informs.
http://pubsonline.informs.org/journal/stsy/

STOCHASTIC SYSTEMS
Vol. 8, No. 2, June 2018, pp. 143–165
ISSN 1946-5238 (print), ISSN 1946-5238 (online)

Heavy-Traffic Limit of the $GI/GI/1$ Stationary Departure Process and Its Variance Function

Ward Whitt,[a] **Wei You**[a]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027

Contact: ww2040@columbia.edu, http://orcid.org/0000-0003-4298-9964 (WW); wy2225@columbia.edu, http://orcid.org/0000-0003-0844-4194 (WY)

Received: August 23, 2017
Revised: December 11, 2017
Accepted: February 5, 2018
Published Online in Articles in Advance:
May 11, 2018

MSC2010 Subject Classification: Primary: 60F17; secondary: 60K25, 90B22

https://doi.org/10.1287/stsy.2018.0011

Abstract. Heavy-traffic limits are established for the stationary departure process from a $GI/GI/1$ queue and its variance function. The limit process is a function of the Brownian motion limits of the arrival and service processes plus the stationary reflected Brownian motion (RBM) limit of the queue-length process. An explicit expression is given for the variance function, which depends only on the first two moments of the interarrival times and service times plus the previously determined correlation function of canonical (drift -1, diffusion coefficient 1) RBM. The limit for the variance function here is used to show that the approximation for the index of dispersion for counts of the departure process used in our new robust queueing network analyzer is asymptotically correct in the heavy-traffic limit.

Funding: This research was supported by National Science Foundation [Grant CMMI 1634133].

Keywords: heavy traffic • departure processes • index of dispersion • queueing networks • robust queueing • stationary process

1. Introduction

In this paper we establish a *heavy-traffic* (HT) limit for the stationary departure process from the stable $GI/GI/1$ queue and its variance function. In doing so, we are primarily motivated by our desire to develop a new *robust queueing network analyzer* (RQNA) for open networks of single-server queues exploiting the *index of dispersion for counts* (IDC) for all arrival processes, which we refer to as the RQNA-IDC.

The HT limit here is also of considerable interest more generally, because the stationary departure process from a $GI/GI/1$ queue is remarkably complicated; for example, it is only a stationary renewal process in the special case of an $M/M/1$ model, when it is Poisson, by Burke's (1956) theorem (also see O'Connell and Yor 2001 and references therein). Indeed, explicit transform expressions for the variance function of the stationary departure process are evidently only available for the $M/GI/1$ and $GI/M/1$ models, due to Takacs (1962) and Daley (1975, 1976); see Bertsimas and Nakazato (1990) and Hu (1996) for related results on $GI/GI/1$. We exploit the $GI/M/1$ and $M/GI/1/$ results here to directly establish HT limits for the variance function in Sections 3 and 4.

The HT limit for the departure process starting empty in the $GI/GI/1$ model and more general multichannel models is an old result, being contained in Iglehart and Whitt (1970b, Theorem 2), but we evidently derive the first HT limits for the stationary departure process and its variance function for any $GI/GI/1$ model except $M/M/1$. The key to the process limit here is the recent HT limits for the stationary vector of queue lengths in a generalized Jackson network in Gamarnik and Zeevi (2006) and the extension in Budhiraja and Lee (2009). The existence of the HT limit for the scaled variance function is a relatively direct consequence.

In particular, Gamarnik and Zeevi (2006) created a Markov process representation by appending the residual interarrival times and service times to the queue length at each node as supplementary variables. Their key step is to identify an appropriate Lyapunov function in that formidable high-dimensional setting. With that approach, they establish HT limits for the sequence of steady-state distributions of the vector of queue lengths. For the relatively simple $GI/GI/1$ model that we consider, the HT limit in Gamarnik and Zeevi (2006) only recovers a variant of the very early direct HT limit for the steady-state waiting time distribution from the early paper by Kingman (1961) and its extension to general $G/G/1$ queues in Szczotka (1990). For $GI/GI/1$, we exploit the *HT process limit* for the entire sequence of steady-state queue-length processes that follows from the Markov process representation. That HT process limit for the stationary queue-length process is important to establish our associated HT process limit for the stationary departure process. (To quickly see that the process part of the HT limit is missing in Kingman 1961 and Szczotka 1990, note that the HT space scaling is used without any time scaling. The importance of both time and space scaling is essential for HT process limits, as discussed in Whitt 2002.) The extension by Budhiraja and Lee (2009) is important too, because they replace a strong moment-generating-function condition on the interarrival-time and service-time distributions in Gamarnik and Zeevi (2006) by the usual $2+\epsilon$ moment conditions, as in Iglehart and Whitt (1970b, Example 3.1).

An interesting and useful contribution here is the explicit form of the limiting variance function and its connection to basic functions of reflected Brownian motion (RBM) in Abate and Whitt (1987, 1988); see (65), which draws upon (24) and (28). For this step, we exploit the transform limits in the $M/GI/1$ and $GI/M/1$ special cases. We apply time-space transformations of the underlying Brownian motions in the $GI/GI/1$ limit to show that the limit with adjusted parameters is the same as for the $M/GI/1$ or $GI/M/1$ special case. Thus, we can identify the explicit form of the limiting variance function for the $GI/GI/1$ model by exploiting the results for the special cases; see the proof of Theorem 5.3. This same formula serves as an approximation more generally; see Section 7.

1.1. The New RQNA-IDC

We now explain a bit more about the new RQNA-IDC. It is a parametric-decomposition approximation (i.e., it treats the individual queues separately) like the *queueing network analyzer* (QNA) in Whitt (1983). Instead of approximately characterizing each flow by its rate and a single variability parameter, we approximately characterize the flow by its rate and the IDC, which is a real-valued function on the positive halfline, in particular the scaled variance function; see Section 6. The need for such an enhanced version of QNA has long been recognized, as can be seen from Kim (2011a, b), Whitt (1995), and Zhang et al. (2005).

This paper is a sequel to Whitt and You (2018a), which developed a *robust queueing* (RQ) approximation for the steady-state workload in a $G/G/1$ queue; see Section 6.1. That RQ algorithm extends an earlier RQ algorithm by Bandi et al. (2015). Indeed, a full RQNA is developed in Bandi et al. (2015), but the RQ approximation in Bandi et al. (2015) is a parametric RQ formulation, based on a single variability parameter, as in Whitt (1983). In contrast, the new RQ algorithm in Whitt and You (2018a) involves a functional formulation that incorporates the variance of the input over time. The advantage of the new formulation is demonstrated by simulation comparisons in Whitt and You (2018a).

A full RQNA-IDC for open networks of $G/GI/1$ single-server queues is outlined in Section 6 of Whitt and You (2018a). There, it is noted that the main challenge in developing an effective RQNA that can better capture dependence in the arrival processes at the different queues is to be able to approximate the IDC of the stationary departure process from a $G/GI/1$ queue. Important theoretical support for the new RQNA-IDC developed and studied in Whitt and You (2018c) is provided by the present paper. Corollary 6.1 shows that the proposed approximation for the IDC of the stationary departure process is asymptotically correct for the $GI/GI/1$ queue in the HT limit.

1.2. On the Heavy-Traffic Scaling

We consider the standard HT scaling of both time and space (discussed in Iglehart and Whitt 1970a, b; Whitt 2002, Chapters 5 and 9) applied to the stationary departure process. That scaling involves the HT scaling *after* the usual time limit to obtain the stationary departure process.

To explain, let $D_{\rho,k}(t)$ denote the number of departures in the interval $[0,t]$ for the $GI/GI/1$ model with traffic intensity $\rho<1$, starting with k customers in the system at time 0 and fully specified remaining times until the first new arrival and the service completion for the customer in service. Then, for each $s>0$, let $D_{\rho,k,s}(t) \equiv D_{\rho,k}(s+t) - D_{\rho,k}(s)$: $t \ge 0$ be that same departure process after time s. Under general regularity conditions (which we do not discuss), for each k, $0 \le k < \infty$, and ρ, $0<\rho<1$, there is convergence in distribution

$$\{D_{\rho,k,s}(t) \colon t \ge 0\} \Rightarrow \{D_\rho(t) \colon t \ge 0\} \quad \text{in } \mathscr{D} \text{ as } s \to \infty, \tag{1}$$

where $\Rightarrow$ denotes convergence in distribution, $\mathscr{D}$ is the usual function space of right-continuous real-valued functions on $[0,\infty)$ with left limits and the J_1 topology as in Whitt (2002), and $D_\rho \equiv \{D_\rho(t) \colon t \ge 0\}$ is a stationary point process. Indeed, this stationary point process D_ρ is the topic of the present paper. In general, the stationary departure process may be well defined even if the limit in (1) does not exist (e.g., because the interarrival-time distribution is deterministic, creating periodic structure).

In this paper we consider the usual HT scaling of space and time applied to the stationary process; in other words, we consider the scaled stationary departure process defined by

$$D^*_\rho(t) \equiv (1-\rho)\big(D_\rho((1-\rho)^{-2}t) - \lambda(1-\rho)^{-2}t\big), \quad t \ge 0, \tag{2}$$

and we let $\rho \uparrow 1$. Observe that convergence $D^*_\rho \Rightarrow D^*$ in $\mathscr{D}$, which we establish in Section 5, is equivalent to the *iterated limit*

$$\lim_{\rho\uparrow 1}\lim_{s\to\infty}\big\{(1-\rho)[D_{\rho,k,s}((1-\rho)^{-2}t) - \lambda(1-\rho)^{-2}t]\colon t \ge 0\big\}, \tag{3}$$

in other words, we first let $s \uparrow \infty$ and then afterwards we let $\rho \uparrow 1$ with the scaling.

1.3. The Variance Function and the BRAVO Effect

Because of our interest in the IDC, we are especially interested in the associated variance function of the stationary departure process, defined in the obvious way by $V_{d,\rho}(t) \equiv \mathrm{Var}(D_\rho(t))$, $t \ge 0$. As a consequence of the HT stochastic-process limit $D^*_\rho \Rightarrow D^*$ and appropriate uniform integrability, we will obtain the associated limit for the appropriately scaled variance functions, i.e.,

$$V^*_{d,\rho}(t) \equiv \mathrm{Var}(D^*_\rho(t)) = (1-\rho)^2 V_{d,\rho}((1-\rho)^{-2}t) \to V^*_d(t) \quad \text{as } \rho \uparrow 1, \tag{4}$$

where $V^*_d(t) \equiv \mathrm{Var}(D^*(t))$, $t \ge 0$.

There is a recent related result for the $M/GI/1$ queue in Hautphenne et al. (2015, Proposition 6), as part of an investigation into the BRAVO (balancing reduces asymptotic variance of outputs) effect; see Nazarathy (2011), Hanbali et al. (2011), the earlier in Berger and Whitt (1992, Theorem 4.1), and Williams (1992). For the stationary $M/GI/1$ model, a two-term asymptotic expansion is developed for the variance function $V_{d,\rho}(t) \equiv \mathrm{Var}(D_\rho(t))$ as $t \to \infty$ for fixed $\rho < 1$. There is no direct heavy-traffic scaling, but the scaling emerges in the parameters. We discuss the connections between our result and this earlier result in Remark 5.3.

1.4. Organization

Here is how this paper is organized: We start in Section 2 by providing a brief review of stationary point processes, focusing especially on the variance function. In Section 3 we use Laplace transforms (LT's) of the stationary departure process in the $GI/M/1$ queue derived by Daley (1975, 1976) to derive the HT limit of its variance function. In Section 4 we use the HT limit for the Palm version of the mean function derived by Takacs (1962) to derive the HT limit of the stationary variance function. (In Section 2.2 we review the application of the Palm-Khintchine equation to express the stationary variance in terms of the Palm mean function.) In Section 5 we establish the HT limit for the stationary departure process in the $GI/GI/1$ queue (Theorem 5.2) and its variance function (Theorem 5.3). In Section 6 we provide a brief overview of the application of Theorem 5.3 to support our RQNA-IDC developed in Whitt and You (2018c), which is briefly outlined in Whitt and You (2018a, Section 6). We discuss extensions in Section 7. Finally, we present postponed proofs in Section 8.

2. Review of Stationary Point Processes

In this section we review basic properties of stationary point processes; see Daley and Vere-Jones (2008) and Sigman (1995) for more background. In Section 2.1 we review renewal processes and their Laplace transforms. In Section 2.2 we review the Palm-Khintchine equation and use it to express the variance function of a stationary point process in terms of the mean function of the Palm version.

2.1. Renewal Processes and the Laplace Transform

We start with a rate-λ renewal process $N \equiv \{N(t)\colon t \ge 0\}$. Let F be the *cumulative distribution function* (cdf) of the interval U between points (the interarrival time in a GI arrival process), having mean $E[U] = \lambda^{-1}$ and finite second moment. As a regularity condition for our queueing application, we also assume that F has a *probability density function* (pdf) f, where $F(t) = \int_0^t f(u)\,du$, $t \ge 0$. Throughout this paper, we assume that the interarrival-time distribution of our renewal arrival processes has a pdf. That pdf assumption ensures that the equilibrium renewal process arises as the time limit of the ordinary renewal process (e.g., see Ross 1996, Sections 3.4 and 3.5).

The stationary or equilibrium renewal process differs from the ordinary renewal process only by the distribution of the first interarrival times. Let F_e be the cdf of the equilibrium distribution, which has pdf $f_e(t) = \lambda(1 - F(t))$. Let $E^e[\cdot]$ denote the expectation under the stationary distribution (with first interval distributed according to F_e) and let $E^0[\cdot]$ denote the expectation under *the Palm distribution* (with first interval distributed as F).

Conditioning on the first arrival, distributed as F under the Palm distribution or as F_e under stationary distribution, the renewal equations for the mean and second moment of $N(t)$, the number of points in an interval $[0,t]$, are:

$$\begin{aligned} m(t) &\equiv E^0[N(t)] = F(t) + \int_0^t m(t-s)\,dF(s),\\ m_e(t) &\equiv E^e[N(t)] = F_e(t) + \int_0^t m(t-s)\,dF_e(s),\\ \sigma(t) &\equiv E^0[N^2(t)] = F(t) + 2\int_0^t m(t-s)\,dF(s) + \int_0^t \sigma(t-s)\,dF(x),\\ \sigma_e(t) &\equiv E^e[N^2(t)] = F_e(t) + 2\int_0^t m(t-s)\,dF_e(s) + \int_0^t \sigma(t-s)\,dF_e(x). \end{aligned}$$

Throughout the paper, we use the Laplace Transform (LT) instead of the Laplace-Stieltjes Transform (LST). The LT of $f(t)$ and the LST of F, denoted by $\mathscr{L}(f)(s) \equiv \hat{f}(s)$, are

$$\hat{f}(s) \equiv \mathscr{L}(f)(s) \equiv \int_0^\infty e^{-st} f(t)\,dt = \int_0^\infty e^{-st}\,dF(t), \tag{5}$$

so that $f(t) = \mathscr{L}^{-1}(\hat{f})(t)$. Throughout the paper, we add a hat to either an LT or an item that appears in LT. The LT of f_e is then

$$\hat{f}_e(s) = \frac{\lambda(1-\hat{f}(s))}{s} \qquad \text{and} \qquad \hat{F}_e(s) = \frac{\hat{f}_e(s)}{s},$$

where $\lambda^{-1} \equiv \int_0^\infty t f(t)\,dt$ is the mean. Applying the LT to the renewal equations, we obtain

$$\hat{m}(s) = \frac{\hat{f}(s)}{s(1-\hat{f}(s))}, \tag{6}$$

$$\hat{m}_e(s) = \frac{\hat{f}_e(s)}{s(1-\hat{f}(s))} = \frac{\lambda}{s^2}, \tag{7}$$

$$\hat{\sigma}(s) = \frac{\hat{f}(s) + 2s\hat{m}(s)\hat{f}(s)}{s(1-\hat{f}(s))} = \frac{\hat{f}(s)(1+\hat{f}(s))}{s(1-\hat{f}(s))^2}, \tag{8}$$

$$\hat{\sigma}_e(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) = \frac{\lambda(1+\hat{f}(s))}{s^2(1-\hat{f}(s))}. \tag{9}$$

From (7), we see that

$$E^e[N(t)] = \lambda t, \quad t \ge 0, \tag{10}$$

as must be true for any stationary point process.

Let $V(t) \equiv \mathrm{Var}^e(N(t))$ be the variance process of $N(t)$ under time-stationary distribution. (We omit the e superscript on $V(t)$ because we will only discuss stationary variance functions.) Combining (9) and (10), we have

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) - \frac{2\lambda^2}{s^3} = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\frac{\hat{f}(s)}{s(1-\hat{f}(s))} - \frac{2\lambda^2}{s^3}. \tag{11}$$

The variance function then can be obtained from the numerical inversion of the Laplace transform (e.g., see Abate and Whitt 1992, 1995, Section 13; Whitt and You 2018b). Term by term inversion shows that we can express $V(t)$ in terms of the renewal function $m(t)$

$$V(t) = \lambda \int_0^t (1 + 2m(u) - 2\lambda u)\,du. \tag{12}$$

In Section 2.2 we show that the Palm-Khintchine equation can be used to derive a generalization of (12) for general stationary and ergodic point processes.

2.2. The Palm-Khintchine Equation

We now consider a continuous-time stationary point process (i.e., having stationary increments). The main idea is the Palm transformation relating continuous-time stationary processes to the associated discrete-time stationary processes. An important manifestation of that relation is the Palm-Khintchine equation; see Daley and Vere-Jones (2008, Theorem 3.4.II). It is important here because it can be applied to generalize the variance formula discussed in Section 2.1; see Daley (1971, Section 2.4) and Daley and Vere-Jones (2008, Section 3.4).

We focus on orderly stationary ergodic point processes with finite intensity. (Orderly means that the points occur one at a time.) Let $N(s,t]$ denote the number of events in interval $(s,t]$, and $N(t) \equiv N(0,t]$.

Theorem 2.1 (Palm-Khintchine Equation). *For an orderly stationary point process of finite intensity λ such that $P^e(N(-\infty,0] = N(0,\infty) = \infty) = 1$, then*

$$P^e(N(t) \le k) = 1 - \lambda \int_0^t q_k(u)\,du = \lambda \int_t^\infty q_k(u)\,du, \quad \textit{for } k = 0,1,2,\ldots, \tag{13}$$

where $q_k(t)$ is the probability of exactly k arrivals in $(0,t]$ under the Palm distribution, i.e.,

$$q_k(t) = \lim_{h \downarrow 0} P(N(t) = k \mid N(-h,0] > 0). \tag{14}$$

Under ergodicity, the Palm distribution is equivalent to the event stationary distribution, so that $q_k(t) = P^0(N(t) = k)$.

We now apply Theorem 2.1 to generalize (12) and (11) to the case of orderly stationary ergodic point process.

Corollary 2.1 (Variance of a Stationary Ergodic Point Process). *For a general stationary ergodic point process with rate λ and finite second moment, the variance function is*

$$V(t) = \lambda \int_0^t (1 + 2m(u) - 2\lambda u)\,du, \quad t \ge 0, \tag{15}$$

where

$$m(t) \equiv E^0[N(t)] = \sum_{k=1}^{\infty} k q_k(t), \quad t \ge 0, \tag{16}$$

and its LT is

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) - \frac{2\lambda^2}{s^3}, \tag{17}$$

where $\hat{m}(s)$ is the LT of $m(t)$.

Proof. Let

$$p_k(t) = P^e(N(t) = k), \quad \text{for } k = 0,1,2,\ldots \tag{18}$$

so that $\sum_{i=1}^k p_i(t) = P^e(N(t) \le k)$. With Theorem 2.1, we can write

$$\begin{aligned} V(t) = \sum_{k=1}^{\infty} k^2 p_k(t) - \lambda^2 t^2 &= \sum_{k=1}^{\infty} k^2 \lambda \int_0^t (q_{k-1}(u) - q_k(u))\,du - \lambda^2 t^2 \\ &= \lambda \int_0^t (1 + 2m(u) - 2\lambda u)\,du, \end{aligned} \tag{19}$$

where $m(t) \equiv E^0[N(t)]$ as in (16). Taking the Laplace Transform, we obtain

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) - \frac{2\lambda^2}{s^3}. \quad \square$$

3. The Departure Variance in the $GI/M/1$ Queue

We now start considering the queueing models. In particular, we focus on the $GI/GI/1$ queue, which has unlimited waiting space, the first-come first-served service discipline and independent sequences of i.i.d. interarrival times and service times distributed as random variables U and V, respectively, where U has a pdf $f(t)$. Let $\lambda \equiv 1/E[U]$ be the arrival rate; let $\hat{f}(s) \equiv E[e^{-sU}]$ be the LT of the interarrival-time pdf $f(t)$; let $\mu \equiv 1/E[V]$ be the service rate; and let $\rho \equiv \lambda/\mu$ be the traffic intensity, assuming that $\rho < 1$.

Daley (1975, 1976) derived the LST of the variance $V_d(t)$ of the stationary departure process in a $GI/M/1$ queue. The associated LT of $V_d(t)$ is

$$\hat{V}_d(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s^3}\left(\mu\delta - \lambda + \frac{\mu^2(1-\delta)(1-\hat{\xi}(s))(\mu\delta(1-\hat{f}(s)) - s\hat{f}(s))}{(s+\mu(1-\hat{\xi}(s)))(s-\mu(1-\delta))(1-\hat{f}(s))}\right) \tag{20}$$

where $\hat{\xi}(s)$ is the root with the smallest absolute value in z of the equation

$$z = \hat{f}(s+\mu(1-z)) \tag{21}$$

and $\delta = \hat{\xi}(0)$ is the unique root in $(0,1)$ of the equation

$$\delta = \hat{f}(\mu(1-\delta)), \tag{22}$$

which appears in the distribution of the stationary queue length in a $GI/M/1$ queue. Useful properties of $\hat{\xi}(s)$ and $\delta = \hat{\xi}(0)$ are contained in Lemma 8.1.

We now establish an HT limit for the departure variance function in the $GI/M/1$ model. To do so, we consider a family of $GI/M/1$ models parameterized by ρ, where $\lambda \equiv 1/E[U]$ and $\mu = \mu_\rho = 1/E[V] \equiv \lambda(1+(1-\rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho\uparrow 1}\gamma_\rho = \gamma > 0$. Note that if $\gamma_\rho = 1/\rho$, then we come to the usual case of $\lambda/\mu = \rho$. We allow this general scaling so that we can gain insight into reflected Brownian motion (RBM) with nonunit drift. Let the HT-scaled variance function be

$$V^*_{d,\rho}(t) \equiv (1-\rho)^2 V_{d,\rho}((1-\rho)^{-2}t), \quad t \geq 0, \tag{23}$$

just as in (4). Throughout the paper, we use the asterisk (*) superscript with ρ subscript to denote HT-scaled items in the queueing model, as in (23), and the asterisk without the ρ subscript to denote the associated HT limit.

As should be expected from established HT limits, e.g., as in Whitt (2002, Section 5.7, Chapter 9), the HT limit of the variance function $V^*_{d,\rho}(t)$ in (23) depends on properties of the normal distribution and RBM. Let $\phi(x)$ be the pdf and $\Phi(x)$ the cdf of the standard normal variable $N(0,1)$. Let $\Phi^c(x) \equiv 1-\Phi(x)$ be the complementary cdf (ccdf). Let $R(t)$ be canonical RBM (having drift -1, diffusion coefficient 1) and let $R_e(t)$ be the stationary version, which has the exponential marginal distribution for each t with mean $1/2$. The correlation function $c^*(t)$ of R_e is defined as

$$\begin{aligned} c^*(t) &\equiv \mathrm{cov}(R_e(0), R_e(t)) = 2(1-2t-t^2)\Phi^c(\sqrt{t}) + 2\sqrt{t}\phi(\sqrt{t})(1+t) = 1 - H^*_2(t) \\ &\equiv 1 - \frac{E[R(t)^2 \mid R(0)=0]}{E[R(\infty)^2]} = 1 - 2E[R(t)^2 \mid R(0)=0], \quad t \geq 0, \end{aligned} \tag{24}$$

where $H^*_2(t)$ is the second-moment cdf of canonical RBM in Abate and Whitt (1987), which has mean 1 and variance 2.5; see Abate and Whitt (1987, Corollaries 1.1.1 and 1.3.4; 1988, Corollary 1). The correlation function $c^*(t)$ has LT

$$\hat{c}^*(s) \equiv \frac{1}{s} - \frac{2}{s^2}\left(1 - \frac{\sqrt{1+2s}-1}{s}\right); \tag{25}$$

see Abate and Whitt (1987, (1.10)). Equivalently, the Gaussian terms in (24) can be re-expressed as $\phi(\sqrt{t}) = e^{-t/2}/\sqrt{2\pi}$ and $\Phi^c(\sqrt{t}) = (1-\mathrm{erf}(\sqrt{t/2}))/2$, where erf is the error function. This LT can be explicitly inverted, yielding

$$c^*(t) = -2(t^2+2t-1)\Phi^c(\sqrt{t}) + 2\phi(\sqrt{t})\sqrt{t}(1+t). \tag{26}$$

By Abate and Whitt (1987, Corollary 1.3.5), the correlation function has tail asymptotics according to

$$c^*(t) = 1 - H^*_2(t) \sim \frac{16}{\sqrt{2\pi t^3}} e^{-(t/2)}, \quad \text{as } t \to \infty. \tag{27}$$

From the correlation function, we define

$$w^*(t) \equiv 1 - \frac{1-c^*(t)}{2t}, \quad t \geq 0. \tag{28}$$

It then follows from the explicit expression of $c^*(t)$ that

$$w^*(t) = \frac{1}{2t}\left((t^2+2t-1)(1-2\Phi^c(\sqrt{t})) + 2\phi(\sqrt{t})\sqrt{t}(1+t) - t^2\right). \tag{29}$$

One may verify that $w^*(t)$ is a increasing function satisfying $0 \le w^*(t) \le 1$. As we shall see in Theorem 3.1, this $w^*(t)$ serves as the weight function that appears in the limiting departure variance function.

We now present the main result for the departure variance in the $GI/M/1$ special case. The idea of the proof is to exploit the explicit form of the LT $\hat{V}^*_{d,\rho}(t)$ of the scaled stationary departure variance and derive its HT limit. We then obtain the convergence of the HT-scaled variance function $V^*_{d,\rho}(t)$ by applying continuity theorem of the LT, see Feller (1971, Chapter XIII, Theorem 2(a)). The proof of the theorem can be found in Section 8.1.

Theorem 3.1 (HT Limit for the $GI/M/1$ Departure Variance). *Consider the $GI/M/1$ model with $1/E[U] = \lambda$ and $1/E[V] = \mu_\rho \equiv \lambda(1+(1-\rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho\uparrow 1}\gamma_\rho = \gamma > 0$. Assume that $E[U^3] < \infty$ so that a two-term Taylor series expansion of the LT $\hat{f}(s)$ about the origin is valid with asymptotically negligible remainder. Then the HT-scaled variance function $V^*_{d,\rho}(t)$ defined in (23) converges as $\rho\uparrow 1$; i.e.,*

$$V^*_{d,\rho}(t) \to V^*_d(t), \quad \textit{as } \rho\uparrow 1, \ \textit{for all } t \ge 0, \tag{30}$$

where $V^*_d(t)$ is a continuous nonnegative real-valued function with LT

$$\hat{V}^*_d(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s^2}\frac{c_a^2-1}{c_a^2+1}\frac{\gamma}{\hat{\xi}^*(s)} = \frac{\lambda c_a^2}{s^2} - \frac{\lambda(c_a^2-1)}{s^2}\left(1 - \frac{2}{c_a^2+1}\frac{\gamma}{\hat{\xi}^*(s)}\right), \tag{31}$$

with $\hat{\xi}^*(s)$ being the unique root with nonnegative real part of the quadratic equation

$$\left(\frac{c_a^2+1}{2}\right)\hat{\xi}^*(s)^2 - \gamma\hat{\xi}^*(s) - \frac{s}{\lambda} = 0. \tag{32}$$

In addition,

$$V^*_d(t) = w^*(\lambda\gamma^2 t/c_x^2)c_a^2\lambda t + (1 - w^*(\lambda\gamma^2 t/c_x^2))c_s^2\lambda t \tag{33}$$

for $w^*(t)$ defined in (28), $c_x^2 \equiv c_a^2 + c_s^2$, $c_a^2 \equiv \mathrm{Var}(U)/E[U]^2$ and $c_s^2 = 1$.

We shall want to relate our HT limit for the departure variance function to associated HT limits for the variance functions of the arrival and service processes. For that step, it is significant that the functional central limit theorem (FCLT) for any stationary point process has the same form as the FCLT for the associated Palm process, as was shown by Nieuwenhuis (1989). Extra uniform integrability is required to get the associated limit for the variance function. To be relatively self-contained, we will directly derive the desired result from the transform of the equilibrium renewal process in (11).

For that purpose, let $N_a(t)$ denote the arrival renewal process and let $V_a(t) \equiv \mathrm{Var}^e(N_a(t))$ denote its variance process under the stationary distribution. Similarly, we define $N_s(t)$ and $V_s(t)$ for the service renewal process. The following lemma states that the terms $c_a^2\lambda t$ and $c_s^2\lambda t$ in (33) can be interpreted as the limiting variance function of the arrival and service renewal processes, respectively. This implies that the limiting departure variance function V^*_d is a convex combination of the arrival and service variance functions with a scaled version of the time-varying weight function $w^*(t)$. This convex combination result is consistent with the more elementary approximation used in QNA; see Whitt (1983, 1984); there the departure variability parameter is approximated by a convex combination of the arrival and service variability parameters.

Lemma 3.1 (Limiting Variance Function of Stationary Renewal Processes). *Let $N(t)$ be a renewal process with rate λ and let c_N^2 be the scv of the inter-renewal distribution. Consider the HT-scaled stationary variance function*

$$V^*_{N,\rho}(t) \equiv \mathrm{Var}^e\left((1-\rho)N((1-\rho)^{-2}t)\right),$$

then

$$V^*_{N,\rho}(t) \to V^*_N(t) \equiv \lambda c_N^2 t, \quad \text{as } \rho\uparrow 1.$$

Proof. Let f denote the inter-renewal distribution. Recall the expression for the LT of a stationary renewal process in (11), we have

$$\begin{aligned}\hat{V}_N^*(s) &= \lim_{\rho\uparrow 1} \hat{V}^*_{N,\rho}(s) \equiv \lim_{\rho\uparrow 1} \mathscr{L}((1-\rho)^2 V_{N,\rho}((1-\rho)^{-2}t)) = \lim_{\rho\uparrow 1}(1-\rho)^4 \hat{V}_{N,\rho}((1-\rho)^2 t)\\ &= \lim_{\rho\uparrow 1}\left(\frac{\lambda}{s^2} + \frac{2\lambda}{s^2}\frac{\hat{f}((1-\rho)^2 t)}{1-\hat{f}((1-\rho)^2 t)} - \frac{2\lambda^2}{(1-\rho)^2 s^3}\right)\\ &= \frac{\lambda}{s^2} + \frac{2\lambda^2}{s^2}\lim_{\rho\uparrow 1}\frac{1}{(1-\rho)^2 s}\left(\frac{\hat{f}((1-\rho)^2 t)}{\lambda((1-\hat{f}((1-\rho)^2 t))/((1-\rho)^2 s))} - 1\right) = \frac{\lambda c_N^2}{s^2}.\end{aligned}$$

The result follows from inverting the LT, i.e., $V_s^*(t) = \lambda c_N^2 t$. □

To derive a pre-limit approximation, define the weight function

$$w_\rho(t) \equiv \frac{V_{d,\rho}(t) - V_s(t)}{V_a(t) - V_s(t)}, \tag{34}$$

where $V_a(t)$ and $V_s(t)$ are the variance functions associated with the equilibrium arrival and service renewal processes, *both* with rate λ. Define the HT-scaled weight function

$$w_\rho^*(t) = w_\rho((1-\rho)^{-2}t). \tag{35}$$

Combining Theorem 3.1 and Lemma 3.1, we obtain

Corollary 3.1 (Limiting Weight Function)**.** *Under the assumptions in Theorem 3.1, we have*

$$w_\rho^*(t) \Rightarrow w^*(\lambda\gamma^2 t/c_x^2)$$

for w^* defined in (28).

This justifies the following approximation for the variance function of the stationary departure process from a $GI/M/1$ queue:

$$V_{d,\rho}(t) = w_\rho(t)V_a(t) + (1-w_\rho(t))V_s(t) \approx w^*((1-\rho)^2\lambda\gamma^2 t/c_x^2)V_a(t) + (1 - w^*((1-\rho)^2\lambda\gamma^2 t/c_x^2))V_s(t). \tag{36}$$

We conclude this section with the tail asymptotic behavior of the variance function. To start, we rewrite V_d^* in terms of c^* and H_2^*,

$$\begin{aligned}V_d^*(t) &= c_a^2\lambda t + (1 - w^*(\lambda\gamma^2 t/c_x^2))(c_s^2 - c_a^2)\lambda t = c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2}H_2^*(\lambda\gamma^2 t/c_x^2)\\ &= c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2}c^*(\lambda\gamma^2 t/c_x^2), \quad t \ge 0.\end{aligned} \tag{37}$$

Combining (27) and (37), we obtain the asymptotic behavior of the departure variance function.

Corollary 3.2 (Asymptotic Behavior of the Departure Variance Function)**.** *Under the assumptions in Theorem 3.1,*

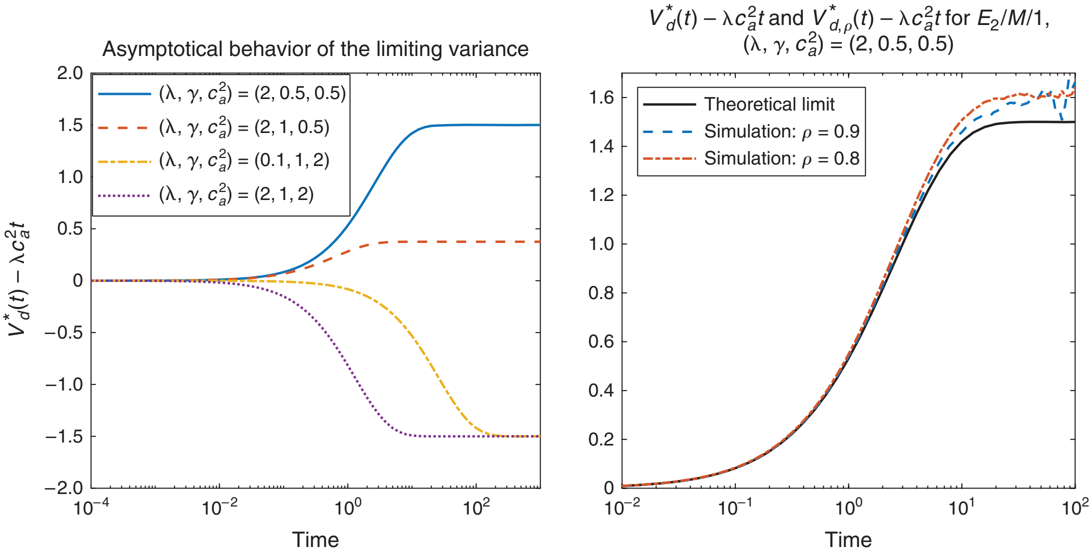
$$\begin{aligned}V_d^*(t) &= c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2}c^*(\lambda\gamma^2 t/c_x^2)\\ &\sim c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{8(c_s^2 - c_a^2)c_x^5}{\gamma^5}\frac{1}{\sqrt{2\pi\lambda^3 t^3}}e^{-(\lambda\gamma^2 t)/(2c_x^2)}, \quad \textit{as } t \to \infty,\end{aligned} \tag{38}$$

where $c_s^2 = 1$.

Example 3.1 (Numerical Experiments)**.** We now evaluate the approximation stemming from Theorem 3.1. First, we use numerical transform inversion as in Abate and Whitt (1992, 1995) to calculate the limiting variance function.

Figure 1 (left) reports $V_d^*(t) - \lambda c_a^2 t$ for four sets of parameters such that the limiting constant $(c_s^2 - c_a^2)c_x^2/2\gamma^2 = (1 - c_a^4)/2\gamma^2$ in Corollary 3.2 will be 1.5, 0.375, −1.5, and −1.5, respectively. Figure 1 (right) confirms Theorem 3.1 by comparing simulation estimates of the HT-scaled and centered departure variance function $V_{d,\rho}^*(t) - \lambda c_a^2 t$ for $\rho = 0.8$ and 0.9 from simulation with the theoretical limit $V_d^*(t) - \lambda c_a^2 t$ for the $E_2/M/1$ model with $\lambda = 2$, $\gamma = 0.5$ and $c_a^2 = 0.5$, showing that the theoretical limit in (31) serves as a good approximation of the HT-scaled variance function.

Figure 1. On the left is $V_d^*(t) - \lambda c_a^2 t$ for four sets of parameters calculated from numerically inverting (31). On the right is the the HT-scaled and centered variance $V_{d,\rho}^*(t) - \lambda c_a^2 t$ for $\rho = 0.8$ and 0.9 in the $E_2/M/1$ model with $\lambda = 2$, estimated by simulation, compared with the theoretical limit $V_d^*(t) - \lambda c_a^2 t$.



4. The Departure Variance in the $M/GI/1$ Queue

We now prove that the HT limit for the stationary departure variance in (33) also holds true for the $M/GI/1$ model. Of course, here we restrict our attention to $c_a^2 = 1$ instead of $c_s^2 = 1$ before. Theorem 5.3 will show that the same formula is valid for $GI/GI/1$ with general c_a^2 and c_s^2.

Recall from (17) that the Laplace Transform of the variance function of a general stationary and ergordic point process is

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) - \frac{2\lambda^2}{s^3}.$$

In the case of the $M/GI/1$ model, Takacs (1962, p. 78) derived an expression for $\hat{m}_d(s)$.

Theorem 4.1 (Laplace Transform of the Palm Mean Function). *For the departure process from a $M/GI/1$ queue,*

$$\hat{m}_d(s) \equiv \int_0^\infty e^{-st} m_d(t)\, dt = \frac{\hat{g}(s)}{s(1-\hat{g}(s))}\left(1 - \frac{s\Pi(\hat{v}(s))}{s + \lambda(1-\hat{v}(s))}\right), \tag{39}$$

where $\hat{g}(s) = E[e^{-sV}]$ is the Laplace Transform of the service time pdf $g(t)$, $\hat{v}(s)$ is the root with the smallest absolute value in z of the equation

$$z = \hat{g}(s + \lambda(1-z)) \tag{40}$$

and

$$\Pi(z) \equiv E[z^Q] = \frac{(1-\lambda/\mu)(1-z)\hat{g}(\lambda(1-z))}{\hat{g}(\lambda(1-z)) - z} \tag{41}$$

is the probability generating function of the distribution of the stationary queue length Q.

Note from (6) that the first part in (39), i.e.,

$$\frac{\hat{g}(s)}{s(1-\hat{g}(s))},$$

is exactly the Laplace Transform of the mean process of the service renewal process.

Now, we state the HT limit in terms of the HT-scaled variance function defined in (23) for the $M/GI/1$ special case. The result parallels that for the $GI/M/1$ case. The proof can be found in Section 8.2.

Theorem 4.2 (HT Limit for the $M/GI/1$ Departure Variance). *Consider an $M/GI/1$ model with $1/E[V] = \mu$ and $1/E[U] = \lambda_\rho \equiv \mu(1-(1-\rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho\uparrow 1}\gamma_\rho = \gamma > 0$. Assume that $E[V^3] < \infty$ so* *that a two-term Taylor series expansion of the LT $\hat{g}(s)$ about the origin is valid with asymptotically negligible remainder. Then the HT-scaled variance function $V^*_{d,\rho}(t)$ defined in (23) converges as $\rho \uparrow 1$, i.e.,*

$$V^*_{d,\rho}(t) \to V^*_d(t), \quad \text{as } \rho \uparrow 1, \text{ for all } t \geq 0, \tag{42}$$

where the limit $V^*_d(t)$ is a continuous nonnegative function with LT

$$\hat{V}^*_d(s) = \frac{\mu c_s^2}{s^2} + \frac{\gamma \mu^2 (1 - c_s^2)}{s^3} \hat{v}^*(s), \tag{43}$$

with $\hat{v}^*(s)$ being the unique root with positive real part of the equation

$$\frac{1 + c_s^2}{2} (\hat{v}^*(s))^2 + \gamma \hat{v}^*(s) - \frac{s}{\mu} = 0. \tag{44}$$

In addition,

$$V^*_d(t) \equiv w^*(\mu\gamma^2 t/c_x^2) c_a^2 \mu t + (1 - w^*(\mu\gamma^2 t/c_x^2)) c_s^2 \mu t, \tag{45}$$

where $w^*(t)$ in (28), $c_x^2 \equiv c_a^2 + c_s^2$, $c_a^2 = 1$ and $c_s^2 \equiv \mathrm{Var}(V)/E[V]^2$.

With the same technique as in Corollary 3.2, one can prove the following corollary, which yields exactly the same asymptotic behavior.

Corollary 4.1 (Asymptotic Behavior of the Departure Variance Curve)**.** *Under the assumptions in Theorem 4.2, we have the limit in (38), except now $c_a^2 = 1$ and c_s^2 is general.*

Hautphenne et al. (2015, Proposition 6) developed a two-term asymptotic expansion for the variance function $V_{d,\rho}(t) \equiv \mathrm{Var}(D_\rho(t))$ as $t \to \infty$ for for the $M/GI/1$ queue and fixed $\rho < 1$. We discuss the connections between our result and this earlier result in Remark 5.3.

5. Heavy-Traffic Limit for the Stationary Departure Process

In this section, we establish an HT limit for the stationary departure process and its variance function in a $GI/GI/1$ queue. To do so, we apply the recent HT results for the stationary queue length (number in system) in Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009) together with the HT limits for the general single-server queue in Whitt (2002, Section 9.3) and the general reflection mapping with nonzero initial conditions in Whitt (2002, Section 13.5). As in Whitt (2002), a major component of the proof is the continuous mapping theorem.

The corresponding limit starting out empty is contained in Iglehart and Whitt (1970b, Theorem 2). There has since been a substantial literature on that case; see Hanbali et al. (2011), Karpelevich and Kreinin (1994), and Whitt (2002). As can be seen from Whitt (2002, Sections 9.3 and 13.5), for the queue length, the key map is the reflection map ψ applied to a potential net-input function x,

$$\psi(x)(t) \equiv x(t) - \zeta(x)(t), \quad t \geq 0, \tag{46}$$

where

$$\zeta(x) \equiv \inf\{x(s) \colon 0 \leq s \leq t\} \wedge 0, \quad t \geq 0 \tag{47}$$

with $a \wedge b \equiv \min\{a, b\}$, so that $\zeta(x) \leq 0$ and $\psi(x)(t) \geq x(t)$ for all $t \geq 0$. The key point is that we now allow $x(0) \neq 0$.

5.1. A General Heavy-Traffic Limit for the $G/G/1$ Model

For the general $G/G/1$ single-server queue with unlimited waiting space and service provided in the order of arrival, we consider a family of processes indexed by the traffic intensity ρ, where $\rho \uparrow 1$. Let $Q_\rho(t)$ be the number of customers in the system at time t; let $A_\rho(t)$ count the number of arrivals in the interval $[0, t]$, which we assume to have rate λ; let $S_\rho(t)$ be a corresponding counting process for the successive service times, applied after time 0, to be applied to the initial $Q_\rho(0)$ customers and to all new arrivals; let $B_\rho(t)$ be the cumulative time that the server is busy in the interval $[0, t]$. Then the queue-length process can be expressed as

$$Q_\rho(t) \equiv Q_\rho(0) + A_\rho(t) - S_\rho(B_\rho(t)), \quad t \geq 0, \tag{48}$$

where the three components are typically dependent. (For simplicity, we assume that $A_\rho(0) = S_\rho(0) = B_\rho(0) = 0$ w.p.1.)

We have in mind that the system is starting in steady-state. Thus the triple $(Q_\rho(0), A_\rho(\cdot), S_\rho(\cdot))$ is in general quite complicated for each ρ. Even in the relatively tractable $GI/GI/1$ cases, which we shall primarily treat, the residual interarrival time and service time at time 0 will be complicated, depending on ρ and $Q_\rho(0)$. We will need to make assumptions ensuring that these are uniformly asymptotically negligible in the HT limit.

By flow conservation, the departure (counting) process can be represented as

$$D_\rho(t) \equiv A_\rho(t) - Q_\rho(t) + Q_\rho(0), \quad t \ge 0. \tag{49}$$

Directly, or by combining (48) and (49),

$$D_\rho(t) \equiv S_\rho(B_\rho(t)), \quad t \ge 0. \tag{50}$$

Let

$$X_\rho(t) \equiv Q_\rho(0) + A_\rho(t) - S_\rho(t), \quad t \ge 0, \tag{51}$$

be a net-input process, acting as if the server is busy all the time, and thus allowing $X_\rho(t)$ to assume negative values. The cumulative busy time $B_\rho(t)$ is then related to $X_\rho(t)$ by

$$B_\rho(t) = t + \zeta(X_\rho)(t), \quad t \ge 0. \tag{52}$$

As a consequence of the assumptions above, $X_\rho(0) = Q_\rho(0)$. Roughly,

$$Q_\rho(t) \approx \psi(X_\rho)(t), \quad t \ge 0, \tag{53}$$

for ψ in (46), but the exact relation breaks down because the service process shuts down when the system becomes idle, so that a new service time does not start until after the next arrival. While (53) does not hold exactly for each ρ, it holds in the HT limit, as shown in Theorem 9.3.4 of Whitt (2002). It would hold exactly if we used the modified system in which we let the continuous-time service process run continuously, so that Equation (53) holds as an equality, as done by Borovkov (1965) and then again in Section 2 of Iglehart and Whitt (1970a). Because the modified system has been shown to be asymptotically equivalent to the original system for these HT limits in Borovkov (1965) and Iglehart and Whitt (1970a), that is an alternative approach.

We now introduce HT-scaled versions of these processes, for that purpose, let

$$\begin{aligned}
X^*_\rho(t) &\equiv (1-\rho)X_\rho((1-\rho)^{-2}t),\\
Q^*_\rho(t) &\equiv (1-\rho)Q_\rho((1-\rho)^{-2}t),\\
A^*_\rho(t) &\equiv (1-\rho)[A_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda t],\\
S^*_\rho(t) &\equiv (1-\rho)[S_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda t/\rho],\\
B^*_\rho(t) &\equiv (1-\rho)[B_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}t],\\
D^*_\rho(t) &\equiv (1-\rho)[D_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda t].
\end{aligned} \tag{54}$$

Let $\mathcal{D}$ be the function space of all right-continuous real-valued functions on $[0,\infty)$ with left limits, with the usual J_1 topology, which reduces to uniform convergence over all bounded intervals for continuous limit functions. Let $\mathcal{D}^k$ be the k-fold product space, using the product topology on all product spaces. Let $\Rightarrow$ denote convergence in distribution. Let e be the identity function in $\mathcal{D}$, i.e., $e(t) \equiv t$, $t \ge 0$.

Theorem 5.1. *If*

$$(Q^*_\rho(0), A^*_\rho, S^*_\rho) \Rightarrow (Q^*(0), A^*, S^*), \quad \textit{in } \mathbb{R} \times \mathcal{D}^2 \textit{ as } \rho \uparrow 1, \tag{55}$$

where A^* and S^* have continuous sample paths with $A^*(0) = S^*(0) = 0$ w.p.1., then

$$(A^*_\rho, S^*_\rho, B^*_\rho, X^*_\rho, Q^*_\rho, D^*_\rho) \Rightarrow (A^*, S^*, B^*, X^*, Q^*, D^*), \tag{56}$$

where the convergence is in $\mathcal{D}^6$ as $\rho \uparrow 1$ and

$$\begin{aligned}
X^* &\equiv Q^*(0) + A^* - S^* - \lambda e,\\
B^* &\equiv \zeta(X^*) < 0,\\
Q^* &\equiv \psi(X^*) = X^* - \zeta(X^*) \quad \textit{and}\\
D^* &\equiv Q^*(0) + A^* - Q^*\\
&= Q^*(0) + A^* - \psi(X^*) = S^* + \lambda e + \zeta(X^*)
\end{aligned} \tag{57}$$

for ψ and ζ in (46) and (47).

Proof. First, note that

$$X_\rho^*(t) = Q_\rho^*(0) + A_\rho^*(t) - S_\rho^*(t) - \lambda t/\rho, \quad t \ge 0, \tag{58}$$

because A_ρ^* and S_ρ^* have different translation terms in (54), ensuring that the potential rate out is λ/ρ, which exceeds the rate in of λ, consistent with a stable model for each ρ, $0 < \rho < 1$. Hence, under the assumption, $X_\rho^* \Rightarrow X^* = Q^*(0) + A^* - S^* - \lambda e$ in $\mathcal{D}$. The limit $B_\rho^* \Rightarrow B^* = \zeta(X^*)$ is obtained by exploiting the relationship in (52). The limits for Q_ρ^* and D_ρ^* then follow from the continuous mapping theorem after carefully accounting for the busy and idle time of the server; see the proof of Theorem 9.3.4 and preceding material in Whitt (2002). □

5.2. A Heavy-Traffic Limit for the Stationary Departure Process

Theorem 5.1 is not easy to apply to establish HT limits for stationary processes because condition (55) is not easy to check and the limit in (56) and (57) is not easy to evaluate.

In order to establish a tractable HT limit for the stationary departure process, we apply the recent HT limits for the stationary queue length in Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009). Their HT limits are for generalized open Jackson networks of queues, which for the single queue we consider reduce to the $GI/GI/1$ model. Following Budhiraja and Lee (2009), we assume that the interarrival times and service times come from independent sequences of i.i.d. random variables with uniformly bounded third moments ($2+\epsilon$ would do).

Theorem 5.2. *For the $GI/GI/1$ model indexed by ρ, assume that* (i) *the interarrival-time cdf has a pdf as in Section 2.1 and* (ii) *the interarrival times and service times have means λ and λ/ρ, scv's c_a^2 and c_s^2, without both being 0, and uniformly bounded third moments. Then*:

(a) *For each ρ, $0 < \rho < 1$, the process Q_ρ^* can be regarded as a stationary process, while the process D_ρ^* can be regarded as a stationary point process* (*with stationary increments*).

(b) *Condition* (55) *in Theorem 5.1 holds with*

$$A^* \equiv c_a B_a \circ \lambda e \qquad \text{and} \qquad S^* \equiv c_s B_s \circ \lambda e, \tag{59}$$

where B_a and B_s are independent standard (*mean 0, variance 1*) *Brownian motions* (*BMs*) *that are independent of $Q^*(0)$, which is distributed as $R_e(0)$ with R_e being a stationary RBM with drift $-\lambda$ and variance $\lambda c_x^2 \equiv \lambda c_a^2 + \lambda c_s^2$, and so an exponential marginal distribution, i.e.,*

$$P(Q^*(0) > x) = e^{-2x/c_x^2}, \quad x \ge 0. \tag{60}$$

(c) *The limits in Theorem 5.1 hold, where*

$$X^* \equiv Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda e \tag{61}$$

with $Q^*(0)$, B_a and B_s being mutually independent.

(d) *We have*

$$D^* = c_a B_a \circ \lambda e + Q^*(0) - Q^* = c_a B_a \circ \lambda e + Q^*(0) - \psi(X^*) \tag{62}$$

for ψ in (46), *or*

$$D^* \equiv S^* + \lambda e + \zeta(X^*) \tag{63}$$

for ζ in (47).

Proof. First, recall that the HT limit as $\rho \to 1$ starting empty is the RBM that converges as $t \to \infty$ to the exponential distribution in (60). We will be applying Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009) to show that the two iterated limits involving $\rho \to 1$ and $t \to \infty$ are equal. Toward that end, we observe that, by Asmussen (2003, Sections X.3–X.4), the queue-length process has a proper steady-state distribution for each ρ. As on p. 63 of Gamarnik and Zeevi (2006), we add the residual interarrival times and service times to the state description for $Q_\rho(t)$ to make it a Markov process that has a unique steady-state distribution for each ρ. These residual interarrival and service times are asymptotically negligible in the HT limit. The associated departure process $D_\rho(t)$ then necessarily is a stationary point process for each ρ. We then can apply (Gamarnik and Zeevi 2006, Theorem 8) to have a limit for the scaled stationary distributions, so that condition (55) holds with (59). Since strong moment-generating-function-condition is imposed in Gamarnik and Zeevi (2006, (1) and (2), p. 62), we apply the extension in Budhiraja and Lee (2009, Theorems 3.1 and 3.2) to cover our moment condition. Hence, we can apply Theorem 5.1 with these special initial distributions to get the associated process limits in the space $\mathcal{D}$. □

We now establish an HT limit for the variance of the stationary departure process. The form of that limit is already given in Theorem 3.1. The proof can be found in Section 8.3.

Theorem 5.3 (Limiting Variance)**.** *Under the conditions of Theorem 5.2 plus the usual uniform integrability conditions, for which it suffices for the interarrival times and service times to have uniformly bounded fourth moments,*

$$V^*_{d,\rho}(t) \equiv \mathrm{Var}(D^*_\rho(t)) = E[D^*_\rho(t)^2] \to E[D^*(t)^2] = \mathrm{Var}(D^*(t)) \equiv V^*_d(t), \quad \textit{as } \rho \uparrow 1, \tag{64}$$

where

$$V^*_d(t) = c_a^2 \lambda t + \frac{c_x^4}{2}(1 - c^*(\lambda t/c_x^2)) - 2\,\mathrm{Cov}(c_a B_a(\lambda t), Q^*(t)) = w^*(\lambda t/c_x^2) c_a^2 \lambda t + (1 - w^*(\lambda t/c_x^2)) c_s^2 \lambda t \tag{65}$$

with $c_x^2 = c_a^2 + c_s^2$, $c^*(t)$ is the correlation function in (24) and $w^*(t)$ is the weight function in (28); i.e., $V^*_d(t)$ is given in (33) with $\gamma = 1$, but allowing general c_s^2. Moreover, we have the covariance formulas

$$\mathrm{Cov}(c_a B_a(\lambda t), Q^*(t)) = (1 - w^*(\lambda t/c_x^2)) c_a^2 \lambda t, \qquad \mathrm{Cov}(c_s B_s(\lambda t), Q^*(t)) = -(1 - w^*(\lambda t/c_x^2)) c_s^2 \lambda t. \tag{66}$$

As a by-product, the covariance formulas in (66) can be generalized to describe the covariance of between a stationary RBM and a BM, where the underlying BMs are correlated.

Corollary 5.1. *Suppose $B = (B_1, B_2)$ is a 2-d Brownian motion with zero drift and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}$. Let $Q = \psi(B_1 + Q^*(0) - \lambda e)$ be the stationary RBM associated with the drifted BM $B_1 - \lambda e$ and $Q^*(0)$ has the stationary distribution of Q^*, which is independent of B_1. Then*

$$\mathrm{cov}(B_2, Q) = (1 - w^*(\lambda^2 t/\sigma_1^2))\sigma_{1,2} t = \frac{\sigma_{1,2}\sigma_1^2}{2\lambda^2}(1 - c^*(\lambda^2 t/\sigma_1^2)).$$

Remark 5.1 (The Quasireversible Case)**.** The limit process

$$(A^*, S^*, X^*, Q^*, D^*),$$

where

$$(A^*, S^*, X^*) = (c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda e),$$

as in Theorem 5.2, can be called the *Brownian queue*; see Harrison (1985), Harrison and Williams (1990, 1992), and O'Connell and Yor (2001). The Brownian queue is known to be quasireversible if and only if $c_a^2 = c_s^2$. In that case, the stationary departure process is a BM and the departures in the past are independent of the steady-state content. Consistent with that theory, $V^*_d(t) = c_a^2 \lambda t, t \ge 0$ in (65) if and only if $c_a^2 = c_s^2$.

Remark 5.2. We considered only the case where $\mu_\rho = \lambda/\rho$ in this section. Now, we list the results for a slightly more general case as in Section 3, where we have $\mu_\rho = \lambda(1 + (1-\rho)\gamma_\rho)$ and $\lim_{\rho\uparrow 1}\gamma_\rho = \gamma$. One can easily check that Theorem 5.1 holds with

$$X^* = Q^*(0) + A^* - S^* - \lambda\gamma e;$$

Theorem 5.2 holds with

$$P(Q^*(0) > x) = e^{-2\gamma x/c_x^2} \qquad \text{and} \qquad X^* = Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda \gamma e;$$

and Theorem 5.3 holds with

$$V^*_d(t) = w^*(\lambda\gamma^2 t/c_x^2) c_a^2 \lambda t + (1 - w^*(\lambda\gamma^2 t/c_x^2)) c_s^2 \lambda t,$$
$$\mathrm{Cov}(c_a B_a(\lambda t), Q^*(t)) = (1 - w^*(\lambda\gamma^2 t/c_x^2)) c_a^2 \lambda t, \qquad \text{and}$$
$$\mathrm{Cov}(c_s B_s(\lambda t), Q^*(t)) = -(1 - w^*(\lambda\gamma^2 t/c_x^2)) c_s^2 \lambda t.$$

We conclude this section with the tail asymptotics of the departure variance function. Just as in Corollaries 3.2 and 4.1, we have

Corollary 5.2 (Asymptotic Behavior of the Departure Variance Function)**.** *Under the assumptions in Theorem 5.3 and Remark 5.2,*

$$V^*_d(t) \sim c_a^2 \lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{8(c_s^2 - c_a^2)c_x^5}{\gamma^5}\frac{1}{\sqrt{2\pi\lambda^3 t^3}} e^{-(\lambda\gamma^2 t)/(2c_x^2)}, \quad \textit{as } t \to \infty. \tag{67}$$

Remark 5.3. Hautphenne et al. (2015) developed explicit expressions for the y-intercept $\bar{b}_\theta$ of the linear asymptote for the variance of the stationary departure from $M/GI/1$

$$V(t) = \bar{v}t + \bar{b}_\theta + o(1), \quad \text{as } t \to \infty;$$

see Proposition 6 there. Their result (i) holds for $M/GI/1$ case; (ii) depends on the third moment of the service distribution; (iii) holds for general traffic intensity. Even though there is no direct heavy-traffic scaling in their result, the scaling parameter emerges in their expression, see the definition of $\bar{b}_\theta$ there. In specific, the scaling constant $\rho/(1-\rho)^2$ coincides (up to a constant ρ) with the one we use in (4).

On the other hand, our result here (i) coincides with their y-intercept (after scaling) in the HT limit in the $M/GI/1$ case, i.e., let $\rho = 1$, $\gamma = 1$ and $c_a = 1$; (ii) holds for general $GI/GI/1$ cases but only under HT limit; (iii) has explicit characterization of the remainder term, again only under HT limit.

6. Application to a Robust Queueing Network Analyzer

We conclude by explaining the important role that Theorem 5.3 plays in our *Robust Queueing Network Analyzer* (RQNA) based on the *index of dispersion for counts* (IDC), which we refer to as RQNA-IDC. In Section 6.1 we briefly review the *robust queueing* (RQ) approximation for the mean steady-state workload of a $G/G/1$ queue developed in Whitt and You (2018a), which requires the IDC of the arrival process as model data in the $G/GI/1$ model. Then, in Section 6.2 we review the approximation for the IDC of the departure process that we propose in Whitt and You (2018c), which is supported by this paper.

6.1. The Mean Steady-State Workload at a $G/G/1$ Queue

In this section we review the RQ approximation for the mean steady-state workload at a $G/G/1$ queue developed in Whitt and You (2018a). We start with a rate-λ stationary arrival process $A \equiv \{A(t)\colon t \ge 0\}$, having stationary increments. Thus, for a renewal arrival process, we work with the associated equilibrium renewal process. We also start with a stationary and ergodic sequence of rate-μ service times $\{V_k\colon k \ge 1\}$ with finite scv c_s^2 and possibly correlated with the arrival process. For model stability, we assume that $\rho = \lambda/\mu < 1$. Let $Y(t)$ be the associated total input process, defined by

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \ge 0, \tag{68}$$

which also has stationary increments. Let $Z_\rho(t)$ be the workload at time t in model ρ with arrival rate ρ and service rate 1, then

$$Z_\rho(t) = \psi(Y-e)(t) = Y(t) - t - \inf_{s \le t}\{Y(s) - s\} = \sup_{s \le t}\{Y(t) - Y(s) - (t-s)\}.$$

As in Whitt and You (2018a, Section 3.1), we can apply a reverse-time construction to write the steady-state workload Z_ρ as a simple supremum.

For the RQ approximation of the mean steady-state workload, we use the associated *index of dispersion for work* (IDW) $I_w \equiv \{I_w(x)\colon x \ge 0\}$ introduced by Fendick and Whitt (1989), which is defined by

$$I_w(t) \equiv \frac{\mathrm{Var}(Y(t))}{E[V]E[Y(t)]}, \quad t \ge 0; \tag{69}$$

see Whitt and You (2018a, Section 4.3) for key properties. Let Z_ρ denote the steady-state workload in the $G/G/1$ model when the traffic intensity is ρ, then the RQ approximation (based on this partial model characterization) is

$$E[Z_\rho] \approx Z_\rho^* \equiv \sup_{x \ge 0}\left\{-(1-\rho)x + \sqrt{2\rho x I_w(x)/\mu}\right\}, \tag{70}$$

for the IDW in (72). The approximation (70) comes from Whitt and You (2018a, (28)), assuming that we set the parameter $b_f = \sqrt{2}$, which makes the approximation asymptotically correct for the $GI/GI/1$ model in both the heavy-traffic and light-traffic limits; see Whitt and You (2018a, Theorem 5). Notice that the approximation in (70) is directly a supremum of a real-valued function and so can be computed quite easily for any given tuple (ρ, I_w).

In what follows, we primarily focus on the $G/GI/1$ special case, where the arrival process is general stationary and ergodic, and the service times are i.i.d. We assume that the arrival process A is partially characterized by its *index of dispersion for counts* (IDC) $I_a(t)$, where

$$I_a(t) \equiv \frac{\mathrm{Var}(A(t))}{E[A(t)]}, \quad t \geq 0; \tag{71}$$

see Cox and Lewis (1966, Section 4.5). (To explain, just as the distribution of a random variable is not fully characterized by its mean and variance, so the distribution of the stationary stochastic process $\{A(t)\colon t \geq 0\}$ is not fully characterized by its mean function $E[A(t)]$ and its variance function $\mathrm{Var}(A(t))$, $t \geq 0$.) In this context,

$$I_w(t) = I_a(t) + c_s^2, \quad t \geq 0, \tag{72}$$

as noted in Whitt and You (2018a, Section 4.3.1). The corresponding RQ approximation is then

$$E[Z_\rho] \approx Z_\rho^* \equiv \sup_{x\geq 0}\{-(1-\rho)x + \sqrt{2\rho x(I_a(x) + c_s^2)/\mu}\}. \tag{73}$$

Hence, to apply the RQ algorithm in the $G/GI/1$ model, we need only develop an approximation for the arrival IDC $I_a(t)$.

6.2. The IDC of a Stationary Departure Process

The main challenge in developing a full RQNA-IDC involving a decomposition approximation is calculating or approximating the required IDC for the arrival process at each queue. For a renewal arrival process, the IDC I_a can be computed by inverting the LT $\hat{V}(s)$ in (11) using the LT $\hat{m}(s)$ of the renewal function $m(t)$ in (6), which only requires the LT $\hat{f}(s)$ of the interarrival-time pdf in (5). Numerical algorithms for calculating and simulations algorithms for estimating the IDC are discussed in Whitt and You (2018b).

As discussed in Section 6 of Whitt and You (2018a), the main challenge is approximating the IDC of a departure process from a $G/GI/1$ queue, partially specified by the tuple $(\lambda, \rho, I_a, c_s^2)$. As in Section 6 of Whitt and You (2018a), we propose approximating the IDC of the departure process, $I_{d,\rho}(t)$ by the weighted average of the IDCs of the arrival and service processes, i.e.,

$$I_{d,\rho}(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(t), \quad t \geq 0, \tag{74}$$

where $I_a(t)$ and $I_s(t)$ are the IDCs associated with the equilibrium arrival and service renewal processes, *both* with rate λ. We thus require $(\lambda, \rho, I_a, I_s)$ as model data. Based on initial study, a candidate weight function $w_\rho(t)$ was suggested in (43) of Whitt and You (2018a).

We now show how the further study in Theorem 5.3 suggests a new weight function

$$w_\rho(t) \equiv w^*((1-\rho)^2 \lambda t/(\rho c_x^2)), \tag{75}$$

where $c_x^2 \equiv c_a^2 + c_s^2$ and w^* is given in (28) with c^* in (24).

To start, let $I_{d,\rho}$ denote the departure IDC and define the weight function

$$w_\rho(t) \equiv \frac{I_{d,\rho}(t) - I_s(t)}{I_a(t) - I_s(t)} = \frac{V_d(t) - V_s(t)}{V_a(t) - V_s(t)}, \tag{76}$$

where $V_a(t)$ and $V_s(t)$ are the variance functions associated with the equilibrium arrival and service renewal processes, *both* with rate λ. Note that this is exactly the same weight function we defined in (28), thus we have the same HT-scaled weight function w_ρ^* as in (35). We then apply Theorem 5.3 to obtain a paralleling corollary.

Corollary 6.1 (Limiting Weight Function). *Under the assumptions in Theorem 5.3, we have*

$$w_\rho^*(t) \Rightarrow w^*(\lambda t/c_x^2)$$

for w^* defined in (28).

To obtain the pre-limit approximation, we rearrange terms in (76), and assume

$$w_\rho(t) \approx w^*((1-\rho)^2 \lambda t/c_x^2). \tag{77}$$

One remaining issue is that the approximation (77) does not automatically yield a correct light-traffic limit, in which case we must have $I_{d,0}(t) = I_a(t)$ since the service time is negligible. As a remedy, we propose to add a constant ρ^{-1} correction in the weight function, so that we have (75) as the final weight function.

We conclude this section with two simulation examples.

Example 6.1 (Simulation Experiments). We now consider two $GI/GI/1$ queues, where neither the interarrival time nor service time has an exponential distribution. Let $H_2(c^2)$ be the H_2 (hyperexponential) distribution with scv c^2 and balanced means, as in (Whitt 1982, (3.7), p. 137). Consider the $H_2(4)/E_2/1$ model with $\lambda = 2$, Figure 2 (top) reports the simulated departure IDCs for three different traffic intensities $\rho = 0.95, 0.8, 0.5$, as well as the approximation (74) with (75). The simulation estimation of the departure IDC is obtained from a single run of length 10^9 time units, with the first 10^6 time units are discarded in order for the system to approach steady-state. The reference IDCs I_a and I_s is calculated by numerically inverting the LT in (11). Figure 2 (bottom) is the corresponding plot for the $E_2/H_2(4)/1$ model with $\lambda = 2$.

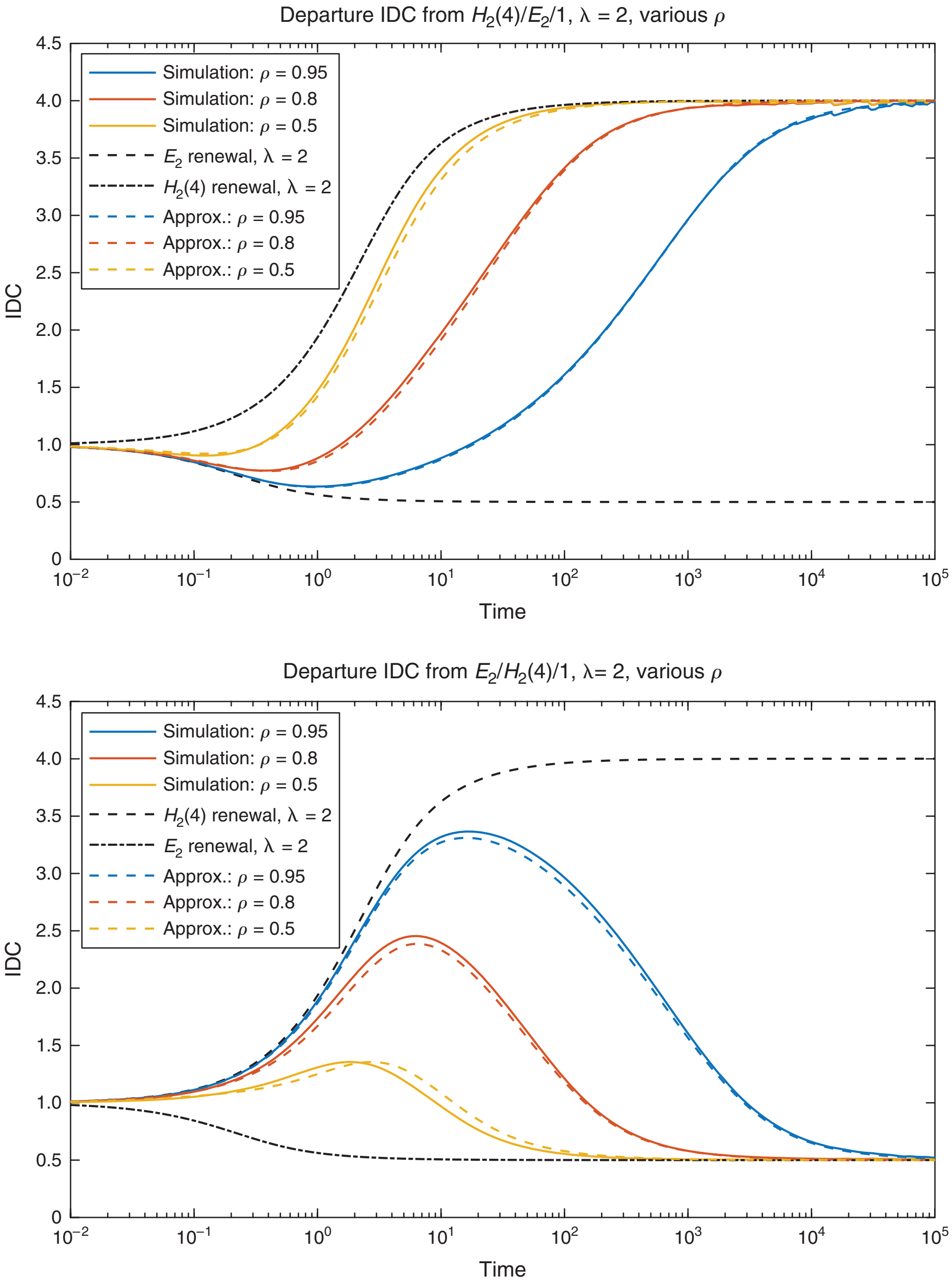


Figure 2. On the top is the IDCs of the departure from $H_2(4)/E_2/1$ model with $\lambda = 2$ and three different traffic intensities, together the two reference IDCs, one for $H_2(4)$ and one for E_2, displayed in broken black lines. The approximation (74) and (75) are displayed in broken colored lines. On the bottom is a similar plot for the $E_2/H_2(4)/1$ model.

7. Extensions

The approximation for the departure IDC $I_d(t)$ in (74) and (75) should be good for much more general models than $GI/GI/1$, with the independence conditions relaxed and more than 1 server. We also conjecture that the HT limit of the variance function in Theorem 5.3 extends to a larger class of models as well. Indeed, we conjecture that the limits established for $GI/GI/1$ extend in that way. First, Theorem 5.1 extends quite directly by exploiting Iglehart and Whitt (1970a, b). For the extension of Theorem 5.2, there is a large class of models for which the HT-scaled arrival and service processes have the limits

$$A^* \equiv c_a B_a \qquad \text{and} \qquad S^* \equiv c_s B_s, \tag{78}$$

where B_a and B_s are independent standard (mean 0, variance 1) Brownian motions (BM's) that are independent of the initial queue length. What is needed is the extension of Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009) to more general models. We conjecture that can be done for $GI/GI/s$ and other models with regenerative structure in the arrival and service processes. For $GI/GI/s$ the queue-length process again becomes a Markov process if we append the s elapsed service times as well as the elapsed interarrival time, but it remains to do the hard technical analysis leading to an appropriate Lyapunov function.

It is also of interest to establish related results for departure processes in models with nonrenewal arrival processes, as in Ferng and Chang (2001) and references therein. It also remains to establish new HT limits for stationary departure processes from a queue within a network, obeying the HT FCLT in Reiman (1984).

The relevant approximation for the stationary departure process from a many-server $GI/GI/s$ queue evidently is quite different, being more like the service process than the arrival process. We conjecture that the relevant many-server heavy-traffic limit for the stationary departure process is a Gaussian process with the covariance function of the stationary renewal processes associated with the service times, as in the CLT for renewal processes in Whitt (2002, Theorems 7.2.1 and 7.2.4). Partial support comes from Aras et al. (2018), Aras et al. (2017b, Apendix F), and Gamarnik and Goldberg (2013).

8. Proofs

8.1. Proof of Theorem 3.1

We now review a useful lemma on the properties of $\hat{\xi}(s)$ and δ, defined in (21) and (22); see Takacs (1962, p. 113) or (Cohen 1982, Appendix 6). (The notation here is slightly different.)

Lemma 8.1 (Takacs' Root Lemma). *If* $\mathrm{Re}(s) \geq 0$, *then the root* $\hat{\xi}(s)$ *of the equation*

$$z = \hat{f}(s + \mu(1 - z))$$

that has the smallest absolute value is

$$\hat{\xi}(s) = \sum_{j=1}^{\infty} \frac{(-\mu)^{j-1}}{j!} \frac{d^{j-1}}{ds^{j-1}} (\hat{f}(\mu + s))^j. \tag{79}$$

This root $\hat{\xi}(s)$ *is a continuous function of* s *for* $\mathrm{Re}(s) \geq 0$. *Furthermore,* $z = \hat{\xi}(s)$ *is the only root in the unit circle* $|z| \leq 1$ *if at least one of two conditions is satisfied* (i) $\mathrm{Re}(s) > 0$, *or* (ii) $\mathrm{Re}(s) \geq 0$ *and* $\lambda/\mu < 1$. *Specifically,* $\delta = \hat{\xi}(0)$ *is the smallest positive real root of the equation*

$$\delta = \hat{f}(\mu(1 - \delta)).$$

If $\lambda/\mu < 1$, *then* $\delta < 1$ *and if* $\lambda/\mu \geq 1$ *then* $\delta = 1$.

Proof of Theorem 3.1. We let $\rho \uparrow 1$ by decreasing the service rate, so that $1/E[U] = \lambda$ is fixed. To allow general drift in the Brownian HT limit, we let $1/E[V] = \mu_\rho \equiv \lambda + (1 - \rho)\lambda\gamma_\rho$ in system ρ, for positive constants $\gamma_\rho \to \gamma$. Under this setting, we have $(\lambda - \mu_\rho)/(1 - \rho) \to -\lambda\gamma$ as $\rho \uparrow 1$. By (20) and (23), we have

$$\hat{V}^*_{d,\rho}(s) = \mathscr{L}((1-\rho)^2 V_{d,\rho}((1-\rho)^{-2} t)) = (1-\rho)^4 \hat{V}_{d,\rho}((1-\rho)^2 s) = \frac{\lambda}{s^2} + \frac{\lambda}{s^2} \hat{W}(s)$$

where

$$\begin{aligned} \hat{W}(s) \equiv \frac{2}{(1-\rho)^2 s} \Bigg(\mu_\rho \left(\delta - \frac{\lambda}{\mu_\rho} \right) &+ \left(\mu_\rho \delta \frac{1 - \hat{f}((1-\rho)^2 s)}{(1-\rho)^2 s} - \hat{f}((1-\rho)^2 s) \right) \\ &\cdot \left(\frac{(1-\rho)^2 s + \mu_\rho (1 - \hat{\xi}((1-\rho)^2 s))}{\mu_\rho (1 - \xi((1-\rho)^2 s))} \cdot \frac{(1-\rho)^2 s - \mu_\rho (1 - \delta)}{\mu_\rho (1 - \delta)} \cdot \frac{1 - \hat{f}((1-\rho)^2 s)}{(1-\rho)^2 s} \right)^{-1} \Bigg). \end{aligned}$$

Then, we write

$$\begin{aligned}\hat{W}(s)&=\frac{2\mu_\rho}{(1-\rho)^2s}\frac{(\delta-\lambda/\mu_\rho)\hat{H}_\rho(s)((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))+\delta((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))-(1/\mu_\rho)\hat{f}((1-\rho)^2s)}{\hat{H}_\rho(s)((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))}\\
&=\frac{2\mu_\rho}{(1-\rho)^2s}\frac{\delta(\hat{H}_\rho(s)+1)((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))-(\lambda/\mu_\rho)\hat{H}_\rho(s)((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))-(1/\mu_\rho)\hat{f}((1-\rho)^2s)}{\hat{H}_\rho(s)((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))}\\
&=\frac{1}{\hat{H}_\rho(s)((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))}\frac{2\mu_\rho}{(1-\rho)^2s}\\
&\quad\cdot\left(\left(\delta-\frac{\lambda}{\mu_\rho}\right)(\hat{H}_\rho(s)+1)\frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s}\,\frac{\lambda}{\mu_\rho}\left(\frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s}-\frac{1}{\lambda}\right)+\frac{1}{\mu_\rho}(1-\hat{f}((1-\rho)^2s))\right)\\
&=\frac{2\mu_\rho}{\hat{H}_\rho(s)((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))}\\
&\quad\cdot\left(\frac{\delta-\lambda/\mu_\rho}{1-\rho}\frac{\hat{H}_\rho(s)+1}{(1-\rho)s}\frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s}+\frac{\lambda((1-\hat{f}((1-\rho)^2s))/((1-\rho)^2s))-1}{\mu_\rho(1-\rho)^2s}+\frac{1-\hat{f}((1-\rho)^2s)}{\mu_\rho(1-\rho)^2s}\right),\end{aligned}$$

where

$$\hat{H}_\rho(s)\equiv\left(\frac{1}{\mu_\rho}\frac{(1-\rho)^2s}{1-\hat{\xi}((1-\rho)^2s)}+1\right)\left(\frac{1}{\mu_\rho}\frac{(1-\rho)^2s}{1-\delta}-1\right).$$

By Lemma 8.1, we know that δ is positive and real, and $\delta<1$ if $\rho<1$ while $\delta=1$ if $\rho=1$. Hence, we may restrict the function $\hat{f}$ to the real axis. Then, expanding $\hat{f}$ in a Taylor series about 0, yields

$$\begin{aligned}\delta=\hat{f}(\mu_\rho(1-\delta))\Rightarrow\delta&=\hat{f}(0)+\hat{f}'(0)\mu_\rho(1-\delta)+\left(\frac{1}{2}\hat{f}''(0)\mu_\rho^2+o(1)\right)(1-\delta)^2\\
&\Rightarrow 0=1-\delta-\frac{\mu_\rho}{\lambda}(1-\delta)+\left(\frac{1}{2}\hat{f}''(0)\mu_\rho^2+o(1)\right)(1-\delta)^2\\
&\Rightarrow 0=\frac{1-\mu_\rho/\lambda}{1-\rho}+\left(\frac{c_a^2+1}{2}\frac{\mu_\rho^2}{\lambda^2}+o(1)\right)\frac{1-\delta}{1-\rho}\\
&\Rightarrow\frac{1-\delta}{1-\rho}=\gamma_\rho\left(\frac{c_a^2+1}{2\rho}+o(1)\right)^{-1}.\end{aligned}\tag{80}$$

This implies that the following limit exist

$$\delta^*\equiv\lim_{\rho\uparrow1}\frac{1-\delta}{1-\rho}=\frac{2\gamma}{c_a^2+1}.\tag{81}$$

Now, let $\hat{\xi}_{\rho,s}\equiv\hat{\xi}((1-\rho)^2s)=\hat{f}((1-\rho)^2s+\mu_\rho(1-\hat{\xi}_{\rho,s}))$, then similarly we have

$$0=\gamma_\rho\frac{1-\hat{\xi}_{\rho,s}}{1-\rho}+\frac{s}{\lambda}-\left(\frac{c_a^2+1}{2\lambda^2}+o(1)\right)\left((1-\rho)s+\mu_\rho\frac{1-\hat{\xi}_{\rho,s}}{1-\rho}\right)^2.\tag{82}$$

Then (82) implies that the following limit exists

$$\hat{\xi}^*(s)\equiv\lim_{\rho\uparrow1}\frac{1-\hat{\xi}_{\rho,s}}{1-\rho},\tag{83}$$

and

$$\frac{c_a^2+1}{2}(\hat{\xi}^*(s))^2-\gamma\hat{\xi}^*(s)-\frac{s}{\lambda}=0.\tag{84}$$

Recall that $\hat{\xi}_{\rho,s}$ is defined to be the root of $z=\hat{f}((1-\rho)^2s+\mu_\rho(1-z))$ with smallest absolute value. By Lemma 8.1, this root is unique and lies in the unit circle unless $s=0$ and $\rho=1$, in which case $\hat{\xi}(0)=1$. Furthermore, it can be proved by Weierstrass Preparation Theorem (see Chow and Hale 2012, Theorem 6.2) that $\hat{\xi}_{\rho,s}$ is continuous in (ρ,s). Hence, we have

$$\mathrm{Re}\left(\frac{1-\hat{\xi}_{\rho,s}}{1-\rho}\right)>0,\quad\text{for all }\rho<1\text{ and }s>0.$$

By taking limit $\rho\uparrow1$, we have $\mathrm{Re}(\hat{\xi}^*(s))\geq0$ for all $s>0$.

As a consequence, we pick the root of (84) with nonnegative real part. In particular, for real s, we have

$$\hat{\xi}^*(s) = \frac{\gamma + \sqrt{\gamma^2 + 2(c_a^2+1)s/\lambda}}{c_a^2+1}. \tag{85}$$

For complex s, the square root in (85) corresponds to two complex roots, which are also the roots of $(\gamma - \sqrt{\gamma^2 + 2(c_a^2+1)s/\lambda})/(c_a^2+1)$, since the polynomial in (84) is of order 2. Hence, we may use the same expression (85) as in the real case, as long as we pick the one with nonnegative real part.

Combining (81) and (83), we obtain

$$\lim_{\rho\uparrow 1} \hat{H}_\rho(s) = -1,$$

and

$$\hat{H}^*(s) \equiv \lim_{\rho\uparrow 1} \frac{\hat{H}_\rho(s)+1}{(1-\rho)s} = \lim_{\rho\uparrow 1}\left(\frac{1}{\mu_\rho}\frac{1-\rho}{1-\delta} - \frac{1}{\mu_\rho}\frac{1-\rho}{1-\hat{\xi}((1-\rho)^2 s)} + O(1-\rho)\right) = \frac{1}{\lambda}\left(\frac{c_a^2+1}{2\gamma} - \frac{1}{\hat{\xi}^*(s)}\right) < \infty, \tag{86}$$

where $\hat{\xi}^*$ is defined in (83). Moreover, we have

$$\lim_{\rho\uparrow 1} \frac{1-\hat{f}((1-\rho)^2 s)}{(1-\rho)^2 s} = -\hat{f}'(0) = E[U] = 1/\lambda,$$

and

$$\lim_{\rho\uparrow 1} \frac{(1-\hat{f}((1-\rho)^2 s))/((1-\rho)^2 s) - 1/\lambda}{(1-\rho)^2 s} = -\frac{\hat{f}''(0)}{2} = -\frac{E[U^2]}{2} = -\frac{c_a^2+1}{2\lambda^2}.$$

Combining everything into the Laplace Transform of $(1-\rho)^2 V_{d,\rho}((1-\rho)^{-2}t)$, we have

$$\hat{V}_d^*(s) \equiv \lim_{\rho\uparrow 1} \hat{V}_{d,\rho}^*(s) = \frac{\lambda}{s^2} - \frac{\lambda(c_a^2-1)}{s^2}\left(\frac{2\gamma\lambda}{c_a^2+1}\hat{H}^*(s) - 1\right) \tag{87}$$

$$= \frac{\lambda}{s^2} + \frac{2\lambda}{s^2}\frac{c_a^2-1}{c_a^2+1}\frac{\gamma}{\hat{\xi}^*(s)}. \tag{88}$$

Plugging in (85), we obtain

$$\hat{V}_d^*(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s^2}\frac{c_a^2-1}{c_a^2+1}\frac{\gamma}{\hat{\xi}^*(s)} = \frac{\lambda}{s^2} + \frac{\lambda}{s^2}\frac{c_a^2-1}{c_a^2+1}\frac{\sqrt{1+2(c_a^2+1)s/(\lambda\gamma^2)} - 1}{s/(\lambda\gamma^2)},$$

where we pick the root such that $(\sqrt{1+2(c_a^2+1)s/(\lambda\gamma^2)} - 1)/(s/(\lambda\gamma^2))$ has nonnegative real part. We used the fact that $\mathrm{Re}(z) \ge 0$ if and only if $\mathrm{Re}(1/z) \ge 0$ for $z \ne 0$.

For the explicit inversion, one can exploit the LT of the correlation function in (25) and note that

$$\mathscr{L}(f(at))(s) = \frac{1}{a}\hat{f}(s/a)$$

for any constant $a \ne 0$ and any function f with LT $\hat{f}$. For our case here, we use $a = \lambda\gamma^2/(c_a^2+1)$. □

8.2. Proof of Theorem 4.2

Proof of Theorem 4.2. To simplify the proof, we consider the HT-scaled difference between departure variance function and service variance function. Let $1/E[V] = \mu$ and $1/E[U] = \lambda_\rho \equiv \mu(1-(1-\rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho\uparrow 1}\gamma_\rho = \gamma > 0$. Under this setting, we have $(\lambda_\rho - \mu)/(1-\rho) \to -\mu\gamma$ as $\rho \uparrow 1$. Let $\hat{V}_{d,\rho}^*(s)$ and $\hat{V}_{s,\rho}^*(s)$ be the LT of $V_{d,\rho}^*(s)$ and $V_{s,\rho}^*(s)$, respectively. Recall that Π was defined in Theorem 4.1. By (39), we have

$$\hat{V}_d^*(s) - \hat{V}_s^*(s) = \lim_{\rho\uparrow 1}(\hat{V}_{d,\rho}^*(s) - \hat{V}_{s,\rho}^*(s)) = \lim_{\rho\uparrow 1}(1-\rho)^4(\hat{V}_{d,\rho}((1-\rho)^2 s) - \hat{V}_{s,\rho}((1-\rho)^2 s))$$

$$= \lim_{\rho\uparrow 1}\left(\frac{\lambda_\rho - \mu}{s^2} + \frac{2(\lambda_\rho - \mu)}{s^2}\frac{\hat{g}((1-\rho)^2 s)}{1-\hat{g}((1-\rho)^2 s)} - \frac{2(\lambda_\rho^2 - \mu^2)}{(1-\rho)^2 s^3}\right)$$

$$- \lim_{\rho\uparrow 1}\left(\frac{2\lambda_\rho}{s^2}\frac{\hat{g}((1-\rho)^2 s)}{1-\hat{g}((1-\rho)^2 s)}\frac{(1-\rho)^2 s\Pi(\hat{v}((1-\rho)^2 s))}{(1-\rho)^2 s + \lambda_\rho(1-\hat{v}((1-\rho)^2 s))}\right)$$

$$
\begin{aligned}
&= \lim_{\rho\uparrow 1} \frac{\lambda_\rho - \mu}{s^2}\left(1 - 2(\lambda_\rho + \mu)\frac{1}{(1-\rho)^2 s}\left(1 - \frac{\hat{g}((1-\rho)^2 s)}{\mu((1-\hat{g}((1-\rho)^2 s))/((1-\rho)^2 s))}\right)\right)\\
&\quad + \lim_{\rho\uparrow 1} \frac{2\lambda_\rho}{s^2}\frac{\hat{g}((1-\rho)^2 s)}{(1-\hat{g}((1-\rho)^2 s))/((1-\rho)^2 s)}\cdot\frac{1}{1-\rho}\left(\frac{\gamma_\rho}{s} - \frac{\Pi(\hat{v}((1-\rho)^2 s))}{(1-\rho)s + \lambda_\rho((1-\hat{v}((1-\rho)^2 s))/(1-\rho))}\right)\\
&= \lim_{\rho\uparrow 1}\hat{F}^{(1)}_\rho(s) + \lim_{\rho\uparrow 1}\frac{2\lambda_\rho\mu}{s^2}\frac{\hat{g}((1-\rho)^2 s)}{\mu((1-\hat{g}((1-\rho)^2 s))/((1-\rho)^2 s))}\frac{1}{(1-\rho)s + \lambda_\rho((1-\hat{v}((1-\rho)^2 s))/(1-\rho))}\cdot\hat{F}^{(2)}_\rho(s)
\end{aligned}
$$

where

$$
\hat{F}^{(1)}_\rho(s) \equiv \frac{\lambda_\rho - \mu}{s^2}\left(1 - 2(\lambda_\rho + \mu)\frac{1}{(1-\rho)^2 s}\left(1 - \frac{\hat{g}((1-\rho)^2 s)}{\mu((1-\hat{g}((1-\rho)^2 s))/((1-\rho)^2 s))}\right)\right)
$$

and

$$
\hat{F}^{(2)}_\rho(s) \equiv \frac{\gamma_\rho}{1-\rho}\left(1-\rho + \frac{\lambda_\rho}{s}\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho} - \frac{1}{\gamma_\rho}\Pi(\hat{v}((1-\rho)^2 s))\right).
$$

One can easily show that $\hat{F}^{(1)}_\rho(s)$ converges to 0 as $\rho \uparrow 1$. Note also that $\hat{g}(0) = 1$ and $\hat{g}'(0) = -E[V] = -1/\mu$, then

$$
\lim_{\rho\uparrow 1}\frac{\hat{g}((1-\rho)^2 s)}{\mu((1-\hat{g}((1-\rho)^2 s))/((1-\rho)^2 s))} = 1.
$$

Furthermore, a Taylor series expansion around $s = 0$ yields

$$
\begin{aligned}
\frac{\hat{v}((1-\rho)^2 s) - 1}{1-\rho} &= \frac{\hat{g}((1-\rho)^2 s + \lambda_\rho(1-\hat{v}((1-\rho)^2 s))) - 1}{1-\rho}\\
&= -\frac{1-\rho}{\mu}s + \frac{\lambda_\rho}{\mu}\frac{\hat{v}((1-\rho)^2 s) - 1}{1-\rho} + \frac{\hat{g}''(0) + o(1)}{2(1-\rho)}((1-\rho)^2 s + \lambda_\rho(1-\hat{v}((1-\rho)^2 s)))^2,
\end{aligned}
$$

which implies that

$$
\lim_{\rho\uparrow 1}\hat{v}((1-\rho)^2 s) = 1 \tag{89}
$$

and

$$
\begin{aligned}
0 &= -\frac{s}{\mu} + \frac{1-\lambda_\rho/\mu}{1-\rho}\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho} + \frac{\hat{g}''(0) + o(1)}{2(1-\rho)^2}((1-\rho)^2 s + \lambda_\rho(1-\hat{v}((1-\rho)^2 s)))^2\\
&= -\frac{s}{\mu} + \gamma_\rho\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho} + \frac{\hat{g}''(0) + o(1)}{2}\left((1-\rho)s + \lambda_\rho\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho}\right)^2\\
&= -\frac{s}{\mu} + \gamma_\rho\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho} + \frac{\lambda_\rho^2}{\mu^2}\frac{c_s^2 + 1}{2}\left(\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho}\right)^2 + o(1),
\end{aligned}
$$

where we used the fact that $\hat{g}''(0) = E[V^2] = (c_s^2 + 1)/\mu^2$. Hence,

$$
\lim_{\rho\uparrow 1}\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho} = \nu^*(s),
$$

where

$$
\frac{1 + c_s^2}{2}(\hat{v}^*(s))^2 + \gamma\hat{v}^*(s) - \frac{s}{\mu} = 0. \tag{90}
$$

With essentially the same argument as in the proof of Theorem 3.1, one can also show that $\nu^*(s)$ is the only root of (90) with positive real part, furthermore

$$
\nu^*(s) = \frac{-\gamma + \sqrt{\gamma^2 + 2(1 + c_s^2)s/\mu}}{1 + c_s^2}. \tag{91}
$$

It remains to show that $\hat{F}_\rho^{(2)}(s)$ converges (pointwise) to a proper limit. To this end, we write

$$\begin{aligned}\hat{F}_\rho^{(2)}(s) &= \frac{\gamma_\rho}{1-\rho}\left(1-\rho+\frac{\lambda_\rho}{s}\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho}-\frac{1}{\gamma_\rho}\Pi(\hat{v}((1-\rho)^2 s))\right)\\ &= \gamma_\rho+\gamma_\rho\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho}\frac{1}{1-\rho}\left(\frac{\lambda_\rho}{s}-\frac{1}{\gamma_\rho}\frac{(1-\lambda_\rho/\mu)(1-\rho)\hat{g}(\lambda_\rho(1-\hat{v}((1-\rho)^2 s)))}{\hat{g}(\lambda_\rho(1-\hat{v}((1-\rho)^2 s)))-\hat{v}((1-\rho)^2 s)}\right)\\ &= \gamma_\rho+\gamma_\rho\frac{1-\hat{v}((1-\rho)^2 s)}{1-\rho}\frac{1}{1-\rho}\left(\frac{\lambda_\rho}{s}-\frac{(1-\rho)^2\hat{g}(\lambda_\rho(1-\hat{v}((1-\rho)^2 s)))}{\hat{g}(\lambda_\rho(1-\hat{v}((1-\rho)^2 s)))-\hat{v}((1-\rho)^2 s)}\right).\end{aligned}$$

Note that

$$\begin{aligned}\hat{g}\big(\lambda_\rho(1-\hat{v}((1-\rho)^2 s))-\hat{v}((1-\rho)^2 s)\big) &= \hat{g}\big(\lambda_\rho(1-\hat{v}((1-\rho)^2 s))\big)-\hat{g}\big((1-\rho)^2 s+\lambda_\rho(1-\hat{v}((1-\rho)^2 s))\big)\\ &= (1-\rho)^2\frac{s}{\mu}-\hat{g}''(0)(1-\rho)^2 s\lambda_\rho\big(1-\hat{v}((1-\rho)^2 s)\big)+O((1-\rho)^4),\end{aligned}$$

one can easily show that

$$\lim_{\rho\uparrow 1}\hat{F}_\rho^{(2)}(s)=\gamma-\gamma\hat{v}^*(s)\frac{\mu}{s}(1+c_s^2\hat{v}^*(s)). \tag{92}$$

Plugging everything into the Laplace Transform of the heavy-traffic scaled difference of the variance functions, we have

$$\hat{V}_d^*(s)=\hat{V}_s^*(s)+\frac{2\mu^2}{s^2}\frac{1}{\mu\hat{v}^*(s)}\left(\gamma-\gamma\hat{v}^*(s)\frac{\mu}{s}(1+c_s^2\hat{v}^*(s))\right)=\frac{\mu c_s^2}{s^2}+\frac{\gamma\mu^2(1-c_s^2)}{s^3}\hat{v}^*(s) \tag{93}$$

where we apply (90) to obtain the simplified expression in (93).

To obtain the explicit inversion, we write

$$\hat{V}_d^*(s)=\frac{\mu c_s^2}{s^2}+\frac{\gamma\mu^2(1-c_s^2)}{s^3}\frac{-\gamma+\sqrt{\gamma^2+2(1+c_s^2)s/\mu}}{1+c_s^2}.$$

Then, one can exploit the LT of the correlation function in (25) and note that $\mathscr{L}(f(at))(s)=\hat{f}(t/a)/a$, for any constant $a\neq 0$ and any function f with LT $\hat{f}$. For our case here, we use $a=\mu\gamma^2/(1+c_s^2)$. □

8.3. Proof of Theorem 5.3

Proof. By combining Theorems 2.1 and 4.2 in Asmussen (2003, Chapter X), we deduce that the kth moment of the steady-state queue length is finite if the $(k+1)$st moments of the interarrival time and service time are finite. We add the extra uniformly bounded fourth moment to provide the uniform integrability needed to get convergence of the moments in the HT limit. We use (49) to obtain the corresponding result for the departure process.

To get (65), combine (64) and (62). Note that

$$\mathrm{Var}(Q^*(t))=\mathrm{Var}(Q^*(0))=\frac{c_x^4}{4},$$

so that

$$\mathrm{Var}(D^*(t))=c_a^2 t+\frac{c_x^4}{2}-2\,\mathrm{Cov}(Q^*(0),Q^*(t))-2\,\mathrm{Cov}(c_a B_a(t),Q^*(t)), \tag{94}$$

where

$$\mathrm{Cov}(Q^*(0),Q^*(t))=\frac{c_x^4}{4}c^*(\lambda t/c_x^2),\quad t\geq 0; \tag{95}$$

see Abate and Whitt (1988, Section 2) or Whitt (2002, Theorem 5.7.11). Inserting (95) into (94) yields the first line in (65) above. To establish the second limit, we do a space-time transformation of the limit, so that the limit is the same as one of the models analyzed directly.

Let us rescale space and time so that the general result is in terms on B_a instead of $c_a B_a$ (assuming that $c_a>0$), so that we can apply the established result for the $M/GI/1$ model. (Essentially the same argument works for

$GI/M/1$.) The first step is to observe that the HT limit for the departure process $\{D^*(t)\colon t \ge 0\}$ can be written as a function $\Psi\colon \mathbb{R} \times \mathcal{D}^3 \to \mathcal{D}$ of the vector process $\{(Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, -\lambda e)\}$; i.e., by (62)

$$D^* = \Psi((Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, -\lambda e)) = Q^*(0) + c_a B_a \circ \lambda e - \psi(Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda e)(t).$$

If we replace the basic vector process $(Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, -\lambda e)$ by another that has the same distribution as a process, then the distribution of D^* will be unchanged.

By the basic time and space scaling of BM, for $c_a > 0$, the stochastic processes have equivalent distributions as follows

$$\begin{aligned}&\{Q^*(0), c_a B_a(\lambda t), c_s B_s(\lambda t), -\lambda t\}\\ &\quad \overset{\mathrm{d}}{=} c_a^2 \left\{ \frac{Q^*(0)}{c_a^2}, B_a(\lambda t/c_a^2), \frac{c_s}{c_a} B_s(\lambda t/c_a^2), -\frac{\lambda t}{c_a^2} \right\} \equiv c_a^2 \left\{ \frac{Q^*(0)}{c_a^2}, B_a(u), \frac{c_s}{c_a} B_s(u), -u \right\},\end{aligned} \tag{96}$$

where $u = \lambda t/c_a^2$. After this transformation, to describe the system at time u, the associated RBM has drift -1 and variance coefficient $1 + (c_s^2/c_a^2) = c_x^2/c_a^2$. Note that the mean of the steady-state distribution associated with the new RBM is the diffusion coefficient divided by twice the absolute value of the drift, which is $c_x^2/(2c_a^2)$. As a result, $Q^*(0)/c_a^2$ is exactly the steady-state distribution needed for the new RBM. From above, we see that

$$\begin{aligned}D^*(t) &\overset{\mathrm{d}}{=} \Psi\big(c_a^2\{Q^*(0)/c_a^2, B_a(u), (c_s/c_a)B_s(u), -u\}\big)\\ &= c_a^2 \Psi\big(\{Q^*(0)/c_a^2, B_a(u), (c_s/c_a)B_s(u), -u\}\big) \equiv c_a^2 \tilde{D}^*(u) = c_a^2 \tilde{D}^*(\lambda t/c_a^2), \quad \text{for } u = \lambda t/c_a^2,\end{aligned}$$

where $\tilde{D}^*(u) \equiv \Psi(\{Q^*(0)/c_a^2, B_a(u), (c_s/c_a)B_s(u), -u\})$, corresponding to the $M/GI/1$ model with service scv c_s^2/c_a^2. Now, let $\tilde{w}^*(t)$ denote the associated weight function in (45) with $(\mu, \gamma, \tilde{c}_x^2) = (\lambda, 1, c_x^2/c_a^2)$, so that

$$\tilde{w}^*(t) = w^*(c_a^2 \lambda t/c_x^2).$$

We now turn to the variance. By applying (45), we obtain

$$V_d^*(t) = c_a^4 \tilde{V}_d^*(u) = c_a^4 \tilde{V}_d^*(\lambda t/c_a^2) = c_a^4 \left(\tilde{w}^*(\lambda t/c_a^2) \frac{\lambda t}{c_a^2} + (1 - \tilde{w}^*(\lambda t/c_a^2)) \frac{c_s^2}{c_a^2} \frac{\lambda t}{c_a^2} \right) = w^*(\lambda t/c_x^2) c_a^2 \lambda t + (1 - w^*(\lambda t/c_a^2)) c_s^2 \lambda t,$$

which agrees with the $GI/GI/1$ formula in (65). Thus, we have proved the variance formula for $GI/GI/1$.

Finally, it remains to establish the covariance formulas. First, by comparing the two lines in (65) and recall (28), we must have

$$\mathrm{Cov}(c_a B_a(\lambda t), Q^*(t)) = (1 - w^*(\lambda t/c_x^2)) c_a^2 \lambda t.$$

Let $\tilde{B}_s(t) = -B_s(t)$, then we have

$$(Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e) \overset{\mathrm{d}}{=} (Q^*(0), c_a B_a \circ \lambda e, c_s \tilde{B}_s \circ \lambda e),$$

so

$$\mathrm{Cov}(c_a B_a(t), Q^*(t)) = \mathrm{Cov}(c_a B_a(t), \psi(Q^*(0) + c_a B_a + c_s \tilde{B}_s - e)(t)) = \mathrm{Cov}(c_a B_a(t), \psi(Q^*(0) + c_a B_a + c_s B_s - e)(t)),$$

and

$$\begin{aligned}\mathrm{Cov}(c_s B_s(\lambda t), Q^*(t)) &= \mathrm{Cov}(-c_s \tilde{B}_s(\lambda t), \psi(Q^*(0) + c_a B_a \circ \lambda e + c_s \tilde{B}_s \circ \lambda e - e)(t))\\ &= -\mathrm{Cov}(c_s B_s(\lambda t), \psi(Q^*(0) + c_a B_a \circ \lambda e + c_s B_s \circ \lambda e - e)(t)).\end{aligned}$$

By symmetry, we thus have

$$\mathrm{Cov}(c_s B_s(\lambda t), Q^*(t)) = -(1 - w^*(\lambda t/c_x^2)) c_s^2 \lambda t. \quad \square$$

References

Abate J, Whitt W (1987) Transient behavior of regulated Brownian motion I: Starting at the origin. *Adv. Appl. Probab.* 19(3):560–598.

Abate J, Whitt W (1988) The correlation functions of RBM and $M/M/1$. *Stochastic Models* 4(2):315–359.

Abate J, Whitt W (1992) The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10(1–2):5–88.

Abate J, Whitt W (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comput.* 7(1):36–43.

Aras AK, Chen C, Liu Y (2018) Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment. *Queueing Systems*, ePub ahead of print March 28, https://doi.org/10.1007/s11134-018-9575-0.

Aras K, Liu Y, Whitt W (2017b) Heavy-traffic limit for the initial content process. *Stochastic Systems* 7(1):95–142.

Asmussen S (2003) *Applied Probability and Queues*, 2nd ed. (Springer, New York).

Bandi C, Bertsimas D, Youssef N (2015) Robust queueing theory. *Oper. Res.* 63(3):676–700.

Berger AW, Whitt W (1992) The Brownian approximation for rate-control throttles and the $G/G/1/C$ queue. *J. Discrete Event Dynam. Systems* 2(1):7–60.

Bertsimas D, Nakazato D (1990) The departure process from a $GI/G/1$ queue and its applications to the analysis of tandem queues. Operations Research Center Report OR 245-91, MIT, Cambridge, MA.

Borovkov AA (1965) Some limit theorems in the theory of mass service, II. *Theor. Probab. Appl.* 10(3):375–400.

Budhiraja A, Lee C (2009) Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.* 34(1):45–56.

Burke PJ (1956) The output of a queueing system. *Oper. Res.* 4(6):699–704.

Chow S-N, Hale JK (2012) *Methods of Bifurcation Theory*, Vol. 251 (Springer, New York).

Cohen JW (1982) *The Single Server Queue*, 2nd ed. (North-Holland, Amsterdam).

Cox DR, Lewis PAW (1966) *The Statistical Analysis of Series of Events* (Methuen, London).

Daley D, Vere-Jones D (2008) *An Introduction to the Theory of Point Processes*, 2nd ed. (Springer, New York).

Daley DJ (1971) Weakly stationary point processes and random measures. *J. Roy. Statist. Soc.* 33(3):406–428.

Daley DJ (1975) Further second-order properties of certain single-server queueing systems. *Stoch. Proc. Appl.* 3(2):185–191.

Daley DJ (1976) Queueing output processes. *Adv. Appl. Prob.* 8(2):395–415.

Feller W (1971) *An Introduction to Probability Theory and Its Applications*, 2nd ed. (John Wiley & Sons, New York).

Fendick KW, Whitt W (1989) Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proc. IEEE* 71(1):171–194.

Ferng H-W, Chang J-F (2001) Departure processes of $BMAP/G/1$ queues. *Queueing Systems* 39(1):109–135.

Gamarnik D, Goldberg DA (2013) Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* 23(6):2382–2419.

Gamarnik D, Zeevi A (2006) Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Ann. Appl. Probab.* 16(1):56–90.

Hanbali AA, Mandjes M, Nazarathy Y, Whitt W (2011) The asymptotic variance of departures in critically loaded queues. *Adv. Appl. Prob.* 43(1):243–263.

Harrison JM (1985) *Brownian Motion and Stochastic Flow Systems* (John Wiley & Sons, New York).

Harrison JM, Williams RJ (1990) On the quasireversibility of a multiclass Brownian service station. *Ann. Probab.* 18(3):1249–1268.

Harrison JM, Williams RJ (1992) Brownian models of feedforward queueing networks: Quasireversibility and product-form solutions. *Ann. Appl. Prob.* 2(2):263–293.

Hautphenne S, Kerner Y, Nazarathy Y, Taylor P (2015) The intercept term of the asymptotic variance curve for some queueing output processes. *Eur. J. Oper. Res.* 242(2):455–464.

Hu JQ (1996) The departure process of the $GI/G/1$ queue and its Maclaurin series. *Oper. Res.* 44(5):810–815.

Iglehart DL, Whitt W (1970a) Multiple channel queues in heavy traffic, I. *Adv. Appl. Probab.* 2(1):150–177.

Iglehart DL, Whitt W (1970b) Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Adv. Appl. Probab.* 2(2):355–369.

Karpelevich FI, Kreinin AY (1994) *Heavy-Traffic Limits for Multiphase Queues*, Vol. 137 (American Mathematical Society, Providence, RI).

Kim S (2011a) Modeling cross correlation in three-moment four-parameter decomposition approximation of queueing networks. *Oper. Res.* 59(2):480–497.

Kim S (2011b) The two-moment three-parameter decomposition approximation of queueing networks with exponential residual renewal processes. *Queueing Systems* 68(2):193–216.

Kingman JFC (1961) The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* 57(4):902–904.

Nazarathy Y (2011) The variance of departure processes: Puzzling behavior and open problems. *Queueing Systems* 68(3–4):385–394.

Nieuwenhuis G (1989) Equivalence of functional limit theorems for stationary point processes and their Palm distributions. *Probab. Theory and Related Fields* 81(4):593–608.

O'Connell N, Yor M (2001) Brownian analogues of Burke's theorem. *Stoch. Proc. Appl.* 96(2):285–304.

Reiman MI (1984) Open queueing networks in heavy traffic. *Math. Oper. Res.* 9(3):441–458.

Ross SM (1996) *Stochastic Processes*, 2nd ed. (John Wiley & Sons, New York).

Sigman K (1995) *Stationary Marked Point Processes: An Intuitive Approach* (Chapman and Hall, New York).

Szczotka W (1990) Exponential approximation of waiting time and queue length for queues in heavy traffic. *Adv. Appl. Probab.* 22(1):230–240.

Takacs L (1962) *Introduction to the Theory of Queues* (Oxford University Press, New York).

Whitt W (1982) Approximating a point process by a renewal process: Two basic methods. *Oper. Res.* 30(1):125–147.

Whitt W (1983) The queueing network analyzer. *Bell Laboratories Tech. J.* 62(9):2779–2815.

Whitt W (1984) Departures from a queue with many busy servers. *Math. Oper. Res.* 9(4):534–544.

Whitt W (1995) Variability functions for parametric-decomposition approximations of queueing networks. *Management Sci.* 41(10):1704–1715.

Whitt W (2002) *Stochastic-Process Limits* (Springer, New York).

Whitt W, You W (2018a) Using robust queueing to expose the impact of dependence in single-server queues. *Oper. Res.* 66(1):184–199.

Whitt W, You W (2018b) Algorithms to compute the index of dispersion of a stationary point process. In preparation, Columbia University, http://www.columbia.edu/~ww2040/allpapers.html.

Whitt W, You W (2018c) A robust queueing network analyzer based on indices of dispersion. In preparation, Columbia University, http://www.columbia.edu/~ww2040/allpapers.html.

Williams RJ (1992) Asymptotic variance parameters for the boundary local times of reflected Brownian motion on a compact interval. *J. Appl. Probab.* 29(4):996–1002.

Zhang Q, Heindle A, Smirni E (2005) Characterizing the $BMAP/MAP/1$ departure process via the ETAQA truncation. *Stochastic Models* 21(2–3):821–846.