Queueing Models and
Service Management
Vol. 1, No. 2, page 079-164, 2018

QMSM
© PU 2018

Time-Varying Queues

Ward Whitt

Industrial Engineering and Operations Research
Columbia University, New York, NY 10027, USA

(Received July 2017; accepted May 2018)

Abstract: Service systems abound with queues, but the most natural direct models are often time-varying queues, which may require nonstandard analysis methods beyond stochastic textbooks. This paper provides an overview of time-varying queues. Most of the recent literature concerns many-server queues, which arise in large-scale service systems, such as in customer contact centers and hospital emergency departments, but there also has been some new work on single-server queues with time-varying arrivals, which arise in some settings, such as airplanes coming to land at an airport, cars coming to a traffic intersection and medical staff waiting for the availability of special operating rooms in a hospital. The understanding of many-server queues and single-server queues is enhanced by heavy-traffic limits, which have been extended to time-varying models as well as stationary models.

Keywords: Heavy traffic, nonhomogeneous Poisson processes, nonstationary queues, staffing to stabilize performance, time-varying arrival rates, time-varying queueing models.

Contents

1. Introduction......81

1.1 The Old Way: Applying Stationary Models in a Nonstationary Way......81

1.2 A New Perspective......82

1.2.1 New Notions of Traffic Intensity......82
1.2.2 Beyond Exponential Distributions and Markov Processes......83
1.2.3 Many Servers Versus Few Servers......84
1.2.4 Relevant Time Scales and Periodic Models......84

1.3 Organization......85

1.4 Related Literature......86

2. Numerical Algorithms for Time-Varying Markov Chains......87

2.1 Discrete-Time Markov Chains......87

2.2 Continuous-Time Markov Chains......88

2.3 Functional Kolmogorov Equations......90

Corresponding author
Email : ww2040@columbia.edu

2.3.1 Closure Approximations 91
2.3.2 Gaussian Closure Approximations 92
2.4 Numerical Algorithms for ODE's..94
2.5 Piecewise-Constant Arrival-Rate Functions 95

3. Time-Varying Deterministic Fluid Models...95
3.1 Fluid Models for TV Markovian Queues...96
3.2 Single-Server Fluid Models 96
3.3 Two-Parameter Fluid Models for Non-Markov Many-Server Queues...................97

4. Time-Varying Infinite-Server Queues 102
4.1 The Poisson Random Measure Representation...102
4.2 Direct IS Approximations with Application to Staffing......................................103
4.3 The Modified-Offered-Load (MOL) Approximation 105
4.4 Different Systems and Performance Measures...107
4.5 Little's Law 112

5. Arrival Process Models...114
5.1 Over-Dispersion and Under-Dispersion 115
5.2 Testing the NHPP Hypothesis 117
5.3 A Composition Construction for Non-Poisson Arrival Processes.........................119

6. Many-Server Heavy-Traffic Limits 122
6.1 The Time-Varying QED MSHT Regime...125
6.2 Scheduling for Multiple Classes in a Time-Varying Environment 127
6.3 The QED^c MSHT Regime 129

7. The Time-Varying Single-Server Queue 130
7.1 The Extended Lemoine Representation of the Workload.....................................130
7.2 Heavy-Traffic Limits for Periodic Single-Server Queues 132
7.3 A Rare-Event Simulation Algorithm for the $GI_t / GI / 1$ Queue...........................135
7.4 Time-Varying Robust Queueing...140
7.5 Service-Rate Controls to Stabilize Performance..145

8. Conclusions……………………………………………………………........................148

1. Introduction

This paper provides an overview of *time-varying* (TV) queueing models. We primarily consider the standard multi-server model with unlimited waiting space and the first-come first-served (FCFS) service discipline, where the servers work independently in parallel and the arrival process has a time-varying rate. We also consider other time-varying model elements, such as the number of servers or the individual service rate, often chosen to stabilize performance over time, as occurs when we want to set staffing levels in a service system to provide nearly constant quality of service at all times of the day. We not only review ways to analyze these models, but we also review important insights about the impact of time-varying arrival rates upon performance.

1.1. The old way: Applying stationary models in a nonstationary way

There is a startling disconnect between stochastic textbooks and many service systems. The stochastic textbooks discuss stationary models and ways to compute their steady-state distribution. In contrast, many service systems have strongly time-varying demand. Nevertheless, there are well developed engineering methods to apply stationary stochastic models to achieve good performance.

First, when there is flexibility in the service capacity to meet the demand, as achieved by TV staffing in call centers, it is often possible to apply stationary models in a nonstationary way in order to achieve good performance. If the arrival rate $\lambda(t)$ changes sufficiently slowly, then we may use a *pointwise-stationary approximation* (PSA), i.e., to use a stationary model in a nonstationary way. In particular, for each time t, we would approximate the performance at time t by the steady-state performance of the stationary model with constant arrival rate equal to $\lambda(t)$; see Green and Kolesar [63] and Whitt [213]. Variants of the PSA have been effective for analyzing the performance of many service systems, as discussed in Green *et al.* [64].

Second, when there is no flexibility in the service capacity to meet the demand, as in the classical telephone networks, and we wish to provide high Quality-of-Service (QoS) at all times, then we may apply stationary models to an appropriate worst case. This is the classic “busy-hour engineering” in telephony. A stationary model then might be applied with the arrival rate equal to the average over the busiest hour of the day. If the capacity indeed cannot be adjusted easily on a short-term basis, then allowance may be needed for growth over a longer time period.

Third, when there is no flexibility in the service capacity to meet the demand, but we do not require high QoS at all times, then we may ignore the performance over short time periods. Then we can ignore local fluctuations and use stationary models with the long-run-average arrival rate.

TV models become important when we want to go beyond one of these relatively simple approaches. We may want to achieve high QoS at all times, but the arrival rate may not change slowly enough. Indeed, we want to decide whether or not the arrival rate changes rapidly or not.

1.2. A new perspective

Moving from stationary models to TV models leads us to focus on time and ask new questions: When is the congestion greatest? How does the congestion build up and decline? How can we effectively control the congestion?

1.2.1. New notions of traffic intensity

For stationary queueing models, the standard approach is to do steady-state analysis. We start by seeing if the model is stable, i.e., if there exists a proper steady state. To do so, we look at the *traffic intensity* ρ, the long-run arrival rate divided by the long-run maximum possible service rate (both assumed to be exogenously specified). We check that $\rho < 1$ and then investigate the long-run steady-state behavior. If $\rho > 1$ ($\rho = 1$), the model is overloaded (critically loaded) and the congestion grows without bound, which can be described in more detail by heavy-traffic limits, e.g., as in Chapters 5 and 8-10 in Whitt [214]. Most attention is given to actually determining the steady-state behavior, assuming that $\rho < 1$, for which many excellent approaches have been developed; e.g., Asmussen [9].

For TV models, that routine starting point changes drastically. We need to reconsider the notions of traffic intensity. Suppose that we consider the Markovian $M_t / M / s$ model with a fixed number of servers, each with a fixed service rate μ. The TV behavior can be understood by looking at the cumulative rates over subintervals $[a,b]$ of $[t_0,\infty)$, defined as

$$\Lambda(a,b) \equiv \int_a^b \lambda(t)dt \quad \text{and} \quad M(a,b) \equiv \int_a^b s(t)\mu(t)dt = s\mu(b-a). \qquad (1.1)$$

($M(a,b)$) represents the service capacity, i.e., the maximum possible total service rate over $[a,b]$.) At time t, we can define a *TV traffic intensity* by

$$\rho^*(t) \equiv \sup_{t_0 \le s \le t} \{\Lambda(s,t) / M(s,t)\}, \qquad (1.2)$$

see Figure 1 of Newell [160] and Theorem 1 of Massey [142].

The following example shows that the TV traffic intensity $\rho^*(t)$ in (1.2) is often more useful than the *instantaneous traffic intensity* $\rho(t) \equiv \lambda(t) / s(t)\mu(t)$.

Example 1.1. (*congestion build up and decline*) To see that the TV traffic intensity provides important new information, it helps to consider a deterministic fluid model, where the rates describe the flows of continuous divisible fluid. For a stochastic model, when the rates are very large, then the stochastic behavior will be similar to the fluid model, by law-of-large-numbers asymptotics.

A relatively simple case is a single-server queue with a TV arrival rate $\lambda(t)$ and a constant service rate μ. If we assume for the fluid model that $\lambda(t)$ smoothly increases from 0 at time t_0, reaching and exceeding μ, and then decreases back down to a level below μ, then we have several important times: (i) t_1, the first time t at which $\lambda(t)=\mu$ (hitting from below), (ii) t_2 the time at which $\lambda(t)$ reaches its maximum value, (iii) t_3, the second time t at which $\lambda(t)=\mu$ (hitting from above), after which it remains below, and (iv) t_4, the time that the queue is first empty.

For the the instantaneous traffic intensity we see that $\rho(t_1)=\rho(t_3)=1$, $\rho(t)>1$ for $t_1<t<t_3$ and $\rho(t)<1$ otherwise. In contrast, for the TV traffic intensity, we see that $\rho^*(t_1)=\rho^*(t_4)=1$, $\rho^*(t)>1$ for $t_1<t<t_4$ and $\rho^*(t)<1$ otherwise. Fluid first waits at time t_1 and the queue first empties at time t_4 and remains empty thereafter.

The time lag. Example 1.1 shows that for TV queues there is a time lag in the impact of congestion after the arrival rate reaches its peak. For Example 1.1, the peak arrival rate occurs at time t_2, while the maximum queue length occurs at time t_3 and the queue first empties at time t_4. For the congestion at time t, we need to look at the past prior to time t, as captured by the TV traffic intensity in (1.2). For the service rate, we ask the question: “What have you done for us lately?” (Extra service capacity when there is no demand is wasted.)

1.2.2. Beyond exponential distributions and Markov processes

Because TV arrival rates are so challenging, it is natural to start looking at the consequence for Markovian stochastic models, but we also want to understand the interaction between non-Markov stochastic variability together with the deterministic variability of a TV arrival-rate function.

Thus, much of our discussion is for the general $G_t/GI/s_t+GI$ model, which has a general arrival process having a TV arrival-rate function (the G_t), a TV number of homogeneous servers (the s_t), unlimited waiting space and abandonment from queue (the $+GI$) and assume that the service times and patience times (times until abandonment after joining the queue) come from independent sequences of independent and identically distributed (i.i.d.) random variables, independent of the arrival process, with service-time cumulative distribution function (cdf) G and patience cdf F.

A good example of the joint impact of the time-varying arrival rate and a non-exponential distribution is the formula for the mean number of busy servers in an $M_t/GI/\infty$ infinite-server (IS) model with a non-homogeneous Poisson process (NHPP) as an arrival process, reviewed in Section 4; in particular,

$$m(t)\equiv E[X(t)]=\int_0^\infty \lambda(t-s)G^*(s)ds=E[\lambda(t-S_e)]E[S],\quad t\ge 0, \qquad (1.3)$$

where $\lambda(t)$ is the deterministic arrival rate at time t, G is the cdf of a service time S, and S_e is a random variable with the service-time stationary-excess cdf; see (4.2). The last expression in (1.3) shows that the mean is the same as in a stationary model ($m = \lambda E[S]$) except for a random time lag by S_e. By doing a Taylor series expansion of the TV arrival rate function $\lambda(t)$, we see that the peak in $m(t)$ tends to lag behind the peak in $\lambda(t)$ by approximately $E[S_e] = E[S^2]/2E[S] = E[S](c_s^2 + 1)/2$, which depends on the second moment of the service time as well as the mean. Equivalently, it depends on the squared coefficient of variation (scv, variance divided by the square of the mean) c_s^2 of the service time as well as the mean. The scv is convenient because it measures the variability independent of the mean. For more on the structure of $m(t)$, see Section 2 and Section 3 of Eick *et al.* [44].

1.2.3. Many servers versus few servers

We only consider a small class of the queueing models in the vast literature on queues. Among those that we do consider, the TV models with many servers have proven much easier to analyze than the TV models with few servers, because we can exploit insights drawn from the $M_t/GI/\infty$ IS model, which is remarkably tractable; e.g., see Eick *et al.* [44, 45] and Massey and Whitt [147]. Indeed, the IS model structure underlies many of the results for many-server queues.

The TV IS model is more tractable because all customers enter service immediately upon arrival, so that there is no waiting. Hence, each customer is in the system over its service time. This simplification breaks down in many server queues, but since the total service rate when many customers are in service is large, the waiting times tend to be relatively short.

In sharp contrast, with few servers, e.g., with one, both the performance and the analysis techniques tend to be very different. For the single-server queue, the waiting times are often longer than the service times. This is especially true when $\rho^*(t) > 1$ for $\rho^*(t)$ in (1.2). Most of our review covers the relatively well-developed theory for the TV many-server queue, but we also discuss recent work on the TV single-server queue in the final Section 7. But that topic is in its infancy; we can expect to learn much more in the future.

1.2.4. Relevant time scales and periodic models

Leaving steady-state and considering TV behavior makes us look more closely at time. We should consider the relevant time scale for performance, which is usually determined by the response (service plus waiting) times. For example, if all service is completed on a given day, as in a telephone call center, then it suffices to look at a representative day. Since the waiting tends to be a consequence of the specific way service is provided, it is often appropriate to consider only the mean service time as the relevant time scale.

In hospitals, patient length of stay may extend over several days. Thus, the relevant time scale tends to be longer than in many other service systems, extending over multiple days. Because the arrival rates vary strongly over each day and also differ substantially by the day of the week, it can be good to use a periodic model of the arrival rate with the week being the length of a period, as in the data-based stochastic model for an emergency department proposed in Whitt and Zhang [230].

A periodic arrival-rate function is a very special case of a TV arrival-rate function, but it is important to note that the model with a periodic arrival-rate function actually is very general, including most TV and stationary models as special cases. If the periodic arrival rate is constant, then the TV model reduces to a stationary model; any TV model over a finite interval can be regarded as a periodic model if we make the length of the period longer than the original interval.

1.3. Organization

We start in Section 2 by reviewing methods for analyzing the $M_t / M_t / s_t + M_t$ TV Markov model, where the arrival process is an NHPP, while the service rate and abandonment rate may now be TV as well. The standard tools are numerical algorithms for *ordinary differential equations* (ODE's). The ODE approach is well illustrated by the studies of airplane landing delays at airports by Koopman [105] and dispatching delays for police patrol cars by Kolesar *et al.* [104]. Especially promising for these TV Markov models are reduced systems of ODE's for summary statistics such as the TV mean obtained via closure approximations. We highlight the recent Gaussian closure approximations developed by Massey and Pender [145, 146], which involve a general framework that seems to be widely applicable.

In Section 3 we review deterministic fluid models, which can serve as useful alternatives to (or approximations for) the TV stochastic models. The deterministic analysis is a natural first-order approach when the deterministic variations tend to be more important than the uncertainty. The basic deterministic analysis is illustrated by the studies of traffic delays at tool booths by Edie [43] and letter delays at post offices by Oliver and Samuel [167]. We highlight the recent two-parameter (or measure-valued) deterministic fluid models for the $G_t / GI / s_t + GI$ model developed by Whitt [221] and Liu and Whitt [123].

In Section 4 we review the $M_t / GI / \infty$ TV infinite-server (TVIS) model and the approximations based on it. The TVIS model ultimately is the source of much that we know about TV models. After reviewing the Poisson random measure representation in Section 4.1, we review the *modified-offered-load* (MOL) approximation in Section 4.3 and its application to set staffing levels to stabilize performance at target levels in Section 4.4. The main ideas about MOL are well covered in Green *et al.* [64], but there has been more work, primarily extending to new models and non-NHPP arrival processes. We also discuss recent work on the TV Little's law ($L = \lambda W$) in Section 4.5 because Little's law is intimately related to the IS queue.

In Section 5 we discuss arrival process models. The standard TV arrival process model is the NHPP, but recent examination of service system data raises questions about the suitability of the NHPP model, usually because of over-dispersion. We highlight the studies by Kim and Whitt [98, 99] of how to test the NHPP model assumption and the associated studies to see if arrival process data are consistent with the NHPP assumption. We also highlight the composition construction of a more general arrival process that has a one-parameter quantification of the level of variability.

In Section 6 we discuss recent (MSHT) limits for TV queues and insights that can be drawn from them. After reviewing the quality-and-efficiency-driven (QED) or Halfin and Whitt [70] MSHT limiting regime, we highlight the more recent complementary-QED (QED^c) regime studied in Liu and Whitt [123, 124, 126], which underlies much of the most useful methods developed so far for TV many-server queues, including the two-parameter fluid model in Liu and Whitt [123] discussed in Section 3.3, the Gaussian approximations in Liu and Whitt [126], the truncated Gaussian approximations in Liu *et al.* [130] and the Gaussian closure approximation of Massey and Pender [145]. We also review the basic QED TV MSHT limit in Mandelbaum *et al.* [140] and Puhalskii [183]. We then highlight a new QED TV MSHT limit by Sun and Whitt [205] for scheduling of multiple classes in a TV setting and the sample-path TV MSHT Little's law that emerges from that TV MSHT limit.

Finally, in Section 7 we discuss recent studies of TV single-server queues. We highlight the new HT limit for periodic queues in Whitt [224] that involves scaling of the arrival rate function in addition to the usual HT scaling. The limit process is reflected periodic Brownian motion (RPBM), the natural periodic analog of the RBM HT limit for stationary models. We highlight a new rare-event simulation algorithm in Ma and Whitt [136] and a new TV robust queueing (TVRQ) algorithm for periodic queues in Whitt and You [227] that can be used to obtain concrete numerical results. We discuss service-rate controls to stabilize performance in TV single-server queues developed in Whitt [225] and studied further in Ma and Whitt [137]. In Section 8 we draw conclusions.

1.4. Related literature

In closing this introduction we point to other related surveys and applications. Broad surveys of the literature on TV queues have recently been provided by Defraeye and van Nieuwenhuyse [39] and Schwarz *et al.* [195]. These are much broader than the earlier surveys in Massey [143], Green *et al.* [64] and Hampshire and Massey [72], which are more directly related to this review. A good account of the remarkable early work on the Erlang models by A. K. Erlang appears in Brockmeyer *et al.* [24], while C. Palm's 1943 early work on time-varying queues appears in Palm [170].

There are broad surveys of the literature on call centers in Gans *et al.* [54] and Aksin *et al.* [5], while Brown *et al.* [25] is the seminal contribution on data analysis for call centers. For hospitals, there is the data analysis by Armony *et al.* [8] and the recent papers by Kim and Whitt [99], Yom-Tov and Mandelbaum [234], Kim *et al.* [95], Shi *et al.* [199], Kim *et al.* [100], Dai and Shi [36], and Whitt and Zhang [230] from which the early literature can be traced.

Time-varying queues also play an important role in communication networks, e.g., see Leung *et al.* [111], Neely *et al.* [155], and Shakkotai *et al.* [198], and in road traffic, e.g., see Newell [159], Ran and Boyce [185], Daganzo [34], and Kurzhanskiy and Varaiya [107].

2. Numerical Algorithms for Time-Varying Markov Chains

The TV behavior of a TV model can be very different from the TV behavior of a stationary model, which is often called the transient behavior, because it describes the dissipating transient impact of initial conditions before the system reaches steady state, where the deterministic performance measures such as the mean queue length do not change over time. Nevertheless, the basic mathematical representations that are the basis for computing the TV behavior of the two kinds of models are essentially the same for Markov chains. To make that clear, we first review the basic theory for *discrete-time Markov chains* (DTMC's) in Section 2.1 and then we review the basic theory for *continuous-time Markov chains* (CTMC's) in Section 2.2. For CTMC's, the TV behavior is characterized by a system of *ordinary differential equations* (ODE's) called the Kolmogorov equations. In Section 2.3 we then review functional Kolmogorov equations for CTMC's, which are greatly simplified systems of ODE's for summary statistics such as the TV moments. In order to get bonafide ODE's for these summary statistics, we need to approximate other TV quantities on the right-hand side of the ODE's. We review early closure approximations for the first few moments of TV queues in Section 2.3.1 and the recent highly successful Gaussian closure approximations developed by Massey and Pender [145] for the TV many-server $M_t / M_t / s_t + M_t$ model in Section 2.3.2.

2.1. Discrete-time Markov chains

A stationary *m*-state DTMC is usually specified by its $m \times m$ transition matrix $P,$ with $P_{i,j}$ representing the probability of making a one-step transition from state i to state j at any time. The *n*-step transition probabilities are then given by the Chapman-Kolmogorov equations

$$P_{i,j}^{(n)} = \sum_{k=1}^{m} P_{i,k}^{(p)} P_{k,j}^{(n-p)} \text{ for each } p,\ 1 \le p \le n-1, \tag{2.1}$$

which is equivalent to the simple matrix product $P^{(n)} = P^n \equiv P \times \cdots \times P$.

The main theoretical result for DTMC's (that are aperiodic irreducible, i.e., for which it is possible to get from each state to any other state in some finite number of transitions) is about the steady state: $P_{i,j}^{(n)}$ converges to a limit π_j as $n \to \infty$, which is independent of the initial state i. The steady-state probability vector π can be calculated as the unique solution to the matrix equation $\pi = \pi P$, which says that π is the unique probability vector that remains unchanged by a single transition under P; i.e., if the initial distribution is π, then the DTMC is a stationary stochastic process.

A TV (inhomogeneous) m-state DTMC is specified by a sequence of transition matrices $\{P(k) : k \ge 1\}$, where $P(k)$ is the $m \times m$ one-step transition probability matrix at discrete time k, with $P_{i,j}(k)$ representing the probability of making a one-step transition from state i at time k to state j at time $k+1$. Paralleling (2.1), the probability of making an n-step transition from state i at time r to state j time $r+n$ is then

$$P_{i,j}^{(n)}(r) = \sum_{k=1}^{m} P_{i,k}^{(p)}(r) P_{k,j}^{(n-p)}(r+p) \text{ for each } p,\ 1 \le p \le n-1, \tag{2.2}$$

or by the matrix product $P^{(n)}(r) \equiv P(r) \times P(r+1) \cdots \times P(r+n-1)$. Obviously, the nice description of the steady-state behavior is lost in the TV setting, but we can still speak of "asymptotic loss of memory" or "weak ergodicity;" see Seneta [196, 197] and Liu and Whitt [122] plus references in and citations to these sources.

If we compute by recursive matrix multiplication, then there is no difference in computing with (2.2) in the TV setting from (2.1) in the stationary setting, but some advanced methods, such as spectral decomposition methods (exploiting eigenvalues), gain computational advantage from stationary representation and additional mathematical structure, e.g., see Latouche and Ramaswami [108] and Stewart [201].

2.2. Continuous-time Markov chains

The story for CTMC's is similar; indeed, it is natural to think of a CTMC as being a DTMC with a very short time between its transitions (with appropriately adjusted transition probabilities), which is exactly how some numerical methods for CTMC's proceed; see Section 2.4. Directly, stationary CTMC's are usually specified by their rate (or infinitesimal generator) matrices Q, with $Q_{i,j}$ giving the rate of a transition from state i to state j at time t; i.e., if $P_{i,j}(t, t+h)$ is the probability of a transition from state i at time t to a state j at time $t+h$, then we assume that

$$P_{i,j}(t+h)-P_{i,j}(t)=Q_{i,j}h+o(h) \quad \text{as} \quad h\downarrow 0 \quad \text{for} \quad i\neq j,$$

where $f(h)$ is $o(h)$ if $f(h)/h\to 0$ as $h\downarrow 0$; e.g., see Chapter 5 of Ross [193]. (We also assume that the rate of leaving state i is $1-P_{i,i}(t)=\nu_i h+o(h)$ as h $\downarrow 0$, where $\nu_i \equiv -Q_{i,i} \equiv \sum_{j:j\neq i} Q_{i,j}$.)

In other words, we regard $Q_{i,j}$ as the derivative of $P_{i,j}(t)$. The transition matrices over positive time intervals t are then characterized by the solution of a a system of ordinary differential equations (ODE's): either the forward Kolmogorov equation $\dot{P}(t)=P(t)Q$ (looking forward to the incremental change over the interval $(t,t+h)$, starting from $P(t)$ at time t) or the backward Kolmogorov equation $\dot{P}(t)=QP(t)$ (looking backward to the incremental change over the interval $(0,h)$ followed by $P(t)$ to get to time $t+h$). For stationary infinite-state CTMC's, the backward Kolmogorov equations are often preferred for the theory; e.g., see Section 5.4 of Ross [193] or Chung [30].

For TV (inhomogeneous) CTMC's, we assume that the rate matrix depends on t, so that $Q_{i,j}(t)$ represents the rate of a transition from state i at time t to state j, while $P_{i,j}(t,t+h)$ is the probability of a transition from state i at time t to a state j at time $t+h$. Just as the Chapman-Kolmogorov equations immediately extend to TV DTMC's, so do the Kolmogorov ODE's immediately extend to TV CTMC's, although the regularity conditions in the theory gets more complicated. For TV CTMC's, it becomes important to work with the forward equations in order to incorporate the TV rates in a TV manner. For some theory formulating TV CTMC's and uniform acceleration approximations for them in terms of time-ordered exponentials, see Section 3 of Massey and Whitt [152].

For most TV Markov queueing models (of a single queue), the number of customers in the system at time t, which we denote by $X(t)$, can be represented as a TV birth-and-death (BD) process on the nonnegative integers; i.e., in state k at time t, arrivals occur according to a birth rate $\lambda_k(t)$ for $k\geq 0$, while departures occur according to a death rate of $\mu_k(t)$ for $k\geq 1$. Let $p_k(t)\equiv P(X(t)=k)$ denote the probability of being in state k at time t and let $\dot{p}_k(t)$ be its derivative with respect to time. For a TV BD process, the (forward) Kolmogorov equations are the system of ODE's

$$\dot{p}_k(t)=-(\lambda_k(t)+\mu_k(t))p_k(t)+\lambda_{k-1}(t)p_{k-1}(t)+\mu_{k+1}(t)p_{k+1}(t), \quad k\geq 1, \quad \text{and}$$
$$\dot{p}_0(t)=-\lambda_0(t)p_0(t)+\mu_1(t)p_1(t). \tag{2.3}$$

Letting $p(t)\equiv(p_0(t),p_1(t),\ldots)$ be the vector of TV state probabilities, the system of ODE's in (2.3) can be represented as a single linear ODE for the vector $p(t)$; ie.,

$$\dot{p}(t)=p(t)Q(t), \tag{2.4}$$

where $Q(t)$ is the TV rate matrix, with $Q_{k,k}(t)\equiv-(\lambda_k(t)+\mu_k(t))$, $Q_{k-1,k}(t)\equiv\lambda_{k-1}(t)$ and $Q_{k+1,k}(t)\equiv\mu_{k+1}(t)$ for $k\geq 1$ and $Q_{0,0}(t)\equiv-Q_{0,1}(t)\equiv-\lambda_0(t)$.

Paralleling the discrete-time setting, if we computed by directly solving the ODE (2.4), then the TV setting is essentially the same as the stationary setting, but some advanced methods, such as directly computing the transition matrix as a matrix exponential ($P(t) = e^{Qt}$), gain computational advantage from stationary representation, despite the need for caution; see Moler and van Loan [154], Hairer *et al.* [69], and Press *et al.* [182].

2.3. Functional Kolmogorov equations

It is natural to look for more elementary ODE's for summary statistics such as the TV mean. For that purpose, we can apply functional versions of the Kolmogorov equation. Let f be a real-valued function of the state. Then we define

$$m_f(t) \equiv E[f(X(t))] = p(t)f \equiv \sum_k p_k(t)f(k) \tag{2.5}$$

and examine the resulting ODE for $\dot{m}_f(t)$. If we let $f(k) = k^p$, $k \geq 0$, then $m_f(t) = E[X(t)^p]$.

The functional Kolmogorov equations for the first few moments take a relatively tractable form for highly structured queueing models such as the $M_t / M_t / s_t + M_t$ model, which has a nonhomogeneous Poisson process as its arrival process with arrival rate $\lambda(t)$, service provided by each busy server at rate $\mu(t)$, $s(t)$ servers and customer abandonment from queue where each waiting customer has an abandonment rate $\theta(t)$. (If the service is M instead of M_t, then the service times are mutually independent exponential random variables, but not more generally; and similarly for the times to abandon.)

For the $M_t / M_t / s_t + M_t$ model, where the number of servers $s(t)$ is TV, we need to specify what happens when all the servers are busy when the staffing is scheduled to decrease. In practice this can be complicated because of specified shifts for the servers; e.g., see Ingolfsson [85]. For simplicity, in this paper we assume that a customer is pushed back to the head of the queue, after which it receives a full new service. Moreover, we do not pay attention to the identity of individual servers and customers.

When $X(t) = k$ at time t, the arrival rate is $\lambda(t)$ and the number of servers is $s(t)$, independent of k, while the total service rate is $(k \wedge s(t))\mu(t)$ and the total abandonment rate is $(k - s(t))^+\theta(t)$ for $x \wedge k \equiv \min\{x, k\}$ and $(x)^+ \equiv \max\{x, 0\}$. Thus, for the $M_t / M_t / s_t + M_t$ model,

$$\begin{aligned}\dot{m}_f(t) = \lambda(t)E[f(X(t)+1) - f(X(t))] + \mu(t)E[(X(t) \wedge s(t))(f(X(t)-1) - f(X(t)))] \\ + \theta(t)E[(X(t) - s(t))^+(f(X(t)-1) - f(X(t)))].\end{aligned} \tag{2.6}$$

For the TV mean $m(t)$, we have $m_f(t)$ in (2.5) for $f(k) = k$, $k \geq 0$, so that

$$\dot{m}(t) = \lambda(t) - \mu(t)E[(X(t) \wedge s(t))] - \theta(t)E[(X(t) - s(t))^+]. \tag{2.7}$$

By combining (2.6) for the first two moments, we get the corresponding ODE for the TV variance, denoted by $v(t)$,

$$\begin{aligned}\dot{v}(t) &= \lambda(t) + \mu(t)E[X(t)\wedge s(t)] + \theta(t)E[(X(t)-s(t))^{+}] \\ &\quad -2(\mu(t)Cov(X(t), X(t)\wedge s(t)) + \theta(t)Cov(X(t), (X(t)-s(t))^{+})). \qquad (2.8)\end{aligned}$$

For special cases of the $M_t / M_t / s_t + M_t$ model, we obtain more concrete results, as we illustrate now.

Example 2.1. (*the* $M_t / M_t / \infty$ *IS Model*) For the $M_t / M_t / \infty$ TVIS model, (2.7) and (2.8) simplify greatly, becoming

$$\dot{m}(t) = \lambda(t) - \mu(t)m(t) \qquad (2.9)$$

and

$$\dot{v}(t) = \lambda(t) + \mu(t)m(t) - 2\mu(t)v(t). \qquad (2.10)$$

By itself, equation (2.9) is an ODE that can be solved for $m(t)$. As we will discuss in Section 4, for the $M_t / M_t / \infty$ model starting empty, $X(t)$ has a Poisson distribution for each t, which implies that $v(t) = m(t)$ for all t. Indeed, we see that ODE (2.10) coincides with ODE (2.9) when $v(t) = m(t)$. Formulas (2.9) and (2.10) are consistent with Theorems 1 and 6 and Corollary 4 in Eick *et al.* [44].

Variations of the simple ODE representations above for the TVIS model extend to Markovian networks of TVIS queues, possibly with phase-type (Ph) distributions or MAP arrival processes; see Section 8 of Massey and Whitt [147], Nelson and Taaffe [157, 158], and Gebhardt *et al.* [57].

Example 2.2. (*the* $M_t / M_t / 1$ *Single-Server Model*) For the $M_t / M_t / 1$ TV single-server model, (2.7) and (2.8) also simplify, becoming

$$\dot{m}(t) = \lambda(t) - \mu(t)(1 - p_0(t)), \qquad (2.11)$$

and

$$\dot{v}(t) = \lambda(t) + \mu(t) - \mu(t)p_0(t)(2m(t)+1), \qquad (2.12)$$

where $p_0(t) \equiv P(X(t)=0)$. Equations (2.11) and (2.12) were first derived and applied by Clarke [32].

2.3.1. Closure approximations

It is evident that equation (2.11) can be approximately *closed* and converted to a single ODE if we approximated $p_0(t)$ by a function of $m(t)$ (and possibly $\lambda(t)$ and $\mu(t)$). For the $M_t / M / 1$ queue, such a closure approximation was suggested by Rider [189]. Rothkopf and Oren [194] later showed that a more effective closure approximation could be obtained from both equations (2.11) and (2.12) . Then we can obtain a system to two ODE's by developing approximations for $p_0(t)$ in terms of $m(t)$ and $v(t)$. They approximated

$p_0(t)$ by fitting the two-parameter negative-binomial distribution to the mean $m(t)$ and variance $v(t)$.

Rothkopf and Oren [194], and then Clark [31] went further and developed a closure approximation for mean and variance in the $M_t / M_t / s$ model, exploiting the fact that the ODE's for $m(t)$ and $v(t)$ can be expressed directly in terms of the s unknown time-varying probabilities $p_0(t)$, $p_1(t), \ldots, p_{s-1}(t)$ as well as $m(t)$ and $v(t)$.

Closure approximations for more complex time-varying Markov models involving time-varying phase-type distributions were developed by Taaffe and Ong [207], Taaffe and Clark [206], Ong and Taaffe [168, 169], and Grier *et al.* [65]. The paper Grier *et al.* [65] is interesting because it applies a closure approximation to reduce a TV two-queue network to two stochastically independent TV queues, providing a TV analog of the reduced-load or Erlang fixed-point approximation in Whitt [211] and Kelly [94].

2.3.2. Gaussian closure approximations

As will be discussed in Section 4, the number in system in the $M_t / M_t / \infty$ TVIS model, starting out empty, is Poisson for all t, and is thus approximately Gaussian for all t provided that the mean is not too small. Thus, it is natural to consider Gaussian closure approximations for TV $M_t / M_t / s_t + M_t$ many-server queues, as proposed by Massey and Pender [145]. It is also significant that this Gaussian closure approximation is supported by the MSHT limit in the complementary-QED (QED^c) MSHT regime, as we will discuss later in Section 6.3.

The *Gaussian variance approximation* (GVA) from Massey and Pender [145] is based on the approximation

$$X(t) \approx m(t) + N(0,1)\sqrt{v(t)} \quad \text{for all} \quad t, \tag{2.13}$$

where $N(0,1)$ is a standard (mean 0, variance 1) Normal (Gaussian) random variable, while $m(t)$ and $v(t)$ are the TV mean and variance satisfying (2.7) and (2.8). The Gaussian approximation is convenient because simple scaling properties of Gaussian distributions can be exploited to calculate the right sides of (2.7) and (2.8) given approximation (2.13) ; i.e., (2.7) and (2.8) can be re-expressed approximately as

$$\dot{m}(t) = \lambda(t) - \mu(t)m(t) - \mu(t)E[N(0,1) \wedge \eta(t))] + \theta(t)E[(N(0,1) - \eta(t))^+] \tag{2.14}$$

and

$$\dot{v}(t) = 2\lambda(t) - \dot{m}(t) - 2\zeta(t)v(t), \tag{2.15}$$

where

$$\eta(t) \equiv [s(t) - m(t)] / \sqrt{v(t)} \quad \text{and}$$

$$\zeta(t) \equiv \mu(t)Cov(N(0,1), N(0,1) \wedge \eta(t)) + \theta(t)Cov(N(0,1), (N(0,1) - \eta(t))^+. \tag{2.16}$$

Massey and Pender [145] exploit Stein's lemma, which states that, if (and only if)

$X \overset{\mathrm{d}}{=} N(0,1)$, then

$$E[Xf(X)] = E[df(X)/dX] \tag{2.17}$$

for all generalized functions f (which includes indicator functions needed in the present setting); see Stein [200]. With the aid of (2.17), (2.14) and (2.15) can be expressed approximately as

$$\dot{m}(t) = \lambda(t) - \mu(t)m(t) - (\mu(t) - \theta(t))(\eta(t)\bar{\Phi}(\eta(t)) - \phi(\eta(t)))\sqrt{v(t)} \tag{2.18}$$

and

$$\dot{v}(t) = 2\lambda(t) - \dot{m}(t) - 2(\mu(t)\Phi(\eta(t)) + \theta(t)\bar{\Phi}(\eta(t)))v(t), \tag{2.19}$$

where $\phi(x)$ is the *probability density function* (pdf) of $N(0,1)$, $\Phi(x) \equiv P(N(0,1) \le x)$ is the associated *cumulative distribution function* (cdf) and $\bar{\Phi}(x) \equiv 1 - \Phi(x)$ is the *complementary cdf* (ccdf). This GVA dynamical system in (2.18) and (2.19) coincides with the approximation developed earlier by Ko and Gautam [103].

The GVA approximation above seems adequate to yield effective numerical approximations for many $M_t / M_t / s_t + M_t$ models in practice, but Massey and Pender [145] in Section 4 go much further and develop a more accurate three-ODE system for the first three TV moments and a systematic framework for developing closure approximations much more generally. This extension can be obtained without much extra work.

The three-ODE refinement is called the *Gaussian skewness approximation* (GSA) because the third moment captures the skewness of non-Gaussian distributions, i.e., $skew(Z) \equiv E[(Z - E[Z])^3]/Var(Z)^{3/2}$. Specifically, the three-ODE system is for the TV mean and the TV second and third central moments. The systematic procedure yields a closure approximation that is a quadratic function of a Gaussian random variable; i.e., instead of (2.13), in their (4.2) the approximation at each time becomes

$$X(t) \approx m(t) + \left(N(0,1)\cos(\xi(t)) + \frac{N(0,1)^2 - 1}{\sqrt{2}}\sin(\xi(t)) \right)\sqrt{v(t)}, \quad \text{for all } t. \tag{2.20}$$

With (2.20), the solution of the three-ODE system directly yields the TV mean, variance and skewness. To apply this scheme, we need not calculate $\xi(t)$, but they give an explicit expression for it.

The general framework for closure approximations for $M_t / M_t / s_t + M_t$ models developed by Massey and Pender [145] represents the distribution at each time t as a polynomial function of a Gaussian distribution. The framework exploits the Hilbert space of Hermite orthogonal polynomials $h_n(x)$ and involves a Hermite polynomial generalization of Stein's lemma, stating that

$$E[f(X)h_n(X)] = E[d^n f(X)/dX^n] \tag{2.21}$$

for any generalized function f. Not only does this paper develop effective closure approximations $M_t / M_t / s_t + M_t$ models, but it develops a systematic approach that has promise for other settings. Indeed, extensions to TV loss models and TV Jackson networks of queues with abandonment have since been treated by Pender [176] and Pender and Massey [178], respectively. Analyses using different orthogonal polynomials have been done in Pender [174, 175].

These Gaussian closure methods require closed-form expressions for the expectations that appear in the functional Kolmogorov forward equations. Pender [177] presents a new sampling algorithm to use with simulation when closed-form expressions are unavailable.

2.4. Numerical algorithms for ODE's

To numerically solve the ODE's discussed above, for the most part it suffices to apply off-the-shelf methods, such as the basic Euler method or the Runge-Kutta fourth-order method. We briefly discuss remaining numerical issues in this section.

Truncating the state space. For infinite-state models, the system of ODE's is infinite, but computation can be done by truncating the state space, as was done to analyze airplane landing delays at airports by Koopman [105] and dispatching delays for police patrol cars by Kolesar *et al.* [104]. Standard ODE algorithms with truncated state space have been used extensively in studies of time-varying queues, e.g., in many of the papers surveyed in Green *et al.* [64]. We do remark that this step can require some care, because the behavior at the truncated boundary can matter. A reasonable approach is to achieve the truncation by approximating by a well-defined finite-state queueing model that can be related to the original model; e.g., approximating by a model with a finite waiting room. Then the impact of the truncation can be understood and managed.

A uniformization algorithm for solving the system of ODE's. Davis *et al.* [38] found that it was effective and convenient to uniformize the TV CTMC to create a TV DTMC over an evenly spaced discrete time grid. Assuming that $K \geq A_{i,i}(t)$ for all i and t, a DTMC for a step size $h < 1/K$ with time-varying transition probabilities can be defined by

$$P_{i,j}(k) = hA_{i,j}(kh), \quad j \neq i \quad \text{and} \quad P_{i,i}(k) = 1 - \sum_{j, j \neq i} P_{i,j}(k). \qquad (2.22)$$

Then we approximate the CTMC at time kh by the DTMC at time k.

Verifying numerical accuracy. It is always important to use simple practical methods to ensure numerical accuracy. For example, at various times t over a target interval $[0, T]$, we can check that the resulting probability vector $p(t)$ is indeed nonnegative with total mass equal to 1. We can correct by either (i) adding missing mass to the highest state or subtracting excess mass from 0 or (ii) subtracting excess mass from the highest state or adding missing mass to 0. Since these two approaches tend to be upper and lower bounds, we can see that the impact of these corrections is negligible by doing both. Similarly, to verify numerical accuracy, it is good to consider several truncation levels and several step sizes to see that further refinement produces negligible change.

2.5. Piecewise-constant arrival-rate functions

Many algorithms have been developed to compute the transient performance of a stationary model for general initial conditions. These algorithms can be applied to calculate the TV performance of a TV model if we approximate the TV arrival-rate function by a piecewise-constant arrival rate function. Then we can recursively compute the TV performance on each interval by letting the initial distribution on each interval be specified by the terminal distribution on the previous interval.

For example, the transient workload process in an $M/GI/1$ queue is a Markov process whose TV distribution can be characterized by a two-dimensional Laplace transform. A numerical inversion algorithm was developed in Choudhury *et al.* [27] for multi-dimensional Laplace transforms, drawing on Abate and Whitt [1, 2, 3], and applied to compute the TV distribution of the transient workload process in the $M/GI/1$ queue. Moreover, that algorithm was applied to compute the performance measures of the TV workload process in the $M_t/GI/1$ queue with a piecewise-constant arrival process in Choudhury *et al.* [28].

Evidently, other algorithms for the transient performance can be used in the same way. For example, Lucantoni *et al.* [134] developed a numerical inversion algorithm for the transient behavior of the $BMAP/GI/1$ queue, having an arrival process that is a batch Markovian arrival process (also known as a Neuts process or versatile Markovian point process); see Lucantoni [133]. In addition, Abate and Whitt [4] developed a numerical inversion algorithm to compute transient blocking probability and other TV performance measures for the $M/M/s/0$ Erlang loss model for general initial conditions.

3. Time-Varying Deterministic Fluid Models

When the deterministic variability in the arrival and departure rates tends to be more important than the stochastic variability about those rates, it may be appropriate to ignore the stochastic part of the model altogether. Moreover, if the number in system varies over a fairly wide range, then we might also ignore the discrete nature of individual customers or jobs. That leads us to continuous deterministic fluid models as alternatives to (or approximations for) discrete stochastic TV queueing models.

These deterministic fluid models make optimization far more tractable; e.g., see Hampshire and Massey [72], Hampshire *et al.* [71], Niyirora and Pender [165], Niyirora and Zhuang [166], and Whitt [222]. We give a simple example in Section 3.2.

3.1. Fluid models for TV Markovian $M_t / M_t / s_t + M_t$ *queues*

We can obtain deterministic fluid models for the TV Markovian $M_t / M_t / s_t + M_t$ queues considered in Section 2 directly from the ODE for the mean in (2.7) by adjusting our interpretation. If we replace the mean $m(t)$ in (2.7) by a deterministic state $x(t)$, then (2.7) becomes the ODE

$$\dot{x}(t) = \lambda(t) - \mu(t)(x(t) \wedge s(t)) - \theta(t)(x(t) - s(t))^+. \tag{3.1}$$

3.2. Single-server fluid models

The single-server fluid model is the natural direct approach when we can assume that we have a TV arrival rate $\lambda(t)$ and a TV service rate $\mu(t)$; see Edie [43], Oliver and Samuel [167], May and Keller [153], and Newell [163] for early examples.

We need a different interpretation than is provided by (3.1). Now we need to keep track of whether or not the single server is busy. Thus, the deterministic analog of (2.11) is the ODE

$$\dot{x}(t) = [\lambda(t) - \mu(t)]1_{\{x(t)>0\}} + [(\lambda(t) - \mu(t))^+]1_{\{x(t)=0\}}, \tag{3.2}$$

where 1_A is the indicator function of the set A, equal to 1 in A and equal to 0 otherwise. The first part of (3.2) states that $x(t)$ evolves according to the simple ODE $\dot{x}(t) = \lambda(t) - \mu(t)$ when $x(t) > 0$. The second part of (3.2) states that, when $x(t) = 0$, $x(t)$ remains at 0 unless $\lambda(t) - \mu(t) > 0$.

This deterministicsingle-server model can often be analyzed by back-of-the-envelope calculations, without any ODE algorithms, as illustrated by the National Cranberry Cooperative case study in Porteus [179, 180, 181]. Here is a short story extracted from the longer detailed narrative: Trucks bring wet cranberries to be processed at a processing facility that has storage capacity for 3200 bbl (barrels). The trucks bring cranberries each day over the 12-hour time interval [7, 19] at a rate of 1050 bbl per hour. If the input exceeds the 3200 bbl storage, the input keeps arriving but waits in the delivery trucks, which is undesirable.

An original plan has the berries processed at the facility at 600 bbl per hour, with processing starting at 11 and continuing until the daily input has been processed. A quick analysis of the deterministic ODE shows that the plan leaves too much waiting for the trucks. A revised plan, yielding no truck waiting, has the processing rate increased from 600 bbl per hour to 800 bbl per hour, with the processing starting at 7 am. Figure 1 shows the inventory level building up and dissipating over a single day with the original plan (above) and the revised plan (below).

Many service system applications require little more than the analysis above; the rest of this paper is intended to help when that is not the case.

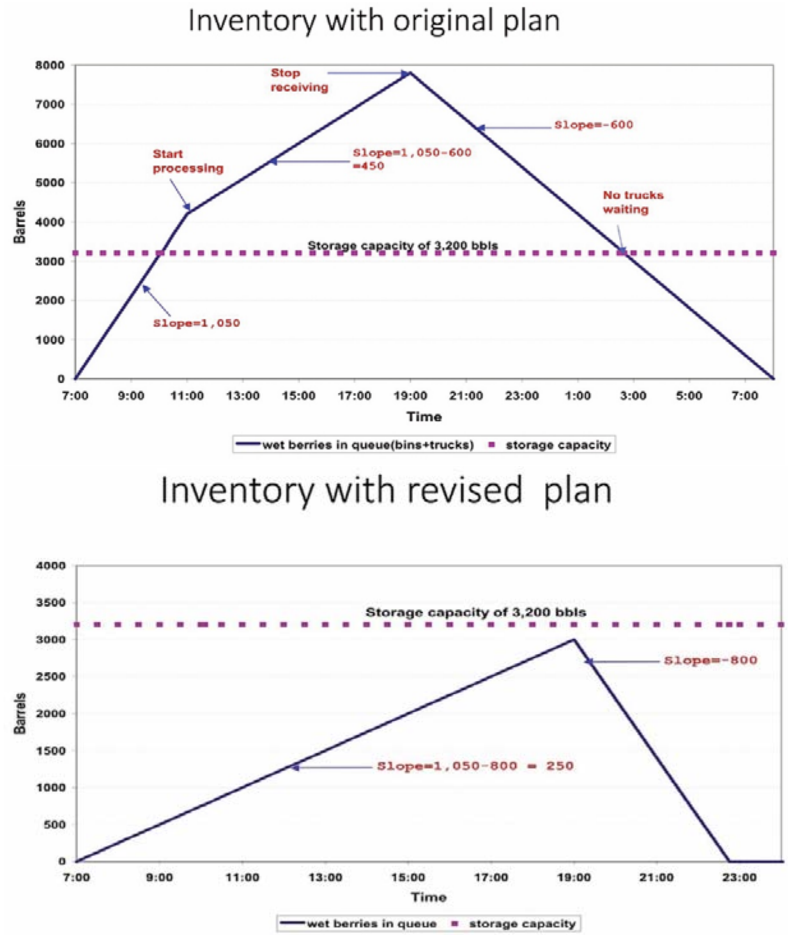

Figure 1. An example of the deterministic single-server fluid model applied to analyze the daily buildup and eventual decline of inventory over the day: a comparison of the original plan with processing at rate 600 bbl per hour starting at 11 (above) with a revised plan having processing at rate 800 bbl per hour starting at 7 (below).

3.3. *Two-parameter fluid models for non-Markov many-server queues*

Turning to something less obvious, but not so difficult after we get acclimated, we now discuss more recent two-parameter fluid models to approximate many-server queues with non-exponential service-time and patience-time distributions. These models are of interest because the service times and patience times often have non-exponential distributions in service systems; e.g., see Armony *et al.* [8], Brown *et al.* [25], and Whitt and Zhang [230]. It is natural to model many of these systems as $M_t / GI / s_t + GI$ queues, where the service times and patience times come from independent sequences of i.i.d. random variables with cdf's G and F, respectively, and associated pdf's g and f. In order to analyze the performance of these systems, it is natural to focus on the pair of two-parameter stochastic processes $(B,Q) \equiv \{(B(t,x),Q(t,x) : x \geq 0, t \geq 0\}$, where $B(t,x)$ $(Q(t,x))$ is the number of customers in service (queue) at time t that have been so for at most time x. This representation is convenient because the function-valued (or measure-valued) stochastic process $\{(B(t,\cdot),Q(t,\cdot)) : t \geq 0\}$ is a Markov process.

Nevertheless, the stochastic process (B,Q) is relatively complicated. Hence, it is natural to approximate the performance by a deterministic fluid model, which was done for the $G/GI/s+GI$ stationary model in Whitt [221] and for the $G_t/GI/s_t+GI$ TV model by Liu and Whitt [123]. The fluid model is specified by the four-tuple of functions (λ,s,g,f), where $\lambda(t)$ is the arrival rate and $s(t)$ is the capacity at time t, while g and f are the service and patience pdf's. For the deterministic fluid model, we interpret the cdf's G and F as proportions; e.g., $G(x)$ is the proportion of fluid that completes service within time x of entering service.

These fluid models also can be viewed as measure-valued functions. In addition, they can be viewed as deterministic limits in MSHT laws of large numbers for the stochastic model; see Section 6 here and Kang and Ramanan [90], Kaspi and Ramanan [91], Liu and Whitt [124, 126], Kang and Pang [89], Zhang [236], and Zuniga [239]. Kang and Pang [89] have shown that four different representations of this fluid model are equivalent.

The impact of the cdf's F and G beyond their means. This deterministic two-parameter fluid model is quite different from the fluid models in (3.1) and Section 3.2 because the cdf's F and G beyond their means can play an important role in system performance. For the overloaded stationary $G/GI/s+GI$ model, Tables 1-3 of Whitt [221] show that the patience cdf F can have a significant impact, but the service cdf G is relatively unimportant. In contrast, for the TV $G_t/GI/s_t+GI$ model, Figure 2 of Liu and Whitt [123] shows that the service cdf G also can have a big impact on the TV performance. That figure compares the fluid content in two $M_t/GI/s+E_2$ systems, one with service time having a mean-1 exponential (M) distribution, having squared coefficient of variation (scv, variance divided by the square of the mean) $c^2 = 1$, and the other having a mean-1 hyperexponential (H_2, mixture of two exponentials) distribution with scv $c^2 = 4$. The sample paths are very different and yet both agree with simulation estimates of the TV mean values in the stochastic models.

An important insight is that lessons learned about the stationary model do not necessarily remain valid for the more general TV setting. For another example, Davis *et al.* [38] showed that the service-time cdf G beyond its mean can have a significant impact on the blocking probability in the TV $M_t/GI/s/0$ loss model, even though the stationary $M/GI/s/0$ model has the celebrated insensitivity property.

Deterministic stochastic models and stochastic deterministic models. The Markovian $M_t/M/s_t+M$ model in Section 2 can be regarded as a *deterministic* *stochastic model*, because the process $X(t)$ being modeled is a stochastic process, evolving randomly over time, but only deterministic features of the model are specified; i.e., the exponential service-time and patience distributions are fully specified by their deterministic means, while the NHPP arrival process is fully specified by its deterministic rate.

In contrast, the $G_t/GI/s_t+GI$ fluid model considered in this section can be regarded as a *stochastic deterministic model*, because the process being modeled is a deterministic process, evolving deterministically over time, but stochastic features of the model are specified; i.e., there is a separate specification of the service-time and patience distributions beyond their means through the cdf's F and G (although the arrival process has no impact on the fluid model performance beyond its deterministic rate).

The evolution of the two-parameter fluid model. The evolution of the $G_t/GI/s_t+GI$ fluid model specified by (λ, s, g, f) can be characterized by the pair $(B(t,x), Q(t,x))$, where $B(t,x)$ $(Q(t,x))$ is the amount of continuous, divisible, deterministic fluid in service (queue) at time t that has been so for at most time x. These in turn can be characterized by the density functions $b(t,x)$ and $g(t,x)$, where

$$Q(t,y) = \int_0^y q(t,x)dx \quad \text{and} \quad B(t,y) = \int_0^y b(t,x)dx, \quad y \geq 0. \tag{3.3}$$

The evolution depends on whether the system is overloaded (all the service capacity $s(t)$ is being used) or whether it is not. The evolution is carefully analyzed in Liu and Whitt [123], to which we refer for more details. A key initial assumption is that the system alternates between *overloaded* (OL) intervals and *underloaded* (UL) intervals. Just as in Section 2.3.2, from the perspective of MSHT limits, this switching between OL and UL intervals again corresponds to being in the complementary-QED (QED^c) MSHT regime, discussed in Section 6.3.

We assume that fluid enters service from queue in order of arrival. As a consequence, at any time t when the system is OL, there will be a lower boundary of the queue-length density

$$w(t) \equiv \inf\{y \geq 0 : q(t,x) = 0, \quad \text{for all} \quad x < y\}. \tag{3.4}$$

The fundamental evolution equations state that fluid in service (queue) that is not served (does not enter service or abandon) remains in service (queue), i.e.,

$$b(t+x, x+u) = b(t,x)\frac{G^c(x+u)}{G^c(x)} \quad \text{and}$$

$$q(t+x, x+u) = q(t,x)\frac{F^c(x+u)}{F^c(x)}, \quad 0 \leq x < w(t) - u, \tag{3.5}$$

where $G^c(x) \equiv 1 - G(x)$ and $F^c(x) \equiv 1 - F(x)$.

The key flows depend on the hazard rates of the service-time cdf G and the patience cdf F, where $h_G(x) \equiv g(x)/G^c(x)$ and $h_F(x) \equiv f(x)/F^c(x)$. In particular, from (3.5) it is evident that the service rate $\sigma(t)$ and abandonment rate $\alpha(t)$ at time t are

$$\sigma(t) = \int_0^\infty b(t,x)h_G(x)dx \quad \text{and} \quad \alpha(t) = \int_0^\infty q(t,x)h_F(x)dx. \tag{3.6}$$

During each UL interval, the system behaves like an $M_t/GI/\infty$ TVIS fluid model, so that $b(t,x)$ evolves according to

$$b(t,x) = G^c(x)\lambda(t-x)1_{\{x \le t\}} + b(0,x-t)[G^c(x)/G^c(x-t)]1_{\{x>t\}}, \tag{3.7}$$

until the first time T that $B(t) > s(t)$, at which point an OL interval starts.

During an OL interval, $b(t,x)$ evolves according to

$$b(t,x) = G^c(x)b(t-x,0)1_{\{x \le t\}} + b(0,x-t)[G^c(x)/G^c(x-t)]1_{\{x>t\}}, \tag{3.8}$$

which is the same as in (3.7) except that the fluid rate entering service, $b(t-x,0)$, replaces the external fluid arrival rate, $\lambda(t-x)$. That is complicated because $b(t-x,0)$ is part of what we are trying to determine. However, it turns out that the function representing the rate fluid enters service, $b(t,0)$, satisfies the following fixed-point equation

$$b(t,0) = \hat{a}(t) + \int_0^t b(t-x,0)g(x)dx, \tag{3.9}$$

where $\hat{a}(t)$ is an explicit function of known quantities; see (19) in Liu and Whitt [123]. Under regularity conditions, the operator specified by the right side of (3.9) is a contraction map, so that equation (3.9) can be solved by successive iteration. The algorithm for each OL interval requires solving the fixed-point equation in (3.8) .

To analyze the queue performance in each OL interval it is convenient to look at the function $\tilde{q}(t,x)$ showing the queue content under the assumption that no fluid enters service from queue. The function $\tilde{q}(t,x)$ evolves the same way $b(t,x)$ does except that the abandonment cdf F plays the role of the service cdf G for $b(t,x)$; i.e., paralleling (3.8),

$$\tilde{q}(t,x) = F^c(x)\lambda(t-x)1_{\{x \le t\}} + q(0,x-t)[F^c(x)/F^c(x-t)]1_{\{t<x\}}, \tag{3.10}$$

The key insight is that, because all fluid enters service from the head of the queues, $q(t,x)$ differs from $\tilde{q}(t,x)$ only for $x < w(t)$; ie.,

$$q(t,x) = F^c(x)q(t-x,0)1_{\{x \le t \wedge w(t)\}} + q(0,x-t)[F^c(x)/F^c(x-t)]1_{\{t<x \le w(t)\}}, \tag{3.11}$$

It then turns out that the boundary function $w(t)$ evolves as an ODE. In particular, under general regularity conditions,

$$\dot{w}(t+) = 1 - \frac{b(t+,0)}{\tilde{q}(t,w(t)-)} \tag{3.12}$$

and any initial value $w(0)$.

Moreover, under regularity conditions, the potential waiting time (the waiting time of a hypothetical infinitely patient arriving atom of fluid at time t), denoted by $v(t)$, is the unique function satisfying the equation

$$v(t-w(t)) = w(t) \quad \text{or, equivalently,} \quad v(t) = w(t+v(t)) \quad \text{for all} \quad t \ge 0. \tag{3.13}$$

It is important to recognize that the $G_t/GI/s_t+GI$ fluid model is a valid mathematical model in its own right. Indeed, Theorems 2-6, Propositions 2, 5 and 6 and Corollaries 3, 5 and 6 in Liu and Whitt [123] present conditions verify that the performance description above is valid under regularity conditions stated there. Moreover, the performance functions, including $(b(t,x), q(t,x), W(t), v(t))$ for each OL interval, can be computed by an algorithm with complexity about the same as for the closure approximations in Section 2.3.1 and Section 2.3.2.

Example 3.1. (*the* $M_t/H_2/s+E_2$ *fluid model with a sinusoidal arrival rate*) Figure 2 illustrates the computational results by showing the plots of six performance functions for an $M_t/H_2/s+E_2$ fluid model with a sinusoidal arrival rate function: $\lambda(t)=1+0.6\sin(t)$, mean service time $1/\mu=1$, mean patience $1/\theta=1$, and fixed service capacity $s=1$, taken from Liu and Whitt [123]. Simulations confirm that these deterministic performance descriptions are effective for approximating the corresponding TV mean values in the n-server stochastic model with arrival rate $n\lambda(t)$ for $n=100$ and smaller n as well.

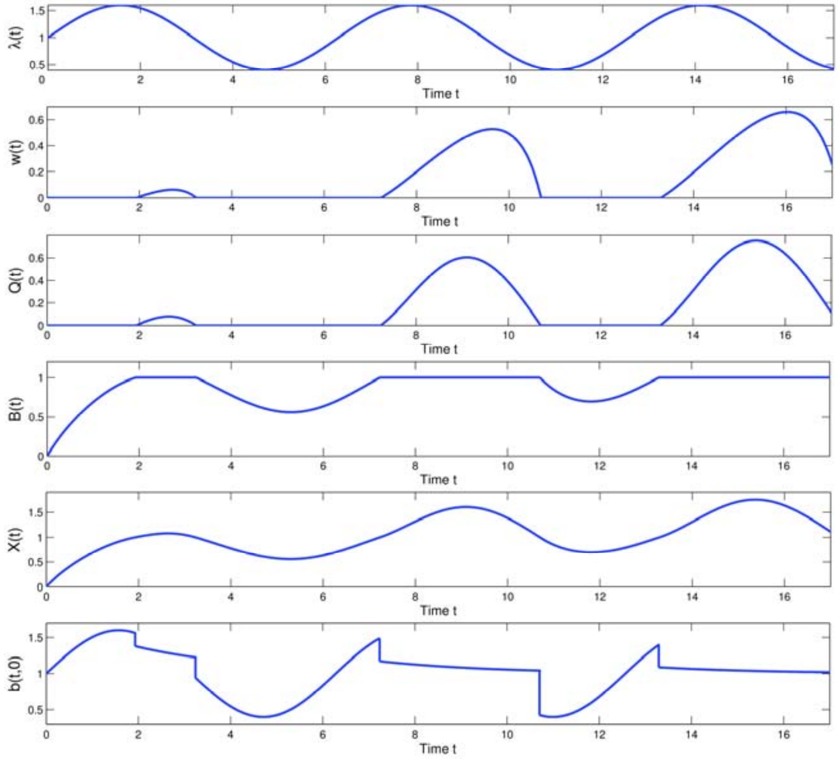

Figure 2. Six performance functions for the $G_t/H_2/s+E_2$ fluid model with sinusoidalarrival-rate function $\lambda(t)=1+0.6\sin(t)$ for service capacity s and mean service and patience times equal to 1: (i) arrival rate $\lambda(t)$; (ii) head-of-line waiting time $w(t)$; (iii) fluid waiting in queue $Q(t)$; (iv) fluid in service $B(t)$; (v) total fluid in system $X(t)$; and (vi) rate into service $b(t,0)$.

More on fluid models. Liu and Whitt [122], and Long and Zhang [131] show that, under regularity conditions, the fluid model has an asymptotic-loss-of-memory property, implying that the performance at time t is asymptotically independent of the initial conditions as t increases. As a consequence, the periodic fluid model has a dynamic periodic steady state, consistent with what we see from performance plots over several cycles. Even if we start the system empty in the algorithm, that limiting behavior is usually evident over a few cycles, as in Figure 2.

This TV fluid model is applied by Ibrahim and Whitt [80, 81] to improve delay announcements in a time-varying environment. Extensions of the analysis to networks of TV fluid queues are contained in Liu and Whitt [121, 127] and Zychlinski *et al.* [240].

There is a larger literature on stationary fluid models, which present directions for TV extensions. Fluid models are used in the analysis of kidney donation by Ata *et al.* [11]. Bassamboo and Randhawa [14] show that fluid models are remarkably effective in setting capacities. Other network fluid models are discussed in Talreja and Whitt [208].

Diffusion process refinements. Two-parameter diffusion models are useful refinements of the two-parameter fluid models, because they can capture the impact of the stochastic variability as well as the deterministic TV arrival-rate function. We will discuss them in Section 6.3.

4. Time-Varying Infinite-Server Queues

Many good methods for analyzing the stochastic behavior of non-Markovian $M_t/GI/s_t+GI$ models exploit the associated $M_t/GI/\infty$ TVIS model, which is remarkably tractable. Indeed, this $M_t/GI/\infty$ TVIS model can be regarded as the “prototype” or ideal form of the $G_t/GI/s_t+GI$ many-server queue that reveals the essential TV behavior, before focusing on the implications of the capacity constraints.

The IS models can also be useful tools even for many-server models that do not directly perform like IS models. For example, IS models can help analyze overloaded models when we regard the patience times as service times; see Liu and Whitt [126] and Aras *et al.* [6]. In the same spirit, the stationary IS model played an important role in analyzing the stationary $GI/GI/s$ model in Reed [186].

4.1. The Poisson random measure representation

Even though the number in system, $X(t)$, in the $M_t/GI/\infty$ TVIS model is *not* a Markov process, it has a Poisson distribution for each t with mean

$$m(t) \equiv E[X(t)] = \int_0^\infty \lambda(t-s)G^c(s)ds = E\left[\int_{t-S}^t \lambda(s)ds\right] = E[\lambda(t-S_e)]E[S], \quad t \ge 0, (4.1)$$

where $\lambda(t)$ is the deterministic arrival rate at time t, G is the cdf of a service time S, $G^c(s) \equiv 1-G(s) \equiv P(S>s)$ and S_e is a random variable with the *stationary-excess* or equilibrium lifetime cdf of G, i.e.,

$$G_e(x) \equiv P(S_e \le x) \equiv \frac{1}{E[S]}\int_0^x G^c(s)ds, \quad x \ge 0. \tag{4.2}$$

This follows because the arrivals together with the service times form a Poisson random measure in the plane; see Theorem 1 of Eick *et al.* [44] and references given there. As a consequence, the departure process is also a Poisson process with departure rate

$$\delta(t) \equiv \int_0^\infty \lambda(t-s)dG(s) = E[\lambda(t-S)], \quad t \ge 0. \tag{4.3}$$

Independence properties hold because the number of points in disjoint sets are independent under a Poisson random measure.

Consistent with Example 2.1, for the Markov $M_t/M_t/\infty$ TVIS model, the mean $m(t)$ satisfies the ODE given there by Theorem 6 and Corollary 4 in Eick *et al.* [44].

The offered load. The TV mean $m(t)$ in (4.1) is called the *offered load*, because with finitely many servers, it represents the expected number of servers needed if we ignored the capacity constraints (considered the associated IS model). As discussed in Eick *et al.* [44], formula (4.1) can be exploited to understand the "physics" of the TVIS queue and, as an approximation, the associated many-server queues.

For example, the final formula in (4.1) shows that the TV offered load coincides with the stationary offered load except for a random time lag by S_e. (In the stationary case, when λ is a constant, $m \equiv m(\infty) = \lambda E[S]$, by Little's law.) To get a rough idea of the impact of the service-time cdf, we can use the mean $E[S_e] = E[S](c_s^2+1)/2$; see (2) in Eick *et al.* [44]. Moreover, we see that there tends to be both a time lag and a space shift in the mean $m(t)$ behind the arrival rate; see equations (14)-(16) of Eick *et al.* [44]. For stationary models starting empty, we get a simple formula for the TV mean starting out empty: $m(t) = m(\infty)P(S_e \le t)$; see (20) in Eick *et al.* [44].

There are extensions to networks of IS queues and more general spatial models; see Duffield *et al.* [41], Massey and Whitt [147, 150], and Leung *et al.* [111].

4.2. Direct IS approximations with application to staffing

As discussed in Section 3 of Jennings *et al.* [87] and Section 4.3 of Green *et al.* [64], if we directly approximate an $M_t/GI/s_t+GI$ TVMS model by a TVIS $M_t/GI/\infty$ model, then we immediately obtain a Poisson distribution, which leads to the Gaussian approximation $X(t) \approx N(m(t), m(t))$, where $m(t)$ is the offered load in (4.1) and $N(m,v)$ denotes a random variable with a normal distribution having mean m and variance v. (We have variance equal to the mean because of the Poisson distribution.) If we choose $s(t)$ so that $P(N(m(t),m(t)) > s(t)) = \alpha$, then we obtain the classical *square-root-safety* (SRS) staffing formula

$$s(t) = m(t) + \gamma\sqrt{m(t)}, \tag{4.4}$$

where $\gamma \equiv P(N(0,1) > \alpha)$ is a *quality-of-service* (QoS) parameter. It is significant that the large subsequent literature primarily provides support for the SRS staffing formula in (4.4) , leading only to adjustments to the QoS parameter γ .

The exact distribution of $X(t)$ is more complicated in the $G_t / GI / \infty$ TVIS model, but fortunately it is often approximately Gaussian, so that it remains to find formulas for the TV mean and variance. The TV mean presents no problem because the mean $m(t) \equiv E[X(t)]$ in the $M_t / GI / \infty$ model in (4.1) remains unchanged if the arrival process is changed to G_t with the same arrival-rate function; see Theorem 2.1 and Remark 2.3 of Massey and Whitt [147] plus (4.20) in Section 4.5 below.

To support the Gaussian approximation and develop an approximation for the TV variance, we can apply the MSHT FCLT for the $G_t / GI / \infty$ TVIS model. For this model, the MSHT limits and the resulting Gaussian approximation tend to follow from the FCLT and Gaussian approximations for the arrival counting process, as can perhaps best be seen from the case of deterministic service times and then extending to service-time distributions that are finite mixtures of these, as in Glynn and Whitt [62]. In particular, for the $G_t / D / \infty$ model, $X(t) = A(t) - A(t - E[S])$ for each t, where A is the arrival process, i.e., the number in system at time t is the number of arrivals over an interval before t of length equal to the service time.

For the stationary model, the MSHT limit was first established by Borovkov [22]. For the $G_t / GI / \infty$ TVIS model, we can apply the two-parameter MSHT limit established by Pang and Whitt [171], which has been supplemented by an improved new chaining proof in Pang and Zhou [173]. The resulting Gaussian approximation has the mean in (4.1) and variance

$$\begin{aligned} v(t) &\equiv \int_0^\infty \lambda(t-s)V(s)ds \quad \text{with} \quad V(s) \equiv \bar{G}(s) + (c_a^2 - 1)\bar{G}(s)^2 \\ &= m(t) + (c_a^2 - 1)\int_0^\infty \lambda(t-s)\bar{G}(s)^2 ds, \end{aligned} \tag{4.5}$$

where c_a^2 is the asymptotic variability parameter for the arrival process, coming from an assumed FCLT for the arrival process, as in (5.6) in Section 5.3 below, so that the ratio of $v(t)$ in (4.5) to $m(t)$ in (4.1) , called the time-varying MSHT peakedness, is

$$z(t) \equiv \frac{v(t)}{m(t)} = 1 + (c_a^2 - 1)m(t)^{-1}\int_0^\infty \lambda(t-s)\bar{G}(s)^2 ds. \tag{4.6}$$

However, because the TV MSHT peakedness formula in (4.6) is complicated, it is natural to approximate it by the MSHT peakedness formula in the associated stationary $G / GI / \infty$ model (letting $m(t)^{-1} \approx \mu / \lambda$ and $\lambda(t-s) \approx \lambda$ in (4.6)) to obtain the MSHT stationary peakedness

$$z \equiv 1 + (c_a^2 - 1)\mu\int_0^\infty \bar{G}(s)^2 ds. \tag{4.7}$$

This analysis leads to our final TV approximation for the $G_t / GI / \infty$ model:

$$X(t) \approx N(m(t), zm(t)), \tag{4.8}$$

where $m(t)$ is given in (4.1) and z is given in (4.7).

Thus, when the arrival process is allowed to be G_t instead of M_t, instead of (4.4), we would staff according to

$$s(t) = m(t) + \gamma\sqrt{v(t)} \approx m(t) + \gamma\sqrt{z}\sqrt{m(t)}, \tag{4.9}$$

where again $\gamma \equiv P(N(0,1) > \alpha)$. This new version is still based on a Gaussian approximation, but involves a change in the variance.

From (4.8), we see that the TV behavior of $X(t)$ is captured by $m(t)$ in (4.1), while the impact of non-Poisson stochastic variability in the arrival process is captured by z in (4.7), which depends on the arrival process only via the parameter c_a^2, but also depends on the entire service-time cdf G. The use of the heavy-traffic approximation for z in delay and loss models is discussed and examined in Pang and Whitt [172], Li and Whitt [114], and He *et al.* [75]; see the references there for earlier work on peakedness.

4.3. The Modified-offered-load (MOL) approximation

The pointwise-stationary approximation (PSA). The MOL is a variation of the *pointwise-stationary approximation* (PSA). The PSA approximation applies to queues with finitely many servers. For any $M_t / GI / s_t + GI$ model, the PSA approximation is based on the associated stationary $M / GI / s + GI$ model. At time t, the PSA approximation for $X(t)$ is the steady-state random number, $\tilde{X} \equiv \tilde{X}(\lambda, G, F, s)$, where we let G and F be the given distributions, but we let $s = s(t)$, the actual number of servers at time t, and we let $\lambda = \lambda(t)$, the actual arrival rate at time t. The PSA approximation tends to be effective if the arrival rate changes slowly during the time of a single service time. The PSA approximation tends to be very effective in call centers when the average call holding time is short, e.g., less than 10 minutes. Asymptotics supporting the PSA approximation and refinements appear in Whitt [213], and Massey and Whitt [152].

The MOL approximation. The MOL approximation is a minor, but important, modification of PSA. The MOL approximation is just like PSA, except that instead of the actual arrival rate at time t, we use the MOL arrival rate

$$\lambda_{mol}(t) \equiv \frac{m(t)}{E[S]}, \tag{4.10}$$

where $m(t)$ is the mean number of busy servers in the associated TVIS model, which is the offered load in (4.1), and $E[S]$ is a mean-service time.

There is a simple logic: If the IS model were stationary at time t, then the offered load would be $m(t) = \lambda(t)E[S]$; the mean $m(t)$ provides a better starting point for a performance approximation than $\lambda(t)$, because it also accounts for the service-time distribution. Most important, MOL proves to be much more accurate than PSA with longer service times, while MOL reduces to PSA for shorter service times.

Example 4.1. (*staffing with MOL in the $M_t / M / s_t$ model*) For the $M_t / M / s_t$ model with TV arrival rate $\lambda(t)$, TV staffing $s(t)$ and constant service rate μ, instead of assuming that $X(t) \approx N(m(t), m(t))$, where $m(t)$ is the offered load in (4.1) as we did in Section 4.2 above, we now assume that $X(t) \approx \tilde{X}(\lambda_{mol}(t), \mu, s(t))$, where $\tilde{X}(\lambda, \mu, s)$ is the steady-state number in system in the stationary $M / M / s$ model with parameter triple (λ, μ, s). Thus, to achieve approximate delay probability α at each time t, we would choose $s(t)$ so that

$$s(t) \equiv \inf\{s \geq 0 : P(\tilde{X}(\lambda_{mol}(t), \mu, s) \leq \alpha\}. \tag{4.11}$$

This approximation is not difficult to implement because, for the stationary Markovian models, algorithms for the steady-state distribution of $\tilde{X}$ are readily available.

However, it is even easier to apply the MSHT limit, which for the stationary $M / M / s$ model comes from Halfin and Whitt [70]. That MSHT limit gives the non-Gaussian approximation

$$P(\tilde{X} \geq m + x\sqrt{m}) \approx HW(x) \equiv 1 / [1 + \Phi(x) / \phi(x)], \tag{4.12}$$

for $m \equiv \lambda / \mu$ not too small, where Φ and ϕ are the cdf and pdf of $N(0,1)$. To staff with delay probability target α, approximation (4.12) dictates that the SRS staffing formula (4.4) should hold with

$$\gamma \equiv HW^{-1}(\alpha), \tag{4.13}$$

where HW^{-1} is the inverse of the "Halfin-Whitt" function HW defined in (4.12). Section 4 of Jennings *et al.* [87] shows that the QoS parameter (4.13) provides an improvement to the Gaussian approximation in (4.4) .

Starting with Jagerman [86]. The MOL method was originated by Jagerman [86] for the $M_t / M / s / 0$ loss model with a fixed number of servers. Consistent with intuition, the MOL approximation tends to be more effective for many-server queues under relative light loading. Theoretical support for the MOL approximation for that model and the more general $M_t / Ph / s / 0$ model were provided in Massey and Whitt [148]. Peak congestion in $M_t / GI / s / 0$ models was studied using TVIS models in Massey and Whitt [151]. The time-varying performance of the nonstationary loss model with fixed staffing was also discussed in Grier *et al.* [65], and Pender [176].

For both delay and loss models, the MOL approximation is an alternative to two natural simple approximations. The first is the PSA discussed above, while the other is the *simple stationary approximation* (SSA), which uses the stationary model with the long-run average arrival rate. The SSA approximation usually exhibits poor performance whenever the arrival rate fluctuates significantly. Figures 1-3 of Jennings *et al.* [87] show the big advantage of the new infinite-server (IS) staffing scheme over PSA and SSA for multi-server delay models with longer service times. (In Jennings *et al.* [87] a direct IS approximation is first proposed, but it is extended to the MOL approximation in Section 4; see the review in Green *et al.* [64].)

As noted above, the MOL approximation tends to be ineffective when the staffing cannot be increased to meet high demand, so that the system becomes seriously overloaded. Then other methods may be needed to describe the performance, such as the Gaussian closure approximation in Section 2.3.2, the fluid models in Section 3, the stationary backlog-carryover approach in Stolletz [203], or extensions of the methods to describe overloaded single-server queues in Section 7.

4.4. Different systems and performance measures

The MOL approximation has been found to be very effective for the practical problem of choosing TV staffing in order to stabilize the performance at target levels. Given that the staffing is chosen by (4.11), which implements MOL, the system tends not ever to be overloaded, which tends to make MOL consistently effective. The application of MOL to set TV staffing levels to stabilize performance, first the delay probability and then other performance measures, was discussed in Jennings *et al.* [87], Green *et al.* [64], Feldman *et al.* [48], Liu and Whitt [125], Defraeye and van Niewenhuyse [40], Liu and Whitt [128], Yom-Tov nad Mandelbaum [234], He *et al.* [75], Liu and Whitt [129], and Liu [120]. Significant new ideas have played a role in the later contributions, including MSHT limits, so that the following discussion overlaps somewhat with Section 6.

Customer abandonment and the ISA. Feldman *et al.* [48] showed that the approach to staffing for the $M_t/M/s_t$ model in Example 4.1 extends to the associated $M_t/M/s_t+M$ model with customer abandonment from queue, using the MSHT limit from Garnett *et al.* [55] instead of the MSHT limit from Halfin and Whitt [70], which leads to the Garnett *et al.* [55] function instead of the Halfin and Whitt [70] function in (4.12).

In addition, Feldman *et al.* [48] developed a simulation-based iterative staffing algorithm (ISA) that can be used for a large class of models, which is useful when the steady-state distribution needed for the MOL approximation is not readily available. The simulation algorithm also confirms the effectiveness of the approximate staffing algorithm while it is being developed. That work provided additional support for the SRS staffing formula in (4.9) based on the offered load by showing that the ISA is consistent with the SRS. That was done by estimating the implied empirical QoS

$$\overline{\gamma}^{ISA}(t) \equiv \frac{\overline{s}^{ISA}(t) - m(t)}{\sqrt{m(t)}}, \quad t \geq 0. \tag{4.14}$$

Plots of the implied empirical QoS were approximately constant across a wide range of target delay probabilities; see Figures 12 and 3 of the e-companion to Feldman *et al.* [48] for the models with and without customer abandonment.

With customer abandonment, higher delay probabilities targets are more reasonable, because abandonment tends to reduce the queues. Without abandonment, the delay probability target might be 0.1, but with abandonment, it might be 0.5.

Stabilizing the abandonment probability and the expected delay. For high QoS (low delay probability targets), Feldman *et al.* [48] found that all performance measures tended to be stabilized using the staffing algorithm with a delay probability target, as discussed above. However, Figure 4 shows that abandonment probabilities are not stabilized at the same time by that approach at low QoS (high delay probability targets). To address, that problem, Liu and Whitt [125] introduced a new MOL staffing algorithm to stabilize abandonment probabilities for all QoS targets. Just like the previous staffing algorithm, this new method tends to stabilize all performance measures at high QoS targets. Indeed, the new method reduces to the previous one as the QoS increases, but the new method differs significantly for lower QoS.

To stabilize the TV abandonment probability (and the expected delay), Liu and Whitt [125] use IS models in a new way. Instead of directly replacing the $M_t / GI / s_t + GI$ model by its $M_t / GI / \infty$ IS counterpart, they introduce a new delayed infinite-server (DIS) model containing two IS facilities in series, one for the waiting room (or the queue), and the other for the service facility, as shown in Figure 3. Customers arrive at the first IS facility according to the given arrival rate $\lambda(t)$, but they remain until they either abandon or enter service. The key assumption is that all customers enter service at time w if they have not yet abandoned, so that the probability of abandoning is $F(w)$.

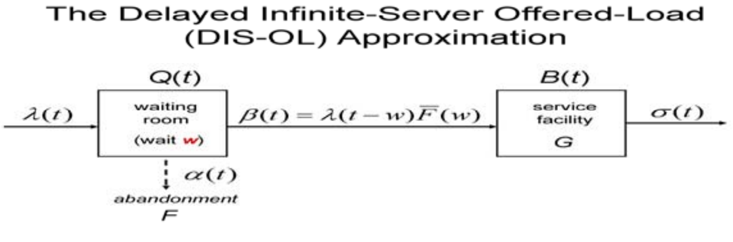

Figure 3. The delayed infinite-server offered-load (DIS-OL) approximation for the $M_t / GI / s_t + GI$ queueing model. The contents $Q(t)$ and $B(t)$ are independent Poisson random variables for each t; the three flows are Poisson processes.

Given that our goal is to stabilize the abandonment probability α, we choose w so that $F(w) = \alpha$. With this DIS model, the two TVIS queues become independent IS queues with NHPP arrival processes. The arrival rate at the second TVIS queue is $\beta(t) \equiv \lambda(t-w)\bar{F}(w)$, where $\bar{F}(w) \equiv 1 - F(w)$. With this DIS model the number of busy servers at time t, $B(t)$, has a Poisson distribution with mean $m_\alpha(t) \equiv E[B(t)]$, which we call the DIS offered load. A direct DIS staffing sets $s(t) = m_\alpha(t)$, but a much better approximation is the DIS-MOL approximation, which uses an MOL approximation withthe OL $m_\alpha(t)$, adjusted by the target abandonment probability, specifically, with

$$\lambda_{MOL,\alpha}(t) \equiv \frac{m_\alpha(t)}{E[S](1-\alpha)}. \tag{4.15}$$

The associated MOL staffing algorithm to stabilize the abandonment probability at α, lets $s_{mol,\alpha}(t)$ be the least staffing level such that the steady-state probability of abandonment in the stationary $M / GI / s + GI$ model with arrival rate $\lambda_{MOL,\alpha}(t)$ is less than or equal to α. We emphasize that this new DIS model is only used to define a new MOL and set the staffing function $s(t)$; it is not intended to directly model the system itself.

Liu and Whitt [125] conduct simulation experiments showing that the new DIS-MOL approximation is effective. They also establish important asymptotic results. First their Theorem 2 proves that both the DIS and DISMOL staffing algorithms are asymptotically correct in the MSHT limit, which puts the system in the ED MSHT limiting regime. Second, they prove that it is impossible to stabilize the mean queue length and the abandonment probability at the same time in the MSHT limit.

Staffing for non-Poisson TV arrival processes. The staffing methods were extended to non-Poisson TV arrival processes and networks of queues in He *et al.* [75], and Liu and Whitt [128, 129]. These new staffing algorithms exploit the methods described above. For the non-Poisson arrival processes, the logic for $G_t / GI / \infty$ models leading to (4.9) is combined with the MOL methods described above, including the MSHT FCLT's.

For the $G_t / GI / s_t$ model, paralleling (4.13) for the $M_t / M_t / s_t$ model, in (3.1) of He *et al.* [75] the SRS staffing formula in (4.4) is used, but with the QoS parameter γ set equal to

$$\gamma \equiv \gamma_\alpha \equiv \gamma_\alpha(1)\sqrt{z}, \tag{4.16}$$

where α is the target delay probability, z is the HT peakedness in (4.7) with c_a^2 there the arrival process asymptotic variability parameter from (5.6) and $\gamma_\alpha(1) \equiv HW^{-1}(\alpha)$ is the inverse of the Halfin and Whitt [70] MSHT delay function HW in (4.12). Section 6 of He *et al.* [75] extends the staffing algorithm to the $G_t / GI / s_t + GI$ model with customer abandonment, using the corresponding Garnett *et al.* [55] MSHT delay function for the $M_t / M / s + M$ model with heuristic extensions for the non-Markov model.

For the networks of queues in Liu and Whitt [128, 129], the DIS model is extended to a larger network of IS queues. In addition, care is taken to address non-Poisson arrival processes within the network that are departures from queues having non-exponential service-time distributions.

Staffing to stabilize the tail probability of delay. Let $V(t)$ be the offered waiting time, the virtual waiting time of a hypothetical arrival at time t if that arrival were infinitely patient. It is evident that it is convenient to focus on the delay probability $P(V(t) > 0)$ because $P(V(t) > 0) = P(X(t) \geq s(t))$, where $X(t)$ is the number of customers in the system, which tends to be easier to analyze or approximate. That was the reason that the delay probability was the main target for the IS staffing in Jennings *et al.* [87], reviewed in Section 4.2.

However, it is more common in practice to focus on the *tail probability of delay* (TPoD). Indeed, for call centers, the classical staffing goal is expressed by the $80-20$ rule, which stipulates that 80% of the customers should be delayed less than 20 seconds before a call is answered. In hospitals, there is concern about the delay after a decision has been made to admit a patient from the emergency department into an internal ward of the hospital. A goal has been to keep this "ED boarding time" below 6 hours; see Shi *et al.* [199].

Motivated by the ED boarding problem, Defraeye and van Niewenhuyse [40] showed that a variant of the ISA in Feldman *et al.* [48] also can be applied to stabilize the TPoD. More recently, Liu [120] has developed explicit analytical formulas to set staffing levels to meet a TPoD target $P(V(t) > w) = \alpha$. To treat this more refined two-parameter target, Liu [120] exploits the MSHT FCLT for the $G_t / M / s_t + GI$ model in Liu and Whitt [126], assuming that the system is overloaded at all times and thus in the ED regime. (The extension to GI service is heuristic.)

In particular, in the sequence of models indexed by n having TV arrival rate $n\lambda(t)$, the staffing is chosen to satisfy

$$s_n(t) = ns^{(1)}(t) + \sqrt{n}s^{(2)}(t), \quad t \geq 0, \tag{4.17}$$

where $s^{(1)}(t)$ is chosen from the fluid limit, while $s^{(2)}(t)$ is chosen more carefully as a "tuning parameter." In particular, $ns^{(1)}(t)$ is the DIS offered load $m_\alpha(t) \equiv E[B(t)]$, in Liu and Whitt [125] depicted in Figure 3. That gets the mean approximately at the right place. (See (6.4) in Section 6.1 for another use of this scaling approach as a tuning parameter and see Section 6.3 for further discussion.)

To get the second component of the staffing, $\sqrt{n}s^{(2)}(t)$, Liu [120] exploits the MSHT FCLT, which shows that the MSHT-scaled waiting time converges to a Gaussian distribution, so that $V(t) \approx N(m_s(t), \sigma_s^2(t))$, where $m_s(t)$ and $\sigma_s^2(t)$ depend on $s^{(2)}(t)$. Hence, it is possible to choose $s^{(2)}(t)$ so that

$$P(V(t) > w) \approx P(N(m_s(t), \sigma_s^2(t)) > w) = \alpha. \tag{4.18}$$

Staffing to stabilize blocking in TV loss models. More recently, MOL has been applied to set staffing levels, when they are flexible, to stabilize blocking probabilities in loss models in Li *et al.* [115], and Whitt and Zhao [233]. When staffing should be regarded as fixed, it is natural to consider controlling the demand instead, e.g., by dynamic pricing, as has been considered in Hampshire and Massey [72], Hampshire *et al.* [73] and references therein. However, there may be more flexibility in staffing than we initially think. For example, loss models are natural for an ambulance base serving several hospitals as in Restrepo *et al.* [188], for the rooms in a hotel as in Levi and Radovanovic [112] and for a bike-sharing system as in Henderson *et al.* [77]. In a short time scale the available resources are fixed, but in a longer time scale adjustments can be made. For example, the number of available ambulances or bicycles may be dynamic, because transfers can be made.

When we do consider staffing in a loss model, the first thing to notice is that it is not possible to stabilize blocking probabilities in loss models with time-varying arrival rates as well as the delay probabilities have been stabilized in corresponding delay models with conventional methods, because the blocking probabilities necessarily jump at the time of any staffing change.

To illustrate the difficulty, we show a simple example in Figure 4 with the sinusoidal arrival rate function

$$\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t), \ \ t \geq 0, \tag{4.19}$$

having average arrival rate $\bar{\lambda}$, amplitude β, and cycle length (or period) T (or equivalently, frequency $\gamma = 2\pi / T$). We let the mean service time be 1 time unit and the blocking probability target be $B = 0.1$. The left side of Figure 4 shows the blocking probability for the parameter triple $(\bar{\lambda}, \beta, T) = (100, 25, 100)$ with a direct application of MOL. The wild fluctuations we see occur because the blocking probability instantaneously jumps with each staffing change. Since the staffing is decreasing in the interval $[25, 75]$, we see jumps up at the staffing changes there, but outside that interval, where the staffing is increasing, we see jumps down. It is evident that the blocking probability immediately drops to 0 after any staffing increase, because there is always free capacity at that instant.

Nevertheless, Li *et al.* [115] and Whitt and Zhao [233] show that good performance can be obtained if we randomize the time of the staffing change in a small interval about each scheduled change time or if we average the probabilities over small intervals. That good performance is illustrated on the right in Figure 4.

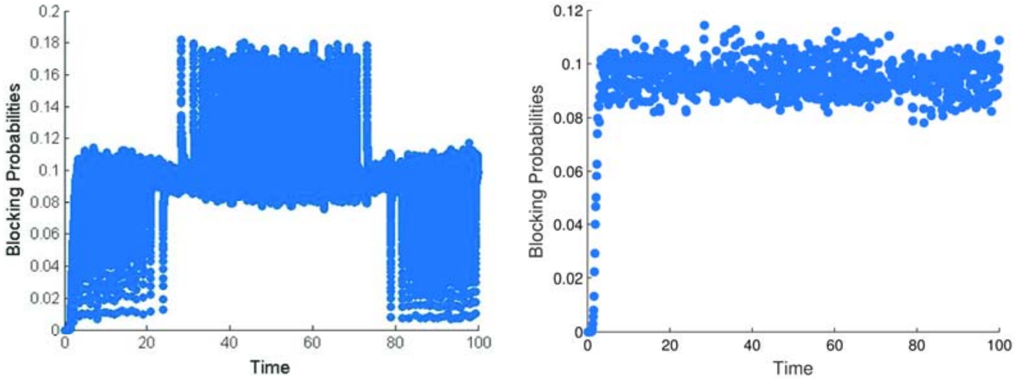

Figure 4. Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model for the sinusoidal arrival rate in (4.19) with blocking probability target $B = 0.1$ and parameter triple $(\bar{\lambda}, \beta, T) = (100, 25, 100)$ with the MOL staffing algorithm without any averaging (left) and with randomization (right).

4.5. Little's law

Little's law and the IS model. We include a discussion of Little's law here in this section on IS queues, because Little's law (LL, $L = \lambda W$), as in Little [116, 117], Stidham [202], and El-Taha and Stidham [46], is intimately connected to (i.e., essentially equivalent to) the infinite-server (IS) queueing model, as emphasized on p. 238 in the early review paper by Whitt [212]: The model for LL can be interpreted as an IS model by interpreting the waiting times as service times of customers who enter service immediately upon arrival.

Renewed interest in LL has occurred because of the important role it can play in interpreting data; see Glynn and Whitt [61], Little and Graves [118], Lovejoy and Desmond [132], Mandelbaum [138], Kim and Whitt [96], and Whitt and Zhang [230, 231].

The TVLL and the TVIS model. The TVIS model is in turn intimately connected to the TV Little's law (TVLL) in Bertsimas and Mourtzinou [18] and Fralix and Riano [52]; also see Kim and Whitt [97]. In fact, the TVLL is identical to the offered load formula in (4.1) extended to the case of a TV service distribution, as in formula (6) in Jennings *et al.* [87]:

$$m(t) \equiv E[X(t)] = \int_0^\infty \lambda(t-s) G^c_{t-s}(s)\,ds = \int_0^\infty \lambda(s) G^c_s(t-s)\,ds. \qquad (4.20)$$

The theoretical results in Fralix and Riano [52] show that this OL formula in (4.20) is valid for very general $G_t/G_t/\infty$ TVIS models. Except for determining the full level of generality, the TVLL in Bertsimas and Mourtzinou [18] follows directly from p. 236 of Whitt [212] and formula (6) in Jennings *et al.* [87].

A periodic Little's law (PLL). We now review a sample-path version of a periodic Little's law (PLL) established in Whitt and Zhang [231], motivated by the data analysis of an Israeli emergency department in Whitt and Zhang [230], in particular, because of the remarkable fit when comparing the stochastic model fit to the data to direct estimates from the data.

Because that data analysis was done in discrete time, the model has discrete time periods (DTP's) indexed by nonnegative integers k. We assume that all arrivals in a DTP occur before any departures. Moreover, we count the number of customers (patients in the ED) in the system in a DTP after the arrivals but before the departures. Thus, each arrival can spend j DTP's in the system for any $j \geq 0$.

With these conventions, just as for the data analysis in Whitt and Zhang [230], we focus on a single sequence, $X \equiv \{X_{k,j} : k \geq 0; j \geq 0\}$, with $X_{k,j}$ denoting the number of arrivals in period k that have length of stay (LoS) j periods. We also could have customers at the beginning, but without lost of generality, we can view them as a part of the arrivals in DTP 0. We define other quantities of interest in terms of X:

$Y_{k,j} \equiv \sum_{i=j}^{\infty} X_{k,i}$: the number of arrivals in DTP k with LoS greater or equal to j, $j \geq 0$,

$A_k \equiv Y_{k,0} = \sum_{j=0}^{\infty} X_{k,j}$: the total number of total arrivals in DTP k,

$Q_k \equiv \sum_{j=0}^{k} Y_{k-j,j} = \sum_{j=0}^{k} A_{k-j} \dfrac{Y_{k-j,j}}{A_{k-j}}, j \geq 0$; the number in system in DTP k,

all for $k \geq 0$. In the last line we understand $0/0 \equiv 1$, so that we properly treat DTP's with 0 arrivals.

With the periodicity in mind, we consider the following averages over n periods:

$$\begin{aligned}
\bar{\lambda}_k(n) &\equiv \frac{1}{n}\sum_{m=1}^{n} A_{k+(m-1)d},\\
\bar{Q}_k(n) &\equiv \frac{1}{n}\sum_{m=1}^{n} Q_{k+(m-1)d} = \frac{1}{n}\sum_{m=1}^{n}\left(\sum_{j=0}^{k+(m-1)d} Y_{k+(m-1)d-j,j}\right),\\
\bar{Y}_{k,j}(n) &\equiv \frac{1}{n}\sum_{m=1}^{n} Y_{k+(m-1)d,j},\ j \geq 0,\\
\bar{F}^c_{k,j}(n) &\equiv \frac{\bar{Y}_{k,j}(n)}{\bar{\lambda}_k(n)} = \frac{\sum_{m=1}^{n} Y_{k+(m-1)d,j}}{\sum_{m=1}^{n} A_{k+(m-1)d}},\ j \geq 0, \quad \text{and}\\
\bar{W}_k(n) &\equiv \sum_{j=0}^{\infty} \bar{F}^c_{k,j}(n), \quad \text{all for} \quad 0 \leq k \leq d-1.
\end{aligned} \tag{4.21}$$

With the framework above, we can state the sample-path version of the PLL. Let $[k] \equiv k \ mod \ d$ be the modulo function, i.e., the remainder when dividing k by d.

Theorem 4.1. (*sample-path PLL (from Whitt and Zhang* [231]*)*) *Assume that the following limits hold:*

$$(A1)\quad \bar{\lambda}_k(n) \to \lambda_k, \quad w.p.1 \quad as \quad n \to \infty, \quad 0 \le k \le d-1,$$

$$(A2)\quad \bar{F}^c_{k,j}(n) \to F^c_{k,j}, \quad w.p.1 \quad as \quad n \to \infty, \quad 0 \le k \le d-1, \quad j \ge 0, \quad and$$

$$(A3)\quad \bar{W}_k(n) \to W_k \equiv \sum_{j=0}^{\infty} F^c_{k,j} \quad w.p.1 \quad as \quad n \to \infty, \quad 0 \le k \le d-1, \tag{4.22}$$

where the limits are deterministic and finite. Then the limits are periodic functions; i.e., for $k \in \mathbb{Z}$ *and* $[k] \equiv k \bmod d$,

$$\lambda_k = \lambda_{[k]}, \quad F^c_{k,j} = F^c_{[k],j}, \quad j \ge 0, \quad and \quad W_k = W_{[k]}, \tag{4.23}$$

and the associated limits hold:

$$\begin{aligned} &(\bar{Q}_k(n), \bar{L}_k(n)) \to (L_k, L_k) \quad w.p.1 \quad as \quad n \to \infty, \quad where \\ &L_k \equiv \sum_{j=0}^{\infty} \lambda_{k-j} F^c_{k-j,j} < \infty \quad and \\ &\bar{L}_k(n) \equiv \sum_{j=0}^{k} \bar{\lambda}_{k-j}(n) \bar{F}^c_{k-j,j}(n) + \sum_{m=1}^{\infty} \sum_{j=1}^{d} \bar{\lambda}_{d-j}(n) \bar{F}^c_{d-j,(m-1)d+j+k}(n) \\ &\qquad = \sum_{j=0}^{\infty} \bar{\lambda}_{[k-j]} \bar{F}^c_{[k-j],j}, \quad n \ge 1 \end{aligned} \tag{4.24}$$

for $0 \le k \le d-1$.

The final line in (4.24) is what we should anticipate given the TVLL in (4.20) . When $d = 1$, the PLL reduces to the LL. Example 1 of Whitt and Zhang [231] shows that the condition on convergence of cdf's in $(A2)$ is needed in Theorem 4.1.

A CLT version of the PLL. Finally, we mention that a central-limit-theorem (CLT) version of the PLL has been established in Whitt and Zhang [232]; it parallels the CLT versions of LL in Glynn and Whitt [58, 59, 60, 61], and Whitt [223].

5. Arrival Process Models

Given that the arrival rate in a service system varies strongly over each day, the NHPP is a natural model for the arrival process, because it makes the queueing models relatively tractable and it tends to be at least roughly realistic. The Poisson property often arises from the independent decisions of many people, each of whom uses the service system only rarely. There is a supporting limit theorem, called the Poisson superposition theorem or the Poisson law of rare events; for example, see Section 11.2 of Daley and Vere-Jones [37] or Section 9.8 of Whitt [214].

5.1. Over-dispersion and under-dispersion

Even though the NHPP is a natural candidate for an arrival process model, there are phenomena that cause deviations from the Poisson property.

Over-dispersion. Indeed, there are several phenomena that tend to make the arrival process more variable than Poisson, leading us to say that there is *over-dispersion* compared to a Poisson process.

A common source of difficulty arises when the service system actually is a network of queues or can be regarded as a queue within such a network. First, when arrivals contain overflows from other service systems, as occurs in hospitals and hotels, the arrivals tend to occur in clusters, when the source system is overloaded. Second, when arrivals are departures from another queue with service distributions more variable than exponential, then the variability of the departure process tends to be greater than Poisson.

Evidently the problem of greatest concern in practice is that there may be uncertainty about the arrival rate, which could stem from the weather, holidays or other special events. Experience indicates that historical arrival data alone may not be adequate to build a good stochastic arrival-process model. Given arrival process data, it is often found that the rate itself needs to be regarded as a stochastic process, so that again the overall arrival process is more variable than Poisson. Evidence of such over-dispersion has been found by Avramidis *et al.* [12], Besbes *et al.* [19], Ibrahim *et al.* [79], Jongbloed and Koole [88], Kim and Whitt [99], and Zhang *et al.* [237]. We illustrate with an example from Mandelbaum [138] based on the US Bank data studied in Brown *et al.* [25].

Example 5.1. (*over-dispersion in call arrivals at a call center*) Figure 5 shows the number of calls arriving at a US bank call center each hour between 7:00 and 23:00 on 25 consecutive Mondays. At first glance, we are impressed by the consistency in the 25 plots, but can these sample paths actually be regarded as samples from i.i.d. NHPP's?

In fact, closer examination reveals significant over-dispersion. For a quick analysis, look around 13:00. We see that the range is approximately [2500, 3200], so that the mean should be about the midpoint, 2850. Given that the width of 700 should correspond to about 5 standard deviations, the standard deviation would be about 140, so that the variance would be about 19,600, which is much greater than the mean. (For an NHPP, the variance would equal the mean.) In fact, detailed data analysis shows that the sample mean is 2842 and the sample variance is 24,500. We can formalize that observation by applying the Poisson dispersion test, as in Section 4.1 of Kim and Whitt [99], to see that an NHPP is inconsistent with the data. Under the Poisson null hypothesis, the test statistic $\bar{D}_n \equiv (n-1)\bar{\sigma}_n^2 / \bar{x}_n = (24)(24500)/2842 = 206.9$ has approximately a chi-square distirbution with $n-1$ degrees of freedom, and so is approximately normal with mean

$n-1=24$ and variance $2(n-1)=48$, which makes the observed value of 206.9 about 26 standard deviations above its mean.

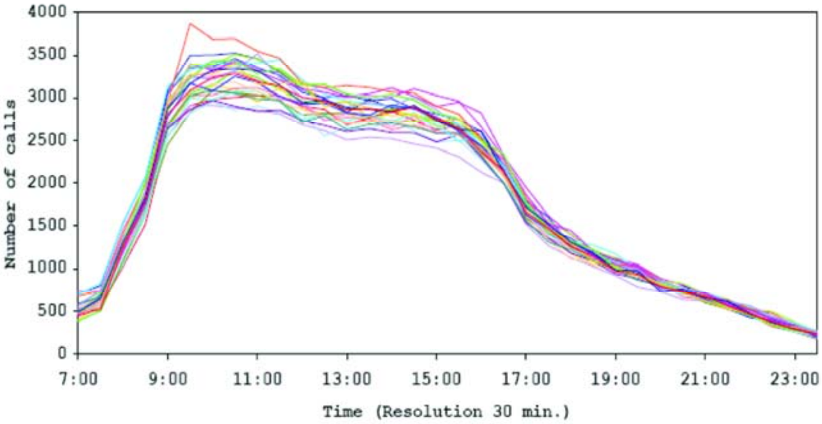

Figure 5. Number of calls arriving at a US bank call center each half hour between 7:00 and 23:00 on 25 consecutive Mondays.

Under-dispersion. On the other hand, there are also phenomena that make arrival processes less variable than Poisson, leading us to say that there is *under-dispersion*: First, there may be forced separation between successive arrivals, as in airplane landings at airports. Such separation may be present even if it is not evident. For example, arrivals at emergency departments may tend to cluster according to the schedule of public transportation. Second, the arrivals at a service system may be filtered, or go through stages, so that the final arrival process is more regular than the exogenous arrival process. That is the case when the arrivals are generated by an appointment system, which is designed to make the arrival process more regular. Moreover, many of the appointments will be for regularly spaced return visits, which again tend to be evenly spaced. Evidence of under-dispersion appears in data analysis of an endocrinology clinic by Kim *et al.* [95], and Kim *et al.* [100].

The relevant timescale for arrival processes. When we consider alternative stochastic models for arrival processes in service systems, we should consider what is the relevant timescale for the arrival process. Expressed more concretely, we should ask: How does the variability in the stochastic process affect the system performance?

Fortunately, heavy-traffic limit theorems for a queueing model of the service system can provide useful insight. The heavy-traffic limits indicate that the performance should be primarily affected by the CLT behavior of the arrival process, which is summarized by the asymptotic variability parameter in the CLT.

Perhaps the main issue is: What stochastic model properly represents the service system? The situation is relatively clear for a large call center, where there may be 1000 servers. The call center is likely to be well modelled by a $G_t / GI / s_t + GI$ model. In this application, interarrival times may be about $1/1000$ of a service time, so that we can infer that the individual interarrival times should not matter much. Indeed, the MSHT HT limits indicate that the performance is indeed primarily affected by the CLT behavior of the arrival process.

However, in other settings, such as in healthcare and web chat, the actual service time tends to be complicated, being composed of many separate pieces. As a consequence, the overall response time may be much longer. In general, it is less clear what is the relevant overall model. Then the congestion experienced by the customers is likely to be influenced more by variability in the arrival process.

Two-time-scale models in healthcare. From data analysis of arrivals to an emergency department (ED) in Whitt and Zhang [230] and to an endocrinology clinic in Kim *et al.* [95], and Kim *et al.* [100], we found that the arrival data were consistent with a two-time scale model in which the successive daily totals are modeled as a discrete-time Gaussian process, while the arrivals within the day, conditioned on the daily total, could be modeled as an NHPP with a deterministic arrival rate. As in Whitt and Zhang [230], both the daily total and the arrival rate function over the day can vary by day of the week, so that a week is a natural period for a periodic model.

The Gaussian-NHPP two-time-scale model implies that, once the daily total and arrival-rate function are given, the specified number of arrivals for that day arrive over the day according to i.i.d. random variables with a density proportional to the arrival rate for that day. This model was found to be effective in historical describing data over many weeks. It clearly can be implemented in simulation experiments.

For the ED data in Whitt and Zhang [230], there was slight over-dispersion compared to Poisson for the daily totals, with average variance to mean ratio of about 1.5, which is much less than in Example 5.1 for the call center. On the other hand, for the clinic arrival data in Kim *et al.* [95], which has arrivals by appointment, there was under-dispersion compared to Poisson for the daily totals, with average variance to mean ratio of about 0.5. For the clinic, the arrival process appears to be intermediate between Poisson and deterministic. In particular, the arrival process appears to be neither Poisson nor the ideal deterministic that appointment systems aim to approach.

This two-time-scale Gaussian-NHPP model would seem adequate for long-term planning, but it has not been tested to use for within-day prediction.

5.2. Testing the NHPP hypothesis

Recent efforts to test whether arrival data are consistent with an NHPP are contained in Kim and Whitt [98, 99], which follows Brown *et al.* [25]. Assuming that the arrival rate can be regarded as approximately piecewise-constant (PC), Brown *et al.* [25] proposed applying the classical conditional uniform (CU) property over each interval where the rate is approximately constant. For a Poisson process (PP), the CU property states that, conditional on the number n of arrivals in any interval $[0,T]$, the n ordered arrival times, each divided by T are distributed as the order statistics of n i.i.d. random variables, each uniformly distributed on the interval $[0,1]$. Thus, under the NHPP hypothesis with a PC arrival-rate function, if we condition in that way, the arrival data over several intervals of each day and over multiple days can all be combined into one collection of i.i.d. random variables uniformly distributed over $[0,1]$.

Thus, the NHPP hypothesis can be tested by applying the Kolmogorov-Smirnov (KS) statistical test to see if the resulting data are consistent with an i.i.d. sequence of uniform random variables. For that purpose, we construct the empirical cdf (ecdf)

$$\bar{F}_n(t) \equiv n^{-1} \sum_{i=1}^{n} 1_{\{X_i \le t\}}, \quad 0 \le t \le 1. \tag{5.1}$$

In this context, the KS test statistic is then

$$\bar{D}_n \equiv \sup_{0 \le t \le 1} \{ | \bar{F}_n(t) - t | \}. \tag{5.2}$$

We call the KS test of a Poisson process (PP) directly after applying the CU property to an NHPP with a piecewise-constant arrival rate the CU KS test.

Given that the CU representation is independent of the rate of the PP on each subinterval, we are able to combine data from separate intervals with different rates on each interval, but the constant rate on each subinterval also could be random; a good test result does not imply that the rate on each subinterval is deterministic. Thus, a random arrival rate remains to be addressed. That shortcoming could have helped Brown *et al.* [25] conclude that their call center arrival data were consistent with an NHPP.

In fact, the statistical testing is even more complicated, because Brown *et al.* [25] actually did not use the CU KS test directly. Instead, they applied a log KS test based on the CU property after performing an additional logarithmic data transformation. Kim and Whitt [98] investigated why an additional data transformation is needed and, if so, what form it should take. They showed through large-sample asymptotic analysis and extensive simulation experiments that the CU KS test of a Poisson process has remarkably little power against alternative processes with nonexponential interarrival-time distributions. That low power evidently occurs because the CU property focuses on the arrival times instead of the interarrival times; i.e., it converts the arrival times into i.i.d. uniform random variables.

Kim and Whitt [98] showed that the log KS test used by Brown *et al.* [25] has much greater power against alternative processes with nonexponential interarrival-time distributions. They also found that Lewis [113] had discovered a different data transformation in Durbin [42] to use after the CU transformation and that the Lewis KS test consistently has more power than the log KS test. In addition, they found that the CU KS test has advantages, because it turns out to be relatively more effective against alternatives with dependent exponential interarrival times. The data transformations evidently make the other methods less effective in detecting dependence because the reordering of the interarrival times weakens the dependence. Hence, Kim and Whitt [98, 99] recommend applying both the Lewis and CU KS tests.

Unfortunately, the Lewis test may be of little relevance for many service systems, because it focuses too much on the local behavior of the arrival process, which is often unimportant. The critical part may be the variability in a longer timescale.

5.3. A Composition construction for non-Poisson arrival processes

A natural way to construct a TV arrival process that goes one step beyond an NHPP is to add a single additional parameter to represent the level of variability. That can be done by applying the CLT, which can be done using a composition construction, which was proposed in Massey and Whitt [149], Gebhardt and Nelson [56], and Nelson and Gebhardt [156], and has been used in all other recent work studying queues with non-Poisson TV arrival processes.

The composition construction and the CLT. Let $A(t)$ count the number of arrivals over the interval $[0,t]$ and let $\lambda(t)$ be its deterministic time-varying arrival-rate function, satisfying $0<\lambda_{LB}\le\lambda(t)\le\lambda_{UB}<\infty$ for positive numbers λ_{LB} and λ_{UB}. Let $\Lambda(t)$ be the cumulative arrival-rate function, i.e.,

$$\Lambda(t)=\int_0^t \lambda(s)ds, \quad t\ge 0. \tag{5.3}$$

We assume that the general nonstationary arrival process A can be represented as the composition of a stochastic counting process N and the cumulative arrival rate function Λ, using the composition function $\circ$, with $(x\circ y)(t)\equiv x(y(t))$, $t\ge 0$; i.e.,

$$A\equiv N\circ\Lambda \quad \text{or, equivalently,} \quad A(t)\equiv N(\Lambda(t)), \quad t\ge 0, \tag{5.4}$$

where N is a stochastic counting process with nondecreasing nonnegative integer-valued sample paths. The construction in (5.4) is standard when A is an NHPP; then N is a rate-1 Poisson process. Then, and more generally, $E[A(t)]=\Lambda(t)$, $t\ge 0$. Ways to fit the arrival-rate function to data were studied in Massey *et al.* [144].

For our heavy-traffic limits, we will want the scaled arrival process based on A to satisfy a functional central limit theorem (FCLT), for which it will suffice for the process N to obey a FCLT, i.e.,

$$\hat{N}_n\equiv n^{-1/2}[N(nt)-nt]\Rightarrow c_a B \quad \text{in} \quad D \quad \text{as} \quad n\to\infty, \tag{5.5}$$

where B is a standard (mean 0, variance 1) Brownian motion (BM), $\Rightarrow$ denotes convergence in distribution and D is the function space of right-continuous real-valued functions on $[0,\infty)$ with the topology of uniform convergence over compact intervals. The asymptotic variability parameter c_a in (5.5) is the single-parameter characterizing the variability, with $c_a = 1$ corresponding to Poisson and $c_a = 0$ corresponding to deterministic.

As an immediate consequence of (5.4) , (5.5) and the continuous mapping theorem, we have a FCLT for the associated sequence of arrival processes $A_n(t) \equiv N(n\Lambda(t))$, $t \ge 0$, $n \ge 1$:

$$\hat{A}_n(t) \equiv n^{-1/2}[A_n(t) - n\Lambda(t)] \Rightarrow c_a B(\Lambda(t)) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty. \qquad (5.6)$$

Renewal and Cox models for N. It is significant that the process N can be very general; many specific models are consistent with the composition construction. First, N could be a renewal process with mean interarrival time 1 as well as its stationary (or equilibrium) version, as in Section V.3 of Asmussen [9], which necessarily satisfy the same FCLT in (5.5) ; e.g., see Nieuwenhuis [164].

However, it need not be either of those, which means that dependence among the interarrival times is allowed (under regularity conditions implying (5.5)). For example, the process N could be a Cox process (doubly stochastic Poisson process), which is a Poisson process where the arrival rate itself is a non-stationary stochastic process, as suggested by Avramidis *et al.* [12], Bassamboo and Zeevi [15], Ibrahim *et al.* [79], and Zhang *et al.* [237].

To represent N as a Cox process, we apply the composition construction again, letting

$$N \equiv M \circ C \quad \text{or, equivalently,} \quad N(t) = M(C(t)), \quad t \ge 0, \qquad (5.7)$$

where M is a stochastic counting process with nondecreasing nonnegative integer-valued sample paths and C is a stochastic cumulative process, expressed as

$$C(t) \equiv \int_0^t Z(s)ds, \quad t \ge 0, \qquad (5.8)$$

with $\{Z(t) : t \ge 0\}$ being a stochastic "rate" process (SRP) with nonnegative sample paths. We assume that the component stochastic processes M and C are mutually independent. Combining representations (5.4) and (5.7) gives a three-fold composition representation for the overall arrival process A: $A = M \circ C \circ \Lambda$.

This representation of N reduces to a stationary Cox process if we assume that M is a Poisson process.The most familiar stationary Cox process is a Markov-modulated Poisson process (MMPP), which arises when the SRP Z is a function of a continuous-time Markov chain (CTMC); see Fischer and Meier-Hellstern [51]. A further special case of an MMPP is an interrupted Poisson process (IPP), which is an MMPP with a two-state environment process, where the rate of the Poisson process is 0 in one of the two environment states. An IPP is equivalent to a renewal process with hyperexponetial (H_2) intervals between renewals; see Kuczura [106] and Section 2.3.1 of Fischer and Meier-Hellstern [51].

Our key stochastic assumption in this new framework is the validity of CLT's for the two stochastic processes M and C. Given that we want N to asymptotically have rate 1 and C to specify the cumulative rate, We assume that $M(t)/t \Rightarrow 1$ and $C(t)/t \Rightarrow 1$ w.p.1 as $t \to \infty$. Our key stochastic assumption in this new framework is the validity of CLT's for the two independent stochastic processes M and C.

$$t^{-1/2}[M(t)-t] \Rightarrow N(0,c_M^2) \quad and \quad t^{-1/2}[C(t)-t] \Rightarrow N(0,c_C^2) \qquad (5.9)$$

These together imply a CLT for N and A as in (5.5) and (5.6) with

$$c_A^2 = c_N^2 = c_M^2 + c_C^2, \qquad (5.10)$$

as in Example 9.6.2 of Whitt [214]. For additional details on the derivation of (5.10) , see Theorem 11.4.4 and Section 13.3 of Whitt [214].

Simulation and Fitting. Efficient simulation algorithms have been developed for both Poisson and non-Poisson TV arrival processes that exploit the composition structure; see Ma and Whitt [135] and Liu *et al.* [119]. Ways to estimate the TV arrival-rate function are discussed in Massey *et al.* [144], Zheng and Glynn [238] and references there.

Limitations. Even though the arrival process model based on the composition in (5.4), possibly with the additional composition in (5.7) , it is quite general, encompassing many specific models, it nevertheless is quite restrictive as well, As emphasized in Remark 2.2 of He *et al.* [75]. Some generality could be gained by allowing the variability parameter c_a^2 to depend on time as well. Despite the generality of this model, it evidently does not directly capture the two-time-scale Gaussian-NHPP model in Section 5.1. It appears that there remains a gap between what we see in arrival data and the TV arrival process models we can analyze.

Estimating the asymptotic variability parameter c_a. Given that we do use the arrival process model in (5.4), and want to apply it to arrival data, it remains to estimate the asymptotic variability parameter c_a in (5.5) and (5.6), The parameter c_a can be estimated by looking at the index of dispersion for counts (IDC), which is a normalized variance-time curve; see Cox and Lewis [33], and Fendick and Whitt [50]. In particular, if $A(t)$ counts the number of arrivals in the interval $[0,t]$, then the IDC is the function

$$I_c(t) \equiv \frac{Var(A(t))}{E[A(t)]}, \quad t \ge 0. \qquad (5.11)$$

If $A(t)$ is an NHPP, then $I_c(t)=1$ for all t.

Even for time-varying arrival processes, under regularity conditions, we can obtain the asymptotic variability parameter c_a^2 from estimates of the IDC over a suitably long time interval; i.e.,

$$c_a^2 = \lim_{t\to\infty} I_c(t). \qquad (5.12)$$

For an NHPP, we have $c_a^2 = 1$; for more (less) variable arrival processes, we have $c_a^2 > (<) 1$.

The IDC has been used to evaluate arrival processes in TVMS queues that are departure processes from other TVMS queues in Section 4 of Liu and Whitt [128]. The bottom-left plots in Figures 6 and 7 there show IDC estimates supporting an NHPP arrival process with $c_a^2 = 1$, whereas the bottom-right plots show IDC estimates suppporting a G_t arrival process with $c_a^2 \approx 3.5$. (This deviation from the NHPP property is partly caused by the previous TVMS queue having H_2 service times with $c_s^2 = 4$.) Thus, we see that network structure as with re-entrant customers in Yom-Tov and Mandelbaum [234] is likely to induce non-Markov arrival processes. Additional discussion of ways to estimate the asymptotic variability parameter c_a^2 are contained in Whitt and You [229].

6. Many-Server Heavy-Traffic Limits

The conventional heavy-traffic (HT) limit for stationary models. The early (conventional) heavy-traffic (HT) limit for stationary $G/G/s$ model in Iglehart and Whitt [83, 84] assumed that the number of servers, s, and the service-time distribution remain fixed, while the arrival rate, λ, increases, causing the traffic intensity ρ to increase toward the critical value 1 from below, which yields the same reflected Brownian motion (RBM) conventional heavy-traffic limit as for the $G/G/1$ single-server, reviewed in Chapter 9 of Whitt [214]. To obtain the limit, we introduce a family of models indexed by ρ and scale space by $1-\rho$ and time by $(1-\rho)^2$ when the traffic intensity is ρ; i.e., the scaled process is

$$\hat{X}_\rho(t) \equiv (1-\rho) X_\rho((1-\rho)^{-2} t), \quad t \geq 0, \tag{6.1}$$

so that the limit applies naturally to congested models (with high ρ) over long time intervals. The time scaling is the square of the space scaling, just as in the classic CLT for random walks, because the HT limit can be regarded as a consequence of Donsker's FCLT for random walks plus the continuous mapping theorem.

It is significant that this RBM limit holds for general $G/G/s$ models beyond the Markovian $M/M/s$ special case, because that special case can be analyzed directly via Section 2 and the classical Erlang results; e.g., see Brockmeyer *et al.* [24], and Whitt [215]. Theorem 2 of Iglehart and Whitt [84] shows that the HT limit for the process in (6.1) holds if the arrival and service processes satisfy a joint FCLT, in which case the RBM limit is only altered by the asymptotic variability parameters appearing in those limits. Thus the HT limit tells us the impact of the non-Markov model properties as well as the limit for the $M/M/s$ Markov model.

MSHT double limits: three limiting regimes. New limiting possibilities arise when there are more parameters that can change simultaneously. In fact, in the early paper on the stationary $M / M / s$ model, Iglehart [82] let s and λ both increase, obtaining the Ornstein-Uhlenbeck (OU) diffusion process MSHT limit when the arrival rate increases to ∞ when either $s = \infty$ or s increases so rapidly that the model is asymptotically equivalent to an IS model. Later, Halfin and Whitt [70] identified three possible limiting regimes for the $GI / M / s$ model as both λ and s increase to ∞, with the service rate fixed at $\mu = 1$ and ρ required to stay below 1, i.e.,

$$\lim\{(1-\rho)\sqrt{s} : s \to \infty, \lambda \to \infty, 0 < \rho < 1\} = \beta. \tag{6.2}$$

The case $\beta = 0$ is the overloaded case, which can be reduced to the HT limit for fixed s; the case $\beta = +\infty$ is the underloaded case, which is the same as for $s = \infty$, as in Iglehart [82]; the case $0 < \beta < \infty$ might be called the critically loaded case; now a useful new limit is obtained, which yields much better approximations in numerical examples, as illustrated by Table 1 in Halfin and Whitt [70].

It is significant that the MSHT process scaling is quite different from (6.1). For MSHT limits, we increase scale by considering a sequence of models indexed by the number of servers, s, and let $s \to \infty$. We again hold the service-time distribution fixed, but let the arrival rate also increase, so that the traffic intensity satisfies (6.2). The scaled process now is very different from (6.1), now being

$$\hat{X}_s(t) \equiv s^{-1/2} X_s(t), \quad t \geq 0. \tag{6.3}$$

By (6.2), we see that the spatial scaling is essentially the same provided that $0 < \beta < \infty$, but in (6.3) there is no additional time scaling. The large scale with many servers and accelerated arrival rate provided by the MSHT regime makes it unnecessary to further accelerate time to obtain a diffusion limit.

In the useful middle MSHT limiting regime, the limiting diffusion process is a hybrid of an RBM and an OU, leading to a corresponding hybrid approximating steady-state distribution of the number in system, being exponential, conditional on all servers being busy, and Gaussian, conditional on all servers not being busy. (For more on such hybrid limits for one-dimensional piecewise-linear diffusion processes, see Browne and Whitt [26].)

The useful middle MSHT limiting regime was called the Halfin-Whitt regime by Puhalskii and Reiman [184], who obtained a multi-dimensional diffusion limit in the case of phase-type service distributions. As illustrated by Puhalskii and Reiman [184], the MSHT limit depends on the service-time distribution in a much more complicated way than when s is held fixed. In particular, the limit process no longer is a tractable one-dimensional diffusion process. Other limits for the $G / G / s$ model with non-exponential service were obtained in Whitt [220], Reed [186] and Kaspi and Ramanan [92]. These papers address the challenging problem of non-exponential service. For the $GI / M / s$ model, Halfin and Whitt [70] obtained MSHT limits for both the transient process and the steady-state distribution.

In contrast, for GI service, the MSHT behavior of the steady-state distribution is more complicated. However, tightness for MSHT scaling has been established by Gamarnik and Goldberg [53]. While this recent work provides important structural insight, much still needs to be done to develop useful approximations; see Whitt [217, 219] and Liu *et al.* [130] for some progress in that direction.

Customer abandonment. For call centers and many other service systems, customer abandonment plays a prominent role. Thus it is significant that similar tractable MSHT limits limits hold for the Markovian $M/M/s+M$ model with customer abandonment, as shown by Garnett *et al.* [55]. With abandonment, we can have $-\infty < \beta < +\infty$ for the QED regime, which is again specified by (6.2) , because the abandonment ensures stability even if $\rho > 1$. The region where $-\infty < \beta < +\infty$ was called the "rationalized" regime in Garnett *et al.*[55] and the quality-and-efficiency-driven (QED) MSHT limiting regime in Section 4.1.1 of Gans *et al.* [54]. The underloaded and overloaded regimes are called the quality-driven (QD) and (QED) driven regimes, respectively. With abandonment, the ED regime becomes more relevant too, as emphasized by Whitt [217]. The abandonment rate θ presents another parameter that we might vary as well; Section 4 of Whitt [217] shows that a tractable OU limit arises as $\theta \to 0$ as well as $\lambda \uparrow \infty$ and $s \uparrow \infty$, provided that $s/\theta \to \infty$. See He [76] for a recent generalization of that result.

With abandonment, MSHT limits again become complicated for nonexponential service times and patience times. Useful approximations were developed directly in Whitt [219] and supplemented by MSHT limits; see Zeltyn and Mandelbaum [235], Mandelbaum and Zeltyn [141], Reed and Tezcan [187], and Liu *et al.* [130] and references therein. See Dai and He [35] and Ward [209] for surveys.

The relevance of the different regimes of course depends on economic factors; see Whitt [216] and Borst *et al.* [23] for discussion.

The QED and QED^c TV MSHT limiting regimes. When we consider TV many-server queues, it is evident that MSHT limits should still play an important role, but the setting is more complicated because the traffic intensity becomes TV as well. We emphasize two relatively simple cases, which have been exploited to advantage: The first, and most natural (or obvious), which we call QED, arises when the TV staffing is chosen so that the key QED condition is maintained at all times. The second, which we now call *the complementary QED* (or QED^c) *MSHT limiting regime*, arises in the complementary case in which the system alternates between overloaded (OL) and underloaded (UL) intervals, instantaneously passing through the critical loading in each transition between OL and UL.

We should hasten to point out that by "complementary" we mean for a single time t, and not for functions, because there are of course many other possibilities. By QED^c, we mean that, for almost all t with respect to Lebesgue measure, the regime is *not* QED. A natural goal is to seek the greatest generality, but we emphasize more restricted goals in order to develop useful insight and algorithms. We next discuss the TV QED MSHT regime; we discuss an application of the QED MSHT regime to scheduling in multi-class queues in Section 6.2; then we discuss the TV QED^c MSHT regime in Section 6.3.

6.1. The time-varying QED MSHT regime

The first MSHT limits for a general class of TV Markov models were obtained in Mandelbaum *et al.* [140]. The framework there is quite general, so that it is a bit difficult to identify the limiting regime. However, the QED regime is imposed through the scaling in (1.13) there. For the TV Markov $M_t / M_t / s_t + M_t$ model, we follow the more recent Puhalskii [183], which presents direct martingale arguments. The QED scaling is clearly set forth there in equations (2.11a)-(2.11d). With that scaling, the QED regime holds for all time.

In fact, following Whitt and Zhao [233] and Sun and Whitt [205], here we will consider the special case consisting of a sequence of $M_t / M / s_t + M$ models indexed by n with fixed service rate $\mu = 1$ and abandonment rate θ, $0 \le \theta < \infty$. Let the arrival rate functions in model n be $\lambda_n(t) \equiv n\lambda(t)$ for a fixed arrival-rate function $\lambda(t)$ with $\lambda(t) = 0$ for all $t < 0$. We write $g(t) = o(t)$ if $g(t)/t \to 0$ as $t \to \infty$. We impose the QED condition by assuming that the staffing functions satisfy

$$s_n(t) = nm(t) + \sqrt{n}c(t) + o(\sqrt{n}) \quad \text{as} \quad n \to \infty, \tag{6.4}$$

where $m(t)$ is the offered load in (4.1) and $c(t)$ is an integrable function for all t, which we think of as a staffing control function. As in Puhalskii [183], when the staffing decreases with all servers busy, let the customers be moved to the end of the queue and let them receive a new full service when they are next assigned. Let $X_n(t)$ be the number of customers in model n at time t.

An important special case of (6.4) arises when $c(t) \equiv c\sqrt{m(t)}$ for some constant c; then (6.4) reduces to the SRS staffing formula in (4.4) , but we allow greater generality. The main point is that the staffing in (6.4) puts the TV $M_t / M / s_t + M$ model into the QED MSHT limiting regime at all times t.

To state the limit, let $X_n(t)$ be the number in system at time t in model n and let the FWLLN and FCLT scaled processes be defined by

$$\bar{X}_n(t) \equiv n^{-1} X_n(t) \quad \textit{and} \quad \hat{X}_n(t) \equiv n^{-1/2}[X_n(t) - nx(t)], \quad t \ge 0, \tag{6.5}$$

where $x(t)$ is the limit in the FWLLN.

Theorem 6.1. (*QED MSHT FCLT in the $M_t / M / s_t + M$ delay model from Mandelbaum et al.* [140], *Puhalskii* [183]) *For the sequence of $M_t / M / s_t + M$ delay models specified above, if $\bar{X}_n(0) \Rightarrow x(0)$ in $\mathbb{R}$ as $n \to \infty$, where $x(0)$ is deterministic, then*

$$\bar{X}_n(t) \equiv n^{-1} X_n(t) \Rightarrow x(t) \quad in \quad D \quad as \quad n \to \infty, \tag{6.6}$$

where $x(t)$ *satisfies the ordinary differential equation* $\dot{x}(t) = \lambda(t) - x(t),$ *so that* $x(t) = m(t)$*, the OL in the* $M_t / M / \infty$ *IS model, provided that it is given consistent initial conditions.*

If, in addition, $\hat{X}_n(0) \Rightarrow \hat{X}(0)$ *in* $\mathbb{R}$ *as* $n \to \infty$*, then*

$$\hat{X}_n(t) \equiv n^{-1/2}[X_n(t) - nx(t)] \Rightarrow \hat{X}(t) \quad in \quad D \quad as \quad n \to \infty, \tag{6.7}$$

where $\hat{X}(t)$ *is a diffusion process satisfying*

$$\hat{X}(t) = \hat{X}(0) - \int_0^t (\hat{X}(s) \wedge c(s))ds - \theta \int_0^t (\hat{X}(s) - c(s))^+ + \int_0^t \sqrt{\lambda(s) + x(s)}dB(s) \tag{6.8}$$

with B *being standard Brownian motion. As a consequence, if* $\lambda(t)$ *is Lipschitz continuous and* $c(t) > 0$ *for all* t*, then*

$$P(X_n(t) \geq s_n(t)) = P(\hat{X}_n(t) \geq c(t) + o(1)) \to P(\hat{X}(t) \geq c(t)) > 0 \quad as \quad n \to \infty \tag{6.9}$$

for all $t > 0$*. Hence, the staffing in (6.4) puts the system asymptotically in the QED MSHT regime for each* $t > 0$*.*

Proof. This is a simplification of Theorems 1 and 2 of Puhalskii [183]. In particular, in the setting there we have: $\gamma_s = \beta_s = 0$, $q_s = \kappa_s = m(s)$, $\alpha_s = \lambda(s)$ and $\delta_s = c(s)$ for all s. Theorem 1 of Puhalskii [183] implies that the limit in (6.6) holds with limit $x(t)$, where $x(t)$ satisfies the ordinary differential equation $\dot{x}(t) = \lambda(t) - x(t)$. However, Corollary 4 of Eick *et al.* [44] implies that the OL $m(t)$ also satisfies the same ODE. Hence, $x(t) = m(t)$, $t \geq 0$. The second limit in (6.7) follows from Theorem 2 of Puhalskii [183]. The Lipschitz continuity of $\lambda(t)$ ensures that the one-dimensional distribution of the diffusion process $\hat{Q}(t)$ has a continuous cdf for each t, which is required for the limit to hold for all $c(t)$ in (6.9) ; see Theorem 3.2.1 of Stroock and Varadhan [204].

Remark 6.1. (*not centered at the natural fluid model*) *It is significant that the FWLLN limit in* (6.6)*, which is the centering term in the FCLT* (6.7)*, is the OL* $m(t)$*, which satisfies the ODE of the associated IS model, instead of the natural direct ODE for the fluid model with abandonment, which would yield*

$$\dot{x}(t) = \lambda(t) - \mu(x(t) \wedge s(t)) - \theta(x(t) - s(t))^+. \tag{6.10}$$

This occurs because we staff to order $O(n)$ *at the scaled OL* $nm(t)$ *in* (6.4)*. In* (2.6) *of* Puhalskii [183] *our scaling yields* $q_s = k_s$ *for* $s \geq 0$*.*

While this statement of the QED MSHT FCLT in the $M_t / M / s_t + M$ delay model is clean, it remains to show that it offers any advantages over Section 2 for actually computing TV performance measures. Nevertheless, in the next section we show how limits in the QED MSHT regime can provide important insight via the sample-path TV MSHT Little's law.

Extensions of Theorem 6.1 to more general non-Markovian models is an important remaining research problem.

6.2. Scheduling for multiple classes in a time-varying environment

We now review recent TV QED MSHT limits for multi-class queues in Sun and Whitt [205] that provide insight into ratio scheduling rules for achieving desired service differentiation with TV arrival rates.

Gurvich and Whitt [66, 67, 68] showed that *fixed-queue-ratio* (FQR) *controls* that schedule (select the next customer to enter service from queue when a server becomes free) and route to different server pools, which we do not consider here, aiming to keep the queue lengths at fixed ratios also are effective for achieving delay-based service-differentiation in stationary large-scale service systems modeled as many-server queues. Indeed, the goals are achieved asymptotically in the QED MSHT regime.

We conducted simulations experiments to see how the FQR control performs with time-varying arrival rates. We found that FQR controls remain quite effective for balancing the queue lengths over time, keeping them near desired ratios, but that the FQR controls can be highly ineffective at the indirect goal of stabilizing delays at fixed ratios. Thus, we investigated alternative *head-of-line-delay-ratio* (HLDR) controls aiming to keep the head-of-line delays at fixed ratios. Figure 6 illustrates by showing simulation results for FQR (left) and HLDR (right) for the two-class $M_t / M / s_t$ model with common service rate $\mu = 1$, no abandonment ($\theta = 0$) and sinusoidal arrival-rate functions $\lambda_i(t) = a_i + b_i \sin(d_i t)$, where $(a_1, b_1, d_1) \equiv (60, -20, 0.5)$ and $(a_2, b_2, d_2) \equiv (90, 30, 0.5)$. We chose the staffing $s(t)$ to be consistent with the SRS staffing formula in (4.4) for the aggregate model, using QoS parameter $\gamma = 0.25$.

The property that causes difficulties for FQR is class-dependent arrival rates, i.e., where the ratios of the arrival rates of two different classes varies strongly over time. It is thus significant that class-dependent arrival rates may indeed occur in applications. For example, Section 3.5 of Whitt and Zhang [230] shows that the proportion of arrivals to the Israeli emergency department (ED) that are admitted to an internal ward of the hospital varied strongly over time. Since the admitted patients tend to be among the more critical patients, we infer that there is likely to be a difference in the arrival rates of patients classified by acuity.

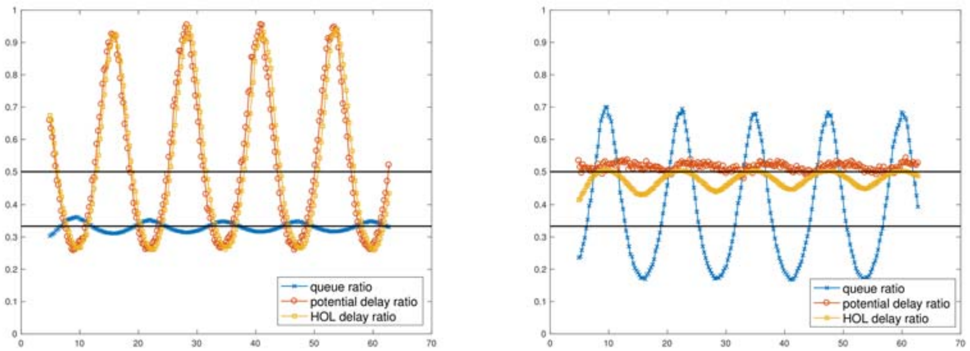

Figure 6. Queue and delay ratios obtained from the FQR rule (left) and HLDR rule (right) for a two-class $M_t / M / s_t$ queue with arrival rate functions $\lambda_1(t) = 60 - 20\sin(t/2), \lambda_2 = 90 + 30\sin(t/2)$, common service rate $\mu = 1$, without abandonment ($\theta_1 = \theta_2 = 0$) and the QoS parameter $\gamma = 0.25$.

Sun and Whitt [205] conclude that the results can be explained by a *sample-path (SP) TV MSHT Little's law (LL)* (SP-TV-MSHT-LL) that is a consequence of a TV QED MSHT limit, which is a generalization of the the SP-MSHT-LL for the stationary model that is a consequence of Theorem 4.3 in Gurvich and Whitt [66] and is discussed after equation (13) in Section 3 of Gurvich and Whitt [68]. In particular, the SP-TV-MSHT-LL states, for large scale systems in the QED MSHT regime, that

$$Q_i(t) \approx \lambda_i(t) V_i(t) \quad \text{for all} \quad t, \tag{6.11}$$

where $Q_i(t)$ is the queue length, $\lambda_i(t)$ is the arrival rate and $V_i(t)$ is the potential delay at time t for class i. It is illustrated for individual sample paths in Figure 5 of Sun and Whitt [205]. The SP-MSHT-LL in (6.11) follows from the TV QED MSHT limit established in Theorem 4.1 of Sun and Whitt [205], which extends Theorem 6.1 above to the multi-class setting. As in Gurvich and Whitt [66, 67, 68], there is great state-space collapse, which makes the limit process for the quantities in (6.11) above for all classes be only one-dimensional.

If we consider ratio

$$QR(t) \equiv Q_1(t) / Q_2(t), \;\; AR(t) \equiv \lambda_1(t) / \lambda_2(t) \;\text{ and }\; DR(t) \equiv V_1(t) / V_2(t),$$

then as a consequence of (6.11) we have

$$QR(t) \approx AR(t) * DR(t) \quad \text{for all} \quad t. \tag{6.12}$$

Given (6.12), $QR(t)$ and $DR(t)$ can both be nearly constant over time only if $AR(t)$ is nearly constant over time. The new SP-MSHT-LL implies that it is impossible to stabilize queue ratios and delay ratios simultaneously with these ratio controls in the MSHT limit when the ratio of the asymptotic arrival rates is time-varying. Otherwise, these ratio controls stabilize both queue ratios and delay ratios; e.g., see Figures 1-4 of Sun and Whitt [205].

6.3. The QED^c MSHT regime

Given that the QED regime has proven so successful for stationary many-server models, it is remarkable that we might elect to discard it entirely by considering the QED^c MSHT regime, which assumes that it never holds or, more precisely, that it holds only at the instantaneous transition points between OL and UL intervals.

The reason is that the QED limit for the $M_t/M/s_t+M$ model in Theorem 6.1 remains relatively intractable, requiring further analysis something like Section 2. On the other hand, the QED^c assumption leads to relatively tractable analysis, which ultimately is based on the IS queue.

It turns out that the behavior on both OL and UL intervals is largely determined by associated IS models. The story for UL intervals is obvious, because during each UL interval the system evolves the same as the corresponding IS model. The OL intervals are more complicated, but it turns out that a similar story holds there as well, provided that we let the patience times play the role of the service times, as discussed in Section 3.3.

In fact, both the Gaussian closure approximations in Section 2.3.2 and the two-parameter fluid model in Section 3.3 can be regarded as a consequence of the QED^c MSHT assumption. That is evident from Liu and Whitt [123], because it is explained at the outset that the model is assumed to alternate between OL and UL intervals. Moreover, the analysis of the OL intervals explicitly exploit the IS perspective in (3.10) . The QED^c role for the Gaussian closure approximations is less apparent, because even in the abstract the authors advertise the QED scaling in Halfin and Whitt [70]. However, further examination of Massey and Pender [145] reveals the QED^c MSHT assumption instead. In particular, that follows from the discussion below (2.6) and the Gaussian distributions claimed for the limit process after (2.8); the Gaussian limit only arises in the QED^c MSHT regime.

That QED^c perspective is maintained in the supporting TV MSHT limits in Liu and Whitt [124, 126] and for the initial content process in Aras *et al.* [7] and the stationary $G/GI/s+GI$ model in Aras *et al.* [6]. The MSHT FCLT during an overloaded interval yields an insightful stochastic differential equation of the form

$$dW(t)=H(t)W(t)+J_1(t)B_1(t)+J_2(t)B_2(t)+J_3(t)B_3(t), \qquad (6.13)$$

where $H(t)$ and $J_i(t)$, $i=1-3$, are deterministic functions, while B_i are mutually independent BMs associated with the arrival process, service times and patience times, respectively. The MSHT FCLT in Liu and Whitt [126] is for the $G_t/M/s_t+GI$ model; the extension to the $G_t/GI/s_t+GI$ model remains an open problem. The results in Aras *et al.* [6, 7] and Pang and Zhou [173] may help in that effort.

The direct QED^c TV MSHT limits lead to Gaussian distributions on both UL and OL intervals, but the resulting direct Gaussian approximations are not very accurate. That leads to a search for refinements, such as the Gaussian closure approximations proposed by Ko and Gautam [103] and Massey and Pender [145] and the truncated Gaussian approximations proposed in (1.2) of Liu and Whitt [126] and examined more thoroughly in Liu *et al.* [130] and extended in Liu [120].

A special case of the QED^c MSHT regime is simply having the overloaded or ED MSHT regime. As demonstrated in Whitt [218, 221] and Liu *et al.* [130], the ED regime is practically relevant even for stationary models. Even in the ED region, it may be important to consider refined staffing as in (4.17) from Liu [120]; see Mandelbaum and Zeltyn [141], Section 10 of Liu and Whitt [126] and Section 6 of Aras *et al.* [6] for other instances.

A complication when we switch between OL and UL intervals is the behavior near the switching points. That is addressed directly in Liu and Whitt [126] and in the limits for the initial content process in Aras *et al.* [7].

It remains to be done for general $G_t / GI / s_t + GI$ models. It also remains to see if new perspectives will yield even better understanding and more effective algorithms.

7. The Time-Varying Single-Server Queue

There is a substantial literature on TV single-server queues: structural results (e.g., definition and existence of processes), as in Harrison and Lemoine [74], Heyman and Whitt [78], Lemoine [109, 110], and Rolski [190, 191, 192], numerical algorithms, as in Section 2, and asymptotic methods and approximations, as in May and Keller [153], Newell [159, 160, 161, 162], Keller [93], Massey [142], Mandelbaum and Massey [139], and Whitt [224, 226].

In Section 7.1 we first review a convenient representation for the workload process introduced by Lemoine [109] that separates all stochastic variability from the deterministic variability of the arrival-rate function. Then in Section 7.2 we review the heavy-traffic (HT) limit for the periodic $G_t / G / 1$ queue established in Whitt [224]. Paralleling the stationary case, the limit is a reflected periodic Brownian motion (RPBM). Since that RPBM limit process is not so tractable, we review further numerical methods and approximations. In Section 7.3 we review the rare-event simulation algorithm developed in Ma and Whitt [136] and in Section 7.4 we review the time-varying robust queueing (TVRQ) from Whitt and You [227]. We conclude in Section 7.5 by reviewing service-rate controls for stabilizing performance in the TV single-server queue from Whitt [225].

7.1. The extended Lemoine representation of the workload

We start by reviewing a convenient representation of the workload process for the $G_t / G / 1$ queue, first discovered for the periodic $M_t / G / 1$ queue by Lemoine [109].

A reverse-time construction. We assume that the system starts out empty at time 0. Let $\{(U_k, V_k)\}$ be the sequence of ordered pairs of nonnegative random variables representing the interarrival times and service times. Let an arrival counting process be defined on the positive halfline by $A(t) \equiv \max\{k \geq 1 : U_1 + \cdots + U_k \leq t\}$ for $t \geq U_1$ and $A(t) \equiv 0$ for $0 \leq t < U_1$, and let the total input of work over the interval $[0,t]$ be the random sum

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0, \tag{7.1}$$

Then the workload (the remaining work in service time) at time t, starting empty at time 0, can be represented using the reflection map as $W(t) = \Psi(Y - e)(t)$, where e is the identity map, i.e., $e(t) \equiv t$, $t \geq 0$. Hence,

$$\begin{aligned} W(t) = \Psi(Y-e)(t) &\equiv Y(t) - t - \inf_{0 \leq s \leq t}\{Y(s) - s\} \\ &= \sup_{0 \leq s \leq t}\{Y(t) - Y(s) - (t-s)\} = \sup_{0 \leq s \leq t}\{Y_t(s) - s\}, \quad t \geq 0, \end{aligned} \tag{7.2}$$

where

$$Y_t(s) \equiv Y(t) - Y(t-s) = \sum_{k=A(t-s)+1}^{A(t)} V_k, \quad 0 \leq s \leq t, \quad t \geq 0, \tag{7.3}$$

is the cumulative input over the interval $(t-s, t]$.

Exploiting the composition construction from Section 5.3. We now exploit the composition construction of the G_t arrival process in Section 5.3. Given that additional model structure, we have

$$\{Y_t(s) : 0 \leq s \leq t\} \overset{\mathrm{d}}{=} \left\{ \sum_{k=1}^{N(\Lambda_t(s))} V_k : 0 \leq s \leq t \right\} \quad \text{for all} \quad t \geq 0, \tag{7.4}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution, which here in (7.4) we mean as stochastic processes, and

$$\Lambda_t(s) \equiv \Lambda(t) - \Lambda(t-s), \quad 0 \leq s \leq t, \quad t \geq 0. \tag{7.5}$$

Assuming a positive bounded arrival-rate function as in (5.3) , the function $\Lambda_t(s)$ in (7.5) is strictly increasing and continuous as a function of s with $\Lambda_t(0) = 0$ for each t, so it has a continuous strictly increasing inverse $\Lambda_t^{-1}(s)$ as a function of s with $\Lambda_t(0) = 0$ for each t.

Hence, we can combine (7.2) and (7.4) to obtain the alternative representation of the workload as

$$W(t) = \sup_{0 \leq s \leq t}\left\{ \sum_{k=1}^{N(\Lambda_t(s))} V_k - s \right\} = \sup_{0 \leq s \leq \Lambda(t)}\left\{ \sum_{k=1}^{N(s)} V_k - \Lambda_t^{-1}(s) \right\}, \tag{7.6}$$

where $\Lambda_t(s)$ is defined in (7.5). The second expression in (7.6) is the Lemoine [109] representation, introduced for the $M_t/G/1$ model. The extended Lemoine representation of the workload for the $G_t/G/1$ queue here is appealing because it has all the stochastic variability in the first term within the supremum but all determinstic variability in the arrival-rate function in the second term within the supremum.

The periodic $G_t/G/1$ queue. If the arrival process and workload process are periodic over the entire real line with period c, then we can obtain an expression for the periodic steady-state workload at time t within the interval $[0,c)$ by letting the system start empty in the distant past. For this periodic steady-state distribution to be well defined, we require that the average arrival rate satisfy

$$\rho = \bar{\lambda} \equiv \Lambda(c)/c < 1, \qquad (7.7)$$

to ensure that the average arrival rate is less than the maximum possible service rate $\mu \equiv 1/E[V] \equiv 1$. We assume that a proper periodic steady-state exists.

Instead of (7.2) for the transient workload, we have the periodic steady-state workload represented as a supremum over the entire real line. In particular, for a fixed position y within a cycle, we have

$$W_y = \sup_{s \ge 0} \left\{ Y_y(s) - s \right\}, \quad 0 \le y < c, \qquad (7.8)$$

where Y_y is defined as in (7.3) .

7.2. Heavy-traffic limits for periodic single-server queues

The seminal papers on HT approximations for the TV single-server queue are Newell [160, 161, 162]. Even though HT limits are not actually discussed, the diffusion approximations discussed there can be obtained via HT limits. Key scaling properties are presented directly. Direct HT limits for the $M_t/M_t/1$ queue were then obtained by Massey [142] and Mandelbaum and Massey [139]. A new perspective on one case is provided by Whitt [226]. That paper shows that there are more possibilities for the scaling.

Here we discuss HT limits for periodic queues. In particular, we review HT limits for the periodic $G_t/GI/1$ queue from Whitt [224]. The HT limit was stated for the queue length in Whitt [224] and then extended to the workload process in Ma and Whitt [136]. A previous HT limit for the $M_t/GI/1$ model was established by Falin [47], but it produced the same limit as for the corresponding $M/GI/1$ model. The key idea in Whitt [224] for obtaining useful new results to expose the TV behavior is to introduce a new HT scaling of the arrival-rate function.

Another key idea in Whitt [224] is to simplify the proof by focusing only on the first-order behavior, in particular, by assuming that the fluid limit is not TV when focusing on the TV FCLT. That is a great simplification over Massey [142] and Mandelbaum and Massey [139], but with that simplification, it is possible to apply the early HT limit for the s-server queue in Theorem 1(a) of Iglehart and Whitt [84]. For the single-server special case, it suffices to apply the HT FCLTs in Theorems 9.3 and 9.4 of Whitt [214]. In either case, these theorems state that a HT FCLT holds jointly for all the standard queueing processes if a FCLT holds jointly for the arrival and service processes. For the case considered here of i.i.d service times, independent of the arrival process, the FCLT for the service process is just Donsker's theorem, so that it suffices to establish a FCLT for the arrival process. Most important, that simplification produces useful results.

The new HT limit for the arrival process depends on a new family of scaled arrival-rate functions, indexed by ρ in (7.7). To avoid notational confusion, we add a superscript d to the diffusion quantities. Given ρ and the limiting cumulative arrival rate function $\Lambda_\gamma^d(t)$ for a periodic arrival-rate function $\lambda_\gamma^d(t)$ with period $1/\gamma$, we let the cumulative arrival-rate function in model ρ be

$$\Lambda_{\gamma,\rho}(t) \equiv \rho t + (1-\rho)^{-1}\Lambda_\gamma^d((1-\rho)^2 t), \quad t \ge 0, \tag{7.9}$$

so that the associated arrival-rate function is

$$\lambda_{\gamma,\rho}(t) \equiv \rho + (1-\rho)\lambda_\gamma^d((1-\rho)^2 t), \quad t \ge 0, \tag{7.10}$$

where

$$\Lambda_\gamma^d(t) \equiv \int_0^t \lambda_\gamma^d(s)ds, \quad \lambda_\gamma^d(t) \equiv h(\gamma t), \quad \text{and} \quad \int_0^1 h(t)dt = 0 \tag{7.11}$$

with $h(t)$ being a periodic function with period 1. Alternatively (more generally), we can assume that

$$\hat{\Lambda}_{\rho,\gamma}(t) \equiv (1-\rho)[\Lambda_{\rho,\gamma}((1-\rho)^{-2}t) - (1-\rho)^{-2}\rho t] \Rightarrow \Lambda_\gamma^d(t) \tag{7.12}$$

uniformly over bounded time intervals.

As a consequence, $\lambda_\gamma^d(t)$ is a periodic function with period $c_\gamma = 1/\gamma$ and $\lambda_{\gamma,\rho}(t)$ is a periodic function with period $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$. To ensure that $\lambda_{\gamma,\rho}$ is nonnegative, we assume that

$$h(t) \ge -\rho/(1-\rho), \quad 0 \le t < 1, \tag{7.13}$$

which will be satisfied for all ρ sufficiently close to the critical value 1 provided that h is bounded below. In fact, we directly assume that

$$-\infty < h^\downarrow \equiv \inf_{0\le t\le 1}\{h(t)\} < \sup_{0\le t\le 1}\{h(t)\} \equiv h^\uparrow < \infty. \tag{7.14}$$

There are two primary cases of interest $h^\uparrow < 1$ and $h^\uparrow > 1$. When $h^\uparrow < 1$, the instantaneous traffic intensity, which is $\lambda_{\gamma,\rho}(t)$, satisfies $\lambda_{\gamma,\rho}(t) < 1$ for all t and ρ. On the other hand, when $h^\uparrow > 1$, $\lambda_{\gamma,\rho}(t) > 1$ for some t. When $\lambda_{\gamma,\rho}(t) > 1$ for some t, the workload can reach very high values when time is scaled, because the cycles are very long. That takes us into the setting of a slowly changing random environment in Choudhury *et al.* [29], to which we refer for additional discussion.

Theorem 3.2 of Whitt [224] and Theorem 2 of Ma and Whitt [136] provide a heavy-traffic limit as $\rho \uparrow 1$ for the arrival process, the queue-length and workload processes at time t starting empty at time 0, but we will focus only on the arrival and workload processes. Let $W_{\gamma,\rho}(t)$ denote the workload process starting empty, but it also applies to the periodic steady-state distribution except for the usual problem of interchanging the order of the limits as $\rho \uparrow 1$ and as $t \uparrow \infty$. We use the periodic steady-state of the limit to approximate the periodic steady-state of the periodic $G_t/GI/1$ queue.

To express the heavy-traffic limits, we use (7.9) and let

$$A_{\gamma,\rho}(t) \equiv N\left(\Lambda_{\gamma,\rho}(t)\right), \quad Y_{\gamma,\rho}(t) \equiv \sum_{k=1}^{A_{\gamma,\rho}(t)} V_k, \quad \text{and} \quad X_{\gamma,\rho}(t) \equiv Y_{\gamma,\rho}(t) - t, \quad t \geq 0. \quad (7.15)$$

Then $X_{\gamma,\rho}(t)$ is the net-input process and $W_{\gamma,\rho}(t)$ is the workload process, which is the image of $X_{\gamma,\rho}$ under the reflection map Ψ, i.e.,

$$W_{\gamma,\rho}(t) = \Psi(X_{\gamma,\rho})(t) = \sup_{0\le s\le t}\{X_{\gamma,\rho}(t) - X_{\gamma,\rho}(t-s)\}. \quad (7.16)$$

For the heavy-traffic functional central limit theorem (FCLT), we introduce the scaled processes

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt], \quad \hat{A}_{\gamma,\rho}(t) \equiv (1-\rho)[A_{\gamma,\rho}\left((1-\rho)^{-2}t\right) - (1-\rho)^2 t],$$
$$\hat{X}_{\gamma,\rho}(t) \equiv (1-\rho)X_{\gamma,\rho}\left((1-\rho)^{-2}t\right) \quad \text{and} \quad \hat{W}_{\gamma,\rho}(t) \equiv (1-\rho)W_{\gamma,\rho}\left((1-\rho)^{-2}t\right), \quad (7.17)$$

for $t \geq 0$.

Let D^k be the k-fold product space of the function space D. Again let e be the identity map in D, i.e., $e(t) \equiv t$, $t \geq 0$.

Theorem 7.1. (*HT FCLT from Theorem 3.2 of Whitt* [224] *and Theorem 2 of Ma and Whitt* [136]) *For the family of* $G_t/GI/1$ *models indexed by* (γ,ρ) *with cumulative arrival-rate functions in* (7.9) *(or in* (7.12)*) and scaled processes in* (7.17)*, if* $\hat{N}_n \Rightarrow c_a B_a$ *as* $n \to \infty$*, where* B_a *is a standard Brownian motion, then*

$$(\hat{A}_{\gamma,\rho}, \hat{X}_{\gamma,\rho}, \hat{W}_{\gamma,\rho}) \Rightarrow (\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \quad \text{in} \quad D \quad \text{as} \quad \rho \uparrow 1, \quad (7.18)$$

where

$$(\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \equiv (c_a B_a + \Lambda_\gamma^d - e, \hat{A}_\gamma + c_s B_s, \Psi(\hat{X}_\gamma)), \quad (7.19)$$

Ψ *is the reflection map in* (7.16)*,* Λ_γ^d *is defined in* (7.11)*, and* B_a *and* B_s *are two independent standard (mean 0 variance 1) Brownian motions; i.e.,* $\hat{W}_\gamma$ *is reflected periodic Brownian motion (RPBM) with*

$$\hat{W}_\gamma = \Psi(c_a B_a + c_s B_s + \Lambda_{d,\gamma} - e) \overset{\mathrm{d}}{=} \Psi(c_x B + \Lambda_{d,\gamma} - e), \quad (7.20)$$

where $c_x^2 = c_a^2 + c_s^2$.

Unfortunately, there evidently are no available performance formulas or algorithms for the RPBM limit in Theorem 7.1. The next two sections present new approaches to remedy that deficiency. We present numerical algorithms to calculate the, exact or approximate, distribution of the periodic steady-state workload in the periodic $G_t/G/1$, which can be applied via the HT limit to compute the distribution of RPBM.

7.3. A rare-event simulation algorithm for the $GI_t/GI/1$ *queue*

Ma and Whitt [136] have shown that the classic rare-event simulation algorithm to efficiently compute the tail probabilities $P(W > b)$ for large b, where W is the steady-state waiting time in the $GI/GI/1$ queue, can be extended to the periodic steady state workload at each place within the periodic cycle in the associated periodic $GI_t/GI/1$ queue, provided that we use the compositon construction of the arrival process in Section 5.3. Just as for the $GI/GI/1$ queue, the algorithm exploits importance sampling using an exponential change of measure, as in Chapter XIII of Asmussen [9] and Chapter VI of Asmussen and Glynn [10].

Moreover, Ma and Whitt [136] show that, by exploiting the HT scaling in Section 7.2, the algorithm can be exploited to compute the tail probabilities and moments of RPBM. By exploiting HT approximations, that can be used to obtain approximations for more general periodic $G_t/G/1$ queues.

We will not review the detailed algorithm, instead referring to Ma and Whitt [136], but we will review the key representation that links the periodic $GI_t/GI/1$ model to the stationary $GI/GI/1$ model. We will also illustrate how the algorithm can be used to compute the performance of RPBM by showing simulation results for several models with ρ increasing toward 1, provided that we exploit the HT scaling.

Based on the Lemoine representation for the TV workload in (7.6), we can obtain a convenient representation of the periodic steady-state workload. Let W_y be defined in terms of the underlying stationary process N and the associated sequence of service times $\{V_k : k \geq 1\}$ via

$$W_y \overset{\mathrm{d}}{=} \sup_{s \geq 0}\{\sum_{k=1}^{N(s)} V_k - \tilde{\Lambda}_y^{-1}(s)\}, \quad 0 \leq y < 1, \tag{7.21}$$

where

$$\tilde{\Lambda}_y(t) \equiv \Lambda(yc) - \Lambda(yc - t), \quad t \geq 0, \tag{7.22}$$

is the *reverse-time cumulative arrival-rate function* starting at time yc within the periodic cycle $[0, c]$, $0 \leq y < 1$, and $\tilde{\Lambda}_y^{-1}$ is its inverse function, which is well defined because $\tilde{\Lambda}_y(t)$ is continuous and strictly increasing.

From the representation in (7.21), it is evident that from each sample path of the underlying stochastic process (N, V), we can generate a realization of W_y in (7.21) for each y, $0 \le y < 1$, by just changing the deterministic function $\tilde{\Lambda}_y^{-1}$. Moreover, from the rare-event construction, we can simultaneously obtain an estimate of $P(W_y > b)$ for all b in the bounded interval $[0, b_0]$ while applying the estimation for the single value b_0. Thus, we can essentially obtain estimates for all *performance parameter pairs* $(y, b) \in [0,1) \times [0, b_0]$ while doing the estimation for only one pair. This efficiency is very useful to conduct simulation studies to expose the way that $P(W_y > b)$ and the other performance measures depend on (y, b).

Bounds on the difference between the periodic and stationary workloads. Fundamentally, the reason that it is possible to exploit the $GI/GI/1$ rare-event algorithm to create a new rare-event algorithm for the periodic $GI_t/GI/1$ queue is because it is possible to bound the difference between the two random quantities. We review that bound now.

We compare the periodic steady-state workload W_y in (7.21) and the associated stationary workload W defined as in (7.21) with $\rho^{-1}s$ replacing $\tilde{\Lambda}_y^{-1}(s)$:

$$W \stackrel{\mathrm{d}}{=} \sup_{s \ge 0} \{ \sum_{k=1}^{N(s)} V_k - \rho^{-1} s \}, \tag{7.23}$$

Note that in both (7.21) and (7.23), N is understood to be a stationary point process. In particular, for the $GI_t/GI/1$ model, N is an equilibrium renewal process with the first inter-renewal time having the equilibrum distribution, therefore W is the stationary workload in the associated $GI/GI/1$ model, which may differ from the stationary waiting time in the same model. We now show that we can bound W_y above and below by a constant difference from the stationary workload W by rewriting (7.21) as

$$W_y = \sup_{s \ge 0} \{ \sum_{k=1}^{N(s)} V_k - \rho^{-1} s - (\tilde{\Lambda}_y^{-1}(s) - \rho^{-1} s) \}. \tag{7.24}$$

From (7.24), we immediately obtain the following lemma.

Proposition 7.1. (*upper and lower bounds on* W_y) *If we construct* W_y *in* (7.21) *and* W *in* (7.23) *using the same service times and base arrival process* N *in the composition construction of the arrival process, then*

$$W_y^- \equiv W - \zeta_y^- \le W_y \le W - \zeta_y^+ \equiv W_y^+ \tag{7.25}$$

where

$$\zeta_y^- \equiv \sup_{0 \le s \le \rho c} \{ \tilde{\Lambda}_y^{-1}(s) - \rho^{-1} s \} = -\rho^{-1} \inf_{0 \le s \le c} \{ \tilde{\Lambda}_y(s) - \rho s \} \ge 0 \quad \text{and}$$

$$\zeta_y^+ \equiv \inf_{0 \le s \le \rho c} \{ \tilde{\Lambda}_y^{-1}(s) - \rho^{-1} s \} = -\rho^{-1} \sup_{0 \le s \le c} \{ \tilde{\Lambda}_y(s) - \rho s \} \le 0 \tag{7.26}$$

Note that the supremum and infimum in (7.26) are over the interval $[0, \rho c]$. Because the average arrival rate is ρ, $\tilde{\Lambda}_y(c) = \Lambda(c) = \rho c$ and thus $\tilde{\Lambda}_y^{-1}(\rho c) = c$. Given that Λ is continuous and strictly increasing, we can use properties of the inverse function as in Section 13.6 of Whitt [214] to determine the alternative second representation of the bounds.

Unified numerical results via heavy-traffic scaling. To produce unified numerical results, we scale the arrival rate function so that the performance measures have heavy-traffic limits as $\rho \uparrow 1$, using the framework in Section 7.2. For the special case of a sinusoidal arrival rate function, we let

$$\lambda_\rho(t) = \rho + (1-\rho)\rho\beta \sin(\gamma(1-\rho)^2 t), \quad t \geq 0, \tag{7.27}$$

so that the cycle length in model ρ is $c_\rho = c^*(1-\rho)^{-2} = 2\pi/(\gamma(1-\rho)^2)$. After scaling, the cycle length is $c^* = 2\pi/\gamma$.

When we consider the periodic steady-state workload, we include spatial scaling by $1-\rho$, so that we consider $P(W_y > b_\rho)$, where $b_\rho = b/(1-\rho)$. Hence, to have asymptotically convergent models, we should choose parameter four-tuples $(\bar{\lambda}_\rho, \beta_\rho, \gamma_\rho, b_\rho)$ indexed by ρ, where

$$(\bar{\lambda}_\rho, \beta_\rho, \gamma_\rho, b_\rho) = (\rho, (1-\rho)\beta, (1-\rho)^2\gamma, (1-\rho)^{-1}b), \tag{7.28}$$

where (β, γ, b) is a feasible base triple of positive constants with $\beta < 1$. (We must constrain $\beta_\rho \leq 1$ so that $\lambda_\rho(t) \geq 0$ for all t.) Hence, we have the ρ-dependent constraint $\rho_b = (1-\rho)\beta \leq 1$. There is no problem if $\beta \leq 1$, but we may want to consider $\beta > 1$. In that case, β_ρ is only well defined for $\rho \geq 1-(1/\beta)$. For example, if $\beta = 5.0$, then we require that $\rho \geq 0.8$.

Example 7.1. (*Using $M_t/M/1$ to estimate the performance of the tail probabilities*) To illustrate how we can apply simulations of the $M_t/M/1$ model with increasing traffic intensities, let the base parameter triple be $(\beta, \gamma, b) = (1.0, 2.5, 4.0)$. Then the parameter 4-tuple for $\rho = 0.8$ is

$$(\bar{\lambda}_\rho, \beta_\rho, \gamma_\rho, b_\rho) = (0.8, (1-0.8)\beta, (1-0.8)^2\gamma, (1-0.8)^{-1}b) = (0.8, 0.2, 0.1, 20.0). \tag{7.29}$$

The associated parameter 4 -tuple for $\rho = 0.9$ is $(0.90, 0.10, 0.025, 40.00)$.

Let W be the steady-state workload in the stationary $M/M/1$ model with the same scaling, which has an exponential distribution except for an atom $1-\rho$ at the origin. Table 1 shows estimates of the ratio $P(W_y > b_\rho)/P(W > b_\rho)$ for 5 different values of $1-\rho$, where we successively divide $1-\rho$ by 2 and 8 different values of the position y within the cycle in the $M_t/M/1$ model with sinusoidal arrival-rate function in (7.27) with the parameter 4-tuple in (7.28) using the base parameter triple $(\beta, \gamma, b) = (1.0, 2.5, 4.0)$. (The paramter 4-tuple for $\rho = 0.8$ and $\rho = 0.9$ are shown above.)

Table 1. Comparison of the ratios $P(W_y > b_\rho) / P(W > b_\rho)$, where W is for the stationary model, for 5 different values of $1-\rho$ and 8 different values of the position y within the cycle in the $M_t/M/1$ model with sinusoidal arrival-rate function in (7.27) with the parameter 4-tuple in (7.28) using the base parameter triple $(\beta, \gamma, b) = (1.0, 2.5, 4.0)$.

y	$1-\rho = 0.16$	$1-\rho = 0.08$	$1-\rho = 0.04$	$1-\rho = 0.02$	$1-\rho = 0.01$
0.000	0.96364	0.96523	0.96424	0.96357	0.96344
0.125	0.97619	0.97686	0.97504	0.97493	0.97482
0.250	1.00456	1.00450	1.00255	1.00251	1.00305
0.375	1.03278	1.03264	1.03035	1.03152	1.03152
0.500	1.04565	1.04470	1.04278	1.04346	1.04405
0.625	1.03213	1.03096	1.03230	1.03150	1.03204
0.750	1.00225	1.00404	1.00425	1.00277	1.00241
0.875	0.97371	0.97696	0.97629	0.97457	0.97545
avg diff	0.00037	0.00112	0.00015	-0.00019	
avg. abs. dif	0.00099	0.00121	0.00081	0.00039	
rmse	0.00116	0.00134	0.00096	0.00049	

Table 1 shows that, for each fixed y, all estimates as a function of ρ serve as reasonable practical approximations for the others as well as for the RPBM limit developed in Section 7.2. The convergence in Table 1 is summarized by showing the average difference, average absolute difference and root mean square error (rmse) of the entry with the corresponding estimate for $\rho = 0.99$ in the final column, taken over 40 evenly spaced values of y in the interval $[0,1)$.

The mean and variance. The tail-integral representations of the mean and higher moments on p. 150 of Feller [49] can be exploited to obtain corresponding rare-event simulations of these related quantities. Recall that, for any nonnegative random variable X, the mean can be expressed as

$$E[X] = \int_0^\infty P(X > t)\,dt, \tag{7.30}$$

while the corresponding representation of the p^{th} moment for any $p > 1$ is

$$E[X^p] = \int_0^\infty p t^{p-1} P(X > t)\,dt. \tag{7.31}$$

To obtain a finite algorithm, it is natural to approximate the integrals for the mean and the second moment by finite sums plus a tail approximation, i.e.,

$$\begin{aligned} E[W_y] &\approx \sum_{k=0}^{n} (P(W_y > k\delta)\delta) + \frac{P(W_y > n\delta)}{\theta^*} \\ E[W_y^2] &\approx \sum_{k=0}^{n} (2P(W_y > k\delta)k\delta) + 2P(W_y > n\delta)\left(\frac{n\delta}{\theta^*} + \frac{1}{(\theta^*)^2}\right). \end{aligned} \tag{7.32}$$

In each case, the second term is based on applying the tail integral formula over $[n\delta,\infty)$ with the approximation

$$P(W_y > n\delta + x) \approx P(W_y > n\delta) e^{-\theta^* x} \tag{7.33}$$

and integrating.

We now illustrate the application of the rare-event simulation algorithm to estimate the mean and standard deviation of W_y and then for RPBM.

Example 7.2. (*Using $M_t/M/1$ to estimate the mean and standard deviation*) Table 2 show estimates of the time varying mean $E[W_y]$ and standard deviation $SD(W_y)$ for the special case of $y = 0.5$ for associated $M_t/M/1$ model with the sinusoidal arrival-rate function for base parameter pair $(\beta,\gamma) = (4, 2.5)$ using the scaling convention in (7.28). The cycle length is $2\pi/\gamma$, which equals $6.28/0.1 = 62.8$ for $\rho = 0.8$. The higher relative amplitude of $\beta = 4$ in Table 2 leads to much larger mean values at $y = 0.5$ than for $\beta = 1$; the cycle midpoint $y = 0.5$ tends to produce the largest values in the cycle.

Table 2. Estimated mean $E[W_y]$ and standard deviation $SD(W_y)$ as a function of $1-\rho$ for five cases of the $M_t/M/1$ queue at $y = 0.5$: $\mu = 1$, $\bar{\lambda} = \rho$ and base parameter pair $(\beta,\gamma) = (4, 2.5)$ having larger relative amplitude

$1-\rho$	0.16	0.08	0.04	0.02	0.01
n_s in (7.32)	40,000	40,000	40,000	40,000	40,000
δ in (7.32)	0.001	0.001	0.001	0.001	0.001
largest b	41	86	173	345	691
$P(W_y > 0)$	0.9728	0.9883	0.9967	0.9965	0.9993
s.e. of $P(W_y > 0)$	3.61E-03	2.69E-03	2.05E-03	1.16E-03	8.52E-04
95% CI of $P(W_y > 0)$	[0.9657, 0.9799]	[0.9831, 0.9936]	[0.9927, 1.0000]	[0.9943, 0.9988]	[0.9976, 1.0000]
$E[W_y]$	15.148	33.583	70.677	145.183	294.222
std of $E[W_y]$	5.58E-02	1.13E-01	2.27E-01	4.59E-01	9.15E-01
95% CI of $E[W_y]$	[15.04, 15.26]	[33.36, 33.81]	[70.23, 71.12]	[144.3, 146.1]	[292.4, 296.0]
$E[W_y \mid W_y > 0]$	15.572	33.980	70.909	145.690	294.437
95% CI of $E[W_y \mid W_y > 0]$	[15.35, 15.80]	[33.58, 34.39]	[70.2, 71.6]	[144.5, 147.0]	[292.4, 296.7]
$E[W_y^2]$	331.868	1528.127	6547.951	27,092.17	110,239.9
std of $E[W_y^2]$	1.023	4.263	17.227	69.632	0.785
95% CI of $E[W_y^2]$	[329.9, 333.9]	[1519.8, 1536.5]	[6514, 6582]	[26,955, 27,228]	[109,691, 110,787]
$SD[W_y]$	10.119	20.007	39.405	77.551	153.861
$P(W_y > 0)/\rho$	1.1581	1.0743	1.0383	1.0169	1.0094
$(1-\rho)E[W_y \mid W_y > 0]$	2.4915	2.7184	2.8364	2.9138	2.9444
$(1-\rho)SD[W_y \mid W_y > 0]$	1.5892	1.5830	1.5704	1.5442	1.5371

The first row shows $1-\rho$, which decreases over successive columns. The next block of three rows shows the parameters for the approximating sums. For the estimates we also shows 95% confidence intervals. The last three rows show scaled estimates that should be converging to RPBM, showing accuracy suitable for engineering applications.

7.4. Time-varying robust queueing

We now review the recent robust optimization approach to approximating the performance in the TV $G_t/G/1$ single-server queue in Whitt and You [227]. Robust optimization is a relatively new approach to difficult stochastic models. As in Bertsimas *et al.* [17], Ben-Tal *et al.* [16], and Beyer and Sendhoff [20]; the main idea is to replace a difficult stochastic model by a tractable optimization problem. We replace an "average-case" expected value by a "worst-case" optimization, where stochastic process sample paths are constrained to belong to uncertainty sets. The paper Whitt and You [227] extends the previous paper, Whitt and You [228], which developed robust queueing (RQ) algorithms to approximate the expected steady-state waiting-time and workload in stationary single-server queues, aiming especially to capture the impact of dependence among interarrival times and service times. In turn that paper builds on the RQ formulation of Bandi *et al.* [13].

The TVRQ optimization problem performs the maximization in (7.2) subject to deterministic constraints placed on the input process $Y(t)$ in (7.1). These constraints convert the stochastic process $W(t)$ in (7.2) into a deterministic approximation as the solution of a deterministic optimization problem. In simulation experiments Whitt and You [227] compare this deterministic approximation to the mean $E[W(t)]$ estimated from multiple independent replications of the model. They show that the detailed structure of the objective function in the PRQ provides insight into the structure of the mean and quantiles of the periodic workload. Thus, they develop a promising new way to obtain new insight into the "physics" of TV single-server queues, paralleling Eick *et al.* [44] for many-server queues.

The general TVRQ formulation. In particular, to formulate the deterministic TVRQ approximation for the time-varying workload $W(t)$ for any $t > 0$, we let

$$W^*(t) \equiv \sup_{X_t \in U_t} \sup_{0 \le s \le t} \{X_t(s)\}, \tag{7.34}$$

where $X_t(s) \equiv Y_t(s) - s$ and U_t is the deterministic uncertainty set, i.e., the set of allowed sample paths $\{X_t(s) : 0 \le s \le t\}$, which we define as

$$\begin{aligned} U_t &\equiv \{X_t(s) \in \mathbb{R} : X_t(s) \le E[Y_t(s) - s] + bSD(Y_t(s) - s), \quad 0 \le s \le t\} \\ &= \{X_t(s) \in \mathbb{R} : X_t(s) \le E[Y_t(s)] - s + bSD(Y_t(s))), \quad 0 \le s \le t\}, \end{aligned} \tag{7.35}$$

With SD being the standard deviation. This uncertainty set requires that the sample path of the reverse-time net-input process $X_t(s) \equiv Y_t(s) - s$ remain within b standard deviations of its mean at all times s, $0 \le s \le t$.

To ensure a finite supremum, we assume that $E[Y(t)^2] < \infty$ for all t. Then, since $Y_t(s) \geq 0$ for all t and s, necessarily $0 \leq W^*(t) < \infty$ for all t. As a consequence, we have the final TVRQ optimization

$$W^*(t) = \sup_{0 \le s \le t} \sup_{X_t \in U_t} \{X_t(s)\} = \sup_{0 \le s \le t} \{E[Y_t(s)] - s + bSD(Y_t(s))\}, \quad t \geq 0. \quad (7.36)$$

The uncertainty set in (7.35) is a natural time-varying generalization of the uncertainty sets in our previous paper, which are similar to the ones used in Bandi *et al.* [13]. The idea is that the uncertainty set (7.35) can be based on a Gaussian approximation, which in turn is supported by central limit theorem (CLT) for $Y_t(s)$ under customary regularity conditions. Nevertheless, providing convincing support for this uncertainty set, even in the stationary setting, is somewhat complicated. Thus the choice may ultimately be justified by its utility, which is demonstrated by establishing connections to the performance of the original queueing model. We refer readers to Section EC.3 and Section EC.4 in Whitt and You [228] for further discussion of the motivation and justification.

Whitt and You [228] give alternative representations for the uncertainty set in (7.35) and the final TVRQ algorithm in (7.36) using indices of dispersion, as in Cox and Lewis [33]. Following Fendick and Whitt [50] and Whitt and You [228], the index of dispersion for work (IDW) is exploited. The IDW, denoted by $I_w(t)$, characterizes the variability of the total input of work $Y(t)$ over the time interval $[0,t]$, independent of its mean. The idea is the same as the squared coefficient of variance (scv, variance divided by the square of the mean), which represents the variability of a single random variable independent of scale.

For the base total input process $\tilde{Y}(t) \equiv \sum_{k=1}^{N(s)} V_k$, the IDW is defined as

$$I_w(t) \equiv \frac{Var(\tilde{Y}(t))}{E[V]E[\tilde{Y}(t)]}, \quad t \geq 0; \quad (7.37)$$

see (1) of Fendick and Whitt [50]. In our setting with mean-1 service times and a rate-1 base process N, the IDW becomes

$$I_w(t) \equiv \frac{Var(\tilde{Y}(t))}{E[\tilde{Y}(t)]} = \frac{Var(\tilde{Y}(t))}{t}, \quad t \geq 0, \quad (7.38)$$

which is just a scaled version of the variance function. For the $M/GI/1$ model, we have $I_w(t) = c_a^2 + c_s^2 = 1 + c_s^2$ with c_a^2 and c_s^2 being the coefficient of variation of the interarrival and service distribution, respectively. We assume that IDW as a function of time is bounded, which is to be anticipated.

As a consequence, the uncertainty set (7.35) in the TVRQ can be written as

$$\mathcal{U}_t = \left\{ X(t) : X(s) \le E\left[\sum_{k=1}^{N(s)} V_k\right] - \Lambda_t^{-1}(s) + b\sqrt{Var\left(\sum_{k=1}^{N(s)} V_k\right)}, 0 \le s \le \Lambda(t) \right\}$$
$$= \left\{ X(t) : X(s) \le s - \Lambda_t^{-1}(s) + b\sqrt{sI_w(s)}, 0 \le s \le \Lambda(t) \right\}$$
$$= \left\{ X(t) : X(s) \le \Lambda_t(s) - s + b\sqrt{\Lambda_t(s) I_w(\Lambda_t(s))}, 0 \le s \le t \right\}. \quad (7.39)$$

Combining (7.34) and (7.39), we have the tractable TVRQ optimization for the $G_t/G/1$ model

$$W^*(t) = \sup_{0 \le s \le t} \left\{ \Lambda_t(s) - s + b\sqrt{\Lambda_t(s) I_w(\Lambda_t(s))} \right\}, \quad t \ge 0, \quad (7.40)$$

with the final expression in (7.40) providing a convenient expression for a computational algorithm because $\Lambda_t(s)$ is usually readily available, whereas $\Lambda_t^{-1}(s)$ in the first expression may not be.

Periodic robust queueing (PRQ). To develop a periodic version of TVRQ, we start in the periodic setting of Section 7.1. For the periodic case, starting empty in the distant past, we consider $y \in [0, c)$. Then periodic RQ (PRQ) for the steady-state workload is just TVRQ in (7.36) except that s is allowed to range over the interval $[0, \infty)$ and that $Y_t(s)$ is replaced by $Y_y(t)$ to emphasize the focus on a fixed location in a cycle. As a consequence, we have the final PRQ optimization

$$W_y^* = \sup_{s \ge 0} \left\{ E[Y_y(s)] - s + bSD(Y_y(s)) \right\}$$
$$= \sup_{s \ge 0} \left\{ \Lambda_y(s) - s + b\sqrt{\Lambda_y(s) I_w(\Lambda_y(s))} \right\}, \quad 0 \le y < c. \quad (7.41)$$

For periodic queues, we establish long-cycle fluid limits for both the original queueing system and the PRQ approximation and we prove that those limits coincide. Simulation experiments show that PRQ can yield useful approximations.

Heavy-traffic limits for PRQ. HT limits are then established for PRQ in the setting of Section 7.2 and compared to Theorem 7.1. Again, we add a subscript y to indicate the place in the cycle. In particular, the workload at fixed place y within a cycle for a system which started empty and has run for t time units is

$$W_{\gamma,\rho,y}(t) \overset{d}{=} \sup_{0 \le s \le t} \left\{ \sum_{k=1}^{A_{\gamma,\rho,y}(t)} V_k - s \right\}, \quad (7.42)$$

where $A_{\gamma,\rho,y}(t) \equiv A_{\gamma,\rho}(y) - A_{\gamma,\rho}(y-t)$, $A_{\gamma,\rho}(t)$ is defined in (7.15) and V_k is a generic service time. in the $G_t/G/1$ setting above, we immediately get the PRQ optimization problem from (7.40) by replacing $\Lambda_t(s)$ with $\Lambda_{\gamma,y,\rho}(s)$

$$W^*_{\gamma,\rho,y} = \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - s + b\sqrt{\Lambda_{\gamma,\rho,y}(s)I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}\right\}. \qquad (7.43)$$

To express the heavy-traffic limit, we define two functions. The first function

$$f(t) \equiv -t + 2\sqrt{t} \qquad (7.44)$$

is a variant of the function to be optimized with the stationary $M / GI / 1$ model, as can be seen from Theorem 1 of Whitt and You [228]. The second function

$$g_{\gamma,\rho,y}(t) \equiv \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} h(s)ds \qquad (7.45)$$

is a periodic function, depending on h in (7.11), that captures the time-varying part of the arrival rate function. The period of $g_{\gamma,\rho,y}(t)$ is $4 / b^2c_x^2\gamma\rho$. When the arrival-rate function is constant, $g_{\gamma,\rho,y}(t) = 0$ because $h(t) = 0$.

Here is the heavy traffic limit for PRQ.

Theorem 7.2. (*Heavy traffic limit for PRQ*) *The heavy traffic limit of the PRQ problem in* (7.43) *for the* $G_t / G / 1$ *model is*

$$\lim_{\rho\uparrow 1}\frac{2}{b^2}\cdot\frac{2(1-\rho)}{\rho c_x^2}\cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0}\left\{f(t) + g_{\gamma,1,y}(t)\right\} \qquad (7.46)$$

for $f(t)$ in (7.44) and $g_{\gamma,\rho,y}(t)$ in (7.45).

Whitt and You [227] show that the HT limits can be usefully combined with the long-cycle perspective to obtain further insight into the TV behavior of periodic queues. They identify three HT cases that can be characterized via the limiting density over a cycle, $\lambda_\gamma^d(t) = h(\gamma t)$ in (7.11). Letting $h^\uparrow \equiv \sup_{t\geq 0}h(t)$, the three HT cases are: underloaded ($h^\uparrow < 1$), overloaded ($h^\uparrow > 1$) and critically loaded ($h^\uparrow = 1$).

They show that the HT limits for PRQ coincide with the HT limit of the pointwise-stationary approximation of the HT limit for the original model in both the underloaded HT and overloaded HT cases; see Theorems 5 and 6 in Whitt and You [227]. They find that the scaling in the critically loaded case agrees with the scaling in Whitt [226]. When the cycle lengths are allowed to grow, there is a great buildup of congestion over time in the overloaded case, but no buildup over time in the underloaded case. For queues in a random environment, the overloaded case is discussed in Choudhury *et al.* [29]. The critically loaded case is discussed in Whitt [226].

Example 7.3. (*a simulation evaluation of PRQ*) Figure 7 illustrates the performance of PRQ in an underloaded case by making comparisons to simulation estimates and the heavy-traffic limit in Theorem 5 of Whitt and You [227]. The model considered is $(H_2(4)_t / LN(1) / 1$

model with the sinusoidal arrival-rate function

$$\lambda_{\gamma,\rho}(t) \equiv \rho + (1-\rho)\beta \sin(2\pi\gamma(1-\rho)^2 t), \quad t \geq 0, \tag{7.47}$$

where $\beta \equiv 0.8$, which is a special underloaded case of (7.10) with $h^{\uparrow} = \beta = 0.8$ for h in (7.11).

In particular, Figure 7 compares three alternative expressions for the normalized mean workload $2(1-\rho)E[W_{\gamma,\rho,y}]/\rho$. The first expression is the solution of the PRQ where the workload $E[W_{\gamma,\rho,y}]$ is calculated from the PRQ optimization in (7.43); the second is the HT limit as $\rho \uparrow 1$ and $\gamma \to 0$ in Theorem 5 of Whitt and You [227]; and the third is the simulation estimate.

Together with the HT limit, the figure on the left shows the PRQ and simulation estimates for three parameter pairs $(\gamma,\rho):(0.7,10^{-2})$, $(0.9,10^{-3})$ and $(0.95,10^{-4})$. Figure 7 confirms Theorem 5 of Whitt and You [227] by showing that both PRQ and the simulation estimates converge to the HT limit as ρ increases (with γ decreasing). Only the simulation estimate for the case $\rho = 0.7$ is not close to the theoretical HT limit. These plots also show that PRQ captures the essential shape of the TV mean workload and can serve as a useful approximation for moderate traffic intensities.

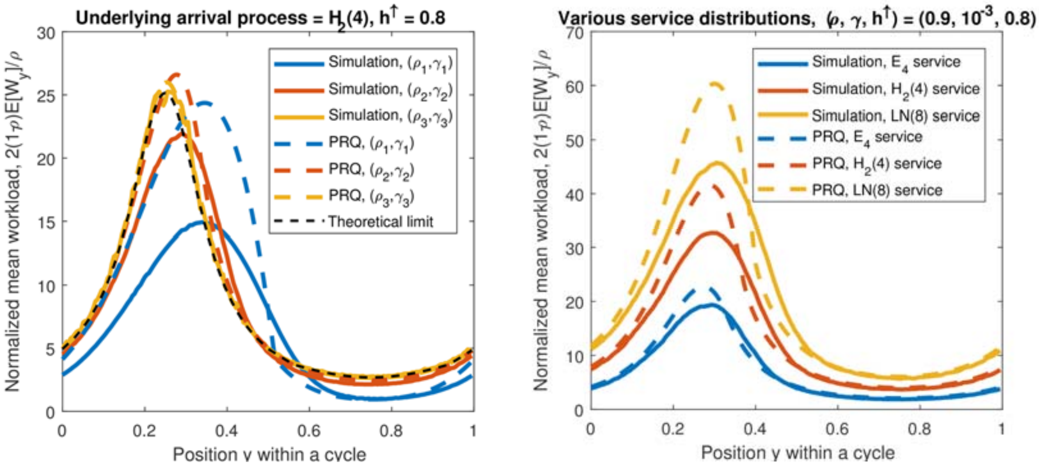

Figure 7. A comparison of the solution to the PRQ problem in (7.43) as a function of the position y within a cycle to simulation estimations and the HT limit in Theorem 5 of Whitt and You [227] for the normalized mean workload $2(1-\rho)E[W_{\gamma,\rho,y}]/\rho$ for $W_{\gamma,\rho,y}$ in (7.42) in the underloaded $(H_2(4)_t/LN(1)/1$ model with arrival-rate function in (7.47) for $(\gamma,\rho) \in \{(0.7,10^{-2}), (0.9,10^{-3}), (0.95,10^{-4})\}$ (left) and for three different service-time distributions (right).

Figure 7 (right) shows the impact of changing variability in the service-time distribution with comparisons between PRQ and simulation for three cases: Erlang E_4 with scv $c_s^2 = 1/4$, hyperexponential $H_2(4)$ with scv $c_s^2 = 4$ and lognormal $LN(8)$ with scv $c_s^2 = 8$.

Consistent with the stationary model, increased variability in the service process tends to increase congestion. However, Davis *et al.* [38] drew the opposite conclusion about the impact of the service-time distribution on the blocking in the time-varying $M_t / GI / n / 0$ loss model.

7.5. Service-rate controls to stabilize performance

As an analog of the staffing problem to stabilize performance in many-server queues with a TV arrival-rate function discussed in Section 4.3 and Section 4.4, we now review service-rate controls to stabilize performance in a single-server queue with a TV arrival-rate function developed in Whitt [225]. For this control, it is assumed that the service rate can be specified as a deterministic function separately from the random service requirements. For example, a customer service requirement might correspond to the size of a message to be transmitted in a communication network, while the service rate might be the processing rate of the message. Thus a service requirement S with a constant service rate μ would lead to a service time of S / μ. With this approach, all randomness appears through the service requirements, which are assumed to be stochastically independent of the arrival process.

Having a single-server queue where the service rate is a continuous deterministic function subject to control is an idealization of what occurs in many service operations, such as hospital surgery rooms and airport security inspection lines. Assigning more doctors and nurses can increase the rate of completed operations; assigning more inspection agents at the airport security line or relaxing the inspection requirements can increase the rate at which passengers are processed through inspection. In these applications, the possible service rate functions may not actually be continuous, or even fully under control. Nevertheless, to better understand the possible benefits of these practical service-rate controls, it is helpful to understand what controls are desirable in the ideal situation when any deterministic continuous service-rate control function is possible.

The model and its service times. We assume that the queue has unlimited waiting space with customers entering service in order of arrival. We let the arrival process be constructed from a rate-1 couting process N and arrival-rate function using the composition composition in Section 5.3. We assume that the $G / G / 1$ model with arrival process N, service-requirement sequence $\{S_k : k \geq 1\}$ and constant service rate $1 / \rho$ for $0 < \rho < 1$, as a number in system $X(t)$ that has a proper steady distribution. We then introduce a periodic arrival-rate function $\lambda(t)$ with average arrival rate $\bar{\lambda} = 1$.

In this setting we consider alternative service-rate controls. We assume that the service-rate control $\mu(t)$ is also a periodic function, with average rate $1 / \rho$, where ρ remains to be specified, subject to $0 < \rho < 1$. To construct the service time V_k determined by the service requirement sequence $\{S_k\}$ and the service-rate control $\mu(t)$, we need to properly relate rates to requirements and time. Assuming that the system starts empty and that B_k is the time customer k enters service, the service time V_k is specified implicitly via the equation

$$S_k = \int_{B_k}^{B_k+V_K} \mu(s)ds, \quad k \ge 1. \tag{7.48}$$

If we let

$$M(t) = \int_0^{Bt} \mu(s)ds, \quad t \ge 0, \tag{7.49}$$

then we see that $M(t)$ is the total amount of service completed in the interval $[0,t]$, assuming that the server is busy continuously. Since M is strictly increasing and continuous, it has an inverse M^{-1}. With that inverse, we obtain an explicit formula for the service times, in particular,

$$V_k = M^{-1}(S_k + M(B_k)) - B_k, \quad k \ge 1. \tag{7.50}$$

Finally, we remark that the Lemoine representation for the workload process in Section 7.1 that has been exploited in previous sections also extends directly to the current setting with time-varying service rate; see Remark 6 and Section Ec.4 of Whitt and You [227].

Three candidate service-rate controls. Whitt [225] considers three different service -rate controls, focusing mostly on the simple *rate-matching control*, which chooses the service rate to be proportional to the arrival rate; i.e., for a given target traffic intensity ρ, we let the service rate be

$$\mu(t) = \frac{\lambda(t)}{\rho}, \quad t \ge 0. \tag{7.51}$$

The rate-matching service-rate control in (7.51) obviously stabilizes the traffic intensity $\rho(t) = \lambda(t)/\mu(t)$ at ρ for all t.

The other two controls are square-root controls similar in spirit to the SRS staffing formulas in (4.4) and (4.9). The first is a variation of the Kleinrock [101] capacity allocation formula for open Jackson queueing networks in steady state, considered by Kleinrock [101], extended for approximations of generalized Jackson networks in Wein [210] and reviewed in Section 5.7 of Kleinrock [102], in Section 7 of Bitran and Dasu [21] and elsewhere. Here, instead of allocating capacity (which corresponds to service rate) to several queues in different locations, we allocate capacity to a single queue at different times. The *first square-root service-rate control* is

$$\mu(t) = \lambda(t) + \xi\sqrt{\lambda(t)}, \quad t \ge 0. \tag{7.52}$$

where ξ is a positive parameter; see Section 7.2 of Whitt [225] for additional discussion.

The *second square-root service-rate control* is

$$\mu(t) = \lambda(t) + \frac{\lambda(t)}{2}\left(\sqrt{1+\frac{\zeta}{\lambda(t)}} - 1\right), \quad t \geq 0, \tag{7.53}$$

where ζ is a positive parameter. It is obtained by assuming that the PSA is appropriate, i.e., that the system is approximately stationary at each time t. In particular, we directly assume that the expected time-varying workload at each time t can be approximated by

$$E[W(t)] \approx \frac{\rho(t)\eta}{\mu(t)(1-\rho(t))}, \quad t \geq 0, \tag{7.54}$$

where η is a variability parameter, as would be appropriate if the $GI_t / GI_t / 1$ model were in steady state at each time t with arrival rate $\lambda(t)$, service rate $\mu(t)$ and traffic intensity $\rho(t) = \lambda(t) / \mu(t) < 1$, where η is a variability parameter which we could take to be $\eta = (c_a^2 + c_s^2)/2$.

We then assume that the goal is to choose the service rate $\mu(t)$ to stabilize $E[W(t)]$ at the target w for all t. Thus, from (7.54), we have the equation

$$w = \frac{\lambda(t)\eta}{\mu(t)^2 - \mu(t)\lambda(t)}, \tag{7.55}$$

which leads to the quadratic equation in $x = \mu(t)$

$$wx^2 - \lambda(t)wx - \lambda(t)\eta = 0; \tag{7.56}$$

See Section 7.3 of Whitt [225] for additional discussion.

The performance of the candidate service-rate controls. Theorem 3.1 of Whitt [225] shows that the rate-matching service-rate control allows us to represent the number in system as a deterministic time transformation of the number in system in a stationary $G / G / 1$ model, implying that it has a proper steady-state distribution, showing that the rate-matching service rate control achieves its objective for the distribution of the number in system.

This good performance is illustrated by the plot on the left in Figure 8 showing simulation estimates of the time-varying mean number in the system for the $M_t / M_t / 1$ model with sinusoidal arrival rate function $\lambda(t) = 1 + \beta \sin \gamma t$ with $\beta = 0.2$ and $\gamma = 0.001$ with the rate-matching control in (7.51) for traffic intensity $\rho = 0.8$. However, the plot on the right in Figure 8 shows that the mean virtual waiting time is not stabilized at the same time with the rate-matching service-rate control.

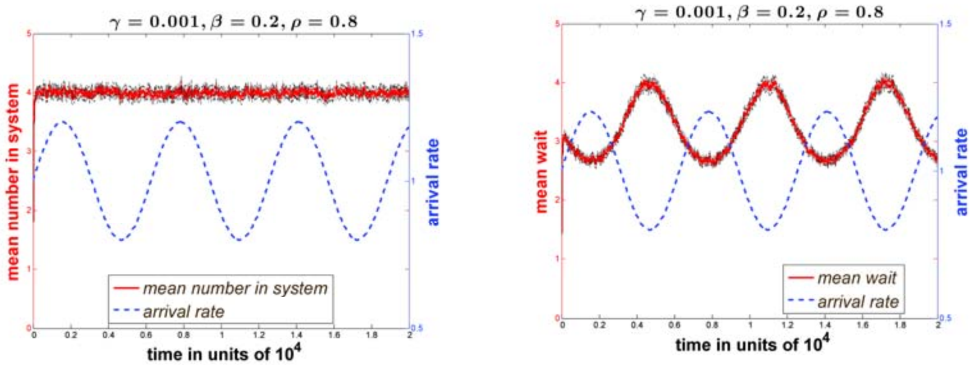

Figure 8. Simulation estimates of the time-varying mean number in the system (left) and the mean virtual waiting time (right), for the $M_t / M_t / 1$ model with sinusoidal arrival rate function $\lambda(t) = 1 + \beta \sin \gamma t$ with $\beta = 0.2$ and $\gamma = 0.001$ with the rate-matching control in (7.51) for traffic intensity $\rho = 0.8$.

Indeed, Theorem 5.2 of Whitt [225] establishes a HT limit for the number in system and virtual waiting time jointly with the rate-matching service-rate control, which implies a SP-TV-HT-LL just like the SP-TV-MSHT-LL in (6.11), showing that the virtual waiting time at each time t is inversely proportional to the arrival rate at time t.

On the other hand, the two square-root service-rate controls in (7.52) and (7.53) are less successful, even though the last in (7.53) does stabilize the mean virtual waiting time when the cycles are very long; e.g., see Figures 2 and 3 of Whitt [225]. The good performance for (7.53) with long cycles is as expected, because it is based on a PSA assumption.

Stabilizing the mean waiting time. Ma and Whitt [137] have shown that it is possible to stabilize the mean waiting time by using a modification of the service-rate control with a time lag and a damping factor. In particular, for arrival rate-functions of the form $\lambda(t) = \rho(1 + s(t))$, where $s(t)$ is periodic with average 0, they consider service-rate functions of the form $\mu(t) = 1 + \xi s(t - \eta)$. They develop a simulation search algorithm to locate the best control parameters (η, ξ). They establish HT limits for the model with these controls.

8. Conclusions

Service systems often can be modeled as many-server queues, at least roughly consistent with the $G_t / GI / s_t + GI$ model, when the arrival rate $\lambda(t)$ and the number of servers, s_t, are not small. When that is the case, we think that the first-order time-varying behavior can be captured by the two-parameter fluid model in Section 3.3 and the $M_t / GI / \infty$ infinite-server queueing model in Sections 4.1 and 4.2. Ways to probe more deeply are contained in Sections 2-6.

Time-varying single-server queues (or queues with relatively few servers) tend to behave differently from time-varying many-server queues. Even though single-server queues tend to have more elementary mathematical models, time-varying single-server queues tend to exhibit more complex behavior, because they allow a buildup of congestion that tends to take longer to dissipate. In a many-server queue, an overload period tends to end when the instantaneous traffic intensity falls below 1, which can be seen from asymptotic and approximate analysis of the QED^c MSHT regime in Section 2.3.2, Section 3.3 and Section 6; in a single-server queue there tends to be a longer recovery period, leading to the different parts of an overload incident, as exposed by Newell [160, 161, 162].

It appears that the first-order time-varying behavior of a single-server queue often can be captured by the methods of Section 3.2 and Section 7.4. To go further, Section 2 can play a key role, but more needs to be done.

We have not discussed networks of time-varying queues. We have mentioned the papers Massey and Whitt [147] and Liu and Whitt [121, 127, 128, 129] for many-server queues, but hardly anything has been done on networks of time-varying single-server queues or networks containing both many-server queues and single-server queues, which are natural models for hospitals and other large-scale service systems.

Acknowledgment

The author is grateful to his collaborators for their important contributions, to Michael R.Taaffe for helpful discussion about Section 2, to Yunan Liu and Xu Sun for helpful feedback, and to the National Science Foundation for support through grant NSF CMMI 1634133.

References

[1] Abate, J., & Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10, 5–88.

[2] Abate, J., & Whitt, W. (1992). Numerical inversion of probability generating functions. *Queueing Systems*, 12, 245–251.

[3] Abate, J., & Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing*, 7, 36–43.

[4] Abate, J., & Whitt, W. (1998). Calculating transient characteristics of the Erlang loss model by numerical transform inversion. *Stochastic Models*, 14, 663–680.

[5] 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.

[6] Aras, K., Chen, X., & Liu, Y. (2018). Many-server Gaussian limits for overloaded non Markovian queues with customer abandonment. *Queueing Systems,* 89, 81-125.

[7] Aras, V., Liu, Y., & Whitt, W. (2017). Heavy-traffic limit for the initial content process. *Stochastic Systems*, 7, 95–142.

[8] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., & Yom-Tov, G. (2015). Patient flow in hospitals: a data-based queueing-science perspective. *Stochastic Systems*, 5, 146–194.

[9] Asmussen, S. (2003). *Applied Probability and Queues*. Springer, New York, second edition.

[10] Asmussen, S., & Glynn, P. W. (2007). *Stochastic Simulation*. Springer, New York, second edition.

[11] Ata, B., Skaro, A., & Tayur, S. (2016). Organjet: Overcoming geographical disparities in access to deceased donor kidneys in the United States. *Management Science*, 62, 146–194.

[12] Avramidis, A. N., Deslauriers, A., & L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50, 896–908.

[13] Bandi, C., Bertsimas, D., & Youssef, N. (2015). Robust queueing theory. *Operations Research*, 63, 676–700.

[14] Bassamboo, A., & Randhawa, R. H. (2010). On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Res*earch, 58, 1398 – 1413.

[15] Bassamboo, A., & Zeevi, A. (2009). On a data-driven method for staffing large call centers. *Operations Research*, 57, 714–726.

[16] Ben-Tal, A., El-Ghaoui, L., & Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press, Princeton, NJ.

[17] Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, 53, 464–501.

[18] Bertsimas, D., & Mourtzinou, G. (1997). Transient laws of nonstationary queueing systems and their applications. *Queueing Systems*, 25, 315–359.

[19] Besbes, O., Phillips, R., & Zeevi, A. (2010). Testing the validity of a demand model: an operations perspective. *Manufacturing and Service Operations Management*, 12, 162–183, 2010.

[20] Beyer, H. G., & Sendhoff, B. (2007). Robust optimization - a comprehensive survey. *Computer Methods in Applied Mechanics and Engineering*, 196, 3190–3218.

[21] Bitran, G. R., & Dasu, S. (1992). A review of open queueing network models of manufacturing systems. *Queueing Systems*, 12, 95–134.

[22] Borovkov, A. A. (1967). On limit laws for service processes in multi-channel systems. *Siberian Mathathematical Journal*, 8, 746–763.

[23] Borst, S. C., Mandelbaum, A., & Reiman, M. (2004). Dimensioning large call centers. *Operations Research*, 52, 17–34.

[24] Brockmeyer, E., Halstrom, H. L., & Jensen, A. (1948). *The Life and Works of A. K. Erlang*. Academy of Technical Sciences, Copenhagen.

[25] 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 the American Statistical Association*, 100, 36–50, 2005.

[26] Browne, S., & Whitt, W. (1995). Piecewise-linear diffusion processes. In J. Dshalalow, editor, *Advances in Queueing*, 463–480, CRC Press, Boca Raton, FL, 1995.

[27] Choudhury, G. L., Lucantoni, D. M., & Whitt, W. (1994). Multi-dimensional transform inversion with applications to the transient M/G/1 queue. *Annals of Applied Probability*, 4, 719–740, 1994.

[28] Choudhury, G. L., Lucantoni, D. M., & Whitt, W. (1997). Numerical solution of piecewise-stationary $M_t / G_t / 1$ queues. *Operations Research*, 45, 451–463.

[29] Choudhury, G. L., Mandelbaum, A., Reiman, M. I., & Whitt, W. (1997). Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models*, 13, 121–146.

[30] Chung, K. L. (1960). Markov Chains with Stationary Transition Probabilities. Springer, Berlin.

[31] Clark, G. M. (1981). Use of Polya distributioons in approximate solutions to nonstationary $M / M / s$ queues. *Commununications of the ACM*, 27, 206–217.

[32] Clarke, A. B. (1956). A waiting line process of Markov type. *Annals of Mathematical Statistic*, 27, 452–459.

[33] Cox, D. R., & Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events*. Methuen, London.

[34] Daganzo, C. F. (1995). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research B: Methodologoical*, 28, 269–287.

[35] Dai, J. G., & He, S. (2012). Many-server queues with customer abandonment: a survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering*, 21, 1–36.

[36] Dai, J. G., & Shi, P. (2015). A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research*, 65, 514–536.

[37] Daley, D. J., & Vere-Jones, D. (2008). *An introduction to the theory of point processes*. Springer, Oxford, UK, second edition.

[38] Davis, J. L., Massey, W. A., & Whitt, W. (1995). Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Science*, 41, 1107–1116.

[39] Defraeye, M., & van Nieuwenhuyse, I. (2016). Staffing and scheduling under nonstationary demand for service: a literature review. *Omega*, 58, 4–25.

[40] Defraeye, M., & van Niewenhuyse, I. (2013). Controlling excessive waiting times in small service systems with time-varying demand: an extension of the ISA algorithm. *Decision Support Systems*, 54, 1558–1567.

[41] Duffield, N. G., Massey, W. A., & Whitt, W. (2001). A nonstationary offered-load model for packet networks. *Telecommunication Systems*, 13, 271–296.

[42] Durbin, J. (1961). Some methods for constructing exact tests. *Biometrika*, 48,41–55.

[43] Edie, L. C. (1954). Traffic delays at toll booths. *Operations Research*, 2, 107–138.

[44] Eick, S. G., Massey, W. A., & Whitt, W. (1993). The physics of the $M_t / G / \infty$ queue. *Operations Research*, 41,731–742.

[45] Eick, S. G., Massey, W. A., & Whitt, W. (1993). $M_t / G / \infty$ queues with sinusoidal arrival rates. *Management Scence*, 39, 241–252.

[46] El-Taha, M., & Stidham, S. (1999). *Sample-Path Analysis of Queueing Systems*. Kluwer, Boston.

[47] Falin, G. I. (1989). Periodic queues in heavy traffic. *Advances in Applied Probability*, 21, 485–487.

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

[49] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*. John Wiley, New York, second edition.

[50] 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*, 71, 171–194.

[51] Fischer, W., & Meier-Hellstern, K. (1992). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18, 149–171.

[52] Fralix, B. H., & Riano, G. (2010). A new look at transient versions of Little's law. *Journal of Applied Probability*, 47, 459–473.

[53] Gamarnik, D., & Goldberg, D. A. (2013). Steady-state $GI / GI / n$ queue in the Halfin-Whitt regime. *Annals of Applied Probability*, 23, 2382–2419.

[54] Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: tutorial, review and research prospects. *Manufacturing and Service Operations Management*, 5, 79–141.

[55] Garnett, O., Mandelbaum, A., & Reiman, M. I. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management*, 4, 208–227.

[56] Gebhardt, I., & Nelson, B. L. (2009). Transforming renewal processes for simulation of non-stationary arrival procsses. *INFORMS Journal on Computing*, 21, 630–640.

[57] Gebhardt, I., Nelson, B. L., & Taaffe, M. R. (2017). The $MAP_t / Ph_t / \infty$ queueing system and and multiclass $\left[MAP_t / Ph_t / \infty\right]^k$ queueing network. *INFORMS Journal on Computing*, 29, 367–376.

[58] Glynn, P. W., & Whitt, W. (1986). A central-limit-theorem version of $L = \lambda W$. *Queueing Systems*, 2, 191–215. (See Correction Note on $L = \lambda W$, *Queueing Systems*, 12, 1992, 431-432. The results are correct; minor but important change needed in proofs.).

[59] Glynn, P. W., & Whitt, W. (1987). Sufficient conditions for functional limit theorem versions of $L = \lambda W$. *Queueing Systems*, 1, 279–287.

[60] Glynn, P. W., & Whitt, W. (1988). Ordinary CLT and WLLN versions of $L = \lambda W$. *Mathematics of Operations Research*, 13, 674–692.

[61] Glynn, P. W., & Whitt, W. (1989). Indirect estimation via $L = \lambda W$. *Operations Research*, 37, 82–103.

[62] Glynn, P. W., & Whitt, W. (1991). A new view of the heavy-traffic limit for infinite-server queues. *Advances in Applied Probability*, 23, 188–209.

[63] Green, L. V., & Kolesar, P. J. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Scence*, 37, 84–97.

[64] 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.

[65] Grier, N., Massey, W. A., McKoy, T., & Whitt, W. (1997). The time-dependent Erlang loss model with retrials. *Telecommunications Systems*, 7, 253–265.

[66] Gurvich, I., & Whitt, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research*, 34, 363–396.

[67] Gurvich, I., & Whitt, W. (2009). Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management*, 11, 237–253.

[68] Gurvich, I., & Whitt, W. (2009). Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research*, 58, 316–328.

[69] Hairer, E., Norsett, S. P., & Wanner, G. (1993). *Solving Ordinary Differential Equations, I: Nonstiff Problems*. Springer, Philadelphia, second edition.

[70] Halfin, S., & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29, 567–588.

[71] Hampshire, R. C., Jennings, O. B., & Massey, W. A. (2009). Dynamic optimization with applications to dynamic rate queues. *Probability in the Engineering and Informational Sciences*, 23, 231–259.

[72] Hampshire, R. C., & Massey, W. A. (2010). Dynamic optimization with applications to dynamic rate queues. *Tutorials in Operations Research, INFORMS 2010*, 27, 208–247.

[73] Hampshire, R. C., Massey, W. A., & Wang, Q. (2009). Dynamic pricing to control loss systems. *Probability in the Engineering and Informational Sciences*, 23, 357–383.

[74] Harrison, J. M., & Lemoine, A. J. (1977). Limit theoorems for periodic queues. *Journal of Applied Probability*, 14, 566–576.

[75] He, B., Liu, Y., & Whitt, W. (2016). Stabilizing performance in nonstationary queues with non-Poisson arrivals. *Probability in the Engineering and Informational Sciences*, 30, 593–621.

[76] He, S. (2017). Diffusion approximation for efficiency-driven queues: a space-time scaling approach. National University of Singapore, working paper.

[77] Henderson, S., O’Mahony, E., & Shmoys, D. B. (2016). (Citi) bike sharing. Cornell University, Ithaca, NY.

[78] Heyman, D. P., & Whitt, W. (1984). The asymptoic behavior of queues with time-varying arrival. *Journal of Applied Probability*, 21, 143–156.

[79] Ibrahim, R., L’Ecuyer, P., Regnard, N., & Shen, H. (2012). On the modeling and forecasting of call center arrivals. *Proceedings of the 2012 Winter Simulation Conference*, 2012, 256–267.

[80] Ibrahim, R., & Whitt, W. (2011). Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, 59, 1106–1118.

[81] Ibrahim, R., & Whitt, W. (2011). Real-time delay estimation based on delay history in many-server queues with time-varying arrivals. *Production and Operations Management*, 20, 654–667.

[82] Iglehart, D. L. (1965). Limit diffusion approximations for the many-server queue and the repairman problem. *Journal of Applied Probability*, 2, 429–441.

[83] Iglehart, D. L., & Whitt, W. (1970). Multiple channel queues in heavy traffic, I. *Advances in Applied Probability*, 2, 150–177.

[84] Iglehart, D. L., & Whitt, W. (1970). Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability*, 2, 355–369.

[85] Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y., & Wu, X. (2007). 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, 201–214.

[86] Jagerman, D. L. (1975). Nonstationary blocking in telephone traffic. *Bell System Technical Journal*, 54, 625–661.

[87] Jennings, O. B., Mandelbaum, A., Massey, W. A., & Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science*, 42, 1383–1394.

[88] Jongbloed, G., & Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17, 307–318.

[89] Kang, W., & Pang, G. (2015). Equivalence of fluid models for $G_t/GI/N+GI$ queues. arXiv preprint arXiv:1502.00346.

[90] Kang, W., & Ramanan, K. (2010). Fluid limits of many-server queues with reneging. *Annals of Applied Probabability*, 20, 2204–2260.

[91] Kaspi, H., & Ramanan, K. (2011). Law of large numbers limits for many-server queues. *Annals of Applied Probabability*, 21, 33–114.

[92] Kaspi, H., & Ramanan, K. (2013). SPDE limits of many-server queues. *Annals of Applied Probabability*, 23, 145–229.

[93] Keller, J. (1982). Time-dependent queues. *SIAM Review*, 24, 401–412.

[94] Kelly, F. (1991). Loss networks. *Annals of Applied Probabability*, 1, 319–378.

[95] Kim, S., Vel, P., Whitt, W., & Cha, W. C. (2015). Poisson and non-Poisson properties of appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters*, 43, 247–253.

[96] Kim, S., & Whitt, W. (2013). Statistical analysis with Little's law. *Operations Research*, 61, 1030–1045.

[97] Kim, S., & Whitt, W. (2013). Estimating waiting times with the time varying Little's law. *Probability in the Engineering and Informational Sciences*, 27, 471–506.

[98] Kim, S., & Whitt, W. (2014). Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics*, 17, 307–318.

[99] Kim, S., & Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Operations Management*, 16, 464–480.

[100] Kim, S., Whitt, W., & Cha, W. C. (2018). A data-driven model of an appointment-generated arrival processes at an outpatient clinic. *INFORMS Journal on Computing*, 30, 181-199.

[101] Kleinrock, L. (1964). *Communication Nets: Stochastic Message Flow and Delay*. Dover, New York.

[102] Kleinrock, L. (1976). *Queueing Systems*, volume 2. Wiley, New York.

[103] Ko, Y. M., & Gautam, N. (2013). Critically loaded time-varying multiserver queues: computational challenges and approximations. *INFORMS Journal of Computing*, 25, 285–301.

[104] Kolesar, P. J., Rider, P. J., Craybill, T. B., & Walker, W. E. (1975). A queueing-linear-programming approach to scheduling police patrol cars. *Operations Research*, 23, 1045–1062.

[105] Koopman, B. O. (1972). Air-terminal queues under time-dependent conditions. *Operations Research*, 20, 1089–1114.

[106] Kuczura, A. (1973). The interrupted Poisson process as an overflow process. *Bell System Technicl Journal*, 52, 437–448.

[107] Kurzhanskiy, A., & Varaiya, P. (2010). Active traffic management on road networks: a macroscopic approach. *Philosophical Transactions of the Royal Society A*, 368, 4607–4626.

[108] Latouche, G., & Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. *SIAM*, Philadelphia.

[109] Lemoine, A. J. (1981). On queues with periodic Poisson input. *Journal of Applied Probability*, 18, 889–900.

[110] Lemoine, A. J. (1989). Waiting time and workload in queues with periodic Poisson input. *Journal of Applied Probability*, 26, 390–397.

[111] Leung, K. K., Massey, W. A., & Whitt, W. (1994). Traffic models for wireless communication networks. *IEEE Journal Selected Areas in Communication*, 12, 1353–1364.

[112] Levi, R., & Radovanovic, A. (2010). Provably near-optimal LP-based policies for revenue managment in systems with reusable resources. *Operations Research*, 58, 503–507.

[113] Lewis, P. A. W. (1965). Some results on tests for Poisson processes. *Biometrika*, 52, 67–77.

[114] Li, A., & Whitt, W. (2014). Approximate blocking probabilities for loss models with independence and distribution assumptions relaxed. *Performance Evaluation*, 80, 82–101.

[115] Li, A., Whitt, W., & Zhao, J. (2016). Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 30, 1–20.

[116] Little, J. D. C. (1961). A proof of the queueing formula: $L = \lambda W$. *Operations Research*, 9, 383–387.

[117] Little, J. D. C. (2011). Little's law as viewed on its 50th anniversary. *Operations Research*, 59, 536–539.

[118] Little, J. D. C., & Graves, S. C. (2008). Little's law. In D. Chhajed and T. J. Lowe, editors, *Building Intuition: Insights from Basic Opertations Management Models and Principles*, Chapter 5, 81–100, Springer, New York.

[119] Liu, R., Kuhl, M. E., Liu, Y., & Wilson, J. R. (2018). Modelling and simulation of nonstationary non-Poisson processes (NNPPs). *INFORMS Journal on Computing*, forthcoming.

[120] Liu, Y. (2018). Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research*, 66, 514–534.

[121] Liu, Y., & Whitt, W. (2011). A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*, 59, 835–846.

[122] Liu, Y., & Whitt, W. (2011). Large-time asymptotics for the $G_t / M_t / s_t + GI_t$ many-server fluid queue with abandonment. *Queueing Systems*, 67, 145–182.

[123] Liu, Y., & Whitt, W. (2012). The $G_t / GI / s_t + GI$ many-server fluid queue. *Queueing Systems*, 71, 405–444.

[124] 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.

[125] Liu, Y., & Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research*, 60, 1551–1564.

[126] Liu, Y., & Whitt, W. (2014). Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability*, 24, 378–421.

[127] Liu, Y., & Whitt, W. (2014). Algorithms for time-varying networks of many-server fluid queues. *INFORMS Journal on Computing*, 26, 59–73.

[128] Liu, Y., & Whitt, W. (2014). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 28, 419–449.

[129] Liu, Y., & Whitt, W. (2017). Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research*, 256, 419–449.

[130] Liu, Y., Whitt, W., & Yu, Y. (2016). Approximations for heavily-loaded $G / GI / N + GI$ queues. *Naval Research Logistics*, 63, 187–2017.

[131] Long, Z., & Zhang, J. (2014). Convergence to equilibrium states for fluid models of many-server queues with abandonment. *Operations Research Letters*, 42, 388–393.

[132] Lovejoy, W. S., & Desmond, J. S. (2011). Little's law flow analysis of observation unit impact and sizing. *Acad, Emergency Medicine*, 18, 183–189.

[133] Lucantoni, D. M. (1991). New results for the single-server queue with a batch markovian arrival process $BMAP/G/1$ queue. *Stochastic Models*, 7, 1–46.

[134] Lucantoni, D. M., Choudhury, G. L., & Whitt, W. (1994). The transient $BMAP/G/1$ queue. *Stochastic Models*, 10, 145–182.

[135] Ma, N., & Whitt, W. (2015). Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters*, 102, 202–207.

[136] Ma, N., & Whitt, W. (2018). A rare-event simulation algorithm for periodic single-server queues. *INFORMS Journal on Computing*, 30, 71–89.

[137] Ma, N., & Whitt, W. (2018). Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems*, forthcoming.

[138] Mandelbaum, A. (2011). Lecture notes, course on service engineering. The Technion, Israel, http://iew3.technion.ac.il/serveng/Lectures/lectures.html.

[139] Mandelbaum, A., & Massey, W. A. (1995). Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20, 33–64.

[140] Mandelbaum, A., Massey, W. A., & Reiman, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems*, 30, 149–201.

[141] Mandelbaum, A., & Zeltyn, S. (2009). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57, 1189–1205.

[142] Massey, W. A. (1985). Asymptotic analysis of the time-varying $M/M/1$ queue. *Mathematics of Operations Research*, 10, 305–327.

[143] Massey, W. A. (2002). The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, 21, 173–204.

[144] Massey, W. A., Parker, G. A., & Whitt, W. (1996). Estimating the parameters of a nonhomogeneous Poisson process with linear rate. *Telecommunication Systems*, 5, 361–388.

[145] Massey, W. A., & Pender, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75, 243–277.

[146] Massey, W. A., & Pender, J. (2018). Dynamic-rate Erlang-A queues. *Queueing Systems*, forthcoming.

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

[148] Massey, W. A., & Whitt, W. (1994). An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Annals of Applied Probability*, 4, 1145–1160.

[149] Massey, W. A., & Whitt, W. (1994). Unstable asymptotics for nonstationary queues. *Mathematics of Operations Research*, 19, 267–291.

[150] Massey, W. A., & Whitt, W. (1994). A stochastic model to capture space and time dynamics in wireless communication systems. *Probability in the Engineering and Informational Sciences*, 8, 541–569.

[151] Massey, W. A., & Whitt, W. (1997). Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems*, 25, 157–172.

[152] Massey, W. A., & Whitt, W. (1998). Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 9, 1130–1155.

[153] May, A. D., & Keller, H. E. M. (1967). A deterministic queueing model. *Transportation Research*, 1, 117–128.

[154] Moler, C., & van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20, 801–836.

[155] Neely, M. J., Modiano, E., & Rohrs, C. E. (2005). Dynamic power allocation and routing for time-varying wireless networks. *IEEE Journal on Selected Areas in Communications*, 23, 89–103.

[156] Nelson, B. L., & Gerhardt, I. (2011). Modeling and simulating renewal nonstationary arrival processes to facilitate analysis. *Journal of Simulation*, 5, 3–8.

[157] Nelson, B. L., & Taaffe, M. (2004). The $Ph_t / Ph_t / \infty$ queueing system: Part I the single node. *INFORMS Journal on Computing*, 16, 266–274.

[158] Nelson, B. L., & Taaffe, M. (2004). The $\left[Ph_t / Ph_t / \infty\right]^K$ queueing system: Part II the multiclass network. *INFORMS Journal on Computing*, 16, 275–283.

[159] Newell, G. F. (1965). Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Review*, 7, 223–240.

[160] Newell, G. F. (1968). Queues with time dependent arrival rates, I: the transition through saturation. *Journal of Applied Probability*, 5, 436–451.

[161] Newell, G. F. (1968). Queues with time dependent arrival rates, II: the maximum queue and the return to equilibrium. *Journal of Applied Probability*, 5, 579–590.

[162] Newell, G. F. (1968). Queues with time dependent arrival rates, III: a mild rush hour. *Journal of Applied Probability*, 5, 591–606.

[163] Newell, G. F. (1982). *Applications of Queueing Theory*. Chapman and Hall/CRC, New York, second edition.

[164] Nieuwenhuis, G. (1989). Equivalence of functional limit theorems for stationary point processes and their Palm distributions. *Probabiity and Related Fields*, 81, 593–608.

[165] Niyirora, J., & Pender, J. (2016). Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics*, 63, 615–630.

[166] Niyirora, J., & Zhuang, J. (2017). Fluid approximations and control of queues in emergency departments. *European Journal of Operational Research*, 261, 1110–1124.

[167] Oliver, R. M., & Samuel, A. H. (1962). Reducing letter delays in post offices. *Operations Research*, 10, 839–892.

[168] Ong, K. L., & Taaffe, M. R. (1988). Approximating $Ph(t)/Ph(t)/1/c$ nonstationary queueing systems. *Mathematics and Computers in Simulation*, 30, 441–452.

[169] Ong, K. L., & Taaffe, M. R. (1989). Nonstationary queues with interrupted Poisson arrivals and unreliable/repairable servers. *Queueing Systems*, 4, 27–46.

[170] Palm, C. (1988). *Intensity Variations in Teletraffic*. North-Holland, Amsterdam. (English translation of 1943 German edition published by Ericsson Technics, vol. 44, 1-189.).

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

[172] Pang, G., & Whitt, W. (2012). The impact of dependent service times on large-scale service systems. *Manufacturing and Service Operations Management*, 14, 262–278.

[173] Pang, G., & Zhou, Y. (2017). Two-parameter process limits for an infinite-server queue with arrival dependent service times. *Probability in the Engineering and Information Sciences*, 127, 1375–1416.

[174] Pender, J. (2014). A Poisson Charlier approximation for nonstationary queues. *Operations Research Letters*, 42, 293–298.

[175] Pender, J. (2014). Gram Charlier expansions for time varying multiserver queues with abandonment. *SIAM Journal of Applied Mathematics*, 74, 1238–1265.

[176] Pender, J. (2015). Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering and Information Sciences*, 29, 27–49.

[177] Pender, J. (2017). Sampling the functional Kolmogorov forward equations for nonstationary queueing networks. *INFORMS Journal on Computing*, 29, 1–17.

[178] Pender, J., & Massey, W. A. (2017). Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probability in the Engineering and Information Sciences*, 31, 1–42.

[179] Porteus, E. (1989). The case analysis section: The National Cranberry Cooperative. *Interfaces*, 19, 29–39.

[180] Porteus, E. (1993). Case analysis: Analyses of the National Cranberry Cooperative – 1. Tactical options. *Interfaces*, 23, 21–39.

[181] Porteus, E. (1993). Case analysis: Analyses of the National Cranberry Cooperative– 2. Envirnmental changes and implementation. *Interfaces*, 23, 81–92.

[182] Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flanner, B. P. (2007). *Numerical Recipes: the Art of Scientific Computing*. Cambridge University Press, third edition.

[183] Puhalskii, A. A. (2013). On the $M_t / M_t / K_t + M_t$ queue in heavy traffic. *Mathematical Methods of Operations Research*, 78, 119–148.

[184] Puhalskii, A. A., & Reiman, M. I. (2000). The multiclass $GI / PH / N$ queue in the halfin-whitt regime. *Advances in Applied Probability*, 32, 564–595.

[185] Ran, B., & Boyce, D. E. (1994). *Dynamic Urban Transportation Network Models*. Springer-Verlag, Berlin, Lecture notes in economics and mathematical systems, 417 edition.

[186] Reed, J. (2009). The $GI / G / N$ queue in the Halfin-Whitt regime. *Annals of Applied Probability*, 19, 2211–2269.

[187] Reed, J., & Tezcan, T. (2012). Hazard rate scaling of the abandonment distribution for the $GI / M / N + GI$ queue in heavy traffic. *Operations Research*, 60, 981–995.

[188] Restrepo, M., Henderson, S. G., & Topaloglu, H. (2009). Erlang loss models for the static seployment of ambulances. *Health Care Management Science*, 12, 67–79.

[189] Rider, K. (1976). A simple approximation to the average queue size in the time-dependent $M / M / 1$ queue. *Journal of the ACM*, 23, 361–367.

[190] Rolski, T. (1981). Queues with nonstationary input stream: Ross's conjecture. *Advances in Applied Probability*, 13, 603–618.

[191] Rolski, T. (1989). Queues with nonstationary inputs. *Queueing Systems*, 5,113–130.

[192] Rolski, T. (1989). Relationships between characteristics in periodic Poisson queues. *Queueing Systems*, 4, 17–26.

[193] Ross, S. M. (1996). *Stochastic Processes*. Wiley, New York, second edition.

[194] Rothkopf, M. H., & Oren, S. S. (1979). A closure approximation for the nonstationary $M / M / s$ queue. *Management Science*, 25, 522–534.

[195] Schwarz, J. A., Selinka, G., & Stolletz, R. (2016). Performance analysis of time-varying queueing systems: survey and classification. *Omega*, 63, 170–189.

[196] Seneta, E. (1973). On the historical development of the theory of finite inhomogeneous Markov chains. *Proceedings of the Cambridge Philosophical Society*, 74, 507–513.

[197] Seneta, E. (1981). *Non-negative matrices and Markov chains*. Springer, New York, second edition.

[198] Shakkotai, S., Srikant, R., & Stolyar, A. L. (2004). Pathwise optimality of the exponential scheduling rule for wireless channels. *Advances in Applied Probability*, 36, 1021–1045.

[199] Shi, P., Chou, M. C., Dai, J. G., Ding, D., & Sim, J. (2016). Models and insights for hospital inpatient operations: time-dependent ED boarding time. *Management Sccience*, 62, 1–28.

[200] Stein, C. M. (1986). *Approximate Computation of Expectations*. IMS, Lecture Notes in Statistics, Hayward, CA.

[201] Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton, NJ, siam classics in applied mathematics edition.

[202] Stidham, S. (1974). A last word on $L = \lambda W$. *Operations Research*, 22,417–421.

[203] Stolletz, R. (2008). Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research*, 190, 478–493.

[204] Stroock, D. W., & Varadhan, S. R. S. (1979). *Multidimensional Diffusion Processes*. Springer, New York.

[205] Sun, X., & Whitt, W. (2018). Delay-based service differentiation with many servers and time-varying arrival rates. *Stochastic Systems*, forthcoming.

[206] Taaffe, M. R., & Clark, G. M. (1988). Approximating nonstationary two-priority nonpreemptive queueing systems. *Annals of Operations Research*, 35, 125–145.

[207] Taaffe, M. R., & Ong, K. L. (1987). Approximating $Ph(t)/M(t)/S/C$ queueing systems. *Annals of Operations Research*, 8, 103–116.

[208] Talreja, R., & Whitt, W. (2008). Fluid models for overloaded multi-class many-server queueing systems with fcfs routing. *Management Science*, 54, 1513–1527.

[209] Ward, A. R. (2012). Asymptotic analysis of queueing systems with reneging: A survey of results for fifo, single class models. *Surveys in Operations Research and Management*, 17, 1–14.

[210] Wein, L. M. (1989). Capacity allocation in generalized Jackson networks. *Operations Research Letters*, 8, 143–146.

[211] Whitt, W. (1985). Blocking when service is required from several facilities simultaneously. *AT&T Technical Journal*, 64, 1807–1856.

[212] Whitt, W. (1991). A review of $L = \lambda W$. *Queueing Systems*, 9, 235–268.

[213] Whitt, W. (1991). The pointwise stationary approximation for $M_t / M_t / s$ queues is asymptotically correct as the rates increase. *Management Science*, 37, 307–314.

[214] Whitt, W. (2002). *Stochastic-Process Limits*. Springer, New York.

[215] Whitt, W. (2002). The Erlang *B* and *C* formulas: problems and solutions. Columbia University, http://www.columbia.edu/~ww2040/allpapers.html.

[216] Whitt, W. (2003). How multiserver queues scale with growing congestion -dependent demand. *Operations Research*, 51, 531–542.

[217] Whitt, W. (2004). A diffusion approximation for the $G / GI / n / m$ queue. *Operations Research*, 52, 922–941.

[218] Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50, 1449–1461.

[219] Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science*, 51, 221–235.

[220] Whitt, W. (2005). Heavy-traffic limits for the $G / H2^* / n / m$ queue. *Mathematics of Operations Research*, 30, 1–27.

[221] Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research*, 54, 37–54.

[222] Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15, 88–102.

[223] Whitt, W. (2012). Extending the FCLT version of $L = \lambda W$. *Operations Research Letters*, 40, 230–234.

[224] Whitt, W. (2014). Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters*, 42, 458–461.

[225] Whitt, W. (2015). Stabilizing performance in a single-server queue with time -varying arrival rate. *Queueing Systems*, 81, 341–378.

[226] Whitt, W. (2016). Heavy-traffic limits for a single-server queue leading up to a critical point. *Operations Research Letters*, 44, 796–800.

[227] Whitt, W., & You, W. (2017). Time-varying robust queueing. under review, Columbia University, Available at: http://www.columbia.edu/~ww2040/allpapers.html.

[228] Whitt, W., & You, W. (2018). Using robust queueing to expose the impact of dependence in single-server queues. Operations Research, 66, 184-199.

[229] Whitt, W., & You, W. (2018). Algorithms to compute the index of dispersion of a stationary point process. in preparation, Columbia University, http://www.columbia.edu/~ww2040/allpapers. html.

[230] Whitt, W., & Zhang, X. (2017). A data-generated queueing model of an emergency department. *Operations Research for Health Care*, 12, 1–15.

[231] Whitt, W., & Zhang, X. (2017). A periodic Little's law. Operations Research, forthcoming.

[232] Whitt, W., & Zhang, X. (2017). A central-limit-theorem version of the periodic Little's law. *Queueing Systems*, forthcoming.

[233] Whitt, W., & Zhao, J. (2017). Many-server loss models with non-Poisson time-varying arrivals. *Naval Research Logistics*, 64, 177–202.

[234] Yom-Tov, G., & Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Operations Management*, 16, 283–299.

[235] Zeltyn, S., & Mandelbaum, A. (2005) Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems*, 51, 361–402.

[236] Zhang, J. (2013). Fluid models of many-server queues with abandonment. *Queueing Systems*, 73, 147–193.

[237] Zhang, X., Hong, L. J., & Glynn, P. W. (2014). Timescales in modeling call center arrivals. Working paper, Stanford University, 2014.

[238] Zheng, Z., & Glynn, P. W. (2017). Fitting continuous piecewise-linear Poisson intensities via maximum likelihood and least squares. In W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, editors, *Proceedings of the 2017 Winter Simulation Conference*, 1740–1749.

[239] Zuniga, A. W. (2014). Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions. *Stochastic Processes and their Applications*, 124, 1436–1468.

[240] Zychlinski, N., Mandelbaum, A., & Momcilovic, P. (2018). Time-varying tandem queues with blocking: modeling, analysis and operational insights via fluid models with reflection. *Queueing Systems*, 89, 15-47.