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

O.J. Boxma[1], O. Kella[2], D. Perry[3], and B.J. Prabhu[1,4]

[1] EURANDOM and Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. (e-mail: boxma@win.tue.nl)

[2] Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel. (e-mail: offer.kella@huji.ac.il)

[3] Department of Statistics, University of Haifa, Haifa 31905, Israel. (e-mail: dperry@stat.haifa.ac.il)

[4] CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. (e-mail: b.j.prabhu@cwi.nl)

Abstract. In this paper 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 impatient or not. We obtain the Laplace-Stieltjes Transform of the joint stationary workload and arrival rate process. This queueing model also captures the interaction between congestion control algorithms and queue management schemes in the Internet.
Keywords. $M/G/1$ queue, impatience, adaptive arrival process, workload.

1 Introduction

In this paper, we study an $M/G/1$ queue with customer impatience and an adaptive arrival process. This model is motivated by data traffic control in the Internet in which each data source adapts its sending rate according to the feedback it receives from the network. This feedback is generated by each link in the network as a function of its input buffer occupancy: the higher the occupancy, the higher is the level of congestion, leading to a larger number of decrease signals, and *vice versa*. An example of a binary feedback signal is packet admission/rejection.

We model and analyze the above described interaction between an adaptive data source and a buffer with a probabilistic admission control policy. The model under investigation in this paper can be seen as a generalization of the $M/G/1$ queue with impatient customers (cf. Perry & Asmussen (1995)) and the $MAP/G/1$ queue with impatient customers (cf. Combé (1994)). In the latter model, the arrival process changes states at each customer (packet, in our model) arrival instant. However, the dynamics of the arrival process

do not depend on whether a customer is impatient or not, which makes our model a generalization of the one studied in Combé (1994).

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 leading to the computation of the Laplace-Stieltjes Transform (LST) of the joint stationary workload and arrival rate process. Finally, we state possible extensions in Section 4.

2 Model description

Consider an adaptive data 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 LST $\tilde{\mathsf{B}}_i(\cdot)$, when the background process is in state i. These packets arrive at a queue, say a router in the Internet, which admits the packets based on the following 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

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

independent 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 more likely not to join the queue.

We shall assume that the data source is informed immediately whether a packet was admitted or rejected. The source then reacts to the feedback by adapting its data rate in the following way. With state i of the source we associate 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, and the change depends on whether that arrival was accepted or not. In an Internet-based 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 Analysis

Let $V_i(t, x)$ denote the distribution function of the workload process at time t when 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.

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

Dividing by δt and letting $\delta t \to 0$ yields a set of integro-differential equations in $V_i(t,x)$. By taking the limit $t\to\infty$, we obtain the following set of integro-differential equations for the joint steady-state distribution $V_i(x)$ of the workload and the input process.

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

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

Taking the LST of (1), with $\tilde{\mathsf{V}}_i(s) := \int_{0^-}^{\infty} e^{-sx}dV_i(x)$, gives

$$\tilde{\mathsf{V}}_i(s) - V_i(0) = \lambda_i\frac{\tilde{\mathsf{V}}_i(s)}{s} - \sum_j p_{ji}\lambda_j\left(\frac{\tilde{\mathsf{V}}_j(s)-\tilde{\mathsf{V}}_j(s+\nu)}{s}\right) - \sum_j p^*_{ji}\lambda_j\frac{\tilde{\mathsf{B}}_j(s)}{s}\tilde{\mathsf{V}}_j(s+\nu), \quad (2)$$

where $V_i(0)$ is the stationary joint probability that the workload is zero and the input process is in state i.

Let $\underline{\tilde{\mathsf{V}}}(s) := [\tilde{\mathsf{V}}_i(s)]$ denote the column vector of the LST of the joint stationary distribution. On rearranging (2), we obtain the following system of recursive equations,

$$\mathbf{A}(s)\underline{\tilde{\mathsf{V}}}(s) = \mathbf{D}(s)\underline{\mathbf{V}}(0) + \mathbf{C}(s)\underline{\tilde{\mathsf{V}}}(s+\nu), \quad (3)$$

where

$$\mathbf{A}(s) = s\mathbf{I} - (\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}, \tag{4}$$

$$\mathbf{D}(s) = s\mathbf{I}, \tag{5}$$

$$\mathbf{C}(s) = (\mathbf{P}^T - \mathbf{P}^{*T}\mathbf{B}(s))\mathbf{\Lambda}, \tag{6}$$

$\mathbf{\Lambda}$ is a diagonal matrix with λ_i as its ith diagonal entry, $\mathbf{B}(s)$ is a diagonal matrix with $\tilde{\mathsf{B}}_i(s)$ as its ith diagonal entry, and $\underline{\mathbf{V}}(0)$ is the column vector with $V_i(0)$ as its ith row element.

We can express $\underline{\tilde{\mathbf{V}}}(s)$ in $\underline{\mathbf{V}}(0)$ by means of the following infinite sum

$$\underline{\tilde{\mathbf{V}}}(s) = \left[\sum_{i=0}^{\infty}\left[\prod_{j=0}^{i-1}\mathbf{A}^{-1}(s+j\nu)\mathbf{C}(s+j\nu)\right]\mathbf{A}^{-1}(s+i\nu)\mathbf{D}(s+i\nu)\right]\underline{\mathbf{V}}(0), \tag{7}$$

where the empty product is assumed to be unity. We next proceed to determine the constants $V_i(0), i = 1, 2, ..., N$, which will then completely characterize $\underline{\tilde{\mathbf{V}}}(s)$.

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

Theorem 1. *The eigenvalues of matrix* $(\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}$ *have positive real parts.*

Proof. Applying Geršgorin's circle theorem, cf. Lancaster and Tismenetsky (1985), every eigenvalue of $(\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}$ lies in at least one of the disks

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

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

Since $\mathbf{A}(s) = s\mathbf{I} - (\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}$, $\mathbf{A}(s)$ is singular at the eigenvalues of $(\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}$, i.e., $\det(\mathbf{A}(s)) = 0$ at $s = \gamma_i, i = 1, 2, ..., N$. However, $\underline{\tilde{\mathbf{V}}}(s)$ is analytic in the plane $\mathrm{Re}(s) \geq 0$, and hence the constants $V_i(0), i = 1, 2, ..., N$, would be such that the RHS of (7) is finite at $s = \gamma_i, i = 1, 2, ..., N$.

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

$$\mathbf{A}(s)\underline{\tilde{\mathbf{V}}}(s) = \left[\sum_{i=0}^{\infty}\left[\prod_{j=0}^{i-1}\mathbf{C}(s+j\nu)\mathbf{A}^{-1}(s+(j+1)\nu)\right]\mathbf{D}(s+i\nu)\right]\underline{\mathbf{V}}(0), \tag{8}$$

which, for simplicity, we rewrite as

$$\mathbf{A}(s)\underline{\tilde{\mathbf{V}}}(s) = \mathbf{M}(s)\underline{\mathbf{V}}(0). \tag{9}$$

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 $(\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}$.

Due to the form of $\mathbf{A}(s)$, every left eigenvector $\overline{\boldsymbol{\alpha}}_i$ of $(\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}$ is also a left eigenvector of $\mathbf{A}(s)$ with eigenvalue $(s - \gamma_i)$. For $i = 2, 3, ..., N$, we left multiply (9) by $\overline{\boldsymbol{\alpha}}_i$ and set $s = \gamma_i$ to get the following $N - 1$ equations

$$0 = \overline{\boldsymbol{\alpha}}_i \mathbf{M}(\gamma_i)\underline{\mathbf{V}}(0), \ i = 2, 3, ..., N. \tag{10}$$

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

$$\mathbf{M}(s) = \mathbf{D}(s) + \mathbf{C}(s)\mathbf{A}^{-1}(s + \nu)\mathbf{M}(s + \nu), \tag{11}$$

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

$$\overline{\boldsymbol{\alpha}}_1 \mathbf{M}(s) = s\overline{\boldsymbol{\alpha}}_1 + \overline{\boldsymbol{\alpha}}_1(\mathbf{I} - \mathbf{B}(s))\mathbf{\Lambda}\mathbf{A}^{-1}(s + \nu)\mathbf{M}(s + \nu) \tag{12}$$

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

$$\lim_{s \to 0} \frac{\overline{\boldsymbol{\alpha}}_1 \mathbf{M}(s)}{s} = \overline{\boldsymbol{\alpha}}_1(\mathbf{I} + \boldsymbol{\mu}^{-1}\mathbf{\Lambda}\mathbf{A}^{-1}(\nu)\mathbf{M}(\nu)), \tag{13}$$

where $\boldsymbol{\mu}$ is a diagonal matrix with μ_i as its ith diagonal entry. We left multiply (9) by $\overline{\boldsymbol{\alpha}}_1$ and use the normalization equation $\overline{\boldsymbol{\alpha}}_1\underline{\tilde{\mathbf{V}}}(0) = 1$ to obtain

$$1 = \overline{\boldsymbol{\alpha}}_1(\mathbf{I} + \boldsymbol{\mu}^{-1}\mathbf{\Lambda}\mathbf{A}^{-1}(\nu)\mathbf{M}(\nu))\underline{\mathbf{V}}(0). \tag{14}$$

Finally, we obtain $\underline{\mathbf{V}}(0)$ as the solution of a system of N linear equations (10) and (14).

3.1 An example with $N = 2$

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

The eigenvalues and the corresponding left eigenvectors of $(\mathbf{I} - \mathbf{P}^T)\mathbf{\Lambda}$ are $\gamma_1 = 0$ with $\overline{\boldsymbol{\alpha}}_1 = [1\ 1]$ and $\gamma_2 = \lambda_2$ with $\overline{\boldsymbol{\alpha}}_2 = [0\ 1]$.

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 (10) and (14) to compute $V_1(0)$ and $V_2(0)$, 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 $\lambda_2 = k\nu + \epsilon$.

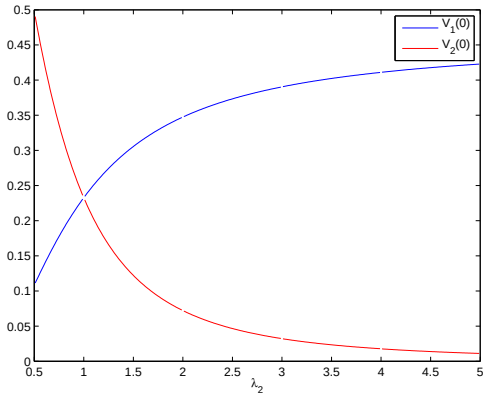

Fig. 1. $V_1(0)$ and $V_2(0)$ as a function of λ_2. $\lambda_1 = 0.5$, $\nu = 1$ and $\mu = 1$.

4 Future work

A variation of this model would be to study packet-based rejection rules instead of workload-based rules. This would be more appropriate for customer-based queues in which an incoming customer can observe the queue length but not the workload. A further extension of this work will involve considering an arrival rate process with infinite support. For the case of exponential service times, we also intend to present a recursive computation of the joint steady-state distribution.

References

Combé, M. (1994) : Impatient customers in the MAP/G/I queue. *Research Report BS-R9413*. CWI, Amsterdam.

Lancaster, P. and Tismenetsky, M. (1985): The Theory of Matrices. *Academic Press*.

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