Operations Research Letters 42 (2014) 458–461

Contents lists available at ScienceDirect

Operations Research Letters

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

Heavy-traffic limits for queues with periodic arrival processes

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, Mail Code 4704, S. W. Mudd Building, 500 West 120th Street, New York, NY 10027-6699, USA

ARTICLE INFO

Article history:
Received 21 June 2014
Received in revised form
31 July 2014
Accepted 4 August 2014
Available online 19 August 2014

Keywords:
Heavy-traffic limits
Queues with periodic arrival rates
Heavy-traffic limits for periodic queues

ABSTRACT

We establish conventional heavy-traffic limits for the number of customers in a $G_t/GI/s$ queue with a periodic arrival process. We assume that the arrival counting process can be represented as the composition of a cumulative stochastic process that satisfies an FCLT and a deterministic cumulative rate function that is the integral of a periodic function. We establish three different heavy-traffic limits for three different scalings of the deterministic arrival rate function. The different scalings capture the three cases in which the predictable deterministic variability (i) dominates, (ii) is of the same order as, or (iii) is dominated by the stochastic variability in the arrival and service processes.

1. Introduction

In this paper we establish heavy-traffic functional weak laws of large numbers (FWLLNs) and functional central limit theorems (FCLTs) for $G_t/GI/s$ queues with arrival processes having periodic arrival rate functions. The model has a fixed number of servers working in parallel, an unlimited waiting space and the first-come first-served service discipline. By "conventional heavy traffic", we mean that we allow the arrival rate to increase, but keep the maximum possible service rate fixed.

Conventional heavy-traffic approximations help understand the performance of complex queueing systems; see [21] for a review (especially Chapters 5 and 9). Highlights are Kingman [9] showing that the steady-state wait in a $GI/GI/1$ queue can be approximated by an exponential random variable and Iglehart and Whitt [6,7] showing that the entire waiting time (and queue length) process in a $G/G/s$ queue can be approximated by a reflected Brownian motion (RBM) with negative drift, which has an exponential steady-state distribution, where the mean of the exponential distribution and the drift and diffusion coefficients of the RBM depend on the basic rate and variability parameters of the arrival and service processes.

It is important to note that Mandelbaum and Massey [13] already developed conventional heavy-traffic approximations for queues with time-varying arrival processes. They analyzed the $M_t/M_t/1$ model and showed that the presence of time-varying arrival rate can introduce major complications; e.g., there is need for care in even properly defining a proper notion of traffic intensity; see § 3 of [13]. In [13] an elaborate theory is developed, which demonstrates both complex performance, e.g., see Figs. 3.1 and 4.1 in [13] and complex proof techniques, including the use of the Skorohod M_1 topology to treat the discontinuities evident in Fig. 4.1 of [13].

In contrast, our goal is to expose the more elementary stories that follow quite directly from [7] if we make additional simplifying assumptions. In particular, we only focus on the first-order performance. If there is an important story in the deterministic fluid approximation stemming from the FWLLN, then we focus on that FWLLN. We only consider the heavy-traffic FCLT when that provides the first-order description of performance, i.e., when the FWLLN is the same as for the model with a constant arrival rate. We then can see if the time-varying arrival rate is insignificant in the heavy-traffic limits by seeing if it plays no role in either the FWLLN or the FCLT.

The theory in [7] implies that, under regularity conditions, a heavy-traffic FWLLN (FCLT) holds for the queue length process whenever FWLLNs (FCLTs) hold for the arrival and service processes. Thus, for periodic arrival processes, previous heavy-traffic FWLLNs and FCLTs can be applied if we have an FWLLN and an FCLT for the periodic arrival process. In particular, we can apply Theorem 1 of [7] and basic continuous mapping arguments to establish conventional heavy-traffic limits for the $G_t/GI/s$ model. This important consequence of [7] no doubt has been recognized, but evidently nothing has been published.

An important role in the conventional heavy-traffic limits is played by the scaling of both time and space. Roughly, the required scaling is the same as needed for a sequence of simple random walks to converge to a Brownian motion with drift: we need to

E-mail address: ww2040@columbia.edu.

http://dx.doi.org/10.1016/j.orl.2014.08.001
0167-6377/© 2014 Elsevier B.V. All rights reserved.

scale time by some factor n and then scale space by $1/\sqrt{n}$, with the mean step being $c/\sqrt{n}$. Since the mean step is related to the traffic intensity of the queue, n should be related to the traffic intensity ρ in the queueing system by $1 - \rho_n = 1/\sqrt{n}$. The important observation is that in terms of the traffic intensity ρ the required time scaling is $(1-\rho)^{-2}$.

As we show here in Corollary 3.1, when time scaling is omitted from the deterministic arrival rate function in the standard heavy-traffic FCLT, the heavy-traffic limit with the time scaling is the same as if the periodic cycles in the periodic arrival rate function are getting shorter in the heavy-traffic limit as $\rho \uparrow 1$. As a consequence, there is still a heavy-traffic limit, but that limit is the same as if the periodic arrival rate were replaced by its long-run average. This phenomenon was first shown for the $M_t/GI/1$ model by Falin [4], but without mentioning any connection to time scaling. When the time scaling is included, the approximation stemming from the heavy-traffic FCLT is a reflection of the usual Brownian motion with drift plus a deterministic cumulative rate function associated with a periodic arrival rate function.

We are especially interested in the time scaling. In the main heavy-traffic FCLT, Theorem 3.2 here, the limit process is relatively complicated so that it is not easy to compute the approximate performance measures. It thus may be necessary to exploit simulation in order to quantify performance. Nevertheless, the heavy-traffic limits can provide useful insight into the simulations. As illustrated by [1], heavy-traffic scaling can help understand numerical performance calculations, because greater regularity is revealed. Indeed, we were motivated to establish these heavy-traffic limits in our study of gray-box modeling of queueing systems in [2,3], in which birth-and-death processes are fit to observations of a queue-length process. Specifically, the present paper arose in the study of that approach applied to periodic queues in [3].

Our approach has an important implication. By focusing on only the first-order performance, we determine when the predictable deterministic variability or the unpredictable stochastic variability dominates. We establish different heavy-traffic limits showing when the predictable deterministic variability (i) dominates, (ii) is of the same order as, or (iii) is dominated by the stochastic variability in the arrival and service processes. The more detailed analysis in [13] shows (i) that there may be different answers at different times and (ii) how to describe the refined second-order performance (diffusion approximation) for the first-order performance (deterministic fluid model) when the deterministic variability dominates.

We make two additional comments about [13]. First, since we tell only a part of the story in [13], it follows that the story can be deduced from the reasoning of [13], extended from the $M_t/M_t/1$ to $G_t/GI/s$ model, but that is a more difficult route. Second, the additional structure revealed in the more general analysis in [13] is also important for understanding the performance of queues with time-varying parameters.

In closing this introduction, we remark that the conventional heavy-traffic regime is quite different from the many-server heavy-traffic regime, which we also briefly discuss in Section 4 for comparison. Since there is no time scaling in the many-server heavy traffic regime, the many-server heavy-traffic approximations are more straightforward in engineering applications. There also is already a significant body of related literature on many-server heavy-traffic approximations for queues with time-varying arrival rates in [14,17,10,19,11]. In both settings, these results are facilitated by previous results concluding that heavy-traffic limits for the queue length depend on the arrival process through its FCLT. Thus in both cases it suffices to establish an FCLT for the arrival process with the appropriate scaling.

2. The arrival process model

We will consider periodic stochastic arrival counting processes defined by

$$A(t) \equiv N(\Lambda(t)), \quad t \geq 0, \tag{1}$$

where N is a stochastic counting process satisfying a functional central limit theorem (FCLT), i.e.,

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt] \Rightarrow c_a B_a(t) \quad \text{in } \mathcal{D} \text{ as } n \to \infty, \tag{2}$$

where $\Rightarrow$ denotes convergence in distribution in the function space $\mathcal{D}$ of right-continuous real-valued functions on the interval $[0, \infty)$ with left limits, as in [21], and B_a is a standard (drift 0, variance 1) Brownian motion (BM), while Λ is a cumulative arrival rate function, satisfying

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0, \tag{3}$$

with λ being a periodic arrival rate function, which is assumed to be integrable over finite intervals with finite long-run average

$$\bar{\lambda} \equiv \lim_{t\to\infty} t^{-1}\Lambda(t). \tag{4}$$

Throughout this paper, we assume that the cumulative arrival rate function Λ in (3) is deterministic, but it is significant that the results here can be extended to cover the case in which the arrival rate function is a stochastic process, which can be important in applications. For example, service system arrival process data often indicate overdispersion caused by day-to-day variation, as discussed in [8].

The construction in (1) is convenient for constructing non-Markov periodic arrival processes. It was suggested by [15] and also used by [5,12] and no doubt others. However, it is important to recognize that, even though it allows very general stochastic processes N, including renewal processes and much more (see § 4.4 of [21]), this model is highly structured, having all unpredictable stochastic variability associated with the process N, with its FCLT behavior captured by the single variability parameter c_a, while all the predictable deterministic variability associated with the deterministic arrival rate function λ and its associated cumulative rate function Λ. More generally, we might contemplate a time-varying variability parameter. In the present context, if the process N is a renewal counting process, then c_a is the square root of c_a^2, the squared coefficient of variation (scv, variance divided by the square of the mean) of an interarrival time. From an engineering perspective, the tractability produced by reducing the impact of the stochastic variability to the single parameter c_a^2 may be essential for drawing useful conclusions about system performance.

3. Conventional heavy-traffic limits for the $G_t/GI/s$ model

In this section we establish heavy-traffic limits for the queue-length process (number in system) in the $G_t/GI/s$ model, which has s homogeneous servers working in parallel, unlimited waiting room and customers entering service in order of arrival. We assume that the service times come from a sequence of independent and identically distributed (i.i.d.) random variables, which is independent of the arrival process. We let the mean service time be s and its scv be c_s^2. This choice of the mean makes the maximum total service rate be 1. We let the arrival process be periodic with the structure in (1)–(4).

We will construct a family of models indexed by the traffic intensity ρ and let ρ increase toward 1, its upper limit for stability. We will let the traffic intensity be determined by the deterministic arrival rate function λ, requiring that $\lambda_\rho = \rho\lambda$ for each ρ, where we