Logarithmic asymptotics for unserved messages at a FIFO.

Ken Duffy and Wayne G. Sullivan,
Communications Network Research Institute,
Dublin Institute of Technology, Rathmines, Dublin 6, Ireland.

October 16, 2002; revised December 12, 2002

Abstract

We consider an infinite–buffered single server First In, First Out (FIFO) queue. Messages arrives at stochastic intervals and take random amounts of time to process. Logarithmic asymptotics are proved for the tail of the distribution of the number of messages awaiting service, under general large deviation and stability assumptions, and formulae presented for the asymptotic decay rate.

1 Introduction

Consider an underloaded single server first-in first-out (FIFO) queue with infinite waiting space. The FIFO queueing discipline means that messages are processed in the order in which they arrive. Assume that message inter–arrival times form a stochastic process which is independent of the times required to process messages, which we call message–sizes. Define ω to be the waiting time a message experiences at the device after it has been running for an infinite period of time and η to be number of requests yet to receive service by that time.

Many authors (for example, see [6, 5, 4]) have established general Large Deviation Principle (LDP) and stability assumptions under which logarithmic asymptotics can be deduced for ω. In this work we focus on η in the G/G/1 queue (general stationary message–sizes independent of general stationary message inter–arrival times). We identify general LDP and stability assumptions under which logarithmic asymptotics can be deduced for the tail of η. In particular, under our assumptions the message–size process can have non–trivial correlation structure, as can the message inter–arrival time process. The method of proof employed is one dimensional, applying estimates directly to the event $\{\eta > \pi\}$.

Our two main results are these:

[1] Keywords: FIFO–system; Logarithmic Asymptotics; Large Deviations.

[2] AMS classification: primary 60K25; secondary 60F10;90B18

[3] Work supported by Science Foundation Ireland under the National Development Plan

- Under the LDP and stability assumptions 2.1, 2.2, 3.1 and 3.2, Theorem 3.1 proves that

$$\lim_{\pi\to\infty}\frac{1}{\pi}\log\mathbb{P}\left[\eta>\pi\right]=-\inf_{z>0}\inf_{y\geq 0}\left(zI_\xi\left(\frac{y}{z}\right)+(1+z)I_\tau\left(\frac{y}{1+z}\right)\right)=:-\delta_\eta, \qquad (1)$$

  where I_ξ is the message–size rate–function and I_τ is the message inter–arrival time rate–function.

- Under the additional assumptions 4.1 and 4.2, on the duality of the scaled cumulant generating function, and the assumptions of Proposition 2 of Glynn and Whitt [6] that ensure the following limit exists:

$$\delta_\omega := -\lim_{q\to\infty}\frac{1}{q}\log\mathbb{P}\left[\omega>q\right],$$

  Theorem 4.1 proves δ_η in Equation (1) is equal to $\lambda_\xi(\delta_\omega)=-\lambda_\tau(-\delta_\omega)$, where λ_ξ and λ_τ are the scaled cumulant generating functions for the message–size and inter–arrival time processes respectively.

In the G/GI/1 setting (i.i.d. service times independent of general stationary inter–arrival times), Glynn and Whitt [6] prove η has logarithmic asymptotics with rate $\delta_\eta = \log\mathbb{E}[\exp(\delta_\omega V)]$, where V is the service time distribution. Theorem 4.1 extends their result to the G/G/1 setting. As their proof is based on scaled cumulant generating functions, they do not get a relationship of the form given in Equation (1) for δ_η in terms of the underlying rate–functions.

Under more restrictive assumptions, Aspandiiarov and Perchersky, [2], prove a stronger result on the joint logarithmic asymptotics of ω and η. In particular, they assume that the message inter–arrival times form an i.i.d. sequence with exponential distribution and the message–sizes form an independent i.i.d. sequence with general distribution, the M/G/1 queue. Their approach uses the sample–path large deviation principle of Dobrushin and Perchersky [3], which relies on the Poisson structure of the inter–arrival times. In [1], Asmussen and Collamore extend the results in [2] to the GI/G/1 queue (i.i.d. service times independent of i.i.d. inter–arrival times) using a direct approach that also gives pre–factors.

In [11] (as an application of results developed in [10]), Russell shows that the message–queue can be written as a random time–change of the message–arrivals by the waiting–time, and uses general time–change arguments to deduce its tail behavior.

The rest of the paper is organized as follows: in Section 2, we set up our notation and define the event of interest; in Section 3, we state large deviation and stability assumptions under which η has logarithmic asymptotics; in Section 4, under assumptions of existence and duality of the scaled Cumulant Generating Function (sCGF), we provide an alternate representation of the rate of decay of the tail of η in terms of the sCGF; in Section 5, a number of examples are presented; all proofs are presented in Appendix A.

2 Self–clocked queues

Define the number of unserved messages to be the number of messages that have arrived and have received no service. We consider the waiting time and number of unserved messages just before each message arrives.

Definition 2.1 *For all $n \in \mathbb{Z}$, define the strictly positive random variable ξ_n to be the amount of time required to process message n, its message–size. For $a, b \in \mathbb{Z}$, define the total time required to process messages a to $b-1$ by $\xi[a,b] := \sum_{i=a}^{b-1} \xi_i$.*

Definition 2.2 *For all $n \in \mathbb{Z}$, define the non–negative random variable τ_n to be the time between the arrival of messages $n-1$ and n. For $a, b \in \mathbb{Z}$, define the total time between the arrival of messages a and b by $\tau[a,b] := \sum_{i=a+1}^{b} \tau_i$.*

We assume that $\{\xi_n\}$ and $\{\tau_n\}$ are independent, as we shall have to consider the message–size and inter–arrival time processes on different scales.

Assumption 2.1 *$\{\xi_n\}$ and $\{\tau_n\}$ are independent.*

Consider the waiting time which is set to be zero before message $-N$ arrives. From the single server queueing recursion, the waiting time before message 0 arrives, ω_N, is given by:

$$\omega_N := \sup_{0 \leq n \leq N} \{\xi[-n,0] - \tau[-n,0]\}.$$

Having started the waiting time to be empty before message $-N$ arrives, no messages are awaiting service before messages $-N$ arrives. Assuming the FIFO queueing discipline, the number of messages awaiting service just before message 0 arrives, η_N, is as many of the recently arrived message that can account for the waiting–time:

$$\eta_N := \sup \left\{ 0 \leq k \leq N : \xi[-k,0] < \sup_{0 \leq n \leq N} \{\xi[-n,0] - \tau[-n,0]\} \right\},$$

which is equivalent to

$$\eta_N := \sup \left\{ k : 0 < \sup_{(k,n):\ 0 \leq k \leq n \leq N} \{\xi[-n,-k] - \tau[-n,0]\} \right\}.$$

Assumption 2.2 *The sequences $\{\xi_n\}$ and $\{\tau_n\}$ are stationary, their difference $\{\xi_n - \tau_n\}$ is ergodic and $\mathbb{E}[\xi_n - \tau_n] < 0$, for all n.*

Theorem 2.1 *Under assumption 2.2, Loynes [8] proves that*

$$\omega := \lim_{N \to \infty} \omega_N = \sup_{n \geq 0} \{\xi[-n,0] - \tau[-n,0]\}$$

exists and is the waiting time just before message 0 arrives.

Lemma 2.1 *As η_N is a non-decreasing function of N, the limit*

$$\eta := \lim_{N\to\infty} \eta_N = \sup\left\{ k : 0 < \sup_{(k,n):\, 0\le k\le n} \{\xi[-n,-k] - \tau[-n,0]\} \right\},$$

exists as an extended real valued random variable.

After the system has been running for a long time, ω is the waiting time just before message 0 arrives and η is the the number of messages yet to receive any service just before message 0 arrives.

3 Logarithmic asymptotics for η

Definition 3.1 *A process, $\{Z_n\}$, taking values in $\mathbb{R}$ satisfies the* Large Deviation Principle *(LDP) with rate–function $I : \mathbb{R} \to \mathbb{R}^+ \cup \{+\infty\}$, if I is lower semi–continuous and, for all Borel sets B,*

$$-\inf_{x\in B^\circ} I(x) \le \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in B] \le \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in B] \le -\inf_{x\in \bar{B}} I(x),$$

where B° denotes the interior of B and $\bar{B}$ denotes the closure of B. A rate–function is good *if its level sets $\{x : I(x) \le \alpha\}$ are compact for all α.*

We identify sufficient conditions under which the following limit exists:

$$\lim_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\eta > \pi] =: -\delta_\eta,$$

and δ_η can be related to the large deviation properties of $\{\xi_n\}$ and $\{\tau_n\}$. Throughout, we assume the following LDPs:

Assumption 3.1 *$\{\xi[-n,0]/n\}$ satisfies the LDP with proper–convex, good rate–function $I_\xi(x)$ and $\{\tau[-n,0]/n\}$ satisfies the LDP with proper–convex, good rate–function $I_\tau(x)$. Assume also that $\inf_{y\ge0}\{I_\xi(y) + I_\tau(y)\} > 0$, so that the system will be stable on the scale of the LDP.*

Define $m_\xi := \mathbb{E}[\xi_n]$ and $m_\tau := \mathbb{E}[\tau_n]$, then $I_\xi(m_\xi) = 0$ and $I_\tau(m_\tau) = 0$ and, by the Loynes assumption 2.2, $m_\xi < m_\tau$.

In the general theory, it is possible that a finite collection of messages could cause arbitrarily large build–up in the number of unserved messages. This happens because the message–size distribution has a long, slow, tail or because it is likely, in the large deviations limit, that the message inter–arrival time will be very big during the arrival of a large number of messages. We provide a general assumption under which the behavior of a finite collection of messages is not the dominating effect.

Assumption 3.2 *There exists $\alpha, \beta > 0$ so that*

$$I_\xi(y) \geq \alpha(y - \beta), \tag{2}$$

for all $y \geq 0$. Given such α and β, define $y^* \in [0, m_\tau]$ by $y^* := \inf\{y : \alpha y \geq I_\tau(y^*)\}$. We assume

$$I_\tau(y^*) > \delta_\eta, \tag{3}$$

where δ_η is defined in Equation (4) below.

If there is a maximum message–size, we can set $y^* = 0$, so that the condition reduces to $I_\tau(0) > \delta_\eta$.

Under assumptions 2.1, 2.2, 3.1 and 3.2, we shall prove that the tail of the distribution of the number of messages awaiting service decays exponentially at rate

$$\delta_\eta = \inf_{z>0} \inf_{y \geq 0} g(z, y), \text{ where } g(z, y) := z I_\xi \left(\frac{y}{z}\right) + (1 + z) I_\tau \left(\frac{y}{1 + z}\right). \tag{4}$$

The function $g(z, y)$, has the following property:

Lemma 3.1 *Under assumption 3.1, $g(z, y)$, as defined in Equation (4), is jointly convex in z and y. Moreover, $\inf_{y \geq 0} g(z, y)$ is convex in z.*

As the tail of number of unserved messages will be determined by a transformation of the assumed LDPs, the following Lemma will prove useful:

Lemma 3.2 *Fix $z > 0$, under assumption 3.1 the process $\{(\xi[-\lceil zn \rceil, 0] - \tau[-\lceil (1+z)n \rceil, 0])/n\}$ satisfies the LDP with rate–function:*

$$I(z; x) := \inf_{y \geq 0} \left\{ z I_\xi \left(\frac{x + y}{z}\right) + (1 + z) I_\tau \left(\frac{y}{1 + z}\right) \right\}. \tag{5}$$

The lower bound follows as a direct consequence of the independence, stability and LDP hypotheses:

Proposition 3.1 *Under assumptions 2.1, 2.2 and 3.1:*

$$\liminf_{\pi \to \infty} \frac{1}{\pi} \log \mathbb{P}[\eta > \pi] \geq - \inf_{z>0} I(z; 0) = - \inf_{z>0} \inf_{y \geq 0} g(z, y) = -\delta_\eta,$$

where $I(z; x)$ is defined in Equation (5) and $g(z, y)$ is defined in Equation (4).

In order to tackle the corresponding upper bound, we use the Principle of the Largest Term, Lemma 2.3 (c) of [7], to break the probability of the event $\{\eta > \pi\}$ into three parts:

Lemma 3.3 *For each* $0 < \underline{c} < \overline{c} < \infty$,

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\eta > \pi] = \min \begin{cases} \limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n:n<\pi\underline{c}} \xi[-n,0] - \tau[-n-\pi,0] > 0]; \\ \limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n:\pi\underline{c}\le n\le\pi\overline{c}} \xi[-n,0] - \tau[-n-\pi,0] > 0]; \\ \limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n:n>\pi\overline{c}} \xi[-n,0] - \tau[-n-\pi,0] > 0]. \end{cases}$$

The first term concerns the tail of $\xi[-n,0]$, for fixed n. The second and third terms are governed by the transformed LDP given in Lemma 3.2.

Proposition 3.2 *Under assumptions 2.1, 2.2, 3.1 and 3.2, there exists* $\underline{c} > 0$ *such that*

$$-\delta_\eta > \limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n:n<\pi\underline{c}} \xi[-n,0] - \tau[-n-\pi,0] > 0],$$

where δ_η *is defined in Equation (4).*

An LDP is intrinsically a statement of pointwise convergence. In order to deal with the middle term we need a sort of uniform convergence on compact intervals. The following observations, and Lemma 3.4, will prove useful in proving Proposition 3.3. For $w \ge x > 0$, consider the following when it is a finite value:

$$f(x,w) := \liminf_{\pi\to\infty} -\frac{1}{\pi} \log \mathbb{P}[\xi[-\lceil \pi x\rceil,0] - \tau[-\lceil \pi w\rceil,0] > 0].$$

Then, for $\lambda > 0$, $f(\lambda x, \lambda w) = \lambda f(x,w)$. Hence, $f(x,1)$ for all $0 < x \le 1$ determines $f(x,w)$ for all $x \le w$. Furthermore, for fixed $\delta > 0$, (x,w) and $\lambda^* > 0$, one can choose N so that $\pi \ge N$ implies

$$-\frac{1}{\pi} \log \mathbb{P}[\xi[-\lceil \lambda\pi x\rceil,0] - \tau[-\lceil \lambda\pi w\rceil,0] > 0] \ge \lambda f(x,w) - \lambda\delta,$$

for all $\lambda > \lambda^*$. As $\{\xi_n\}$ and $\{\tau_n\}$ are non–negative, it follows that $f(x^*,w) \le f(x,w) \le f(x,w^*)$ for $x < x^*$ and $w < w^*$. By arguments similar to those used in Lemma 3.1, $f(x,1) = \inf_{y>0}(xI_\xi(y/x) + I_\tau(y))$ is convex in x. Hence (see Theorem 10.1, page 82, of [9]) it is continuous on the interior of set upon which it is finite. These remarks lead us to the following:

Lemma 3.4 *Given* $0 < c \le d$ *and* $\delta > 0$, *there exists* N *so that*

$$-\frac{1}{\pi} \log \mathbb{P}[\xi[-\lceil \pi x\rceil,0] - \tau[-\lceil \pi w\rceil,0] > 0] \ge f(x,w) - \delta,$$

for all $\pi > N$ *and all* $c \le x \le w \le d$.

Proposition 3.3 *Under assumptions 2.1 and 3.1,*

$$-\delta_\eta \geq \limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n:\pi\underline{c}<n<\pi\overline{c}} \xi[-n,0] - \tau[-n-\pi,0] > 0],$$

for all $0 < \underline{c} < \overline{c} < \infty$.

The final term is governed by the fact that for large n, it is very unlikely that $\xi[-n,0] - \tau[-n,0] > 0$, because of the LDP stability assumption.

Proposition 3.4 *Under assumptions 2.1 and 3.1, there exists* $\overline{c}$ *such that*

$$-\delta_\eta > \limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n:n>\pi\overline{c}} \xi[-n,0] - \tau[-n-\pi,0] > 0].$$

Combining Propositions 3.1, 3.2, 3.3, 3.4 and Lemma 3.3, we have the result:

Theorem 3.1 *Under assumptions 2.1, 2.2, 3.1 and 3.2, the asymptotic rate of decay of the number of unserved messages is given by:*

$$\lim_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\eta > \pi] = -\inf_{z>0} \inf_{y\geq 0} \left(zI_\xi \left(\frac{y}{z}\right) + (1+z) I_\tau \left(\frac{y}{1+z}\right)\right) =: -\delta_\eta. \tag{6}$$

4 An alternative representation: the sCGF

It is often the case that when the scaled Cumulant Generating Functions (sCGF) exists and is dual to the rate–function, it is easier to evaluate. We provide an alternate expression for δ_η in terms of the associated sCGFs, under the assumption of their existence and duality.

Assumption 4.1 *The sCGF,* $\lambda_\xi(\theta)$, *for the process* $\{\xi[-n,0]/n\}$, *exists and is the Legendre–Fenchel transform of the rate–function* $I_\xi(x)$. *That is,*

$$\lambda_\xi(\theta) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{\theta\xi[-n,0]}\right],$$

exists as an extended real number for all θ *and, for all* $x \geq 0$, $I_\xi(x) = \sup_\theta(\theta x - \lambda_\xi(\theta))$. *The sCGF,* $\lambda_\tau(\theta)$, *for the process* $\{\tau[-n,0]/n\}$ *exists and is the Legendre–Fenchel transform of the rate–function* $I_\tau(x)$.

Proposition 4.1 *Under assumptions 2.1, 3.1 and 4.1, for fixed* $z > 0$, *the process* $\{(\xi[-\lceil zn\rceil, 0] - \tau[-\lceil(1+z)n\rceil, 0])/n\}$ *has sCGF,* $\lambda(z;\theta)$, *given by*

$$\lambda(z;\theta) = z\lambda_\xi(\theta) + (1+z)\lambda_\tau(-\theta).$$

As, under the additional assumption 3.2, $\delta_\eta = \inf_{z>0} I(z;0)$, *we have:*

$$\delta_\eta = \inf_{z>0} \sup_\theta \{-\lambda(z;\theta)\} = -\sup_{z>0} \inf_\theta \lambda(z;\theta). \tag{7}$$

One can express the LDP stability assumption in the following form:

Assumption 4.2 *$\lambda_\tau(-\theta)$ is strictly decreasing where it is finite for $\theta \geq 0$ and for some θ_1, θ_2, $0 < \theta_1 < \theta_2$, λ_ξ and λ_τ are finite and*

$$\lambda_\xi(\theta_1) + \lambda_\tau(-\theta_1) < 0, \ \lambda_\xi(\theta_2) + \lambda_\tau(-\theta_2) > 0.$$

Lemma 4.1 *Under assumption 4.2, there exists a unique $\theta^* > 0$ so that $\lambda_\xi(\theta^*) + \lambda_\tau(-\theta^*) = 0$.*

Note that, under the assumptions of Proposition 2 of [6], the waiting time has an exponential tail with rate δ_ω and $\delta_\omega = \theta^*$.

Theorem 4.1 *Under assumptions 2.1, 3.1, 3.2, 4.1 and 4.2:*

$$\delta_\eta = \lambda_\xi(\theta^*),$$

where θ^* is the unique strictly positive root of $\lambda_\xi(\theta) + \lambda_\tau(-\theta) = 0$. That is, under Glynn and Whitt's [6] assumptions,

$$\delta_\eta = \lambda_\xi(\delta_\omega) = -\lambda_\tau(-\delta_\omega). \tag{8}$$

5 Examples

Example 1, constant inter-arrival time: Assume that a new message arrives at intervals of length c, then,

$$I_\tau(x) = \begin{cases} 0 & \text{if } x = c, \\ \infty & \text{otherwise.} \end{cases}$$

Hence,

$$\inf_{x>0} \inf_{y \geq 0} \left(x I_\xi \left(\frac{y}{x} \right) + (1+x) I_\tau \left(\frac{y}{1+x} \right) \right) = \inf_{x>0} \left(x I_\xi \left(\frac{c(1+x)}{x} \right) \right).$$

Thus, the rate of decay of the number of unserved messages is given by

$$\delta_\eta = \inf_{x>0} x I_\xi \left(\frac{c}{x} + c \right) = c\, \delta_\omega,$$

where δ_ω is the rate of decay of the waiting time given in Proposition 2 of [6]. That is, the tails of both ω and η decay at the same rate rescaled by the dimensional constant c, as one expects.

Example 2, exponentially distributed message–sizes and inter–arrival times: The logarithmic asymptotics in this example can be deduced from, and agree with, Theorem 1 of [2]. Let the message inter–arrival time be exponentially distributed with mean $1/\beta$. This corresponds to a rate–function

$$I_\tau(y) = \beta y - 1 - \log(\beta y).$$

Assume the message–size to be exponentially distributed with mean $1/\alpha$ so that the message–size rate–function is

$$I_\xi(y) = \alpha y - 1 - \log(\alpha y).$$

Next we consider

$$\inf_{y>0} \left\{ z\, I_\xi\left(\frac{y}{z}\right) + (1+z) I_\tau \left(\frac{y}{1+z} \right) \right\}.$$

These functions are strictly convex, so there will be a unique minimizing value of y. Differentiation with respect to y yields

$$\alpha - \frac{z}{y} + \beta - \frac{(1+z)}{y} = 0 \Longrightarrow y^* = \frac{1+2z}{\alpha+\beta}$$

at the minimum, where the value is

$$J(z) := (1+z) \log \frac{1+z}{\beta y^*} + z \log \frac{z}{\alpha y^*}.$$

Now

$$\frac{d}{dz} J(z) = \log \frac{1+z}{\beta y^*} + \log \frac{z}{\alpha y^*} + 2 - \left(\frac{1+z}{y^*} + \frac{z}{y^*} \right) \frac{2}{\alpha+\beta} .$$

The non log terms add to zero so that

$$\frac{d}{dz} J(z) = 0 \implies \log \frac{1+z}{\beta y^*} + \log \frac{z}{\alpha y^*} = 0,$$

which can be rewritten as

$$\frac{\alpha}{\alpha+\beta} \frac{\beta}{\alpha+\beta} = \frac{z}{1+2z} \frac{1+z}{1+2z}.$$

As $\beta < \alpha$, we must have equality of the ratios $\beta : \alpha = z : (1+z)$, so that

$$z = \frac{\beta}{\alpha-\beta},\ 1+z = \frac{\alpha}{\alpha-\beta}, \implies \delta_\eta = J\left(\frac{\beta}{\alpha-\beta} \right) = \log \frac{\alpha}{\beta} .$$

Relation to δ_ω: The sCGF for $\{\xi[-n,0] - \tau[-n,0])/n\}$ is

$$\phi(\theta) = \log \left[\frac{\alpha}{\alpha-\theta} \frac{\beta}{\beta+\theta} \right].$$

The asymptotic rate δ_ω is the positive zero of $\phi(\theta)$, $\phi(\delta_\omega) = 0$, so

$$\delta_\omega = \alpha - \beta.$$

Example 3, 2–state Markov message–sizes and exponential inter–arrival times: Let the message–size process form a 2–state Markov chain taking values $\{A, B\}$, where $A < B$. Let a be the probability of going from state A to state B given that the chain is in

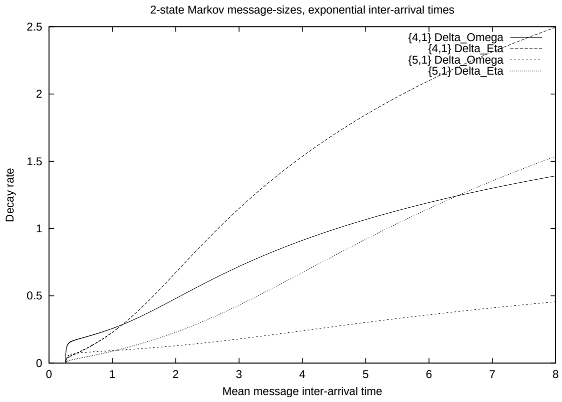

Figure 1: δ_ω and δ_η versus m_τ for 2–state Markov message–sizes with exponential inter–arrivals.

state A and let d be the probability of going from state B to state A given that the chain is in state B. If the chain is irreducible, then the sCGF is equal to the largest eigenvalue of the tilted transition matrix, which is obtained from the transition matrix by multiplying the first column by $\exp(\theta A)$ and the second column by $\exp(\theta B)$,

$$\begin{pmatrix} (1-a)e^{\theta A} & ae^{\theta B} \\ de^{\theta A} & (1-d)e^{\theta B} \end{pmatrix}$$

The mean message–size is $m_\xi = (aB + dA)/(a + d)$.

Let the inter–arrival times be a exponentially distributed with rate τ. In order for the system to be stable, the mean inter–arrival time must be greater that the mean message–size, $m_\tau > m_\xi$. The sCGF for the inter–arrival time process is given by:

$$\lambda_\tau(\theta) = \begin{cases} \log\left(\frac{\tau}{\tau-\theta}\right) & \text{if } \theta < \tau, \\ +\infty & \text{otherwise.} \end{cases}$$

δ_ω in terms of the sCGFs forms a transcendental equation. It is, however, readily solved numerically. Set $a = 1/1000$ and $d = 1/4$, so that the Markov chain is positively correlated. Figure 1 shows how δ_ω and δ_η change as the mean inter–arrival time, m_τ, is increased for two chains: in one chain the message–sizes are 0.25 and 2 and in the other chain they are 0.25 and 4. Note that, depending on the mean inter–arrival time, the tail of the waiting time distribution can decay quicker or slower than the tail of the number of unserved requests.

A Proofs

Proof of Lemma 3.1:

PROOF $g(z,y)$ is jointly convex in z and y if and only if $g(\alpha z_1+(1-\alpha)z_2, \alpha y_1+(1-\alpha)y_2) \leq \alpha g(z_1,y_1)+(1-\alpha)g(z_2,y_2)$ for all $z_1,y_1,z_2,y_2 \in (0,\infty)$, $\alpha \in [0,1]$. As the sum of two convex functions is convex, we need only prove that $h(z,y) := zI_\xi(y/z)$ is convex.

$$h(\alpha z_1+(1-\alpha)z_2, \alpha y_1+(1-\alpha)y_2) = (\alpha z_1+(1-\alpha)z_2)I_\xi\left(\frac{\alpha y_1+(1-\alpha)y_2}{\alpha z_1+(1-\alpha)z_2}\right).$$

Set $\gamma = \alpha z_1/(\alpha z_1+(1-\alpha)z_2) \in [0,1]$ and note that

$$\frac{\alpha y_1+(1-\alpha)y_2}{\alpha z_1+(1-\alpha)z_2} = \frac{\gamma y_1}{z_1}+\frac{(1-\gamma)y_2}{z_2}.$$

Using the convexity of $I_\xi(x)$,

$$h(\alpha z_1+(1-\alpha)z_2, \alpha y_1+(1-\alpha)y_2) \leq \alpha z_1 I_\xi\left(\frac{y_1}{z_1}\right)+(1-\alpha)z_2 I_\xi\left(\frac{y_2}{z_2}\right),$$

as required.

For the convexity, in z, of $\inf_{y\geq 0} g(z,y)$, note that we need only consider $\inf_{y\geq 0} h(z,y)$ and that

$$\inf_{y\geq 0} h(\alpha z_1+(1-\alpha)z_2, y) = \inf_{y_1,y_2\geq 0} h(\alpha z_1+(1-\alpha)z_2, \alpha y_1+(1-\alpha)y_2).$$

The result follows using the joint convexity of $h(z,y)$.

■

Proof of Lemma 3.2:

PROOF As $\{\xi[-n,0]/n\}$ satisfies a LDP with rate–function $I_\xi(x)$, by dilation of scale, $\{\xi[-\lceil zn\rceil,0]/n\}$ satisfies a LDP with rate–function $zI_\xi(x/z)$. Similarly, $\{\tau[-\lceil(1+z)n\rceil,0])/n\}$ satisfies a LDP with rate–function $(1+z)I_\tau(x/(1+z))$. As subtraction is continuous, the result follows applying the contraction principle, Theorem 6.4 of [7].

■

Proof of Proposition 3.1:

PROOF For fixed z,

$$\mathbb{P}[\eta > \pi] \geq \mathbb{P}\left[\xi[-\lceil z\pi\rceil,0]-\tau[-\lceil(1+z)\pi\rceil,0] > 0\right].$$

By Lemma 3.2,

$$\liminf_{\pi\to\infty}\frac{1}{\pi}\log\mathbb{P}[\eta > \pi] \geq -\inf_{x>0} I(z;x) = -I(z;0),$$

as I is convex and, by assumption 2.2, $\mathbb{E}\left[\xi[-\lceil z\pi\rceil,0]-\tau[-\lceil(1+z)\pi\rceil,0]\right] < 0$. As this is true for all $z > 0$, the result follows.

■

Proof of Proposition 3.2:

PROOF For each $\pi, \underline{c} > 0$,

$$\mathbb{P}[\sup_{n<\underline{c}\pi} \xi[-n,0] - \tau[-n-\pi,0] > 0] \leq \underline{c}\pi \max_{n<\underline{c}\pi} \mathbb{P}[\xi[-n,0] - \tau[-n-\pi,0] > 0].$$

Hence as $\limsup_{\pi\to\infty} \log(\underline{c}\pi)/\pi = 0$,

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n<\underline{c}\pi} \xi[-n,0]-\tau[-n-\pi,0] > 0] \leq \limsup_{\pi\to\infty} \frac{1}{\pi} \log \max_{n<\underline{c}\pi} \mathbb{P}[\xi[-n,0]-\tau[-n-\pi,0] > 0].$$

For each π and $0 < x \leq \underline{c}$,

$$\begin{array}{rcl} \mathbb{P}[\xi[-\lceil\underline{c}\pi\rceil,0] - \tau[-\pi,0] > 0] & \geq & \mathbb{P}[\xi[-\lceil x\pi\rceil,0] - \tau[-\pi,0] > 0] \\ & \geq & \mathbb{P}[\xi[-\lceil x\pi\rceil,0] - \tau[-\lceil(1+x)\pi\rceil,0] > 0], \end{array}$$

by positivity of the processes $\{\xi_n\}$ and $\{\tau_n\}$. Thus it suffices to prove for some $\underline{c}$ that

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\xi[-\lceil\underline{c}\pi\rceil,0] - \tau[-\pi,0] > 0] < -\delta_\eta.$$

Using assumption 3.1 we have:

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\xi[-\lceil\underline{c}\pi\rceil,0] - \tau[-\pi,0] > 0] \leq -\inf_{y\geq 0} \left\{\underline{c} I_\xi\left(\frac{y}{\underline{c}}\right) + I_\tau(y)\right\}.$$

Using assumption 3.2, there exists α such that

$$\alpha y + I_\tau(y) \geq I_\tau(y^*)$$

for all $y \geq 0$, as $I_\tau(y) \geq I_\tau(y^*)$ for all $0 \leq y \leq y^*$ and $\alpha y > I_\tau(y^*)$ for $y > y^*$. Using Equations (2) and (3), we deduce the existence of $\underline{c}$ such that

$$\inf_{y\geq 0} \left\{\underline{c} I_\xi\left(\frac{y}{\underline{c}}\right) + I_\tau(y)\right\} \geq \inf_{y\geq 0} \{\alpha y - \beta\underline{c} + I_\tau(y)\} \geq I_\tau(y^*) - \beta\underline{c} > \delta_\eta.$$

■

Proof of Proposition 3.3:

PROOF Consider

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log \mathbb{P}[\sup_{n:\pi\underline{c}<n<\pi\bar{c}} \xi[-n,0] - \tau[-n-\pi,0] > 0],$$

which is less that or equal to:

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log(\overline{c} - \underline{c})\pi \max_{n:\pi\underline{c}<n<\pi\overline{c}} \mathbb{P}[\xi[-n,0] - \tau[-n-\pi,0] > 0],$$

which is equal to:

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log \max_{n:\pi\underline{c}<n<\pi\overline{c}} \mathbb{P}[\xi[-n,0] - \tau[-n-\pi,0] > 0].$$

Given $\epsilon > 0$, by Lemma 3.4, there exists N such that for each $n \in [\underline{c}\pi, \overline{c}\pi]$

$$-\frac{1}{\pi} \log \mathbb{P}[\xi[-n,0] - \tau[-n-\pi,0] > 0] \geq I\left(\frac{n}{\pi}, 0\right) - \epsilon \geq \inf_{x\in[\underline{c},\overline{c}]} I(x;0) - \epsilon,$$

for all $\pi > N$. The result follows taking ϵ arbitrarily close to zero.

■

Proof of Proposition 3.4:

PROOF We have:

$$\begin{aligned} \mathbb{P}[\sup_{n\geq\overline{c}\pi} \{\xi[-n,0] - \tau[-n-\pi,0]\} > 0] &\leq \sum_{n\geq\overline{c}\pi} \mathbb{P}[\xi[-n,0] - \tau[-n-\pi,0] > 0] \\ &\leq \sum_{n\geq\overline{c}\pi} \mathbb{P}[\xi[-n,0] - \tau[-n,0] > 0]. \end{aligned}$$

Using the contraction principle, Theorem 6.4 of [7], we know that:

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\xi[-n,0] - \tau[-n,0] > 0] \leq -\inf_{y\geq 0} \{I_\xi(y) + I_\tau(y)\} =: -\gamma$$

and that γ is positive by the LDP stability assumption 3.1. Given $0 < \epsilon < \gamma$, there exists N_ϵ so that

$$\mathbb{P}[\xi[-n,0] - \tau[-n,0] > 0] \leq e^{-n\{\gamma-\epsilon\}},$$

for all $n > N_\epsilon$. Fix $0 < \epsilon < \gamma$ and let $\overline{c}\pi > N_\epsilon$, then

$$\sum_{n\geq\overline{c}\pi} \mathbb{P}[\xi[-n,0] - \tau[-n,0] > 0] \leq \sum_{n\geq\overline{c}\pi} e^{-n(\gamma-\epsilon)} \leq \int_{x\geq\overline{c}\pi-1} e^{-x(\gamma-\epsilon)} dx = e^{-(\overline{c}\pi-1)(\gamma-\epsilon)}.$$

Hence,

$$\limsup_{\pi\to\infty} \frac{1}{\pi} \log \sum_{n\geq\overline{c}\pi} \mathbb{P}[\xi[-n,0] - \tau[-n-\pi,0] > 0] \leq -\overline{c}(\gamma - \epsilon),$$

and we can choose $\overline{c}$ such that $\overline{c}(\gamma - \epsilon) > \delta_\eta$.

■

Proof of Proposition 4.1:

PROOF For fixed $z > 0$,

$$\begin{aligned}\lambda(z;\theta) &:= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\exp\left(\theta(\xi[-\lceil zn\rceil, 0] - \tau[-\lceil(1+z)n\rceil, 0])\right)\right] \\ &= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{\theta(\xi[-\lceil zn\rceil,0])}\right] + \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{-\theta(\tau[-\lceil(1+z)n\rceil,0])}\right] \\ &= z\lambda(\theta) + (1+z)\lambda(-\theta),\end{aligned}$$

because of independence.

Under the additional assumptions, $\delta_\eta = \inf_{z>0} I(z;0) = \inf_{z>0}\sup_\theta\{-\lambda(z;\theta)\}$, by the Legendre–Fenchel transform formula.

■

Proof of Lemma 4.1:

PROOF The assumption 4.2 and convexity ensures the root θ^* is between θ_1 and θ_2. Since the convex function $\lambda_\xi(\theta) + \lambda_\tau(-\theta)$ has roots $\theta = 0$ and $\theta = \theta^*$, and $\lambda_\xi(\theta_1) + \lambda_\tau(-\theta_1) < 0$, there are no other roots.

■

Proof of Theorem 4.1:

PROOF Let θ^* be the unique strictly positive root of $\lambda_\xi(\theta) + \lambda_\tau(-\theta) = 0$, then

$$\inf_\theta (z\lambda_\xi(\theta) + (1+z)\lambda_\tau(-\theta)) \le z\lambda_\xi(\theta^*) + (1+z)\lambda_\tau(-\theta^*) = \lambda_\tau(-\theta^*) = -\lambda_\xi(\theta^*).$$

Since the right hand side is independent of z, $\lambda_\xi(\theta^*)$ is a lower–bound for δ_η.

For the upper–bound: let a be a subgradient slope to $\lambda_\xi(\theta)$ at θ^*; $-b$, to $\lambda_\tau(-\theta)$ at θ^*:

$$\lambda_\xi(\theta^* + t) \ge \lambda_\xi(\theta^*) + a\,t, \quad \lambda_\tau(-\theta^* - t) \ge \lambda_\tau(\theta^*) - b\,t.$$

That $a > b > 0$ follows from assumption 4.2, the convexity of $\lambda_\xi(\theta)$, $\lambda_\tau(\theta)$, and the strict monotonicity of $\lambda_\tau(-\theta)$ for $\theta > 0$. Define

$$z^* := \frac{b}{a-b} \quad \text{so that} \quad 1 + z^* = \frac{a}{a-b}.$$

For this fixed value of z^*, consider

$$\inf_\theta(z^*\,\lambda_\xi(\theta) + (1+z^*)\lambda_\tau(-\theta)) = \inf_\theta\left(\frac{b}{a-b}\lambda_\xi(\theta) + \frac{a}{a-b}\lambda_\tau(-\theta)\right).$$

With $\theta = \theta^* + t$,

$$\frac{b}{a-b}\lambda_\xi(\theta) + \frac{a}{a-b}\lambda_\tau(-\theta) \ge \frac{b}{a-b}(\lambda_\xi(\theta^*) + a\,t) + \frac{a}{a-b}(\lambda_\tau(-\theta^*) - b\,t)$$

This shows that

$$\inf_\theta\{\, z^*\,\lambda_\xi(\theta) + (1+z^*)\lambda_\tau(-\theta)\} \ge \lambda_\tau(-\theta^*).$$

From (7) we have $\delta_\eta \le -\lambda_\tau(-\theta^*) = \lambda_\xi(\theta^*)$.

■

References

[1] S. Asmussen and J. F. Collamore. Exact asymptotics for a large deviations problem for the GI/G/1 queue. *Markov Processes and Related Fields*, 5(4):451–476, 1999.

[2] S. Aspandiiarov and E. A. Perchersky. Large deviations problem for compound Poisson processes in queueing theory. *Markov Processes and Related Fields*, 3(3):333–366, 1997.

[3] R. L. Dobrushin and E. A. Perchersky. Large deviations for random processes with independent increments on finite intervals. *Problems of Information Transmission*, 34:354–384, 1998.

[4] K. Duffy, J. T. Lewis, and W. G. Sullivan. Logarithmic asymptotics for the supremum of a stochastic process. To appear in the Annals of Applied Probability.

[5] A. J. Ganesh and N. O'Connell. A large deviation principle with queueing applications. *Stochastics and Stochastic Reports*, 73(1–2):25–35, 2002.

[6] P. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, 31A:413–430, 1994.

[7] J. T. Lewis and C. E. Pfister. Thermodynamic probability theory: some aspects of large deviations. *Russian Mathematical Surveys*, 50(2):279–317, 1995.

[8] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proceedings of the Cambridge Philosphical Society*, 58:497–520, 1962.

[9] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[10] R. Russell. *The Large Deviations of Random Time Changes*. PhD thesis, Trinity College Dublin, 1997.

[11] R. Russell. Private communication, 2001.