Performance Evaluation 80 (2014) 82–101

Contents lists available at ScienceDirect

Performance Evaluation

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

Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed

Andrew A. Li *, Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, United States

ARTICLE INFO

Article history:
Available online 13 August 2013

Keywords:
Loss models
Blocking probability
Infinite-server queueing models
Heavy traffic
Time congestion
Parcel blocking

ABSTRACT

Effective approximations are developed for the blocking probability in a general stationary loss model, where key independence and exponential-distribution assumptions are relaxed, giving special attention to dependence among successive service times, not studied before. The new approximations exploit recent heavy-traffic limits for the steady-state number of busy servers in the associated infinite-server model with the same arrival and service processes. In addition, a new heavy-traffic approximation is developed for the long-run proportion of time that all servers are busy. These new approximations are then combined to develop new approximations for the separate blocking probabilities of individual arrival streams in multi-class loss models. Simulation experiments show that these approximations are effective.

1. Introduction

There is growing interest in the performance evaluation of service systems, such as hospitals and call centers. Just as for communication systems, stochastic loss models are proving to be useful; they have been used in healthcare to manage the overflow of intensive care patients [1], to model ambulance deployment [2], to determine the required number of hospital beds [3] and to plan capacity for neonatal units [4]. They have also been used in revenue management to model reusable resources such as hotel rooms and rental vehicles [5].

In this paper, we study the stationary $G/G/s/0$ loss model, allowing non-exponential distributions and dependence among successive interarrival times and among successive service times. Throughout we assume that the service times are independent of the arrival process. Since we allow dependence, we include an I in the Kendall queueing model notation if either the interarrival times or service times are i.i.d. Since the steady-state distribution of the number of busy servers in the $MI/GI/s/0$ loss model is insensitive to the service-time distribution beyond its mean (Section 5.7.2 of [6]), an extension to general service-time distributions alone is not needed. It is well known that a non-Poisson arrival process alters the blocking probabilities. The $GI/MI/s/0$ loss model has also been thoroughly analyzed [7,8], but other model generalizations make exact analysis difficult; otherwise simulation can be used.

There is a long history of studies developing approximations for the blocking probability in loss models with general non-Poisson arrival processes, primarily motivated by the bursty arrival processes arising when overflows from one or more loss systems are forwarded to receive service at a secondary loss system [9,10]. Such overflows commonly occur in alternative routing schemes, in which the traffic that finds no capacity on an initial path is allowed to seek capacity on a succession of alternative paths. These overflows have a big impact on the blocking. Since the overflows only occur when the initial system is full, they tend to occur in clumps, making the overflow process more “bursty” than a Poisson process. An effective

Abbreviations: Infinite server, (IS); Heavy traffic, (HT); Quality-and-efficiency-driven, (QED).

* Correspondence to: Operations Research Center, MIT, Cambridge, MA 02139, United States. Tel.: +1 845 392 4586.
E-mail addresses: aali@mit.edu (A.A. Li), ww2040@columbia.edu (W. Whitt).

0166-5316/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.peva.2013.08.004

early approximation scheme is the equivalent random method [10], which makes use of the specific structure of an overflow process of a Poisson process to a secondary system with service times distributed the same as in the initial system. Overflows can play a key role in service applications as well; e.g., hotel booking services suggest alternative hotels when a current selection is unavailable.

1.1. The first contribution: dependent service times

The substantial literature on approximations for the blocking probabilities in loss models with general non-Poisson arrival processes is restricted to the $G/GI/s/0$ model. Our first contribution here is to develop and evaluate approximate performance formulas for $G/G/s/0$ loss models, which include dependent service times. We use simulation to evaluate the accuracy.

Dependence among service times can occur in a variety of settings. For example, response times of ambulances from a centrally located base (as studied in [2]) can be influenced by the previous responses if an ambulance is sent to respond to a call directly from a hospital rather than first returning to the base. In the management of a reusable resource like hotel rooms (as studied in [5]), multiple reservations may be for some major event located near the hotel and are therefore similar in length.

It thus can be important to be able to predict the impact of dependence among service times on loss model performance. It is even necessary to ask how the dependence should be characterized, because dependence is a complicated notion. Our characterization of the dependence will exploit the familiar correlations among successive service times, but it remains to reduce this sequence of measurements to single measurements that are important for performance, and it remains to see how the dependence in the service times interacts with the arrival process. Measurements of service systems have been receiving considerable attention, e.g., see [11] and the many citations to it, but dependence evidently has not yet been considered, evidently because (i) a convenient characterization of dependence has been lacking and (ii) it was not known how it could be used in performance evaluation.

Our first contribution here is to show that a heavy-traffic limit theorem for the corresponding $G/G/\infty$ infinite-server model from [12], stated as Theorem 1 here, can be used to obtain a concise partial characterization of the dependence among the service times and its impact upon the performance of the loss model. Previously, Pang and Whitt [13] showed that the HT peakedness could be used to generate effective approximations for associated delay models with dependent service times.

By exploiting the associated infinite-server model with the same arrival and service processes, we follow a long tradition in teletraffic engineering [9,14–22]. We too use the *peakedness*, the ratio of the variance to the mean of the steady-state number of busy servers in the IS model. In particular, as in [19,20,12,13], we use the heavy-traffic (HT) limit of the peakedness, which appears in formula (3). The HT limit is crucial for producing a relatively parsimonious characterization of the dependence, capturing the main impact on performance.

We show that the HT peakedness can be used to generate effective approximations for the performance of loss models with dependent service times provided that the number of servers is not too small. In particular, we find that both the normal approximation obtained from the HT limit for the IS model (Section 4.1) and the Hayward approximation (Section 4.2) continue to produce quite accurate approximations for the blocking probability with dependent service times. In support, we show that these two approximations are asymptotically equivalent as the scale increases (in the QED regime; see Theorem 5 and following discussion). This too, while not difficult, is an important contribution, because it helps unify the literature.

1.2. The second contribution: time congestion and parcel blocking probabilities

We also consider the $\sum_i G_i/G/s/0$ model with multiple independent arrival streams and approximate the separate blocking probability of each stream, called the *parcel blocking probabilities*. As a basis for the parcel blocking approximation, and for its own sake, we also study the long-run proportion of time that all servers are busy in the general $G/G/s/0$ model or, equivalently, the probability that all servers are busy at an arbitrary time, often called the *time congestion*. The time congestion directly describes the system as seen by an outside observer, and approximately describes the blocking experienced by a class of rare arrivals that itself contributes negligibly to the overall system performance. By the celebrated *Poisson Arrivals See Time Averages* (PASTA) property [23], the time congestion coincides with the call congestion (blocking probability) when the arrival process is Poisson, but it does not more generally. Thus, the time congestion and parcel blocking require additional analysis and have received considerable attention [24,16,25,26,21], but not yet for dependent service times.

We expose a defect in the literature on heavy-traffic approximations for loss systems, evidently not exposed before. Our simulation experiments show that an associated normal approximation for the time congestion ((19) here) from [19], based on a HT limit theorem stated without proof in [27], performs badly for higher blocking probabilities with higher peakedness, even with i.i.d. exponential service times. Our second contribution is to, first, establish a new (different) HT limit for the time congestion in the $GI/MI/s/0$ model (Corollary 4), which implies that the theorem stated without proof in [27] is in fact invalid, and, second, develop a new approximation for the time congestion based on an approximation for the ratio of the call congestion to the time congestion, called the *congestion ratio*. We show that our new approximation remedies this problem and we apply the new approximation for the time congestion to obtain corresponding good approximations for parcel blocking probabilities.

1.3. The third contribution: review of exact peakedness and a new approximation

We also provide an overview of the performance approximations for loss models based on peakedness, which are restricted to independent service times. We review the exact peakedness in the $G/GI/s/0$ model from Eckberg [17] (Theorem 2 below) and give explicit formulas for special cases. We then establish a new refined second-order HT approximation for the peakedness in the $GI/MI/\infty$ model (Theorem 3 below), which shows when the ordinary HT approximation should perform well and when it should break down.

1.4. Organization of the paper

We begin in Section 2 with a review of infinite-server results that we will use in our multi-server loss approximations, including the recent heavy-traffic result by [12] and the peakedness properties of the general $G/GI/\infty$ model developed by Eckberg [17]. In Section 3 we review two models of dependent random variables that we use for service times and interarrival times. In Section 4, we review two approximations for the blocking probability in loss models based on peakedness; then in Section 5 we use simulation to evaluate their performance in our more general setting. In Section 6, we expose the shortcomings of the existing HT approximation for the time congestion and propose a significantly better approximation based on the congestion ratio. Finally, in Section 7 we develop and evaluate the approximation for the parcel blocking probabilities. Conclusions are drawn in Section 8. Additional material appears in the appendices.

2. Review of infinite-server results

We will develop effective approximations for the steady-state blocking probability in the stationary $G/G/s/0$ loss model, which has a sequence of stationary and possibly dependent service times, each with mean μ^{-1}, that is independent of a general stationary arrival process with arrival rate λ. To do so, we will exploit the steady-state number of busy servers, N, in the corresponding stationary $G/G/\infty$ infinite-server (IS) model with the same arrival process and service times. By Little's law, $E[N] = \alpha \equiv \lambda/\mu$, the *offered load*. In addition to the steady-state mean $E[N]$, we will exploit the steady-state variance $\mathrm{Var}(N)$ via the ratio $z^e \equiv z^e_{G/G} \equiv \mathrm{Var}(N)/E[N]$, which is called the *peakedness*. (The superscript e denotes "exact"; we will use z for the heavy-traffic approximation in Section 2.1. We will relate z and z^e.) The reference case is the $MI/GI/\infty$ model, where N has a Poisson distribution with mean, and thus variance, equal to α. Thus, the peakedness in the reference case is $z^e_{MI/GI} = 1$. Assuming i.i.d. service times, arrival processes with $z^e_{G/GI} > 1$ such as overflow processes are called "bursty", while arrival processes with $0 \leq z^e_{G/GI} < 1$ are called "smooth", but more generally the peakedness depends on both the arrival process and the service times.

2.1. The heavy-traffic peakedness

The heavy-traffic (HT) peakedness is the limit of the peakedness as the arrival rate λ is allowed to increase or, equivalently, as the mean service time μ^{-1} or the offered load α is allowed to increase. Let $A(t)$ count the number of arrivals in the interval $[0, t]$ and let S be a generic service time. It is necessary to specify how the models change in such a limit; we assume that these changes occur by simple scaling. In particular, starting with a rate-1 arrival process, denoted by $A^{(1)}(t)$ and service times that have mean 1, denoted by $S^{(1)}$, we consider the associated scaled arrival process $A^{(\lambda)}(t) \equiv A^{(1)}(\lambda t)$, which has rate λ, and the associated scaled service times with $S^{(\mu)} \equiv S^{(1)}/\mu$, which has mean $1/\mu$. In this setting, the HT approximation is obtained as $\lambda \uparrow \infty$, as $\mu \downarrow 0$ or as $\alpha \uparrow \infty$.

The HT peakedness can be obtained as a consequence of a more general stochastic HT limit theorem. For the most general $G/G/\infty$ model, we rely on recent results in [12], where references can be found to the previous results for the $G/GI/\infty$ special case, the seminal one being by Borovkov [28]. For the HT limit theorem, it suffices to assume only very general regularity conditions. Very roughly, it suffices to assume that a functional central limit theorem is valid for the arrival process and service times separately. For practical purposes, this means that a normal approximation is valid for the arrival process over large intervals, i.e.,

$$\frac{A^{(\lambda)}(t) - \lambda t}{\sqrt{\lambda c_a^2 t}} \approx \mathcal{N}(0, 1) \quad \text{for all sufficiently large } t, \tag{1}$$

where c_a^2 is a constant characterizing the variability (in the limit), $\mathcal{N}(\mu, \sigma^2)$ is a normal random variable with mean μ and variance σ^2, and $\approx$ means approximately equal in distribution. For a renewal arrival process, the variability parameter c_a^2 is the *squared coefficient of variation* (SCV, variance divided by the square of the mean) of an interarrival time. In general,

$$c_a^2 = \lim_{t\to\infty} \frac{\mathrm{Var}(A(t))}{E[A(t)]}. \tag{2}$$

Explicit formulas for c_a^2 are available for a wide array of arrival process models; e.g., see Sections 7 and 9 of [29]. When there is dependence among successive interarrival times, we need weak dependence as in Section 4.4 of [29], which is formally characterized by various mixing conditions; see [12,29] and references therein, but that is present in most applications, e.g., with overflow processes. Violations of this condition are usually associated with infinite values of c_a^2, which is usually easy to detect via exceptionally high estimates in data analysis.

The service times are assumed to be independent of the arrival process, but they are allowed to be mutually dependent. However, just like the interarrival times, the service times must be only weakly dependent. To approximately characterize that dependence, following [12], let $H_k(t_1, t_2) \equiv P(S_j \le t_1, S_{j+k} \le t_2)$ be the joint (bivariate) cdf of two service times separated by k indices, which is independent of j because of the assumed stationarity. Let $J_k \equiv E[S_j \wedge S_{j+k}]/E[S_j]$, with $\wedge$ the minimum, and $I_1 \equiv E[S_1 \wedge_{\text{indep}} S_2]/E[S]$ with $S_1 \wedge_{\text{indep}} S_2$ being the minimum of two independent random variables distributed as S.

Theorem 1 (*HT Peakedness from [12]*)**.** *Under regularity conditions, as the offered load α increases, the scaled steady-state number N_α of busy servers in the $G/G/\infty$ model becomes approximately normally distributed, i.e., $(N_\alpha - \alpha)/\sqrt{\alpha z} \approx \mathcal{N}(0, 1)$, where the constant z, called the HT peakedness, has the explicit form*

$$z \equiv z_{G/G}(c_a^2, G, H) = 1 + \mu(c_a^2 - 1)\int_0^\infty [1 - G(t)]^2 dt + 2\mu \int_0^\infty \left(\sum_{k=1}^\infty \left(H_k(t, t) - G(t)^2\right)\right) dt,$$
$$= 1 + (c_a^2 - 1)I_1 + 2\sum_{k=1}^\infty (J_k - I_1), \qquad (3)$$

with c_a^2 being the arrival process variability parameter in (1), G being the cdf of a generic service time with mean μ^{-1} and (H_k, I_1, J_k), $k \ge 1$, defined above.

The third term in formula (3) for the HT peakedness characterizes the impact of the dependence among the service times; it drops out if the service times are mutually independent, because then $J_k = I_1$ for all $k \ge 1$. The second term captures the consequence of a non-Poisson arrival process; it drops out if $c_a^2 = 1$, which occurs for Poisson arrival processes. Thus, we obtain $z = 1$ for the $MI/GI/\infty$ model. For further discussion, see [13].

Note that the arrival process is characterized beyond its rate (which appears via the offered load) by the single constant c_a^2, whereas the service times are characterized by the constant I_1 and the sequence $\{J_k : k \ge 1\}$. Proposition 3 of [13] gives a simple approximation for z that is exact in some instances, namely,

$$z \approx 1 + (c_a^2 - 1)I_1 + 2(1 - I_1)\Sigma_\rho, \qquad (4)$$

where Σ_ρ is the sum of all correlations, i.e., $\Sigma_\rho \equiv \sum_{k=1}^\infty \text{Corr}(S_j, S_{j+k})$. The dependence parameter Σ_ρ is intimately connected to the asymptotic variability parameter c_s^2 of the service times, defined as in (1) and (2) or, equivalently, via the CLT for associated partial sums; see Section 7.3 of [29]. In particular, Theorem 4.4.1 of [29] implies that $\Sigma_\rho = [(c_s^2/c_{s,rp}^2) - 1]/2$, where $c_{s,rp}^2$ is the SCV of a single service time and thus the asymptotic variability parameter in a renewal process with interrenewal times distributed as a single service time. Since the common form of dependence is for the service times to be positively correlated, typically $\Sigma_\rho \ge 0$ and $c_s^2/c_{s,rp}^2 \ge 1$. Thus well established ways to estimate $c_{s,rp}^2$ and c_s^2 from the data yield estimates of Σ_ρ. One way to estimate Σ_ρ is to estimate $\text{Corr}(S_j, S_{j+k})$ for a modest number of k and fit a single functional form, such as to $\text{Corr}(S_j, S_{j+k}) \approx \rho^k$ for some ρ with $0 < \rho < 1$. In that case, $\Sigma_\rho \approx \rho/(1 - \rho)$.

2.2. Exact peakedness in the $G/GI/\infty$ model

An expression for the exact peakedness z^e is not yet known for the $G/G/\infty$ IS model with dependent service times, but a nice account for the $G/GI/\infty$ model with independent service times was provided by Eckberg [17]. In many cases the exact peakedness of a general $G/GI/\infty$ model can be computed, but it often suffices to use asymptotic approximations, as shown in Tables 1, 2 and 4 of [13] and as we will show here. We introduce a refined second-order HT approximation that reveals how the HT peakedness differs from the exact peakedness, and can be used to improve the HT approximation for smaller offered loads (and thus in loss models with fewer servers).

A first rough approximation for the number of servers needed in the loss model is the offered load, because the offered load is the expected number of busy servers in the IS model. (Little's law implies that the expected number of busy servers in the loss model is $\lambda(1 - B)/\mu = (1 - B)\alpha$, where B is the steady-state blocking formula. In order for B to be suitably small, the actual number of servers must be roughly $\alpha + \beta\sqrt{\alpha}$ by the HT limit in Section 2.1, which tends to be not too much greater than α.)

In this section only, without loss of generality, we assume that the arrival rate is 1 and a generic service time is S/μ where S is a mean-1 random variable with cdf $G(x) \equiv P(S \le x)$, so that the offered load is $\alpha = \lambda/\mu = 1/\mu$. Let $U(x)$ be the expected number of arrivals in an interval of length x after an arbitrary arrival in the rate-1 arrival process, which we refer to as the *mean function*. For a renewal process, the mean function $U(x)$ is the familiar renewal function, but we allow dependence among successive interarrival times. Let $G^c(x) \equiv 1 - G(x)$ be the complementary cdf (ccdf).

Theorem 2 (*Exact Peakedness from [17]*). *For the $G/GI/\infty$ model, the peakedness is*

$$z^e(\mu) \equiv z^e_{G/GI}(\mu) \equiv \frac{\mathrm{Var}(N_\mu)}{E[N_\mu]} = 1 + 2\int_{0-}^{\infty}\left(\int_{\mu x}^{\infty} G^c(u)G^c(u-\mu x)\,du\right)dU(x) - \mu^{-1}. \tag{5}$$

For the case of i.i.d. exponential service times, the peakedness takes a simple form because the inner integral over u in (5) reduces to $e^{-\mu x}/2$. Let $\hat{U}_s(s)$ be the Laplace–Stieltjes transform of the mean function U of the rate-1 arrival process, i.e.,

$$\hat{U}_s(s) \equiv \int_{0-}^{\infty} e^{-st}\,dU(t). \tag{6}$$

Corollary 1 (*Exact Peakedness with MI Service from [17]*). *For the $G/MI/\infty$ model,*

$$z^e_{G/MI}(\mu) \equiv 1 + \hat{U}_s(\mu) - \mu^{-1}. \tag{7}$$

By the same reasoning, we can obtain the corresponding result for mixtures of exponential random variables, i.e., i.i.d. hyperexponential (H_kI) service times, as we show for H_2 in Appendix B. For general stationary arrival processes, Theorem 2 and Corollary 1 are not easy to apply because the mean function U and its Laplace–Stieltjes transform $\hat{U}_s(s)$ in (6) are not easy to determine.

However, in the special case of renewal arrival processes Laplace–Stieltjes transform $\hat{U}_s(s)$ in (6) can be expressed directly in terms of the Laplace transform of the interarrival-time density. Thus, for MI and H_2I service, the peakedness can easily be computed for an arbitrary renewal arrival process, provided that we can compute the Laplace transform of an interarrival time pdf, as we illustrate now for the MI case. Let f be the pdf of an interarrival time and let $\hat{f}(s)$ be its Laplace transform, i.e.,

$$\hat{f}(s) \equiv \int_0^{\infty} e^{-st} f(t)\,dt. \tag{8}$$

Corollary 2 (*Exact Peakedness in the $GI/MI/\infty$ Model*). *For the $GI/MI/\infty$ model having interarrival time pdf f with mean 1 and i.i.d. exponential service times with mean $1/\mu$,*

$$z^e_{GI/MI}(\mu) = \frac{\hat{f}(\mu) - 1 + \mu}{\mu(1-\hat{f}(\mu))}. \tag{9}$$

Proof. Since $\hat{U}_s(s) = \hat{f}(s)/(1-\hat{f}(s))$ for a renewal process with interarrival time pdf f, we can apply Corollary 1. □

We now develop a refined second-order HT approximation for the peakedness in the $GI/MI/\infty$ model. Let m_k be the kth moment of the interarrival-time pdf f, assuming that $m_1 = 1$ as before. Recall that a function $h(\mu)$ is $o(\mu)$ if $h(\mu)/\mu \to 0$ as $\mu \downarrow 0$.

Theorem 3 (*Refined Second-Order HT Peakedness in the $GI/MI/\infty$ Model*). *For the $GI/MI/\infty$ model, if the interarrival time pdf f has finite third moment, then the exact peakedness has the asymptotic form*

$$z^e_{GI/MI}(\mu) = \gamma_2 + (\gamma_2^2 - \gamma_3)\mu + o(\mu) \quad \text{as } \mu \downarrow 0, \tag{10}$$

where $\gamma_2 \equiv z = m_2/2 = (c_a^2+1)/2$ and $\gamma_3 \equiv m_3/6$ with $m_1 = 1$.

Proof. Assuming that the first k moments m_k are finite, the Laplace transform admits the Taylor series expansion

$$\hat{f}(s) = \sum_{j=0}^{k} (-1)^j \frac{m_j s^j}{j!} + o(s^k); \tag{11}$$

e.g., see Chapter 6 of [30]. Thus, we first obtain asymptotic expansions separately for the numerator and denominator in (9):

$$\begin{aligned} \hat{f}(\mu) - 1 + \mu &= \gamma_2\mu^2 - \gamma_3\mu^3 + o(\mu^3) \quad \text{as } \mu \downarrow 0, \\ \mu(1-\hat{f}(\mu)) &= \mu^2 - \gamma_2\mu^3 + \gamma_3\mu^4 + o(\mu^4) \quad \text{as } \mu \downarrow 0, \end{aligned} \tag{12}$$

from which we immediately obtain (10) from the asymptotic expansion of the ratio of two series expansions, exploiting the algebra of power series. □

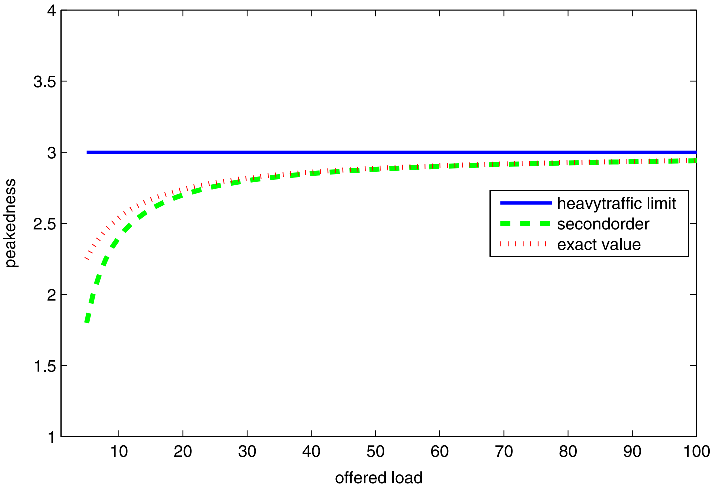

Fig. 1. A comparison of the heavy-traffic and the second-order heavy-traffic peakedness approximations to the exact peakedness for the $H_2I/MI/\infty$ model where the hyperexponential interarrival times have balanced means and SCV $c_a^2 = 5$.

As a quick check on Theorem 3, note that $c_a^2 = \gamma_2 = \gamma_3 = 1$ for an exponential interarrival time, so that $z^e_{MI/MI}(\mu) = 1$ for all μ. Indeed, the steady-state distribution of N_μ in the $MI/GI/\infty$ model is insensitivity to the service-time distribution beyond its mean, so that $z^e_{MI/GI}(\mu) = 1$ for all service-time distributions and all μ. More generally, Theorem 3 is useful to show when the first-order HT approximation of the peakedness (γ_2 in (10)) is likely to perform well and when it is likely to break down. It is more likely to break down when $\gamma_2^2 - \gamma_3$ is far from 0.

Example 1 (*The $H_2I/MI/\infty$ Model*)**.** To illustrate, we consider a renewal arrival process with a hyperexponential (H_2) interarrival-time pdf f. We let the H_2 distribution have balanced means as in (3.7) on p. 137 of [31]; we let the SCV be $c_a^2 = 5$ and the mean interarrival time be $m_1 = 1$. This H_2 density has the explicit form (B.1) for $p_1 = 0.9082$, $\lambda_1 = 1.8165$, $p_2 = 0.0918$, $\lambda_2 = 0.1835$. The associated parameters in Theorem 3 are $\gamma_2 \equiv z = (c_a^2 + 1)/2 = 3$ and $\gamma_3 \equiv m_3/6 = 15$. Hence, formula (9) yields $z^e_{H_2I/MI}(\mu) = 3 - 6\mu + o(\mu)$ as $\mu \downarrow 0$. Fig. 1 compares the HT peakedness $z_{H_2I/MI} = 3$ and the associated second-order HT approximation $z_{H_2I/MI,sec}(\mu) \equiv \gamma_2 + (\gamma_2^2 - \gamma_3)\mu = 3 - 6\mu$ to the exact peakedness $z^e_{H_2I/MI}(\mu)$ in Corollary 1. Fig. 1 shows that the HT peakedness and the second-order approximation are accurate to within about 10% and 1%, respectively, when the offered load $\alpha = 1/\mu$ (and thus the staffing) exceeds 30. The second-order approximation performs well for $10 \le \alpha \le 30$, while the exact peakedness is needed for $\alpha \le 10$. □

We can also easily analyze the case with deterministic service times, as we show in Appendix B. This section should be viewed as illustrative of what is possible; e.g., explicit results as in Theorem 7 can also be obtained for many non-renewal processes, e.g., see [32]. This section also provides strong motivation for the more elementary HT approximation provided by Theorem 1.

3. Models of dependence

For applications with system measurements, it is significant that the HT peakedness approximation in Section 2.1 can be applied without directly constructing a specific loss model or providing a detailed model of the dependence, because we can estimate the HT peakedness from (system or simulation) data, as was done in Section 5.4 of [13]. However, in order to evaluate the performance of the approximations using simulation, we need concrete models of the dependent service times. Fortunately, there are many several models in the literature. We will use two of these models from [13], which we quickly review here.

3.1. *Two models of dependence*

The first dependence model is the *Randomly Repeated Service* (RRS) process. In this model, we start with a sequence $\{B_n, n \ge 1\}$ of i.i.d. service times and a probability p of repeating the previous service time. Letting $\{I_n, n \ge 2\}$ be a sequence of i.i.d. Bernoulli random variables with mean p and $A_1 = B_1$, the RRS process $\{A_n, n \ge 1\}$ is defined by

$$A_n = I_n A_{n-1} + (1 - I_n) B_n, \quad n \ge 2. \tag{13}$$

This RRS process has a simple asymptotic variability parameter, $c_s^2 = c_B^2 \left(1 + \frac{2p}{1-p}\right)$, where c_B^2 is the SCV of B_1. Proposition 4 of [13] implies that the simple HT peakedness approximation in (4) is exact for the RRS process. This occurs primarily because the third term in (3) has a simple analytic expression for RRS, i.e.

$$2\mu \int_0^\infty \left(\sum_{k=1}^\infty \left(H_k(t,t) - G(t)^2 \right) \right) dt = \frac{2p}{1-p} \left(1 - \mu \int_0^\infty [1-G(t)]^2 dt \right). \tag{14}$$

The second dependence model is the *Exponential Autoregressive–Moving Average* (EARMA) process introduced by [33]. The model is specified by three parameters: μ, β, and ρ. To construct the process, we begin with three independent sequences of i.i.d. random variables $\{X_n : n \geq 0\}$, $\{U_n : n \geq 1\}$, and $\{V_n : n \geq 1\}$, with X_n exponentially distributed with rate μ, and U_n and V_n Bernoulli random variables with probabilities β and ρ of being equal to 0. The EARMA process $\{S_n : n \geq 1\}$ is then defined as

$$S_n = \beta X_n + U_n Y_{n-1} \quad \text{and} \quad Y_n = \rho Y_{n-1} + V_n X_n, \quad n \geq 2. \tag{15}$$

The EARMA process has an exponential marginal distribution with rate μ. Its correlation structure is identical to the Autoregressive–Moving Average ARMA(1,1) model:

$$\rho_k \equiv \mathrm{Corr}(S_j, S_{j+k}) = \gamma \rho^{k-1} \quad \text{and} \quad \gamma = \beta(1-\beta)(1-\rho) + (1-\beta)^2 \rho. \tag{16}$$

Since we do not have an expression for the HT peakedness (3) for EARMA, we use simulation to obtain a good estimate, as in Section 5.4 of [13].

3.2. The peakedness approximation with dependent service times

The two models of dependent service times can be used to evaluate the approximation of the exact peakedness by the HT peakedness. Such a study was conducted in Section 5 of [13], to which we refer. For a large class of models, an error of less than 1% (10%) could be achieved with an offered load of $\alpha = 100$ ($\alpha = 10$). However, for highly variable models (which we would regard as pathological), the required offered loads were much higher, e.g., $\alpha = 1000$ ($\alpha = 100$);

4. Approximations for the blocking probability

In this section we review two approximations for the blocking probability based on peakedness, which have been found to be quite accurate for the $G/GI/s/0$ model, without dependent service times. We will be using these same approximations with dependent service times, using the HT peakedness in Theorem 1.

4.1. The IS HT approximation: a normal approximation plus the conditioning heuristic

Let Y_α be the steady-state number of busy servers $G/G/s/0$ model at an arbitrary time. Following [19], we can use Theorem 1 plus a simple conditioning heuristic to generate the approximation

$$P(Y_\alpha = r) \approx P(N_\alpha = r)/P(N_\alpha \leq s) \quad \text{for } 0 \leq r \leq s. \tag{17}$$

This relation is exact for the $MI/GI/s/0$ system. We then apply Little's law with (17), Theorem 1 and the property $E[\mathcal{N}(0,1) | \mathcal{N}(0,1) \leq c] = \phi(c)/\Phi(c)$, where $\phi(x)$ and $\Phi(x)$ are the pdf and cdf of $\mathcal{N}(0,1)$, respectively, to obtain the approximation for the blocking probability, which we denote B_C (with subscript C for *call congestion*):

$$B_C = 1 - \frac{E[Y_\alpha]}{\alpha} \approx \sqrt{\frac{z}{\alpha}} \left(\frac{\phi\left((s-\alpha)/\sqrt{\alpha z}\right)}{\Phi\left((s-\alpha)/\sqrt{\alpha z}\right)} \right). \tag{18}$$

From Theorem 1 and (17), we can also approximate the distribution of Y_α. For example, we have the following approximation for the time congestion, denoted B_T:

$$B_T \equiv P(Y_\alpha = s) \approx \frac{P(s - 0.5 \leq \mathcal{N}(\alpha, \alpha z) \leq s + 0.5)}{P(\mathcal{N}(\alpha, \alpha z) \leq s + 0.5)} \approx \frac{\phi\left((s-\alpha)/\sqrt{\alpha z}\right)}{\sqrt{\alpha z}\,\Phi\left((s-\alpha+0.5)/\sqrt{\alpha z}\right)}. \tag{19}$$

The IS HT approximations (18) and (19) suggest that $B_T \approx B_C/z$. This is exactly true if the arrival process is Poisson and service times are independent, for which $z = 1$, due to the familiar PASTA property from [23]. However, we will see that this relation is not accurate in general, so that we will need to develop an improved approximation for the time congestion.

An important theoretical reference point is an early heavy-traffic limit.

Theorem 4 (*HT Limit from Borovkov [27]*). *For the $GI/MI/s/0$ model, as $\alpha \to \infty$ with $(s-\alpha)/\sqrt{\alpha} \to \beta$ for any constant β, $-\infty < \beta < +\infty$,*

$$\sqrt{\alpha} B_C \to \sqrt{z}\phi(\beta/\sqrt{z})/\Phi(\beta/\sqrt{z}), \tag{20}$$

where $z \equiv (c_a^2 + 1)/2$, the HT peakedness.

The scaling in Theorem 4 produces the familiar *Quality-and-Efficiency-Driven* (QED) many-server heavy-traffic regime. Thus, we expect the IS HT approximation to perform better in the QED regime than in the ED regime, where $(s-\alpha)/\sqrt{\alpha} \to -\infty$ or, for practical purposes $\beta < -2\sqrt{z}$, or the QD regime, where $(s-\alpha)/\sqrt{\alpha} \to +\infty$ or, for practical purposes $\beta > 2\sqrt{z}$. Borovkov [27] also gives a theorem without proof, cited by [19], stating that (19) is also asymptotically correct in the QED regime, but we will present a new theorem in Section 6 showing that must be incorrect.

4.2. The Hayward approximation

Our second approximation for the blocking probability is attributed to Walter Hayward of Bell Laboratories, and given a heuristic explanation by Fredericks [15]. The approximation makes use of the well-known Erlang loss formula, $B(s,\alpha) = (\alpha^s/s!)/(\sum_{i=0}^{s}(\alpha^i/i!))$. Hayward's approximation for $G/G/s/0$ systems uses the peakedness z to scale both s and α, yielding

$$B_C \equiv B_C(s,\alpha,z) \approx B\left(s/z, \alpha/z\right). \tag{21}$$

Since the heuristic development in [15] provides helpful intuition, we briefly review it here. Consider a loss system with constant service times each of size $1/\mu$, a batch Poisson arrival process with total arrival rate $\alpha\mu$, so that the offered load is α, where the batches sizes are all z, assumed to be an integer, and s servers with s a multiple of z. Thus the s servers can be split into z groups with each group handling one arrival from each batch. Since the service times are deterministic, the groups are always identical. Hence, each group behaves as an $MI/D/(s/z)/0$ model with arrival rate $\alpha\mu/z$ and offered load α/z, so that the blocking probability of each group, which equals the total blocking probability, is given by the Erlang loss formula with parameters s/z and α/z, exactly as in (21). (Recall the insensitivity of the blocking to the service time distribution beyond its mean in the $MI/GI/s/0$ model.)

Now let us consider the associated IS model. The infinitely many servers can also be divided into z groups. The steady-state number in each group is distributed exactly as Poisson with mean α/z and thus distributed approximately as $\mathcal{N}(\alpha/z, \alpha/z)$. Hence, the steady state number of customers in the entire IS system is distributed as $z\mathcal{N}(\alpha/z,\alpha/z)$, which is distributed the same as $\mathcal{N}(\alpha, z\alpha)$, implying that the exact peakedness is z.

A rough generalization of this idea is that the servers in a loss system facing bursty traffic ($z > 1$) can be divided into groups, and the arrivals can be allocated in such a way that each group has approximately the same number of busy servers, the same blocking probability, and peakedness one. Eq. (21) is then approximately correct for each group, and the total blocking probability is approximately that of the groups. Loss systems facing smooth traffic ($z < 1$) are treated similarly, but the original system is viewed as the result of splitting a larger system as before.

Unlike the IS HT approximation, which gives an approximate blocking probability of the exact system, the Hayward approximation uses the exact blocking probability of an approximate system. However, the two very different approximations are tightly linked, and thus each provides support for the other. As a consequence of the asymptotic behavior of the Erlang loss function from [34] or the HT limit by [27], we have the following result.

Theorem 5 (*Asymptotic Equivalence of the Two Approximations*). *Suppose that the assumptions on the arrival and service processes are as in Theorem 1. If $\alpha \to \infty$ with $(s-\alpha)/\sqrt{\alpha} \to \beta$, $-\infty < \beta < \infty$, then the Hayward and IS HT approximations for the scaled blocking probability $\sqrt{\alpha}B_C$ in the $G/G/s/0$ with either the HT peakedness or the exact peakedness both converge to the same nondegenerate limit $\sqrt{z}\phi(\beta/\sqrt{z})/\Phi(\beta/\sqrt{z})$, and so their difference is asymptotically negligible.*

Proof. First, by Theorem 1 the exact peakedness converges to the HT peakedness under the assumptions. Then observe that both approximations for the blocking probability remain the same if we divide all components of the vector (s,α,z) by z, yielding $(s/z, \alpha/z, 1)$. For the Hayward approximation, we apply asymptotics for the $MI/GI/\infty$ model or the Erlang loss formula. □

In practice, the parameter s/z will often not be an integer, so the Erlang loss formula cannot be used directly, but the continuous extension in [34], based on the integral representation of $B(\alpha,s)^{-1}$, has been found to be very effective. We use

$$B\left(s,\alpha\right) = \left[\frac{\Gamma\left(s+1\right)}{e^{-\alpha}\alpha^s} - \sum_{k=1}^{\infty}\frac{\alpha^k}{\left(s+1\right)\cdots\left(s+k\right)}\right]^{-1}, \tag{22}$$

with Γ denoting the gamma function, from Theorem 5 of [34]. For very large arguments, we can use asymptotic expansions, which essentially means the IS HT approximation.

5. Evaluating the extension to dependent service times

In this section we apply both the IS HT approximation for the blocking probability B_C in (18) and the Hayward approximation in (21) with dependent service times, using the HT peakedness in Theorem 1. We also evaluate the HT approximation for the time congestion B_T in (19). We find that the blocking approximations (call congestion) consistently performs well, but the associated IS time congestion approximation in (19) does not, even with i.i.d. service times. In Section 6 we explain the difficulty with the time congestion and develop a new approximation for it.

Table 1
Comparison of the blocking probability and time congestion approximations to simulated values for a Poisson arrival process. Shown are the minimum number of servers required to achieve the given target.

System	α	Target	Sim block	IS block	Hayward	Sim time	IS time	Ratio time
MI/GI	10	0.001	21	20	21	21	20	20
$z = 1$		0.01	18	18	18	18	18	18
		0.1	13	13	13	13	13	13
	50	0.001	71	71	71	71	71	71
		0.01	64	64	64	64	64	64
		0.1	51	52	51	51	51	52
	100	0.001	128	128	128	128	128	128
		0.01	117	117	117	117	117	117
		0.1	97	97	97	97	97	97
$MI/RRS(M)$	10	0.001	26	25	27	26	*24*	25
$z = 2$		0.01	21	21	22	21	20	21
$p = 0.5$		0.1	14	*16*	15	14	13	16
	50	0.001	81	80	82	81	*78*	80
		0.01	70	71	71	70	*67*	71
		0.1	54	*56*	55	54	**47**	*56*
	100	0.001	141	141	142	141	**137**	141
		0.01	126	127	127	126	**122**	127
		0.1	100	*103*	101	100	**87**	*103*
$MI/EARMA$	10	0.001	24	23	25	24	*22*	23
$z \approx 1.526$		0.01	20	20	20	20	19	20
$(\beta, \rho) = (0.5, 0.75)$		0.1	14	15	14	14	13	15
	50	0.001	76	76	77	76	75	76
		0.01	67	68	68	67	66	68
		0.1	52	*54*	53	52	*49*	*54*
	100	0.001	135	135	136	135	*133*	135
		0.01	121	*123*	*123*	121	120	*123*
		0.1	98	*100*	99	98	**92**	*100*

5.1. Results of simulation experiments

We test the accuracy of these approximations using simulation to estimate the true values. The obvious method is to fully specify a system, including the number of servers, and compare the simulated blocking probability to the approximation. However, in practice we are primarily interested in staffing under a constraint on the blocking probability, so we adopt that view; i.e., we select a target blocking probability and find the minimum number of servers such that the blocking probability (simulated or approximate) is below the target.

Besides being practical, this method of evaluating the approximations has an additional advantage over simply comparing blocking probabilities. Since the number of servers can only take discrete integer values, the blocking probabilities can only take finitely many values, which tend to differ greatly with few servers. As a consequence, a direct comparison of blocking probabilities will often be overly pessimistic, whereas the staffing approach does not have that problem.

The previous evaluations of the blocking probability approximations in [19] were performed with analytic results rather than simulation and are limited to $GI/MI/s/0$ models. For those cases, those experiments showed that both the IS HT and Hayward approximations are accurate for the blocking probability. Here we expand upon those experiments using simulation. Besides the usual exponential distribution for interarrival and service times, we will use the Erlang (E_k) and hyperexponential (H_k) distributions. The E_k distribution is less variable than the exponential, with $SCV = 1/k$; we use E_4. The H_k distribution is more variable than exponential and admits any $SCV > 1$. As before, we use H_2 with balanced means as on p. 137 of [31] and $SCV = 4$.

Tables 1–4 show the results of our simulation experiments for a number of systems using the staffing approach. Errors are emphasized. A difference of 1 server is not considered as an error. An error of 2 or 3 servers is indicated by showing the value in italics, while an error of more than 3 servers is indicated by boldface values. Most of the serious errors occur in the time congestion, which we discuss in the next section. The quality of the approximations depends on the case. The quality tends to deteriorate as the peakedness increases, with the peakedness increasing in the arrival process variability and the dependence (with positive correlations in our dependence models), but decreasing in the service time SCV.

First, Table 1 shows that there is no error (>1 server) in the staffing by the IS HT approximation for the MI/GI model, In this case, the HT peakedness coincides with the exact peakedness $z = 1$ and the Hayward approximation is exact. For the $MI/RRS(M)$ model, Table 1 shows that the dependence significantly increases the required staffing. Using the simple MI/GI model would result in unacceptable errors in all cases. Table 1 shows that the Hayward approximation has no error, while the IS HT approximation overemphasizes the required staffing with the high blocking target of 0.1. At least the IS HT approximation is conservative in its staffing recommendation.

Table 2
Comparison of various approximations for the blocking probability and time congestion to simulated values. Values indicate the minimum number of servers required to achieve a blocking probability/time congestion below the given target. All H_2 distributions have balanced means and $SCV = 4$.

System	α	Target	Sim block	IS block	Hayward	Sim time	IS time	Ratio time
E_4I/E_4I	10	0.001	17	17	17	18	17	17
$z = 0.46$		0.01	15	15	15	16	16	15
		0.1	12	12	12	13	13	13
	50	0.001	64	63	64	65	65	64
		0.01	59	58	58	60	60	60
		0.1	49	49	48	51	*53*	51
	100	0.001	119	118	118	120	120	119
		0.01	111	110	110	113	113	112
		0.1	94	94	94	99	*101*	98
E_4I/MI	10	0.001	18	18	19	19	18	19
$z = 0.63$		0.01	16	16	16	17	16	17
		0.1	12	12	12	13	13	13
	50	0.001	66	66	66	67	67	67
		0.01	60	60	60	61	62	62
		0.1	49	50	49	52	52	52
	100	0.001	121	121	121	123	123	123
		0.01	113	113	113	115	115	115
		0.1	95	95	95	99	100	99
E_4I/H_2I	10	0.001	19	19	20	20	19	19
$z = 0.74$		0.01	16	16	17	17	17	17
		0.1	12	13	12	13	13	14
	50	0.001	67	67	68	69	68	68
		0.01	61	61	61	62	62	63
		0.1	49	50	50	52	52	53
	100	0.001	123	123	123	125	124	125
		0.01	114	114	114	116	116	116
		0.1	95	96	95	99	99	100
$E_4I/EARMA$	10	0.001	22	21	22	22	21	22
$z \approx 1.151$		0.01	18	18	19	19	18	19
$(\beta, \rho) = (0.5, 0.75)$		0.1	13	14	13	14	13	15
	50	0.001	72	72	73	73	72	74
		0.01	64	65	65	65	64	*67*
		0.1	50	*52*	*52*	53	*50*	*55*
	100	0.001	129	130	*131*	131	*129*	132
		0.01	117	*119*	*119*	119	118	*122*
		0.1	96	*98*	*98*	100	**95**	*103*

Second, Table 2 shows excellent performance with the smooth E_4I arrival process. Again the departure from the basic MI/GI model leads to significantly different staffing, but with the smooth E_4I arrival process, the required staffing is less than with a Poisson arrival process.

Next, Table 3 shows that the bursty H_2I arrival process leads to significantly different staffing than for MI arrivals, but now much greater staffing, consistent with the higher peakedness. Notice that the case H_2I/E_4I is especially difficult, combining a bursty arrival process with a low-variability service distribution, which makes even higher peakedness. However, on the positive side, note that the IS HT staffing is consistently high by from 2 to 4 servers across all offered loads and performance targets, showing that a simple correction can be consistently applied. In general, Table 3 shows that the bursty H_2I arrival process is the most difficult. However, in general, because of our staffing perspective, the approximations perform quite consistently across all three offered loads, Thus, in practice, one could use some rule-of-thumb adjustment to the approximations that depends only on the variability of the arrival process and service distribution.

The $H_2I/MI/s/0$ case in Table 3 has extra data in parentheses that represents values if actual peakedness is calculated from Corollary 2 and used instead of the heavy-traffic peakedness. Consistent with Fig. 1, we see no change at all for $\alpha \geq 50$, but we do for $\alpha = 10$. The tables show that the HT peakedness is reasonable for the approximations, but some improvement can be expected by using the exact peakedness for lower offered loads. For the performance of the even smaller offered loads of $\alpha = 1$ and 5, see Appendix E.1.

Finally, Table 4 shows corresponding results with the non-renewal $RRS(M)$ arrival process. We see that the approximations continue to behave as before in this more bursty setting, with one exception. A significant error is seen in the IS HT approximation for high peakedness with the high offered load $\alpha = 100$ and low quality-of-service (QoS) target 0.1. This might seem inconsistent with the HT limit in Theorem 4, but it actually is not, because the scaled blocking probability $\sqrt{\alpha}B_C$ tends to converge to a proper limit in the QED HT regime. As a consequence, for the high offered load $\alpha = 100$ and the low

Table 3
Comparison of various approximations for the blocking probability and time congestion to simulated values. Values indicate the minimum number of servers required to achieve a blocking probability/time congestion below the given target. All H_2 distributions have balanced means and $SCV = 4$.

System	α	Target	Sim block	IS block	Hayward	Sim time	IS time	Ratio time
H_2I/E_4I	10	0.001	27	*29*	**33**	26	27	*28*
$z = 3.18$		0.01	23	*25*	*26*	22	22	*24*
		0.1	16	*18*	17	14	*12*	*17*
	50	0.001	86	*89*	**92**	85	84	*87*
		0.01	75	*78*	*78*	73	*70*	*75*
		0.1	57	*60*	58	53	**41**	*55*
	100	0.001	150	*153*	**155**	147	*145*	*150*
		0.01	133	*136*	*136*	130	**124**	*133*
		0.1	104	**108**	*106*	97	**75**	*100*
H_2I/MI	10	0.001	26	27(26)	**30**(*28*)	25	25(24)	26(25)
$z = 2.5$		0.01	22	23(22)	*24*(23)	21	21(20)	22(21)
		0.1	15	*17*(16)	16(16)	14	13(13)	15(15)
	50	0.001	83	84(83)	*86*(*86*)	81	80(80)	*83*(82)
		0.01	73	74(74)	*75*(74)	71	*69*(*69*)	72(72)
		0.1	56	57(57)	56(56)	51	**44**(**45**)	*53*(*53*)
	100	0.001	145	146(146)	*148*(*148*)	143	*141*(*140*)	144(144)
		0.01	130	131(131)	131(131)	126	*123*(*123*)	*128*(*128*)
		0.1	103	*105*(*105*)	103(103)	96	**82**(**82**)	*98*(*98*)
H_2I/H_2I	10	0.001	25	25(24)	*27*(26)	24	24(23)	25(23)
$z = 2.05$		0.01	21	21(21)	22(21)	20	20(19)	21(20)
		0.1	15	16(15)	15(15)	14	13(13)	15(14)
	50	0.001	80	81(80)	*82*(*82*)	79	78(78)	79(79)
		0.01	71	71(71)	72(71)	69	68(*67*)	69(69)
		0.1	55	56(56)	55(55)	51	**46**(**47**)	52(52)
	100	0.001	141	141(141)	*143*(142)	139	*137*(*137*)	139(139)
		0.01	127	127(127)	128(127)	124	*122*(*122*)	125(124)
		0.1	102	103(103)	102(101)	95	**86**(**87**)	96(96)
$H_2I/RRS(H_2)$	10	0.001	31	30	*33*	30	*27*	29
$z = 3.35$		0.01	25	25	*27*	24	*22*	24
$p = 0.5$		0.1	16	*18*	*18*	15	*12*	*17*
	50	0.001	91	90	*93*	89	**84**	88
		0.01	78	79	79	75	**70**	76
		0.1	58	*60*	59	54	**40**	*56*
	100	0.001	155	154	*157*	152	**146**	152
		0.01	136	137	*138*	133	**125**	134
		0.1	105	*108*	106	99	**73**	*101*
$H_2I/EARMA$	10	0.001	29	29	*32*	28	27	28
$z \approx 3.026$		0.01	23	24	*26*	22	22	23
$(\beta, \rho) = (0.5, 0.75)$		0.1	16	*18*	17	14	13	*16*
	50	0.001	87	88	**91**	85	*83*	86
		0.01	75	*77*	*78*	73	*70*	*75*
		0.1	56	*59*	*58*	52	**42**	*55*
	100	0.001	149	*151*	**154**	147	*144*	*149*
		0.01	133	*135*	*135*	129	**124**	*132*
		0.1	104	**107**	105	97	**77**	*100*

QoS target 0.1, the system tends to be in the ED many-server heavy-traffic regime, where $B_C \to 1 - \rho^{-1}$ as $\alpha \to \infty$ with $\alpha/s = \rho > 1$ held fixed, independent of z; for partial support, see Section 6.3 of [19].

6. Time congestion

In this section, we study the time congestion. We begin in Section 6.1 by analyzing the performance of the infinite-server time congestion approximation (19). Then in Section 6.2 we present an improved approximation based on the ratio of the blocking probability to the time congestion, called the *congestion ratio*.

6.1. The IS HT time congestion approximation

We can see from Tables 1–4 that the IS HT time congestion approximation (19) performs well except for bursty models with $z > 1$. Moreover, for $z > 1$, the IS HT approximation for B_T tends to perform worse as the offered load increases, raising

Table 4
Comparison of various approximations for the blocking probability and time congestion to simulated values. Values indicate the minimum number of servers required to achieve a blocking probability/time congestion below the given target. All H_2 distributions have balanced means and $SCV = 4$.

System	α	Target	Sim block	IS block	Hayward	Sim time	IS time	Ratio time
$RRS(M)/E_4I$	10	0.001	28	27	29	27	*25*	26
$z = 2.453$		0.01	22	23	*24*	21	21	22
$p = 0.5$		0.1	15	*17*	16	14	13	*16*
	50	0.001	84	84	*86*	82	*80*	83
		0.01	72	*74*	*74*	70	69	*72*
		0.1	55	*57*	56	51	**45**	**55**
	100	0.001	145	146	*148*	142	*140*	*144*
		0.01	129	*131*	*131*	126	*123*	*129*
		0.1	101	**105**	*103*	96	**82**	**100**
$RRS(M)/MI$	10	0.001	27	*25*	27	25	24	25
$z = 2$		0.01	21	21	22	20	20	21
$p = 0.5$		0.1	15	16	15	13	13	*15*
	50	0.001	81	80	82	79	78	79
		0.01	70	71	71	68	67	*70*
		0.1	54	*56*	55	51	**47**	*53*
	100	0.001	141	141	142	139	*137*	139
		0.01	126	127	127	123	122	*125*
		0.1	101	*103*	101	96	**87**	*99*
$RRS(M)/H_2I$	10	0.001	26	*24*	25	24	23	23
$z = 1.7$		0.01	21	20	21	19	19	20
$p = 0.5$		0.1	14	15	15	13	13	14
	50	0.001	79	78	79	77	76	77
		0.01	69	69	69	67	66	68
		0.1	53	*55*	54	50	*48*	*52*
	100	0.001	138	137	138	136	135	136
		0.01	124	124	124	121	121	*123*
		0.1	100	101	100	95	**90**	*97*
$RRS(M)/EARMA$	10	0.001	29	30	*27*	27	*25*	26
$z \approx 2.526$		0.01	23	24	23	21	21	22
$p = 0.5$		0.1	15	16	*17*	14	13	*16*
$(\beta, \rho) = (0.5, 0.75)$	50	0.001	84	*87*	84	83	*81*	83
		0.01	73	*75*	74	71	*69*	*73*
		0.1	55	56	*58*	51	**44**	**55**
	100	0.001	146	*148*	146	143	*141*	*145*
		0.01	129	*132*	*131*	126	*123*	*129*
		0.1	101	*103*	**105**	96	**82**	**101**

doubts about the claimed HT limit. In fact, we will prove that the claimed HT limit is not correct. The IS HT approximation is especially bad for high performance targets, being remarkably bad for the low QoS target 0.1.

We offer two heuristic explanations: first, in these difficult cases with $z > 1$, we have $B_C > B_T$, Hence, staffing to meet B_T will require fewer agents, but that lower level of staffing will cause additional blocking, and that additional blocking may in fact smooth the arrival process, making the carried arrival process (consisting of the non-blocked arrivals) less bursty than the original arrival process. Second, the lower staffing may in fact push the system out of the QED regime into the ED regime, where a different asymptotic behavior occurs. However, in any case, we will show that a different HT limit holds.

That analysis suggests a relatively simple heuristic adjustment to the IS HT time congestion approximation. Assuming that some traffic smoothing is taking place, we can replace the given peakedness $z > 1$ with $z \approx 1$ whenever the traffic intensity is above some threshold. We propose to make this adjustment whenever the traffic intensity is greater than 1, i.e., when $\alpha > s$. To see that this is effective, note that the problem cases in Tables 1–4 occurring when the target for B_T is 0.1 and the offered loads are 50 and 100 would all be staffed at 51 and 97 (taken from the $MI/GI/s/0$ system in Table 1), respectively, which gives an error of at most three servers among all the systems. Nevertheless, we next develop a new approximation, which performs better.

6.2. The congestion ratio

The underlying idea of the new approximation for the time congestion is to find it indirectly by leveraging the accurate approximation for the blocking probability. Specifically, we approximate the blocking probability B_C and the ratio of the blocking probability to the time congestion $B_R \equiv B_C/B_T$, called the *congestion ratio*. If both values are accurate, then the ratio of the two approximations should give a good approximation for B_T. Here we use the IS HT approximation (18) for B_C.

The congestion ratio is not an especially intuitive performance measure, but it has been successfully analyzed for $GI/MI/s/0$. We give a proof that suggests an approximation for the more general $G/G/s/0$ model. Let $\hat{U}_s(s)$ be the Laplace transform of the mean function U as in (6) and let f be the interarrival time pdf with $\hat{f}(s)$ being its Laplace transform as in (8).

Theorem 6 (*Congestion Ratio for GI/MI/s/0 from [8,7]*). *For $GI/MI/s/0$ model with arrival rate 1 and individual service rate $\mu = 1/\alpha$,*

$$B_R \equiv \frac{B_C}{B_T} = \frac{s\hat{f}(s/\alpha)}{\alpha(1-\hat{f}(s/\alpha))} = (s/\alpha)\hat{U}_s(s/\alpha). \tag{23}$$

We give an alternate proof, which is useful for generating an approximation more generally.

Proof. For the $GI/MI/s/0$ model, the instances that all servers become busy constitute regeneration times. Thus, there is an alternating renewal process of full times distributed as X and non-full times distributed as Y. By the renewal reward theorem, we can write

$$B_T = \frac{E[X]}{E[X+Y]} \quad \text{and} \quad B_C = \frac{E[A(X)]}{E[A(X+Y)]}. \tag{24}$$

Thus,

$$B_R \equiv \frac{B_C}{B_T} = \left(\frac{E[A(X)]}{E[X]}\right)\left(\frac{[X+Y]}{E[A(X+Y)]}\right). \tag{25}$$

Since X is exponential with mean α/s, $E[X] = s/\alpha$. By the renewal reward theorem again, $E[A(X+Y)]/E[X+Y] = \lambda = 1$. Finally, the $GI/MI/s/0$ structure allows us to deduce that the number of arrivals during a full period is geometrically distributed with parameter $p = \int_0^\infty f(x)e^{-(s/\alpha)x}\,dx$, so that

$$E[A(X)] = \frac{p}{1-p} = \frac{\hat{f}(s/\alpha)}{1-\hat{f}(s/\alpha)} = \hat{U}_s(s/\alpha). \quad \square \tag{26}$$

We can immediately apply Theorem 6 to obtain a many-server HT limit for the congestion ratio B_R

Corollary 3 (*Many-Server HT Limit for the Congestion Ratio*). *If $\alpha \to \infty$ with $\alpha/s \to 1$ in the $GI/MI/s/0$ model, then*

$$B_R \equiv B_R(f, s, \alpha) \to B_R^* \equiv \hat{U}_s(1) = \frac{\hat{f}(1)}{1-\hat{f}(1)}, \tag{27}$$

for $\hat{U}_s$ in (6) where $U(x)$ is the renewal function.

As a consistency check, note that $B_R = 1$ for all s and α when f is exponential. If the approximation (19) were asymptotically correct as $\alpha \to \infty$ with $\alpha/s \to 1$, then we should have $B_R^* = z = (c_a^2+1)/2$, but that does not hold. We can combine Theorem 4 and Corollary 3 to obtain the following limit for the time congestion.

Corollary 4 (*Many-Server HT Limit for the Time Congestion*). *If $\alpha \to \infty$ with $(\alpha - s)/\sqrt{\alpha} \to \beta, -\infty < \beta < \infty$, in the $GI/MI/s/0$ model, then*

$$\sqrt{\alpha}B_T \to \frac{\sqrt{z}\phi(\beta/\sqrt{z})}{\hat{U}_s(1)\Phi(\beta/\sqrt{z})}. \tag{28}$$

Remark 1 (*The Role of the Arrival Process*). Corollaries 3 and 4 show that, unlike the HT peakedness z and the associated approximation for the blocking probability, the congestion ratio and the time congestion depend on the arrival process through more than the arrival rate and the asymptotic variability parameter c_a^2.

Example 2 (*Renewal Processes with H_2^b and E_4 Interrenewal Times*). For the H_2^b distribution with balanced means, mean 1, where necessarily $c_a^2 > 1$, from p. 137 of [31] we have

$$\hat{f}(1) = \frac{2c_a^2+2}{3c_a^2+5} > \frac{1}{2} \quad \text{and} \quad B_R^*(c^2; H_2^b) = \frac{2c_a^2+2}{c_a^2+3} > 1, \tag{29}$$

with $\hat{f}(1) = 1/2$ and $B_R^* = 1$ in the limiting exponential case when $c_a^2 = 1$. We have $B_R^* = z \equiv (c_a^2+1)/2$ if and only if $c_a^2 = 1$.

For the E_k distribution with mean 1, we have

$$\hat{f}_k(1) = (k/(k+1))^k \geq \frac{1}{2} \quad \text{and} \quad B_R^*(E_k) < 1, \tag{30}$$

with $\hat{f}(1) = 1/2$ and $B_R^* = 1$ in the limiting exponential case when $k = 1$. Since $\hat{f}_k(1)/\hat{f}_{k+1}(1) > 1$, we see that $\hat{f}_k(1)$ is decreasing in k approaching $\hat{f}_\infty(1) = \hat{f}_D(1) = e^{-1}$. □

We base our proposed approximation for the congestion ratio in the more general $G/G/s/0$ loss model on the following conjecture.

Conjecture 1. *The first relation in Corollary 3 also holds for the more general $G/G/s/0$ loss model, with $U(x)$ more generally being the mean function when the arrival process is non-renewal.*

Supporting reasoning: first, and most important, we can approximate the full time X by an exponential random variable with mean α/s, exactly as in the $GI/MI/s/0$ case, by using the fact that a superposition of mutually independent stationary point processes is asymptotically Poisson; see Theorem 9.8.1 of [29]. For this step, we think of all s servers being busy, which should be approximately correct if s is not too small; then the departure process can be viewed as the superposition of the s service-completion processes for the separate servers. If the service times are i.i.d., then this is a superposition of independent renewal processes, but the limit holds in greater generality. We act as if this limit is valid.

That approximation gives us $E[X] \approx \alpha/s$ as $\alpha \to \infty$ with $\alpha/s \to 1$. Second, we can use the ergodic theory associated with stationary processes satisfying suitable mixing conditions instead of the renewal reward theorem to again obtain $E[A(X+Y)]/E[X+Y] = \lambda = 1$. Finally, paralleling (26), we use the approximation

$$E[A(X)] \approx \hat{U}_s(s/\alpha). \tag{31}$$

Here we are assuming that the mean function at the beginning of a full period is approximately the same as at an arbitrary arrival, which need not be the case. □

When the arrival process is not renewal, we can estimate $\hat{U}_s(x)$ for $0.5 < x < 3$ to use with the approximation. (See Appendix C.3 for our specific method.) The main approximation using the exact $\hat{U}_s(s/\alpha)$ is referred to as "Ratio Time" in Section 5, where it is shown to be quite effective. This approximation tends to perform worst for non-exponential service times under light loads, where the exponential approximation for the full time is not well justified.

7. Parcel blocking

The term *parcel blocking* comes from telecommunications, though the concept can apply to more diverse settings. The general heterogeneous blocking problem is a loss system facing k mutually independent individual arrival streams, each having the same service distributions, and the goal is to find the blocking probability for each stream, called the *parcel blocking probabilities*. This problem has been studied in the past, and there are exact results for certain special cases. For example, the parcel blocking probabilities are given exactly for the $GI + MI/MI/s/0$ system in [25], though the solution is inconvenient to compute. Approximations have also been found that are at the same time accurate and easy to calculate, and it is these that we focus on.

We number the arrival streams from 1 to k, and for each arrival stream i, we denote the offered load and peakedness in relation to the service distribution by α_i and z_i, respectively. We also let α and z be the offered load and peakedness of the entire superposition arrival process; α is just the sum of the α_i's, and the independence of the separate streams gives

$$z = \frac{\sum_{i=1}^{k} \alpha_i z_i}{\sum_{i=1}^{k} \alpha_i}. \tag{32}$$

The approximation we use here is simple and reasonably precise. Here it is presented in the form given in [21]:

$$B_i \approx B_T + \frac{z_i - 1}{z - 1}(B_C - B_T). \tag{33}$$

Note that this approximation requires the blocking probability and time congestion of the entire system, but these can be approximated with our previous methods. We discuss the derivation of (33) in Appendix D.

Table 5 shows the results of some experiments with a variety of systems with two independent arrival streams. Just as in Tables 1–4, we staff according to a target overall blocking probability. However, we then measure the other probabilities directly at this staffing level and present them as ratios of the blocking probability to show the differences and the accuracy of the approximations more directly. Further simulation data for these systems can be found in Appendix E.2. The total blocking probability is approximated with the infinite-server approximation (18), and the time congestion is approximated

Table 5
Comparison of approximate parcel blocking probabilities to simulated values. Staffing level is set according to target B_C, and B_T and B_1 are given as ratios to B_C. Both arrival streams have equal rates, and the H_2I distribution has balanced means with $SCV = 4$. B_C is approximated with the infinite-server approximation (18), B_1 is approximated with (33), and the B_T approximation uses the congestion ratio (23).

System	α	Target	Sim B_C	Approx	Sim B_T/B_C	Approx	Sim B_1/B_C	Approx
$H_2I + MI/MI$	10	0.001	24	24	0.82	0.89	1.21	1.11
$z = 1.75$		0.01	20	20	0.82	0.89	1.19	1.11
$z_1 = 2.5$		0.1	14	15	0.83	0.89	1.17	1.11
$z_2 = 1$	50	0.001	77	78	0.82	0.89	1.18	1.11
		0.01	68	69	0.83	0.89	1.18	1.11
		0.1	53	*55*	0.84	0.89	1.16	1.11
	100	0.001	137	138	0.82	0.89	1.18	1.11
		0.01	124	125	0.83	0.89	1.17	1.11
		0.1	100	101	0.84	0.89	1.16	1.11
$E_4I + MI/MI$	10	0.001	20	19	1.26	1.24	0.75	0.76
$z = 0.8125$		0.01	17	17	1.23	1.22	0.78	0.78
$z_1 = 0.625$		0.1	13	13	1.18	1.18	0.82	0.82
$z_2 = 1$	50	0.001	69	68	1.19	1.19	0.81	0.81
		0.01	62	62	1.18	1.18	0.83	0.82
		0.1	50	51	1.15	1.15	0.85	0.85
	100	0.001	125	125	1.17	1.18	0.83	0.82
		0.01	115	115	1.17	1.17	0.84	0.83
		0.1	96	96	1.14	1.15	0.86	0.85
$H_2I + MI/H_2I$	10	0.001	23	23	0.81	0.89	1.19	1.11
$z = 1.525$		0.01	19	20	0.82	0.89	1.18	1.11
$z_1 = 2.05$		0.1	14	15	0.83	0.89	1.17	1.11
$z_2 = 1$	50	0.001	76	76	0.82	0.89	1.18	1.11
		0.01	67	68	0.83	0.89	1.18	1.11
		0.1	53	54	0.84	0.89	1.16	1.11
	100	0.001	135	135	0.82	0.89	1.18	1.11
		0.01	122	123	0.83	0.89	1.17	1.11
		0.1	99	100	0.84	0.89	1.16	1.11
$H_2I + MI/RRS(M)$	10	0.001	28	28	0.82	0.89	1.17	1.05
$p = 0.5$		0.01	22	23	0.83	0.89	1.17	1.05
$z = 2.75$		0.1	15	*17*	0.85	0.89	1.15	1.05
$z_1 = 3.5$	50	0.001	85	86	0.84	0.89	1.16	1.05
$z_2 = 2$		0.01	74	75	0.85	0.89	1.16	1.05
		0.1	55	*58*	0.86	0.89	1.14	1.05
	100	0.001	148	149	0.85	0.89	1.15	1.05
		0.01	131	*133*	0.85	0.89	1.15	1.05
		0.1	102	**106**	0.87	0.89	1.13	1.05

with the congestion ratio (23). Table 5 shows that the parcel blocking probability can be significantly different from the total blocking probability and that the approximation (33) works well for different systems including those with dependent service times. It should be noted that, just as the performance of a loss system with Poisson arrivals is the same for any independent sequence of service times, the first three systems in Table 5 are not affected if the arrivals from the Poisson stream face a different independent service process.

8. Conclusions

We have applied simulation to show that the truncated-normal and Hayward approximations in (18) and (21) for the blocking probability in the general $G/G/s/0$ stationary loss model remain effective when there is dependence among successive service times as well as among successive interarrival times, and non-exponential distributions, when the dependence is captured by the heavy-traffic (HT) peakedness z in (3), provided that the offered load is not too small. (The approximations may also be useful for low offered loads as well, as illustrated by Table E.1.) We have shown in Theorem 5 that these two approximations are asymptotically equivalent in the QED many-server heavy-traffic regime. In Section 2.2 we have also reviewed the exact peakedness for the $G/GI/s/0$ model from [17] and shown how it is related to the HT peakedness, in part via the refined second-order heavy-traffic approximate peakedness with i.i.d. exponential service times in Theorem 3.

In Section 6 we have shown that the corresponding normal approximation for the time congestion from [27,19] is not accurate and developed a new HT approximation based on the congestion ratio, which we showed is effective. Significantly, the new approximation for the time congestion depends on the arrival process beyond its rate and asymptotic variability parameter c_a^2 in (1) and (2). We then applied this new approximation for the time congestion to develop a new approximation for the parcel blocking in multi-class loss models in Section 7, which we showed is also effective.

Among the many directions for future research, it remains to (i) exhibit the exact peakedness for the $G/G/s/0$ model considered here, (ii) extend Theorem 1 and the approximations here to cover the dependence between the service times and the arrival process, (iii) study blocking approximations in $G_t/G/s_t/0$ models with time-varying arrivals and staffing, drawing on Section 7 of [13], and (iv) it remains to measure the dependence in arrival and service processes in applications, hopefully exploiting Theorem 1.

Acknowledgments

This research was supported by NSF grant CMMI 1066372. This research was conducted while the first author was an undergraduate in the IEOR Department at Columbia University.

Appendix A. Overview

In these appendices we present additional material supplementing the main paper. First, in Appendix B we supplement Section 2.2 by giving additional results about the exact peakedness for hyperexponential (H_2) and deterministic (D) service times. Next, in Appendix C we describe our simulation procedure in more detail. In Appendix D we review one derivation for the parcel blocking approximation in (33). Finally, in Appendix E we present additional results of the simulation experiments.

Appendix B. More on the exact peakedness

We now supplement Section 2.2 with a few additional results about the exact peakedness for the $G/GI/\infty$ model. We first give the exact peakedness for H_2 service times; i.e., let the mean-1 random variable S have probability density function (pdf)

$$f(t) = p_1\lambda_1 e^{-\lambda_1 t} + p_2\lambda_2 e^{-\lambda_2 t}, \quad t \geq 0, \tag{B.1}$$

with $E[S] = (p_1/\lambda_1) + (p_2/\lambda_2) = 1$. We apply Theorem 2 to obtain the following corollary.

Corollary 5 (*Exact Peakedness with H_2I Service*)**.** *For the $G/H_2I/\infty$ model with service pdf in (B.1),*

$$z^e_{G/H_2I}(\mu) \equiv 1 + \left(\frac{p_1^2}{\lambda_1} + \frac{2p_1p_2}{\lambda_1+\lambda_2}\right)\hat{U}_s(\lambda_1\mu) + \left(\frac{p_2^2}{\lambda_2} + \frac{2p_1p_2}{\lambda_1+\lambda_2}\right)\hat{U}_s(\lambda_2\mu) - \mu^{-1} \tag{B.2}$$

where $\hat{U}_s(s)$ is the Laplace–Stieltjes transform of the rate-1 arrival process mean function in (6).

We can also easily analyze the case with deterministic service times. Recall that the arrival counting process A is assumed to be a stationary point process (with stationary increments), which is the equilibrium renewal process if the interarrival times are i.i.d. (which is a delayed renewal process associated with the given renewal arrival process in which the first renewal is distributed according to the interarrival-time stationary-excess distribution) if the interarrival times are i.i.d.

Theorem 7 (*Exact and Second-Order HT Peakedness with D Service*)**.** *For the $G/D/\infty$ model, $z^e_{G/D}(\mu) = \mu\mathrm{Var}(A(\mu^{-1}))$. For the special case of a renewal arrival process with interarrival time pdf f having finite third moment,*

$$z^e_{GI/D}(\mu) = \mu\,\mathrm{Var}(A(\mu^{-1})) = c_a^2 + 2(\gamma_2^2 - \gamma_3)\mu + o(\mu) \quad \text{as } \mu \downarrow 0, \tag{B.3}$$

where $\gamma_2 = m_2/2 = (c_a^2+1)/2$ and $\gamma_3 = m_3/6$ with $m_1 = 1$.

Proof. We exploit the fact that N_μ coincides with $A(t) - A(t-\mu^{-1})$ when the service times are all deterministic taking the value $1/\mu$. Hence, $z^e_{G/D}(\mu) = \mu\,\mathrm{Var}(A(\mu^{-1}))$, as stated above. Then, for GI arrivals, $\mu\,\mathrm{Var}(A(\mu^{-1}))$ satisfies (B.3) by (18) on p. 58 of [35]. □

As a quick check on Theorem 7, note that $c_a^2 = \gamma_2 = \gamma_3 = 1$ for an exponential interarrival time, so that $z^e_{MI/D}(\mu) = 1$ for all μ, which again is consistent with the fact that $z^e_{MI/GI}(\mu) = 1$ for all service-time distributions and all μ.

Example 3 (*The $E_2I/D/\infty$ Model*)**.** For an explicit example, consider a renewal arrival process with interarrival times that are Erlang E_2 with mean 1. For the $E_2I/D/\infty$ model the exact peakedness is

$$z^e_{E_2I/D}(\mu) = \mu\,\mathrm{Var}(A(\mu^{-1})) = \frac{1}{2} + \frac{\mu}{8} - \frac{\mu e^{-4\mu^{-1}}}{8} = \frac{1}{2} + \frac{\mu}{8} + o\left(e^{-4\mu^{-1}}\right) \quad \text{as } \mu \downarrow 0; \tag{B.4}$$

see p. 57 of [35]. In this case, $m_1 = 1$, $m_2 = 3/2$ and $m_3 = 3$, so that $\gamma_2 = 3/4$ and $\gamma_3 = 1/2$, from which we see that (B.4) is consistent with (B.3), with the error in the second-order approximation decreasing exponentially. □

Appendix C. Description of simulation procedures in Sections 5.1 and 7

Here we give a more detailed description of the procedures used in our loss model simulation experiments, the results of which are found in Tables 1–4 of Section 5.1 and Table 5 of Section 7.

C.1. Individual replications

In all cases (each defined by an arrival process, a service process, and the number of servers), a single replication starts with an empty system and runs for $150/\mu$ units of time, where $1/\mu$ is the mean service time. To ensure that the performance is measured in steady state, the data is only taken in the time interval $[50/\mu, 150/\mu]$; we have verified for each case that $50/\mu$ units of time is long enough to reach steady state by plotting the estimated mean and variance of the system size over time in for 1000 simulated runs and ensuring that these variables appear constant graphically. The estimated blocking probability in a replication is the number of blocked arrivals over total arrivals, and when there are multiple arrival streams as in the parcel blocking experiment in Section 7, separate blocking probabilities are estimated for each stream in the same way. The time congestion is estimated by the time the system spends full divided by the total observation ! time $(100/\mu)$.

C.2. Multiple replications and confidence intervals

To estimate a performance measure, we execute four independent sets of independent replications and take the sample mean in each set. We then take the mean of these four independent estimates for our final estimate, and a 95% confidence interval (CI) is calculated using the t-distribution with three degrees of freedom (if $\hat{\sigma}^2$ is the sample variance, then the confidence interval has half-width $3.182\hat{\sigma}/\sqrt{4}$). For the staffing levels given in Tables 1–5, enough replications are executed so that the CI lies entirely below the target at the given number of servers and entirely above the target when one more server is included. Up to 5000 replications for each of the four sets are performed to achieve the necessary precision, and in the event that this is not enough replications, the lowest staffing level that produces a CI containing the target is listed. When a performance measure is directly estimated as in Table 5!, 10,000 replications are done for each set so that there are at least two significant digits.

C.3. Estimation of the mean function

As shown in Eq. (31), an important component of our congestion ratio approximation is the use of $\hat{U}_s$, the Laplace–Stieltjes transform in (6) of the mean function $U(x)$ to estimate the expected arrivals during a full period. For a renewal arrival process, this is easily calculated with the Laplace transform of the interarrival time pdf, as shown in (26), but it presents difficulty for more general arrival processes. As mentioned in Section 6.2, our method throughout was to directly estimate the mean function $U(x)$ using simulation. To do this, we performed one million independent replications of the rate-1 arrival process. Each replication was 10 time units long, and we measured the number of arrivals every 0.01 time units so that each replication consisted of 1000 data points. We averaged all of the replications to estimate $U(x)$, and then we estimated the derivative $U'(x)$ linearly: $U'(x) \approx (U(x+0.01) - U(x-0.01))/0.02$. The final integral was approximated! with rectangles.

Appendix D. Derivation of the parcel blocking approximation in (33)

The approximation (33) has been derived a number of ways, including a birth–death argument with state dependent birth rates by [16]. [24] also provide a nice derivation that considers a special case where the relation is exact, which we now review. Let there be $k = 3$ arrival streams. Arrival streams 1 and 2 are identical general processes, while arrival stream 3 is a Poisson process. According to the PASTA property, $B_3 = B_T$, which is also given by (33). It remains to find the parcel blocking probability B_1 for stream 1, which is equal to B_2, and the blocking probability of the superposition of streams 2 and 3, which we denote B_{23}. In general, we must have the balance equation

$$2\alpha_1 B_1 + \alpha_3 B_3 = \alpha B_C, \tag{D.1}$$

which can be rearranged after B_T is substituted for B_3 to give

$$B_1 = B_T + \frac{\alpha}{2\alpha_1}(B_C - B_T). \tag{D.2}$$

This is in fact identical to (33), which can be seen by substituting (32) into (33). To find B_{23}, we use an additional balance equation:

$$\alpha_1 B_1 + (\alpha_2 + \alpha_3) B_{23} = \alpha B_C. \tag{D.3}$$

Table E.1
Comparison of various approximations for the blocking probability and time congestion to simulated values in the case of smaller offered loads, specifically for $\alpha = 1$ and 5. Values indicate the minimum number of servers required to achieve a blocking probability/time congestion below the given target. All H_2 distributions have balanced means and $SCV = 4$.

System	α	Target	Sim block	IS block	Hayward	Sim time	IS time	Ratio time
M/GI	1	0.001	6	5	6	6	5	5
$z = 1$		0.01	5	4	5	5	4	4
		0.1	3	3	3	3	3	3
	5	0.001	14	13	14	14	13	13
		0.01	11	11	11	11	11	11
		0.1	8	8	8	8	8	8
H_2/M	1	0.001	7	7(6)	*10*(7)	7	7(6)	7(6)
$z = 2.5$		0.01	6	6(5)	*8*(6)	5	6(5)	6(5)
		0.1	4	5(4)	5(4)	3	4(3)	4(3)
	5	0.001	17	17(16)	*20*(18)	16	16(15)	17(16)
		0.01	14	15(14)	*16*(14)	13	13(13)	14(13)
		0.1	9	*11*(10)	10(10)	9	8(8)	10(9)
H_2/H_2	1	0.001	7	7(*5*)	*9*(7)	7	6(*5*)	6(*5*)
$z = 2.05$		0.01	6	6(5)	7(5)	5	5(5)	5(4)
		0.1	4	4(4)	4(3)	3	4(3)	4(3)
	5	0.001	16	16(15)	*18*(17)	16	15(15)	16(15)
		0.01	13	14(13)	*15*(13)	13	13(12)	13(12)
		0.1	9	10(9)	10(9)	8	8(8)	9(9)

Combining (D.2) and (D.3), we get a formula similar to (D.2),

$$B_{23} = B_T + \frac{\alpha}{2(\alpha_2 + \alpha_3)}(B_C - B_T), \tag{D.4}$$

and this is also given by (33) when the system is viewed as having two arrival streams (stream 1 and the superposition of streams 2 and 3).

Appendix E. Additional simulation data

E.1. Performance of the approximations at small offered loads

Most of the approximations presented throughout are supported by heavy-traffic limits, and we have shown that they are accurate for high offered loads ($\alpha = 100, 500$). We also used the case of $\alpha = 10$ to demonstrate that the approximations work well even for lower offered loads. Here we consider even smaller offered loads to address the question of how far the approximations can be pushed.

Table E.1 is formatted identically to Tables 1–4 except the offered loads are $\alpha = 1, 5$. In addition, for all three of the systems presented, the exact peakedness is known in addition to the heavy-traffic peakedness, so we can isolate the effects of approximating the exact peakedness with the heavy-traffic peakedness. It seems the approximation methods retain their accuracy in low traffic, as the cases are off by at most three servers. It is not clear from the data which step is the bigger cause of the error, and while the approximations are not well supported by heavy-traffic limits at these smaller offered loads, the data suggest they may still be useful.

E.2. Parcel blocking experiments extending Section 7

Table E.2 is a complement to the data provided in Table 5 of Section 7 using the same pairs of arrival and service processes. Unlike the previous experiments, the performance measures are directly given here rather than taking the staffing approach. Offered loads range from 10 to 500, and the number of servers was chosen so that the performance measures would fall approximately in our range of interest.

At the higher offered load of 500, all of the approximations perform well including the parcel blocking approximation (33). The approximations are only slightly worse as the offered load drops to 100, but they do not appear to be accurate at an offered load of 10. In fact, even the blocking probability approximation (18) seems to perform poorly at this offered load, though the experiments in Section 5.1 suggested that this approximation produces accurate staffing levels at offered loads as low as 10 for a variety of systems. This highlights the point made in Section 5.1 that direct comparison of the performance measures can be overly strict as a method of evaluating approximations, particularly when staffing is the more common application.

Table E.2
Comparison of approximate parcel blocking probabilities to simulated values. Both arrival streams have equal rates, and the H_2I distribution has balanced means with $SCV = 4$. The total blocking probability is approximated with the infinite-server approximation (18), the parcel blocking probabilities are approximated with (33), and the time congestion approximation uses the congestion ratio (23).

System	α	s	Sim block	Approx	Sim time	Approx	Sim block$_1$	Approx
$H_2I + MI/MI$	500	450	0.12	0.12	0.10	0.11	0.14	0.14
$z = 1.75$		500	0.045	0.047	0.038	0.042	0.052	0.052
$z_1 = 2.5$		550	0.0053	0.0059	0.0044	0.0053	0.0061	0.0066
$z_2 = 1$	100	80	0.24	0.26	0.21	0.23	0.28	0.28
		100	0.095	0.11	0.080	0.094	0.11	0.12
		120	0.015	0.018	0.013	0.016	0.018	0.020
	10	10	0.25	0.33	0.21	0.30	0.28	0.37
		15	0.062	0.092	0.051	0.082	0.073	0.10
		20	0.0067	0.0097	0.0055	0.0086	0.0080	0.011
$E_4I + MI/MI$	500	450	0.11	0.11	0.13	0.13	0.10	0.097
$z = 0.8125$		500	0.031	0.032	0.036	0.037	0.027	0.027
$z_1 = 0.625$		550	0.00078	0.00075	0.00090	0.00087	0.00066	0.00063
$z_2 = 1$	100	80	0.22	0.23	0.25	0.26	0.20	0.20
		100	0.069	0.072	0.079	0.083	0.059	0.061
		120	0.0033	0.0031	0.0038	0.0036	0.0027	0.0026
	10	10	0.20	0.23	0.22	0.26	0.17	0.19
		15	0.025	0.025	0.030	0.031	0.020	0.020
		20	0.00068	0.00024	0.00086	0.00030	0.00053	0.00018
$H_2I + MI/H_2I$	500	450	0.12	0.12	0.10	0.11	0.14	0.13
$z = 1.525$		500	0.044	0.044	0.037	0.039	0.050	0.049
$z_1 = 2.05$		550	0.0045	0.0044	0.0037	0.0040	0.0052	0.0049
$z_2 = 1$	100	80	0.24	0.25	0.21	0.23	0.27	0.28
		100	0.092	0.099	0.077	0.088	0.11	0.11
		120	0.013	0.014	0.011	0.013	0.015	0.016
	10	10	0.24	0.31	0.21	0.28	0.27	0.34
		15	0.055	0.076	0.046	0.068	0.065	0.085
		20	0.0052	0.0059	0.0042	0.0052	0.0061	0.0066
$H_2I + MI/RRS(M)$	500	450	0.13	0.13	0.11	0.12	0.14	0.14
$p = 0.5$		500	0.054	0.059	0.047	0.053	0.062	0.062
$z = 2.75$		550	0.012	0.013	0.010	0.012	0.013	0.014
$z_1 = 3.5$	100	80	0.25	0.28	0.22	0.25	0.28	0.29
$z_2 = 2$		100	0.11	0.13	0.10	0.12	0.13	0.14
		120	0.029	0.036	0.024	0.032	0.033	0.038
	10	10	0.26	0.42	0.23	0.37	0.29	0.44
		15	0.086	0.16	0.073	0.14	0.10	0.17
		20	0.019	0.035	0.016	0.031	0.022	0.037

References

[1] N. Litvak, M. van Rijsbergen, R.J. Boucherie, M. van Houdenhoven, Managing the overflow of intensive care patients, European J. Oper. Res. 185 (2008) 998–1010.
[2] M. Restrepo, S.G. Henderson, H. Topaloglu, Erlang loss models for the static deployment of ambulances, Health Care Manag. Sci. 12 (2009) 67–79.
[3] A.M. de Bruin, R. Bekker, L. van Zanten, G.M. Koole, Dimensioning hospital wards using the Erlang loss model, Ann. Oper. Res. 178 (2010) 23–43.
[4] M. Asaduzzaman, T.J. Chaussalet, N.J. Tobertson, A loss network model with overflow for capacity planning of a neonatal unit, Ann. Oper. Res. 178 (2010) 67–76.
[5] R. Levi, A. Radovanovic, Provably near-optimal LP-based policies for revenue management in systems with reusable resources, Oper. Res. 58 (2010) 503–507.
[6] S.M. Ross, Stochastic Processes, second ed., Wiley, New York, 1996.
[7] J.W. Cohen, The full availability group of trunks with an arbitrary distribution of the interarrival times and a negative exponential holding time distribution, Simon Stevin 31 (1957) 169–181.
[8] L. Takacs, On the generalization of Erlang's formula, Acta Math. Hungar. 7 (1956) 419–433.
[9] R. Wilkinson, Theories of toll traffic engineering in the USA, Bell Syst. Tech. J. 35 (1956) 421–514.
[10] R.B. Cooper, Introduction to Queueing Theory, second ed., North Holland, Amsterdam, 1981.
[11] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao, Statistical analysis of a telephone call center: a queueing-science perspective, J. Amer. Statist. Assoc. 100 (2005) 36–50.
[12] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues with dependent service times, Queueing Syst. 73 (2013) 119–146.
[13] G. Pang, W. Whitt, The impact of dependent service times on large-scale service systems, Manuf. Serv. Oper. Manage. 14 (2012) 262–278.
[14] J.M. Holtzman, The accuracy of the equivalent random method with renewal inputs, Bell Syst. Tech. J. 52 (1973) 1673–1679.
[15] A.A. Fredericks, Congestion in blocking systems—a simple approximation technique, Bell Syst. Tech. J. 59 (1980) 805–827.
[16] A.A. Fredericks, Approximating parcel blocking via state dependent birth rates, in: Proc. 10th Int. Teletraffic Congress, Montreal, Canada, June 1983.
[17] A.E. Eckberg, Generalized peakedness of teletraffic processes, in: Proceedings 10th Int. Teletraffic Congress, Montreal, Canada, June 1983. Paper 4.4.b.3.
[18] A.E. Eckberg, Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness, in: Proc. 11th Int. Teletraffic Congress, Kyoto, Japan, 1985, pp. 331–335.
[19] W. Whitt, Heavy traffic approximations for service systems with blocking, AT & T Bell Lab. Tech. J. 63 (1984) 689–708.
[20] W. Whitt, A diffusion approximation for the $G/GI/n/m$ queue, Oper. Res. 52 (2004) 922–941.
[21] B. Sanders, E.A. van Doorn, Estimating time congestion from traffic parameters, IEEE Trans. Commun. 35 (1987) 856–862.
[22] D.L. Jagerman, B. Melamed, Burstiness descriptors of traffic streams: indices of dispersion and peakedness, in: Proc. 1994 Conf. Informations Sciences and Systems, Vol. 1, Princeton University, Princeton, NJ, 1994, pp. 24–28.

[23] R.W. Wolff, Poisson arrivals see time averages, Oper. Res. 30 (1982) 223–231.
[24] H. Akimaru, H. Takahashi, An approximate formula for individual call losses in overflow systems, IEEE Trans. Commun. 31 (1983) 808–811.
[25] A. Kuczura, Loss systems with mixed renewal and Poisson inputs, Oper. Res. 21 (1973) 787–795.
[26] K.S. Meier-Hellstern, The analysis of a queue arising in overflow models, IEEE Trans. Commun. 37 (1989) 367–372.
[27] A.A. Borovkov, Stochastic Processes in Queueing Theory, Springer, New York, 1976.
[28] A.A. Borovkov, On limit laws for service processes in multi-channel systems, Sib. Math. J. 8 (1967) 746–763.
[29] W. Whitt, Stochastic-Process Limits, Springer, New York, 2002.
[30] K.L. Chung, A Course in Probability Theory, third ed., Academic Press, New York, 2001.
[31] W. Whitt, Approximating a point process by a renewal process, Oper. Res. 30 (1982) 125–147.
[32] M.F. Neuts, A versatile Markovian point process, J. Appl. Probab. 16 (1979) 764–779.
[33] P.A. Jacobs, P.A.W. Lewis, A mixed autoregressive-moving average exponential sequence and point process (EARMA 1,1), Adv. Appl. Probab. 9 (1977) 87–104.
[34] D.L. Jagerman, Some properties of the Erlang loss function, Bell Syst. Tech. J. 53 (1974) 525–551.
[35] D.R. Cox, Renewal Theory, Methuen, London, 1962.

Andrew A. Li is a doctoral student in the Operations Research Center at M.I.T. This research was started while he was an undergraduate in the Department of Industrial Engineering and Operations Research at Columbia University.

Ward Whitt is the Wai T. Chang Professor of Industrial Engineering and Operations Research at Columbia University. He received his doctorate from the School of Operations Research and Information Engineering at Cornell University in 1969. After teaching at Stanford University and Yale University, he joined Bell Laboratories in 1977. He spent twenty-five years in research at Bell Labs and AT&T Labs before joining Columbia University in 2002. His research focuses on stochastic models and their applications to performance evaluation.