*Probability in the Engineering and Informational Sciences*, **26**, 2012, 197–220.
doi:10.1017/S0269964811000337

# INFINITE-SERVER QUEUES WITH BATCH ARRIVALS AND DEPENDENT SERVICE TIMES

Guodong Pang
*Harold and Inge Marcus Department of Industrial and Manufacturing Engineering*
*Pennsylvania State University, University Park, PA 16802*
*E-mail: gup3@psu.edu*

Ward Whitt
*Department of Industrial Engineering and Operations Research*
*Columbia University, New York, NY 10027*
*E-mail: ww2040@columbia.edu*

Motivated by large-scale service systems, we consider an infinite-server queue with batch arrivals, where the service times are dependent within each batch. We allow the arrival rate of batches to be time varying as well as constant. As regularity conditions, we require that the batch sizes be i.i.d. and independent of the arrival process of batches, and we require that the service times within different batches be independent. We exploit a recently established heavy-traffic limit for the number of busy servers to determine the performance impact of the dependence among the service times. The number of busy servers is approximately a Gaussian process. The dependence among the service times does not affect the mean number of busy servers, but it does affect the variance of the number of busy servers. Our approximations quantify the performance impact upon the variance. We conduct simulations to evaluate the heavy-traffic approximations for the stationary model and the model with a time-varying arrival rate. In the simulation experiments, we use the Marshall–Olkin multivariate exponential distribution to model dependent exponential service times within a batch. We also introduce a class of Marshall–Olkin multivariate hyperexponential distributions to model dependent hyper-exponential service times within a batch.

## 1. INTRODUCTION

This paper is a sequel to Pang and Whitt [15,16]. In [15] we established heavy-traffic (HT) stochastic-process limits for the queue length process (number in system) in the infinite-server (IS) queueing model $G_t/G^D/\infty$, having a general arrival process with time-varying arrival rate (the $G_t$) and weakly dependent (satisfying a $\phi$-mixing or $S$-mixing condition, see Berkes, Hörmann, and Schauer [1]) service times (the $G^D$). To do so, we applied functional central limit theorems (FCLTs) for sequential empirical processes (Berkes [1], Berkes and Philipp [2]) driven by dependent service times. From the HT limits, we observe that dependence among the services times does not affect the mean queue length, but it can affect the variance of queue length significantly. However, the variance formula takes a complicated form, depending on the joint bivariate distribution of each pair of service times.

In [16], we began to extract the engineering significance of the HT limit established in [15]. We showed how the variance formula can be effectively computed, and carried out computations in several specific models. In particular, we considered (i) a class of exponential autoregressive moving-average (EARMA, Jacobs and Lewis [8])-dependent exponential service times and (ii) a class of randomly repeated service (RRS) times, which allows for non-exponential-dependent service times. Both classes have a geometric decay of correlations between service times of arrivals $j$ and $j+k$ as a function of $k$. We conducted simulations to show that the heavy-traffic approximations are remarkably accurate in both stationary models and models with time-varying arrival rates.

Here we introduce a more specific $G_t/G^D/\infty$ model, motivated by an idea about how the dependence among service times should naturally arise. In particular, in many service systems there may be multiple service requests in response to a common event. As mentioned in [16], in a hospital emergency room, there may be multiple patients associated with the same medical incident. Several people may be victims of a single highway accident or food poisoning at the same restaurant. There may be rapid spread of a contagious disease. The common causes may lead to dependent service times. However, in all these examples the dependence arises in a particular way. As a first approximation, we have batch arrivals, where all dependence is confined to the service times of the customers (or jobs) within the same batch. Another example is the component ordering process in assemble-to-order systems, where components for a product are often ordered in batches and their production processes can be dependent. Thus, it is interesting to understand the impact of dependence among the service times within batches.

In order to address the more specific batch phenomenon, we introduce a new IS model, denoted as $G_t^B/G^D/\infty$, in which customers arrive in batches, where the batch sizes are i.i.d. and independent of the arrival process of batches, with all dependence among service times limited to customers in the same batch, and moreover, the bivariate distributions of each two service times within the same batch are the same. Thus, within the model the dependence among the service times is determined by two model features: (i) the batch-size distribution and (ii) the bivariate distribution of any two service times within a batch. Because we wish to consider a stationary version of the entire service-time sequence, it turns out that an important role is played by the associated batch-size stationary-excess (or equilibrium residual lifetime) distribution; see Eq. (2.1).

This new $G_t^B/G^D/\infty$ IS batch model can be regarded as a special case of the previous $G_t/G^D/\infty$ IS model, but the new structure leads to new performance formulas. For this new IS batch model, we show how the performance depends on the parameters of: (i) the arrival process of batches, (ii) the batch-size distribution, (iii) the service distribution for each customer, and (iv) the dependence assumed for the service times of the customers in the same batch. The impact of the dependence is determined by the bivariate distribution of any pair of service times in the batch, see Eqs. (2.11)–(2.15).

To illustrate our general results, we consider two special classes of dependent service times, multivariate Marshall–Olkin (MO) exponential distributions (Marshall and Olkin [12]), and newly defined multivariate MO hyperexponential distributions (Definition 5.1). For the stationary batch model, it is remarkable that the peakedness (steady-state variance divided by the mean of queue length) is (almost) linear in the single correlation parameter between any pair of service times within each batch (Proposition 3.2). Moreover, such a linearity relationship is exact when the service times within each batch have a multivariate MO exponential distribution (Proposition 5.1). For the batch model with time-varying arrival rates, we give an explicit expression for the HT approximation of the variance function in terms of mean values of the minimum of two independent and dependent service times and their associated stationary excesses (Proposition 4.1). We also give two approximations to the HT variance formula, based on a Taylor series approximation and a recent average arrival rate. We give the explicit expressions for all these HT approximations when the arrival rate is sinusoidal.

Of course, IS queues with batch arrivals have been considered before, for example, see Liu, Kashyap, and Templeton [9] and Shanbhag [17] and references therein, but it is standard to assume that the service times are mutually independent. However, there are some notable exceptions: Liu and Templeton [10] study the autocorrelation properties of an IS queueing model with multi-classes of arrivals where arrivals are modulated by a Markov renewal process and batch sizes and service times depend on the customer class, while service times are mutually independent conditional on customer class. Falin [7] considers an IS queue with Poisson arrivals of batches where each batch has a fixed number of classes that have correlated service times, independent of arrivals, while service times among different batches are independent. Our detailed model is different than the models in these previous papers. Moreover, we aim for relatively tractable formulas based on HT limits. We aim to expose the consequence of the dependence in a way that will provide insight and be more useful for engineering applications.

Here is how the rest of this paper is organized: In Section 2 we specify the $G_t^B/G^D/\infty$ model and present the HT approximation following from [15]. In Section 3 we present the HT approximations for the peakedness measure in the stationary model. In Section 4.1 we give alternative HT approximations for the mean and variance functions for the model with time-varying arrivals, and show how these expressions simplify when the arrival rate function for the batches is a sinusoidal function. In Section 4.3, we also give two approximations for these expressions based on Taylor series approximations and by applying a recent average arrival rate. In Section 4.2 and Section 4.4, we give the corresponding explicit approximation formulas when the arrival rate function is sinusoidal. In Section 5 we conduct simulations to evaluate the approximations for the batch model, including stationary models and models with sinusoidal arrival rates. We use the MO multivariate exponential distributions to model dependent exponential service times within a batch in Section 5.1. In Section 5.2, we first define the class of multivariate MO hyperexponential distributions, and then use it to model dependent hyperexponential service times within a batch. We conclude in Section 6.

## 2. THE $G_T^B/G^D/\infty$ MODEL AND ITS HT APPROXIMATION

We now introduce the batch model $G_t^B/G^D/\infty$ and specify the HT approximation following from [15].

### 2.1. The Model

The arrival process of batches is general with a time-varying arrival rate; more is mentioned below. The successive batch sizes come from a sequence $\{B_k : k \geq 1\}$ of i.i.d. random variables that is independent of the arrival process. Each random variable $B_k$ is distributed as a random variable $B$ that has probability mass function $\{p_k\}$, mean $m_B$, and variance $\sigma_B^2$, and thus SCV $c_B^2 \equiv \sigma_B^2/m_B^2$. Let the service times be independent of the arrival process. Let the service times all have the common marginal cdf $F$ with mean $m_S$, variance $\sigma_S^2$, and SCV $c_S^2 \equiv \sigma_S^2/m_S^2$. However, we now assume that the service times within a batch are dependent, while the service times in different batches are independent. Moreover, we assume that the bivariate cdf's for all pairs of customers in the same batch are identical, denoted by $H$, where $H(x,\infty) = H(\infty,x) = F(x)$.

Since we want to consider a stationary version of the service times, when we look at an arbitrary customer in steady state, we need to use the stationary-excess distribution of the batch-size distribution, that is, we need to use a new discrete random variable $B^*$ with probability mass function

$$p_k^* \equiv P(B^* = k) \equiv \frac{1}{m_B}\sum_{j=k}^{\infty} p_j, \tag{2.1}$$

which has mean

$$m_{B^*} \equiv E[B^*] = \frac{E[B^2] + m_B}{2m_B} = \frac{m_B(c_B^2 + 1) + 1}{2}. \tag{2.2}$$

See Whitt [18] for more on the batch-size stationary-excess distribution.

## 2.2. The Heavy-Traffic Limit

Following common practice for many-server HT limits (Pang, Talreja, and Whitt [14]), we consider a sequence of these $G_t^B/G^D/\infty$ models indexed by $n$ and let $n \to \infty$. In this sequence of models we only change the arrival process, letting the arrival rate be proportional to $n$. Specifically, the arrival rate in model $n$ at time $t$ is a function $n\lambda_B^*(t)$, where $\lambda_B^*(t)$ is an integrable function. Let the process $N_n \equiv \{N_n(t) : t \geq 0\}$ count the arrivals of batches in model $n$. We assume that the sequence of arrival processes of batches satisfies FCLT, that is,

$$(N_n(t) - n\Lambda_B(t))/\sqrt{n} \Rightarrow W(c_{a,B}^2 \Lambda_B(t)) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{2.3}$$

where $\Lambda_B(t)$ is the continuous function defined by

$$\Lambda_B(t) \equiv \int_0^t \lambda_B^*(s)\, ds, \qquad t \geq 0. \tag{2.4}$$

$W$ is a standard Wiener process or Brownian motion (BM) and the limit holds in $\mathcal{D} \equiv \mathcal{D}([0,\infty),\mathbb{R})$, the space of real-valued functions on the interval $[0,\infty)$ that are right continuous with left limits, (see e.g., Whitt [20]), and the variability parameter $c_{a,B}^2$ is a constant, which is the SCV of an interarrival time when the arrival processes of batches are renewal processes.

As a consequence of the assumptions above plus the results in Section 7.4 and Section 13.3 of Whitt [20], the overall arrival process $A_n \equiv \{A_n(t) : t \geq 0\}$ is

$$A_n(t) \equiv \sum_{k=1}^{N_n(t)} B_k, \qquad t \geq 0, \tag{2.5}$$

and it satisfies the FCLT

$$\frac{A_n(t) - n\Lambda(t)}{\sqrt{n}} \Rightarrow \sqrt{c_a^2} W(\Lambda(t)) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{2.6}$$

where $\Lambda(t) \equiv m_B \Lambda_B(t)$, $W(t)$ is a standard BM, and the overall arrival-process variability parameter is

$$c_a^2 \equiv m_B(c_B^2 + c_{a,B}^2). \tag{2.7}$$

Let $Q_n \equiv \{Q_n(t) : t \geq 0\}$ be the queue-length process for model $n$. Here we assume that the system starts empty at time 0. We state the HT limit theorem for $Q_n$ in the following theorem, following from Theorem 3.2 and Proposition 3.2 in [15].

THEOREM 2.1 (HT limits in the IS batch model): *In the IS batch model $G_t^B/G^D/\infty$ above,*

$$\hat{Q}_n(t) \equiv \frac{Q_n(t) - nq(t)}{\sqrt{n}} \Rightarrow \hat{Q}(t) \quad in \quad \mathcal{D} \quad as \quad n \to \infty, \tag{2.8}$$

*where*

$$q(t) = m_B \int_0^t \lambda_B^*(t-s) F^c(s), \tag{2.9}$$

$$\hat{Q}(t) = \int_0^t F^c(t-s)\sqrt{c_a^2}\, dW(\Lambda(t)) + \int_0^t \int_0^\infty \mathbf{1}(s+x>t)\, d\hat{K}(\Lambda(s),x), \tag{2.10}$$

*the process $\hat{K}(s,x)$ is a generalized Kiefer process* (Berkes and Phillipp [2])*, and the double integral in Eq.* (2.10) *is defined in the mean-square limit sense. The limit process $\hat{Q}$ is Gaussian process with mean 0 and variance function*

$$\mathrm{Var}(\hat{Q}(t)) = \int_0^t \Lambda(t-s)\Big(F^c(s) + (c_a^2 - 1)(F^c(s))^2 + \Gamma(s)\Big)\, ds, \tag{2.11}$$

*where*

$$\Gamma(s) = 2(E[B^*] - 1)(H^c(s,s) - F^c(s)^2). \tag{2.12}$$

*Approximations for IS batch models using HT limits.* When we consider an IS batch model with the time-varying arrival rate $\lambda_B(t) \approx n\lambda_B^*(t)$ for large $n$, by the FCLT above, we obtain the following HT approximation for the queue length at time $t$, $Q(t)$:

$$Q(t) \approx \mathrm{N}(m(t), v(t)), \qquad t \geq 0, \tag{2.13}$$

where $\mathrm{N}(a,b)$ is a random variable with normal distribution of mean $a$ and variance $b$,

$$m(t) = m_B \int_0^t \lambda_B(t-s) F^c(s)\, ds, \qquad t \geq 0, \tag{2.14}$$

and

$$v(t) = m_B \int_0^t \lambda_B(t-s)\Big(F^c(s) + (c_a^2 - 1)(F^c(s))^2 + \Gamma(s)\Big)\, ds, \qquad t \geq 0. \tag{2.15}$$

Moreover, for a stationary IS batch model with $\lambda_B(t) = \lambda_B$ for all $t \geq 0$, we have a normal approximation of the steady-state queue length

$$Q(\infty) \approx \mathrm{N}(m^*, v^*), \tag{2.16}$$

where

$$m^* = \lambda_B m_B m_S, \tag{2.17}$$

and

$$v^* = \lambda_B m_B \left[ m_S + (c_a^2 - 1)\int_0^\infty (F^c(s))^2\, ds + \int_0^\infty \Gamma(s)\, ds \right]. \tag{2.18}$$

The rest of this paper is primarily devoted to obtaining alternative expressions for the variance in Eq. (2.15).

## 3. PEAKEDNESS IN THE STATIONARY BATCH MODEL

As indicated in Section 3.1 of [16], it is appealing to focus on the peakedness measure, defined by the ratio of the steady-state variance and mean of the queue length for the stationary IS batch model. The peakedness has been an effective measure of burstiness caused by non-Poisson arrival processes in associated loss models, see Eckberg [4], Mark, Jagerman, and Ramamurthy [11], Massey and Whitt [13], Whitt [19], and references therein. We obtain the following proposition characterizing the peakedness measure in the stationary batch model from Eqs. (2.17) and (2.18).

PROPOSITION 3.1: *For the stationary $G^B/G^D/\infty$ batch model, the peakedness is given by*

$$z \equiv z(G^B/G^D) \equiv z(c_a^2, F, H) = 1 + (c_a^2 - 1)I_1 + I_2, \tag{3.1}$$

*where $c_a^2$ is given in Eq.* (2.7),

$$I_1 \equiv I_1(F) \equiv \frac{\int_0^\infty F^c(s)^2\,ds}{m_S}, \tag{3.2}$$

$$I_2 \equiv I_2(F,H) = \frac{2(m_{B^*} - 1)}{m_S}\int_0^\infty (H^c(s,s) - F^c(s)^2)\,ds. \tag{3.3}$$

From Proposition 3.1, we see that the peakedness depends on three parameters; $c_a^2$, $I_1$, and $I_2$. From Eq. (2.7), we see that the first parameter $c_a^2$ is the variability of the overall arrival process, which in turn depends on three parameters: the variability parameter of the arrival process of batches, $c_{a,B}^2$, the mean match size, $m_B$, and the SCV of the batch sizes, $c_B^2$. The two quantities $I_1$ and $I_2$ in Eqs. (3.1)–(3.3) depend only on the service times, so the contributions of the arrival process and the service times and the way they interact have been fully identified.

The two quantities $I_1$ and $I_2$ in Eqs. (3.1)–(3.3) depend on the marginal distribution and the bivariate joint distribution (capturing dependence) of the service times. The quantities $I_1$ and $I_2$ in (3.1) can be written as the mean of the minimum of two independent or dependent service times within a batch, that is,

$$I_1 = \frac{E[S_1 \wedge_{\text{ind}} S_2]}{m_S}, \tag{3.4}$$

$$I_2 = (m_B(c_B^2 + 1) - 1)(J_1 - I_1), \tag{3.5}$$

where

$$J_1 \equiv \frac{E[S_1 \wedge_{\text{dep}} S_2]}{m_S}, \tag{3.6}$$

where $S_1$ and $S_2$ in Eq. (3.4) are regarded as two independent service times with distribution function $F$, while $S_1$ and $S_2$ in Eq. (3.6) are understood as two different service times in the same batch of our IS batch model.

It is significant that, for this batch model, the two integral terms $I_1$ and $I_2$ appearing in the general peakedness formula, Eq. (3.1), are fully expressed in terms of only four mean values:

$$m_{B^*}, \quad m_S, \quad E[S_1 \wedge_{\text{ind}} S_2] \quad \text{and} \quad E[S_1 \wedge_{\text{dep}} S_2]. \tag{3.7}$$

This representation can be usefully exploited in model fitting from system data or simulations, because we can directly estimate the four means in Eq. (3.7).

If we choose to work with correlations or, equivalently, if we choose to work with the special bivariate cdf's in Section 4 of [16], then we can specify the single correlation $\rho$ of the bivariate cdf $H$ and then approximate $H$ by $\tilde{H}_\rho$ as in Eq. (29) in [16], that is,

$$\tilde{H}_\rho(x,y) \equiv \rho F(x \wedge y) + (1-\rho)F(x)F(y). \tag{3.8}$$

Thus, we have an analog of Proposition 3 in [16]:

Proposition 3.2: *If we approximately characterize the bivariate distribution of the service times within a batch by the bivariate cdf $\tilde{H}_\rho$ in Eq.* (3.8) *based on a specified marginal cdf F and a non-negative correlation $\rho$ for service times within a batch, then we obtain the approximation $J_1 \approx \rho + (1-\rho)I_1$, which leads to the simple formula from Eqs.* (3.5) *and* (3.6),

$$I_2 \approx 2(m_{B^*} - 1)\rho\,(1 - I_1) = (m_B(c_B^2 + 1) - 1)\rho(1 - I_1). \tag{3.9}$$

Corollary 3.1: *If, in addition to the condition of Proposition 3.2, the marginal service time is exponential, then $I_1 = 1/2$,*

$$I_2 \approx \frac{(m_B(c_B^2+1) - 1)\rho}{2} \tag{3.10}$$

*and*

$$z \approx \frac{m_B(c_{a,B}^2 + c_B^2) + 1}{2} + \rho\frac{m_B(c_B^2+1) - 1}{2}. \tag{3.11}$$

Corollary 3.2: *If, in addition to the conditions of Corollary 3.1, $c_{a,B}^2 = 1$, which occurs when the arrival process of batches is Poisson, that is, for the $M^B/M^D/\infty$ model, then the approximate peakedness in Eq.* (3.11) *simplifies to*

$$z \approx z(M^B/M^D) = 1 + \frac{m_B(c_B^2+1) - 1}{2}(1+\rho). \tag{3.12}$$

It is remarkable that in the stationary batch model, the HT approximation of peakedness, and thus the variance of steady-state queue length, is *linear* in the correlation $\rho$ between service times within a batch if the bivariate distribution of service times within a batch is approximately characterized by the bivariate cdf $\tilde{H}_\rho$ in Eq. (3.8). In fact, such a linearity relationship is exact for some special service time distributions, for example, the MO multivariate exponential distribution, see Proposition 5.1 below.

## 4. APPROXIMATIONS FOR BATCH MODELS WITH TIME-VARYING ARRIVALS

As observed in [17], formulas derived for the time-varying mean in the $M_t/GI/\infty$ model apply first to the $G_t/GI/\infty$ model and then also the more general $G_t/G^D/\infty$. Thus they also apply to our $G_t^B/G^D/\infty$ model.

### 4.1. Exact Expressions for General Time-Varying Arrival Rates

First, we review exact expressions for the time-varying mean from Eick, Massey, and Whitt [5] and then develop analogs for the time-varying variance. Theorem 1 of Eick et al. [5] gives two alternative expressions for the mean in Eq. (2.14), that is,

$$m(t) = E\left(\int_{t-S}^{t} m_B\lambda_B(u)\,du\right) = m_B E[\lambda_B(t-S_{\mathrm{e}})]m_S, \tag{4.1}$$

where $S_{\mathrm{e}}$ has the *stationary-excess cdf*

$$F_{\mathrm{e}}(x) \equiv P(S_{\mathrm{e}} \le x) \equiv \frac{1}{m_S}\int_0^x F^c(x)\,dx, \tag{4.2}$$

and $E[S_{\mathrm{e}}] = m_S(c_S^2+1)/2$.

The first formula in Eq. (4.1) expresses $m(t)$ as the integral of the arrival rate over the interval $[t-S,t]$ of random length $S$ ending at $t$. The second formula expresses $m(t)$ as the *pointwise-stationary approximation* (PSA) $\lambda_B(t)m_Bm_S$ modified by a random time shift by the stationary-excess random variable $S_{\mathrm{e}}$. Analogous to the mean function expression in Eq. (4.1), we obtain the following expression for the time-varying variance in Eq. (2.15).

PROPOSITION 4.1: *An alternative* (*exact*) *expression for the time-varying variance* (*of the HT limit*) *in* (2.15) *for the batch model $G_t^B/G^D/\infty$ is*

$$\begin{aligned} v(t) &= E[\lambda_B(t-S_{\mathrm{e}})]m_Bm_S + m_B(m_B(c_{a,B}^2+c_B^2)-1) \\ &\quad\times E[\lambda_B(t-(S_1\wedge_{\mathrm{ind}}S_2)_{\mathrm{e}})]E[S_1\wedge_{\mathrm{ind}}S_2] \\ &\quad+ m_B(m_B(c_B^2+1)-1)\Big(E[\lambda_B(t-(S_1\wedge_{\mathrm{dep}}S_2)_{\mathrm{e}})]E[S_1\wedge_{\mathrm{dep}}S_2] \\ &\quad- E[\lambda_B(t-(S_1\wedge_{\mathrm{ind}}S_2)_{\mathrm{e}})]E[S_1\wedge_{\mathrm{ind}}S_2]\Big). \end{aligned} \tag{4.3}$$

### 4.2. Exact Expressions for Sinusoidal Arrivals

In Eick, Massey, and Whitt [6] exact formulas are given for the mean with a sinusoidal arrival-rate function and in [16] we give exact formulas for the mean and variance of the IS model with sinusoidal arrival-rate function. We can construct corresponding exact formulas for the time-varying mean and variance for the batch model. Suppose the arrival rate function for batches is

$$\lambda_B(t) = \bar{\lambda}_B + \beta \sin(\gamma t), \qquad t \geq 0. \tag{4.4}$$

Theorem 4.1 of Eick et al. [6] gives the following expression for the mean:

$$m(t) = \Big(\bar{\lambda}_B + \beta\,(\sin(\gamma t)E[\cos(\gamma S_{\mathrm{e}})] - \cos(\gamma t)E[\sin(\gamma S_{\mathrm{e}})])\Big) m_B m_S. \tag{4.5}$$

Following Proposition 4.1, we obtain a corresponding exact expression for the variance function.

PROPOSITION 4.2: *An alternative* (*exact*) *expression for the time-varying variance* (*of the HT limit*) *in Eq.* (2.15) *for the batch model* $G_t^B/G^D/\infty$ *when the arrival-rate function is sinusoidal as in Eq.* (4.4) *and mean service time in* $m_S$ *is*

$$\begin{aligned} v(t) = {} & m_B m_S\Big[\bar{\lambda}_B + \beta\Big(\sin(\gamma t)E[\cos(\gamma S_{\mathrm{e}})] - \cos(\gamma t)E[\sin(\gamma S_{\mathrm{e}})]\Big)\Big] \\ & + (m_B(c_{a,B}^2 + c_B^2) - 1)m_B\Big[\bar{\lambda}_B + \beta\Big(\sin(\gamma t)E[\cos(\gamma (S_1 \wedge_{\mathrm{ind}} S_2)_{\mathrm{e}})] \\ & - \cos(\gamma t)E[\sin(\gamma (S_1 \wedge_{\mathrm{ind}} S_2)_{\mathrm{e}})]\Big)\Big]E[S_1 \wedge_{\mathrm{ind}} S_2] \\ & + (m_B(c_B^2 + 1) - 1)m_B\Big\{\Big[\bar{\lambda}_B + \beta\Big(\sin(\gamma t)E[\cos(\gamma (S_1 \wedge_{\mathrm{dep}} S_2)_{\mathrm{e}})] \\ & - \cos(\gamma t)E[\sin(\gamma (S_1 \wedge_{\mathrm{dep}} S_2)_{\mathrm{e}})]\Big)\Big]E[S_1 \wedge_{\mathrm{dep}} S_2] \\ & - \Big[\bar{\lambda}_B + \beta\Big(\sin(\gamma t)E[\cos(\gamma (S_1 \wedge_{\mathrm{ind}} S_2)_{\mathrm{e}})] \\ & - \cos(\gamma t)E[\sin(\gamma (S_1 \wedge_{\mathrm{ind}} S_2)_{\mathrm{e}})]\Big)\Big]E[S_1 \wedge_{\mathrm{ind}} S_2]\Big\}. \end{aligned} \tag{4.6}$$

### 4.3. Approximations for General Time-Varying Arrivals

We consider two types of approximations. First, we apply a Taylor series approximation in the time-varying mean and variance formulas in Eqs. (4.1) and (4.3), assuming that the arrival rate is suitably smooth and that the successive derivatives are suitably small so that the Taylor approximation is justified. Following Eq. (15) of Eick et al. [5], we obtain

$$m(t) \approx \lambda_B(t - E[S_{\mathrm{e}}])m_B m_S + \frac{\lambda_B''(t)}{2}\mathrm{Var}(S_{\mathrm{e}})m_B m_S. \tag{4.7}$$

The analog of approximation Eq. (4.7) for $v(t)$ in Proposition 4.1 is obtained by again applying a two-term Taylor series approximation to the arrival-rate function $\lambda_B(t)$

$$\begin{aligned} v(t) \approx{}& \lambda_B(t - E[S_e])m_B m_S + (m_B(c_B^2 + c_{a,B}^2) - 1)m_B\lambda_B(t - E[(S_1 \wedge_{\text{ind}} S_2)_e]) \\ &\times E[S_1 \wedge_{\text{ind}} S_2] + (m_B(c_B^2 + 1) - 1)m_B\Big(\lambda_B(t - E[(S_1 \wedge_{\text{dep}} S_2)_e]) \\ &\times E[S_1 \wedge_{\text{dep}} S_2] - \lambda_B(t - E[(S_1 \wedge_{\text{ind}} S_2)_e])E[S_1 \wedge_{\text{ind}} S_2]\Big) \\ &+ \frac{1}{2}m_B\lambda_B''(t)\text{Var}(S_e)m_S \\ &+ \frac{1}{2}m_B(m_B(c_B^2 + c_{a,B}^2) - 1)\lambda_B''(t)\text{Var}((S_1 \wedge_{\text{ind}} S_2)_e)E[S_1 \wedge_{\text{ind}} S_2] \\ &+ \frac{1}{2}m_B(m_B(c_B^2 + 1) - 1)\lambda_B''(t)\Big(\text{Var}[(S_1 \wedge_{\text{dep}} S_2)_e]E[S_1 \wedge_{\text{dep}} S_2] \\ &- \text{Var}[(S_1 \wedge_{\text{ind}} S_2)_e]E[S_1 \wedge_{\text{ind}} S_2]\Big). \end{aligned} \tag{4.8}$$

Second, we exploit the formulas and approximations for the stationary model, after replacing the time-varying arrival rate function in Eq. (2.15) by its time-varying average prior to $t$. So we obtain the following alternative approximations:

$$m(t) \approx \hat{\lambda}_B(t)m_B \int_0^\infty F^c(s)\,ds = \hat{\lambda}_B(t)m_B m_S \tag{4.9}$$

and

$$\begin{aligned} v(t) \approx \hat{\lambda}_B(t)\Big[&m_B m_S + (m_B(c_B^2 + c_{a,B}^2) - 1)m_B E[S_1 \wedge_{\text{ind}} S_2] \\ &+ (m_B(c_B^2 + 1) - 1)m_B\Big(E[S_1 \wedge_{\text{dep}} S_2] - E[S_1 \wedge_{\text{ind}} S_2]\Big)\Big], \end{aligned} \tag{4.10}$$

where

$$\hat{\lambda}_B(t) \equiv \int_0^\infty \lambda_B(t - s)\delta e^{-\delta s}\,ds, \tag{4.11}$$

with $\delta$ being a weighting factor that can be selected. A natural choice is $\delta = 1/E[S_e] = 2E[S]/E[S^2] = 2/(E[S](c_S^2 + 1))$, because $S_e$ is the random time lag and $E[S_e]$ is the approximate time lag. We remark that in the approximations for the mean and variance functions in Eqs. (4.9) and (4.10) give us a constant approximation of peakedness

$$z(t) \approx \frac{v(t)}{m(t)} = 1 + (m_B(c_B^2 + c_{a,B}^2) - 1)I_1 + I_2. \tag{4.12}$$

### 4.4. Approximations for Sinusoidal Arrivals

The corresponding approximations in Eqs. (4.7) and (4.8) for the batch model $G_t^B/G^D/\infty$ with sinusoidal arrival rate in Eq. (4.4) are given by

$$m(t) \approx \Big([\bar{\lambda}_B + \beta \sin(\gamma(t - E[S_{\rm e}]))] - \frac{1}{2}\beta\gamma^2 \sin(\gamma t)\mathrm{Var}(S_{\rm e})\Big) m_B m_S, \qquad \textbf{(4.13)}$$

and

$$\begin{aligned} v(t) \approx{} & [\bar{\lambda}_B + \beta \sin(\gamma(t - E[S_{\rm e}]))] m_B m_S \\ & + (m_B(c_B^2 + c_{a,B}^2) - 1) m_B [\bar{\lambda}_B + \beta \sin(\gamma(t - E[(S_1 \wedge_{\rm ind} S_2)_{\rm e}]))] E[S_1 \wedge_{\rm ind} S_2] \\ & + (m_B(c_B^2 + 1) - 1) m_B \Big( [\bar{\lambda}_B + \beta \sin(\gamma(t - E[(S_1 \wedge_{\rm dep} S_2)_{\rm e}]))] E[S_1 \wedge_{\rm dep} S_2] \\ & - [\bar{\lambda}_B + \beta \sin(\gamma(t - E[(S_1 \wedge_{\rm ind} S_2)_{\rm e}]))] E[S_1 \wedge_{\rm ind} S_2] \Big) \\ & - \frac{1}{2} m_B \beta \gamma^2 \sin(\gamma t) \mathrm{Var}(S_{\rm e}) m_S \\ & - (m_B(c_B^2 + c_{a,B}^2) - 1) \frac{1}{2} m_B \beta \gamma^2 \sin(\gamma t) \mathrm{Var}((S_1 \wedge_{\rm ind} S_2)_{\rm e}) E[S_1 \wedge_{\rm ind} S_2] \\ & - (m_B(c_B^2 + 1) - 1) \frac{1}{2} m_B \beta \gamma^2 \sin(\gamma t) \Big( \mathrm{Var}[(S_1 \wedge_{\rm dep} S_2)_{\rm e}] E[S_1 \wedge_{\rm dep} S_2] \\ & - \mathrm{Var}[(S_1 \wedge_{\rm ind} S_2)_{\rm e}] E[S_1 \wedge_{\rm ind} S_2] \Big). \end{aligned} \qquad \textbf{(4.14)}$$

For approximations in Eqs. (4.9) and (4.10), we replace $\hat{\lambda}_B(t)$ in Eq. (4.11) by

$$\hat{\lambda}_B(t) = \int_0^t [\bar{\lambda}_B + \beta \sin(\gamma(t - s))] \delta e^{-\delta s}\, ds. \qquad \textbf{(4.15)}$$

## 5. SIMULATION EXPERIMENTS

In this section, we conduct simulations to evaluate the approximations for the batch model. For the service times within a batch, we will exploit the MO multivariate exponential distribution and the multivariate hyperexponential distributions constructed from the MO exponential distributions. We will evaluate the heavy-traffic approximations for the stationary model and the model with time-varying arrival rates with both types of dependent service times.

### 5.1. Dependent Exponential Service Times

A concrete multivariate distribution for the service times to use in the batch model is the MO multivariate exponential distribution (Marshall and Olkin [12]). (Other forms of multivariate exponential distributions appear in Bladt and Nielsen [3] and Jacobs and Lewis [8].) The MO bivariate exponential distribution function $H(x,y)$ for the random vector $(S_1, S_2)$ is defined by

$$H^c(x,y) = P(S_1 > x, S_2 > y) = \exp(-\mu_1 x - \mu_2 y - \mu_{12}(x \vee y)), \qquad x,y \geq 0, \tag{5.1}$$

with three positive parameters $\mu_1$, $\mu_2$, and $\mu_{12}$. The marginals of $S_1$ and $S_2$ are exponential with rates $\mu_1 + \mu_{12}$, and $\mu_2 + \mu_{12}$, respectively, and the correlation between $S_1$ and $S_2$ is given by

$$\rho \equiv \rho_{S_1,S_2} = \mu_{12}/(\mu_1 + \mu_2 + \mu_{12}) \in [0,1]. \tag{5.2}$$

Note that this class of bivariate exponential distributions can have only non-negative correlation, which is all that we wish to consider.

It is significant that there are multivariate generalizations of this bivariate distribution such that each pair of random variables has this bivariate marginal distribution. We can obtain a specific bivariate distribution with exponential marginal cdf with rate $\mu$ and specified correlation $\rho$ by choosing the parameters

$$\mu_1 = \mu_2 = \mu - \mu_{12}, \qquad \mu_{12} = \frac{2\mu\rho}{1+\rho}, \tag{5.3}$$

where $\rho \in [0,1]$ is the specified correlation coefficient between each pair of service times, and $\mu$ is the rate of exponential service times. Note that for any pair $(S_1, S_2)$ of service times in a batch, $E[S_1 S_2] = E[S_1]E[S_2] + \rho\text{Var}[S_1] = \mu^{-2}(1+\rho)$. It is easy to check that $I_1 = 1/2$ and $J_1 = (1+\rho)/2$, so that Eq. (3.5) gives

$$I_2 = (m_B(c_B^2 + 1) - 1)\rho/2, \tag{5.4}$$

which is the same as the approximation given by Eq. (3.9).

Proposition 5.1: *For the $G^B/M^D/\infty$ model, where the bivariate service-time distribution is the MO distribution specified above, the exact formulas for $I_2$ and the peakedness $z$ coincide with the approximations in Corollary* 3.1. *For the $M^B/M^D/\infty$ model, the exact peakedness coincides with Corollary* 3.2.

### *5.1.1. Evaluating the Stationary Batch Model*

We can use the MO multivariate exponential algorithm to generate the MO multivariate exponential random variables with bivariate marginals in Eq. (5.1) of parameters in Eq. (5.2): (i) generate independent exponential random variables $Y_1, Y_2, \ldots, Y_n, W$, where $Y_i \sim \text{Exp}(\mu - \mu_{12})$ and $W \sim \text{Exp}(\mu_{12})$, (ii) set $X_1 = \min\{Y_1, W\}, \ldots, X_n = \min\{Y_n, W\}$. Then each $X_i \sim \text{Exp}(\mu)$ and each pair $(X_i, X_j)$ has joint cdf in Eq. (5.1) with correlation $\rho$.

We compare the heavy-traffic peakedness approximation in Eq. (3.11) with simulations for five models, where the results are shown in Table 1. In all models we choose the mean of the marginals of service times to be 1, and simulate for correlation parameter $\rho = 0, 0.25, 0.5, 0.75, 1$. We calculate the peakedness approximation using

**Table 1.** Comparison of the HT peakedness approximation in Eq. (3.11) and simulations in stationary IS batch models with dependent exponential service times

| Model/Correlation ($\rho$) | | 0 | 0.25 | 0.50 | 0.75 | 1 |
|---|---|---|---|---|---|---|
| $M^B/M^D/\infty$ | Approx. | 2 | 2.25 | 2.5 | 2.75 | 3 |
| $B \sim$ Geom(0.5) | Sim. | 2.012 | 2.258 | 2.511 | 2.760 | 3.006 |
| | $\lambda_B = 100$ | ±0.014 | ±0.017 | ±0.028 | ±0.012 | ±0.007 |
| | Sim. | 1.987 | 2.259 | 2.484 | 2.736 | 2.996 |
| | $\lambda_B = 10$ | ±0.024 | ±0.037 | ±0.013 | ±0.056 | ±0.022 |
| $M^B/M^D/\infty$ | Approx. | 10 | 12.25 | 14.5 | 16.75 | 19 |
| $B \sim$ Geom(0.1) | Sim. | 10.009 | 12.243 | 14.501 | 16.762 | 19.006 |
| | $\lambda_B = 100$ | ±0.038 | ±0.047 | ±0.062 | ±0.035 | 0.072 |
| | Sim. | 10.013 | 12.262 | 14.486 | 16.786 | 19.042 |
| | $\lambda_B = 10$ | ±0.063 | ±0.118 | ±0.101 | ±0.064 | ±0.062 |
| $M^B/M^D/\infty$ | Approx. | 6 | 7.25 | 8.5 | 9.75 | 11 |
| $B$ Mixture | Sim. | 6.027 | 7.240 | 8.497 | 9.737 | 11.098 |
| | $\lambda_B = 100$ | ±0.045 | ±0.036 | ±0.040 | ±0.084 | ±0.104 |
| | Sim. | 5.932 | 7.227 | 8.485 | 9.767 | 11.070 |
| | $\lambda_B = 10$ | ±0.107 | ±0.072 | ±0.130 | ±0.121 | ±0.156 |
| $E_2^B/M^D/\infty$ | Approx. | 5.5 | 6.75 | 8 | 9.25 | 10.5 |
| $B$ Mixture | Sim. | 5.517 | 6.747 | 7.961 | 9.249 | 10.504 |
| | $\lambda_B = 100$ | ±0.028 | ±0.095 | ±0.040 | ±0.060 | ±0.065 |
| | Sim. | 5.485 | 6.735 | 8.077 | 9.257 | 10.460 |
| | $\lambda_B = 10$ | ±0.081 | ±0.093 | ±0.076 | ±0.085 | ±0.082 |
| $H_2^B/M^D/\infty$ | Approx. | 6.667 | 7.917 | 9.167 | 10.417 | 11.667 |
| $B$ Mixture | Sim. | 6.669 | 7.920 | 9.147 | 10.403 | 11.673 |
| | $\lambda_B = 100$ | ±0.068 | ±0.071 | ±0.022 | ±0.069 | ±0.037 |
| | Sim. | 6.574 | 7.908 | 9.116 | 10.436 | 11.693 |
| | $\lambda_B = 10$ | ±0.029 | ±0.094 | ±0.043 | ±0.094 | ±0.125 |

formula in Eq. (3.11), which are listed in the row of "Approx." for each model. To estimate the peakedness at each time point, we conducted 2000 (or in some cases 3000, 5000) independent replications up to time 30, starting with an empty system. In each simulation run, we collected data over the time interval $[5, 30]$ and formed the time average. (The system tends to reach steady state in a few service times.) To estimate the halfwidth of the 95% confidence interval, we conducted four more independent simulations and used Student's $t$-distribution with three degrees of freedom. (The halfwidth is $3.183S_4/\sqrt{4}$, where $S_4$ is the sample deviation.)

In the first two models, we consider the $M^B/M^D/\infty$ model with Poisson arrivals of batches and batch size $B$ of geometric distribution, and we set the parameters for the geometric distributions to be 0.5 and 0.1, respectively. (Note that here the geometric random variables take values $1, 2, 3, \ldots$ .) So, $c_{a,B}^2 = 1$ for both models and $m_B = 2$, $\mathrm{Var}[B] = 2$, $c_B^2 = 0.5$ for the first model while $m_B = 10$, $\mathrm{Var}[B] = 90$, $c_B^2 = 0.9$ for the second model.

In the third model, we consider the $M^B/M^D/\infty$ model with Poisson arrivals of batches and batch size $B$ of a mixture distribution, where $B = 1$ w.p. 8/9 and

$B \sim \mathrm{Geom}(0.1)$ w.p. 1/9. So, $c_{a,B}^2 = 1$, $m_B = 2$, $\mathrm{Var}[B] = 18$ and $c_B^2 = 4.5$. We simulate the cases for arrival rates of batches $\lambda_B = 10, 100$ for these three models.

In the fourth model, we consider the $E_2^B/M^D/\infty$ model where the arrival process of batches follows a renewal process with interarrival times of an $E_2$ distribution, and the batch size $B$ of a mixture distribution as in the third model. We simulate the cases of $E_2$ distribution of parameters equal to 20 and 200, so that the arrival rates of batches $\lambda_B = 10, 100$ and $c_{a,B}^2 = 0.5$.

In the fifth model, we consider the $H_2^B/M^D/\infty$ model where the arrival process of batches follows a renewal process with interarrival times of an $H_2$ distribution, and the batch size $B$ of a mixture distribution as in the third model. We simulate the cases of $H_2$ distribution of two parameter sets: a mixture of exponential of rate 5 w.p. 0.25 and exponential of rate 15 w.p. 0.75, and a mixture of exponential of rate 50 w.p. 0.25 and exponential of rate 150 w.p. 0.75, so that the arrival rates of batches are equal to 10 and 100, respectively, and $c_{a,B}^2 = 5/3$ for both parameter sets.

We observe that in all the five models with various parameter sets and correlation values, the HT approximation of the peakedness in Eq. (3.11) is remarkably accurate, even when the arrival rates are relatively small, equal to 10. The halfwidths of the confidence intervals of all estimates are approximately 1%.

### *5.1.2. Evaluating the Batch Model with Sinusoidal Arrival Rates*

With MO mutlivariate exponential distributions for the service times within each batch, we can write down explicit expressions for the variance formulas in Eqs. (4.3) and (4.6) in Propositions 4.1 and 4.2. Both $S_1 \wedge_{\mathrm{ind}} S_2$ and $(S_1 \wedge_{\mathrm{ind}} S_2)_{\mathrm{e}}$ are exponential with mean $m_S/2$. By Eqs. (5.1) and (5.3), both $S_1 \wedge_{\mathrm{dep}} S_2$ and $(S_1 \wedge_{\mathrm{dep}} S_2)_{\mathrm{e}}$ are exponential with mean $(1+\rho)m_S/2$. Note that for an exponential random variable $S$ with mean $m_S$,

$$E[\sin(\gamma S)] = E[\sin(\gamma S_{\mathrm{e}})] = \frac{\gamma m_S}{1+\gamma^2 m_S^2}, \tag{5.5}$$

$$E[\cos(\gamma S)] = E[\cos(\gamma S_{\mathrm{e}})] = \frac{1}{1+\gamma^2 m_S^2}. \tag{5.6}$$

Thus, by Proposition 4.2, we obtain the following corollary for the explicit formula of variance.

COROLLARY 5.1: *An alternative* (*exact*) *expression for the time-varying variance* (*of the HT limit*) *in Eq.* (2.15) *for the batch model $G_t^B/M^D/\infty$ when the arrival-rate function is sinusoidal as in Eq.* (4.4) *and service times within each batch have MO mutlivariate exponential distribution with mean $m_S$ and correlation $\rho$ is*

$$v(t) = m_B m_S[\bar{\lambda}_B + \beta(1+\gamma^2 m_S^2)^{-1}(\sin(\gamma t) - \gamma m_S \cos(\gamma t))]$$
$$+ \frac{1}{2}(m_B(c_{a,B}^2 + c_B^2) - 1)m_B m_S[\bar{\lambda}_B + \beta(1+\gamma^2 m_S^2/4)^{-1}$$

$$
\begin{aligned}
&\times (\sin(\gamma t) - (\gamma m_S/2)\cos(\gamma t))] + \frac{1}{2}(m_B(c_B^2 + 1) - 1)m_B m_S \\
&\times \{(1+\rho)[\bar{\lambda}_B + \beta(1 + \gamma^2(1+\rho)^2 m_S^2/4)^{-1} \\
&\times (\sin(\gamma t) - (\gamma(1+\rho)m_S/2)\cos(\gamma t))] \\
&- [\bar{\lambda}_B + \beta(1 + \gamma^2 m_S^2/4)^{-1}(\sin(\gamma t) - (\gamma m_S/2)\cos(\gamma t))]\}. \qquad \textbf{(5.7)}
\end{aligned}
$$

Moreover, approximations in Eqs. (4.14) and (4.10) become

$$
\begin{aligned}
v(t) \approx{}& [\bar{\lambda}_B + \beta\sin(\gamma(t - m_S))]m_B m_S \\
&+ (m_B(c_B^2 + c_{a,B}^2) - 1)[\bar{\lambda}_B + \beta\sin(\gamma(t - m_S/2))]m_B m_S/2 \\
&+ (m_B(c_B^2 + 1) - 1)\Big([\bar{\lambda}_B + \beta\sin(\gamma(t - (1+\rho)m_S/2))](1+\rho) \\
&- [\bar{\lambda}_B + \beta\sin(\gamma(t - m_S/2])))]\Big)m_B m_S/2 \\
&- \frac{1}{2}\beta\gamma^2\sin(\gamma t)m_B m_S^3 - \frac{1}{16}(m_B(c_B^2 + c_{a,B}^2) - 1)\beta\gamma^2\sin(\gamma t)m_B m_S^3 \\
&- \frac{1}{16}(m_B(c_B^2 + 1) - 1)\beta\gamma^2\sin(\gamma t)m_B\Big((1+\rho)^3 m_S^3 - m_S^3\Big) \qquad \textbf{(5.8)}
\end{aligned}
$$

and

$$
v(t) \approx \hat{\lambda}_B(t)m_B m_S\left(\frac{m_B(c_B^2 + c_{a,B}^2) + 1}{2} + \rho\frac{m_B(c_B^2 + 1) - 1}{2}\right), \qquad \textbf{(5.9)}
$$

where $\hat{\lambda}_B(t)$ is given by Eq. (4.15) with $E[S_e]$ equal to $m_S$.

To evaluate the above approximations, we conducted simulations for the model $M_t^B/M^D/\infty$ with the following parameter values for the $\lambda_B(t)$ function

$$
\bar{\lambda}_B = 80, \qquad \beta = 20, \qquad \gamma = 0.5, \qquad \rho = 0.25
$$

and with the batch size distribution being a mixture as in the third model in Table 1.

We plot the simulated variance for an experiment with 5000 replications for this model, and compare with the heavy-traffic exact variance formula in Eq. (5.7), and the Taylor series approximation in Eq. (5.8) and the recent-average-arrival-rate approximation in Eq. (5.9), denoted as "HT exact," "Approx.1," and "Approx.2", respectively, in Figure 1. We remark that the Taylor series approximation will not be as good when the frequency parameter $\gamma$ is large, as shown by extensive simulations in Section 7 of [16].

## 5.2. Dependent Hyperexponential Service Times

In this section, we consider the model $G^B/H_k^D/\infty$ with batch arrivals and dependent hyperexponential ($H_k$, mixture of $k$-exponentials) service time distributions.

**TABLE 2.** Comparison of the HT peakedness approximation in Proposition 5.2, the HT peakedness approximation based on correlation in (3.9), and simulations in stationary IS batch models with dependent MO $H_2$ service times

| Model | $\lambda_B$ | Corr. | HT Approx. | Sim. | 95% CI | HT Approx. based on $\rho$ |
|---|---|---|---|---|---|---|
| $M^B/H_2^D/\infty$ | 100 | 0.521 | 11.620 | 11.627 | ±0.068 | 11.325 |
| $B \sim$ Geom(0.1) | 10 | | | 11.690 | ±0.109 | |
| $M^B/H_2^D/\infty$ | 100 | 0.360 | 2.136 | 2.137 | ±0.006 | 2.146 |
| $B \sim$ Geom(0.5) | 10 | | | 2.137 | ±0.013 | |
| $M^B/H_2^D/\infty$ | 100 | 0.248 | 6.506 | 6.508 | ±0.030 | 6.509 |
| $B$ mixture | 10 | | | 6.531 | ±0.063 | |
| $E_2^B/H_2^D/\infty$ | 100 | 0.168 | 6.047 | 6.044 | ±0.060 | 6.071 |
| $B$ mixture | 10 | | | 6.063 | ±0.103 | |
| $H_2^B/H_2^D/\infty$ | 100 | 0.632 | 8.897 | 8.891 | ±0.040 | 8.993 |
| $B$ mixture | 10 | | | 8.847 | ±0.093 | |

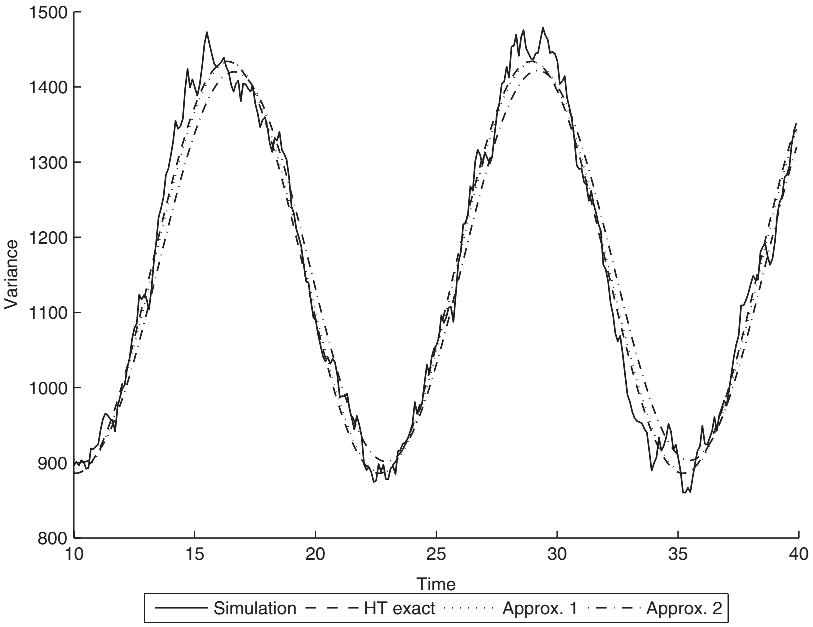

**FIGURE 1.** Comparison of simulated variance and its approximation functions in the $M_t^B/M^D/\infty$ queue with $B$ as a mixture, $B = 1$ w.p. 8/9 and $B \sim$ Geom(0.1) w.p. 1/9. $m_S = 1$ and $\rho = 0.25$. $\lambda_B(t) = 80 + 20\sin(0.5t)$.

We first define a class of dependent $H_k$ distributions from MO multivariate exponential distributions.

DEFINITION 5.1: *A vector of random variables* $(X_1, \ldots, X_n)$ *is said to have a* multivariate Marshall–Olkin hyperexponential distribution, *denoted by* MO $\mathrm{H_k}$, *if each* $X_i$ *has a hyperexponential marginal distribution,* $H_k(\alpha_1, \mu_1, \ldots, \alpha_k, \mu_k)$, *with cdf*

$$F(x) = \alpha_1 F_1(x) + \cdots + \alpha_k F_k(x), \qquad x \geq 0, \tag{5.10}$$

*where* $\alpha_i \in [0,1]$ *such that* $\alpha_1 + \cdots + \alpha_k = 1$, $F_i$*'s are the cdf's of exponential random variables with rate* $\mu_i$, $i = 1, \ldots, k$, *and each pair of* $X_i$ *and* $X_j$ *has a joint cdf*

$$H(x,y) = \sum_{i=1}^{k} \alpha_i H_i(x,y), \qquad x, y \geq 0, \tag{5.11}$$

*where* $H_i(x,y)$ *is the MO bivariate exponential distributions with parameters*

$$\mu_i^{(1)} = \mu_i^{(2)} = \mu_i - \mu_i^{(12)}, \qquad \mu_i^{(12)} = \frac{2\mu_i \rho_i}{1+\rho_i}, \qquad \rho_i \in [0,1]. \tag{5.12}$$

This class of multivariate MO $H_k$ distributions can be easily generated by adopting the algorithm to generate multivariate MO exponential random variables, that is, with probability $\alpha_i$, we generate a vector $(X_1, \ldots, X_n)$ with the multivariate MO exponential distributions. It is easy to check that each pair of random variables $(X_i, X_j)$ in Definition 5.1 has a common correlation

$$\rho = \frac{\sum_{i=1}^{k} \alpha_i (1+\rho_i) \mu_i^{-2} - \mu^{-2}}{2\sum_{i=1}^{k} \alpha_i \mu_i^{-2} - \mu^{-2}}, \tag{5.13}$$

where $\mu \equiv (\sum_{i=1}^{k} \alpha_i \mu_i^{-1})^{-1}$.

Now we assume that the service times for each batch follow a MO $H_k$ distribution with marginals $H_k(\alpha_1, \mu_1, \ldots, \alpha_k, \mu_k)$ and the joint cdf of any two service times in a batch is $H(x,y)$ in Eq. (5.11). Then the service times have mean $m_S \equiv \sum_{i=1}^{k} \alpha_i \mu_i^{-1} = \mu^{-1}$ and variance $\sigma_S^2 \equiv 2\sum_{i=1}^{k} \alpha_i \mu_i^{-2} - \mu^{-2}$, and the correlation between any pair of service times within a batch is given by Eq. (5.14), that is,

$$\rho = 1 - \left( \sum_{i=1}^{k} \alpha_i (1-\rho_i) \mu_i^{-2} \right) \Bigg/ \sigma_S^2. \tag{5.14}$$

Note that $\rho$ can take values in $[0,1]$. When $\alpha_i = 1$ for some $i$, $\rho = \rho_i \in [0,1]$.

PROPOSITION 5.2: *For the stationary $G^B/H_k^D/\infty$ batch model with the service times within a batch as a MO $H_k$ distribution in Definition 5.1, the peakedness is given by Eq.* (3.1) *in Proposition 3.1, where*

$$I_1 = m_S^{-1} \sum_{i,j=1,\dots,k} \frac{\alpha_i\alpha_j}{\mu_i+\mu_j}, \tag{5.15}$$

*and*

$$I_2 = \left(m_B(c_B^2+1)-1\right) m_S^{-1} \left[\sum_{i=1}^{k} \frac{\alpha_i(1+\rho_i)}{2\mu_i} - \sum_{i,j=1,\dots,k} \frac{\alpha_i\alpha_j}{\mu_i+\mu_j}\right]. \tag{5.16}$$

*5.2.1. Evaluating the stationary batch model* We conduct simulations to compare the HT peakedness approximation in Proposition 5.2 with simulation for five models, and also with the peakedness approximation based solely on correlations in Eq. (3.9). These results are shown in Table 2. We follow the same procedure for the estimation of peakedness and the 95% confidence intervals as in Section 5.1.1. It is remarkable that in all models, the approximations of peakedness based on correlations in Eq. (3.9) are very close to the HT approximation in Proposition 5.2 and the simulation results, all within 1–3% errors.

In the first model, we consider the $M^B/H_2^D/\infty$ queue with Poisson arrivals of batches and batch size $B$ of geometric distribution of parameter 0.1. We set the following parameter values for the $H_2$ distribution, $\alpha_1 = 0.1, \mu_1 = 1/9, \mu_2 = 9, \rho_1 = 0.1, \rho_2 = 0.9$. Simple calculation gives us $m_S = 1$, $\sigma_S^2 = 15.222$, and $\rho = 0.521$.

In the second model, we consider the same model as the first, but with the following different parameter values. The batch size $B$ is geometric with parameter 0.5, and the $H_2$ distribution has $\alpha_1 = 0.5, \mu_1 = 1, \mu_2 = 10, \rho_1 = 0.1, \rho_2 = 0.5$. We have $m_S = 0.55$, $\sigma_S^2 = 0.708$, and $\rho = 0.360$.

In the third model, we consider the $M^B/H_2^D/\infty$ queue with the batch size $B$ of a mixture distribution as in the third model of Section 5.1.1, and the $H_2$ distribution of $\alpha_1 = 0.8, \mu_1 = 1, \mu_2 = 100, \rho_1 = 0.1, \rho_2 = 0.6$. We have $m_S = 0.802$, $\sigma_S^2 = 0.957$, and $\rho = 0.248$.

In the fourth model, we consider the $E_2^B/H_2^D/\infty$ queue with a renewal arrival process with the same parameters as in the fourth model of Section 5.1.1, and the batch size $B$ is of the same mixture distribution as above, and the $H_2$ distribution has $\alpha_1 = 0.9$, $\mu_1 = 1, \mu_2 = 10, \rho_1 = 0.1, \rho_2 = 0.3$. we have $m_S = 0.910$, $\sigma_S^2 = 0.974$, and $\rho = 0.168$.

In the fifth model, we consider the $H_2^B/H_2^D/\infty$ queue with a renewal arrival process with the same parameters as in the fifth model of Section 5.1.1, and the batch size $B$ is of the same mixture distribution as above, and the $H_2$ distribution has $\alpha_1 = 0.6, \mu_1 = 0.5, \mu_2 = 10, \rho_1 = 0.5, \rho_2 = 0.2$. We have $m_S = 1.240$, $\sigma_S^2 = 3.270$, and $\rho = 0.632$.

*5.2.2. Evaluating the batch model with sinusoidal arrival rates* First, we note that for a random variable $S$ with the $H_k$ distribution in Eq. (5.10), its stationary excess $S_e$ also has a $H_k(\beta_1, \mu_1, \ldots, \beta_k, \mu_k)$ distribution, where $\beta_i = (\alpha_i/\mu_i)/m_S \in [0,1]$, $i = 1, \ldots, k$. Thus,

$$E[S_e] = m_S^{-1} \sum_{i=1}^{k} \alpha_i \mu_i^{-2}, \qquad \mathrm{Var}[S_e] = 2m_S^{-1} \sum_{i=1}^{k} \alpha_i \mu_i^{-3} - E[S_e]^2. \tag{5.17}$$

Moreover,

$$E[\sin(\gamma S)] = \sum_{i=1}^{k} \frac{\gamma \alpha_i/\mu_i}{1+\gamma^2/\mu_i^2}, \qquad E[\cos(\gamma S)] = \sum_{i=1}^{k} \frac{\alpha_i}{1+\gamma^2/\mu_i^2}, \tag{5.18}$$

$$E[\sin(\gamma S_e)] = \sum_{i=1}^{k} \frac{\gamma \beta_i/\mu_i}{1+\gamma^2/\mu_i^2} = m_S^{-1} \sum_{i=1}^{k} \frac{\gamma \alpha_i/\mu_i^2}{1+\gamma^2/\mu_i^2} \tag{5.19}$$

and

$$E[\cos(\gamma S_e)] = \sum_{i=1}^{k} \frac{\beta_i}{1+\gamma^2/\mu_i^2} = m_S^{-1} \sum_{i=1}^{k} \frac{\alpha_i/\mu_i}{1+\gamma^2/\mu_i^2}. \tag{5.20}$$

If two independent random variables $S_1$ and $S_2$ have the $H_k$ distribution in Eq. (5.10), then their minimum $S_1 \wedge_{\mathrm{ind}} S_2$ has

$$E[S_1 \wedge_{\mathrm{ind}} S_2] = \sum_{i,j=1}^{k} \frac{\alpha_i \alpha_j}{\mu_i + \mu_j}, \qquad E[(S_1 \wedge_{\mathrm{ind}} S_2)^2] = \sum_{i,j=1}^{k} \frac{2\alpha_i \alpha_j}{(\mu_i + \mu_j)^2}, \tag{5.21}$$

$$E[(S_1 \wedge_{\mathrm{ind}} S_2)^3] = \sum_{i,j=1}^{k} \frac{6\alpha_i \alpha_j}{(\mu_i + \mu_j)^3}, \tag{5.22}$$

and its stationary excess $(S_1 \wedge_{\mathrm{ind}} S_2)_e$ has

$$E[(S_1 \wedge_{\mathrm{ind}} S_2)_e] = \frac{E[(S_1 \wedge_{\mathrm{ind}} S_2)^2]}{2E[S_1 \wedge_{\mathrm{ind}} S_2]}, \qquad E[(S_1 \wedge_{\mathrm{ind}} S_2)_e^2] = \frac{E[(S_1 \wedge_{\mathrm{ind}} S_2)^3]}{3E[S_1 \wedge_{\mathrm{ind}} S_2]}. \tag{5.23}$$

Moreover, we have

$$E[\sin(\gamma(S_1 \wedge_{\mathrm{ind}} S_2))] = \sum_{i,j=1}^{k} \frac{\gamma \alpha_i \alpha_j/(\mu_i + \mu_j)}{1+\gamma^2/(\mu_i + \mu_j)^2}, \tag{5.24}$$

$$E[\cos(\gamma(S_1 \wedge_{\mathrm{ind}} S_2))] = \sum_{i,j=1}^{k} \frac{\alpha_i \alpha_j}{1+\gamma^2/(\mu_i + \mu_j)^2}, \tag{5.25}$$

$$E[\sin(\gamma(S_1 \wedge_{\mathrm{ind}} S_2)_e)] = \frac{1}{E[S_1 \wedge_{\mathrm{ind}} S_2]} \sum_{i,j=1}^{k} \frac{\gamma \alpha_i \alpha_j/(\mu_i + \mu_j)^2}{1+\gamma^2/(\mu_i + \mu_j)^2} \tag{5.26}$$

and

$$E[\cos(\gamma(S_1 \wedge_{\text{ind}} S_2)_{\text{e}})] = \frac{1}{E[S_1 \wedge_{\text{ind}} S_2]} \sum_{i,j=1}^{k} \frac{\alpha_i\alpha_j/(\mu_i+\mu_j)}{1+\gamma^2/(\mu_i+\mu_j)^2}. \qquad \textbf{(5.27)}$$

In the calculations, we use the identities that $\int_0^\infty \sin(\gamma x)e^{-\mu x}\,dx = (\gamma/\mu^2)/(1+\gamma^2/\mu^2)$ and $\int_0^\infty \cos(\gamma x)e^{-\mu x}\,dx = (1/\mu)/(1+\gamma^2/\mu^2)$, and also the formula to calculate expectations for functions of non-negative random variables $X$, $E[g(X)] = g(0) + \int_0^\infty g'(x)P(X>x)\,dx$ for any differentiable real-valued function $g$.

If two random variables $S_1$ and $S_2$ are dependent with joint distribution function $H$ in Eq. (5.11), then their minimum $S_1 \wedge_{\text{dep}} S_2$ has a $H_k(\alpha_1, 2\mu_1/(1+\rho_1), \ldots, \alpha_k, 2\mu_k/(1+\rho_k))$ distribution, and its stationary excess $(S_1 \wedge_{\text{dep}} S_2)_{\text{e}}$ has a $H_k(\hat{\alpha}_1, 2\mu_1/(1+\rho_1), \ldots, \hat{\alpha}_k, 2\mu_k/(1+\rho_k))$ distribution with $\hat{\alpha}_i = (\alpha_i(1+\rho_i)/(2\mu_i))(\sum_{i=1}^k \alpha_i(1+\rho_i)/(2\mu_i))^{-1}$, $i = 1, \ldots, k$. Thus,

$$E[S_1 \wedge_{\text{dep}} S_2] = \sum_{i=1}^{k} \frac{\alpha_i(1+\rho_i)}{2\mu_i},$$

$$\text{Var}[S_1 \wedge_{\text{dep}} S_2] = \sum_{i=1}^{k} \frac{\alpha_i(1+\rho_i)^2}{2\mu_i^2} - E[S_1 \wedge_{\text{dep}} S_2]^2. \qquad \textbf{(5.28)}$$

$$E[(S_1 \wedge_{\text{dep}} S_2)_{\text{e}}] = \sum_{i=1}^{k} \frac{\hat{\alpha}_i(1+\rho_i)}{2\mu_i} = E[S_1 \wedge_{\text{dep}} S_2]^{-1} \sum_{i=1}^{k} \frac{\alpha_i(1+\rho_i)^2}{4\mu_i^2}, \qquad \textbf{(5.29)}$$

$$\text{Var}[(S_1 \wedge_{\text{dep}} S_2)_{\text{e}}] = \sum_{i=1}^{k} \frac{\hat{\alpha}_i(1+\rho_i)^2}{2\mu_i^2} - E[(S_1 \wedge_{\text{dep}} S_2)_{\text{e}}]^2. \qquad \textbf{(5.30)}$$

Moreover, we have

$$E[\sin(\gamma(S_1 \wedge_{\text{dep}} S_2))] = \sum_{i=1}^{k} \frac{\gamma\alpha_i(1+\rho_i)/(2\mu_i)}{1+\gamma^2(1+\rho_i)^2/(2\mu_i)^2}, \qquad \textbf{(5.31)}$$

$$E[\cos(\gamma(S_1 \wedge_{\text{dep}} S_2))] = \sum_{i=1}^{k} \frac{\alpha_i}{1+\gamma^2(1+\rho_i)^2/(2\mu_i)^2}, \qquad \textbf{(5.32)}$$

$$E[\sin(\gamma(S_1 \wedge_{\text{dep}} S_2)_{\text{e}})] = \sum_{i=1}^{k} \frac{\gamma\hat{\alpha}_i(1+\rho_i)/(2\mu_i)}{1+\gamma^2(1+\rho_i)^2/(2\mu_i)^2}, \qquad \textbf{(5.33)}$$

$$E[\cos(\gamma(S_1 \wedge_{\text{dep}} S_2)_{\text{e}})] = \sum_{i=1}^{k} \frac{\hat{\alpha}_i}{1+\gamma^2(1+\rho_i)^2/(2\mu_i)^2}. \qquad \textbf{(5.34)}$$

When the arrival rate is time-varying as given in Eq. (4.4) in the model $M_t^B/H_2^D/\infty$, we conduct simulations to evaluate the approximations in Eqs. (4.6),

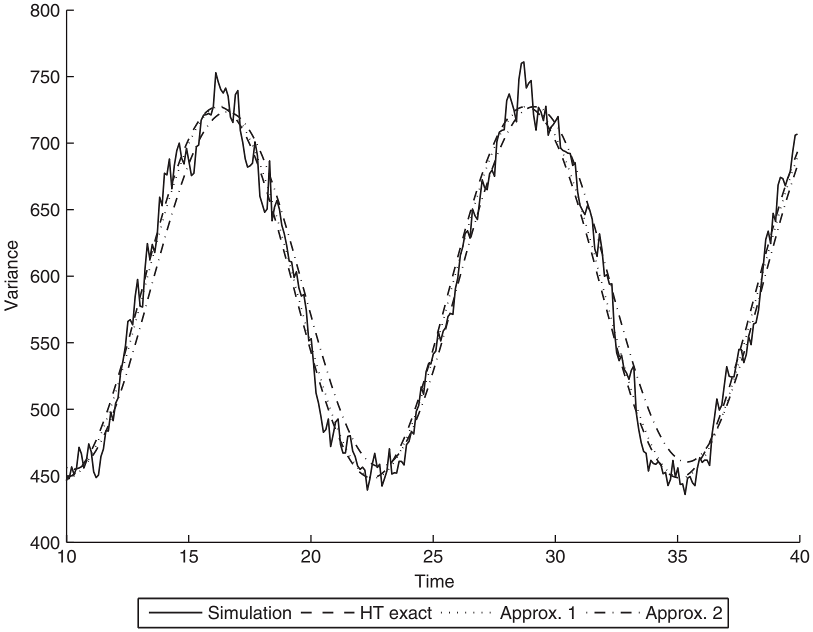

**FIGURE 2.** Comparison of simulated variance and its approximation functions in the $M_t^B/H_2^D/\infty$ queue with $B$ as a mixture, $B = 1$ w.p. 8/9 and $B \sim \text{Geom}(0.1)$ w.p. 1/9 and $H_2$ service times with parameters $\alpha_1 = \alpha_2 = 0.5, \mu_1 = 1, \mu_2 = 10,$ $\rho_1 = 0.1, \rho_2 = 0.5$, so that $m_S = 0.55$ and $\rho = 0.360$. $\lambda_B(t) = 80 + 20\sin(0.5t)$.

(4.14), and (4.10) with $\hat{\lambda}_B$ replaced by Eq. (4.15) and $\delta = 1/E[S_e]$. We consider the following parameter set:

$$\bar{\lambda}_B = 80, \qquad \beta = 20, \qquad \gamma = 0.5,$$

$$H_2 : \alpha_1 = \alpha_2 = 0.5, \qquad \mu_1 = 1, \qquad \mu_2 = 10, \qquad \rho_1 = 0.1, \qquad \rho_2 = 0.5,$$

and the batch size $B$ as a mixture of deterministic and geometric distributions as in the third model in Table 1. We have $m_S = 0.55$, $\sigma_S^2 = 0.708$, and $\rho = 0.360$. We plot the simulated variance for an experiment with 5000 replications for this model, and compare with the heavy-traffic exact variance formula in Eq. (4.6), and the Taylor series approximation in Eq. (4.14), and the recent-average-arrival-rate approximation in Eq. (4.10), denoted as "HT exact," "Approx. 1," and "Approx. 2", respectively, in Figure 2.

## 6. CONCLUSION

In this paper, we have introduced and studied the $G_t^B/G^D/\infty$ infinite-server model with batch arrivals and dependence among the service times within each batch. In order to capture the main effects, we have assumed that all pairs of service times within the same batch have identical bivariate distributions. From our earlier work in [15,16], we already know that the HT limit for the queue-length process (number of busy servers) is a Gaussian process, where the mean is independent of the dependence, while the variance depends on the dependence, but in a relatively complicated way.

In this paper, we have quantified the impact of the dependence among the service times on the variance of the queue length, both in stationary models and in models with time-varying arrival rates. For the stationary model, Proposition 3.1 dramatically shows the variance as a function of all model parameters. All quantities appearing there are either model parameters or can be expressed as mean values of random variables; see Eq. (3.7). Proposition 3.2 shows an even more elementary formula if we approximate the bivariate distribution by a special bivariate distribution, depending only on a given correlation. Proposition 5.1 shows that the resulting simple approximation based on the correlation parameter is actually exact for the MO bivariate exponential distribution, used in later experiments.

In Section 4 we showed that effective approximations for the time-varying variance of the queue-length process can also be developed for time-varying arrival rates. In Section 5 we conducted simulation experiments evaluating the approximations. To do so, we needed to introduce specific models of service times that are dependent within batches. For that purpose we relied on the MO multivariate exponential distribution and introduced a generalization to multivariate hyperexponential distributions. The tables and plots show that the approximations are remarkably accurate.

There are many directions for future research. An important one is to empirically investigate the presence of batch arrivals in service systems and dependence among the service times within these batches. More generally, it would be good to estimate the dependence among service times in service systems. In the batch model, it is natural for the customers in a batch not to arrive precisely at the same instant. Instead, each customer in the batch may arrive after a random delay following a "common triggering incident". To capture that effect, we propose considering a more general model, in particular, two IS stations in series, with the first IS station representing the extra delay, while the second is the actual service facility. A basis for such investigations lies in [15]. We hope to be able to report on progress in these directions in the future.

## References

1. Berkes, I., Hörmann, S., & Schauer, J. (2009). Asymptotic results for the empirical process of stationary sequences. *Stochastic Process & Their Applications* 119: 1298–1324.
2. Berkes, I. & Philipp, W. (1977). An almost sure invariance principle for empirical distribution function of mixing random variables. *Zeitschrift Wahrscheinlichkeitstheorie und Verwanate Gebiete* 41: 115–137.
3. Bladt, M. & Nielsen, B.F. (2010). On the construction of bivariate exponential distributions with an arbitrary correlation coefficient. *Stochastic Models* 26: 295–308.
4. Eckerg, A.E. (1983). Generalized peakedness of teletraffic processes. Proceedings of 10th International Teletraffic Congress. Montreal, Canada.

5. Eick, S.G., Massey, W.A., & Whitt, W. (1993). The physics of The $M_t/G/\infty$ queue. *Operations Research* 41: 731–742.
6. Eick, S.G., Massey, W.A., & Whitt, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* 39: 241–252.
7. Falin, G. (1994). The $M^k/G/\infty$ batch arrival queue with heterogeneous dependent demands. *Journal of Applied Probability* 31: 841–846.
8. Jacobs, P.A. & Lewis, P.A.W. (1977). A mixed autoregressive-moving average exponential sequence and point process (EARMA 1,1). *Advances in Applied Probability* 9: 87–104.
9. Liu, L., Kashyap, B.R.K., & Templeton, J.G.C. (1991). On the $GI^X/G/\infty$ system. *Journal of Applied Probability* 27: 671–683.
10. Liu, L. & Templeton, J.G.C. (1993). Autocorrelations in infinite server batch arrival queues. *Queueing Systems* 14: 313–337.
11. Mark, B.L., Jagerman, D.L., & Ramamurthy, G. (1997). Peakedness measures for traffic characterization in high-speed networks. INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Vol. 2, 427–435.
12. Marshall, A.W. & Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association* 62(317): 30–44.
13. Massey, W.A. & Whitt, W. (1996). Stationary-process approximations for the nonstationary Erlang loss model. *Operations Research* 44(6): 976–983.
14. Pang, G., Talreja, R., & Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* 4, 193–267.
15. Pang, G. & Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* 65: 325–364.
16. Pang, G. & Whitt, W. (2011). The impact of dependent service times on large-scale service times. *Manufacturing & Service Operations Management* http://dx.doi.org/10.1287/msom.1110.0363.
17. Shanbhag, D.N. (1966). On infinite server queues with batch arrivals. *Journal of Applied Probability* 9: 208–213.
18. Whitt, W. (1983). Comparing batch delays and customer delays. *Bell System Technical Journal* 62 (7): 2001–2009.
19. Whitt, W. (1984). Heavy traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal* 63, 689–708.
20. Whitt, W. (2002). *Stochastic-process limits*. New York: Springer.