion control protocols. In *Proceeding of INFOCOM*, 2005.

[2] E. Altman, K.E. Avrachenkov, and B.J. Prabhu. A singular perturbation approach to analysing a RED queue. In *Proceedings of HET-NETs*, 2004.

[3] R. Bekker, S.C. Borst, O.J. Boxma, and O. Kella. Queues with workload-dependent arrival and service rates. *Queueing Syst. Theory Appl.*, 46(3/4):537–556, 2004.

[4] T. Bonald, M. May, and J. Bolot. Analytic evaluation of RED performance. In *Proceedings of the IEEE INFOCOM*, 2000.

[5] O.J. Boxma, O. Kella, D. Perry, and B.J. Prabhu. Analysis of an M/G/1 queue with impatience and an adaptive arrival process. In *Proceedings of IWAP*, 2008.

[6] M. Combé. Impatient customers in the MAP/G/1 queue. Research Report BS-R9413, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1994.

[7] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, August 1993.

[8] E. Kuumola, J.A.C. Resing, and J. Virtamo. Joint distribution of instantaneous and averaged queue length in an M/M/1/K system. In P. Tran-Gia and J. Roberts, editors, *Proceedings of the 15th ITC Specialist Seminar "Internet Traffic Engineering and Traffic Management"*, pages 58–67, July 2002.

[9] P. Kuusela, P. Lassila, J. Virtamo, and P. Key. Modeling RED with idealised TCP sources. In *Proceedings of IFIP ATM and IP*, 2001.

[10] P. Lancaster and M. Tismenetsky. *The Theory of Matrices.* Academic Press, New York, 1985.

[11] S. Low. A duality model of TCP and queue management algorithms. *IEEE/ACM Transactions on Networking*, 11(4):525–536, 2003.

[12] K. Maulik and B. Zwart. An extension of the square root law of TCP. *Annals of Operations Research*, 170(1):217–232, 2009.

[13] V. Mishra, W. Gong, and D. Towsley. Fluid based analysis of a network of AQM routers supporting TCP Flows with an application to RED. In *Proceedings of the ACM SIGCOMM*, 2000.

[14] R. Pan, B. Prabhakar, and K. Psounis. CHOKe: A stateless active queue management scheme for approximating fair bandwidth allocation. In *Proceedings of the IEEE INFOCOM*, 2000.

[15] D. Perry and S. Asmussen. Rejection rules in the M/G/1 queue. *Queueing Systems - Theory and Applications*, 29:105–130, 1995.

[16] N. van Foreest, M. Mandjes, and W.R.W. Scheinhardt. Analysis of a feedback fluid model for heterogeneous TCP sources. *Stochastic Models*, 19(3):299–324, 2003.

[17] Y.T. Wang and B. Sengupta. Performance analysis of a feedback congestion control policy under non-negligible propagation delay. In *Proceedings of ACM SIGCOMM*, pages 149–157, 1991.