MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1523-4614 | EISSN 1526-5498 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

The Impact of Dependent Service Times on Large-Scale Service Systems

Guodong Pang
Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802; gup3@psu.edu, http://www2.ie.psu.edu/pang/index.html

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027; ww2040@columbia.edu, http://www.columbia.edu/ww2040

This paper investigates the impact of dependence among successive service times upon the transient and steady-state performance of a large-scale service system. That is done by studying an infinite-server queueing model with time-varying arrival rate, exploiting a recently established heavy-traffic limit, allowing dependence among the service times. That limit shows that the number of customers in the system at any time is approximately Gaussian, where the time-varying mean is unaffected by the dependence, but the time-varying variance is affected by the dependence. As a consequence, required staffing to meet customary quality-of-service targets in a large-scale service system with finitely many servers based on a normal approximation is primarily affected by dependence among the service times through this time-varying variance. This paper develops formulas and algorithms to quantify the impact of the dependence among the service times upon that variance. The approximation applies directly to infinite-server models, but also indirectly to associated finite-server models, exploiting approximations based on the peakedness (the ratio of the variance to the mean in the infinite-server model). Comparisons with simulations confirm that the approximations can be useful to assess the impact of the dependence.

Key words: large-scale service systems; dependence among service times; stochastic models; infinite-server queueing models; peakedness; time-varying arrival rates
History: MSOM-11-062; first draft March 5, 2011; revisions August 2, September 23, October 8, 2011

1. Introduction

Performance analysis models for managing service systems, e.g., setting staffing in call centers, usually assume that the required service times are mutually independent random variables, but successive service times may in fact be dependent. For example, in a technical support telephone call center responding to service calls, a product defect can lead to many calls concerning that same product after the product is first introduced, with these tending to require longer-than-usual handling times. That would increase the average handling time during this time period, but that also would make the call handling times positively correlated. There are well developed methods to study the impact of average service times, but the impact of the dependence, for given mean, has evidently not been studied before. We will show that positive correlation among service times typically produces additional congestion, reducing the quality of service during that time period, unless staffing is increased. Moreover, we will quantify the impact. This phenomenon and our results apply both to the transient system performance after any one new product is introduced and the steady-state performance as a succession of new products are introduced over time.

For another example, 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 of serious problems may lead to multiple patients with longer-than-usual service times. Again, that would increase the average service time during this time period, but also would make the service times positively correlated. As in the first example, this dependence affects both the transient system performance associated with a single incident and the steady-state performance as a succession of incidents occur over time.

In this paper, we investigate the performance impact of dependence among the service times in a queueing model with a large number of homogeneous servers. We treat transient effects as well as steady-state effects. We do so by establishing results for transient behavior, allowing time-varying arrival rates. We directly develop analytical formulas and numerical algorithms to expose the approximate performance impact of dependent service times for infinite-server (IS) models. The results have useful applications for service systems with only finitely many servers in two ways. First, IS models can be directly applied to understand and control the performance of large-scale service systems. Second, the new approximation for IS models can be applied to yield corresponding performance approximations for models with only finitely many servers; we elaborate below. In both cases, analytical formulas and numerical algorithms can usefully complement and supplement computer simulation. Analytical formulas provide important insight; e.g., see Proposition 3 and the following discussion.

Previous work has shown that IS models can be directly applied to effectively approximate and control the performance of large-scale service systems; see Jennings et al. (1996), Feldman et al. (2008) and Green et al. (2007) for applications of the IS model to staffing. For systems with finitely many servers, where customers wait if they cannot be served immediately upon arrival, we can directly approximate the number of customers in the system by the corresponding number in the IS model. For example, if $s(t)$ is the number of servers in the actual system at time t and $X(t)$ is the random number in system in the IS model at time t, then the probability that an arrival at time t would have to wait before starting service can be approximated by $P(X(t) \geq s(t))$.

For the IS model, we determine the performance impact of the dependence among the service times by approximating the distribution of the number of customers in the system, allowing the dependence, and comparing the result with and without the dependence. To do that, we apply an approximation for the the distribution of the number of customers in an IS model based on a many-server heavy-traffic limit established in Pang and Whitt (2011). That analysis shows that: (i) the number of customers in the IS model at each time is approximately normally distributed, (ii) the average number in the IS model at each time is unaffected by dependence among the service times, and (iii) the variance of the number in the IS model at each time *is* affected by dependence among the service times, and it can be quantified.

Hence, to characterize the performance impact of the dependence among the service times, it suffices to examine the expression for the variance of the number in the IS system. However, the expressions for the variance in Pang and Whitt (2011) are quite complicated. Even the steady-state variance formula is complicated; see (6) below, which uses (3) and (5). Our main contribution is to show that useful engineering approximations can be extracted from the results in Pang and Whitt (2011) and to conduct simulation experiments showing that the approximations are effective.

Our results about the variance translate quite directly into implications for staffing. As discussed in Jennings et al. (1996), if we aim to set staffing to achieve a target probability of delay, then the normal approximation dictates that the staffing level be set at the mean plus some constant multiple of the standard deviation. Suppose that it has been decided to set the staffing level at the mean plus a constant q times the standard deviation. If dependence among the service times changes the variance of the number in system at time t from $\sigma^2(t)$ to $\eta(t)\sigma^2(t)$, where $\eta(t) > 1$, then the required staffing should increase from $E[X(t)] + q\sigma(t)$ to $E[X(t)] + q\sqrt{\eta(t)}\sigma(t)$, which is an increment of $(\sqrt{\eta(t)} - 1)q\sigma(t)$.

We indicated that effective approximations for the performance in finite-server models can be based on the performance of associated IS models. For this purpose, the peakedness – the ratio of the variance to the mean of the number of busy servers in the IS model – has proven to be very useful; see Eckberg (1983), Whitt (1984), Jagerman and Melamed (1994), Massey and Whitt (1996), Mark et al. (1997), Whitt (2004) and references therein. Since the dependence in the service times in the IS model does not affect the mean number of busy servers at all, our approximation for the variance of the number of busy servers translates directly into an associated approximation for the peakedness. Thus we can obtain new approximations for the performance in finite-server models with dependent service times by simply substituting our new peakedness for the old peakedness without dependence into the previous approximation formulas; see (34). Such new approximations need to be carefully examined, because they have not considered previously. We demonstrate the potential of the new approximations by reporting results from a simulation experiment for a finite-server model with dependent service times. Table 5 shows that the new approximation for the steady-state delay probability is remarkably effective.

Related literature. Even for the relatively elementary IS model with Poisson arrivals, relatively little work has been done previously on dependent service times; one exception is Falin (1994), who provided an algorithmic approach for the exact distribution with Poisson arrivals. Another exact numerical approach could be based on replacing the Ph_t service times in the $Ph_t/Ph_t/\infty$ model in Nelson and Taaffe (2004) by an associated Markovian service process (MSP_t), which has the structure of a Markovian arrival process, admitting dependence among the service times; e.g., see Asmussen (2003).

Unlike for the many-server models considered here, much is known about the performance impact of dependence among the service times, as well as among the interarrival times and between interarrival times and service times, in single-server queueing models and related models with few servers. That impact is clearly revealed in conventional heavy-traffic approximations, where the traffic intensity is allowed to increase toward its critical value 1 from below; see Theorems 9.3.3 and 9.3.4 and §9.6 of Whitt (2002) for a detailed treatment of the case of a single-server queue. The impact of the dependence is captured via the sums of all the pairwise correlations, as shown in (6.11) on p. 308 of Whitt (2002). The three forms of dependence - among interarrival times, among service times, and between interarrival times and service times - can all be important as shown for a packet queue example in Fendick et al. (1989), reviewed in Example 9.6.1 of Whitt (2002). Our results here indicate that the impact of the dependence is less dramatic for many-server queues, but it still can be significant; e.g., see the end of §6.

Organization of the paper. In §2 we review the result from Pang and Whitt (2011). In §3 we restrict attention to the the steady-state distribution of one stationary IS model, and develop effective representations for the terms in the steady-state variance formula. In §4 we develop approximations for the variance of the steady-state number in the system based on the correlations of successive service times. We show that this approximation is realized exactly via a model of randomly repeated service times, which is a special case of a first-order discrete autoregressive process, DAR(1), studied by Jacobs and Lewis (1978, 1983). In §5 we consider examples and compare the approximations to simulations. As an example of geometrically decaying correlations, we include a simulation using the special autoregressive moving-average stationary sequence of dependent service times with exponential marginals, the so-called EARMA(1,1) process, from Jacobs and Lewis (1977). In §6 we evaluate the performance of the approximation for the delay probability in the finite-server model with the same EARMA(1,1) service process. In §7 we apply the approximations developed for the stationary model to develop an approximation for the time-varying variance in the model with time-varying arrival rates. In §8 we conduct simulations to evaluate the approximations for time-varying arrival rates, considering the special case of sinusoidal time-varying arrival rates. Finally, in §9 we draw conclusions.

2. Review of the Heavy-Traffic Limit

The new approximation for the performance of the $G_t/G/\infty$ model from Pang and Whitt (2011), stated in (5) below, is obtained from a many-server heavy-traffic limit. The arrival process is assumed to satisfy a FCLT of the form

$$(A_n(t) - n\Lambda(t))/\sqrt{n} \Rightarrow \sqrt{c_a^2} B(\Lambda(t)) \quad \text{as} \quad n \to \infty, \tag{1}$$

in the function space D (see Whitt (2002)), where

$$\Lambda(t) \equiv \int_0^t \lambda^*(s)\, ds, \quad t \geq 0, \tag{2}$$

and B is a standard Brownian motion. Thus, asymptotically, the arrival process is characterized by the time-varying arrival-rate function $\lambda^*(t)$ and the variability parameter c_a^2, which is determined by the limit (1). Dependence among the interarrival times is captured by the parameter c_a^2, e.g., see (7) below. For the principal case in which $A_n(t)$ is a nonhomogeneous Poisson process, $c_a^2 = 1$. Consequently, in model n, the arrival rate at time t is approximately $n\lambda^*(t)$, while the number of arrivals in the interval $[0,t]$, $A_n(t)$, is approximately distributed as $N(n\Lambda(t), nc_a^2\Lambda(t))$, where $N(m,\sigma^2)$ denotes a random variable normally distributed with mean m and variance σ^2. Asymptotically, the arrival process has independent (but not necessarily stationary) increments.

Unlike the arrival process, we know from Krichagina and Puhalskii (1997) and Pang and Whitt (2010, 2011) that the service times affect the many-server heavy-traffic limit, not via their counting process or partial sums, but instead via the sequential empirical process. Let $\{S_i : i \geq 1\}$ be the sequence of service times of successive customers. The (fluid-scaled) *sequential empirical process* is $\bar{K}_n(t,x) \equiv n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}(S_i \leq x)$, where $\equiv$ denotes equality by definition and $\mathbf{1}(A)$ is the indicator function of the event A, equal to 1 on A and 0 elsewhere. The sequential empirical process takes a horizontal (or sideways) view of the service times instead of the customary vertical view. Let all service times be distributed as the random variable S having cumulative distribution function (cdf) F with finite mean m_S. With independent service times, $\bar{K}_n(t,x) \to tF(x)$ as $n \to \infty$ by the law of large numbers. The FCLT for the scaled process $\hat{K}_n(t,x) \equiv \sqrt{n}(\bar{K}_n(t,x) - tF(x))$ is the basis for the heavy-traffic limit for the IS model; the limit process is $\hat{K}(t,x) = U(t, F(x))$, where $U(t,x)$ is the Kiefer process. The key tool for dependent service times in Pang and Whitt (2011) is the FCLT for the sequential empirical process of weakly dependent random variables in Berkes and Philipp (1977) and Berkes et al. (2009). The service times are assumed to be independent of the arrival process, but the service times can be mutually dependent. To be able to apply Berkes et al. (2009) and Berkes and Philipp (1977), Pang and Whitt (2011) assume that the service times come from a stationary sequence of random variables, satisfying appropriate mixing conditions (producing weak dependence), which we assume prevails. Let the stationary sequence be extended to a two-sided stationary sequence (which always can be done).

The many-server heavy-traffic impact of the dependence among the service times is determined by the bivariate cdf of service times j and $j+k$, $H_k(x,y) \equiv P(S_j \leq x, S_{j+k} \leq y)$, where $H_k(x,\infty) = H_k(\infty,x) = F(x)$ for all $k \geq 1$ and all $x \geq 0$. In particular, the bivariate cdf's H_k appear via the function

$$\Gamma(s) \equiv 2\sum_{k=1}^{\infty}(H_k(s,s) - F(s)^2) = 2\sum_{k=1}^{\infty}(H_k^c(s,s) - F^c(s)^2), \tag{3}$$

where $F^c(s) \equiv 1 - F(s)$ is the complementary cdf. The last relation in (3) holds because

$$H_k(s,s) = H_k^c(s,s) + 2F^c(s) - 1 \quad \text{and} \quad F(s)^2 = F^c(s)^2 + 2F^c(s) - 1, \tag{4}$$

so that $H_k(s,s) - F(s)^2 = H_k^c(s,s) - F^c(s)^2$.

Assuming that the system started empty in the distant past, we obtain the following heavy-traffic approximation from the FCLT in Pang and Whitt (2011):

$$Q(t) \approx N(m(t), v(t)), \quad \text{where} \quad m(t) \equiv \int_0^\infty \lambda(t-s) F^c(s)\, ds, \quad t \ge 0, \quad \text{and}$$
$$v(t) \equiv \int_0^\infty \lambda(t-s) V(s)\, ds, \quad V(s) \equiv F^c(s) + (c_a^2 - 1) F^c(s)^2 + \Gamma(s), \quad s \ge 0, \tag{5}$$

with Γ in (3), understanding $\lambda(t) = n\lambda^*(t)$. If we want the system to start at time 0 instead of in the distant past, then we can simply set $\lambda(t) = 0$ for $t < 0$ in (5).

From above, it follows that the desired approximate time-dependent variance $v(t)$ is a function of the arrival process through the arrival-rate function $\lambda(t)$ and the variability parameter c_a^2, and of the the service times through the bivariate cdf's $H_k(x,y)$. If the bivariate cdf's $H_k(x,y)$ were known and understood, then the story would be complete above, but that typically is not the case. Much of the following is devoted to developing effective ways to represent and estimate $v(t)$ without directly calculating or estimating $H_k(x,y)$ for all (k,x,y). After doing so, we conduct simulations to show that the resulting approximations are effective.

3. An Effective Representation for the Stationary Model

Associated steady-state formulas are obtained by simply replacing the time-varying arrival rate function $\lambda(t)$ in (5) by the constant λ. The corresponding steady-state formulas are

$$m(\infty) = \lambda \int_0^\infty F^c(s)\, ds = \lambda m_S, \quad v(\infty) = \lambda \int_0^\infty V(s)\, ds = \lambda v_\infty, \quad v_\infty \equiv \int_0^\infty V(s)\, ds, \tag{6}$$

where $V(s)$ is given in (5).

In the stationary setting, it is customary to work with a single fixed arrival process $A(t)$ with rate λ^* and let $A_n(t) \equiv A(nt)$, $t \ge 0$, $n \ge 1$. Then the FCLT (1) holds with $\Lambda(t) = \lambda^* t$ in (2). Then any dependence among the interarrival times is captured by the variability parameter c_a^2. Let $\{U_i\}$ be the sequence of interarrival times in A, assumed to be strictly stationary. As in §4.4 of Whitt (2002), the standard case is

$$c_a^2 = (\lambda^*)^2 \sigma_a^2, \quad \text{where} \quad \sigma_a^2 = Var(U_1) + 2\sum_{k=1}^\infty cov(U_1, U_k). \tag{7}$$

The series is required to converge in order to have (1). For a renewal process, $c_a^2 = Var(U_1)/(E[U_1])^2$, the squared coefficient of variation (SCV) of an interarrival time U.

Clearly, dependence among the service times affects the performance differently. If there is no dependence among the service times, then $\Gamma(s) = 0$, so that the third term in the integrand $V(s)$ in (6) and (5) drops out. If the arrival process is Poisson or if only the arrival process satisfies a FCLT with variability parameter $c_a^2 = 1$, then $c_a^2 - 1 = 0$, so that the second term $(c_a^2 - 1)F^c(s)^2$ in the integrand $V(s)$ in (6) and (5) drops out.

3.1. Peakedness

As we indicated in the introduction, for stationary IS models, it is revealing to focus on the peakedness, $z \equiv v(\infty)/m(\infty)$. The Markovian $M/M/\infty$ IS model is the reference case; then $z \equiv z(M/M) = 1$ because the distribution is Poisson. Here we understand z to be the *heavy-traffic approximation for the peakedness* (which can be shown to be the limit of $z(\lambda)$ as $\lambda \to \infty$ by limit interchange and uniform integrability arguments). For the $G/M/\infty$ model in heavy traffic, where the stationary arrival process ($\Lambda(t) = \lambda^* t$ in (2)) satisfies a FCLT with variability parameter c_a^2, and the service times are i.i.d., $z \equiv z(G/M) = (c_a^2 + 1)/2$ for c_a^2 in (7). For the more general model with general possibly dependent, service times, from (3), after dividing and multiplying by $m_S \equiv 1/\mu$, and recalling the tail integral formula used in (6), we obtain a revealing alternative expression.

Proposition 1. *For the general stationary model allowing dependent service times, the (heavy-traffic) peakedness can be represented as*

$$z \equiv z(G/G) \equiv z(c_a^2, F, \{H_k\}) = 1 + (c_a^2 - 1)I_1 + I_2, \tag{8}$$

where

$$I_1 \equiv I_1(F) \equiv \frac{\int_0^\infty F^c(s)^2\, ds}{m_S} \tag{9}$$

and

$$I_2 \equiv I_2(\{H_k\}) = \frac{\int_0^\infty \Gamma(s)\, ds}{m_S} = \frac{2\int_0^\infty \left(\sum_{k=1}^\infty (H_k^c(s,s) - F^c(s)^2)\right) ds}{m_S}. \tag{10}$$

3.2. Representation of Integrals as Mean Values

Unfortunately, formulas (6), (9) and (10) are still complicated, requiring that we somehow determine or estimate the entire functions F^c and Γ. However, we actually only require the integrals of these functions. We now show that the integrals have convenient expressions as means of random variables. That facilitates both analysis and statistical estimation. In particular, with simulation or system data, we can estimate these mean values directly via standard statistical methods for estimating means. That provides a convenient simplification for application of these results.

We obtain the new representation by exploiting tail integrals, as in (6). Let $S_1 \wedge_{ind} S_2$ be the minimum of two independent random variables each distributed as a single service time random variable S with cdf F, so that its complementary cdf is $P(S_1 \wedge_{ind} S_2 > s) = F^c(s)^2$. Hence, the integral term I_1 in (8) and (9) can be expressed as the ratio of two mean values:

$$I_1 = E[S_1 \wedge_{ind} S_2]/E[S_1], \tag{11}$$

where S_1 and S_2 are independent random variables with common cdf F. As a consequence, $0 \le I_1 \le 1$, with $I_1 = 1$ when F in the deterministic case in which F is the distribution of a single point mass. More generally, I_1 tends to decrease as the distribution of F gets more variable. The exponential case is an intermediate case, yielding $I_1 = 1/2$.

Similarly, $H_k^c(s,s) = P(S_j \wedge S_{j+k} \ge s)$, where these random variables have their given joint distribution. Hence, we can also obtain an alternative representation of I_2 exploiting tail integrals. Paralleling (11), we can write

$$J_k \equiv \frac{E[S_j \wedge S_{j+k}]}{E[S_j]}, \quad k \ge 1, \quad \text{and} \quad I_2 = 2\sum_{k=1}^\infty (J_k - I_1). \tag{12}$$

Just like I_1, we have $0 \le J_k \le 1$. Since we are considering positive dependence, we expect to have $J_k \ge I_1$ for all k. At first glance, there is an issue about convergence for I_2 in (12), but it can be expected because we should have $J_k - I_1 \to 0$ as $k \to \infty$.

We summarize our conclusions in

Proposition 2. *The two integral terms I_1 and I_2 defined in (9) and (10) and appearing in the peakedness formula (8) can be expressed in terms of mean values of the random variables S_j, $S_1 \wedge_{ind} S_2$ and $S_j \wedge S_{j+k}$ via (11) and (12).*

4. Approximations Based on Correlations

In applications, it is common to specify dependence through correlations as opposed to the full bivariate cdf H_k or the mean values of the random variables S_j, $S_1 \wedge_{ind} S_2$ and $S_j \wedge S_{j+k}$. Thus, in this section we develop an approximation that depends only on the correlations. In the next section we investigate how well it works. We cannot apply any assumption about the correlations directly, because the approximation formula for the key integral term I_2 in (10) and (12) depends on the full cdf's H_k, and not just the correlations. Thus we now provide a way to approximate the bivariate cdf's H_k given the partial information provided by their correlation.

4.1. Exploiting Extremal Bivariate cdf's

We will construct the approximating cdf by exploiting extremal bivariate cdf's with the given marginal cdf F; see Whitt (1976). The maximum correlation 1 is achieved when the two service times are identical; the joint cdf is $\tilde{H}_1(x,y) \equiv F(x \wedge y)$. The minimum nonnegative (zero) correlation is achieved when the random variables are independent; the joint cdf is $\tilde{H}_0(x,y) \equiv F(x)F(y)$. (It is possible to construct multivariate distributions with negative correlation, but that does not seem realistic for the present application.) A specific cdf with correlation ρ, $0 \le \rho \le 1$, is achieved by taking a convex combination of these two cdf's, i.e.,

$$\tilde{H}_\rho(x,y) \equiv \rho \tilde{H}_1(x,y) + (1-\rho)\tilde{H}_0(x,y) = \rho F(x \wedge y) + (1-\rho)F(x)F(y). \tag{13}$$

Given that two service times have correlation ρ and marginal cdf F, we can let $\tilde{H}_\rho$ be the joint cdf in (13). It has marginal cdf F and correlation ρ. We can thus compute an approximation to I_2 based on the partial characterization of the joint cdf H_k by its marginal cdf F and correlation ρ. In particular, as a first step, from (13) and (4), we get

$$\tilde{H}^c_\rho(s,s) - F^c(s)^2 = \rho(F^c(s) - F^c(s)^2) \quad \text{and} \quad \tilde{H}^c_\rho(s,s) = \rho F^c(s) + (1-\rho)F^c(s)^2, \quad s \ge 0. \tag{14}$$

Using the second relation in (14), we see that the function $H^c_\rho(s,s)$ of the single variable s coincides with the ccdf of a random variable, say Y_ρ, that is a mixture of random variables with ccdf's $F^c(s)$ and $F^c(s)^2$. On the other hand, we can apply the first relation in (14) with (10) to directly obtain a new approximate expression for the integral I_2 as a function of all the pairwise correlations, in particular we get

$$I_2 \approx 2(1-I_1)\Sigma_\rho, \quad \Sigma_\rho \equiv \sum_{k=1}^{\infty} \rho_k \quad \text{and} \quad \rho_k \equiv Corr(S_j, S_{j+k}). \tag{15}$$

We assume that the sum Σ_ρ in (15) is finite. Given (13), we also obtain a simple representation for the function V in (5). For these bivariate cdf's, we have

$$V(s) = (1+2\Sigma_\rho)F^c(s) + (c_a^2 - 1 - 2\Sigma_\rho)F^c(s)^2, \quad s \ge 0. \tag{16}$$

We now summarize our conclusions.

Proposition 3. *If we fit an approximating bivariate cdf H to a specified marginal cdf F and a nonnegative correlation ρ via* (13), *then the peakedness in Proposition* 1 *becomes*

$$z = 1 + (c_a^2 - 1)I_1 + 2(1-I_1)\Sigma_\rho = (1+2\Sigma_\rho) + (c_a^2 - 1 - 2\Sigma_\rho)I_1, \tag{17}$$

where Σ_ρ is the sum of all correlations in (15).

Approximation (17) is very useful to obtain a basic understanding of the causes of peakedness. There are three relevant distinct parameters in approximation (17): the arrival process variability parameter c_a^2, the marginal service-time variability parameter I_1 depending only on F, and the service-time dependence factor Σ_ρ in (15).

From the first expression in (17), we see that the peakedness z is *linearly increasing* in the two variables c_a^2 and Σ_ρ for any value of I_1. For c_a^2, the growth factor is I_1; for Σ_ρ, the growth factor is $2(1-I_1)$. The first growth factor increases in I_1, ranging from 0 to 1, while the second decreases in I_1, ranging from 2 to 0. The second expression in (17) shows that z is *linearly increasing* (*unaffected by* or *linearly decreasing*) in I_1 when $c_a^2 - 1 - 2\Sigma_\rho >$ ($=$ or $<$)0.

In the cases of D service, M service and highly variable service, we have $I_1 = 1$, $I_1 = 1/2$ and $I_1 \approx 0$, respectively. Thus, for D service, $z = c_a^2$; for M service, $z = ((c_a^2+1)/2) + \Sigma_\rho$; and for highly variable service (characterized by $I_1 \equiv 0$), $z = 1 + 2\Sigma_\rho$. Thus, with D service, only the arrival process matters; with M service, we add Σ_ρ to the $G/M/\infty$ peakedness expression; with highly variable service (as represented by small I_1), the arrival process variability as captured by the parameter c_a^2 plays no role.

4.2. Randomly Repeated Service Times

We now introduce a model for a stationary sequence of service times for which the bivariate cdf's H_k coincide with the special bivariate cdf's $\tilde{H}_{\rho_k}$ in (13). As a consequence, for these models, the approximation for the (heavy-traffic) peakedness in Proposition 3 is exact.

A simple one-parameter model. We start with an initial simple single-parameter model for the sequence of service times with marginal cdf F. We let each successive service time be a mixture of the previous service time with probability p or a new independent service time having cdf F, with probability $1-p$. (This is a special case of a first-order discrete autoregressive process, DAR(1), studied by Jacobs and Lewis (1978, 1983).) In other words, we can let S_1 be distributed according to F, $\{X_k : k \geq 2\}$ be a sequence of i.i.d. random variables, each with cdf F, and then $\{Z_k : k \geq 2\}$ be a sequence of i.i.d. random variables with $P(Z_k = 1) = 1 - P(Z_k = 0) = p$. Then, given S_1, we construct the sequence of service times $\{S_k : k \geq 1\}$ by stipulating that

$$S_k = Z_{k-1}S_{k-1} + (1 - Z_{k-1})X_k, \quad k \geq 2. \tag{18}$$

This model produces independent groups or "batches" of identical service times, where the batch sizes are geometric. In this model, we have the correlations

$$corr(S_j, S_{j+k}) = p^k \tag{19}$$

and we have all the bivariate cdf's

$$H_k(x,y) \equiv P(S_j \leq x, S_{j+k} \leq y) = \tilde{H}_{\rho_k}(x,y) \quad \text{where} \quad \rho_k = p^k. \tag{20}$$

Hence $\Sigma_\rho = p/(1-p)$ in this case.

General batch-size distribution. We can extend the model to allow batches of identical service times with non-geometric distributions. The new model for the sequence of service times with marginal cdf F has each successive service time be a mixture of the previous service time with probability p_k or a new independent service time having cdf F, with probability $1-p_k$, where the probability p_k depends on how many repeated service times have occurred so far. We get probability p_k if there have been k successive identical service times previously

In other words, as before, we can let S_1 be distributed according to F, we can let $\{X_k : k \geq 2\}$ be a sequence of i.i.d. random variables, each with cdf F. But now include the counter variables N_k. We initially set $N_1 = 1$. Now let $\{Z_k : k \geq 1\}$ be a sequence of random variables that are conditionally independent given the sequence $\{N_k : k \geq 1\}$, with

$$P(Z_k = 1 | N_j, j \leq k) = 1 - P(Z_k = 0 | N_j, j \leq k) = p_{N_k}, \tag{21}$$

where $\{p_k : k \geq 1\}$ is a sequence of probabilities ($0 \leq p_k \leq 1$), where $\prod_{k=1}^{\infty} p_k = 0$. Then we can recursively construct the sequences of service times $\{S_k : k \geq 1\}$ and the counting variables $\{N_k : k \geq 1\}$ by stipulating that

$$S_k = Z_{k-1}S_{k-1} + (1 - Z_{k-1})X_k, \quad k \geq 2, \tag{22}$$

and $N_k = Z_{k-1}N_{k-1} + 1$, $k \geq 2$; i.e., $N_k = N_{k-1} + 1$ if $Z_{k-1} = 1$ and $N_k = 1$ if $Z_{k-1} = 0$.

We have specified the distribution of each successive batch size via the conditional probabilities $P(B = k | B \geq k-1) = p_{k-1}$, $k \geq 2$. We have directly assumed that B has a proper distribution: $P(B < \infty) = 1$. We will also want to require that the mean $E[B]$ is finite as well. Hence, we assume that

$$E[B] = \sum_{k=1}^{\infty} P(B \geq k) = \sum_{k=1}^{\infty} \prod_{i=1}^{k-1} (1 - p_i) < \infty. \tag{23}$$

We also want the service times we consider to come from a stationary sequence. In order to achieve that, we need to start with a stationary batch, denoted by B^*. We want to assume that the age of the initial batch is distributed according to the batch-size stationary excess distribution. In particular, we assume that

$$P(N_1 = k) = P(B^* = k) \equiv p^*_k = \frac{P(B \geq k)}{\sum_{k=1}^{\infty} P(B \geq k)} = \frac{P(B \geq k)}{E[B]}; \tag{24}$$

which has mean

$$m_{B^*} \equiv E[B^*] = \frac{E[B^2] + m_B}{2m_B} = \frac{m_B(c_B^2 + 1) + 1}{2}. \tag{25}$$

See Whitt (1983) for more on the batch-size stationary-excess distribution.

For this more general model, we again have the bivariate distributions as in (20), but now, for $k \geq 1$,

$$\rho_k \equiv Corr(S_j, S_{j+k}) = \sum_{j=1}^{\infty} p^*_j P(B \geq j + k | B \geq j) = \frac{\sum_{j=1}^{\infty} P(B \geq j + k)}{E[B]} = P(B^* > k). \tag{26}$$

Since the bivariate cdf's are the same as in §4, the approximation in Proposition 3 is again exact for this model.

PROPOSITION 4. *For the $G/RRS/\infty$ IS model with random batch-size B having finite first two moments, the exact heavy traffic peakedness is as in Proposition* 3 *with*

$$\Sigma_\rho \equiv \sum_{k=1}^{\infty} \rho_k = \sum_{k=1}^{\infty} P(B^* > k) = m_{B^*} - P(B^* \geq 1) = m_{B^*} - 1 = \frac{m_B(c_B^2 + 1) - 1}{2} < \infty. \tag{27}$$

Proof. Apply (26) and (25) to compute Σ_ρ. ■

Under regularity conditions, we can also construct an RRS model to have given correlations.

PROPOSITION 5. *Suppose that the stationary sequence of service times has the sequence of correlations $\{\rho_k : k \geq 1\}$ with $\Sigma_\rho < \infty$. If the associated sequence $\{\rho_{k-1} - \rho_k : k \geq 1\}$, where $\rho_0 \equiv 1$, is nonincreasing, then we can construct an RRS model with the given correlation sequence. The mean batch size is $m_B = (1 - \rho_1)^{-1}$. The batch-size distribution is specified by having $P(B \geq 1) = 1$ and*

$$P(B \geq k) = \frac{\rho_{k-1} - \rho_k}{1 - \rho_1}, \quad k \geq 1. \tag{28}$$

For the associated $G/RRS/\infty$ IS model, the exact heavy-traffic peakedness is as in Proposition 3.

Proof. From (26), $1 - \rho_1 = P(B^* = 1) = P(B \geq 1)/m_B = \frac{1}{m_B}$, implying that $m_B = (1 - \rho)^{-1}$. Also from (26),

$$\rho_{k-1} - \rho_k = P(B^* = k) = \frac{P(B \geq k)}{m_B} = (1 - \rho_1) P(B \geq k), \quad k \geq 1. \quad ■$$

5. Examples and Simulation Comparisons

In this section we consider some examples and make comparisons with simulation. In addition to the model with random repeated service times in §4.2, we also consider the EARMA service times introduced by Jacobs and Lewis (1977), extended by Lawrence and Lewis (1980) and Sim (1990), and studied further in queueing models by Jacobs (1980). We consider a wide range of correlations, including quite high values, which seem less realistic, but indicate the limits of the approximations.

5.1. Evaluating the Approximation with Random Repeated Service Times

We first describe simulations to evaluate the heavy-traffic peakedness approximation for the $M/RRS/\infty$ model, having random repeated service times, as in §4.2. We let the service times all be exponentially distributed with mean 1. We consider the simple one-parameter model in (22) specified by the parameter p. We consider 5 values of p: 0.1, 0.25, 0.50, 0.75 and 0.90, with higher values indicating higher correlations.

For the stationary model, from (16), we get the associated peakedness in (17). In the case of a Poisson arrival process and exponential service times having mean 1, we get

$$V(s) = (1 + 2\Sigma_\rho)e^{-s} - 2\Sigma_\rho e^{-2s}. \tag{29}$$

For the stationary model, we get the associated heavy-traffic peakedness

$$z = (1 + 2\Sigma_\rho) - \Sigma_\rho = 1 + \Sigma_\rho = \frac{1}{1-p}. \tag{30}$$

Results for our simulation experiments are shown in Table 1. In our simulation experiments we considered a range of arrival rates. For each value of p, we show values of λ that approximately yield 1% and 10% error for each case (discovered through simulation experiments). We see that the required value of λ increases as the dependence (measured by p) increases. When p is not extremely large, the heavy-traffic peakedness provides a remarkably good approximation for a wide range of λ. For the RRS examples in Table 1, the heavy-traffic peakedness seems to be an upper bound that is approached monotonically as λ increases.

p	λ	HT Approx.	simul.	95%$c.i.$	Time Int.	N reps.
0.10	10.0	1.111	1.103	±0.009	$[20, 100]$	1000
0.10	0.1	1.111	1.007	±0.018	$[20, 100]$	3000
0.25	25.0	1.333	1.320	±0.007	$[20, 100]$	1000
0.25	2.0	1.333	1.201	±0.009	$[20, 100]$	1000
0.50	100.0	2.000	1.976	±0.012	$[20, 100]$	2000
0.50	10.0	2.000	1.839	±0.021	$[20, 100]$	2000
0.75	800.0	4.000	3.976	±0.012	$[20, 100]$	5000
0.75	80.0	4.000	3.861	±0.014	$[20, 100]$	5000
0.90	1000.0	10.000	9.899	±0.018	$[20, 1000]$	5000
0.90	80.0	10.000	8.999	±0.014	$[200, 1000]$	5000

Table 1 **Comparison of the heavy-traffic peakedness for the** $M/RRS/\infty$ **model in Propositions 3 and 4 to simulation estimates. Here we use the single-parameter randomly repeated exponential service times with heavy-traffic peakedness equal to** $1 + \Sigma_\rho = 1/(1-p)$ **from (17) and (19). We consider five cases for the RRS parameter** p: 0.10, 0.25, 0.50, 0.75 **and** 0.90. **For each case, two arrival rates are considered, the higher one yielding about** 1% **error and the lower one yielding about** 10% **error.**

5.2. Non-Exponential Distributions

We now extend the last subsection by considering dependence in the interarrival times as well as the service times and non-exponential distributions. (We use independent RRS models for both the interarrival times and the service times.) In addition to exponential marginal distributions, now we also consider hyperexponential (H_2, a mixture of two exponential) marginal distributions, using SCV $c^2 = 4$ and balanced means to fix the parameters; see (3.7) on p. 137 of Whitt (1982). As before, the mean service time is 1. For this H_2 distribution, $I_1 = 0.3500$. As in Table 1, we consider random repeated service times, using the single-parameter model with $p = 1/2$. The RRS asymptotic variability parameter for the arrival process is $c_a^2 = c^2(1+2\Sigma_\rho) = 3c^2$, where c^2 is the SCV for a single interarrival times; it is obtained by combining Theorems 4.4.1 and 7.3.2 of Whitt (2002). The arrival rate was initially set at $\lambda = 100$, but then increased to $\lambda = 1000$ in the two cases with the more variable RSS(H_2) arrival process. The results in Table 2 show that the good performance extends to this greater level of generality, again provided that the arrival process is sufficiently large.

arrival	service	c_a^2	I_1	λ	HT Approx. (17)	simul.	95%*c.i.*
M	RRS(M)	1.00	0.500	100	2.000	1.976	±0.012
M	RRS(H_2)	1.00	0.350	100	2.300	2.274	±0.019
RRS(M)	RRS(M)	3.00	0.500	100	3.000	2.944	±0.024
RRS(M)	RRS(H_2)	3.00	0.350	100	3.000	2.974	±0.018
RRS(H_2)	RRS(H_2)	12.00	0.350	100	6.150	5.580	±0.032
				1000		6.085	±0.023
RRS(H_2)	RRS(M)	12.00	0.500	100	7.500	6.876	±0.058
				1000		7.417	±0.031

Table 2 **Comparison of the approximate heavy-traffic peakedness for infinite-server models with dependence and non-exponential distributions to simulation estimates. As in Table 1, random repeated service times are used, but now with hyperexponential marginals, RRS(H_2), as well as with exponential marginals, RRS(M). Hence, the peakedness is given in Propositions 3 and 4. The H_2 distributions have balanced means and SCV $c^2 = 4.0$. We use the single-parameter RRS model with $p = 1/2$, so that $\Sigma_\rho = 1$. The arrival rate is $\lambda = 100$ and the mean service time is 1. The first case is from Table 1.**

5.3. EARMA Service Times

The EARMA sequence of random variables is stationary with exponential marginal distributions and the correlation structure of an autoregressive moving average process, ARMA(1,1), called EARMA(1,1) in Jacobs and Lewis (1977) and simply EARMA here. The EARMA variables are random linear combinations of i.i.d. exponentials with the same mean. Specifically, we can start with three independent sequences of i.i.d. random variables $\{X_n : n \geq 0\}$, $\{U_n : n \geq 1\}$, and $\{V_n : n \geq 1\}$, where X_n is exponentially distributed with mean m, while

$$P(U_n = 0) = 1 - P(U_n = 1) = \beta \quad \text{and} \quad P(V_n = 0) = 1 - P(V_n = 1) = \rho. \tag{31}$$

The EARMA sequence $\{S_n : n \geq 1\}$ is defined recursively by

$$S_n = \beta X_n + U_n Y_{n-1} \quad \text{and} \quad Y_n = \rho Y_{n-1} + V_n X_n, \quad n \geq 1. \tag{32}$$

The serial correlation has geometric decay. Specifically,

$$\rho_k \equiv Corr(S_j, S_{j+k}) = \gamma \rho^{k-1}, \quad \text{where} \quad \gamma = \beta(1-\beta)(1-\rho) + (1-\beta)^2 \rho. \tag{33}$$

5.4. Simulations with EARMA Service Times

To evaluate the heavy-traffic approximation in Propositions 1 and 2, we simulated several $M/EARMA/\infty$ models with different EARMA service time sequences. Without loss of generality (since we are always free to choose the units to measure time), let the mean service time be 1. There are two remaining EARMA service-time parameters: β and ρ. In our simulations we consider five cases: $(0.75, 0.50)$, $(0.50, 0.50)$, $(0.50, 0.75)$, $(0.00, 0.75)$, $(0.25, 0.90)$. The cumulative correlations Σ_ρ increase over these five cases: 0.25, 0.50, 1.00, 3.00 and 5.25. For each of these five EARMA models, we used simulation to estimate the first 100 values of J_k in (12) in order to compute I_2 in (12) and

k	$(0.75, 0.50)$	$(0.50, 0.50)$	$(0.50, 0.75)$	$(0.00, 0.75)$	$(0.25, 0.90)$
1	0.523542	0.557248	0.567959	0.799895	0.699137
2	0.516550	0.533666	0.550338	0.695559	0.672134
3	0.508623	0.517945	0.537533	0.633580	0.648603
4	0.504265	0.508890	0.528097	0.593897	0.628662
5	0.502138	0.504377	0.520967	0.567253	0.611864
10	0.500073	0.500093	0.504739	0.514462	0.558417
20	0.500038	0.499998	0.500192	0.500792	0.518249
40	0.500046	0.499984	0.499920	0.499993	0.502008
100	0.500026	0.499977	0.499936	0.499988	0.499885

Table 3 **Simulation estimates of J_k for the five EARMA examples with parameters (β, ρ) considered in Table 4.**

thus the exact heavy-traffic peakedness in (8). (These estimates of I_2 apply to all arrival rates.) From the estimates of J_k, we see that the 100 (or much fewer) values are adequate; see Table 3.

Since the approximation is based on the heavy-traffic limit in which $\lambda \to \infty$, we consider a range of λ values. We consider a Poisson arrival process with five different arrival rates: $\lambda = 200, 100, 20, 10, 3$. From (17) and (33), the approximate peakedness based solely on correlations here is $z = 1 + \Sigma_\rho = 1 + \gamma/(1-\rho)$. From (33), in the five cases we have $\gamma = 1/8, 1/4, 1/4, 3/4, 21/40$.

To estimate the peakedness at each time point, we performed 2000 (or in some cases 5000) independent replications, starting the system empty. In each simulation run we collected data over the time interval $[20, 100]$ 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 intervals, we performed 4 further independent replications and used the Student t distribution with three degrees of freedom. (The halfwidth is $3.182 S_4/\sqrt{4}$, where S_4^2 is the sample variance.) The halfwidths of the confidence intervals of all estimates are approximately 1%. Table 4 shows the results.

The first thing to observe from Table 4 is that the dependence has a significant impact in these examples. The peakedness would simply be 1 if there were no dependence among the service times. The peakedness is higher by 11%, 25%, 53%, 200% and 325% in these five examples. Even if we dismiss the last cases as extreme cases (chosen to test the approximation), it is evident that the impact of the dependence can be significant.

From Table 4, we see that the heavy-traffic approximation for the peakedness (HT Approx.) is remarkably accurate, provided that the arrival rate λ is not too small. For very small values of λ, such as the value $\lambda = 3$ for the last three cases, the exact peakedness falls significantly below the heavy-traffic approximation in (8), but for moderate values of λ such as 10, the approximation is remarkably accurate. The exact value evidently first increases as λ increases, even slightly passing the heavy-traffic limit and then decreases toward that limit (as can be seen from the cases $\lambda = 100$ and 200).

Unfortunately, the appealing simple approximation based solely on the correlations alone in (17) is not very accurate. The errors in the five cases are: 12%, 20%, 31%, 33% and 42%. Evidently a reasonable rough approximation can be obtained from the correlations alone if the correlations are not too large, but the quality of the approximation degrades seriously as the correlations increase. From the results, it appears that $1 + \Sigma_\rho/2$ is a pretty good rough approximation for the $M/EARMA/\infty$ IS model, except in the last case. This experiment suggests that the approximation based on correlations may yield an upper bound. (Other experiments confirm this too.)

Case 4 in Tables 4 and 1 both have the correlation structure $Corr(S_1, S_{1+k}) = p^k$ for $p = 0.75$. Hence, the heavy-traffic approximation in Table 1 coincides with the approximation based on the correlations in Table 4, both yielding an approximate peakedness of 4.00. The simulation estimate is quite close to that approximation for the $M/RRS/\infty$ model in Table 1, but not for the

(β, ρ)	λ	simul.	95%$c.i.$	HT Approx.	95%$c.i.$	Corr. Approx.
$(0.75, 0.50)$	200	1.116	±0.003	1.119	±0.009	1.250
	100	1.116	±0.010	1.119		1.250
	20	1.108	±0.004	1.119		1.250
	10	1.101	±0.008	1.119		1.250
	3	1.080	±0.011	1.119		1.250
$(0.50, 0.50)$	200	1.251	±0.009	1.249	±0.025	1.500
	100	1.255	±0.008	1.249		1.500
	20	1.245	±0.011	1.249		1.500
	10	1.242	±0.008	1.249		1.500
	3	1.198	±0.005	1.249		1.500
$(0.50, 0.75)$	200	1.533	±0.011	1.526	±0.019	2.000
	100	1.535	±0.008	1.526		2.000
	20	1.533	±0.009	1.526		2.000
	10	1.519	±0.015	1.526		2.000
	3	1.379	±0.010	1.526		2.000
$(0.00, 0.75)$	200	2.988	±0.013	2.951	±0.021	4.000
	100	3.012	±0.020	2.951		4.000
	20	3.130	±0.017	2.951		4.000
	10	3.127	±0.017	2.951		4.000
	3	2.591	±0.025	2.951		4.000
$(0.25, 0.90)$	200	4.335	±0.030	4.240	±0.027	6.250
	100	4.390	±0.019	4.240		6.250
	20	4.357	±0.029	4.240		6.250
	10	3.936	±0.065	4.240		6.250
	3	2.454	±0.043	4.240		6.250

Table 4 **Comparison of (i) the heavy-traffic peakedness in (8), using simulation to estimate J_k and I_2 in (12) and (12), and (ii) the approximation in (17) based on the correlations, to simulation estimates of the peakedness for $M/EARMA/\infty$ examples specified by the parameters (β, ρ). For each model, five arrival rates are considered: $\lambda = 200, 100, 20, 10, 3$. Halfwidths of 95% confidence intervals are shown.**

$M/EARMA/\infty$ model in Table 4. That illustrates that correlations alone are not sufficient for a good peakedness approximation.

6. Approximations for the Delay Probability in the $G/G/n$ Model

We now consider the stationary $G/G/n$ queue with dependent service times, n servers and unlimited waiting room, in which customers enter service from queue in order of arrival. Let W be the steady-state waiting time experienced by an arrival. Formula (1.5) in Whitt (2004) approximates the steady-state probability $P(W > 0)$ in the $G/GI/n$ model, which has i.i.d. service times, by

$$P(W > 0) \approx \alpha(\beta^*/\sqrt{z}), \tag{34}$$

where (i) $\alpha(\beta^*)$ can be taken to be either the exact steady-state probability of delay in the elementary Markovian $M/M/n$ model with same arrival rate, service rate, number of servers and parameter $\beta^* = \sqrt{n}(1 - \rho^*)$, with ρ^* being the traffic intensity, or the many-server heavy-traffic approximation of it, given by the so-called Halfin-Whitt delay function (see Halfin and Whitt (1981)),

$$\alpha(\beta^*) \equiv [1 + \beta^* \Phi(\beta^*)/\phi(\beta^*)]^{-1}, \tag{35}$$

with Φ and ϕ being the standard normal cdf and probability density function (pdf) and (ii) z is the peakedness in the associated $G/GI/\infty$ IS model. As a new approximation for the more general model with dependent service times, here we propose the identical formula (34), but with our new peakedness z accounting for the dependent service times instead of the previous one based on i.i.d. service times.

To evaluate this new approximation, we consider the same EARMA service times as in the previous section, considering the first three cases with two different numbers of servers, with the arrival rate λ chosen according to the many-server heavy-traffic scaling $\lambda = \mu n(1-(\beta^*/\sqrt{n}))$, using $\rho^* \equiv \lambda/n\mu$, where β^* is a quality-of-service (QoS) parameter; see Halfin and Whitt (1981).

EARMA case	(QoS,servers)	simulation	95% conf.	heavy-traffic	percentage
(β,ρ)	(β^*,n)	estimate	interval	approx.	error
$(0.75, 0.50)$	(1, 25)	0.2310	±0.0003	0.2459	6.5%
	(1,400)	0.2411	±0.0002		0.7%
	(0.25, 4)	0.7449	±0.0008	0.7345	−1.4%
	(0.25, 16)	0.7423	±0.0006		−1.1%
$(0.50, 0.50)$	(1, 25)	0.2551	±0.0002	0.2683	5.2%
	(1, 400)	0.2705	±0.0002		−0.8%
	(0.25, 4)	0.7542	±0.0003	0.7472	−0.9%
	(0.25, 16)	0.7555	±0.0005		−1.1%
$(0.50, 0.75)$	(1, 25)	0.2976	±0.0003	0.3009	4.2%
	(1, 400)	0.3142	±0.0005		−1.4%
	(0.25, 4)	0.7627	±0.0003	0.7690	0.8%
	(0.25, 16)	0.7755	±0.0002		−0.8%

Table 5 **Comparison of the approximation for the delay probability in** (34) **using the Halfin-Whitt function** (35) **to simulation estimates for** $M/EARMA/n/\infty$ **examples, using the EARMA service times from the previous subsection for the first three cases of** (β,ρ)**:** $(0.75, 0.50)$**,** $(0.50, 0.50)$ **and** $(0.50, 0.75)$**. The arrival rate was chosen according to the many-server heavy-traffic scaling with QoS parameter** β^* **From Table 4, the peakedness values in these three cases estimated by simulation with** $n = 200$ **(approximation) are, respectively,** 1.116 **(**1.119**),** 1.251 **(**1.249**) and** 1.533 **(**1.526**).**

Table 5 shows the results. Clearly, the accuracy is again remarkably good. Higher n is needed for good accuracy as β^* increases. Anticipated improvement in accuracy is seen as n increases in all three EARMA cases with $\beta^* = 1.0$. This example demonstrates that the new peakedness approximations can be useful for finite-server models.

To put these results in perspective, note that that the delay probabilities based on (34) for the case $z = 1$ (with no dependence) are 0.7209 and 0.2234 when $\beta^* = 0.25$ and 1.0, respectively. With low peakedness, as in the first case with $(\beta,\rho) = (0.75, 0.50)$, there is little difference, but with higher peakedness, as in the third case, the difference is significant. However, the impact is much less than in Fendick et al. (1989).

7. Time-Varying Arrival Rates

To treat transient phenomena and time-varying arrival rates in service systems, we now develop approximations for the time-varying variance $v(t)$ in (5) and the time-varying peakedness $z(t) \equiv v(t)/m(t)$ to go with the mean $m(t)$ in (5) and various approximations for it, such as those based on Taylor series approximations; see Eick et al. (1993b) and Green et al. (2007).

7.1. Exact Expressions

First, we review exact expressions for the time-varying mean from Eick et al. (1993b) and then develop analogs for the time-varying variance. To express these, recall that for any nonnegative random variable S with cdf F and finite mean m_S, we can define a random variable S_e with the *stationary-excess cdf*

$$F_e(x) \equiv P(S_e \le x) \equiv \frac{1}{m_S}\int_0^x F^c(x)\,dx, \tag{36}$$

where $F^c(x) \equiv P(S > x)$ is again the ccdf, which has mean $E[S_e] = m_S(c_S^2+1)/2$ and k^{th} moment

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)m_S}, \quad k \ge 1. \tag{37}$$

Theorem 1 of Eick et al. (1993b) gives two alternative expressions for the mean in (5), namely,

$$m(t) = E\left(\int_{t-S}^{t} \lambda(u)\,du\right) = E[\lambda(t-S_e)]m_S. \tag{38}$$

The first formula in (38) 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) $\lambda(t)m_S$ modified by a random time shift by the stationary-excess random variable S_e.

In Eick et al. (1993b), these expressions are shown to be exact for the number in system at time t in the $M_t/GI/\infty$ model with nonhomogeneous Poisson arrival process (the M_t) and a sequence of i.i.d. service times independent of the arrival process, but the same reasoning shows that the mean formula is also exact for the more general $G_t/G/\infty$ model with the same arrival rate function and stationary service times independent of the arrival process. The extension to G_t arrivals is noted and explained in Remark 2.3 of Massey and Whitt (1993). The formulas then remain exact in the heavy-traffic limit.

It is immediate from the representations (11) and (12) that the same constructions can be used for the time-varying variance formula in (5). We now give an analog for the final relation in (38), with the understanding that an analog of the other can be obtained as well.

Proposition 6. *An alternative* (*exact*) *expression for the time-varying variance* (*of the heavy-traffic limit*) *in* (5) *is*

$$\begin{aligned} v(t) =\ & E[\lambda(t-S_e)]m_S + (c_a^2-1)E[\lambda(t-(S_1 \wedge_{ind} S_2)_e)]E[S_1 \wedge_{ind} S_2] \\ & +2\sum_{k=1}^{\infty}(E[\lambda(t-(S_1 \wedge S_{1+k})_e)]E[S_1 \wedge S_{1+k}] - E[\lambda(t-(S_1 \wedge_{ind} S_2)_e)]E[S_1 \wedge_{ind} S_2]). \end{aligned} \tag{39}$$

The associated (*heavy-traffic*) *peakedness is* $z(t) = v(t)/m(t)$, *combining* (39) *with* (38).

Proof. The key is to write $V(s)$ in (5) in terms of stationary-excess random variables S_e, $(S_1 \wedge_{ind} S_2)_e$ and $(S_1 \wedge S_{1+k})_e$ associated with the random variables S, $S_1 \wedge_{ind} S_2$ and $S_1 \wedge S_{1+k}$, defined as in (36). We get

$$\begin{aligned} V(s) =\ & f_{S_e}(s)E[S] + (c_a^2-1)f_{(S_1 \wedge_{ind} S_2)_e}(s)E[S_1 \wedge_{ind} S_2] \\ & +2\sum_{k=1}^{\infty}\left(f_{(S_1 \wedge S_{1+k})_e}(s)E[S_1 \wedge S_{1+k}] - f_{(S_1 \wedge_{ind} S_2)_e}(s)E[S_1 \wedge_{ind} S_2]\right). \end{aligned} \tag{40}$$

We now can integrate in (5). ■

To provide illustrative examples and insight, it has become standard to study sinusoidal arrival rates. In Eick et al. (1993a) exact formulas are given for the mean with a sinusoidal arrival-rate function. Paralleling §7, we can construct corresponding exact formulas for the time-varying variance. Suppose the arrival rate function is

$$\lambda(t) = \bar{\lambda} + \beta\sin(\omega t), \quad t \ge 0, \tag{41}$$

where $\bar{\lambda}$ is the average arrival rate, β is the *amplitude*, $\beta/\bar{\lambda}$ is the *relative amplitude*, $\omega = 2\pi/T$ is the *frequency* and T is the *period.* Theorem 4.1 of Eick et al. (1993a) gives the following expression for the mean

$$m(t) = \bar{\lambda} m_S + \beta\left(\sin(\omega t) E[\cos(\omega S_e)] - \cos(\omega t) E[\sin(\omega S_e)]\right) m_S. \qquad (42)$$

Following Proposition 6, we obtain a corresponding exact expression for the variance.

Proposition 7. *An alternative* (*exact*) *expression for the time-varying variance* (*of the heavy-traffic limit*) *in* (5) *when the arrival-rate function is sinusoidal as in* (41) *is*

$$\begin{aligned} v(t) = {} & \bar{\lambda} m_S + \beta\Big(\sin(\omega t) E[\cos(\omega S_e)] - \cos(\omega t) E[\sin(\omega S_e)]\Big) m_S \\ & + (c_a^2 - 1)\Big[\bar{\lambda} + \beta\Big(\sin(\omega t) E[\cos(\omega (S_1 \wedge_{ind} S_2)_e)] - \cos(\omega t) E[\sin(\omega (S_1 \wedge_{ind} S_2)_e)]\Big)\Big] E[S_1 \wedge_{ind} S_2] \\ & 2\sum_{k=1}^{\infty}\Bigg(\Big[\bar{\lambda} + \beta\Big(\sin(\omega t) E[\cos(\omega (S_1 \wedge S_{1+k})_e)] \\ & \qquad - \cos(\omega t) E[\sin(\omega (S_1 \wedge S_{1+k})_e)]\Big)\Big] E[S_1 \wedge S_{1+k}] \\ & - \Big[\bar{\lambda} + \beta\left(\sin(\omega t) E[\cos(\omega (S_1 \wedge_{ind} S_2)_e)] - \cos(\omega t) E[\sin(\omega (S_1 \wedge_{ind} S_2)_e)]\right)\Big] E[S_1 \wedge_{ind} S_2]\Bigg) \qquad (43) \end{aligned}$$

The associated (*heavy-traffic*) *peakedness is* $z(t) = v(t)/m(t)$, *combining* (43) *with* (42).

From Proposition 7, we can deduce that the heavy-traffic peakedness can be effectively computed with sinusoidal arrival rates. Paralleling Proposition 2 for the stationary model, we can conclude the following for sinusoidal arrival rates.

Corollary 1. *If the arrival-rate function is sinusoidal as in* (41), *then the heavy-traffic approximations for the time-varying mean, variance and peakedness in* (42) *and Proposition 7 can be computed in terms of the mean values:* $E[S_j]$, $E[S_1 \wedge_{ind} S_2]$, $E[S_j \wedge S_{j+k}]$, $E[\cos(\omega S_j)]$, $E[\cos(\omega(S_1 \wedge_{ind} S_2))]$, $E[\cos(\omega(S_1 \wedge S_{1+k}))]$, $E[\sin(\omega S_j)]$, $E[\sin(\omega(S_1 \wedge_{ind} S_2))]$ *and* $E[\sin(\omega(S_1 \wedge S_{1+k}))]$, $k \geq 1$.

As noted in §5 and §7 of Eick et al. (1993a), nice explicit formulas can be obtained in the case of exponential and hyperexponential service times, because if S is exponential with mean m_S, S_e is distributed the same as S and $E[\sin(\omega S_e)] = m_S \omega/(1 + m_S^2 \omega^2)$ and $E[\cos(\omega S_e)] = 1/(1 + m_S^2 \omega^2)$. From (15) of Eick et al. (1993b), the mean has the expression

$$m(t) = \Big(\bar{\lambda} + \beta(1 + \omega^2 m_S^2)^{-1}\Big(\sin(\omega t) - \omega m_S \cos(\omega t)\Big)\Big) m_S. \qquad (44)$$

Hence, we can obtain the following explicit formula in the case of RRS exponential service times. We exploit equation (16).

Proposition 8. *An alternative* (*exact*) *expression for the time-varying variance* (*of the heavy-traffic limit*) *in* (5) *when the arrival-rate function is sinusoidal as in* (41) *and the service times are RRS exponential with mean* $m_S = 1$ *is*

$$\begin{aligned} v(t) = {} & (1 + 2\Sigma_\rho) m_S \Big(\bar{\lambda} + \beta(1 + \omega^2 m_S^2)^{-1}\Big(\sin(\omega t) - \omega m_S \cos(\omega t)\Big)\Big) \\ & + \frac{1}{2}(c_a^2 - 1 - 2\Sigma_\rho) m_S \Big[\bar{\lambda} + \beta(1 + \omega^2 m_S^2/4)^{-1}\left(\sin(\omega t) - (\omega m_S/2)\cos(\omega t)\right)\Big]. \qquad (45) \end{aligned}$$

The associated (*heavy-traffic*) *peakedness is* $z(t) = v(t)/m(t)$, *combining* (45) *with* (44).

The formulas in Proposition 8 can serve as an approximation based on the correlations alone, paralleling Proposition 3.

7.2. Approximations

Various approximations for the mean are given in Eick et al. (1993b) and reviewed in §4.4 of Green et al. (2007).

Taylor-series approximations. A simple effective approximation for the mean is obtained by applying a Taylor series approximation in the final formula in (38), assuming that the arrival rate is suitably smooth and that the successive derivatives are suitably small so that the Taylor approximation is justified. Eick et al. (1993b) observe that a quadratic variant of a two-derivative approximation is revealing and often effective. It produces a time lag and a space shift, yielding

$$m(t) \approx \lambda(t - E[S_e])m_S + \frac{\lambda''(t)}{2} Var(S_e) m_S. \tag{46}$$

The analog of approximation (46) for $v(t)$ is obtained by again applying a two-term Taylor series approximation to the arrival-rate function λ. For $v(t)$, we obtain the approximation

$$\begin{aligned} v(t) \approx\ & \lambda(t - E[S_e])m_S + (c_a^2 - 1)\lambda(t - E[(S_1 \wedge_{ind} S_2)_e])E[S_1 \wedge_{ind} S_2] \\ & + 2\sum_{k=1}^{\infty} (\lambda(t - E[(S_1 \wedge S_{1+k})_e]E[S_1 \wedge S_{1+k}] - \lambda(t - E[(S_1 \wedge_{ind} S_2)_e]E[S_1 \wedge_{ind} S_2]) \\ & + \frac{\lambda''(t)}{2} Var(S_e) m_S + \frac{(c_a^2 - 1)\lambda''(t)}{2} Var(S_1 \wedge_{ind} S_2)_e])E[S_1 \wedge_{ind} S_2] \\ & + \lambda''(t) \sum_{k=1}^{\infty} (Var(S_1 \wedge S_{1+k})_e)E[S_1 \wedge S_{1+k}] - Var(S_1 \wedge_{ind} S_2)_e)E[S_1 \wedge_{ind} S_2]). \end{aligned} \tag{47}$$

The associated peakedness approximation is $z(t) \approx v(t)/m(t)$ for $v(t)$ in (47) and $m(t)$ in (46).

Approximations based on a recent average arrival rate. It may suffice to even use a more elementary approximation, exploiting the formulas and approximations for the stationary model, after replacing the time-varying arrival rate function in (5) by its time-varying average prior to t. (Ways to apply results for stationary models to describe the average performance of models with periodic arrival rates were proposed in Massey and Whitt (1996).) In particular, we propose the alternative approximations (again assuming that the system starts in the distant past)

$$\begin{aligned} m(t) &\approx \hat{\lambda}(t) \int_0^{\infty} (F^c(s))\, ds = \hat{\lambda}(t) m_S, \\ v(t) &\approx \hat{\lambda}(t) \int_0^{\infty} \left(F^c(s) + (c_a^2 - 1)F^c(s)^2 + \Gamma(s)\right) ds, \\ z(t) &\approx \frac{v(t)}{m(t)} = 1 + (c_a^2 - 1)I_1 + I_2 \end{aligned} \tag{48}$$

for $t \geq 0$, where

$$\hat{\lambda}(t) \equiv \int_0^{\infty} \lambda(t - s)\delta e^{-\delta s}\, ds, \tag{49}$$

with δ 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. From these formulas, we can deduce that the stationary model approaches steady state over a few service times; e.g., see (20) in Eick et al. (1993b). It is significant that, the approximate peakedness in (48) is not time-varying; it has precisely the same form as in Proposition 1. Thus, even though the variance may be strongly time-varying, we expect its fluctuations to be largely cancelled out by the mean. Equation (48) tells us to expect that the peakedness should be nearly constant, assuming nearly the same value as in the case of a constant arrival rate.

Paralleling the mean in Eick et al. (1993b), the simple stationary approximation (SSA) for $v(t)$ replaces $\hat{\lambda}(t)$ in (48) by the long-run average $\bar{\lambda}$; the PSA replaces $\hat{\lambda}(t)$ in (48) by $\lambda(t)$. Both SSA and PSA predict a constant peakedness, just as in (48).

7.3. Random Repeated Service Times

It is significant that we can directly compute all the quantities in approximation (47) in the setting of §4. To evaluate the terms $E[S_1 \wedge S_{1+k}]$, $E[(S_1 \wedge S_{1+k})_e]$ and $Var((S_1 \wedge S_{1+k})_e)$, we can exploit the first line of (14), noting that the random variable $S_1 \wedge S_{1+k}$ is the mixture of two random variables, having pdf

$$f_{S_1 \wedge S_{1+k}}(x) = \rho_k f_{S_1}(x) + (1-\rho_k) f_{S_1 \wedge_{ind} S_2}(x), \quad x \geq 0. \tag{50}$$

Hence, the first three moments are

$$\begin{aligned} E[S_1 \wedge S_{1+k}] &= \rho_k E[S_1] + (1-\rho_k) E[S_1 \wedge_{ind} S_2], \\ E[(S_1 \wedge S_{1+k})^2] &= \rho_k E[S_1^2] + (1-\rho_k) E[(S_1 \wedge_{ind} S_2)^2], \\ E[(S_1 \wedge S_{1+k})^3] &= \rho_k E[S_1^3] + (1-\rho_k) E[(S_1 \wedge_{ind} S_2)^3], \end{aligned} \tag{51}$$

where, by integration by parts, p. 150 of Feller (1971),

$$\begin{aligned} E[S_1 \wedge_{ind} S_2] &= \int_0^\infty F^c(s)^2\, ds, \\ E[(S_1 \wedge_{ind} S_2)^2] &= 2\int_0^\infty sF^c(s)^2\, ds \\ E[(S_1 \wedge_{ind} S_2)^3] &= 3\int_0^\infty s^2 F^c(s)^2\, ds; \end{aligned} \tag{52}$$

The moments of the associated stationary-excess random variable are then determined by (37).

A relatively simple case arises when S is exponential with mean m. Then $S_1 \wedge_{ind} S_2$ is exponential with mean $m/2$, and so $S_1 \wedge S_{1+k}$ is H_2 (hyperexponential of order 2, a mixture of two exponentials), taking the value of an exponential with mean m with probability ρ_k and taking the value of another exponential with mean $m/2$ with probability $1-\rho_k$. Moreover, the stationary-excess distribution associated with an exponential is again that same exponential, whereas the stationary-excess cdf associated with an H_2 cdf is again H_2, with the same mixing probabilities but new exponential means. In this case, the first three moments of S_1 are m, $2m^2$ and $6m^3$; the first three moments of $S_1 \wedge_{ind} S_2$ are $m/2$, $m^2/2$ and $3m^3/4$; and the first three moments of $S_1 \wedge S_{1+k}$ are $(1+\rho_k)m/2$, $(1+3\rho_k)m^2/2$ and $(3+21\rho_k)m^3/4$. Finally, $E[(S_1 \wedge S_{1+k})_e] = (1+3\rho_k)m/(2(1+\rho_k))$ and $E[((S_1 \wedge S_{1+k})_e)^2] = (3+21\rho_k)m^2/(6(1+\rho_k))$.

We now elaborate on how to compute the last term in (47) in the case of random repeated service times when S is exponential. Suppose that $m=1$. Then

$$\begin{aligned} Var((S_1 \wedge S_{1+k})_e) &= \frac{1+10\rho_k+5\rho_k^2}{4(1+\rho_k)^2} \quad \text{and} \\ Var((S_1 \wedge_{ind} S_2)_e) &= Var(S_1 \wedge_{ind} S_2) = \frac{1}{4}, \end{aligned} \tag{53}$$

so that

$$\begin{aligned} U_k &\equiv Var((S_1 \wedge S_{1+k})_e) E[(S_1 \wedge_{dep} S_{1+k})] = \frac{1+10\rho_k+5\rho_k^2}{8(1+\rho_k)}, \\ W &\equiv Var((S_1 \wedge_{ind} S_2)_e) E[S_1 \wedge_{ind} S_2] = \frac{1}{8} \quad \text{and} \\ U_k - W &= \left(\frac{\rho_k}{8}\right)\left(\frac{9+5\rho_k}{1+\rho_k}\right). \end{aligned} \tag{54}$$

Hence, the sum in the last term of (47) is finite when $\sum_{k=1}^\infty \rho_k < \infty$; i.e., the last term is

$$\lambda^{''}(t)\sum_{k=1}^{\infty}(U_k - W) < \infty, \tag{55}$$

where $U_k - W$ is given in (54) above.

8. Simulation Experiments for Sinusoidal Arrival Rates

We now conduct simulation experiments to evaluate the approximations with time-varying arrival rates. We restrict attention to sinusoidal arrival rates as in (41), letting the relative amplitude be fixed at $\beta/\bar{\lambda} = 0.25$.

Our first example is for the $M_t/RRS/\infty$ model, with nonhomogeneous Poisson arrival process and the same randomly repeated exponential service times with mean 1 as in Table 1. We consider 4 cases of the single parameter p: 0.10, 0.25, 0.50 and 0.75. For each p, we let the average arrival rate $\bar{\lambda}$ coincide with the larger constant arrival rate in Table 1, yielding about 1% error with the constant arrival rates. As a base case, we consider a period $T = 2\pi/\omega = 10$. Afterwards, we consider the much longer period $T = 50$ and the much shorter period $T = 1$.

We consider four different approximations for the time-varying variance $v(t)$. We first consider the exact heavy-traffic value in (45) (and, more generally, in (43)). We also consider the "recent" approximation in (48) using $\delta = 1/E[S_e] = 1/E[S] = 1$, the Taylor series approximation in (47) and the pointwise stationary approximation (PSA), obtained by replacing $\hat{\lambda}(t)$ in (48) by $\lambda(t)$.

We show the results for the single case $p = 0.75$, $\bar{\lambda} = 800$ and $T = 10$ in Figure 1. As before, the simulation is conducted using multiple replications, without any averaging over time. Further independent replications were used to confirm that the confidence 95% intervals are about 1%. From Figure 1, it is apparent that the new approximations are better than PSA, with exact and recent approximations in (45) and (48) being noticeably better than the Taylor approximation in (47).

Table 6 gives a mote details for other cases, all with period $T = 10$. Table 6 shows that the approximation effectiveness has the clear ordering $Exact > Recent > Taylor > PSA$, where $>$ means "better than." The average absolute errors for PSA are quite large, about 10%, but all the other approximations are quite effective. As in Table 1, the approximations consistently overestimate the actual values, but only by a relatively small value, ranging from 0.5% to 1.5%.

We next consider the longer period $T = 50$ in Table 7; the other parameters are the same. In this case with very slowly varying arrival rate, the system can be regarded as being approximately in a "local" steady state at each time, so that it is natural to use PSA, and it performs quite well. Nevertheless, even in this case, Table 7 shows that the other methods are superior to the PSA approximation (even though the difference is not likely to matter for applications). In this case, the Taylor series approximation in (47) gets better than in Table 6, performing slightly better than the Recent approximation in (48).

We now turn to the case of very short periods, letting $T = 1$, but keeping all other parameters the same. Table 8 reports results for the sinusoidal arrival rate with period $T = 1$. With such short periods, the Taylor series approximation makes no sense at all. It should not be surprising that it gives exceptionally bad results. That is because the successive derivatives fail to get smaller. In this case, the PSA approximation is also very bad. The average absolute errors for PSA in Table 8 are approximately 10 times bigger than with the exact heavy traffic approximation in (45). The recent approximation in (48) performs quite well here; it is only slightly worse for lower values of p, but has twice the average absolute error for $p = 0.75$.

The recent approximation in (48), which produces constant approximate peakedness, indicates that there should be benefits from looking at the time-varying peakedness instead of the time-varying variance. Approximation (48) suggests that errors in the time-varying mean and variance may cancel when we divide. We thus look at direct estimates for the time-varying peakedness for the same case in Figure 2 and Table 6 when the period is $T = 10$. Paralleling Figure 1, we show the results for the single case $p = 0.75$, $\bar{\lambda} = 800$ and $T = 10$ in Figure 2. Figure 2 shows that the peakedness values for this case are all within ±8% of the average value, and mostly within ±4%, so that the constant approximation should not be so bad. More detailed results for more cases, all with period $T = 10$, are shown in Table 9.

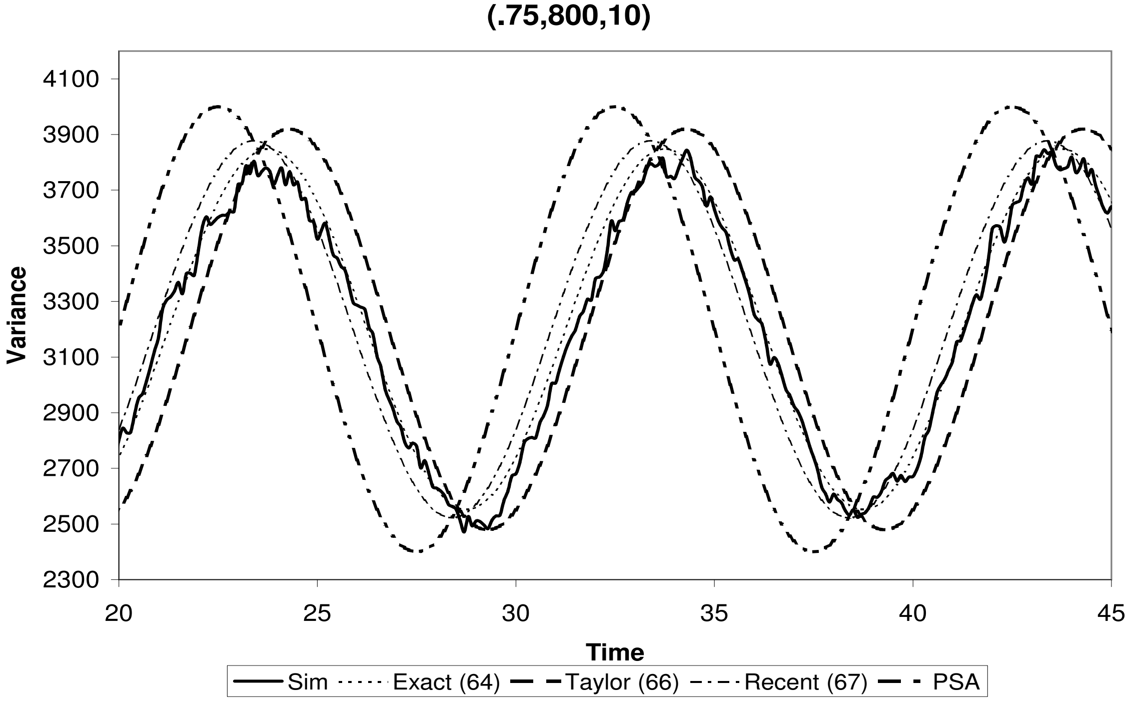

Figure 1 **Comparison of the approximations for the time-varying variance $v(t)$ in the $M_t/RRS/\infty$ model with sinusoidal arrival rates to simulation estimates. Here the RRS service times have mean 1 and parameter $p = 0.75$. The average arrival rate is $\bar{\lambda} = 800$, the relative amplitude is $\beta/\bar{\lambda} = 0.25$ and the period is $T = 2\pi/\omega = 10$.**

In Table 9, we use the same method for approximating the mean as we use for the variance, in order to increase the likelihood of errors cancelling. That is shown directly for the recent and PSA approximations in (48). Thus, the Exact peakedness approximation is the variance in (45) divided by the corresponding exact mean in (44), while the Taylor approximation is the Taylor variance approximation in (47) divided by the corresponding Taylor approximation for the mean in (46). In a separate detailed comparison, we found that this Taylor approximation for the peakedness performs substantially worse when we divide by the exact mean in (44) than when we divide by the Taylor mean in (46). The simulation estimate for the peakedness is the estimated variance divided by the exact mean (which is exact in the stochastic model, as noted before).

9. Conclusions

In this paper we have explored the applied consequences of a recent heavy-traffic limit established by Pang and Whitt (2011). That research tells us that, under regularity conditions, the distribution of the number of customers in an infinite-server (IS) model at time t becomes approximately Gaussian as the arrival rate increases, even in the presence of dependence among the service times. Moreover, the dependence among the service times affects the normal distribution only through the variance or, equivalently, the peakedness (the variance divided by the mean).

Parameters	Approximation	Errors		
$(p,\bar{\lambda},T)$	Method	Avg.	Avg. Abs.	Max. Abs.
$(0.10, 10, 10)$	Exact HT (45)	0.14	0.19	0.66
$(\bar{v} \approx 10.97)$	Recent (48)	0.14	0.20	0.69
	Taylor (47)	0.14	0.38	1.08
	PSA	0.14	1.00	1.99
$(0.25, 25, 10)$	Exact HT (45)	0.47	0.56	1.95
$(\bar{v} \approx 32.85)$	Recent (48)	0.47	0.60	2.15
	Taylor (47)	0.47	1.25	3.76
	PSA	0.48	3.15	6.50
$(0.50, 100, 10)$	Exact HT (45)	1.59	2.64	9.34
$(\bar{v} \approx 198.3)$	Recent (48)	1.60	4.42	13.14
	Taylor (47)	1.56	8.74	21.32
	PSA	1.65	20.73	38.80
$(0.75, 800, 10)$	Exact HT (45)	16.03	47.12	169.02
$(\bar{v} \approx 3183)$	Recent (48)	16.26	114.28	296.19
	Taylor (47)	15.14	148.67	372.24
	PSA	17.14	373.55	691.20

Table 6 **Comparison of the approximations for the time-varying variance $v(t)$ in the $M_t/RRS/\infty$ model with sinusoidal arrival rates to simulation estimates. As in Table 1, we use the single-parameter randomly repeated exponential service times with mean 1 and parameter p, here considering 4 values of p. The average variance $\bar{v}$ is shown in each case. The sinusoidal arrival rate function is as in (41) with relative amplitude $\beta/\bar{\lambda} = 0.25$ and period $T = 2\pi/\omega = 10$. We let the average arrival rate $\bar{\lambda}$ for each p be the higher level in Table 1, yielding about 1% error in Table 1.**

Parameters	Approximation	Errors		
$(p,\bar{\lambda},T)$	Method	Avg.	Avg. Abs.	Max. Abs.
$(0.10, 10, 50)$	Exact HT (45)	0.11	0.17	0.79
$(\bar{v} \approx 10.54)$	Recent (48)	0.11	0.17	0.79
	Taylor (47)	0.11	0.17	0.79
	PSA	0.08	0.28	0.86
$(0.25, 25, 50)$	Exact HT (45)	0.36	0.47	1.75
$(\bar{v} \approx 31.58)$	Recent (48)	0.35	0.46	1.76
	Taylor (47)	0.36	0.48	1.77
	PSA	0.27	0.74	2.77
$(0.50, 100, 50)$	Exact HT (45)	1.67	2.55	9.34
$(\bar{v} \approx 190.1)$	Recent (48)	1.53	2.71	13.14
	Taylor eqnnew42	1.68	2.55	21.32
	PSA	1.02	5.52	38.80
$(0.75, 800, 50)$	Exact HT (45)	4.66	41.00	195.29
$(\bar{v} \approx 3064)$	Recent (48)	1.25	47.11	204.10
	Taylor (47)	4.44	40.97	194.33
	PSA	−6.83	96.06	250.83

Table 7 **Comparison of the approximations for the time-varying variance $v(t)$ in the $M_t/RRS/\infty$ model with slowly varying sinusoidal arrival rates to simulation estimates. The model is the same as in Table 6 except now the period T is changed from 10 to 50.**

Here we have shown that the exact heavy-traffic variance and peakedness with dependent service times can often be effectively computed. Explicit expressions are given for the general stationary

Parameters	Approximation	Errors		
$(p,\bar{\lambda},T)$	Method	Avg.	Avg. Abs.	Max. Abs.
$(0.10,10,1)$	Exact HT (45)	0.08	0.16	0.57
$(\bar{v}\approx 11.03)$	Recent (48)	0.08	0.16	0.57
	Taylor (47)	0.13	36.44	56.89
	PSA	0.08	1.76	3.15
$(0.25,25,1)$	Exact HT (45)	0.47	0.57	2.19
$(\bar{v}\approx 32.86)$	Recent (48)	0.47	0.59	2.04
	Taylor (47)	0.64	124.86	194.87
	PSA	0.46	5.28	9.84
$(0.50,100,1)$	Exact HT (45)	2.06	3.22	10.85
$(\bar{v}\approx 197.9)$	Recent (48)	2.05	3.79	11.97
	Taylor (47)	3.32	881.94	1373.84
	PSA	2.03	31.92	60.19
$(0.75,800,1)$	Exact HT (45)	11.20	39.13	147.83
$(\bar{v}\approx 3188)$	Recent (48)	10.99	64.57	213.50
	Taylor (47)	34.05	16077.45	24978.46
	PSA	10.70	516.70	906.70

Table 8 **Comparison of the approximations for the time-varying variance $v(t)$ in the $M_t/RRS/\infty$ model with rapidly varying sinusoidal arrival rates to simulation estimates. The model is the same as in Table 6 except now the period is $T=1$.**

Parameters	Approximation	Errors		
$(p,\bar{\lambda},T)$	Method	Avg.	Avg. Abs.	Max. Abs.
$(0.10,10,10)$	Exact HT (45)	0.013	0.019	0.060
$(\bar{z}\approx 1.100)$	Taylor (46), (47)	0.013	0.019	0.057
	Recent (48)&PSA	0.013	0.019	0.057
$(0.25,25,10)$	Exact HT (45)	0.018	0.023	0.079
$(\bar{z}\approx 1.315)$	Taylor (46), (47)	0.018	0.024	0.093
	Recent (48)&PSA	0.018	0.024	0.091
$(0.50,100,10)$	Exact HT (45)	0.016	0.027	0.103
$(\bar{z}\approx 1.986)$	Taylor (46), (47)	0.014	0.038	0.130
	Recent (48)&PSA	0.014	0.044	0.127
$(0.75,800,10)$	Exact HT (45)	0.0067	0.054	0.215
$(\bar{z}\approx 3.988)$	Taylor (46), (47)	0.0062	0.054	0.215
	Recent (48)&PSA	0.0022	0.062	0.205

Table 9 **Comparison of the approximations for the time-varying peakedness $z(t)$ in the $M_t/RRS/\infty$ model with sinusoidal arrival rates to simulation estimates. We use the designated approximation for the variance, divided by the time-varying mean in (44). As in Table 1, we use the single-parameter randomly repeated exponential service times with parameter p, here considering 4 values of p. The average peakedness $\bar{z}$ is shown in each case. The sinusoidal arrival rate function is as in (41) with relative amplitude $\beta/\bar{\lambda}=0.25$ and period $T=2\pi/\omega=10$. We let the average arrival rate $\bar{\lambda}$ for each p be the higher level in Table 1, yielding 1% error in Table 1.**

model in Propositions 1 and 2, the stationary model with randomly repeated service times in Propositions 3 and 4, and for the model with time-varying arrival rates in Propositions 6 and 7, the last being for the case of a sinusoidal arrival rate. Simulation experiments for the stationary model in §5 and for the case of sinusoidal arrival rates in §8 show that the explicit heavy-traffic expressions are quite accurate for the examples of EARMA service times and randomly repeated

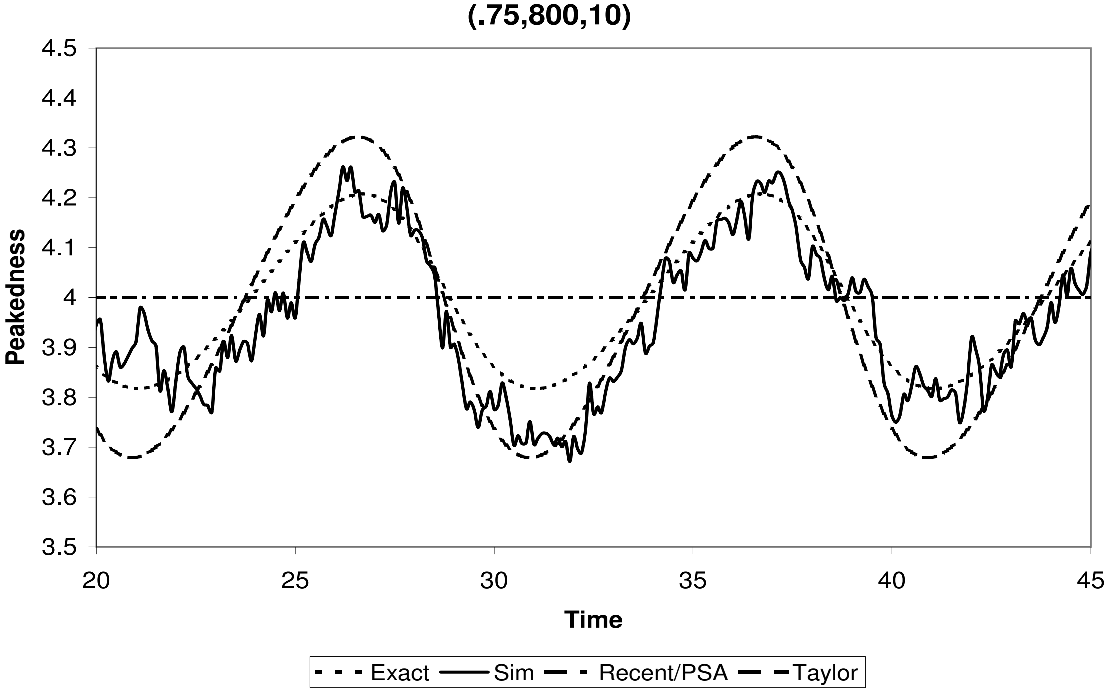

Figure 2 **Comparison of the approximations for the time-varying variance $v(t)$ in the $M_t/RRS/\infty$ model with sinusoidal arrival rates to simulation estimates. Here the RRS service times have mean 1 and parameter $p = 0.75$. The average arrival rate is $\bar{\lambda} = 800$, the relative amplitude is $\beta/\bar{\lambda} = 0.25$ and the period is $T = 2\pi/\omega = 10$.**

service (RRS) times. The simulation experiment in §5.2 shows that the approximations remain effective with dependence among the interarrival times as well as the service times and for non-exponential marginal distributions. These experiments show that the dependence can have a big impact, with the impact increasing in the amount of dependence.

Since the explicit heavy-traffic formulas are complicated, it is of interest to develop even more elementary rough approximations that can provide insight. From that perspective, the approximations based on the correlations alone in Proposition 3 for the stationary model and in Proposition 8 for sinusoidal arrival rates are especially interesting. The contributions of the dependence to the peakedness are clearly visible from these formulas. It is significant that these approximations are accurate for the RRS model, for which they are the exact heavy-traffic formulas. However, Tables 4 and 1 showing results for EARMA and RRS service times shows that the approximation based on correlations alone only provide a rough approximation more generally.

For the case of time-varying arrival rate, we found that the constant approximate peakedness of the corresponding stationary model performs remarkably well. This is the "recent" approximation in (48). Table 9 shows that its performance is not so much below the exact heavy-traffic peakedness (which is not constant). More generally, approximation errors in the mean and variance tend to cancel when dividing to calculate the peakedness; the Taylor approximation performs better when dividing by the Taylor approximation for the mean than when dividing by the exact mean.

There are many directions for future research. In a sequel to this paper, Pang and Whitt (2012), we ourselves have examined an alternative model with batch arrivals and dependence among service times only within the same batch. It remains to consider the performance impact of the dependence between arrival times and service times. In Fendick et al. (1989) all three forms of dependence were shown to play a significant role, even though the dependence between interarrival times and service times was least important. In §6 we showed that the new peakedness approximations with dependent service times can be used to extend previous approximations for finite-servers based on peakedness for independent service times, but it remains to more thoroughly explore such approximations for queues with only finitely many servers. It remains to consider more general network models, with dependence among the service times at different queues. It remains to conduct empirical studies to estimate the level of dependence among service times in applications.

Acknowledgments
The authors thank Columbia University undergraduate Andrew Li for his valuable assistance, both in conducting the simulation experiments reported here and in providing helpful comments on the paper. The second author was supported by NSF CMMI Grants 0948190 and 1066372.

References

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

Berkes, I., S. Hörmann, J. Schauer. 2009. Asymptotic results for the empirical process of stationary sequences. *Stoch. Proc. Appls.* **119** 1298–1324.

Berkes, I., W. Philipp. 1977. An almost sure invariance principle for empirical distribution function of mixing random variables. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **41** 115–137.

Eckberg, A. E. 1983. Generalized peakedness of teletraffic processes. *Proceedings of* 10th *International Teletraffic Congress*. Montreal, Canada.

Eick, S. G., W. A. Massey, W. Whitt. 1993a. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** 241–252.

Eick, S. G., W. A. Massey, W. Whitt. 1993b. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731–742.

Falin, G. 1994. The $M^k/G/\infty$ batch arrival queue with heterogeneous dependent demands. *J. Appl. Prob.* **31** 841–846.

Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2) 324–338.

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*. Second edition ed. John Wiley, New York.

Fendick, K. W., V. R. Saksena, W. Whitt. 1989. Dependence in packet queues. *IEEE Trans. Commun.* **37** 1173–1183.

Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16** 13–29.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Jacobs, P. A. 1980. Heavy traffic results for single-server queues with dependent (EARMA) service and interarrival times. *Adv. Appl. Prob.* **12** 517–529.

Jacobs, P. A., P. A. W. Lewis. 1977. A mixed autoregressive-moving average exponential sequence and point process (EARMA 1,1). *Adv. Appl. Prob.* **9** 87–104.

Jacobs, P. A., P. A. W. Lewis. 1978. Discrete time series generated by mixtures i. correlational and run properties. *J. Roy. Statist., Ser. B* **40** 94–105.

Jacobs, P. A., P. A. W. Lewis. 1983. Stationary discrete moving average autoregressive time series generated by mixtures. *J. Times Series Analysis* **4** 19–36.

Jagerman, D. L., B. Melamed. 1994. Burstiness descriptors of traffic streams: indices of dispersion and peakedness. *Proceedings of the 1994 Conference on Information Sciences and Systems*. Princeton, New Jersey.

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

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

Lawrence, A. J., P. A. W. Lewis. 1980. The exponential autoregressive-moving average EARMA(p,q) process. *J. Roy. Stat. Soc. Ser. B* **42**(2).

Mark, B. L., D.L. Jagerman, G. Ramamurthy. 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.

Massey, W. A., W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* **13**(1) 183–250.

Massey, W. A., W. Whitt. 1996. Stationary-process approximations for the nonstationary Erlang loss model. *Oper. Res.* **44**(6) 976–983.

Nelson, B. L., M. R. Taaffe. 2004. The $Ph_t/Ph_t/\infty$ queueing system: Part i - the single node. *INFORMS J. Computing* **16**(3) 266–274.

Pang, G., W. Whitt. 2010. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65** 325–364.

Pang, G., W. Whitt. 2011. Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems* .

Pang, G., W. Whitt. 2012. Infinite-server queues with batch arrivals and dependent service times. Tech. rep., Columbia University, http://www.columbia.edu/~ww2040/allpapers.html.

Sim, C. H. 1990. First-order autoregressive models for gamma and exponential processes. *J. Appl. Prob.* **27**(2).

Whitt, W. 1976. Bivariate distributions with given marginals. *Ann. Statistics* **4** 1280–1289.

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

Whitt, W. 1983. Comparing batch delays and customer delays. *Bell System Tech. J.* **62** 2001–2009.

Whitt, W. 1984. Heavy traffic approximations for service systems with blocking. *AT&T Bell Lab. Tech. J.* **63** 689–708.

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

Whitt, W. 2004. A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* **52** 689–708.