Queueing Systems
https://doi.org/10.1007/s11134-022-09769-4

Moderate deviation asymptotics of the $GI/G/n$ queue in the Halfin–Whitt regime

Anatolii A. Puhalskii[1]

Received: 2 February 2022 / Accepted: 28 February 2022

1 Introduction

The Halfin–Whitt heavy-traffic scaling for a multiserver queue with a large number of servers was introduced in a seminal paper by Halfin and Whitt [4] in order to account for the situation where the probability of a customer's delay lies strictly between 0 and 1. It has since gained great popularity and the literature on the subject is enormous. The bulk of the work done assumes that the service times are exponentially distributed. There are few results concerning general service time distributions. Even when they are available, the limit process is complex and the hypotheses introduced are often heavily affected by the specific techniques used, see, e.g., Aghajani and Ramanan [1], Gamarnik and Goldberg [3], Kaspi and Ramanan [5], Puhalskii and Reed [8]. Notably, obtaining approximations for the distribution of the steady-state number of customers has been problematic. This paper attempts to get a handle on the latter distribution by establishing, first, a large deviation principle in the scaling of moderate deviations for the transient distribution and, then, by evaluating a quasipotential which is related to the asymptotics of the stationary distribution. Deviation functions for moderate deviations being quadratic, the variational problem of evaluating the quasipotential may be tractable. Furthermore, it is speculated below that the quasipotential may furnish a candidate for the density of a stationary Halfin–Whitt limit, which, in turn, would shed light on the stationary queue length distribution. Generally, the use of asymptotics for producing approximations is discussed in Whitt [9].

2 Submission

The model is as follows. Assume as given a sequence of many-server queues indexed by n, where n represents the number of servers. Service is performed on a first-come-first-served basis. The service times of customers that are either in the queue at time 0 or arrive after time 0 have distribution function $F(t)$, for all n. It is assumed that $F(t)$ is continuous, $F(0) = 0$, and the mean $\mu^{-1} = \int_0^\infty (1 - F(s))\, ds$ is finite. The

✉ Anatolii A. Puhalskii
puhalski@iitp.ru

[1] Institute for Problems in Information Transmission (IITP), Moscow, Russia

Published online: 29 March 2022

residual service times of customers in service at time 0 have distribution function $F_0(t) = \mu \int_0^t (1 - F(s))\, ds$. The arrival process is a renewal process of rate λ_n. The usual independence assumptions are made. For the n-th queue, let $Q_n(t)$ denote the number of customers that are either in the queue or in service at time t, let $A_n(t)$ denote the number of arrivals by time t and let $\rho_n = \lambda_n/(n\mu)$.

The statements below use the following definition. Given a sequence $r_n \to \infty$, as $n \to \infty$, a sequence P_n of probability distributions on the Borel σ-algebra of a metric space M and a $[0,\infty]$-valued function I on M such that the sets $\{m \in M : I(m) \le \gamma\}$ are compact for all $\gamma \ge 0$, the sequence P_n is said to obey a large deviation principle (LDP) for rate r_n with deviation function I provided $\lim_{n\to\infty} 1/r_n \ln P_n(B) = -\inf_{m\in B} I(m)$, for every Borel set B such that the infima of I over the interior of B and over the closure of B agree.

Conjecture 1 Let $b_n \to \infty$ and $b_n/\sqrt{n} \to 0$, as $n \to \infty$. Let $\beta \in \mathbb{R}$ and $z_0 \in \mathbb{R}$. Suppose that, as $n \to \infty$, $\sqrt{n}(1-\rho_n)/b_n \to \beta$ and the sequence of random variables $\sqrt{n}/b_n\, (Q_n(0)/n - 1)$ obeys an LDP in $\mathbb{R}$ for rate b_n^2 with deviation function $I_{z_0}(y)$ such that $I_{z_0}(z_0) = 0$ and $I_{z_0}(y) = \infty$, for $y \ne z_0$. Suppose that the sequence of processes $\bigl(\sqrt{n}/b_n(A_n(t)/n - \mu\rho_n t),\, t \ge 0\bigr)$ obeys an LDP in $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ for rate b_n^2 with deviation function $I^A(a)$ such that $I^A(a) = 1/(2\sigma^2) \int_0^\infty a'(t)^2\, dt$, provided $a = (a(t)\,, t \ge 0)$ is an absolutely continuous nondecreasing function with $a(0) = 0$, and $I^A(a) = \infty$, otherwise, where $\sigma > 0$. Then the sequence $(\sqrt{n}/b_n\, (Q_n(t)/n - 1)\,, t \ge 0)$ obeys an LDP in $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ for rate b_n^2 with deviation function

$$I^Q(z) = \frac{1}{2} \inf\{\int_0^1 v'(x)^2\, dx + \int_0^\infty w'(t)^2\, dt + \int_0^\infty \int_0^1 k'(t,x)^2\, dx\, dt\},$$

the inf being taken over $v = (v(x)\,, x \in [0,1]) \in \mathbb{D}([0,1], \mathbb{R})$, $w = (w(t)\,, t \ge 0) \in \mathbb{D}(\mathbb{R}_+, \mathbb{R})$ and $k = ((k(t,x)\,, x \in [0,1])\,, t \ge 0) \in \mathbb{D}(\mathbb{R}_+, \mathbb{D}([0,1], \mathbb{R}_+))$ such that $v(0) = v(1) = 0$, $w(0) = 0$, $k(0,x) = k(t,0) = k(t,1) = 0$, v, w and k are absolutely continuous with respect to Lebesgue measures on $[0,1]$, $\mathbb{R}_+$ and $\mathbb{R}_+ \times [0,1]$, respectively, and

$$\begin{aligned} z(t) = {} & (1 - F(t)) z_0^+ - (1 - F_0(t)) z_0^- + \int_0^t z(t-s)^+\, dF(s) - \beta F_0(t) + v(F_0(t)) \\ & + \int_0^t \bigl(1 - F(t-s)\bigr)\sigma\, w'(s)\, ds + \int_{\mathbb{R}_+^2} \mathbf{1}_{\{s+x\le t\}}\, k'(\mu s, F(x))\mu\, ds\, dF(x)\,, \end{aligned} \qquad (1)$$

where $v'(x)$, $w'(t)$ and $k'(t,x)$ represent the respective Radon–Nikodym derivatives and superscripts $^+$ and $^-$ signify the positive and negative parts of a real number, respectively. If v, w and k as indicated do not exist, then $I^Q(z) = \infty$.

A proof seems possible along the lines of the arguments in Puhalskii and Reed [8] by capitalizing on the analogy between weak convergence and LDP. In order that the LDP for $\bigl(\sqrt{n}/b_n(A_n(t)/n - \mu\rho_n t),\, t \ge 0\bigr)$ in the statement holds, it suffices that $E(nu_n) \to 1/\mu$, $\mathrm{Var}(nu_n) \to \sigma^2/\mu^3$ and that either $\sup_n E(nu_n)^{2+\epsilon} < \infty$, for some $\epsilon > 0$, and $\sqrt{\ln n}/b_n \to \infty$ or $\sup_n E \exp(\alpha(nu_n)^\delta) < \infty$ and $n^{\delta/2}/b_n^{2-\delta} \to \infty$, for some $\alpha > 0$ and $\delta \in (0,1]$, where u_n represents a generic interarrival time for the n-th queue, cf. Puhalskii [6].

The next conjecture concerns the stationary distribution of $Q_n(t)$. If $\beta > 0$, then $\rho_n < 1$, provided n is great enough, so that, in some generality, there exists a unique stationary distribution of $Q_n(t)$, see, e.g., Asmussen [2]. Also, the dynamical system associated with (1)

$$z(t) = (1 - F(t))z_0^+ - (1 - F_0(t))z_0^- + \int_0^t z(t-s)^+ \, dF(s) - \beta F_0(t)$$

has a unique equilibrium $z(t) = -\beta$. For $y \in \mathbb{R}$, let a quasipotential be defined as

$$V(y) = \lim_{t\to\infty} \inf\{I^Q(z) : z(0) = -\beta, z(t) = y\}.$$

Conjecture 2 Let the hypotheses of Conjecture 1 hold, $\beta > 0$ and the $Q_n(t)$ be stationary in t. Then the sequence of the distributions of $(Q_n(t) - n)/(b_n\sqrt{n})$ obeys an LDP in $\mathbb{R}$ for rate b_n^2 with deviation function $V(y)$.

It seems as though a proof can be obtained by applying the results in Puhalskii [7].

3 Closing remarks

Calculations for the $M/M/n$ queue, suggested by the referee, yield $V(y) = \beta y + (y^-)^2/2$. Comparison with Theorem 1 in Halfin and Whitt [4] shows that the moderate deviation scaling $(Q_n(t) - n)/(b_n\sqrt{n})$ under the "moderate" heavy-traffic condition $\sqrt{n}/b_n\,(1 - \rho_n) \to \beta$ captures the exponent in the distribution density that arises for the weak convergence scaling $(Q_n(t) - n)/\sqrt{n}$ under the heavy-traffic condition $\sqrt{n}\,(1 - \rho_n) \to \beta$. The same pattern emerges in the setup of the $GI/G/1$ queue, see Puhalskii [6]. Could the quasipotential be linked similarly to a stationary distribution of a Halfin–Whitt heavy-traffic limit for the $GI/G/n$ queue ?

References

1. Aghajani, R., Ramanan, K.: The limit of stationary distributions of many-server queues in the Halfin-Whitt regime. Math. Oper. Res. **45**(3), 1016–1055 (2020)
2. Asmussen, S.: Applied Probability and Queues. volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability
3. Gamarnik, D., Goldberg, D.: Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. Ann. Appl. Probab. **23**(6), 2382–2419 (2013)
4. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**(3), 567–588 (1981)
5. Kaspi, H., Ramanan, K.: SPDE limits of many-server queues. Ann. Appl. Probab. **23**(1), 145–229 (2013)
6. Puhalskii, A.: Moderate deviations for queues in critical loading. Queueing Syst. Theory Appl. **31**(3–4), 359–392 (1999)
7. Puhalskii, A.: Large deviation limits of invariant measures. arxiv preprint arxiv:2006.16456v2, (2021)
8. Puhalskii, A., Reed, J.: On many-server queues in heavy traffic. Ann. Appl. Probab. **20**(1), 129–195 (2010)
9. Whitt, W.: Stochastic-Process Limits. Springer Series in Operations Research. Springer-Verlag, New York, 2002. An introduction to stochastic-process limits and their application to queues

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.