Exact Asymptotics in a Multi-class $M/G/1$ Queue *

Jiyeon Lee† · André Dabrowski‡ · David R. McDonald§

Abstract

Consider a multitype queue where queued customers are served in their order of arrival at a rate which depends on the customer type. Here we calculate the sharp asymptotics of the probability the total number of customers in the queue reaches a high level before emptying. The natural state space to describe this queue is a tree whose branches increase in length as the number of customers in the queue grows. Consequently it is difficult to prove a large deviation principle. Moreover, since service rates depend on the customer type the stationary distribution is not of product form so there is no simple expression for the stationary distribution. Instead, we use a change of measure technique which increases the arrival rate of customers and decreases the departure rate thus making large deviations common.

Keywords: Rare events, change of measure, h transform, quasi-stationarity

1 Introduction

In this paper, we derive an exact asymptotic expression for a large deviation of the total number of customers in a multiclass FIFO queue. Suppose customers are of different classes in a set C. We assume class c customers arrive independently at the server according to a Poisson process with rate λ_c. We assume a class c customer at the head of line is served at rate μ_c. Without loss of generality, assume

$$\sum_{c\in C}(\lambda_c + \mu_c) = 1.$$

A state of the queue is given by a branch x in a tree representing the classes of customers in the queue. $x = (x_0, x_1, \ldots x_{n-1})$ and $|x| = n$ if there are n customers in

*This research was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)(KRF-2005-204-C00016)

†Associate Professor, Department of Statistics, Yeungnam University, Kyeongsan, 712-749, Korea E-mail: leejy@yu.ac.kr

‡Professor, Department of Mathematics and Statistics, University of Ottawa, Canada

§Professor, Department of Mathematics and Statistics, University of Ottawa, Canada

the queue and the head of line customer has class x_0 and the next in line is x_1 and so on. Let ϕ denote the state of the empty queue.

The $M|G|1$ multiclass queue with FIFO service is quasi-reversible (Baskett et al. 1975; Walrand 1988) if each class served is at the same rate μ. Our queue is *not* quasi-reversible because of fact that the service rate depends on the customer class. Consequently the stationary distribution Π is not of product form and has no simple representation. There do exist representations of the steady state using transform techniques. See the recent paper by Choi et al.(2000) for a detailed bibliography (Choi et al.(2000) even allows feedback). Our goal here is to describe the asymptotics of $\Pi(x)$ for $|x| = \ell$ as $\ell \to \infty$ as well as the mean hitting time and hitting distribution when the number of customers reaches a high level ℓ.

We can view this queue as one where the customer type is only determined at the moment service starts. Then the system is an $M|G|1$ queue with an arrival rate $\lambda = \sum_{c\in C} \lambda_c$. The service time density and the associated generating function are

$$g(s) = \sum_{c\in C} \frac{\lambda_c}{\lambda} \mu_c \exp(-\mu_c s) \text{ for } s \geq 0, \text{ and } \phi_G(\theta) = \sum_{c\in C} \frac{\lambda_c}{\lambda} \frac{\mu_c}{\mu_c - \theta}.$$

Such an $M|G|1$ queue is stable if and only if $\rho = \lambda m_G < 1$ where $m_G = \sum_{c\in C} \lambda_c/(\lambda\mu_c)$ is the mean service time. It follows the queue is stable if and only if $\rho = \sum_{c\in C} \lambda_c/\mu_c < 1$.

In fact from Corollary 1 of McDonald and Théberge (2000) we have the asymptotics of the number of customers in an $M|G|1$ queue. As $\ell \to \infty$,

$$\Pi(\{x : |x| = \ell\}) \sim C_0 e^{-\Gamma\ell} \tag{1.1}$$

where Γ solves $\phi_G(\lambda(e^{\Gamma} - 1)) = \exp(\Gamma)$ and

$$C_0 = (1 - \lambda m_G) \frac{(e^{\Gamma} - 1)}{(\lambda \phi_G'(\lambda(e^{\Gamma} - 1)) - 1)}. \tag{1.2}$$

We would, however, like more detailed information about the queue and the trajectory when there are ℓ customers in the queue for the first time.

For any point $x = (x_0, \ldots, x_{n-1})$ let $N_c(x) = \#\{x_k = c, k \geq 0\}$: thus $N_c(x)$ counts the number of customers of class c waiting in the queue or currently being served. Define H at the empty queue to be 1 and for x representing the content of a non-empty queue,

$$H(x) = \prod_{c\in C} \left(e^{\gamma_c \cdot N_c(x)}\right).$$

Note that H is harmonic on $\{x : |x| > 0\}$ as long as for all $a \in C$

$$\sum_{c\in C} \lambda_c e^{\gamma_c} + \mu_a e^{-\gamma_a} = \sum_{c\in C} \lambda_c + \mu_a$$

or equivalently, if

$$\sum_{c\in C} \lambda_c e^{\gamma_c} + \mu_a e^{-\gamma_a} + \sum_{c\neq a} \mu_c = 1.$$

A non-trivial solution for the γ_a exists. In fact, by taking differences of the above constraint we find that $e^{-\gamma_a} = (e^{-\gamma_1}\mu_1 + \mu_a - \mu_1)/\mu_a$. Thus solving for the γ_a reduces to solving for γ_1 in the following manner.

$$\begin{aligned}
0 &= e^{\gamma_1}\lambda_1 + \left(\sum_{d\neq 1} e^{\gamma_d}\lambda_d + \sum_{d\neq 1}\mu_d - 1\right) + e^{-\gamma_1}\mu_1 \\
&= (e^{\gamma_1}-1)\lambda_1 + \left(\sum_{d\neq 1}(e^{\gamma_d}-1)\lambda_d\right) + (e^{-\gamma_1}-1)\mu_1 \\
&= (1-e^{-\gamma_1})\mu_1\left(\sum_d \frac{\lambda_d}{(e^{-\gamma_1}-1)\mu_1 + \mu_d}\right) + (e^{-\gamma_1}-1)\mu_1 \\
1 &= \sum_d \frac{\lambda_d}{(e^{-\gamma_1}-1)\mu_1 + \mu_d}. \qquad (1.3)
\end{aligned}$$

Solving this last equality provides γ_1, and then the other γ_a are obtained as above. We can also interpret (1.3) as

$$1 = \sum_d \frac{\lambda_d}{\widetilde{\mu}_d}. \qquad (1.4)$$

We investigate asymptotics of Π by performing a change of measure associated with this harmonic function H. If a class c customer is being served at the head of the queue, at the next transition of the twisted walk a class a customer arrives with probability $\widetilde{\lambda}_a := \lambda_a e^{\gamma_a}$ and the head-of-the-line (HOL) class c customer is served with probability $\widetilde{\mu}_c := \mu_c e^{-\gamma_c}$. The twisted total customer arrival rate is $\widetilde{\lambda} := \sum_{d\in C}\widetilde{\lambda}_d$. The twisted load of class c customers is $\widetilde{\rho}_c := \widetilde{\lambda}_c/\widetilde{\mu}_c$ and the twisted total load is $\widetilde{\rho} := \sum_{c\in C}\widetilde{\rho}_c$. This allows us to state the following multi-type version of (1.1).

Theorem 1.1 *In the above notation,*

$$e^{\Gamma} = \widetilde{\lambda}/\lambda,$$

and so

$$C_0 = \frac{1-\rho}{\widetilde{\rho}-1}\left(\frac{\widetilde{\lambda}}{\lambda} - 1\right).$$

As $\ell \to \infty$,

$$\pi(\ell, a) \equiv \Pi(\{x : |x| = \ell, x_0 = a\}) \sim \left(\frac{\lambda_a}{\widetilde{\mu}_a}\right) C_0 e^{-\Gamma \ell}. \tag{1.5}$$

Example 1 *Take* $C = \{1, 2\}$, $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\mu_1 = 0.3$ *and* $\mu_2 = 0.4$. *Then* $m_G = 2.78$ *and* $\rho = 5/6 < 1$. *The queue is stable and we can solve* $\phi_G(\lambda(e^{\Gamma} - 1)) = \exp(\Gamma)$ *to obtain* $e^{\Gamma} \simeq 1.1953$ *in (1.1). Solving for the twisted system yields* $e^{\gamma_1} \simeq 1.243$, $e^{\gamma_2} \simeq 1.172$, $\widetilde{\lambda}_1 \simeq 0.1243$, $\widetilde{\lambda}_2 \simeq 0.2343$, $\widetilde{\mu}_1 \simeq 0.2414$, $\widetilde{\mu}_2 \simeq 0.3414$, $\widetilde{\rho}_1 \simeq 0.515$, $\widetilde{\rho}_2 \simeq 0.686$ *and* $\widetilde{\rho} \simeq 1.201$.

Note that in view of (1.4), summing (1.5) over a yields (1.1). Theorem 1.1 also specifies the equilibrium distribution of clients at the head of the line similarly to line (7.11) of McDonald(2004), which states the equilibrium composition (i.e. relative numbers of each type of client) for the multiclass M/M/1 queue. We can provide a complete analog to that result as well.

Lemma 1.2 $\Pi((a, x_1, \ldots x_{n-1})) = \pi(n, a) \prod_{k=1}^{n-1} \frac{\lambda_{x_k}}{\lambda}$ *and* $\Pi(\phi) = \pi(\phi)$.

Since we have the asymptotics of $\pi(\ell, a)$ we can summarize:

Corollary 1.3 *Fix* k *to be a positive integer and* $(a_0, \ldots, a_k)$. *As* $\ell \to \infty$,

$$\Pi(\{x : |x| = \ell,\ x_i = a_i\ i = 0 \ldots k\}) \sim \left(\frac{\lambda_{a_0}}{\widetilde{\mu}_{a_0}}\right) \prod_{i=1}^{k} \left(\frac{\lambda_{a_i}}{\lambda}\right) C_0 e^{-\Gamma \ell}. \tag{1.6}$$

Let τ_ℓ denote the first $t > 0$ that the multiclass queue reaches size ℓ.

Theorem 1.4 *For an* $M|G|1$ *queue let* $P_0(H_\ell)$ *denote the probability the queue size grows from 0 to* ℓ *without returning first to 0, and let* $E_0\tau_\ell$ *denote the mean time for the queue to reach level* ℓ. *Then*

$$P_0(H_\ell) \sim e^{-\Gamma \ell} e^{\Gamma} \frac{1 - \lambda m_G}{\lambda \phi_G'(\lambda(e^{\Gamma} - 1)) - 1}(e^{\Gamma} - 1)$$

and

$$E_0\tau_\ell \sim e^{\Gamma \ell} \frac{1}{\lambda(1 - \lambda m_G)} e^{-\Gamma} \frac{\lambda \phi_G'(\lambda(e^{\Gamma} - 1)) - 1}{1 - \lambda m_G}(e^{\Gamma} - 1)^{-1},$$

where $\phi_G(t) = \int_0^\infty e^{tx} dG(x)$, m_G *is the mean of* G *and* Γ *is the unique solution to* $\phi_G(\lambda(e^{\Gamma} - 1)) = \exp(\Gamma)$.

If we apply Theorem 1.4 to the $M|M|1$ multiclass case with $\mu_b = \mu$ for all b, then $e^{\Gamma} = \mu/\lambda$. By substitution,

$$E_0\tau_\ell \sim e^{\Gamma\ell}\frac{\mu}{(\mu-\lambda)^2}.$$

The time for a birth and death process with birth rate λ and death rate μ to reach level ℓ can be solved exactly as in Ch. XIV.3 of Feller(1968),

$$\begin{aligned} E_0\tau_\ell &= \frac{\mu}{(\mu-\lambda)^2}((\frac{\mu}{\lambda})^\ell - 1)) - \frac{\ell}{\mu-\lambda} \\ &= e^{\Gamma\ell}\frac{\mu}{(\mu-\lambda)^2} + o(\ell). \end{aligned}$$

The approximation error of Theorem 1.4 grows only linearly with ℓ in the $M|M|1$ case.

References

[1] Baccelli F., McDonald D. (2005). Rare Events for Stationary Processes. Stochastic Processes and their Applications Vol. 89. 141-173.

[2] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G. (1975). Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. Journal of the ACM (JACM), 22, no. 2, 248-260.

[3] Choi, B.D., Kim, B, and Choi, S.H. (2000). On the $M|G|1$ Bernoulli Feedback Queue with Multi-class Customers. Computers & Operations Research, 27, 269-286.

[4] Feller W. (1968). An Introduction to Probability Theory and Its Applications, Volume I. John Wiley and Sons, New York.

[5] Kesten H. (1974). Renewal Theory for Functionals of a Markov Chain with General State Space. Ann. Probab. 2, no. 3, 355-386.

[6] McDonald D. (2004). Elements of Applied Probability for Engineering, Mathematics and Systems Science. World Scientific, River Edge N.J.

[7] McDonald D., Théberge, F. (2000). Cell Loss Probabilities for $M|G|1$ and Time-Slotted Queues. J. Appl. Probab. 37, no. 4, 1149–1156.

[8] Walrand J. (1988). An Introduction to Queuing Networks, GL Jordan, Ed., Prentice Hall, England Cliffs, NJ .