A Poisson Limit for the Departure Process from a Queue with Many Busy Servers

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

Abstract

We establish a limit theorem supporting a Poisson approximation for the departure process from a multi-server queue that tends to have many busy servers. This limit can support approximating a flow out of such a queue in a complex queueing network by an independent Poisson source. The main ideas are: (i) to scale time so that previous many-server heavy-traffic limits can be applied and (ii) for time-varying arrival-rate functions, to scale (spread out) time by a large factor about each fixed time.

Keywords: Poisson approximations, departure processes, output processes, nonhomogeneous Poisson processes, queueing networks, many-server heavy-traffic limits for queues

1. Introduction

Complex queueing systems are typically networks of queues, with arrival processes at individual queues being composed of departures and overflows from other queues, with the service-time cumulative distribution functions (cdf's) often being not nearly exponential. Thus an arrival process at an internal queue usually can not be assumed to be exactly a Poisson process; e.g., see [1]. Nevertheless, a Poisson approximation may be reasonable.

Example 1.1. *final checkout in online shopping.* Suppose that we want to develop a stochastic arrival process model for the final checkout in a complex online shopping system. Many separate people shop online until they are ready for final checkout, To illustrate, we model the checkout as the second queue in a two-queue $G_t/GI/\infty \to \cdot/GI/1$ network, in which the first queue

is an infinite-server (IS) model with a general arrival process having a time-varying arrival-rate function $\lambda(t)$, which is independent of service times that are independent and identically distributed (iid) with a general cdf F having a continuous probability density function (pdf) f with $F(t) = \int_0^t f(s)\,ds$, $t \geq 0$. The output of the IS queue is the arrival process to a final single-server (SS) checkout queue, with general service cdf, unlimited waiting room and service in order of arrival. The exact form of the departure-rate function from the IS queue is

$$\delta(t) = \int_0^\infty f(y)\lambda(t-y)\,ds, \tag{1}$$

as given in Theorem 1 of [2]; it is the same for G_t as for M_t; see §5 of [3]. In this setting we provide support for approximating the final SS queue by an $M_t/GI/1$ queue, where the arrival process is a nonhomogeneous Poisson process (NHPP) with arrival-rate function $\delta(t)$ in (1). An efficient algorithm to calculate performance measures when $\lambda(t)$ is periodic is given in [4].

For a concrete simulation, consider the stationary $GI/GI/\infty \to \cdot/GI/1$ model in which all service times are iid and the external arrival process is a renewal process. To introduce extra variability, we assume that all three GI components have the hyperexponential cdf (H_2, mixture of two exponentials) with squared coefficient of variation (scv, variance divided by the square of the mean) $c^2 = 4$ and balanced means as on p. 137 of [5]; that leaves only the mean or its reciprocal, the rate, to be specified. We let the arrival rate be $\lambda = 100$ and the service rates at the two queues be $\mu_1 = 1$ and $\mu_2 = 200$. By Little's law, these rates make the mean steady-state number of busy servers in the IS queue be 100, which we regard as moderately large scale. In actual online checkout, the mean number of busy shoppers is likely to be much larger, and the difference between the two service rates is likely to be even greater.

In this context, we suggest that the performance at the final SS queue can be approximated by the $M/H_2/1$ model, for which the mean steady-state waiting time before starting service has the Pollaczek-Khintchine (PK) formula $EW = \rho\mu_2^{-1}(1+c^2)/2(1-\rho) = 0.0125$ for $\rho = 0.50$, $\mu_2 = 200$ and $c^2 = 4$. The intuition is that, with many busy servers, the departure process from the IS queue is much like the superposition of iid renewal processes, one for each server, for which the limit is Poisson, as discussed in §9.8 of [6]. Of course, the servers do not remain busy all the time and the number of busy servers is random, varying over time, so that that representation is

only approximate. Thus, there remains something to prove for departure processes.

A simulation experiment was conducted for this example. It shows that the interarrival-time cdf at the second queue is approximately exponential with mean 0.01 and that the estimated mean wait EW is only 8% above the PK formula for M arrivals; see the appendix for more details.

We conclude this example by mentioning that part of the justification for the $M/H_2/1$ approximation with a Poisson arrival process for the SS queue is the relatively low traffic intensity at the SS queue, because the departure process from the $H_2/H_2/\infty$ IS queue with many busy servers is only approximately Poisson over a short time scale. For example, the central limit theorem for the departure process will not have the same variability parameter as for a Poisson process. As discussed in §9.8 of [6], there is different variability at different time scales. As $\rho \uparrow 1$, the ratio of the actual mean $EW(\rho)$ to the mean with Poisson arrivals increases. We found that the $M/H_2/1$ approximation for the mean EW was 27% low when the service rate at the second queue was decreased so that $\rho_2 = 0.90$. See [7] for a related superposition process example. ∎

In [8] we previously established a limit theorem supporting the Poisson approximation for the departure process in the simulated example; our purpose here is to extend the result to a larger class of models. First, for infinite-server models, we extend the result established for the $GI/Ph/\infty$ model in [9] to the $G_t/GI/\infty$ model, having a general service-time distribution (the GI) instead of Ph and from a renewal arrival process (GI) to general (allowing non-renewal) arrival process with a time-varying rate (the G_t). The proof is similar, except now we apply the two-parameter MSHT FWLLN for the $G_t/GI/\infty$ model reviewed in [10] instead of the single-parameter FWLLN for the $GI/Ph/\infty$ model in [9].

We are also interested in establishing a result that applies to models with finitely many servers, perhaps including customer abandonment and feedback. A concrete example of a closed network of two $\cdot/GI/s$ queues which could be used in this way is contained in [11]. In that model there is one SS station with state-dependent service rate and one IS station. In the same spirit, our approach provides the basis for an alternate proof of a Poisson limit for a queue with delayed feedback (which can be regarded as a $\cdot/GI/\infty$ IS queue) in [12]; they established the Poisson limit using a coupling technique.

The Poisson limit in [8] was established using martingale methods The "martingale method" means that we focus on the stochastic departure rate or intensity of the departure process and its integral, called the compensator, which depends on a specification of the history or filtration; see [13] and [14] for introductions and [15] and [16] for advanced accounts. We will establish the Poisson limit, independent of the history of the queueing system, by showing that the compensators approach a deterministic limit; e.g., see Theorem VIII.4.10 in [16] and Problem 1 on p. 360 of [15].

We have special interest in many-server queues with time-varying arrival-rate functions. To obtain useful Poisson limits for those models, we will introduce a new scaling method, spreading out time about a fixed reference time. The Poisson limit then provides support for approximating the departure process by an NHPP. For the required MSHT FWLLN's in $G_t/GI/\infty$ and $G_t/GI/s_t + GI$ models with general nonstationary arrival processes, we can apply [10, 17] and [18, 19], respectively. These limits exploit a random-measure or two-parameter framework. We present our results with minimum technicalities; we refer to those papers for the details.

In §2 we review the MSHT FWLLN in a $G_t/GI/\infty$ model and establish the required FWLLN for the departure rate process in Theorem 2.1. In §3 we establish the main result, Theorem 3.1, which provides general conditions for the desired Poisson limit in terms of associated MSHT limits. We present additional supplementary material on the simulation for Example 1.1 and a direct NHPP approximation for the departure process in an appendix, which is available from the author's website.

2. Review of the MSHT FWLLN for $G_t/GI/\infty$ Queues

We start by reviewing the MSHT FWLLN in Theorem 3.1 in [10], because we will use established properties as conditions in our new theorem for other models.

Let $\Rightarrow$ denote convergence in distribution and let $D \equiv D(I, \mathbb{R})$ be the usual Skorohod space of right-continuous real-valued functions with left limits on a subinterval I of the entire real line $\mathbb{R}$, possibly $\mathbb{R}$ itself [6, 15, 16]. In our setting with a continuous limits, convergence in the Skorohod J_1 topology is equivalent to uniform convergence over bounded subintervals of I.

We consider a sequence of queueing models indexed by n. Let the arrival process have a well-defined arrival rate for each n; i.e., let $A_n(t_1, t_2)$ be the

number of arrivals in model n in the time interval $(t_1, t_2]$ and assume that

$$E[A_n(t_1, t_2)] = n\Lambda(t_1, t_2), \quad \text{where} \quad \Lambda(t_1, t_2) \equiv \int_{t_1}^{t_2} \lambda(s)\, ds \tag{2}$$

for $-\infty < t_1 < t_2 < +\infty$, with $\equiv$ denoting equality by definition. This can be achieved by scaling (accelerating) time in a fixed arrival process. Thus, the arrival rate in model n is

$$\lambda_n(t) = n\lambda(t), \quad -\infty < t < +\infty. \tag{3}$$

As a regularity condition, we also assume that $0 \le \lambda(t) \le \lambda_U < \infty$. We furthermore assume that the system starts empty at time $-t_0 \le 0$. That avoids having to carefully treat the initial conditions, but for a way to do so, see [20]. Let $\bar{A}_n(t_1, t_2) \equiv n^{-1} A_n(t_1, t_2)$. We assume a FWLLN is valid for the arrival processes; i.e.,

$$\sup_{t_L \le t_1 < t_2 \le t_U} |\bar{A}_n(t_1, t_2) - \Lambda(t_1, t_2)| \Rightarrow 0 \quad \text{as} \quad n \to \infty$$

for all t_L and t_U with $-\infty < -t_0 \le t_L < t_U < \infty$ (weak convergence uniformly over bounded intervals).

Assumption 1 of [10] allows a general sequence of arrival processes, but they are required to satisfy a functional central limit theorem (FCLT) because the primary concern was establishing the MSHT FCLT. That FCLT condition can be weakened to having only a FWLLN, because Theorem 3.1 only requires the MSHT FWLLN conclusion. The proof of the FWLLN for the number of busy servers under the weaker FWLLN condition is not discussed in [10], but it is discussed in [14]; see Theorem 3.6 and §§3.4, 4.3, 5.2, 6.1 and 6.2.

Assumption 2 of [10] stipulates that the service times come from a single i.i.d. sequence, independent of n and the arrival processes, distributed as a random variable S having a general cdf F. In addition, we require that the cdf F have a continuous pdf f in terms of which we can write $F(t) = \int_0^t f(s)\, ds$, $t \ge 0$, for $F^c(t) \equiv 1 - F(t)$, and a failure-rate function $h(t) \equiv f(t)/F^c(t)$ that is bounded over finite intervals. In [10] the system starts empty at time 0. Without loss of generality, we assume that the system starts empty at time $-t_0 < 0$. We then can let $t_0 \to \infty$ to obtain the simple approximation formula in (1).

Let $N_n^e(t,y)$ be the number of customers in service at time t in model n that have been so for at most time y. Let $\bar{N}_n^e$ be the FWLLN-scaled version $\bar{N}_n^e(t,y) \equiv n^{-1}N_n^e(t,y)$. A variant of (3.5) and (3.7) of Theorem 3.1 of [10] then implies that

$$\sup_{t_L \le t \le t_U, y_L \le y \le y_U} |\bar{N}_n^e(t,y) - N^e(t,y)| \Rightarrow 0 \quad \text{as} \quad n \to \infty \tag{4}$$

for all t_L and t_U with $-\infty < -t_0 \le t_L < t_U < \infty$ and for all y_L and y_U with $-\infty < y_L < y_U < \infty$ (again weak convergence uniformly over bounded intervals), where

$$N^e(t,y) \equiv \int_0^y F^c(s)\lambda(t-s)\,ds. \tag{5}$$

Let $D_n(t) \equiv A_n(t) - N_n^e(t, t+t_0)$ be the associated departure counting process in model n and let $\bar{D}_n(t) \equiv n^{-1}D_n(t)$ be the fluid-scaled version. Along with (4), we also have the limit

$$\bar{D}_n \Rightarrow D \quad \text{in} \quad D([-t_0,\infty) \quad \text{as} \quad n \to \infty, \tag{6}$$

where

$$D(t) \equiv \Lambda(t) - N^e(t, t+t_0) = \int_0^{t+t_0} F(s)\lambda(t-s)\,ds, \quad t \ge -t_0. \tag{7}$$

For the new part, let $\Delta_n(t)$ be the stochastic departure rate at time t in model n. The departure rate can be expressed as a stochastic integral (which is just a random sum) via

$$\Delta_n(t) = \int_0^{t+t_0} h(y)\,d_y N_n^e(t-,y)\,dy, \quad t \ge -t_0. \tag{8}$$

As in (2.1) of [8], we use the left limit $t-$ in (8) to make $\Delta_n(t)$ be the predictable stochastic intensity with respect to the appropriate history that includes the ages of all the customers in service and the history of the arrival process at each time t; see §1.3 of [13] and [14]. That can be understood and justified by a discretization argument, dividing the interval $[-t_0, t]$ into k subintervals, doing a discrete-time analysis and then letting $k \to \infty$. A detailed proof is given in §5.2 of [17]; see Lemma 5.4.

To elaborate, $\Delta_n(t)$ being a stochastic intensity means that the centered process $D_n(t) - C_n(t)$ is a martingale with compensator

$$C_n(t) = \int_{-t_0}^t \Delta_n(s)\,ds, \quad t \ge -t_0, \tag{9}$$

again with respect to the full system history at time t.

Let $\bar{\Delta}_n \equiv n^{-1}\Delta_n$ for in (8) be the FWLLN-scaled departure rate process. We first establish a bound on the expectations.

Lemma 2.1. (*expectation bound*) *Under the assumptions above for the sequence of $G_t/GI/\infty$ models,*

$$E[\bar{\Delta}_n(t)] \le K \max\{1, t+t_0\} \sup_{0\le s\le t+t_0}\{h(s)\} < \infty \tag{10}$$

for all n and t.

Proof. Since $N_n(t) \equiv N_n^e(t,\infty) \le A_n(-t_0, t)$ we can apply (2). Since the failure rate function h is bounded over bounded intervals, we can replace it by a constant outside the integral. ∎

Theorem 2.1. (*MSHT limit for the departure rate*) *For the $G_t/GI/\infty$ model under the assumptions above,*

$$\bar{\Delta}_n \equiv n^{-1}\Delta_n \Rightarrow \delta \quad in \quad D([-t_0,\infty),\mathbb{R}) \quad as \quad n\to\infty, \tag{11}$$

where

$$\delta(t) \equiv \int_0^{t+t_0} h(y) d_y N(t,y), \quad t \ge -t_0, \tag{12}$$

so that

$$\delta(t) = \int_0^{t+t_0} f(y)\lambda(t-y)\,dy \quad and \quad D(t) = \int_{-t_0}^{t} \delta(s)\,ds, \quad t \ge -t_0. \tag{13}$$

Proof. We first apply Lemma 2.1 to get bounded expectations. Then we apply the Skorohod representation theorem, Theorem 3.2.2 of [6], to reduce the argument to a deterministic one, but use the same notation. We establish the desired uniform convergence over bounded intervals by showing, for any t in a bounded an interval and any sequence $\{t_n\}$ with $t_n \to t$ as $n\to\infty$, that $n^{-1}\Delta_n(t_n) \to \delta(t)$ as $n\to\infty$. To do that, we exploit the fact that the convergence in (4) corresponds to the weak convergence of finite measures, where we regard $\bar{N}_n^e(t,y)$ as a function of y as a cdf. Hence, we can show, for each $t \ge -t_0$ that we have the associated convergence of the integrals

$$\begin{aligned} n^{-1}\Delta_n(t_n) &= \int_0^{t_n+t_0} h(y)\, d_y \bar{N}_n^e(t_n,y) \\ &\to \int_0^{t+t_0} h(y) F^c(y)\lambda(t-y)\,dy \quad \text{as} \quad n\to\infty. \end{aligned}$$

We use the fact that h is continuous and bounded on the interval $[0, t + t_0]$. The limiting integral simplifies, yielding

$$\int_0^{t+t_0} h(y)F^c(y)\lambda(t-y)\,dy = \int_0^{t+t_0} f(y)\lambda(t-y)\,dy$$

by the simple relation $h(y)F^c(y) = f(y)$. That convergence implies that $\bar{\Delta}_n \to \delta$ in $D(\mathbb{R}, \mathbb{R})$ as $n \to \infty$, which implies the weak convergence for the original processes. ∎

Remark 2.1. (*starting empty in the distant past*) In many papers on IS queues, the system is assumed to start empty in the distant past (at $-\infty$). That is tantamount to letting $t_0 \to \infty$. As $t_0 \to \infty$, $\delta(t)$ in Theorem 2.1 approaches (1), the departure rate $E[\lambda(t - S)]$ in the $M_t/GI/\infty$ model in equation (4) of Theorem 1 in [2] and in the associated $G_t/GI/\infty$ fluid model; see §4 of [21].

3. The Supporting Limit for a Poisson Approximation

We now establish the Poisson limit for the departure process from a general $G_t/GI/\infty$ model. At the same time, we provide a framework for treating many other models. To do so, we assume some of the conclusions deduced for the $G_t/GI/\infty$ model is §2 rather than specify the detailed model. Thus, we now consider a more general multi-server queue. As before, we assume that the servers work independently in parallel having an individual remaining service-time failure rate function h. However, the queue may be in the middle of a complex network and there may be customer abandonment and feedback.

As in §2, we consider a sequence of models indexed by n in a MSHT framework. That typically means that the arrival rate is allowed to grow without bound as in (2) and if the there are finitely many servers, then that number is allowed to grow as well. We directly assume that the processes $N_n^e(t, y)$, $D_n(t)$, $C_n(t)$ and $\Delta_n(t)$ are well defined with the same meaning as in §2, but we do not fully specify the system; e.g., we do not specify the arrival process. We directly assume that the stochastic departure rate can be defined by the stochastic integral in (8) and that $D_n(t) - C_n(t)$ is a martingale with respect to the system history up to time t, where $C_n(t)$ is the compensator and is the integral of $\Delta_n(t)$ as in (9). We also assume that the limits in (4) and (8) hold, but without assuming the explicit form of the limits $N^e(t, y)$

and $D(t)$ in (5) and (7). Finally, we assume that the bound in (10) holds. Under these assumptions, we also have the conclusions of Theorem 2.1 with the limit in (12), but without the explicit limit in (13), because the same proof applies. For example, these assumptions apply to the $G_t/GI/s+GI$ model with finitely many servers and customer abandonment, for which a FWLLN was established in [21, 22].

Paralleling [8], we will do an additional slow-time scaling in order to establish the supporting Poisson limit. However, in order to capture the time-varying arrival rate appropriately, instead of simply undoing the MSHT scaling in (2), we do the time scaling about an arbitrary time t, which we regard as fixed.

For this purpose, we introduce two-parameter processes

$$\begin{aligned} D_n(t,u_2) - D_n(t,u_1) &\equiv D_n(t+u_2/n) - D_n(t+u_1/n), \\ C_n(t,u_2) - C_n(t,u_1) &\equiv C_n(t+u_2/n) - C_n(t+u_1/n), \\ \Delta_n(t,u) &\equiv \Delta_n(t+u/n)/n, \quad -\infty < u_1 < u_2 < +\infty. \end{aligned} \tag{14}$$

Note that the definitions for $C_n(t,u)$ and $\Delta_n(t,u)$ follow from the definition for $D_n(t,u)$. With these definitions and the assumptions above,

$$C_n(t,u_2) - C_n(t,u_2) = \int_{u_1}^{u_2} \Delta_n(t,v)\,dv, \quad -\infty < u_1 < u_2 < +\infty, \tag{15}$$

$\{D_n(t,s) - C_n(t,s) : s \geq u_1\}$ is a martingale and $\Delta_n(t,u)$ is a predictable stochastic intensity with respect to the system history.

With this preparation, we are able to establish our desired result. In our setting, weak convergence of the processes with nondecreasing sample paths to a Poisson process in $D(I,\mathbb{R})$ is equivalent to convergence of all finite-dimensional distributions; see VI.3.37 of [16].

Theorem 3.1. (*Poisson limit*) *Under the assumptions in this section above,*

$$D_n(t,\cdot) \Rightarrow \Pi_{\delta(t)}(\cdot) \quad \textit{in} \quad D(\mathbb{R},\mathbb{R}) \quad \textit{as} \quad n \to \infty, \tag{16}$$

where Π_c is a homogeneous Poisson process with constant rate c and $\delta(t)$ is the limit in (12)*; i.e., for any integer k, any k-tuple of disjoint subintervals $((u_{i,1},u_{i,2}] : 1 \leq i \leq k)$ and any k-tuple of nonnegative integers $(j_i : 1 \leq i \leq k)$,*

$$P\left(D_n(t,u_{i,2}) - D_n(t,u_{i,1}) = j_i : 1 \leq i \leq k\right) \to \prod_{i=1}^{k} \frac{e^{-\mu_i(t)}\mu_i(t)^{j_i}}{j_i!}$$

as $n \to \infty$, *where* $\mu_i(t) \equiv \delta(t)(u_{i,2} - u_{i,1})$.

Proof. The proof is similar to the proof of Theorem 2 in [8]. The limit in (11) implies that

$$\sup_{u_L<u<u_U} |n^{-1}\Delta_n(t + (u/n)) - \delta(t)| \Rightarrow 0 \quad \text{as} \quad n \to \infty$$

for all u_L and u_U with $-\infty < u_L < u_U < +\infty$. Then, paralleling the proof of Theorem 2 in [8], we write

$$\begin{aligned} C_n(t + (u_2/n)) - C_n(t + (u_1/n)) &= \int_{u_1/n}^{u_2/n} \Delta_n(t + v)\, dv \\ &= \int_{u_1}^{u_2} n^{-1}\Delta_n(t + v/n)\, dv \\ &\Rightarrow \int_{u_1}^{u_2} \delta(t)\, dv = \delta(t)(u_2 - u_1) \quad \text{as} \quad n \to \infty. \end{aligned} \tag{17}$$

Combining (17) with (14), we have the analog of Corollary 2 of [8], i.e.,

$$C_n(t, u_2) - C_n(t, u_1) \Rightarrow \delta(t)(u_2 - u_1) \quad \text{as} \quad n \to \infty.$$

That implies that the limit (16) holds, as claimed, by Theorem VIII.4.10 of [16]. ∎

Remark 3.1. (*supporting an NHPP approximation*) The statement of Theorem 3.1 may seem a bit paradoxical, because it states that the departure process is asymptotically a *homogeneous* Poisson process but with the *time-varying rate* $\delta(t)$ in (12). That dichotomy arises because of our scaling about the fixed time t. For applications, we interpret the limit as supporting an NHPP approximation with time-varying rate $\delta(t)$.

Remark 3.2. (*the stationary case*) For a stable stationary model without abandonment, the rate out equals the rate in, so that the departure rate must equal the constant arrival rate. Consistent with that basic property, we see that $\delta(t) = \lambda$ for all t if the arrival process has a constant arrival rate λ.

Remark 3.3. (*models with finitely many servers*) For the stationary $GI/M/s$ and the $M/M/s + M$ models, the papers [23] and [24] can be applied to establish analogs of Theorem 2.1. For the quality-and-efficiency-driven (QED)

and efficiency-driven (ED) MSHT regimes, $\delta(t) = \mu s$ for all t. The FWLLN follows immediately from the MSHT FCLTs established in those papers. These result can be extended to general arrival processes using §7.3 of [14]. Extensions to the $G/G/s$ and $G/GI/s + GI$ follow from [17, 18].

We can also apply [19] to obtain the analog of Theorem 2.1 for the $G_t/M/s_t+GI$ Model with customer abandonment, which alternates between overloaded intervals and underloaded intervals. With exponential service times, it suffices to look at $N(t)$, the number of customers in service at each time, instead of the more complicated two-parameter process $N^e(t, y)$. The departure rate at time t is simply $\mu \min\{X(t), s(t)\}$, where μ is the fixed service rate, $X(t)$ is the number of customers in the system and $s(t)$ is the number of servers at time t. The FWLLN is given for overloaded intervals in (4.2) of Theorem 4.1 and §3 of [19]; then $\delta(t) = s(t)\mu$. The FWLLN is given for underloaded intervals in (5.1) and (5.2) of Theorem 5.1 of [19]; except for the initial conditions, $\delta(t)$ is the same as in an IS system. Extensions to GI service follow from [22].

Acknowledgement. The author thanks Vahid Sarhangian for suggesting that it would be good to extend [8], Guodong Pang and an anonymous referee for helpful comments, Jingtong Zhao for conducting the supporting simulation, and NSF for research support (CMMI 1265070).

References

[1] R. L. Disney, D. Konig, Queueing networks: a survey of their random processes, SIAM Review 27 (3) (1985) 335–403.

[2] S. G. Eick, W. A. Massey, W. Whitt, The physics of the $M_t/G/\infty$ queue, Oper. Res. 41 (1993) 731–742.

[3] O. B. Jennings, A. Mandelbaum, W. A. Massey, W. Whitt, Server staffing to meet time-varying demand, Management Sci. 42 (1996) 1383–1394.

[4] N. Ma, W. Whitt, A performance algorithm for periodic queues, columbia University, working paper. (2016).

[5] W. Whitt, Approximating a point process by a renewal process: two basic methods, Oper. Res. 30 (1982) 125–147.

[6] W. Whitt, Stochastic-Process Limits, Springer, New York, 2002.

[7] K. Sriram, W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, IEEE Journal on Selected Areas in Communications SAC-4 (6) (1986) 833–846.

[8] W. Whitt, Departures from a queue with many busy servers, Mathematics of Operations Research 9 (4) (1984) 534–544.

[9] W. Whitt, On the heavy-traffic limit theorem for $GI/G/\infty$ queue, Advances in Applied Probability 14 (1) (1982) 171–190.

[10] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues, Queueing Systems 65 (2010) 325–364.

[11] E. V. Krichagina, A. A. Puhalskii, A heavy-traffic analysis of a closed queueing system with a GI/∞ service center, Queueing Systems 25 (1997) 235–280.

[12] E. A. Pekoz, N. Joglekar, Poisson traffic flow in a general feedback queue, Journal of Applied Probability 39 (2002) 630–636.

[13] P. Bremaud, Point Processes and Queues: Martingale Dynamics, Springer, New York, 1981.

[14] G. Pang, R. Talreja, W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues, Probability Surveys 4 (2007) 193–267.

[15] S. N. Ethier, T. G. Kurtz, Markov Processes: Characterization and Convergence, Wiley, New York, 1986.

[16] J. Jacod, A. N. Shiryaev, Limit Theorems for Stochastic Processes, Springer, New York, 1987.

[17] H. Kaspi, K. Ramanan, Law of large number limit for many-server queues, Ann. Appl. Prob. 20 (6) (2011) 2204–2260.

[18] W. Kang, K. Ramanan, Law of large number limit for many-server queues, Ann. Appl. Prob. 21 (1) (2010) 33–114.

[19] Y. Liu, W. Whitt, Many-server heavy-traffic limits for queues with time-varying parameters, Annals of Applied Probability 24 (1) (2014) 378–421.

[20] K. Aras, Y. Liu, W. Whitt, Heavy-traffic limit for the initial content process, columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2014).

[21] Y. Liu, W. Whitt, The $G_t/GI/s_t + GI$ many-server fluid queue, Queueing Systems 71 (2012) 405–444.

[22] Y. Liu, W. Whitt, A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading, Oper. Res. Letters 40 (2012) 307–312.

[23] S. Halfin, W. Whitt, Heavy-traffic limits for queues with many exponential servers, Operations Research 29 (3) (1981) 567–588.

[24] O. Garnett, A. Mandelbaum, M. I. Reiman, Designing a call center with impatient customers, Manufacturing and Service Oper. Management 4 (3) (2002) 208–227.

[25] K. W. Fendick, V. Saksena, W. Whitt, Dependence in packet queues, IEEE Trans Commun. 37 (1989) 1173–1183.

Appendix A. The Simulation Experiment for Online Checkout

In this first section we describe the simulation experiment related to Example 1.1.

Appendix A.1. The $H_2/H_2/200 \to \cdot/H_2/1$ Model

To provide a concrete illustration in the setting of the online checkout example in Example 1.1, we simulated the $H_2/H_2/200 \to \cdot/H_2/1$ model with external arrival rate $\lambda = 100$, $\mu_1 = 1$ and $\mu_2 = 200$ specified there. The interarrival times and service times come from three mutually independent sequences of iid random variables, each with H_2 distributions having scv $c^2 = 4$ and balanced means as on p. 137 of [5]. As a consequence, the traffic intensity at both queues is $\rho_i = 100/200 = 0.50$, $i = 1, 2$.

Because we never see all servers busy in a long simulation, even though that event necessarily has (small) positive probability in the model, the first queue behaves like an $H_2/H_2/\infty$ model. So the model illustrates both Theorems 2.1 and 3.1.

The important point for the relevance of the theory in this paper is that the service rate at the first many-server queue is much lower than the arrival rate and the service rate at the second queue. By Little's law, the mean number of busy servers at the first queue is $\lambda/\mu_1 = 100$. From [10] and earlier papers, we know that the distribution is approximately normal. Thus, this example illustrates a many-server queue that tends to have many busy servers, for which Theorem 3.1 is intended. In fact, for the online shopping example, the actual parameters λ and μ_2 are likely to be many times larger relative to μ_1, which we have taken as our unit to measure time, so that this is far from an extreme example.

Because of the H_2 distributional assumptions, the departure process from the first queue is not exactly Poisson. Nevertheless, we suggest that the steady-state performance of the second queue may be well approximated by the steady-state distribution of an $M/H_2/1$ queue with an independent Poisson arrival process, for which the mean waiting time is given by the Pollaczek-Khintchine formula

$$EW = \frac{\tau\rho(1 + c_s^2)}{2(1 - \rho)} = \frac{(0.005)(0.5)(1 + 4.0)}{2(1 - 0.5)} = 0.01250 \qquad \text{(A.1)}$$

where $\tau = 1/200 = 0.005$ is the mean service time, $\rho = 0.5$ and $c_s^2 = 4.0$. In contrast, if the arrival proess were a renewal process with an H_2 distribution, like the service-time distribution at the first queue, then a common

approximation for the waiting time would be

$$EW \approx \frac{\tau\rho(c_a^2 + c_s^2)}{2(1-\rho)} = \frac{(0.005)(0.5)(4.0+4.0)}{2(1-0.5)} = 0.02000 \qquad \text{(A.2)}$$

Our simulation estimate of the mean waiting time at the second single-seerver queue is $EW = 0.01350$, which is only 8% above the exact Poisson value.

Appendix A.2. The Experimental Design

The simulation experiment consisted of 100 iid replications of the model over the time interval $[0, 1020]$, where we used the data over $[10, 1010]$, an interval of length 1000 to estimate the interdeparture distribution. Since the arrival rate is 100, there was a total sample of $100 \times 100 \times 1000 = 10^7$ arrivals and thus essentially the same number of departures and interdeparture times. We delete initial and terminal intervals of length 10, having about $100 \times 10 = 1,000$ arrivals and departures, to avoid end effects, and to allow the departure process approach steady state (stationarity). From the very large sample size, it is evident that we should have extraordinarily high precision. (We found the answers are significantly distorted if the end effects are not properly removed; the last interdeparture times can be quite large.)

It is important to estimate the interdeparture-time variance carefully, because there is dependence among the interdeparture times. Thus, we estimate the second moment by the sample average, just like we estimate the mean. We then estimate the variance by $\hat{\sigma}^2 = \hat{m}_2 - (\hat{m}_1)^2$, where $\hat{m}_k$ is the direct estimate of the k^{th} moment as a sample average (using the fact that the mean of a sum is always the sum of the means, whether or not there is dependence).

Appendix A.3. The Interdeparture-Time Distribution

The estimated mean and variance of an interdeparture time from the first many-server queue were 0.0100 and 0.00010466, respectively, so that the estimated scv of one interval is $c_d^2 = 1.05$, which is close the Poisson value $c^2 = 1.00$. By visual comparison, the histogram of the interdeparture-time distribution matches the exponential distribution perfectly, and is very different from the corresponding H_2 distribution with scv $c^2 = 4$. We show two views of histograms of the interarrival-time distributions in Figures A.1 and A.2; then we show two views of histograms of the interdeparture-time distributions in Figures A.3 and A.4, showing that the interdeparture-time distribution is very nearly exponential.

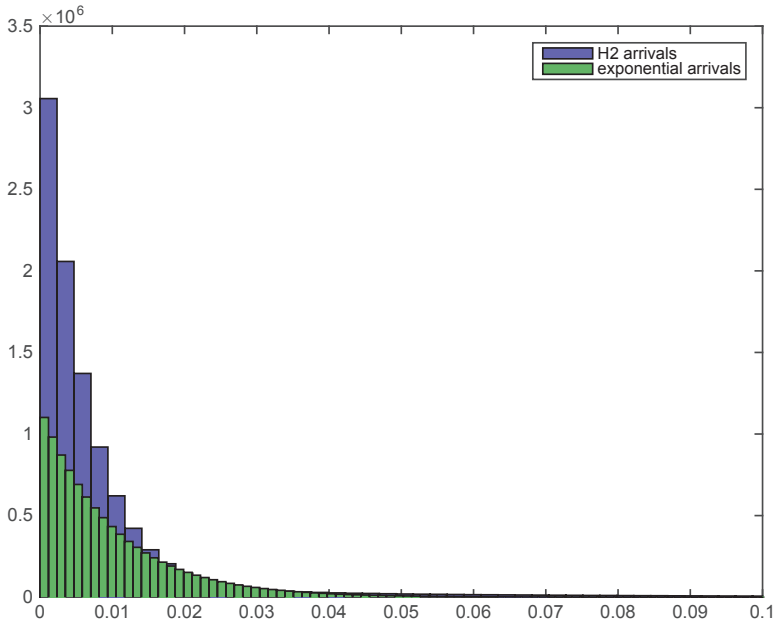

Figure A.1: First estimated interarrival-time histogram from an H_2 renewal process compared to an exponential distribution

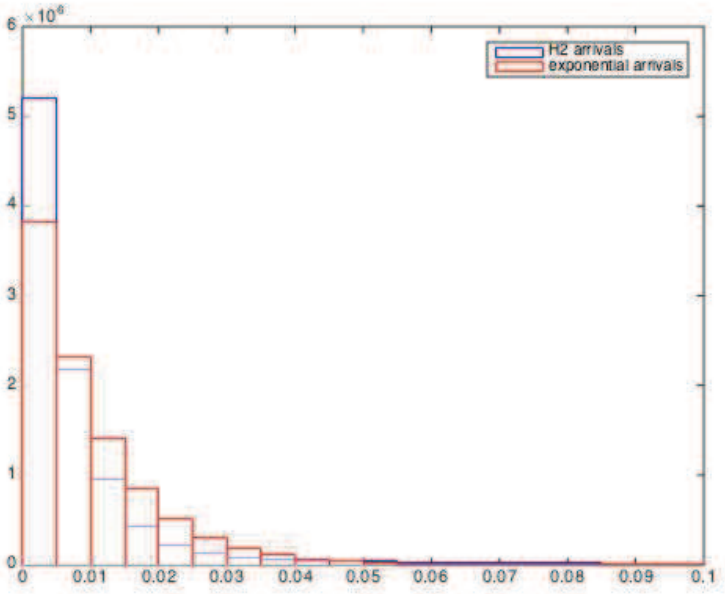

Figure A.2: Second estimated interarrival-time histogram from an H_2 renewal process compared to an exponential distribution

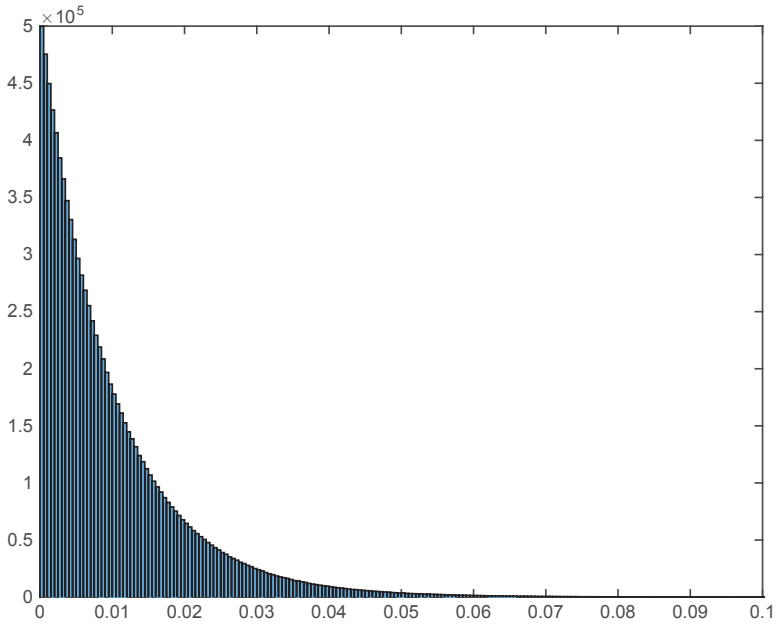

Figure A.3: First estimated interdeparture-time histogram from the $H_2/H_2/200$ model with $\lambda = 100$ compared to an exponential pdf

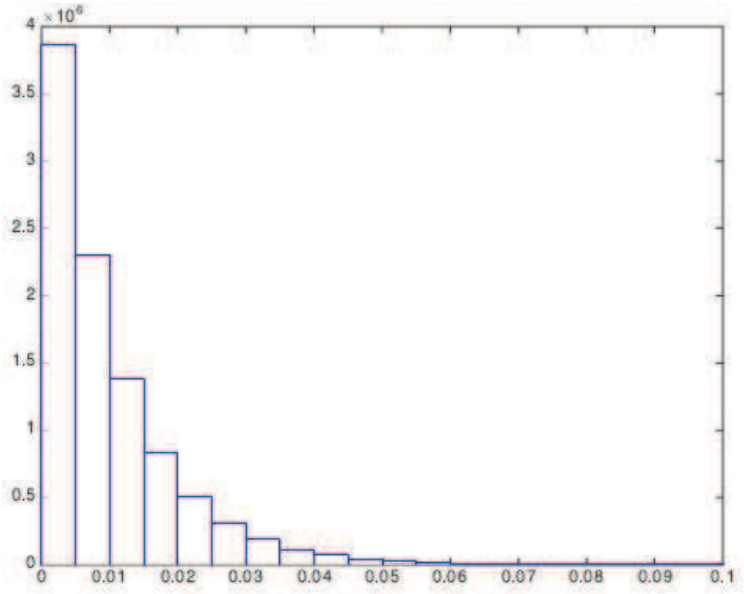

Figure A.4: Second estimated interdeparture-time histogram from the $H_2/H_2/200$ model with $\lambda = 100$ compared to an exponential pdf

Appendix A.4. Performance at the Second Single-Server Queue

The performance at the second queue is not only affected by the distribution of an interarrival time to that queue, which of course is just an interdeparture time from the previous many-server queue, but also depends on the dependence among successive interarrival times. It is important to note that there is indeed dependence among successive interdeparture times from the first queue. However, the dependence among a nearby interdeparture times tends to be quite small.

An important reference point for understanding is the conventional heavy-traffic limit theorem for the single-server queue in [6] and, in particular, the heavy-traffic bottleneck phenomenon; see Example 9.9.1 on p. 335 of [6] and the references cited there. As the traffic intensity of the second queue ρ_2 increases toward the critical value 1, the performance at the second queue will approach the performance of that queue with the external arrival process, as if the first queue were not even there; i.e., in this case it will approach (A.2) with $c_a^2 = 4$ by virtue of the heavy-traffic bottleneck phenomenon. However, the dependence among successive interdeparture times tends to be quite small; it is only the cumulative impact of all the interdeparture times that captures that heavy-traffic bottleneck phenomenon, and that dependence over many interarrival times only occurs in heavy traffic.

We estimated the mean waiting time at the second queue with $\mu_2 = 200$ and $\rho_2 = 0.50$ as $EW \approx 0.0135$, which is 0.0010 more than the 0.01250 exact value for the $M/H_2/1$ queue in (A.1); thus the Poisson approximation is 7.4% too low. When we decreased the service rate at the second queue to $\mu_2 = 140$, to obtain $\rho_2 = 0.714$, the estimated mean mean waiting time at the second queue as $EW \approx 0.0520$, that is 0.0065 more than the 0.0446 exact value for the $M/H_2/1$ queue in (A.1); thus the Poisson approximation is 12.5% too low.

In general, if we increased the scale at the first queue by multiplying the arrival rate and number of servers by 10 or 100, we would see that the Poisson approximation for the arrival process at the second queue improve. On the other hand, if we decrease the service rate μ_2 toward 100, so that the traffic intensity ρ_2 increases toward 1, then we would see the $H_2/H_2/1$ approximation at the second queue become good. In general, the departure process behaves like the superpostion of renewal processes, one for each server, for which a discussion can be found in §9.8 of [6].

To further expose the heavy-traffic effect, we also decreased the service rate at the second queue to $\mu_2 = 111.11$, to obtain $\rho_2 = 0.900$. Then the

estimated mean mean waiting time at the second queue increases to $EW \approx 0.278$, that is 0.076 more than the 0.2025 exact value for the $M/H_2/1$ queue in (A.1); thus the Poisson approximation is 27% too low. On the other hand, the simulation estimate is only 0.046 below the heavy-traffic approximation 0.324 in (A.2); the HT approximation is 16.5 too high. At $\rho = 0.9$, the heavy-traffic approximation is closer than the Poisson approximation.

However, in the onliine checkout application, we are likely to have a much external larger arrival rate and a lower traffic intensity ot the checkout queue, so the Poisson approximation is likely to be appropriate. However, the heavy-traffic approximation is likely to become relevant in overload, as in [7, 25].

The theory here and in [8] provides a useful theoretical reference point, along with the heavy-traffic limits, which cover the case in which $\rho_2 \uparrow 1$.

Appendix B. A Direct Poisson Approximation

In this section we directly develop a Poisson approximation for the departure process from a $G_t/GI/s$ many-server queue. In particular, consider a queueing model with a large number s of servers, each with i.i.d. service times, independent of the arrival process, and distributed according to a random variable S with cdf F having a continuoous pdf f with $F(t) = \int_0^t f(s)\,ds$, $t \geq 0$.

Let $N^e(t, y)$ be the number of customers in service at time t that have been so for at most time y. Let $N(t) \equiv N^e(t, \infty)$ be the total number of customers in service at time t.

Even though we have not yet defined the arrival process, we can conclude that (under regularity conditions) there should be a well defined departure rate at time t. (As one regularity condition, we assume that there is a well-defined arrival rate function, so that the probability of an arrival at any specific time is 0.) The departure rate can be expressed as a stochastic integral (which is just a random sum) via

$$\Delta(t) \equiv \int_0^\infty h(y)\, d_y N^e(t-, y), \tag{B.1}$$

where $\equiv$ denotes equality by definition, $h(t) \equiv f(t)/F^c(t)$ is the failure (or hazard) rate and $F^c(t) \equiv 1 - F(t)$. As in (2.1) of [8], we use the left limit $t-$ in (B.1) to make $\Delta(t)$ be the predictable stochastic intensity with respect to the appropriate history; see §1.3 of [13].

It should be noted that the departure rate $\Delta(t)$ in (B.1) is in general stochastic, depending on the stochastic process $\{N^e(t,y) : y \geq 0\}$, which in turn depends on the model history up to time t. Nevertheless, we propose (B.1) as a basis for a tractable deterministic approximation for the departure rate, independent of the history, in the case that the stochastic process $N^e(t,y)$ has relatively low variability, as often occurs when (i) the number of customers in service is relatively large and (ii) the service times are relatively long. In that case, we propose *approximating* the departure process be an NHPP with time-varying rate

$$\delta_{ap,1}(t) \equiv E[\Delta(t)] = \int_0^\infty h(y)\, d_y E[N^e(t-,y)]. \tag{B.2}$$

Of course, the expectation $E[N^e(t,y)]$ appearing in (B.2) is typically difficult to compute, but it readily can be estimated by simulation.

For a more elementary analytical deterministic approximation, we can exploit a MSHT FWLLN. Our main example is the $G_t/GI/\infty$ model with time-varying arrival rate function $\lambda \equiv \lambda(t)$. We can exploit the FWLLN in Theorem 3.1 of [10]. Assuming that the system started empty in the distant past, in addition to the other conditions there, that leads to the NHPP approximation with rate

$$\delta_{ap,2}(t) = \int_0^\infty h(y)\, d_y \int_0^y F^c(s)\lambda(t-s)\, ds = \int_0^\infty f(y)\lambda(t-y)\, ds \tag{B.3}$$

with the final relation in (B.3) holding because of the simple relation $h(y)F^c(y) = f(y)$. If the arrival rate is constant λ, so that we have the stationary $G/GI/\infty$ model, then $\delta(t) = \lambda$ and the approximating NHPP is homogeneous Poisson with rate λ, the same as the arrival rate. Our asymptotic results support an NHPP approximation with the time-varying rate (1) for the $G_t/GI/\infty$ model.