Operations Research Letters 44 (2016) 602–606

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

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

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form
15 May 2016
Accepted 18 June 2016
Available online 6 July 2016

Keywords:
Poisson approximations
Departure processes
Output processes
Nonhomogeneous Poisson processes
Queueing networks
Many-server heavy-traffic limits for queues

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.

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 cannot be assumed to be exactly a Poisson process; e.g., see [3]. 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 (i.i.d.) 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 [4]; it is the same for G_t as for M_t; see §5 of [9]. 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 [16].

For a concrete simulation, consider the stationary $GI/GI/\infty \to \cdot/GI/1$ model in which all service times are i.i.d. 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 in p. 137 of [21]; 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

E-mail address: ww2040@columbia.edu.

http://dx.doi.org/10.1016/j.orl.2016.06.006
0167-6377/

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 i.i.d. renewal processes, one for each server, for which the limit is Poisson, as discussed in §9.8 of [23]. Of course, the servers do not remain busy all the time and the number of busy servers is random, varying over time, so 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 [23], 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 [20] for a related superposition process example. ■

In [22] 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 [24] 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 [18] instead of the single-parameter FWLLN for the $GI/Ph/\infty$ model in [24].

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 [12]. 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 [19]; they established the Poisson limit using a coupling technique.

The Poisson limit in [22] 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 [2,17] for introductions and [5,8] 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 [8] and Problem 1 on p. 360 of [5].

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 [11,18,10,15], 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 Section 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 Section 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 (http://www.columbia.edu/~ww2040/allpapers.html).

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

We start by reviewing the MSHT FWLLN in Theorem 3.1 in [18], 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 [5,8,23]. 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

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

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 [1]. 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.,

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

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 [18] 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 [18], but it is discussed in [17]; see Theorem 3.6 and §§3.4, 4.3, 5.2, 6.1 and 6.2.

Assumption 2 of [18] 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 [18] the system starts empty at time 0. Without loss of generality, we assume that the system