A large deviations characterization of the fixed point of a $\cdot/G/1$ queue

A. J. Ganesh, Neil O'Connell, and Balaji Prabhakar†

BRIMS, Hewlett-Packard Laboratories, Filton Road, Bristol BS12 6QZ, UK, e-mail: ajg,noc@hplb.hpl.hp.com
† Laboratory for Information and Decision Systems, MIT, Cambridge, MA 02139, USA, e-mail: balaji@lids.mit.edu

Abstract — **This paper characterizes the large deviations behaviour of the fixed point of a $\cdot/G/1$ queue. Given a general service time distribution with mean 1 and any $\alpha < 1$, the large deviation rate function, I_α, of the fixed point with mean arrival rate α is derived. I_α is shown to be identical to the rate function of an *exponential tilting* of the service distribution. An implication of this result is that the fixed point has *minimum relative entropy* with respect to the service process over all processes satisfying the constraint that the mean arrival rate is α.**

I. Introduction

Burke's theorem says that if the arrival process to a $\cdot/M/1$ queue is Poisson with rate less than the service rate, then the departure process in equilibrium is also Poisson of the same rate. In other words, a Poisson process of rate α is a fixed point of the $\cdot/M/1$ queue with service rate 1, for every $\alpha < 1$. It has recently been shown that a similar result holds for single–server queues with a general service time distribution [7, 10]. More precisely, given a service time distribution with mean 1, and any $\alpha < 1$, there is a stationary and ergodic arrival process with law μ_α and mean arrival rate α such that the equilibrium departure process also has law μ_α. Moreover, μ_α is unique for each $\alpha < 1$. However, an explicit description of the laws $\{\mu_\alpha, 0 < \alpha < 1\}$ is not known.

In this paper, given a general service time distribution with mean 1, and any $\alpha < 1$, we derive the large deviation rate function, I_α, of the fixed point with mean arrival rate α. I_α has a simple description in terms of the rate function of the service time distribution; in fact, I_α is identical to the rate function of an exponential tilting of the service distribution. An implication of this result is that the fixed point has minimum relative entropy with respect to the service process over all processes satisfying the constraint that the mean arrival rate is α.

Consider a queue with arrivals having rate function I_{α_1} and service times having rate function I_{α_2}, where $\alpha_1 < \alpha_2$. Then, our results imply that the departure process in equilibrium has the same rate function, I_{α_1}, as the arrivals. The analogue of this property in the $\cdot/M/1$ context is that a Poisson process of rate α_1 is a fixed point of the queue with Exp(α_2) service times, for every $\alpha_1 < \alpha_2$.

II. Model and Preliminaries

The results in this paper are derived in the context of a discrete time queueing model which we now describe. The queue has arrival process denoted $\{A_n, n \in \mathbb{Z}\}$, and service process $\{S_n, n \in \mathbb{Z}\}$, assumed stationary and ergodic. A_n denotes the amount of work arriving in the n^{th} time slot and S_n denotes the maximum amount of work that can be completed in the n^{th} time slot. The queue is assumed to be work-conserving, so the evolution of the workload process, $\{W_n\}$, is described by Lindley's recursion: $W_{n+1} = \max\{W_n + A_n - S_n, 0\}$. The amount of work departing in time slot n is given by

$$D_n = A_n + W_n - W_{n+1} = \min\{W_n + A_n, S_n\}. \quad (1)$$

If A_n and S_n are integer–valued for all n, then the workload can be thought of as the number of customers in the queue. The fixed point problem can now be posed as follows: given the law of the service process, $\{S_n\}$, can we find a law μ such that, if the arrival process $\{A_n\}$ is stationary with law μ, then so is the departure process, $\{D_n\}$? In this paper, we address the somewhat simpler question of finding a large deviation rate function such that both the arrival and departure process obey a large deviation principle (LDP) with this same rate function.

We shall say that a real-valued stationary process $\{X_n, n \in \mathbb{Z}\}$ satisfies an LDP with good rate function I if I is a non-negative lower semicontinuous function with compact level sets, and for all "nice" sets $B \subseteq \mathbb{R}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in B\right) = -\inf_{x\in B} I(x).$$

A set B is nice if it is Borel-measurable and if I has the same infimum on the interior as on the closure of B. The Gärtner-Ellis theorem [3] provides sufficient conditions for a process to satisfy an LDP; these conditions are quite mild and are satisfied by most processes commonly encountered in queueing applications.

Suppose that the arrival process $\{A_n\}$ and the service process $\{S_n\}$ satisfy the assumptions of the Gärtner-Ellis theorem, with limiting logarithmic moment generating functions denoted Λ_A and Λ_S respectively. Then $\{A_n\}$ and $\{S_n\}$ satisfy LDPs with *convex, good* rate functions I_A and I_S that are the convex duals of Λ_A and Λ_S respectively. In other words,

$$I_A(x) = \sup_{\theta\in\mathbb{R}} [\theta x - \Lambda_A(\theta)], \quad I_S(x) = \sup_{\theta\in\mathbb{R}} [\theta x - \Lambda_S(\theta)], \quad (2)$$

and a similar relation holds with the roles of I_A and Λ_A or I_S and Λ_S reversed.

In the usual large deviations scaling, processes such as arrivals, services and departures are modeled as fluids with stochastic flow rate; the large deviations rate function of the process describes (the negative logarithm of) the probability that the flow rate has a specified value. If an event of interest can be described as a "continuous" function of the sample paths of the arrival and service processes, then the contraction principle says that its probability is approximately equal to the probability of the most likely way that this event can happen.

III. The Rate Function for Fixed Points

It has been shown by several authors (see, e.g., [2, 4]) that the tail of the workload distribution in equilibrium is exponential with parameter δ, i.e.,

$$\lim_{b\to\infty} \frac{1}{b} \log P(W > b) = -\delta, \quad (3)$$

0-7803-4408-1/98/$10.00

where

$$\delta = \inf_{T>0} TI_W(1/T) \tag{4}$$

and, for $w > 0$,

$$I_W(w) = \inf_{a \geq w} \left[I_A(a) + I_S(a-w) \right]. \tag{5}$$

δ has the following alternative characterization:

$$\delta = \sup\{\theta : \Lambda_A(\theta) + \Lambda_S(-\theta) \leq 0\}. \tag{6}$$

The above equations have the following interpretation. In order for the workload to build up at rate w over a long period of time, arrivals over this period must occur at some rate a exceeding the service rate by w; the most likely way for this to happen is found by minimizing the expression in (5) over all possible choices of a. Large workloads occur by the queue building up at rate $1/T$ over a period of (scaled) length T, chosen optimally according to (4).

The large deviations rate function of the equilibrium departure process has been derived by the second author [8]; here, we describe and use his result intuitively. Assume without loss of generality that the mean service rate is 1 and that the mean arrival rate is $\alpha < 1$. What is the most likely way that the departures over a long period of time have mean rate d, for some $d \leq 1$? Intuitively, this would require the arrivals to have mean rate d over this period and the services to have mean rate 1. Since $d \leq 1$, all arrivals depart (recall that on the large deviations scale we use fluid models for all processes), so that the departure process has the desired mean rate d. Consequently, the large deviations rate function for departures evaluated at d must be given by

$$I_D(d) = I_A(d) + I_S(1) = I_A(d).$$

It was shown by the first two authors [5] that this intuition isn't always correct. It is correct under a certain condition on the rate functions of the arrival and service processes, namely, that $I_A(x) \leq I_S(x)$ for all $x \leq \alpha$. It is easy to see that this condition is satisfied by the fixed point rate function defined below.

Next, what is the most likely way to obtain departures at mean rate $d > 1$? One possibility is that both arrivals and departures occur at mean rate d. This would imply that the departure rate function is given by $I_D(d) = I_A(d) + I_S(d)$. But then $I_A(d) \neq I_D(d)$ since $I_S(d) > 0$. Hence, such an arrival process can't be a fixed point. The other possibility is that we start with a non-empty queue, have arrivals at rate a and services at rate $d > a$. Then departures will have rate d provided that the queue doesn't empty during the period in consideration (which is scaled to have length 1), i.e., if the initial queue size is at least $d-a$. By (3), this has probability approximately equal to $\exp -\delta(d-a)$, and so we obtain for the departure rate function, the formula

$$\begin{aligned} I_D(d) &= \inf_{0 \leq a \leq d} \left[\delta(d-a) + I_A(a) + I_S(d) \right] \\ &= \delta d + I_S(d) - \Lambda_A(\delta), \end{aligned} \tag{7}$$

where (see (2)) $\Lambda_A(\delta) = \sup_a \ \delta a - I_A(a)$. We want $I_D = I_A$ and $I_A(\alpha) = 0$ (since arrivals are assumed to have mean rate α). By (7), this means that $\Lambda_A(\delta) = \delta\alpha + I_S(\alpha)$, and

$$I_A(d) = \delta(d-\alpha) + I_S(d) - I_S(\alpha). \tag{8}$$

Now, for I_A to be non-negative, its minimum should be achieved at α, since $I_A(\alpha) = 0$. It follows from (8) that

$$\delta = -I_S'(\alpha), \quad I_A(d) = I_S(d) - I_S(\alpha) - I_S'(\alpha)(d-\alpha), \tag{9}$$

provided I_S is differentiable at α (otherwise, we show in [6] that $-\delta$ is the infimum of the subdifferential of I_S at α). It can be verified that if I_A is defined by (9), then the value of δ given by (6) is the same as that given in (9), and that $I_D = I_A$. Does this imply that I_A is the rate function of the fixed point? We show in [6], using results of Mairesse and Prabhakar [7], that it does.

By taking duals in (9), we get

$$\Lambda_A(\theta) = \Lambda_S(\theta + I_S'(\alpha)) + I_S(\alpha) - \alpha I_S'(\alpha),$$

i.e., the rate function of the fixed point corresponds to an exponential tilting of the service process. This generalizes a well-known result for the $\cdot/M/1$ queue.

It was shown by Anantharam [1] that the most likely behaviour leading to the build-up of large delays in a $G/G/1$ queue involves the arrival and service processes having empirical distributions equal to exponential tiltings of their true distributions, with tilting parameters δ and $-\delta$ respectively, for δ given by (6). When the arrival rate function is given by (9), this implies that the most likely path leading to large delays has the service process looking like the true arrivals process, and the arrival process looking like the true service process. This switching of roles between arrivals and services is exactly how large delays build up in an $M/M/1$ queue (see, for example, [9]).

IV. Conclusions

Our main result is that the fixed point of a $\cdot/G/1$ queue has large deviations behaviour identical to an exponential tilting of the service process. In this paper, we have confined ourselves to an intuitive sketch of the key ideas behind this result. Full details can be found in [6].

References

[1] V. Anantharam, "How large delays build up in a $GI/G/1$ queue," *Queueing Systems* vol. 5, pp. 345–368, 1988.

[2] C. S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control* vol. 39, pp. 913–931, 1994.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, 1993.

[4] N. Duffield and Neil O'Connell, "Large deviations and overflow probabilities for the general single server queue, with applications," *Math. Proc. Camb. Phil. Soc.* vol. 118, pt. 1, 1995.

[5] A. J. Ganesh and Neil O'Connell, "The linear geodesic property is not generally preserved by a FIFO queue," to appear in *Ann. Appl. Prob.*, 1998.

[6] A. J. Ganesh, Neil O'Connell and Balaji Prabhakar, "A large deviations characterization of the fixed point of a $\cdot/G/1$ queue," in preparation.

[7] Jean Mairesse and Balaji Prabhakar, "On the existence of fixed points for the $\cdot/GI/1$ queueing operator," In preparation.

[8] Neil O'Connell, "Large deviations for departures from a shared buffer," *J. Appl. Prob.*, vol. 34, pt. 3, pp. 753–766, 1997.

[9] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues," *IEEE Trans. Autom. Control* vol. 34, pp. 54–66, 1989.

[10] B. Prabhakar, "On the attractiveness of the fixed points of $\cdot/GI/1$ queues," In preparation.