Probability in the Engineering and Informational Sciences, **28**, 2014, 419–449.
doi:10.1017/S0269964814000084

STABILIZING PERFORMANCE IN NETWORKS OF QUEUES WITH TIME-VARYING ARRIVAL RATES

YUNAN LIU AND WARD WHITT
Department of Industrial Engineering, North Carolina State University
Raleigh, NC 27695, USA
Department of Industrial Engineering and Operations Research, Columbia University
New York, NY 10027, USA
E-mails: yliu48@ncsu.edu; ww2040@columbia.edu

This paper investigates extensions to feed-forward queueing networks of an algorithm to set staffing levels (the number of servers) to stabilize performance in an $M_t/GI/s_t + GI$ multi-server queue with a time-varying arrival rate. The model has a non-homogeneous Poisson process (NHPP), customer abandonment, and non-exponential service and patience distributions. For a single queue, simulation experiments showed that the algorithm successfully stabilizes abandonment probabilities and expected delays over a wide range of Quality-of-Service (QoS) targets. A limit theorem showed that stable performance at fixed QoS targets is achieved asymptotically as the scale increases (by letting the arrival rate grow while holding the service and patience distributions fixed). Here we extend that limit theorem to a feed-forward queueing network. However, these fixed QoS targets provide low QoS as the scale increases. Hence, these limits primarily support the algorithm with a low QoS target. For a high QoS target, effectiveness depends on the NHPP property, but the departure process never is exactly an NHPP. Thus, we investigate when a departure process can be regarded as approximately an NHPP. We show that index of dispersion for counts is effective for determining when a departure process is approximately an NHPP in this setting. In the important common case when all queues have high QoS targets, we show that both: (i) the departure process is approximately an NHPP from this perspective and (ii) the algorithm is effective.

1. INTRODUCTION

Many large-scale service systems arising in healthcare, judicial and penal systems, and both front-office and back-office operations in business systems can be viewed as networks of multi-server queues with time-varying arrival rates [1,4,21–23,26]. The successful design and management of these systems requires allocating critical resources, such as the number of beds, nurses and associated equipment in a hospital ward, each of which can be represented generically as the number of servers at each queue. Moreover, this needs to be done in a dynamic way in order to respond effectively to the time-varying demand. As reviewed in [16], coping with time-varying arrival rates can be difficult for longer service times, because the level of time-varying demand extends after the arrival times by the service times of those arriving customers. We contribute here by developing an effective algorithm to set staffing

 0269-9648/14 $25.00

levels to stabilize performance at each queue within a network of queues, each of which may be given its own Quality-of-Service (QoS) performance target. We do so in a quite general setting, allowing customer abandonment from each queue and non-exponential service-time and patience-time distributions at each queue. It is important to include non-exponential service and patience distributions as well as time-varying arrivals because they commonly occur [4,5].

1.1. The Delayed-Infinite-Server Modified-Offered-Load (DIS-MOL) Approximation

This paper extends [30], which introduced a new framework to analyze the staffing problem for a single queue. An algorithm was developed to set time-dependent staffing levels (the number of servers) in order to stabilize abandonment probabilities and expected delays at specified QoS targets in the $M_t/GI/s_t + GI$ model, having arrivals according to a *non-homogeneous Poisson process* (NHPP, the M_t) with arrival-rate function $\lambda(t)$, independent and identically distributed (i.i.d.) service times with a general distribution (the first GI), a time-varying number of servers (the s_t, to be determined), i.i.d. patience times with a general distribution (times to abandon from queue, the final $+GI$), unlimited waiting space and the first-come first-served service discipline.

The DIS-MOL algorithm exploits *infinite-server* (IS) queues and is a MOL approximation (reviewed in Section 2). The key DIS idea is to obtain the OL by considering *two* IS queues in series, the first representing the waiting room and the second representing the service facility. In this artificial construction for generating an appropriate OL, each arrival is required to stay a constant waiting time w in the waiting room if that customer does not elect to abandon. Given that F is the patience time cumulative distribution function (cdf), each arrival abandons with probability $\alpha = F(w)$, so that the abandonment target α is linked to the constant w, which should correspond to the expected waiting time target. The expected number of busy servers in the second IS queue, $m_\alpha(t)$, serves as the new OL to be used in the new MOL approximation.

An initial staffing algorithm, called the DIS algorithm, simply staffs according to the OL itself, letting $s_\alpha(t) = \lceil m_\alpha(t) \rceil$, the least integer greater than or equal to $m_\alpha(t)$. The DIS algorithm was shown to be effective for suitably high OLs and abandonment-probability targets (low QoS), but the DIS algorithm is not successful in stabilizing performance at more typical abandonment probability targets occurring in well managed systems (high QoS).

To treat the important high QoS cases, the new MOL approximation, DIS-MOL, uses an approximation for the performance in the corresponding stationary $M/GI/s + GI$ model from [42] to set staffing at each time t to meet the new abandonment and delay targets (the usual minimum number of servers such that the QoS target is met), where the arrival rate at time t depends on the DIS OL $m_\alpha(t)$, in particular,

$$\lambda_\alpha^{\mathrm{mol}}(t) \equiv \frac{m_\alpha(t)}{(1-\alpha)E[S]}, \tag{1.1}$$

where $E[S]$ is the mean service time. (See Section 2.)

In [30], a heavy-traffic limit theorem was proved showing that both DIS and DIS-MOL are asymptotically correct as the scale increases for fixed targets. However, these fixed QoS targets provide low QoS as the scale increases. Hence, these limits only support the algorithms with a low QoS target. Simulation experiments showed that DIS-MOL staffing tends to coincides with DIS staffing for low QoS targets, where both work well. Thus, there is no need for DIS, but it is appealing for its simplicity.

In contrast, DIS-MOL is needed for high QoS. Simulation experiments in [30] confirmed that the DIS-MOL algorithm is effective in stabilizing abandonment probabilities and expected delays over a wide range of QoS targets, ranging from low QoS (high abandonment probability targets) to high QoS (low abandonment probability targets).

1.2. Extending the DIS-MOL Algorithm to Feed-Forward Networks

The purpose of the present paper is to investigate if the DIS and DIS-MOL algorithms can be extended to feed-forward networks of many-server queues, where there may be different performance targets at the different queues. We assume that we have a feed-forward $M_t/GI/s_t + GI$ network, by which we mean that all external arrival processes are mutually independent NHPPs, the service and patience times at all queues come from mutually independent sequences of i.i.d. random variables with general (queue-dependent) distributions, and each queue has unlimited waiting space and the first-come first-served service discipline. For this $M_t/GI/s_t + GI$ network, our goal is to determine staffing functions at each queue to stabilize abandonment probabilities and expected waiting times at targets set for each queue.

It is not difficult to *implement* a generalization of the algorithm in [30] for these networks, because we can simply apply the previous algorithm iteratively to each queue one at a time. It suffices to calculate a good approximation for the net arrival-rate function at each queue in the network. That is not difficult because the single-queue algorithm already calculates an approximate departure (service-completion) rate function, which simulation shows is very accurate. However, the effectiveness of the DIS and DIS-MOL staffing algorithms is not immediate. The primary question we address is: *Can the DIS and DIS-MOL staffing algorithms be effective for networks of many-server queues and, if so, when?*

To address that question, here we primarily focus on the special case of an $M_t/GI/s_t + GI$ network with two queues in series, which evidently embodies the primary difficulties among feed-forward networks. We assume an M_t arrival process with arrival-rate function λ, service-time cdf's G_1 and G_2, patience cdf's F_1 and F_2, delay targets $w_1 > 0$ and $w_2 > 0$, and abandonment probability targets $\alpha_1 \equiv F_1(w_1)$ and $\alpha_2 \equiv F_2(w_2)$.

First, we obtain a strongly supporting asymptotic result for general feed-forward networks. In Section 6 we prove that both the DIS and DIS-MOL algorithms achieve the objective asymptotically as the scale increases in general feed-forward $G_t/GI/s_t + GI$ networks with fixed QoS targets. However, with fixed abandonment probability and expected delay targets, the limit puts the model in the overloaded ED many-server heavy-traffic regime [14], where the DIS and DIS-MOL algorithms are asymptotically equivalent, both approaching the corresponding staffing algorithm for a limiting fluid model [27–29]. In that limit with increasing scale, the fixed QoS targets become low QoS targets. Our simulation results confirm that the proposed algorithm performs well for *all* arrival processes for such low QoS targets.

Of course, the systems of primary interest in applications tend to have high QoS targets. From [30], we know that DIS-MOL is needed for high QoS targets. From extensive simulation experiments, we find that DIS-MOL is *not* always effective for two queues in series when the second has a high QoS target. Fortunately, however, we find that bad behavior only occurs when the first queue has a low QoS. The simulations show that DIS-MOL is effective in the important common cases when *both* queues have high QoS targets. See Section 8.1 for a summary of our detailed conclusions.

1.3. Statistical Tests of Departure Processes

Since the previous DIS-MOL algorithm was shown to be effective for the $M_t/GI/s_t + GI$ model with an NHPP arrival process, and since the DIS model produces a good

approximation for the departure rate function, it is evident that the DIS-MOL algorithm should remain effective at the second queue if the departure process from the first queue is approximately an NHPP. We are thus led to ask the question: *When is the departure process from a many-server $M_t/GI/s_t + GI$ queue approximately an NHPP?*

Important insight can be gained by considering the associated $M_t/GI/\infty$ IS queue. The good simulation results for high QoS evidently can be attributed to the fact that the departure process from an $M_t/GI/\infty$ IS queue is an NHPP; see Theorem 1 of [11].

More generally, to gain insight into departure processes from queues with time-varying arrival rates, it is natural to start by asking the more elementary question: *When is the stationary departure process from a stationary many-server $M/GI/s + GI$ model approximately a Poisson process?*

First, it is known that the stationary departure process from the $M/GI/s$ model (without customer abandonment) is Poisson if and only if the service distribution is exponential, but even in the more general $G/GI/s$ system, if we let both s and the mean service time increase, then the departure process approaches a Poisson process; see [40] and references therein. Unfortunately, the addition of abandonment does not help, as can be seen by considering the limiting $M/GI/s/0$ loss model, corresponding to very fast abandonment. It is well known that, because of the blocking, the departure process of served customers (the carried traffic) is smoother than Poisson; see [6,25] and references therein. Even the departure process from the Markovian $M_t/M/s_t + M$ system is not exactly an NHPP.

Hence, it is natural to statistically analyze data from departure processes, either from system data or simulation of mathematical models. At first glance, we might think that more elementary question about the stationary model would be easily resolved by simply looking at the histogram of inter-departure times and seeing if it is nearly exponential. However, we show that the departure process from an $M/GI/s + GI$ many-server queue routinely passes that test and yet can be far from Poisson.

We find that the NHPP property can be effectively studied with the *index of dispersion for counts* (IDC),

$$I(t) \equiv \frac{\mathrm{Var}(D(t))}{E[D(t)]}, \quad t \geq 0, \tag{1.2}$$

for suitably large t, just as in [7,13,37].

It is significant that the IDC also applies nicely to point processes with time-varying rates. For any NHPP, $I(t) = 1$ for all t. When $I(t)$ differs significantly from 1, we judge that the departure process is not nearly an NHPP. We estimate the IDC of each departure process by performing independent replications in the simulations. We show that the IDC allows us to determine whether or not the departure process can be regarded as an NHPP in the performance prediction. This test is also important for a single queue to check if arrival data are consistent with the NHPP assumption in [30].

The statistical analysis of departure processes conducted here is intimately connected to recent work on statistical tests of a Poisson process and an NHPP in [5,19,20] and earlier work on indices of dispersion [7,13,37]. We compare these methods in Section 4.

1.4. Organization of the Paper

Here is how the rest of this paper is organized: We start in Section 2 by reviewing the MOL approximations. In Section 3, we present the results of our basic experiment for two many-server queues in series. In Section 4, we conduct statistical tests of the departure processes from the first queue to see if it is approximately NHPP. We show that the IDC is effective in predicting when the DIS-MOL approximation will be effective. In Section 5,

we conduct additional experiments to gain more insight; for example, we consider examples with smaller scale and different service-time distributions. In Section 6, we establish the asymptotic effectiveness of DIS and DIS-MOL (which coincide in the limit) as the scale increases. In Section 7, we give the DIS performance functions at the second queue of two queues in series. We draw conclusions in Section 8. Additional material appears in the e-companion and an online appendix.

2. BACKGROUND

There is a substantial literature on queues with time-varying arrival rates, which tends to be split between the two cases: (i) high QoS targets and (ii) low QoS targets. Of course, there is no fixed boundary between these two cases, and the definition depends on the scale (typical numbers of busy servers). Nevertheless, the classification is useful.

2.1. High QoS Targets

Not only are customers more satisfied with high QoS, but performance is easier to predict, so that the system is more easily managed; see the review [16]. With shorter service times, it is usually possible to apply familiar stationary models in a non-stationary way, using a variant of the pointwise stationary approximation (PSA), but for longer service times the PSA becomes highly inaccurate and alternative approximations are needed, such as MOL [12,18], the simulation-based iterative staffing algorithms (ISA) [8,12], the stationary backlog carry-over approach [38] and the Gaussian skewness method [34].

2.1.1. MOL and DIS-MOL. The main idea with an OL approach is to see how many resources (servers) would be used if there were no limit on their availability, which is achieved by using an IS model. If the system starts empty in the distant past, then the OL is the mean number of busy servers in the $M_t/GI/\infty$ model, which is

$$m_0(t) = \int_{-\infty}^{t} \lambda(s)P(S > t-s)\,ds = E[\lambda(t-S_e)]E[S] = E\left[\int_{t-S}^{t} \lambda(u)\,du\right], \qquad \textbf{(2.1)}$$

where S is a generic service time and S_e is a random variable with associated stationary-excess cdf; see [11]. The basic MOL approximation for the $M_t/GI/s_t + GI$ model is to staff aiming to stabilize the probability of delay using the stationary $M/GI/s + GI$ model (or an approximation of it, for example, [14,42]) at each time t, but with the arrival rate

$$\lambda^{\mathrm{mol}}(t) \equiv \frac{m_0(t)}{E[S]}, \qquad \textbf{(2.2)}$$

at time t. (That is done because in a stationary model the OL is $m \equiv \lambda E[S]$.) For general performance prediction, the MOL approximation tends to be effective if the targeted QoS is not too low. When the MOL approximation is used to set staffing to stabilize performance, the MOL approximation becomes more effective, because the performance becomes consistent over time. In [12,18] it is shown that MOL is effective in stabilizing delay probabilities across a wide range of performance targets.

The corresponding DIS-MOL algorithm from [30] aims to stabilize the expected waiting time at $E[W(t)] = w$ and the time-dependent abandonment probability at $\alpha = F(w)$, where

F is the patience cdf and $W(t)$ is the time an arrival at time t would wait if that arrival had infinite patience. Since networks of IS models are easy to analyze [35], there are explicit formulas for the DIS OL, $m_\alpha(t)$. As indicated in Section 1, $m_\alpha(t)$ is $E[B(t)]$, the mean number of busy servers in the second of two IS queues in series, with the first representing the waiting room and the second representing the service facility; that is,

$$m_\alpha(t) \equiv E[B(t)] = (1 - F(w))E[\lambda(t - w - S_e)]E[S], \tag{2.3}$$

which differs from $m_0(t)$ in Eq. (2.1) by the two places w appears: the multiplication by $1 - F(w)$ and the time shift by w itself; see Section 3 of [30]. As the QoS increases (as α and w decrease), $m_\alpha(t) \to m_0(t)$ in Eq. (2.1). Experience indicates that both MOL and DIS-MOL successfully stabilize all performance measures with a high QoS [12,30] (assuming an NHPP arrival process), but for low QoS they each achieve their separate goals, but not the other.

2.1.2. Extension to Networks. It is significant that the MOL approximation can easily be implemented in networks of queues without customer abandonment, because the $M_t/GI/\infty$ network (where again all external arrival processes are NHPPs) has been extensively studied under a variety of routing schemes among the queues in [35]. There it is proved that all the departure processes and arrival processes are NHPP's. Moreover, explicit formulas are given there for the arrival-rate function and the mean number of busy servers at each queue inside the network. Thus, the MOL approximation can be applied to networks by applying it to each queue using the exact expression for the OL for each queue. A special case of this MOL approximation has been applied to treat retrials arising in healthcare in [43].

The extension of the OL and MOL approximations to networks is more complicated with customer abandonment because the abandonment alters the departure rate from those queues where it occurs and thus alters the arrival rates at the queues. Thus it is natural to calculate the departure rate functions and the arrival-rate functions iteratively, treating one queue at a time, as discussed in Section 1. This is of course easily done recursively (without iteration) in feed-forward networks, as we will illustrate for two queues in series. Thus, the MOL, DIS and DIS-MOL staffing algorithms extend quite directly to feed-forward networks of $M_t/GI/s_t + GI$ queues, as specified in Section 1. We primarily focus on the special case of two $M_t/GI/s_t + GI$ queues in series, for which the DIS model has four IS queues in series, as depicted in Figure 1. Performance functions for the second queue in this DIS model are given in Theorem 7.1 and Corollary 7.1.

2.1.3. The Effectiveness of the Extension to Networks. While the implementation of MOL and DIS-MOL in feed-forward networks of many-server queues is straightforward, the effectiveness is not. Since the MOL and DIS-MOL algorithms employ stationary models with Poisson arrival processes, their effectiveness may well depend on the assumed NHPP (M_t) arrival process for the model. However, as discussed in Section 1.2, this M_t property does not propagate forward to the departure process. The effectiveness of the MOL approximations for networks of many-server queues was investigated to some extent in [43], but they restricted attention to the special case of the Markovian model with exponential service times and without abandonment. We find that the special case of exponential service times is significantly better behaved than others (see Section 5.2), but based on this study we conclude that the NHPP property propagates forward *approximately* and both MOL and DIS-MOL ought to perform well provided that the QoS targets are consistently high at all queues.

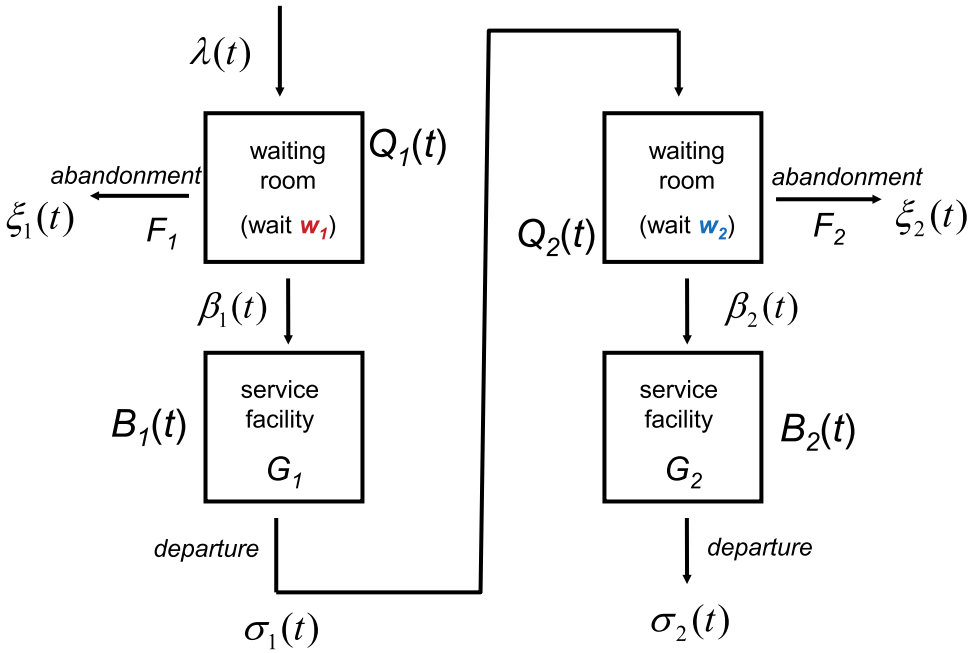

FIGURE 1. (Color online) The DIS approximation for two queues in series with delay targets w_1 and w_2.

2.2. Low QoS Targets

Low QoS targets lead to heavily loaded, even overloaded, queues, operating in the so-called *efficiency-driven* (ED) regime [14]. The same MOL and DIS-MOL algorithms can be applied, but as discussed in [12,30], abandonment probabilities are not stabilized by MOL with the objective of stabilizing the probability of delay, while delay probabilities are not stabilized with DIS-MOL with the objective of stabilizing abandonment probabilities.

For these overloaded queues, and more generally for queues that only experience some periods of overloading, the essential behavior of DIS-MOL and DIS is captured by deterministic fluid approximations and diffusion process refinements, arising from direct modeling or many-server heavy-traffic limits as in [28,29,31–34]. The fluid model provides important insight into the good performance we find for the DIS and DIS-MOL staffing algorithms with low QoS targets. We draw on those previous results to establish a new *functional weak law of large numbers* (FWLLN) in Section 6 showing that the DIS and DIS-MOL staffing algorithms are effective asymptotically as the scale increases with fixed QoS targets.

3. THE SIMULATION EXPERIMENT FOR TWO QUEUES IN SERIES

Our main experiment is the simulation of two queues in series with non-exponential service-time distributions and an NHPP arrival process with a sinusoidal arrival-rate function at the first queue.

3.1. The Design of the Experiment

We consider an $M_t/H_2/s_t + M$ network with two queues in series. We let the service-time distributions at both queues be a hyperexponential (H_2, mixture of two exponentials)

distribution with mean $E[S] = 1$, squared coefficient of variation (scv, variance divided by the square of the mean) $c^2 \equiv \mathrm{Var}(S)/(E[S])^2 = 4$ and balanced means, as on p. 137 of [39]. This distribution is significantly more variable than an exponential distribution and yet the variability is not extremely large. (We will discuss other service-time distributions in Section 5.)

We use the sinusoidal arrival-rate function

$$\lambda(t) = \bar{\lambda}(1 + r\sin(t)) = 100(1 + r\sin(t)), \quad t \geq 0, \tag{3.1}$$

with relative amplitude $r = 0.4$ We let A be a generic patience time with mean $\theta^{-1} = E[A] = 2$. Since the average OL is $\bar{m}(t) = \bar{\lambda}E[S] = 100$, the staffing will fluctuate around 100. We let the system start empty. (We will discuss lower OLs and staffing in Section 5.)

In each case, the simulation estimates are based on 1000 independent replications of the system over the time interval $[0, 20]$, starting empty. (In each run sampling is done over intervals of length 0.1.) Thus, there are approximately 2000 external arrivals in each run and 2×10^6 arrivals for each case. However, with low abandonment probabilities, the total abandonment rate is much less, such as about 1 when $\alpha = 0.01$. Hence, over any subinterval of length 1 the abandonment probability estimate is based on about 1000 observations. Further details about the way the time-dependent performance functions are estimated appear in the e-companion.

The simulation imposes a real system constraint: when the staffing level is scheduled to decrease with all servers busy, service in progress is completed before a server is allowed to leave, but server assignments can be switched when a server is scheduled to leave. Hence, when the staffing is scheduled to decrease with all servers are busy, a server is released when any one of the busy servers first becomes free. With a large number of servers, service switching greatly reduces the remaining time until a server can depart (roughly dividing it by the number of servers) [17].

3.2. Performance Results in the Four Cases

There are four main cases for the $M_t/H_2/s_t + M$ network, corresponding to all combinations of (i) a high QoS target (lower abandonment probability targets, $\alpha = 0.005, 0.010, 0.015, 0.020$) and (ii) a low QoS target (higher abandonment probability targets, $\alpha = 0.05, 0.10, 0.15, 0.20$). We show selected results for all these cases.

3.2.1. Low QoS (High Abandonment Probability) Targets at Both Queues. First, Figure 2 shows the estimated performance functions at the two queues, with the first queue on the left and the second on the right, for low QoS targets (relatively high α, that is, $\alpha = 0.05$, 0.10, 0.15 and 0.20) and DIS staffing ($s_\alpha(t) = \lceil m_\alpha(t) \rceil$) at both queues. (The performance with DIS-MOL is essentially the same.) For these plots, the same targets are used at both queues, so we reduce the total number of cases considered from $4 \times 4 = 16$ to 4. The dashed red lines are the targets and the DIS approximation for the mean queue length. Recall that the mean service time has been set at $E[S] = 1$, so that the units for delays are mean service times.

The first plots on the top show the arrival-rate function to that queue (which is the departure rate from the first queue on the right), while the sixth (bottom) plots show the DIS staffing functions. The third and fifth plots show the time-dependent abandonment probability and the expected delay (the average virtual waiting time, that is, the time that an arrival at time t would wait, if that arrival had unlimited patience) in each case, because these are the performance functions that the algorithm is designed to stabilize. In addition, we see that $\alpha \approx \theta EW$, because $w = -/\theta \log(1 - \alpha) \approx \alpha/\theta$.

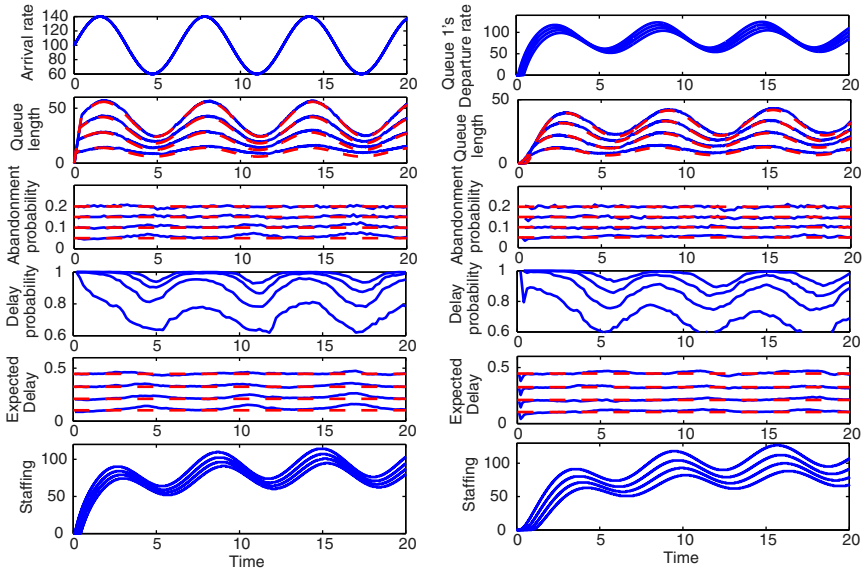

FIGURE 2. (Color online) Performance functions in the $M_t/H_2/s_t + M$ network the sinusoidal arrival rate in Eq. (3.1) for $\bar{\lambda} = 100$ and $r = 0.4$: the cases of low QoS targets ($\alpha = 0.05$, 0.10, 0.15 and 0.20) and simple DIS staffing at both queues.

In the second and fourth plots, we also show that the average number in queue (the "queue length") and the delay probability (the probability that an arrival would have to wait in queue before starting service), are not directly stabilized. The second plot shows the average queue length, which agrees closely with the analytical approximation formula in [30] and Section 7 in each case, has substantial variations. The fourth plot shows the delay probability, which is also not stabilized. The delay probability starts off at 1 at time 0, because the staffing algorithm does not start staffing until time w_i. (In practice, this feature of DIS staffing is likely not to be used.)

3.2.2. High QoS (Low Abandonment Probability) Targets at Both Queues. Figure 3 shows the corresponding estimated performance functions at the two queues for high QoS targets (lower abandonment probability targets, $\alpha = 0.005$, 0.010, 0.015 and 0.02) and DIS-MOL staffing at both queues. Again, the same targets are used at both queues, so we reduce the total number of cases considered from $4 \times 4 = 16$ to 4.

Here we see that all performance functions become more stable as the abandonment probability target decreases. At the lowest value $\alpha = 0.005$, all performance measures are stabilized remarkably well. At the highest abandonment probability target here, $\alpha = 0.02$, the abandonment probability and expected delay are stabilized quite well, clearly much better than the expected queue length and the delay probability. It is significant that the performance functions at the second queue behave much like they do at the first.

For perspective, it is helpful to examine the impact of a single agent in the staffing, as was done in Section EC.11 of [30]. Table EC.5 there shows that, for arrival rate 100 and mean service time 1, a single agent changes the abandonment probability about 9% when the abandonment probability target is 0.10 and about 19% when the abandonment probability target is 0.01. Very roughly, the observed fluctuations in the stabilizing cases are within this range.

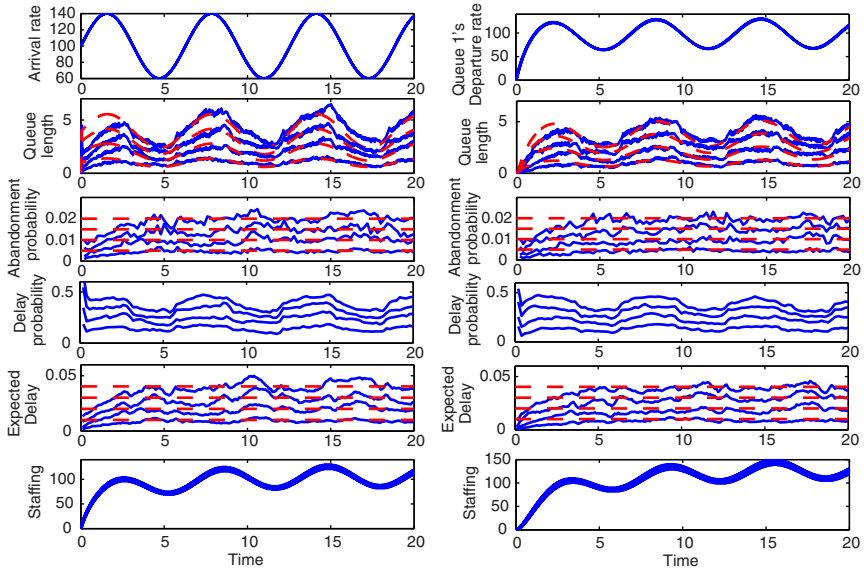

FIGURE 3. (Color online) Performance functions in the $M_t/H_2/s_t + M$ network with the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$: the cases of high QoS targets ($\alpha = 0.005$, 0.01 and 0.02) and DIS-MOL staffing at both queues.

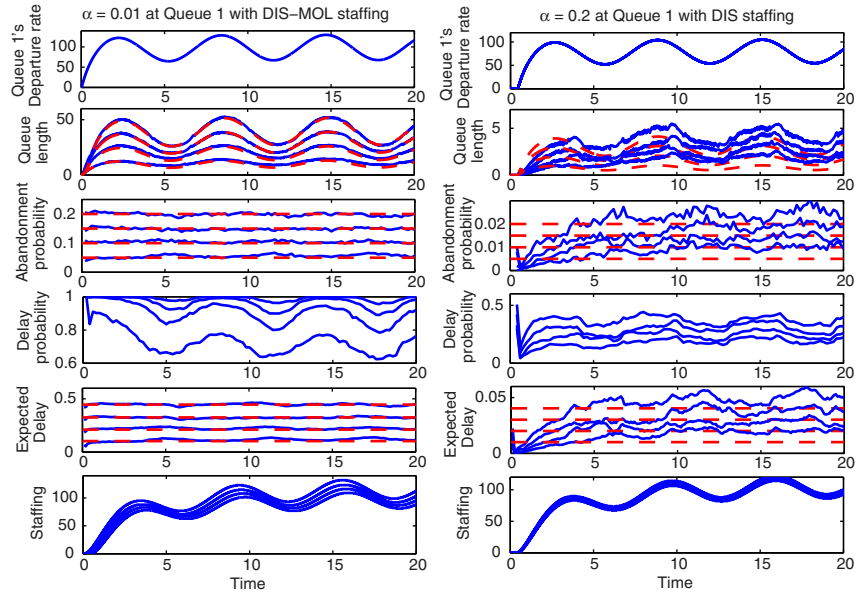

FIGURE 4. (Color online) Performance functions at the second queue in a $M_t/H_2/s_t + M$ network with the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$: the cases of low QoS targets ($\alpha = 0.05$, 0.10, 0.15 and 0.2) and DIS staffing at the second queue with fixed target $\alpha = 0.010$ at the first queue on the left, and high QoS targets ($\alpha = 0.005$, 0.010, 0.015 and 0.020) and DIS-MOL staffing at the second queue with fixed target $\alpha = 0.20$ at the first queue on the right.

3.2.3. The Two Mixed Cases. Figure 4 shows results for the mixed cases with low QoS targets (high abandonment probability target, $\alpha = 0.05$, 0.10, 0.15 and 0.20) and DIS staffing at one queue, but high QoS targets (low abandonment probability targets, $\alpha = 0.005$, 0.010, 0.015 and 0.020) and DIS-MOL staffing at the other queue. Since we already have shown the performance at the first queue in these cases in Figures 2 and 3, we now show only the performance at the second queue. Figure 4 shows on the left the performance measures at the second queue with low QoS targets and DIS staffing, always using the high QoS target $\alpha = 0.01$ and DIS-MOL staffing at queue 1. Figure 4 shows on the right the performance measures at the second queue with high QoS targets and DIS-MOL staffing, always using the low QoS target $\alpha = 0.20$ and DIS staffing at queue 1. (The fixed values at the first queue were chosen to be unambiguously high and low QoS, respectively.)

Figure 4 shows that the performance (abandonment probability and expected delay) is remarkably good for DIS at the second queue on the left, but significantly worse for DIS-MOL at the second queue on the right. On the right, the performance prediction might be judged adequate for practical engineering purposes, but clearly the abandonment probability is neither stabilized nor centered at its target. In the next section, we will show that the poor performance of DIS-MOL on the right can be explained by the fact that the departure process is not nearly NHPP, as required for the MOL refinement using the stationary $M/GI/s + GI$ model. The higher abandonment probabilities evidently occur because the departure process is more bursty (variable) than Poisson in this example.

4. DIRECT ANALYSIS OF DEPARTURE PROCESS SIMULATION DATA

In order to better understand when the DIS-MOL approximation will be effective at a downstream queue in a feed-forward network, in this section we carefully examine the departure process from the $M_t/H_2/s_t + M$ model with sinusoidal arrival rate in Eq. (3.1) that we are using for the first queue. We do so by analyzing the departure process data obtained from the simulation experiments. It is natural to expect that this departure process with time-varying rate would be approximately an NHPP if the stationary departure process from the associated stationary $M/H_2/s + M$ model is approximately a Poisson process. For the stationary model, we use the long-run average constant arrival rate $\bar{\lambda} = 100$ (obtained by letting the relative amplitude be $r = 0$), but all other parameters kept fixed. Hence, we first look at that more elementary stationary model to gain insight. We then directly examine the departure process from the $M_t/H_2/s_t + M$ model with sinusoidal arrival rate in Eq. (3.1) for non-zero relative amplitudes.

4.1. The Stationary Departure Process from the Stationary Model

We start by examining the stationary departure process from the stationary $M/H_2/s + M$ model with the same parameters except for the arrival process, which now is given as the long-run average constant arrival rate $\bar{\lambda} = 100$, obtained by setting $r = 0$ in Eq. (3.1). Again, we do so by simulating this stationary model over the time interval $[0, 20]$. Since the simulation starts with the queue empty, we collect departure process data over the interval $[6, 20]$ to allow the system to approach steady state. This is confirmed by plots of the estimated departure rate function over the interval $[0, 20]$ (in the appendix). We conduct multiple independent replications to obtain large samples, for example, of order 10^6. The sample size in each run is approximately $\bar{\lambda}(1 - \alpha)T = 100(20 - 6)(1 - \alpha) = 1400(1 - \alpha)$.

4.1.1. The Interdeparture-Time Distribution. A common way to investigate whether or not a constant-rate point process can be regarded as a Poisson process is to estimate the distribution of the times between successive points and see if it is approximately exponential. That can be done in various ways, a simple one being to look at the histogram.

We find that *all* the stationary departures processes from the stationary $M/H_2/s+M$ model pass this test of a Poisson process with flying colors, as illustrated by Figure 5 for the three abandonment probability targets $\alpha = 0.5, 0.05, 0.005$. Corresponding plots for other cases appear in the appendix. The plots on the left show the histograms of the interdeparture times for different α; the plots on the right compare the estimated probability density function (p.d.f.) f (normalized histogram) to (i) the exponential p.d.f. with the estimated mean and the scaled H_2 service-time distribution with the estimated mean. Also plotted on the right is an estimate of the hazard (or failure) rate function $h(x) \equiv f(x)/\bar{F}(x)$, where $\bar{F}(x) = 1 - F(x)$, which will be constant if and only if the p.d.f. is exponential. Here the estimation for $h(x)$ becomes less accurate for $x > 0.07$ due to the lack of samples with extremely large service times ($>0.07 \times 100 = 7$). Clearly, Figure 5 and the similar figures for the other cases show that the interdeparture-time distribution in the stationary departure process from the stationary $M/H_2/s+M$ model with OL $\bar{\lambda}E[S] = 100$ is very closely approximated by an exponential distribution.

4.1.2. The Lag-k Correlations. Those interdeparture-time distribution tests are so convincing that we might be inclined to stop there, being fully convinced, but we have not yet looked at the *joint* distribution of several successive intervals. A quick way to check on the dependence is to estimate the lag-k correlations for a few k, for example, $k = 1, 2$. When we do this, we see that these correlations are consistently very small. Thus we should be even more convinced.

Moreover, there is a limit theorem for the departure processes in $M/GI/s$ queues in [40], showing that the departure process converges to a Poisson process as the mean service time

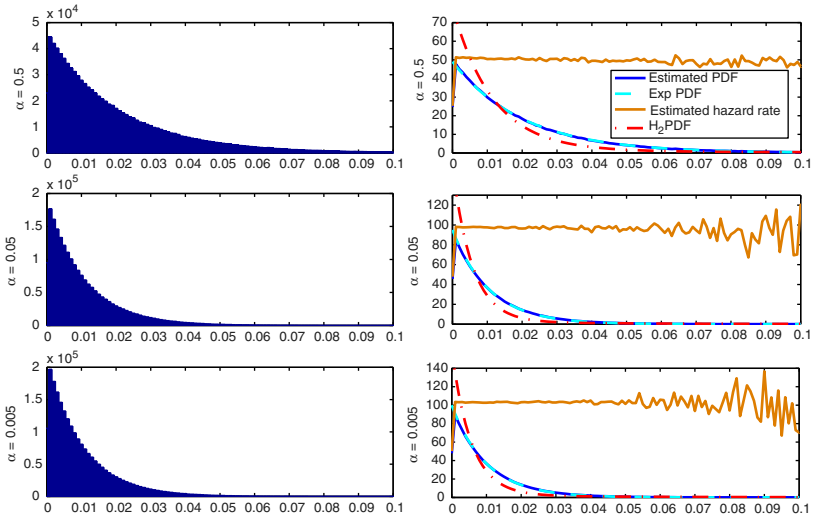

Figure 5. (Color online) Histograms of the interdeparture times from the stationary $M/H_2/s+M$ model (on the left) and fitted densities and hazard rate functions (on the right): the cases of low QoS targets and DIS staffing: a range of abandonment probability targets $\alpha = 0.5, 0.05, 0.005$.

and number of servers increases. However, that convergence is expressed in the topology of uniform convergence over bounded intervals. That implies that the departure process should look like a Poisson process *locally* as the scale increases appropriately.

4.1.3. The IDC. However, when the servers are often all busy, as will be the case with higher abandonment-probability targets, the departure process is similar to the superposition of i.i.d. renewal processes, each having service times as the i.i.d. interrenewal times. Experience with superposition arrival processes, for example, in [2,3,13,37] indicates that the process may look different in a longer time scale. Indeed, for any fixed m, the superposition of m i.i.d. renewal processes tends to behave just like a single renewal process in a sufficiently long time scale. For example, it has the same central limit theorem behavior; see [39] and Sections 9.4 and 9.8 of [41].

Thus, just as for superposition processes, to look at the departure process across a wide range of time scales, it is helpful to look at the IDC, $I(t) \equiv \mathrm{Var}(D(t))/E[D(t)]$ as a function of time t, where $D(t)$ is the departure counting process, as discussed in [7,13,37]. This description of point processes is also appealing because it extends naturally to nonstationary point processes with time-varying arrival-rate functions, with $I(t) = 1$ for all t for an NHPP.

And, indeed, we obtain a very different view when we look at the IDC $I(t)$. We estimate $I(t)$ by taking multiple replications, again starting to collect data at time 6 in each run. Estimates of $I(t)$ for $0 \le t \le 14$ are shown in Figure 6, again for stationary $M/H_2/s + M$ model, in the cases of high and low QoS targets, on the left and right, respectively. For the high QoS targets on the left, Figure 6 shows that $I(t) \approx 1$, indicating that the departure process is approximately Poisson across a range of time scales. Indeed, we see that $I(t)$ is actually somewhat *less* than 1, presumably because there is some smoothing caused by the customer abandonment.

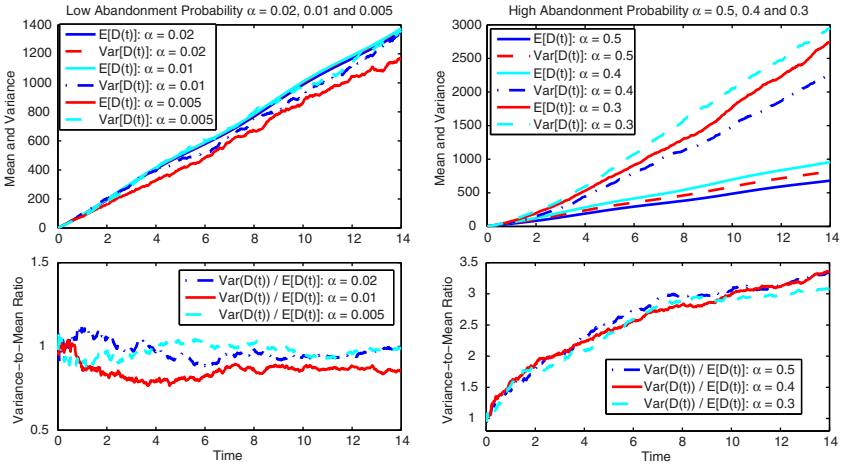

FIGURE 6. (Color online) Estimates of the mean $ED(t)$, variance $\mathrm{Var}(D(t))$ and IDC $I(t) \equiv \mathrm{Var}(D(t))/ED(t)$ for the departure counting process from the stationary $M/H_2/s + M$ queue with arrival rate 100: the cases of DIS-MOL staffing and high QoS targets (low α, on the left) and DIS staffing and very low QoS targets (high α, on the right).

A theoretical reference point for these good results is the associated $M_t/GI/\infty$ IS model. It is well known that the departure process is *exactly* an NHPP in the IS model; see Theorem 1 of [11]. Since the $M_t/GI/s_t + GI$ model tends to be quite similar to this ideal IS model if the QoS target is sufficiently high, these results for high QoS targets should not be too surprising.

However, in stark contrast, for low QoS targets, the plots on the right in Figure 6 show that $I(t)$ is much greater than 1, increasing towards 4, the limiting value of the IDC for a single H_2 renewal process, which would be the limiting value of the IDC of the counting process associated with the superposition of the fixed number 100 i.i.d. H_2 renewal processes with $c^2 = 4$; see Section 9.8 of [13,41]. We thus conclude that, just as in superposition processes, the cumulative impact of many small correlations over many interdeparture times prevents the departure process from being approximately a Poisson process over a longer time scale when the QoS target is low.

The IDC provides information about the lag correlations. Since the arrival rate is 100, the departure rate is nearly 100. Thus we see approximately twice the sum of the first 100 correlations in $I(1)$, indicating that the sum of the first 100 lag-k correlations is about 0.25, which averages to 0.0025. As discussed in [7,37], we can look directly at the cumulative impact of the correlations among interdeparture times by looking at the corresponding *index of dispersion for intervals* (IDI), which is $I_i(n) \equiv nc^2_{D_n} \equiv n\mathrm{Var}(D_n)/(E[D_n])^2$, where D_n is the nth departure time, that is, the sum of n consecutive interdeparture times. The large-n values of $I_i(n)$ agree with the large-t values of the IDC $I(t)$. The cumulative impact of many small positive correlations is much easier to see by looking at the indices of dispersion than trying to estimate the small individual correlations.

4.2. The Index of Dispersion of the $M_t/H_2/s_t + M$ Departure Process

Having established the importance of the IDC of an arrival process for understanding performance in the queue, and seeing that the IDC applies naturally to point processes with time-varying rate functions as well as constant arrival-rate functions, with $I(t) = 1$ for all $t \geq 0$ for an NHPP, we now look directly at the IDC of the departure processes in the $M_t/H_2/s_t + M$ model with sinusoidal arrival rate in Eq. (3.1) and relative amplitude $r = 0.4$ at the first queue that we obtained from our simulation experiments in Section 3.

Figure 7 shows the results for the sinusoidal arrival-rate function. Figure 7 is consistent with Figure 6 for the stationary model. We see that the IDC is again consistently near 1 when the QoS is high, but is increasing toward 4 when the QoS is low. Thus, we conclude that, from the perspective of the IDC, the departure process is approximately Poisson for high QoS targets at the first queue, but it is not approximately Poisson for low QoS targets at the first queue. From Section 3, we see that the IDC is able to predict whether or not the DIS-MOL staffing algorithm will be effective at the second queue.

4.3. Kolmogorov–Smirnov (KS) Statistical Tests of an NHPP

We conclude this section by considering KS statistical tests of an NHPP, as first proposed by Brown et al. [5] and then subsequently studied further in [19,20]. These KS tests apply if the arrival-rate function can be regarded as approximately constant over appropriate subintervals. (The question of how to choose subintervals so that the arrival-rate function can be regarded as approximately constant over each subinterval is studied in [19].) Given intervals over which the rate is approximately constant, we combine the data from these subintervals after exploiting the classical conditional uniform (CU) property over each subinterval. The CU property of a Poisson process states that, conditional on the number of points observed

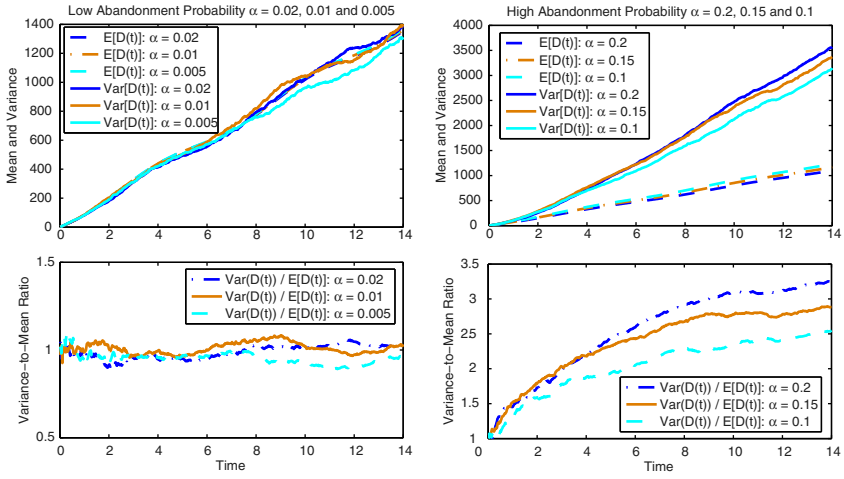

FIGURE 7. (Color online) Estimates of the mean $ED(t)$, variance $\mathrm{Var}(D(t))$ and IDC $I(t) \equiv \mathrm{Var}(D(t))/ED(t)$ for the departure counting process from the $M_t/H_2/100 + M$ queue with sinusoidal arrival-rate function in Eq. (3.1) having relative amplitude $r = 0.4$: the cases of high QoS targets (low α, on the left) and DIS-MOL staffing and very low QoS targets (high α, on the right) with DIS staffing.

in the subinterval, these values divided by the length of the interval are distributed as i.i.d. random variables uniformly distributed over $[0, 1]$. Thus, assuming that the piecewise-constant approximation is appropriate, under the NHPP hypothesis, all the data can be combined into one sequence of i.i.d. random variables uniformly distributed over $[0, 1]$. The first test is the CU KS test of the uniform distribution applied to the data that would be i.i.d. random variables uniformly distributed over $[0, 1]$ if the arrival process were an NHPP with piecewise-constant arrival-rate function.

However, Kim and Whitt, [20] showed that the CU KS test has remarkably little power against point processes with different marginal distributions. Hence, in [5,20] alternative KS tests are suggested based on additional transformations of the data. In [20], a KS test proposed by Lewis [24] based on a transformation due to Durbin [9] was found to have relatively high power against non-exponential marginal distributions. Thus, here we consider the Lewis KS test from [19,20], but we also consider the CU KS test, because it has more power than the Lewis test against departures from a Poisson process due to non-stationarity and dependence, which we have just shown turn out to be important in the present context.

Table 1 shows the results of the CU and Lewis KS tests of an NHPP applied to the departure process data over the interval $[6, 20]$ for several cases.

The cases include (i) high and low QoS targets and (ii) three sinusoidal arrival-rate tests with relative amplitudes $r = 0.0$ (constant), $r = 0.2$ (moderate fluctuations) and $r = 0.6$ (very high fluctuations). In order to apply the CU property we considered equally spaced subintervals of length L (over each of which the rate is treated as being approximately constant) for $L = 0.5$, 2.0 and 14.

As emphasized by [5,19], it is important to check if the data are rounded and, if so, appropriately unround the data by adding small i.i.d. uniform random variables to the observations. The present departure data were in fact rounded, so we first unrounded the data here. Both the raw (unrounded) and rounded data are shown in the appendix.

TABLE 1. The CU and Lewis KS tests applied to the departure processes over $[6, 20]$ from the $M_t/H_2/s_t + M$ model with the sinusoidal arrival-rate function in Eq. (3.1) in 18 cases: three relative amplitudes [$r = 0$ (constant), 0.2 and 0.6] and six abandonment probability targets, three low QoS [$\alpha = 0.5$, 0.4 and 0.3] and three high QoS [$\alpha = 0.02$, 0.01 and 0.005]. The KS tests are applied 20 times, once for each 25 replications in six cases: with rounded data and three subinterval lengths L: 0.5, 2 and 14.

Arrival Rate Fct.	Aband. Prob. Target	Result	Sample Size # n	$L = 0.5$ CU	$L = 0.5$ Lewis	$L = 2$ CU	$L = 2$ Lewis	$L = 14$ CU	$L = 14$ Lewis
$r = 0$	$\alpha = 0.5$	p-val	17,269	0.49	0.65	0.41	0.59	0.25	0.53
		# pass		19	20	20	20	12	20
Const.	$\alpha = 0.4$	p-val	20,652	0.49	0.55	0.52	0.43	0.30	0.45
		# pass		19	18	19	19	14	19
	$\alpha = 0.3$	p-val	24,292	0.45	0.59	0.59	0.47	0.16	0.48
		# pass		16	18	20	19	14	18
	$\alpha = 0.02$	p-val	33,863	0.43	0.37	0.64	0.34	0.40	0.36
		# pass		20	17	19	17	15	17
	$\alpha = 0.01$	p-val	34,272	0.42	0.29	0.47	0.30	0.28	0.22
		# pass		16	18	19	16	15	17
	$\alpha = 0.005$	p-val	34,453	0.56	0.38	0.52	0.35	0.33	0.39
		# pass		19	16	19	16	16	18
$r = 0.2$	$\alpha = 0.5$	p-val	17,018	0.53	0.41	0.50	0.44	0.00	0.18
		# pass		20	19	18	19	0	10
sine	$\alpha = 0.4$	p-val	20,482	0.55	0.49	0.46	0.61	0.00	0.11
		# pass		20	20	20	19	0	9
	$\alpha = 0.3$	p-val	23,966	0.51	0.48	0.39	0.60	0.00	0.16
		# pass		19	20	18	18	0	13
	$\alpha = 0.02$	p-val	33,782	0.43	0.48	0.12	0.47	0.00	0.27
		# pass		18	19	7	16	0	15
	$\alpha = 0.01$	p-val	34224	0.36	0.53	0.11	0.36	0.00	0.34
		# pass		17	19	9	18	0	16
	$\alpha = 0.005$	p-val	34,372	0.44	0.52	0.16	0.53	0.00	0.34
		# pass		19	19	9	18	0	17
$r = 0.6$	$\alpha = 0.5$	Avg p-val	16,523	0.52	0.47	0.42	0.05	0.00	0.00
		# pass		19	19	18	7	0	0
sine	$\alpha = 0.4$	p-val	19,968	0.27	0.38	0.09	0.02	0.00	0.00
		# pass		18	17	8	2	0	0
	$\alpha = 0.3$	p-val	23,358	0.33	0.40	0.03	0.01	0.00	0.00
		# pass		18	15	5	0	0	0
	$\alpha = 0.02$	p-val	33,757	0.42	0.29	0.00	0.00	0.00	0.00
		# pass		18	14	0	0	0	0
	$\alpha = 0.01$	p-val	34,180	0.49	0.55	0.00	0.00	0.00	0.00
		# pass		20	20	0	0	0	0
	$\alpha = 0.005$	p-val	34,344	0.28	0.38	0.00	0.00	0.00	0.00
		# pass		14	18	0	0	0	0

Table 1 shows results of 20 KS tests, each applied to the data from 25 replications, for the rounded data. Specifically, the number of tests out of 20 that pass at an 0.05 significance level and the p value, that is, the significance level at which the KS test would reject the Poisson hypothesis. For the first constant arrival-rate case, there is evidence that the CU

test detects the dependence for $L = 14$, but it does not consistently reject the Poisson hypothesis.

The remaining cases involve *both* non-stationarity (time dependence) and stochastic dependence. Unfortunately, the two effects of (i) non-stationarity and (ii) stochastic dependence are confounded. In order to have intervals where the rate is approximately constant, we would like to chose L relatively small (the cases with $L \leq 2.0$), but in order to see the full impact of the dependence (the cumulative impact of the many small correlations revealed by the IDC), we need to have L large.

Table 1 shows that for the very short intervals with $L = 0.5$, both KS tests of an NHPP consistently accept the NHPP hypothesis for the rounded data. However, that is consistent with our previous analysis, because we see only dependence over times less than L in the KS test; we do not see the cumulative impact of many small correlations. On the other hand, Table 1 shows that a significant deviation from the NHPP is detected in the case $L = 14$, especially by the CU KS test. Table 1 indicates the non-stationarity is a more significant departure from the Poisson property than the stochastic dependence, especially when $r = 0.6$.

We observe that the CU KS test is somewhat less conclusive than the IDC in rejecting the NHPP hypothesis, but each KS test is based on the data of only a 25 simulation runs, which involves a much smaller sample size than used to estimate the IDC. Overall, we conclude that these KS tests are consistent with the previous analysis of the departure process, but here (i) the CU test seems more effective than the Lewis test (for the reasons mentioned above) and (ii) the IDC evidently is more effective in detecting whether or not the departure process should be regarded as an NHPP, but we observe that it requires much more data. That data are routinely not difficult to obtain with simulation, because we can perform multiple replications. However, useful system data are much harder to obtain. In order to have suitable sample sizes from service system data, it is natural to combine data from multiple days, but we need to be cautious about overdispersion, caused by a random rate function on each day, as discussed in [19].

5. ADDITIONAL EXPERIMENTS

In this section, we present the results of additional experiments to add additional insight.

5.1. Lower Arrival Rates and Staffing

It is known that the performance of the MOL and DIS-MOL staffing algorithms improves as the scale increases. Nevertheless, these approximations can be useful for much smaller OLs, as shown by [43]. We illustrate by showing in Figure 8 the analog of Figure 3 for the same model except $\bar{\lambda}$ is reduced from 100 to 20.

See the appendix for more results. As the scale decreases, the discretization becomes a more and more serious issue. Thus there is a limit to the stabilization that can be achieved with very small scale. Here we increase the number of replications to 5000.

5.2. The Totally Markovian First Queue

To put the previous results in Section 3 in perspective, we also simulated the same $M_t/GI/s_t + M$ network with the sinusoidal arrival rate in Eq. (3.1) having relative amplitude $r = 0.4$ and $\bar{\lambda} = 100$ after changing the service-time distribution at the first queue from

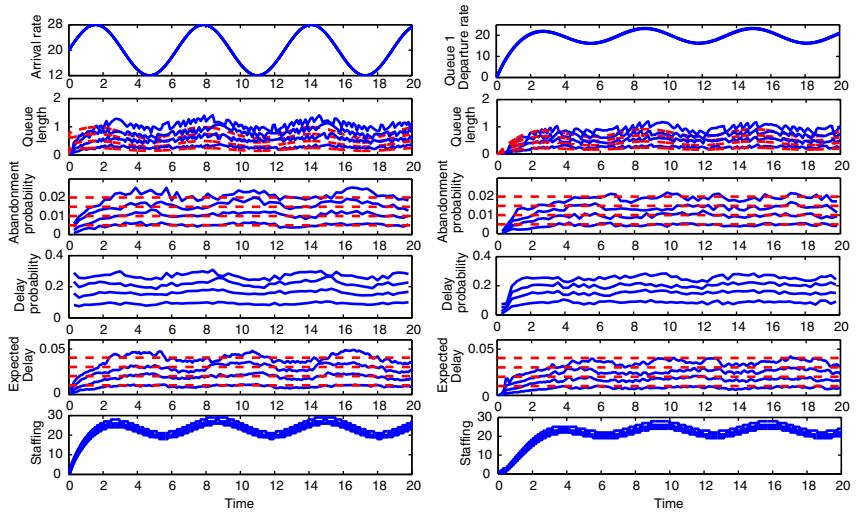

FIGURE 8. (Color online) Performance functions in the $M_t/H_2/s_t + M$ network with the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$ and $\bar{\lambda} = 20$: the cases of high QoS targets ($\alpha = 0.005$, 0.01 and 0.02) and DIS-MOL staffing at both queues.

H_2 to M, still with mean 1. We let the service-time distribution remain H_2 at the second queue.

The principal case of interest has a low QoS target at the first queue and a high QoS target at the second queue, which produces the bad results at the second queue for the H_2 service-time distribution at the first queue on the right in Figure 4. Thus we display the results for this case in Figure 9 below. As in Figure 4, we fix the abandonment probability

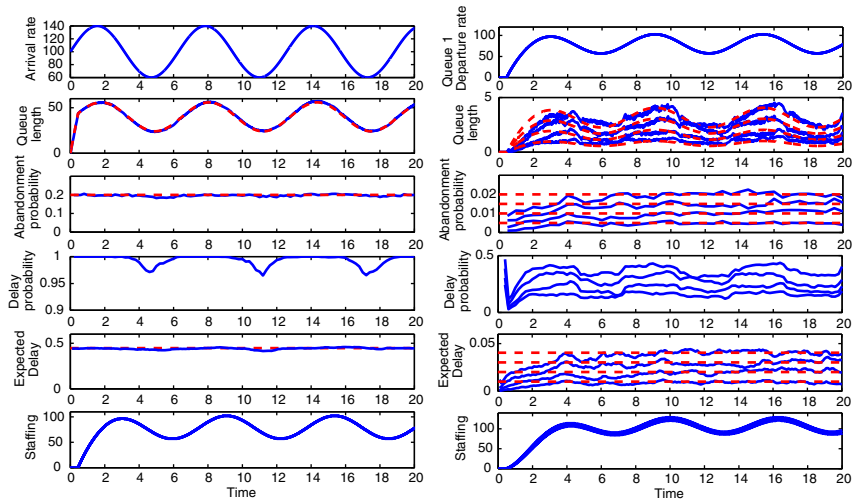

FIGURE 9. (Color online) Performance functions in the $M_t/GI/s_t + M$ network with the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$ and $\bar{\lambda} = 100$ when the service-time distribution is M at the first queue and H_2 at the second queue: the case of a fixed low QoS targets ($\alpha = 0.20$) and DIS staffing at the first queue and high QoS targets ($\alpha = 0.005$, 0.010, 0.015 and 0.020) and DIS-MOL staffing at the second queue.

target at $\alpha = 0.2$ for the first queue, so that we have one case at the first queue and four at the second queue.

Comparing Figure 9 to Figure 4, we see that the performance at the second queue with DIS-MOL staffing is now good instead of bad. In particular, the performance at the second queue is essentially the same as the performance at both queues in Figure 3.

5.3. Lognormal Service-Time Distributions

We also did experiments with lognormal (LN) service-time distributions instead of the H_2 distributions in Section 3, which are of interest because they have been found to fit service system data [4,5]. We found that the LN distribution with scv $c^2 = 4$ behaved like the H_2 distribution with $c^2 = 4$ in Section 3. However, we found that the LN distribution with $c^2 = 1$, which is similar to the data fitting results, behaved much like the M service-time distribution in Section 5.2, producing performance supporting DIS-MOL and IDC's supporting an NHPP approximation for the departure process. To illustrate, we plot the analog of Figure 9 for the case in which both service time distributions are LN with $c^2 = 1$ in Figure 10. See the appendix for other cases.

5.4. The Impact of a Non-NHPP External Arrival Process

We have observed that DIS-MOL is likely to perform poorly at the second queue with a high QoS (low-abandonment-probability) target there when the departure process from the first queue is not approximately an NHPP. We now show that the same problem occurs at a single $G_t/H_2/s_t + M$ queue when the external arrival process is not nearly M_t. Such an external arrival process has *both* strongly time-varying arrival rate and non-NHPP stochastic behavior. In particular, we consider the $H_2(t)/M/s_t + M$ network of two queues in series,

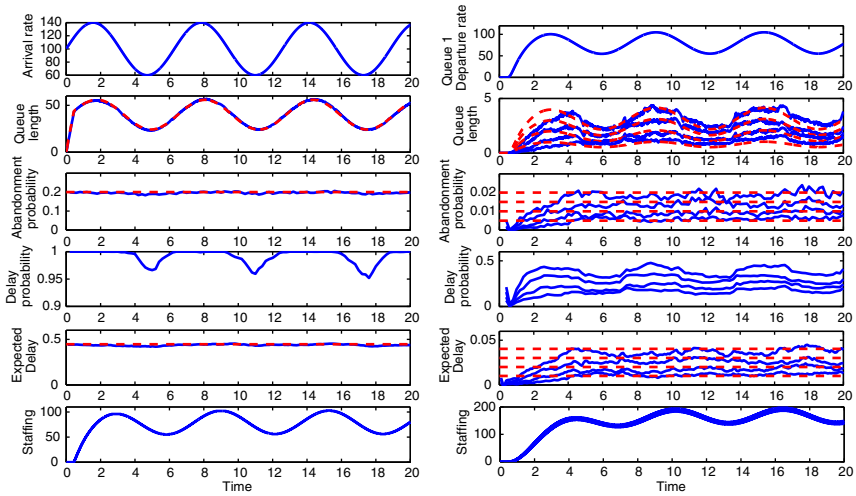

FIGURE 10. (Color online) Performance functions in the $M_t/LN/s_t + M$ network with the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$ and $\bar{\lambda} = 100$ when the service-time distribution is LN with scv $c^2 = 1$ at both queues: the case of a fixed low QoS targets ($\alpha = 0.20$) and DIS staffing at the first queue and high QoS targets ($\alpha = 0.005$, 0.010, 0.015 and 0.020) and DIS-MOL staffing at the second queue.

which differs from the totally Markovian $M_t/M/s_t + M$ network only by having an external arrival process that is a time-varying version of a renewal process with H_2 interarrival times.

5.4.1. Constructing a Non-NHPP Process with Time-Varying Arrival Rate. We use a standard construction to construct the $H_2(t)$ arrival process: Given any arrival-rate function $\lambda(t)$, let the associated cumulative arrival-rate function be defined by

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0. \tag{5.1}$$

This construction is a special case of the construction in Section 7 of [36]; it is used again in [15]. Let $A^e(t)$ be a rate-1 *equilibrium renewal process* (ERP), that is a standard renewal process with the first cycle replaced by the equilibrium version of the interarrival times. For the cumulative arrival-rate function Λ associated with any given arrival-rate function λ, which we take to be the specified sinusoidal arrival-rate function in Eq. (3.1) having $r = 0.4$, and an ERP $A^e(t)$ with H_2 interarrival times (constructed from H_2 random variables with mean 1 and $c^2 = 4$, just like the service-time distribution before), the $H_2(t)$ counting process we consider is defined by the simple composition

$$A(t) \equiv A^e(\Lambda(t)), \quad t \geq 0. \tag{5.2}$$

The stochastic process $A \equiv \{A(t) : t \geq 0\}$ inherits the time-dependence through Λ and the stochastic dependence through A^e. Since A^e is a stationary process, we have $E[A(t)] = \Lambda(t)$ for all $t \geq 0$.

5.4.2. Performance at the Queue. Since $c^2 = 4$, the IDC of the arrival processes $A_2(t)$ and $A(t)$ approach 4 as t increases. The conclusions in Section 3 about the performance at the *second* queue when the departure process is not nearly an NHPP now apply to the *first* queue, because the external arrival process is itself not nearly an NHPP. Figure 11 confirms that DIS-MOL is not effective at the first queue with high QoS targets because the arrival process is not nearly M_t, while DIS is effective at the first queue with low QoS targets because only the rate of the arrival process matters.

5.5. Three-Queue Models

It is interesting to see if the conclusions drawn in Section 3 extend to bigger feed-forward networks. We conduct two three-queue experiments in this subsection.

5.5.1. Three Queues in Series. We first show extensions of the experiment in Section 3 to the corresponding $M_t/H_2/s_t + M$ network with three queues in series, each with an H_2 service-time distribution, but now different means: 1.0, 0.8 and 1.2. The arrival-rate function is again sinusoidal as in Eq. (3.1) with relative amplitude $r = 0.4$. Figures 12 and 13 show the results, which are consistent with the good performance seen in Figures 2 and 3 before.

5.5.2. Three-Queue Distribution Model. In order to take into account the Markovian routing features in a general feedforward $M_t/GI/s_t + GI$ network, we next design a network of three $M_t/H_2/s_t + M$ queues, with Queue 1 feeding Queues 2 and 3 with probabilities $p_{1,2} = 1 - p_{1,3} = 0.6$. Each queue has an H_2 service-time distribution with mean 1 and

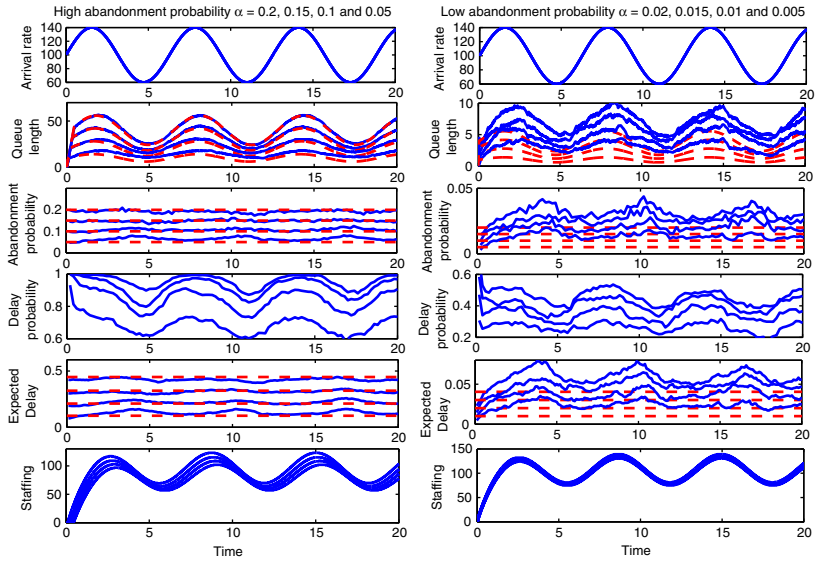

FIGURE 11. (Color online) Performance functions at a $H_2(t)/M/s_t + M$ queue with $H_2(t)$ arrival process having the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$ and $\bar{\lambda} = 100$: the cases of low QoS (high abandonment probability targets), $\alpha = 0.05$, 0.10, 0.15 and 0.20) and DIS staffing on the left and high QoS (low abandonment probability targets), $\alpha = 0.005$, 0.01, 0.015 and 0.02) and DIS-MOL staffing on the right.

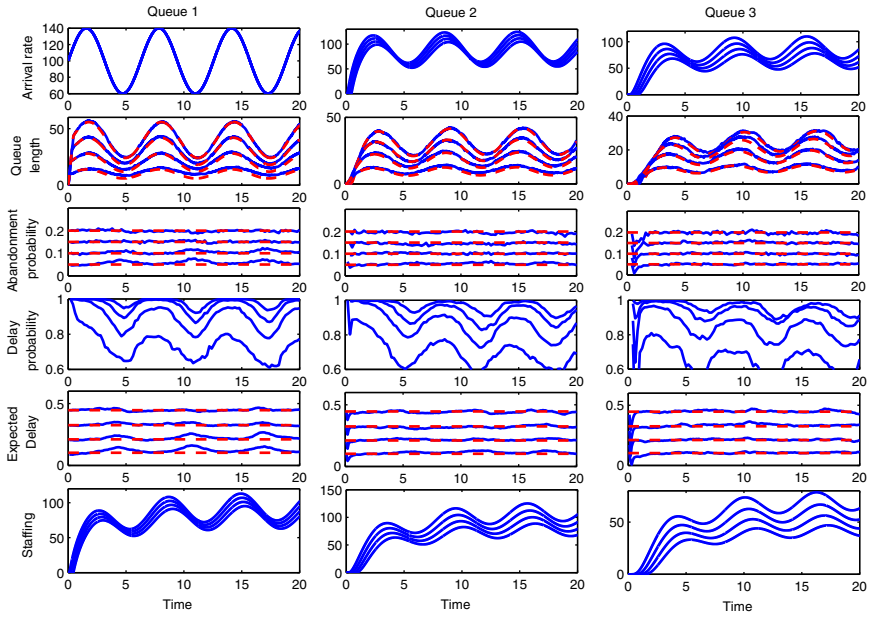

FIGURE 12. (Color online) Performance functions in the $M_t/H_2/s_t + M$ network with three queues in series and the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$ and mean service times 1.0, 0.8 and 1.2: the cases of identical low QoS targets ($\alpha = 0.05$, 0.10, 0.15 and 0.20) and simple DIS staffing at all queues.

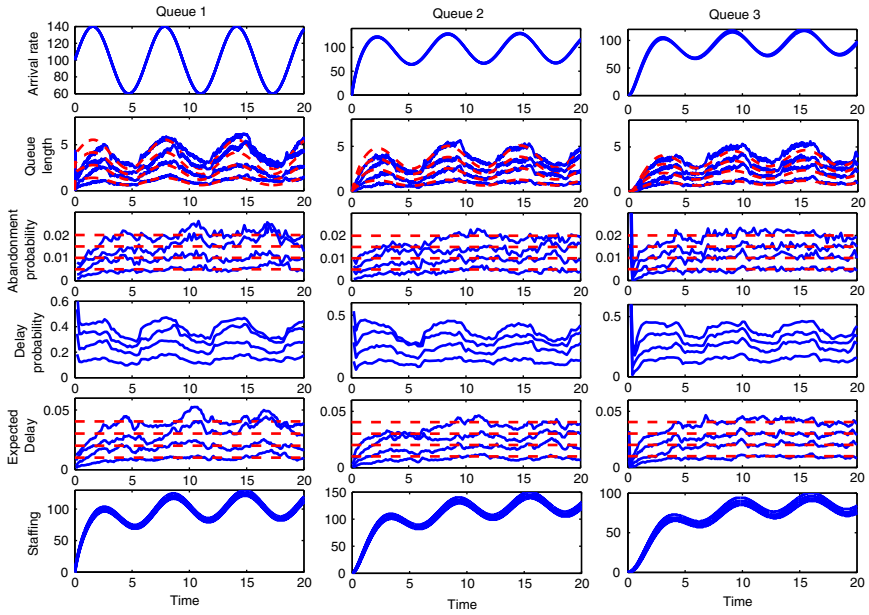

FIGURE 13. (Color online) Performance functions in the $M_t/H_2/s_t + M$ network with three queues in series and the sinusoidal arrival rate in Eq. (3.1) for $r = 0.4$ and mean service times 1.0, 0.8 and 1.2: the cases of identical high QoS targets ($\alpha = 0.005$, 0.010, 0.015 and 0.020) and DIS-MOL staffing at all queues.

$c^2 = 4$. The arrival-rate function is again sinusoidal as in Eq. (3.1) with relative amplitude $r = 0.4$. Figures in the appendix show good performance of DIS (for low QoS) and DIS-MOL (for high QoS), that are consistent with Figures 2 and 3 before.

6. ASYMPTOTIC EFFECTIVENESS IN FEED-FORWARD NETWORKS

Theorem 2 of [30] established that the simple DIS staffing algorithm with $s_\alpha(t) = m_\alpha(t)$ is effective asymptotically for any expected delay target $w > 0$ and abandonment probability targets $\alpha > 0$ related by $\alpha = F(w)$ as the scale increases in the $M_t/GI/s_t + GI$ model. We now extend that asymptotic result to all queues within a feed-forward $G_t/GI/s_t + GI$ network of queues, with fixed i.i.d. routing decisions when choices are available. In particular, we assume that departures from queue i will be routed to subsequent "downstream" queue j with probability $p_{i,j}$, independent of the history up to that time. For each queue, the sum of these routing probabilities is less than or equal to 1, with any excess probability corresponding to routing out of the network. Hence, each departure process with downstream queues may have its departure process split, and sent in multiple directions, possibly out of the network. Similarly, each queue with "upstream" queues that can send arrivals to it has an arrival process that is a superposition of its external arrival process and the "internal" arrival processes sent from upstream queues.

The first essential insight is that the performance targets involving positive values of w and α force the system to operate in the overloaded ED many-server heavy-traffic regime as the scale increases, as discussed for stationary models in [14]. As the scale increases, to achieve the targets it suffices to have a proportion α of the arrivals at any time eventually abandon, just as in the corresponding fluid model.

The second essential insight is that, while the main queueing model is quite complicated, the approximating DIS IS model depicted in fig1, for which formulas are given in Theorem 1 of [30] and Theorem 7.1 here, Has performance identical, except for interpretation, to the fluid model performance when we stabilize waiting times as in Section 10 of [28]. The expected values in the stochastic DIS IS model coincide with the deterministic values in the fluid model. The fluid model with general staffing has a much more complicated performance involving a fixed-point equation for the rate fluid enters service during each overloaded interval, but that complexity disappears when we stabilize the waiting times at a positive target. For both the fluid model and the DIS approximation, the performance is easily extended to feed-forward networks.

The third essential insight is that there is a scale proportionality for these IS and fluid models, discussed in Section 4 of [30], that implies that the performance scales by n if we multiply the arrival rates and staffing by n. Hence, we can start with external arrival-rate functions and staffing functions that coincide for the DIS and fluid models. Then we create a sequence of stochastic models by simply multiplying the arrival-rate function and staffing functions by n, and then discretizing the staffing function. As the scale n increases, the discretization becomes negligible.

Finally, we establish a FWLLN to show that the associated sequence of scaled performance processes in the original model converges to the fluid model. As in [30], we apply the many-server heavy-traffic FWLLN theorem from [29]. That FWLLN involves a sequence of $G_t/GI/s_t + GI$ queueing models indexed by n. We consider a corresponding sequence of feed-forward $G_t/GI/s_t + GI$ networks indexed by n, with a fixed number m of queues and fixed i.i.d. routing for departures from each queue, independent of n. We let the service and patience distributions at the queues be independent of n. At queue i, there are i.i.d. service times distributed as a generic random variable S_i with cdf G_i, and i.i.d. patience times of successive customers distributed as a random variable T_i with general cdf F_i. The cdf's G_i and F_i are differentiable, with positive p.d.f.'s g_i and f_i.

The arrival process $N_n(t)$ was assumed to be NHPP in [30], but greater generality is allowed by [29]. In order to simplify the proof, we exploit the essential insights above by letting the external arrival-rate functions in model n be scaled versions of fixed external arrival-rate functions.

Scaling the arrival-rate function works directly if we assume that the external arrival processes are NHPPs, but to allow greater generality we assume a specific process representation. We assume that each queue i that has an external arrival process has a *base* external arrival counting process that can be expressed as

$$N^{(i,e)}(t) = N^{(i,1)}(\Lambda_{i,e}(t)), \quad t \geq 0, \tag{6.1}$$

where $\Lambda_{i,e}(t)$ is a differentiable cumulative rate function with

$$\Lambda_{i,e}(t) \equiv \int_0^t \lambda_{i,e}(s)\, ds \tag{6.2}$$

and $N^{(i,1)} \equiv \{N^{(i,1)}(t) : t \geq 0\}$ is a rate-1 stationary point process satisfying a FWLLN, that is,

$$n^{-1}N^{(i,1)}(n\cdot) \Rightarrow e \quad \text{in} \quad D \quad \text{as} \quad n \to \infty, \tag{6.3}$$

where $e(t) \equiv t$ and $\Rightarrow$ denotes convergence in distribution in the function space D with the topology of uniform convergence over bounded subintervals of the domain $[0, \infty)$ as in [41].

In that framework, we then define the external arrival process at queue i in model n by letting

$$N_n^{(i,e)}(t) \equiv N^{(i,1)}(n\Lambda_{i,e}(t)), \quad t \geq 0, \tag{6.4}$$

which gives it cumulative arrival rate function $\Lambda_n^{(i,e)}(t) = n\Lambda_{i,e}(t)$, a simple multiple of the base arrival-rate function. On account of this construction and assumption Eq. (6.3), we deduce that $N_n^{(i,e)}$ also obeys the FWLLN

$$n^{-1}N^{(i,e)}(n\cdot) \Rightarrow \Lambda_{i,e} \quad \text{in} \quad D \quad \text{as} \quad n \to \infty. \tag{6.5}$$

Since the external arrival rates have been constructed by simple scaling, the associated DIS staffing can be constructed by simple scaling as well. Hence, in model n, we can use a time-varying number of servers $s_{n,\alpha}^{(i)}(t) \equiv \lceil n\, s_\alpha^{(i)}(t) \rceil$, which we assume is set by the DIS staffing algorithm, which is a scaled version of the staffing in the associated fluid model with cumulative arrival rate $\Lambda_{i,e}$.

We define the following performance functions for the nth model: Let $N_n^{(i)}(t)$ be the total number (external plus internal) arrivals at queue i in the interval $[0,t]$; let $Q_n^{(i)}(t)$ be the number of customers waiting in queue i at time t; let $W_n^{(i)}(t)$ be the corresponding potential waiting time, that is, the virtual waiting time at time t if there were an arrival at time t at queue i, assuming that arrival had unlimited patience; let $A_n^{(i)}(t)$ be the number of customers that have abandoned from queue i in the interval $[0,t]$; let $A_n^{(i)}(t,u)$ be the number of customers among arrivals to queue i in $[0,t]$ that have abandoned in the interval $[0,t+u]$; let $D_n^{(i)}(t)$ be the number of customers to complete service from queue i in the interval $[0,t]$; let $D_n^{(i,j)}(t)$ be the number of customers to complete service from queue i and be routed to queue j in the interval $[0,t]$. Define associated FWLLN-scaled processes: by letting $\bar{N}_n^{(i,e)}(t) \equiv n^{-1}N_n^{(i,e)}(t)$, and similarly for the other processes except the process $W_n^{(i)}(t)$ is not scaled.

We assume that the limiting arrival-rate functions $\lambda_{i,e}$, staffing function $s_\alpha^{(i)}(t)$, and cdf's G_i and F_i satisfy the assumptions of the fluid model in [28]. We assume that the regularity conditions in [28,29] are satisfied, in particular, the model elements $\Lambda_{i,e}$, F_i and G_i are differentiable functions with piecewise-continuous derivatives $\lambda_{i,e}$, f_i and g_i. We assume in addition that the service times have finite second moments. Let 1_C be the indicator variable, which is equal to 1 if event C occurs and is equal to 0 otherwise.

THEOREM 6.1 (Asymptotic Stability): *Consider a sequence of feed-forward $G_t/GI/s_t + GI$ networks with external arrival processes defined as in Eq. (6.4) and the many-server heavy-traffic scaling specified above. Suppose that these systems start empty at time 0, the regularity conditions in [28,29,32] are satisfied and $E[S_i^2] < \infty$ for all queues i. Then, for any set of abandonment-probability targets $\alpha_i > 0$ for all queues i (or related expected waiting time targets w_i with $\alpha_i = F_i(w_i)$), use the DIS staffing $s_{n,\alpha}^{(i)}(t) \equiv \lceil n\, s_\alpha^{(i)}(t) \rceil$, where*

$$s_\alpha^{(i)}(t) = m_\alpha^{(i)}(t) \equiv E[B_\alpha^{(i)}(t)] = \bar{F}_i(w_i) \int_0^{t-w_i} \bar{G}_i(x)\lambda_i(t-w_i-x)dx \cdot 1_{\{t>w_i\}}, \tag{6.6}$$

which involves first staffing queue i at time w_i, where the total arrival-rate function at queue i, λ_i, is obtained recursively via the equation

$$\lambda_j(t) = \lambda_{j,e}(t) + \sum_i p_{i,j}\sigma_i(t), \quad t \geq 0, \tag{6.7}$$

with $\sigma_i(t)$ being the departure rate function from queue i, that is,

$$\sigma_i(t) = \bar{F}_i(w_i)\left(\int_0^{t-w_i} \lambda_i(t-w_i-x)g_i(x)\,dx\right)1_{\{t>w_i\}}, \tag{6.8}$$

which necessarily is positive only after time w_i. Then, as $n \to \infty$, the expected delays and abandonment probabilities are stabilized at their targets w_i and α_i for all i for any time b_i with $w_i < b_i < \infty$. In particular,

$$\sup_{0\le t\le b_i}\{|\bar{Q}_n^{(i)}(t) - E[Q^{(i)}(t)]|\} \Rightarrow 0, \quad \sup_{0\le t\le b_i}\{|W_n^{(i)}(t) - w_i|\} \Rightarrow 0, \quad E[W_n^{(i)}(t)] \to w_i, \quad t \ge 0,$$
$$\sup_{0\le t\le b_i}\{|\bar{A}_n^{(i)}(t) - A^{(i)}(t)|\} \Rightarrow 0 \quad \textit{and} \quad \sup_{0\le t\le b_i,\, w_i<u<b_i}\{|\bar{A}_n^{(i)}(t,t+u) - A^{(i)}(t,u)|\} \Rightarrow 0 \tag{6.9}$$

as $n \to \infty$, where

$$E[Q^{(i)}(t)] = E[Q^{(i)}(t,0)] \equiv \int_0^{w_i} \lambda_i(t-x)\bar{F}_i(x)\,dx, \quad A^{(i)}(t) \equiv \int_0^t \xi_i(s)\,ds$$
$$\xi_i(t) \equiv \int_0^{w_i} \lambda_i(t-x)f_i(x)\,dx \quad \textit{and} \quad A^{(i)}(t,u) \equiv \Lambda_i(t)\alpha_i, \quad u > w_i. \tag{6.10}$$

The limit functions are the performance functions of the associated fluid network, constructed recursively from [28].

PROOF: Like the result, the proof is an extension of the proof of Theorem 2 of [30]. We can apply the results in [28,29] recursively and inductively. We exploit the deterministic limits of all assumed and established FWLLN's to obtain joint convergence of all individual limits established, invoking Theorem 11.4.5 of [41]. We also exploit the feed-forward assumption. In a feed-forward network there necessarily are some queues with only external arrival processes. Clearly, for these queues, $N_n^{(i)}(t) = N_n^{(i,e)}(t)$ and the assumed FWLLN for $N_n^{(i,e)}$ produces the FWLLN for $N_n^{(i)}$ with $\Lambda_i(t) = \Lambda_{i,e}(t)$ and $\lambda_i(t) = \lambda_{i,e}(t)$ for all t. For any queues for which the FWLLN for $N_n^{(i)}$ is established, we can obtain the FWLLN from the single-server result, Theorem 1 of [29]. That limit includes a limit for the associated total departure process $D_n^{(i)}$, that is, $\bar{D}_n^{(i)} \Rightarrow D^{(i)}$, with

$$D^{(i)}(t) = \int_0^t \sigma_i(s)\,ds, \quad t \ge 0, \tag{6.11}$$

where the service completion (departure) rate function $\sigma_i(t)$ given in Eq. (6.8) above, as in Theorem 8 of [28], which in turn follows from the general formula (9) of [28]. The key point is that the FWLLN for $D_n^{(i)}$ follows directly from [28,29,32] once the limit for $N_n^{(i)}(t)$ is established.

We now indicate how to obtain the FWLLN limit for $N_n^{(j)}(t)$ when the FWLLN limit for $D_n^{(i)}$ has been established for all upstream queues. An important step is actually *constructing* the net arrival process $N_n^{(j)}(t)$ at each queue j. For that purpose, let $\{X_{n,i,k} : k \ge 1\}$ be a sequence of routing i.i.d random variables with $X_{n,i,k} = j$ if the kth departure from queue

i is routed to queue j. Then we can represent $N_n^{(j)}(t)$ explicitly as

$$N_n^{(j)}(t) = N_n^{(j,e)}(t) + \sum_i \sum_{k=1}^{D_n^{(i)}(t)} 1_{\{X_{n,i,k}=j\}}, \quad t \geq 0, \tag{6.12}$$

and the associated scaled version as

$$\bar{N}_n^{(j)}(t) = \bar{N}_n^{(j,e)}(t) + \sum_i \bar{Z}_{n,i,j}(t) \circ \bar{D}_n^{(i)}(t), \quad t \geq 0, \tag{6.13}$$

where $\circ$ is the composition function and

$$\bar{Z}_{n,i,j}(\cdot) \equiv \frac{1}{n} \sum_{k=1}^{\lfloor n\cdot \rfloor} 1_{\{X_{n,i,k}=j\}} \Rightarrow p_{i,j} e \quad \text{in} \quad D \tag{6.14}$$

by the FWLLN for partial sums of i.i.d. random variables. Hence, given the FWLLN assumed for $N_n^{(j,e)}(t)$, the FWLLN for $\bar{Z}_{n,i,j}$ in Eq. (6.14) and the recursively established FWLLN limit for $D_n^{(i)}(t)$, it follows that a FWLLN holds for $N_n^{(j)}(t)$ with

$$N_n^{(j)}(t) \Rightarrow \Lambda_j(t) = \Lambda_{j,e}(t) + \sum_i p_{i,j} D^{(i)}(t), \quad t \geq 0. \tag{6.15}$$

In particular, we apply the continuous mapping theorem in Section 3.4 of [41] for the continuous addition and composition functions appearing in Eq. 6.13; see Theorems 12.7.3 and 13.2.1 of [41]. In this way we recursively obtain the FWLLN limits for all the external arrival processes. As in [30], the assumption that $E[S_i^2] < \infty$ for all i is used to justify the uniform integrability implying the convergence of the expected waiting times. Since we are applying the DIS algorithm to stabilize the expected waiting time, the limiting formulas follow Theorem 8 of [28]. ■

Consistent with Figure 2, from the representation of the DIS approximating mean queue length $E[Q_\alpha^{(i)}(t)]$ in Theorem 6.1, just as in Corollary 1 of [30].

7. THE DIS PERFORMANCE AT THE SECOND OF TWO QUEUES IN SERIES

For the $M_t/GI/s_t + GI$ two-queue series network, the DIS model consists of four IS queues in series, as depicted in Figure 1. Note that this is a legitimate mathematical model in its own right. The performance functions at the first two-queue DIS model are given in Theorem 1 of [30]. The performance functions at the second two-queue DIS model follow from that, using the departure rate from the first two-queue DIS model as the arrival rate to the second two-queue DIS model. To facilitate application of the extension to two queues in this paper, we now give explicit expressions for the DIS performance functions at the second two-queue DIS model in terms of the model parameters. This theorem has been used to compute the DIS OL $m_\alpha(t)$ at the second queue and the mean queue length, for example, as shown by the red dashed lines in Figure 2. For any random variable X, let X_e have the associated stationary-excess distribution, just as in Eq. (2.1).

THEOREM 7.1 (DIS performance at the second two-queue DIS model): *Consider the four-queue DIS model serving as am approximation for the two-queue $M_t/GI/s_t + GI$ series*

network with arrival-rate function λ, service-time cdf's G_1 and G_2, patience cdf's F_1 and F_2, delay targets $w_1 > 0$ and $w_2 > 0$, and abandonment probability targets $\alpha_1 \equiv F_1(w_1)$ and $\alpha_2 \equiv F_2(w_2)$. Let $T_i \equiv \min(A_i \wedge w_i)$ where A_i is a generic abandonment time at Queue i, $i = 1, 2$. Then the time-dependent random numbers of customers in the waiting room and service facility of the second two-queue DIS model, $Q_2(t)$ and $B_2(t)$, are independent Poisson random variables having means

$$E[Q_2(t)] = \alpha_1 E\left[\int_{t-T_2}^{t} \lambda(x - w_1 - S_1)\, dx\right] = \alpha_1 E\left[\lambda(t - w_1 - S_1 - T_{2,e})\right] \cdot E[T_2],$$

$$\begin{aligned} m_2(t) \equiv E[B_2(t)] &= \alpha_1\alpha_2 E\left[\int_{t-w_2-S_2}^{t-w_2} \lambda(x - w_1 - S_1)\, dx\right] \\ &= \alpha_1\alpha_2 E\left[\lambda(t - w_1 - w_2 - S_1 - S_{2,e})\right] \\ &= \alpha_1\alpha_2 \int_{-\infty}^{t}\int_{0}^{t} \lambda(x - y - w_1 - w_2)\, \bar{G}_2(t - x)\, dG_1(y)\, dx. \end{aligned}$$

At the second two-queue DIS model, the processes counting the numbers of customers abandoning, entering service and completing service are independent NHPPs with rate functions ξ_2, β_2 and σ_2, where

$$\begin{aligned} \xi_2(t) &= \alpha_1 E\left[\lambda(t - w_1 - S_1 - T_2) | T_2 < w_2\right], \\ \beta_2(t) &= \alpha_1\alpha_2 E\left[\lambda(t - w_1 - w_2 - S_1)\right] E[S_2], \\ \sigma_2(t) &= \alpha_1\alpha_2 E\left[\lambda(t - w_1 - w_2 - S_1 - S_2)\right]. \end{aligned}$$

The following Corollary draws on [10] and Theorem 6.3 of [35].

COROLLARY 7.1 The special case of sinusoidal arrival rate: *In the setting of Theorem 7.1, suppose that the arrival-rate function is the sinusoidal function $\lambda(t) = a + b \cdot \sin(ct + \psi)$, starting in the indefinite past. Then for $i = 1, 2$, $Q_i(t)$ and $B_i(t)$ are independent Poisson random variables having sinusoidal means*

$$E[Q_i(t)] = a_i^Q + b_i^Q \cdot \sin(ct + \psi_i^Q) \qquad \textit{and} \qquad m_i(t) \equiv E[B_i(t)] = a_i^B + b_i^B \cdot \sin(ct + \psi_i^B),$$

where

$$\begin{aligned} a_1^Q &\equiv a\, E[T_1], \quad b_1^Q \equiv b\, E[T_1]\, \gamma_c(T_{1,e}), \quad \psi_1^Q \equiv \psi - \theta_c(T_{1,e}), \\ a_2^Q &\equiv \alpha_1\, a\, E[T_2], \quad b_1^Q \equiv \alpha_1\, b\, E[T_2]\, \gamma_c(S_1)\, \gamma_c(T_{2,e}), \\ \psi_2^Q &\equiv \psi - (\theta_c(S_1) + \theta_c(T_{2,e}) + c\, w_1), \\ a_1^B &\equiv \alpha_1\, a\, E[S_1], \quad b_1^B \equiv \alpha_1\, b\, E[S_1]\, \gamma_c(S_{1,e}), \quad \psi_1^B \equiv \psi - (\theta_c(S_{1,e}) + c\, w_1), \\ a_2^B &\equiv \alpha_1\alpha_2\, a\, E[S_2], \quad b_1^B \equiv \alpha_1\alpha_2\, b\, E[S_2]\, \gamma_c(S_1)\, \gamma_c(S_{2,e}), \\ \psi_2^B &\equiv \psi - (\theta_c(S_1) + \theta_c(S_{2,e}) + c(w_1 + w_2)), \end{aligned}$$

$$\gamma_c(X) \equiv \sqrt{(E[\sin(cX)])^2 + (E[\cos(cX)])^2}, \quad \textit{and} \quad \theta_c(X) \equiv \arctan\left(\frac{E[\sin(cX)]}{E[\cos(cX)]}\right),$$

for a non-negative random variable X. The departure processes are NHPPs with sinusoidal rate functions

$$\sigma_1(t) = a_1^\sigma + b_1^\sigma \cdot \sin(ct + \psi_1^\sigma) \qquad \textit{and} \qquad \sigma_2(t) = a_2^\sigma + b_2^\sigma \cdot \sin(ct + \psi_2^\sigma),$$

where

$$a_1^\sigma \equiv \alpha_1\, a, \quad b_1^\sigma \equiv \alpha_1\, b\, \gamma_c(S_1), \quad \psi_1^\sigma \equiv \psi - (\theta_c(S_1) + c\, w_1)\,,$$
$$a_2^\sigma \equiv \alpha_1\alpha_2\, a, \quad b_1^\sigma \equiv \alpha_1\alpha_2\, b\, \gamma_c(S_1)\, \gamma_c(S_2), \quad \psi_2^\sigma \equiv \psi - (\theta_c(S_1) + \theta_c(S_2) + c(w_1 + w_2))\,.$$

8. CONCLUSIONS

We have examined the extension of the DIS-MOL and more elementary DIS staffing algorithms for one $M_t/GI/s_t + GI$ many-server queue with time-varying arrivals in [30] to a feed-forward network of such queues. We have shown that these algorithms do not necessarily remain as effective in a network context. In particular, we found that there can be significant performance degradation at a queue with a high QoS target, when a preceding queue has both a non-exponential service distribution and a low QoS target; see the right-hand plot in Figure 4. This same performance degradation can occur at a single queue if the external arrival process is not nearly an NHPP, as shown in Figure 11. Otherwise (which includes the important common case of an NHPP external arrival process and high QoS targets at all queues), the extension to networks performed well.

We established the asymptotic effectiveness as the scale increases with fixed QoS targets in Theorem 6.1, but that only supports good performance at each queue with a low QoS. We relied on extensive simulation experiments to study the performance of DIS-MOL and whether the departure process from an $M_t/GI/s_t + GI$ many-server queue can be regarded as an HNPP (M_t). We saw that the interdeparture-time distribution in a stationary departure process can be nearly exponential without the departure process being approximately Poisson. For both stationary and non-stationary departure processes, we conclude that the IDC is effective in predicting whether or not a departure process can be regarded as approximately an NHPP for the purpose of this application.

We next summarize our main conclusions about the effectiveness of DIS-MOL to stabilize performance in feed-forward networks of $M_t/GI/s_t + GI$ queues, each with a sufficiently high OL and sufficiently many servers (which may not need to be so large, see Section 5.1) and an NHPP external arrival process (whenever that queue has an external arrival process). Then we discuss the implications for other MOL approximations, which apply to systems with or without customer abandonment. Finally, we discuss directions for future research.

8.1. The Main Conclusions about DIS-MOL

(i) (first good news) If the *targeted QoS is high at all queues*, then the departure process from each queue and the net (internal) arrival process to each queue should be approximately NHPP's and DIS-MOL should be effective at all queues. (See Section 3.)

(ii) (second good news) For a *totally Markovian network* model (if the external arrival processes are all NHPP's and if the service times are all exponential), the internal arrival processes should be approximately NHPP's and the DIS-MOL staffing algorithm should be effective at all queues. (See Section 5.2.) Indeed, it may suffice to have the scv of the service times be $c^2 \approx 1$. (See Section 5.3.)

(iii) (third good news) If the *targeted QoS is low at any queue*, then both DIS-MOL and the more elementary DIS algorithm (staffing at the OL $m_\alpha(t)$ itself) should be effective at that queue, even if the arrival process is not approximately an NHPP. (See Sections 3 and 6.)

(iv) (conditional good news) If the targeted QoS is high at a queue, then the DIS-MOL approximation should be effective if the arrival process is approximately an NHPP, which should occur if the IDC $I(t)$ remains near 1 for all t in the range of interest. (See Sections 3 and 4.)

(v) (bad news) For two queues in series, if (i) the second queue has a high targeted QoS, (ii) the first queue has a low targeted QoS and (iii) the first queue has a service-time distribution that is not nearly exponential, then the departure process from the first queue is likely not be approximately an NHPP and DIS-MOL is likely to be ineffective in stabilizing performance at the second queue at the specified target. (See Figures 4 and 7.) For a single queue, DIS-MOL is likely to be ineffective in stabilizing performance if the arrival process is not nearly an NHPP. (See Figure 11.)

8.2. Implications for the Standard MOL Staffing Algorithm

In this paper, we focused on network extensions of the DIS and DIS-MOL staffing algorithms from [30], but much of this paper is also applicable to the MOL approximation for stabilizing delay probabilities, with or without customer abandonment, where the QoS target involves the probability of delay, as in [8,12,18,43]. It is significant that DIS-MOL approximation is consistent with the MOL approximations. As the abandonment-probability target α decreases to 0, the OL $m_\alpha(t)$ approaches the OL $m_0(t)$ without any abandonment in Eq. (2.1). Since both the MOL and DIS-MOL tend to stabilize *all* performance measures at high QoS targets, all of our main conclusions above except for the third should apply directly to the standard MOL staffing algorithm, which applies equally well to systems without customer abandonment. At any queue where abandonment is negligible or is not of primary concern, it is natural to use MOL with a delay probability target.

8.3. Future Research

We conjecture that the difficulty in the final bad-news case in Section 8.1 can be addressed by extending DIS-MOL by using an appropriate approximation for the stationary $G/GI/s + GI$ model instead of the current stationary $M/GI/s + GI$ model, but that remaining problem is complicated by the fact that the arrival process in the stationary $G/GI/s + GI$ model should deviate from a Poisson process by the cumulative impact of many small correlations; it is more complicated than a time-transformed renewal process, as shown in Section 4. We conjecture that methods such as in [13,37,39] and Section 9.8 of [41] can be brought to bear; that remains a topic for future research.

Acknowledgment

The first author gratefully acknowledges support from NSF grant CMMI 1362310. The second author gratefully acknowledges support from NSF grants CMMI 1066372 and 1265070.

References

1. Aksin, O.Z., Armony, M. & Mehrotra, V. (2007). The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management* 16: 665–688.
2. Albin, S.L. (1982). On Poisson approximations for superposition arrival processes in queues. *Management Science* 28(2): 126–137.
3. Albin, S.L. (1984). Approximating a point process by a renewal process, II: superposition arrival processes to queues. *Operations Research* 32(5): 1133–1162.

4. Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y. & Yom-Tov, G. (2011). Patient flow in hospitals: a data-based queueing-science perspective. New York University, http://www.stern.nyu.edu/om/faculty/armony/.
5. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. & Zhao, L. (2005). Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of American Statistics Association* 100: 36–50.
6. Chevalier, P. & Tabordon, N. (2003). Overflow analysis and cross-trained servers. *International Journal of Production Economics* 86(1): 47–60.
7. Cox, D.R. & Lewis, J.A.W. (1966). *The Statistical Analysis of Series of Events*. London: Methuen.
8. Defraeye, M. & van Nieuwenhuyse, I. (2013). Controlling excessive waiting times in small service systems with time-varying demand: an extension of the ISA algorithm. *Decision Support Systems* 54(4): 1558–1567.
9. Durbin, J. (1961). Some methods for constructing exact tests. *Biometrika* 48(1): 41–55.
10. Eick, S.G., Massey, W.A. & Whitt, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* 39: 241–252.
11. Eick, S.G., Massey, W.A. & Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research* 41: 731–742.
12. Feldman, Z., Mandelbaum, A., Massey, W.A. & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science* 54(2): 324–338.
13. Fendick, K.W. & Whitt, W. (1989). Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE* 77(1): 171–194.
14. Garnett, O., Mandelbaum, A. & Reiman, M.I. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management* 4(3): 208–227.
15. Gebhardt, I. & Nelson, B.L. (2009). Transforming renewal processes for simulation of non-stationary arrival processes. *INFORMS Journal on Computing* 21: 630–640.
16. Green, L.V., Kolesar, P.J. & Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* 16: 13–29.
17. Ingolfsson, I., Akhmetshina, E., Li, Y. & Wu, X. (1979). A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. *INFORMS Journal on Computing* 19(2): 201–214.
18. Jennings, O.B., Mandelbaum, A., Massey, W.A. & Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science* 42: 1383–1394.
19. Kim, S.-H. & Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Operations Management*, forthcoming.
20. Kim, S.-H. & Whitt, W. (2014). Choosing arrival process models for service systems: tests of a nonhomogeneous Poisson process. *Naval Research Logistics* 61(1): 66–90.
21. Koizumi, N., Kuno, E. & Smith, T.E. (2005). Modeling patient flows using a queueing network with blocking. *Health Care Management Science* 8: 49–60.
22. Korporaal, R., Ridder, A., Kloprogge, P. & Dekker, R. (2000). An analytical model for capacity planning of prisons in the Netherlands. *The Journal of the Operational Research Society* 51(11): 1228–1237.
23. Larson, R.C., Cahn, M.F. & Shell, M.C. (1990). The New York city arrest-to-arraignment system. *Interfaces* 23(1): 76–96.
24. Lewis, P.A.W. (1965). Some results on tests for Poisson processes. *Biometrika* 52(1): 67–77.
25. Li, A. & Whitt, W. (2014). Approximate blocking probabilities for loss models with independence and distribution assumptions relaxed. *Performance Evaluation*.
26. Liu, Y., Gorton, I. & Zhu, L. (2007). Performance prediction of service-oriented applications on an enterprise service bus. *Thirty-first Computer Software and Applications Conference (COMPSAC 2007)* 36: 1–8.
27. Liu, Y. & Whitt, W. (2011). A network of time-varying many-server fluid queues with customer abandonment. *Operations Research* 59: 835–846.
28. Liu, Y. & Whitt, W. (2012). The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71: 405–444.
29. Liu, Y. & Whitt, W. (2012). A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Operations Research Letters* 40: 307–312.
30. Liu, Y. & Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research* 60: 1551–1564.
31. Liu, Y. & Whitt, W. (2014). Algorithms for time-varying networks of many-server fluid queues. *Informs Journal on Computing* 26: 59–73.

32. Liu, Y. & Whitt, W. (2014). Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability* 24: 378–421.
33. Mandelbaum, A., Massey, W.A. & Reiman (1998). Strong approximations for Markovian service networks. *Queueing Systems* 30: 149–201.
34. Massey, W.A. & Pender, J. (2013). Gasussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* 75: 243–277.
35. Massey, W.A. & Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13(1): 183–250.
36. Massey, W.A. & Whitt, W. (1994). Unstable asymptotics for nonstationary queues. *Mathematics of Operations Research* 19(2): 267–291.
37. Sriram, K. & Whitt, W. (1986). Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications* SAC-4(6): 833–846.
38. Stolletz, R. (2008). Approximation of the nonstationary $M(t)/M(t)/c(t)$ queue using stationary models: the stationary backlog-carryover approach. *Biometrika* 190(2): 478–493.
39. Whitt, W. (1982). Approximating a point process by a renewal process: two basic methods. *Operations Research* 30: 125–147.
40. Whitt, W. (1984). Departures from a queue with many busy servers. *Mathematics of Operations Res* 9(4): 534–544.
41. Whitt, W. (2002). *Stochastic-Process Limits*. New York: Springer.
42. Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science* 51: 221–235.
43. Yom-Tov, G. & Mandelbaum, A. (2010). The Erlang-R queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. Working paper, the Technion.