Telecommunication Systems 21:2–4, 173–204, 2002

The Analysis of Queues with Time-Varying Rates for Telecommunication Models

WILLIAM A. MASSEY wmassey@princeton.edu
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

Abstract. Time dependent behavior has an impact on the performance of telecommunication models. Examples include: staffing a call center, pricing the inventory of private line services for profit maximization, and measuring the time lag between the peak arrivals and peak load for a system. These problems and more motivate the development of a queueing theory with time varying rates. Queueing theory as discussed in this paper is organized and presented from a communications perspective. Canonical queueing models with time-varying rates are given and the necessary mathematical tools are developed to analyze them. Finally, we illustrate the use of these models through various communication applications.

Keywords: loss models, delay models, offered load, carried load, fluid approximations, diffusion approximations, multiserver queues, wireless networks, packet networks, server staffing, private line services, circuit switched networks

Table of contents

1. Introduction

There are many telecommunication motivations for the study of queueing systems with time-varying rates. First, real life is nonstationary. The number of telephone calls made during the five minute interval of 2:07 pm to 2:12 pm on a Wednesday afternoon is considerably larger than the number of calls made during the five minute interval of 3:46 am to 3:51 am Monday morning. Second, the fields of voice and data communications have been the major sources of motivation for the growth and creation of queueing theory throughout the entire twentieth century, starting in 1917 with the Erlang blocking formula [Erlang, 9]. As we increase our mathematical understanding of how time dependent behavior affects queueing models, then we also increase our understanding of how nonstationary phenomena affect the performance of communication systems.

The greater mathematical complexity of time-varying rate problems has resulted in far less literature on these types of queues compared to what has been written on the equilibrium behavior of queues with constant rates. [Hall, 13], for example, is one of the rare textbooks that devotes an entire chapter to the subject of nonstationary queues. Many of the theoretical tools such as equilibrium probabilities for Markov chains, matrix geometric solutions, and Laplace transforms are not available or directly applicable for queues with time varying rates. This means that new analytical tools continually need to be invented. The challenge of working in this field is creating the new tools for the analysis of queues with time varying rates. The reward of working in this field is helping to determine which mathematical tools are needed to advance the theory of nonstationary queues. Creating such a new theory provides new formulas and algorithms to employ in the performance modelling of communication systems.

This paper is an overview of the author's work on queues with time-varying rates. The papers by Rothkopf and Oren [47] and Newell [39] are two early works that inspired the author's Ph.D. thesis [Masey, 31] in nonstationary queues, under the direction of Joseph B. Keller. Another significant influence is the work of [Jagerman, 17].

Understanding the impact of time-varying behavior on communication systems is a primary goal of this paper, so this perspective shapes our presentation of queueing theory. We start with *traffic models* in section 2, which describe the arrivals of customers *requesting* communication services, and progress to *offered load models,* which describe the total number of personal communication resources *requested* by arriving customers (like channels, bandwidth, radio frequencies). This section is a summary of papers [Duffield et al., 6; Eick et al., 7,8; Leung et al., 26; Massey et al., 33; Massey and Whitt, 34,36].

We then progress from the offered load models of section 2 to discuss the *loss models* in section 3. Here, we realistically assume that there are a limited number of resources to be used in parallel (like virtual voice circuits or video on demand). Customers requesting service resources currently in use either leave impatiently or are blocked from accessing the system. This section is a summary of papers [Davis et al., 5; Grier et al., 12; Hampshire et al., 14,15; Jennings et al., 18–20; Lanning et al., 25; Massey and Whitt, 35].

Finally, in section 4, we transform the offered load model into a *delay model,* when customers, arriving to limited resources already in use, patiently wait in a buffer until they can acquire their communication resources. This section is a summary of papers [Mandelbaum and Massey, 27; Mandelbaum et al., 28,29; Massey, 31,32]. E-mail servers and call centers are examples of applications for delay models.

We associate a canonical queueing example with each type of model and then proceed to introduce the time-varying rate analogue to this canonical example. The classical examples for traffic, offered load and loss models are, respectively, Poisson processes, the $M/G/\infty$ queue and the $M/G/L/L$ queue. The canonical examples for delay models are both the $M/M/1$ and $M/M/L$ queues. Next, we proceed to develop the time-varying analogues to the classical models and briefly present the theoretical tools needed to understand the analysis of each time-varying queueing system. The use of these formulas are illustrated through various communication applications.

Throughout the paper, we discuss the applications of this time-varying rate queueing theory to areas such as server staffing, circuit switched networks, private line services, call centers, packet networks and wireless communications.

2. Traffic and offered load models

Performance modelling of telecommunication systems starts with capturing the behavior of the call *arrival traffic.* In this section, we make a case for using the nonhomogeneous Poisson process as the natural model for customer arrivals. The arriving customers request some specific amount of resource (circuit, bandwidth, radio channel, etc.) to facilitate their communication service. In classical telephony, the collective amount of resources requested by customers at a given time is referred to as the *offered load.* We discuss the $Mt/G/\infty$ queue as a canonical model for the offered load process, as first presented by Palm [42].

What is motivating much of the underlying mathematics in this section is the fact that communication services is wireless and packet networks expand our sense of what "service" means in a queueing theoretic context. Instead of directly using the theory of point process or Poisson random measures such as in [Daley and Vere-Jones, 4] or [Prékopa, 43,44], we appeal to the theory of stochastic integration with respect to a Poisson process to construct the appropriate measure. This machinery allows us to generalize this $M_t/G/\infty$ model and helps to construct the offered load processes for wireless and packet network systems.

2.1. *Poisson processes and connection level traffic*

The Poisson process is the canonical traffic process model. We omit the definition of the stationary version and immediately define the nonstationary version. The stochastic process $A = \{A(t) \mid -\infty < t < \infty\}$ is *nonhomogeneous Poisson* with *mean rate function* λ if it has independent Poisson increments. This means that for all $s < t$ we assume that $\int_s^t \lambda(\tau)\,\mathrm{d}\tau < \infty$, which is equivalent to saying that λ is *locally integrable.* Moreover, we assume that the *increment* $A(t) - A(s)$ for the interval $(s, t]$ has a Poisson distribution or

$$\mathsf{Pr}\big(A(t) - A(s) = n\big) = \mathrm{e}^{-\int_s^t \lambda(\tau)\,\mathrm{d}\tau} \frac{\left(\int_s^t \lambda(\tau)\,\mathrm{d}\tau\right)^n}{n!}, \tag{1}$$

for all non-negative integers n. Finally, we assume that the process has the *independent increment property,* i.e. for all mutually disjoint intervals $(s_1, t_1]$, $(s_2, t_2], \ldots, (s_k, t_k]$, the random variables

$$\big\{A(t_i) - A(s_i) \mid i = 1, \ldots, k\big\} \tag{2}$$

are mutually independent.

While these three assumptions constitute a standard definition for Poisson processes, we can show that the last two assumptions are redundant by citing the following theorem due to Prekopa [43,44].

Theorem 2.1 [Prékopa, 43]. A simple point (counting) process A with a (locally integrable) mean rate function X is nonhomogeneous Poisson if and only if it has independent increments.

This is a simple characterization of nonhomogeneous Poisson processes that sheds light on when we are justified to model call arrival traffic by a nonhomogeneous Poisson process. A case can be made for Poisson modelling at the connection level, since the arriving units are large numbers of people who act independently of each other. However, the reasonableness of this assumption is not as clear at the packet or burst level, since a stream of packets may arrive from the same file transfer. Statistical studies of real packet traffic data has verified both of these assumptions (see [Willinger and Paxson, 50]). This is the insight that theory gives to modelling.

To make the transition from traffic models to offered load models, we must prepare by using the right mathematical tools. We do so by defining a special stochastic calculus for nonhomogeneous Poisson processes.

2.2. *Stochastic integrals, cumulant moments, stationary excess distributions*

We now want to move from mathematically modelling arrival traffic to describing the collective amount of communication resources requested at any given time. In classical telephony, this is referred to as the *offered load.* This is in contrast to the *carried load* which is the collective amount of communication resources used at any given time. As

network capacity increases and communication resources become more abundant (like optical networks), we may view the offered load of current networks as the carried load of future networks. Moreover, offered load models tell us what are the largest number of resources needed at any given time. They also tell a service provider what is the maximal revenue that can be obtained from the provided services, no matter how efficiently they are allocated. We assume that customers do not request their resources until they arrive and their use of the resources is independent but identically distributed to any other customer.

Mathematically, we want to have the ability to sum over the customer arrival times, the amount of resources in use and for how long. This can easily be expressed as a stochastic integral but with a special class of integrands.

Let $\{S_n \mid n = 1, 2, \ldots\}$ be an independent and identically distributed sequence of random variables and we define $\phi : \mathbb{R}^2 \to \mathbb{R}$ to be an *integrand* if it is a non-negative measurable function. We then define our stochastic integral with respect to a nonhomogeneous Poisson process A to be

$$\int_s^t \phi(S_{A(\tau)}, \tau)\, \mathrm{d}A(\tau) \equiv \sum_{n=1}^{A(t)-A(s)} \phi\big(S_n, \hat{A}_n\big) \tag{3}$$

where $\mathrm{d}A(\tau) = A(\tau) - A(\tau-)$ and $\hat{A}_n$ is the time of the nth arrival in the interval $(s, t]$. Unlike stochastic integration over Brownian motion, this can be defined as a sample path integration. Moreover, it can be shown that

$$\mathsf{E}\left[\int_{-\infty}^t \phi(S_{A(\tau)}, \tau)\, \mathrm{d}A(\tau)\right] = \int_{-\infty}^t \mathsf{E}\big[\phi(S, \tau)\big]\lambda(\tau)\, \mathrm{d}\tau, \tag{4}$$

where S has the same distribution as all of the S_n.

In addition to stochastic integrals, we introduce an important class of moments that are useful for random variables with distributions "close" to the Poisson distribution. Let X be some non-negative random variable with $\mathsf{E}[\mathrm{e}^{\theta X}] < \infty$ for some $\theta > 0$. The *cumulant moments* of X, denoted $\mathsf{C}^{(n)}[X]$ for positive integers n, are defined by the generating function relation

$$\log\, \mathsf{E}\big[\mathrm{e}^{\theta X}\big] = \sum_{n=1}^{\infty} \frac{\theta^n}{n!} \mathsf{C}^{(n)}[X]. \tag{5}$$

Note that this quantity is used in the definition of effective bandwidth (for a good discussion of this important topic, see [Kelly, 23] as well as [Shwartz and Weiss, 49]). All these cumulant moments uniquely characterize the distribution of X. The first four cumulant moments are the following:

$$\mathsf{C}^{(1)}[X] = \mathsf{E}[X], \qquad \mathsf{C}^{(2)}[X] = \mathsf{E}\big[\widehat{X}^2\big] = \mathsf{Var}[X], \tag{6}$$

$$\mathsf{C}^{(3)}[X] = \mathsf{E}\big[\widehat{X}^3\big] \quad \text{and} \quad \mathsf{C}^{(4)}[X] = \mathsf{E}[\widehat{X}^4\big] - 3\mathsf{E}\big[\widehat{X}^2\big]^2, \tag{7}$$

where $\widehat{X} = X - \mathsf{E}[X]$. Cumulant moments have the following properties:

Additivity and homogeneity. If X and Y are independent, then

$$\mathsf{C}^{(n)}[X+Y] = \mathsf{C}^{(n)}[X] + \mathsf{C}^{(n)}[Y] \tag{8}$$

and for all constants λ,

$$\mathsf{C}^{(n)}[\lambda X] = \lambda^n \mathsf{C}^{(n)}[X]. \tag{9}$$

Poisson test. The distribution of X is Poisson if and only if for all $n = 1, 2, \ldots$

$$\mathsf{C}^{(n)}[X] = \mathsf{E}[X]. \tag{10}$$

Gaussian test. The distribution of X is Gaussian (normal) if and only if for all $n = 3, 4, \ldots$

$$\mathsf{C}^{(n)}[X] = 0. \tag{11}$$

Independence test. Given k random variables $X_1, \ldots, X_k$, they are mutually independent if and only if for all real constants $a_1, \ldots, a_k$ we have

$$\mathsf{C}^{(n)}\left[\sum_{i=1}^{k} a_i X_i\right] = \sum_{i=1}^{k} a_i^n \mathsf{C}^{(n)}[X_i]. \tag{12}$$

Below is the fundamental result for this special class of stochastic integrals. The proof of this result can be found in [Duffield, Massey and Whitt, 6].

Theorem 2.2 [Duffield, Massey and Whitt, 6]. If

$$Z_\phi(t) = \int_{-\infty}^{t} \phi(S_{A(\tau)}, \tau)\, \mathrm{d}A(\tau) \tag{13}$$

and $\int_{-\infty}^{t} \mathsf{E}[\mathrm{e}^{\theta\phi(S,\tau)} - 1]\, \lambda(\tau)\, \mathrm{d}\tau < \infty$, for some $\theta > 0$, then

$$\mathsf{C}^{(n)}\big[Z_\phi(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[\phi(S,\tau)^n\big]\lambda(\tau)\, \mathrm{d}\tau, \tag{14}$$

for all $n = 1, 2, \ldots,$ which is equivalent to

$$\log\, \mathsf{E}\big[\mathrm{e}^{\theta Z_\phi(t)}\big] = \int_{-\infty}^{t} \mathsf{E}\big[\mathrm{e}^{\theta\phi(S,\tau)} - 1\big]\lambda(\tau)\, \mathrm{d}\tau. \tag{15}$$

From this theorem follows a host of stochastic integral formulas and properties.

Corollary 2.3 [Massey and Whitt, 36]. For all t, we have mean, variance, and covariance formulas: for integrands ϕ and ψ, we have for all $t \geqslant 0$

$$\mathsf{E}\big[Z_\phi(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[\phi(S,\tau)\big]\lambda(\tau)\, \mathrm{d}\tau, \tag{16}$$

$$\mathsf{Var}\big[Z_\phi(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[\phi(S,\tau)^2\big]\lambda(\tau)\, \mathrm{d}\tau \tag{17}$$

and

$$\mathsf{Cov}\big[Z_\phi(t), Z_\psi(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[\phi(S,\tau)\psi(S,\tau)\big]\lambda(\tau)\,\mathrm{d}\tau. \tag{18}$$

Poisson thinning. The process $\{Z_\phi(t) \mid -\infty < t < \infty\}$ is Poisson if and only if ϕ is a binary (0 or 1) valued integrand. Moreover, if ϕ and ψ are both binary valued, then Z_ϕ and Z_ψ are independent Poisson processes if and only if $\phi + \psi$ is binary also.

Given a non-negative random variable X, where $\mathsf{E}[X] < \infty$, we say that the random variable X_e has the *stationary excess distribution* of X if for all $x \geqslant 0$

$$\mathsf{Pr}(X_\mathrm{e} \leqslant x) \equiv \frac{1}{\mathsf{E}[X]} \int_0^x \mathsf{Pr}(X > y)\,\mathrm{d}y. \tag{19}$$

A simple way to motivate X_e is to consider an i.i.d. sequence of random variables $\{X_n \mid n \geqslant 1\}$ where each X_n has the same distribution as X. Given a fixed time x, the amount of time that the nth renewal lives before age x is $\min(X_n, x)$. The fraction of time that a lifetime of less than x time units is observed after the first n renewals is then the ratio of $\sum_{i=1}^n \min(X_i, x)$ divided by $\sum_{i=1}^n X_i$. By the strong law of large numbers, the limiting time average for this event is then

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \min(X_i, x)}{\sum_{i=1}^n X_i} = \frac{\mathsf{E}[\min(X, x)]}{\mathsf{E}[X]} = \mathsf{Pr}(X_\mathrm{e} \leqslant x). \tag{20}$$

We can simplify many useful formulas by using the following distributional transformation. Note that X and X_e have the same distribution or $X \stackrel{\mathrm{d}}{=} X_\mathrm{e}$ if and only if X is exponentially distributed. If X is constant, then X_e is uniformly distributed on $[0, X]$. We can use the stationary excess distribution to give a probabilistic version of the mean value theorem.

Theorem 2.4 [Massey and Whitt, 34]. For all constants and random variables $X \geqslant 0$, if f is differentiable on $[0, \infty)$ and suitably integrable, then we have

$$\frac{\mathsf{E}[f(t + X)] - f(t)}{\mathsf{E}[x]} = \mathsf{E}\big[f'(t + X_\mathrm{e})\big]. \tag{21}$$

An immediate application of this theorem is that for any integer $n \geqslant 0$ where $\mathsf{E}[X^{n+1}] < \infty$, then

$$\mathsf{E}[X_\mathrm{e}^n] = \frac{\mathsf{E}(X^{n+1})}{(n+1)\mathsf{E}[X]}. \tag{22}$$

Now we proceed to construct the canonical example for offered load models.

2.3. *The $M_t/G/\infty$ queue*

The $M/G/\infty$ queue, which has Poisson arrivals, generally distributed holding times, and an infinite number of servers, is the canonical example of an offered load model. If $\mathsf{E}[S] < \infty$ and Q_∞ is the steady state number in the $M/G/\infty$ queue, then for all non-negative integers n

$$\mathsf{Pr}(Q_\infty = n) = \mathrm{e}^{-\lambda \mathsf{E}[S]} \frac{(\lambda \mathsf{E}[S])^n}{n!} \quad \text{and} \quad \mathsf{E}[Q_\infty] = \lambda \mathsf{E}[S], \tag{23}$$

where λ is the mean Poisson arrival rate and S is the random service time. Observe that the distribution of Q_∞ is *insensitive* to the distribution of S for fixed mean $\mathsf{E}[S]$.

The mathematical tools of the previous subsection give us the machinery to construct the $M_t/G/\infty$ queue. If we let S_n equal the i.i.d. random service time for nth arrival, we can then construct the following quantities,

$$Q_\infty(t) = \text{number of calls in progress at time } t$$

and

$$D_\infty(t) = \text{number of terminated calls before time } t,$$

using stochastic integration, i.e.

$$Q_\infty(t) = \int_{-\infty}^{t} \mathbb{1}_{\{S_{A(\tau)} > t-\tau\}} \, \mathrm{d}A(\tau) \tag{24}$$

and

$$D_\infty(t) = \int_{-\infty}^{t} \mathbb{1}_{\{S_{A(\tau)} \leqslant t-\tau\}} \, \mathrm{d}A(\tau). \tag{25}$$

These sample path constructions also describe the mechanics of the $M_t/G/\infty$ queue. The quantity $\mathrm{d}A(\tau)$ equals 1 only when τ is the time of a customer arrival, otherwise it is 0. If τ is the time of the nth arrival then $n = A(\tau)$ and its connection holding time is S_n. If $S_n > t - \tau$, then the nth arrival is still in the system a time t, otherwise it has departed the system. Thus $Q_\infty(t)$ is counting all the arrivals still in service and $D_\infty(t)$ is counting all the arrivals with service completions. We can interpret the $M_t/G/\infty$ as the offered load model for the classical telephone trunkline system. From our general theory for these stochastic integrals, we obtain

Theorem 2.5 [Eick, Massey and Whitt, 8]. The $M_t/G/\infty$ queue has the following properties:

1. For all t, $Q_\infty(t)$ has a Poisson distribution with

$$\mathsf{E}\big[Q_\infty(t)\big] = \int_{-\infty}^{t} \mathsf{Pr}(S > t - \tau) \, \lambda(\tau) \, \mathrm{d}\tau. \tag{26}$$

2. For all $t < u$,

$$\mathsf{Cov}\big[Q_\infty(t), Q_\infty(u)\big] = \int_{-\infty}^{t} \mathsf{Pr}(S > u - \tau) \, \lambda(\tau) \, \mathrm{d}\tau. \tag{27}$$

3. The process $\{D_\infty(t) \mid -\infty < t < \infty\}$ is Poisson with

$$\mathsf{E}\big[D_\infty(t)\big] = \int_{-\infty}^{t} \mathsf{Pr}(S \leqslant t - \tau)\,\lambda(\tau)\,\mathrm{d}\tau. \tag{28}$$

4. For all t, $Q_\infty(t)$ and $D_\infty(t)$ are independent random variables.

The corollary to this theorem yields results that we refer to [Eick et al., 8] as the "physics" of the $M_t/G/\infty$ queue. Notice how the transient result for the departure process D_∞ anticipates the Poisson-in-Poisson-out result of Burke's theorem [Burke, 1] for the $M/M/1$ queue in steady state.

Corollary 2.6 [Eick, Massey and Whitt, 8]. We have the following identities for the mean:

$$\mathsf{E}\big[Q_\infty(t)\big] = \mathsf{E}\bigg[\int_{t-S}^{t} \lambda(\tau)\,\mathrm{d}\tau\bigg] = \mathsf{E}\big[\lambda(t - S_\mathrm{e})\big]\mathsf{E}[S]. \tag{29}$$

Moreover if $\lambda(t) = \int_{-\infty}^{\infty} \hat{\lambda}(k)\,\mathrm{e}^{\mathrm{i}kt}\,\mathrm{d}k$, then

$$\mathsf{E}\big[\widehat{Q}_\infty(k)\big] = \hat{\lambda}(k)\mathsf{E}[S] \cdot \mathsf{E}\big[\mathrm{e}^{\mathrm{i}kS_\mathrm{e}}\big]. \tag{30}$$

If $\lambda(t) = a + bt + ct^2$, then

$$\mathsf{E}\big[Q_\infty(t)\big] = \big(\lambda(t - \mathsf{E}[S_\mathrm{e}]\big) + c \cdot \mathsf{Var}[S_\mathrm{e}]\big)\mathsf{E}[S]. \tag{31}$$

Finally, if λ is general, but $\mathsf{Pr}(S > t) = \mathrm{e}^{-\mu t}$ for all $t \geqslant 0$, then

$$\mathsf{E}\big[\widehat{Q}_\infty(k)\big] = \frac{\hat{\lambda}(k)}{\mu + \mathrm{i}k} = \frac{\hat{\lambda}(k)}{\mu} \cdot \frac{\mathrm{e}^{\mathrm{i}\arctan(k/\mu)}}{\sqrt{1 + k^2/\mu^2}}. \tag{32}$$

The formula for the mean offered load in (29) illustrates a recurring theme in queues with time-varying rates: there is a phase lag between the times of peak arrivals (i.e. $\lambda(t)$) and the times of peak load (i.e. $\mathsf{E}[Q_\infty(t)]$). Examining the Fourier transform of $\mathsf{E}[Q_\infty(t)]$ in (30), we not only see this phase lag but we also see a damping of all the amplitudes of $\mathsf{E}[Q_\infty(t)]$ compared to $\lambda(t)\mathsf{E}[S]$. This behavior is more explicit in (31) for the special case of a quadratic arrival rate. Here, the phase lag is precisely $\mathsf{E}[S_\mathrm{e}]$ and the amplitude is "pushed down" by the amount $c \cdot \mathsf{Var}[S_\mathrm{e}] \cdot \mathsf{E}(S)$ when c is negative which corresponds to a λ that attains a finite maximum. Similarly, the amplitude is "pushed up" by the amount $c \cdot \mathsf{Var}[S_\mathrm{e}] \cdot \mathsf{E}[S]$ when c is positive which corresponds to a λ that attains a finite minimum. Moreover, the quadratic results suggest that for fixed $\mathsf{E}[S_\mathrm{e}]$, the damping effect is amplified by increasing $\mathsf{Var}[S_\mathrm{e}]$.

Moreover, since the phase lag behaves more like $\mathsf{E}[S_\mathrm{e}]$ than $\mathsf{E}[S]$, we know by (31) that $\mathsf{E}[S_\mathrm{e}] = \mathsf{E}[S^2]/(2\mathsf{E}[S])$ which for fixed mean is dependent on the distribution of S through its second moment. This has major implications for holding times with heavy tail distributions. They can have reasonable means but create enormous phase lags between the times of peak arrivals and peak loads. In the next section, we use these stochastic integration tools to model wireless networks.

2.4. *Offered load models for wireless networks*

Now we show how this stochastic integration theory can easily be adapted to construct the offered load model for a wireless communication network. These offered load wireless models are developed in the trilogy of papers [Massey and Whitt, 34,36; Leung et al., 26]. Wireless communications inspires a new way of looking at the notion of "service" for queueing models. In queueing theory, service is traditionally viewed as an interval of time and in communications modelling the service time models the conversation time or connection time. However, wireless mobiles move during their connection times. This brings both temporal and spatial dimensions to the notion of service so that now it becomes a "path through a location space." Moreover, if a node in a queueing network is defined to be a place where a customer receives service, then a wireless network is a concrete setting for defining an "infinite node" network. Every point in space should be a place to receive service so we can easily motivate the location space to be a subset of $\mathbb{R}$.

Let Γ represent a *base station cell* as the subset of the location space and

$$L_n(\tau, t) = \text{ location at time } t \text{ for } n\text{th mobile},$$

given that it initiates a call at time τ, where the L_n are mutually independent random processes in t. We can then construct the following quantity:

$$Q_\Gamma(t) = \text{ number of calls in progress in cell } \Gamma \text{ at time } t$$

using stochastic integration, i.e.

$$Q_\Gamma(t) = \int_{-\infty}^{t} \mathbb{1}_{\{L_{A(\tau)}(\tau,t)\in\Gamma\}}\, \mathrm{d}A(\tau). \tag{33}$$

In [Massey and Whitt, 34], we refer to this as a *Poisson Arrival Location Model* (PALM). The stochastic calculus gives it the following properties.

Theorem 2.7. The following results hold for the PALM process

1. For all t, $Q_\Gamma(t)$ has a Poisson distribution and
$$\mathsf{E}\big[Q_\Gamma(t)\big] = \int_{-\infty}^{t} \mathsf{Pr}\big(L(\tau, t) \in \Gamma\big)\, \lambda(\tau)\, \mathrm{d}\tau. \tag{34}$$
2. If Δ is another cell, then $Q_\Gamma(t)$ and $Q_\Delta(t)$ are independent random variables if and only if Γ and Δ are "disjoint."
3. For all $t < u$, cell Γ, and cell Δ, we have
$$\mathsf{Cov}\big[Q_\Gamma(t), Q_\Delta(u)\big] = \int_{-\infty}^{t} \mathsf{Pr}\big(L(\tau, t) \in \Gamma, L(\tau, u) \in \Delta\big)\lambda(\tau)\, \mathrm{d}\tau. \tag{35}$$

The last two results look almost contradictory. One result states that Q_Γ and Q_Δ are independent random variables, for two disjoint cells Γ and Δ, but the other result gives

a formula for the covariance between the two random variables. The critical difference between these two results involves *when* we compare the two random variables. Saying that Q_Γ and Q_Δ are independent at the same time means only that for offered load traffic all customers and their service requirements are independent entities and so no one person can exist at two different places at the *same* time. However, this means that comparing these two random variables at two *different* times means that it is then possible for one customer to move from one cell to another in the elapsed time and thus produce a nonzero correlation.

This type of independence result for the transient behavior of the PALM model anticipates the "product form" results for stochastic networks as formulated by Jackson [16] and Kelly [22]. A special case of the PALM model is the finite node version of infinite server networks. These models have been discussed in the papers of Keilson and Servi [21].

2.5. *Offered load models for packet network links*

Now we show how the stochastic integration theory can be adapted to construct the offered load version of the total bandwidth process for a link in a packet network (see [Duffield et al., 6]). Packet networks take advantage of the fact that customers do not use a fixed amount of resources (like bandwidth) for the duration of their connection time. The asynchronous behavior of the resource requests, both due to when customers arrive and their variations per customer in the amount resources used during the connection time, contributes to the multiplexing gain of systems like ATM. Given

$$B_n(\tau, t) = \text{ bandwidth used at time } t \text{ by the } n\text{th connection},$$

which was initiated at time τ, where the B_n are mutually independent random processes in t. We can then construct the quantity

$$R(t) = \text{ total amount of bandwidth in use at time } t,$$

using stochastic integration, i.e.

$$R(t) = \int_{-\infty}^{t} B_{A(\tau)}(\tau, t)\, \mathrm{d}A(\tau). \tag{36}$$

We can interpret $R(t)$ to be the random packet arrival rate at time t. In this manner we construct a non-Poisson, packet level, traffic arrival process out of a Poisson, connection level, traffic arrival process. The stochastic integral properties give us

Theorem 2.8 [Duffield, Massey and Whitt, 6]. The total bandwidth model yields the following set of formulas:

1. For all t and $n = 1, 2, \ldots$ we have,

$$\mathsf{C}^{(n)}[R(t)] = \int_{-\infty}^{t} \mathsf{E}\big[B(\tau, t)^n\big]\lambda(\tau)\, \mathrm{d}\tau. \tag{37}$$

2. In particular we have for all t,

$$\mathsf{E}\big[R(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[B(T,t)\big]\lambda(\tau)\,\mathrm{d}\tau \tag{38}$$

and

$$\mathsf{Var}\big[R(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[B(\tau,t)^2\big]\lambda(\tau)\,\mathrm{d}\tau. \tag{39}$$

3. For all $t < u$, we have

$$\mathsf{Cov}\big[R(t), R(u)\big] = \int_{-\infty}^{t} \mathsf{E}\big[B(\tau,t)B(\tau,u)\big]\lambda(\tau)\,\mathrm{d}\tau. \tag{40}$$

Now we consider a simple bandwidth function model. Only the connection time is random and we assume that all customers agree to use no more that a specified amount of bandwidth that is a function of their connection time.

Theorem 2.9 [Duffield, Massey and Whitt, 6]. If $B(\tau,t) = b(t-\tau)$ when $S_{A(\tau)} > t-\tau$ and zero otherwise, where b is a deterministic, non-negative function, then

$$\mathsf{C}^{(n)}\big[R(t)\big] = \mathsf{E}\big[b(S_\mathrm{e})^n\lambda(t-S_\mathrm{e})\big]\mathsf{E}[S] \tag{41}$$

and

$$\int_{-\infty}^{\infty} \mathsf{C}^{(n)}\big[R(t)\big]\mathrm{e}^{\mathrm{i}kt}\,\mathrm{d}t = \hat{\lambda}(k)\mathsf{E}[S]\cdot\mathsf{E}\big[b(S_\mathrm{e})^n\big]\cdot\mathsf{E}_{b^n}\big[\mathrm{e}^{\mathrm{i}kS_\mathrm{e}}\big]. \tag{42}$$

Just like the $M_t/G/\infty$ queue, the total bandwidth model has its own laws of "physics".

Theorem 2.10 [Duffield, Massey and Whitt, 6]. If $\lambda(t) = a + bt + ct^2$, then

$$\mathsf{C}^{(n)}\big[R(t)\big] = \big(\lambda(t-\mathsf{E}_{b^n}[S_\mathrm{e}]\big) + c\mathsf{Var}_{b^n}[S_\mathrm{e}])\mathsf{E}\big[b(S_\mathrm{e})^n\big]\mathsf{E}[S], \tag{43}$$

where for any bounded continuous function f we define $\mathsf{E}_{b^n}[f(S_\mathrm{e})] \equiv \mathsf{E}[b(S_\mathrm{e})^n f(S_\mathrm{e})]/\mathsf{E}[b(S_\mathrm{e})^n]$. Finally, combining this quadratic behavior with an increasing b gives us

$$\mathsf{E}[S_\mathrm{e}] \leqslant \mathsf{E}_b[S_\mathrm{e}] \leqslant \mathsf{E}_{b^2}[S_\mathrm{e}] \leqslant \cdots \leqslant \mathsf{E}_{b^n}[S_\mathrm{e}] \leqslant \cdots. \tag{44}$$

but if b is decreasing, then

$$\cdots \leqslant \mathsf{E}_{b^n}[S_\mathrm{e}] \leqslant \cdots \leqslant \mathsf{E}_{b^2}[S_\mathrm{e}] \leqslant \mathsf{E}_b[S_\mathrm{e}] \leqslant \mathsf{E}[S_\mathrm{e}]. \tag{45}$$

Moreover, if λ is a general function of time but $b(t) = b$ when t belongs to some subset of time A and $b(t) = 0$ otherwise, then $R(t)/b$ has a Poisson distribution with

$$\mathsf{E}\left[\frac{R(t)}{b}\right] = \mathsf{E}\big[\lambda(t-S_\mathrm{e});\ S_\mathrm{e}\in A\big]\cdot\mathsf{E}[S]. \tag{46}$$

Combining this with quadratic λ, we have a phase lag equal to $\mathsf{E}[S_\mathrm{e} \mid S_\mathrm{e} \in A]$.

The tail behavior of the bandwidth process can be analyzed by the following simple upper bound to its own Chernoff bound.

Theorem 2.11. For x and t such that $x \geqslant \mathsf{E}[R(t)]$, we have

$$\log \mathsf{Pr}(R(t) > x) \leqslant \frac{-1}{|B(U_\lambda(t), t)|_\infty} \int_{\mathsf{E}[R(t)]}^{x} \log\left(1 + \left|B(U_\lambda(t), t)\right|_\infty \frac{y - \mathsf{E}[R(t)]}{\mathsf{Var}[R(t)]}\right) \mathrm{d}y,$$

where $U_\lambda(t)$ is independent of B with

$$\mathsf{Pr}\big(U_\lambda(t) \leqslant s\big) = \frac{\int_{-\infty}^{s} \lambda(r)\,\mathrm{d}r}{\int_{-\infty}^{t} \lambda(r)\,\mathrm{d}r} \tag{47}$$

and for any random variable X, $|X|_\infty = \inf\{x \mid \mathsf{Pr}(|X| \leqslant x) = 1\}$.

Now we apply much of the insight found in the exact analysis of offered load models to do an approximate analysis of loss models.

3. Loss models

The offered load of communication resources requested by the customers may exceed the amount of resources that the communication service really has. One way of dealing with this finite amount of resource is to assume that the queueing system is a loss model. We then assume that customers arriving to request resources currently in use are rejected and leave the system. This is a natural model for realtime communication services like voice calls and video on demand. Our canonical nonstationary queueing model here is the $M_t/G/L/L$ queue which is the time varying analogue to the classical Erlang loss model [Erlang, 9].

We compare the various approximation methods for the $M_t/G/L/L$ queue as found in [Green and Kolesar, 11; Jagerman, 17]. We also discuss the various approximate algorithms that this time varying analysis inspires for telecommunication applications.

3.1. The $M_t/G/L/L$ queue

Our canonical model for a loss system is the $M/G/L/L$ queue, which has Poisson arrivals, generally distributed holding times, L servers, and no buffer, so arriving customers are lost or blocked when all the servers are in use. If Q_L is the steady state number in the $M/G/L/L$ queue, then for all $n = 0, 1, \ldots, L$

$$\mathsf{Pr}(Q_L = n) = \frac{\rho^n/n!}{\sum_{j=0}^{L}(\rho^j/j!)}, \tag{48}$$

where $\rho = \lambda \mathsf{E}[S]$, λ is the mean Poisson arrival rate and S is the random service time. Notice that this distribution is only influenced by the random holding time S through its

mean $\mathsf{E}[S]$. Thus the $M/G/L/L$ has the same insensitivity property that the $M/G/\infty$ queue has, where the corresponding Q_∞ has a Poisson steady state distribution with $\mathsf{E}[Q_\infty] = \rho$.

If we let β_L equal the function

$$\beta_L(x) = \frac{x^L/L!}{\sum_{j=0}^{L}(x^j/j!)}, \tag{49}$$

then $\beta_L(\rho) = \mathsf{Pr}(Q_L = L)$ which is the classical *Erlang B formula* and

$$\beta_L(\rho) = \mathsf{Pr}(Q_L = L) = \mathsf{Pr}(Q_\infty = L \mid Q_\infty < L). \tag{50}$$

Finally,

$$\mathsf{E}[Q_L] = \rho\big(1 - \beta_L(\rho)\big), \tag{51}$$

which is referred to in classical telephony as the *carried load*. A theme that is explicit here in these exact formulas is one that is the crux of the approximation methods for loss models with time varying rates. We use the exact analysis of the offered load models with time varying rates to approximate the blocking behavior and mean carried load of the loss models.

First we discuss how time varying rates force us to rethink what is meant by blocking. Consider the following "metrics:"

$$\mathsf{Pr}\big(Q_L(t)\big) = L\big), \qquad \frac{1}{T}\int_0^T \mathsf{Pr}\big(Q_L(t) = L\big)\,\mathrm{d}t, \quad \text{and} \quad \frac{\int_0^T \lambda(t)\mathsf{Pr}(Q_L(t) = L)\,\mathrm{d}t}{\int_0^T \lambda(t)\,\mathrm{d}t}.$$

These metrics are, respectively, the probability that all channels are in use at time t, the fraction of time during $[0, T]$ that all channels are in use, and finally, the ratio of the average number of customers blocked to the average number of arriving customers during $[0, T]$. For constant rate steady state analysis, all these metrics equal the Erlang blocking formula. In the world of time varying rates we must distinguish between them and select the most appropriate metric in the context of some specific communications problem.

In [36], we proved that the last metric can be viewed as the expectation of a ratio. If we sample the blocking probabilities by a Poisson process of the same rate but independent of the $M_t/G/L/L$ queue, then

$$\mathsf{E}\left[\frac{\int_0^T \mathsf{Pr}(Q_L(t) = L)\,\mathrm{d}A(t)}{A(t)} \,\middle|\, A(T) > 0\right] = \frac{\int_0^T \lambda(t)\mathsf{Pr}(Q_L(t) = L)\,\mathrm{d}t}{\int_0^T \lambda(t)\,\mathrm{d}t}. \tag{52}$$

3.2. The modified offered load approximation

Two standard approximations used for the $M_t/G/L/L$ queue are *modified offered load* (MOL) (see [Jagerman, 17]):

$$\mathsf{Pr}\big(Q_L(t) = L\big) \approx \beta_L\big(\mathsf{E}\big[Q_\infty(t)\big]\big) \tag{53}$$

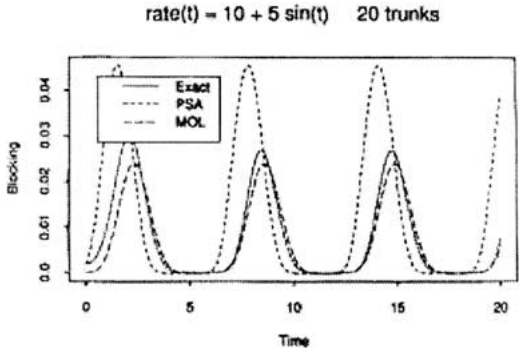

Figure 1. Modified offered load vs. pointwise stationary approximations.

and the *pointwise stationary* (PS) (see [Green and Kolesar, 11]):

$$\mathsf{Pr}\big(Q_L(t) = L\big) \approx \beta_L\big(\lambda(t)\mathsf{E}[S]\big). \tag{54}$$

In this section we show that for most modelling cases of interest, the MOL approximation is superior to the PSA approximation. For example, compare the curves in figure 1. Much of the behavior here can be explained by the physics of the $M_t/G/\infty$ queue as discussed in the previous section as well as [Eick et al., 7,8]. The exact blocking probabilities in figure 1 (solid line) are for the case of $\lambda(t) = 10 + 5\sin t$, S is exponentially distributed with $\mathsf{E}[S] = 1$, and $L = 20$. Here the largest blocking value is 0.02 hence the MOL approximation (long dashed lines) is close to the exact answer whereas the PS approximation (short dashed lines) is not even close. The Erlang blocking formula $\beta_L(\cdot)$ is an increasing function of the offered load. This means that since the PS approximation is β_L applied to $\lambda(t)\mathsf{E}[S]$, then PS peaks the same times that the arrival rate peaks. Thus we see in figure 1 that there is a lag between the times of peak arrivals and peak blocking. Since the exact blocking probabilities are well approximated by MOL or β_L applied to $\mathsf{E}[Q_\infty(t)]$, the offered load results of corollary 6 explain this lag. Moreover, (2.27) shows the amplitude damping of $\mathsf{E}[Q_\infty(t)]$ compared to $\lambda(t)\mathsf{E}[S]$ explains why the PSA curve in figure 1 has a much larger amplitude than either the MOL or the exact probability curve. Finally, notice that the PS and MOL curves intersect at the peak of the MOL curve. Observing that from (32), the case of exponentially distributed holding times gives us

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[Q_\infty(t)\big] = \lambda(t) - \mu\mathsf{E}\big[Q_\infty(t)\big], \tag{55}$$

it is clear that $\mathsf{E}[Q_\infty(t)]$ equals $\lambda(t)\mathsf{E}[S]$ (recall that $\mathsf{E}[S] = 1/\mu$) exactly when the derivative of $\mathsf{E}[Q_\infty(t)]$ is zero. These are precisely the times of extreme values for $\mathsf{E}[Q_\infty(t)]$.

Here we have numerical evidence of how the MOL approximation serves as a bridge between offered load models and loss models. This now makes the analysis of offered load models are relevant to the study of loss models. We have also buttressed this approximation technique with exact bounds for the error between MOL and the ex-

act distribution. One of the results is the following for the $M_t/M/L/L$ queue. For the proof, we refer the reader to [Massey and Whitt, 35].

Theorem 3.1 (Massey and Whitt, 1994). If λ is differentiate and bounded on $[0, \infty)$, its derivative λ' is also bounded on $[0, \infty)$, and $\mathsf{E}[Q_\infty(0)] = \lambda(0)/\mu$, then

$$\sup_{t\geqslant 0} \left|\mathsf{E}\big[Q_L(t)\big] - \mathsf{E}\big[Q_\infty(t)\big]\big(1 - \beta_L\big(\mathsf{E}\big[Q_\infty(t)\big]\big)\big)\right| \leqslant \frac{|\lambda| \cdot |\lambda'|}{\mu^3} \cdot \beta_{L-1}\frac{|\lambda|}{\mu},$$

where $|\lambda| = \sup_{t\geqslant 0} |\lambda(t)|$.

3.3. Server staffing for call centers

Now we consider the application of server staffing for the $M_t/M_t/L_t/L_t$ queue. By stochastic ordering, if $Q_L(0) = Q_\infty(0)$, then $\mathsf{Pr}(Q_L(t) \geqslant L) \leqslant \mathsf{Pr}(Q_\infty(t) \geqslant L)$. Given ε, let the relation $(1/\sqrt{2\pi}) \int_{\psi(\varepsilon)}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \varepsilon$ define $\psi(\varepsilon)$. Now we define $L_t(\varepsilon)$ where

$$L_t(\varepsilon) = \Big\lceil q(t) + \psi(\varepsilon)\sqrt{q(t)}\Big\rceil \tag{56}$$

and $q(t) = \mathsf{E}[Q_\infty(t)]$. The motivation for $L_t(\varepsilon)$ comes from the fact that $Q_\infty(t)$ has a Poisson distribution. Thus we have $q(t)$, the mean of $Q_\infty(t)$, plus $\psi(\varepsilon)$ times $\sqrt{q(t)}$, the standard deviation of $Q_\infty(t)$. We then have for all $t \geqslant 0$,

$$\begin{aligned}\mathsf{Pr}\big(Q_\infty(t) \geqslant L_t(\varepsilon)\big) &\leqslant \mathsf{Pr}\big(Q_\infty(t) \geqslant q(t) + \psi(\varepsilon)\sqrt{q(t)}\big)\\ &= \mathsf{Pr}\left(\frac{Q_\infty(t) - q(t)}{\sqrt{q(t)}} \geqslant \psi(\varepsilon)\right) \approx \varepsilon.\end{aligned}$$

The same argument as above gives a lower bound for the delay probability of the $M_t/M_t/L_t$ queue. Using $L_t(\varepsilon)$ approximately gives ε for both the delay and blocking probabilities. For more details on this server staffing algorithm, see [Jennings et al., 18]. This work also led to a Lucent Technologies patent (U.S. patent number 5,923,873).

3.4. Private line services

Now consider the application of loss systems to capacity management process flows for leased service as discussed in [20]. We can model the number of DSO lines in inventory as an $M_t/M/L/L$ queue. For this application $\lambda(t)$ equals average customer arrival rate at time t, $1/\mu$ the average circuit holding time, L the total number of DS0 circuits in stock, and $Q_L(t)$ the random number of leased DS0 circuits at time t. Whereas the average holding time for the $M_t/G/L/L$ in the context of telephone trunklines is 5 min, an average holding time for private line services may be 2 years. Thus we have a communications example where steady state analysis has no relevance.

Given a fixed time interval $[0, T]$, the cost of providing a unit circuit for a unit time a_c, the revenue from a unit circuit for a unit time a_r, and the initial number of

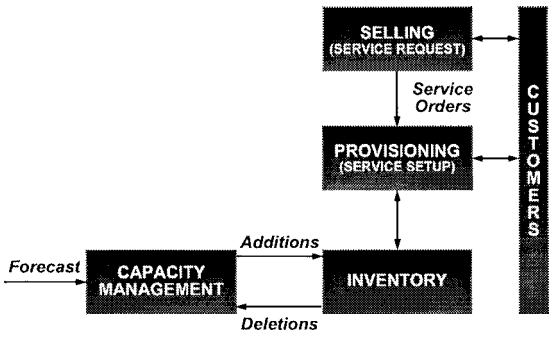

Figure 2. Diagram of capacity management process flows.

DS0 circuits $Q_L(0)$, the primary goal now is to determine the inventory level L of DS0 circuits needed to maximize the profit function Π where

$$\Pi(L) \equiv a_r \int_0^T \mathsf{E}\big[Q_L(t)\big]\,\mathrm{d}t - a_c LT. \tag{57}$$

We use the knowledge of offered load models to construct an approximate profit function which we call the *sorted offered load* approximation (SOL):

$$\Pi_{\mathrm{SOL}}(L) = a_r \int_0^T \min\big(\mathsf{E}\big[Q_\infty(t)\big], L\big)\,\mathrm{d}t - a_c LT. \tag{58}$$

The SOL approximation serves as an upper bound for Π and is significantly easier to compute. It is sufficiently accurate for solving the profit optimality problem. Its optimal solution can be expressed analytically which yields simple rule of thumb interpretations of optimality. Finally, the simple form of its optimal solution leads to a fast algorithm for computing it.

Below is the fundamental result for the sorted offered load approximation.

Theorem 3.2. If we extend Π_{SOL} to be a function of a continuous variable, then it achieves its maximum at

$$L_{\max} = q_*\left(\frac{a_c}{a_r}\,T\right), \tag{59}$$

where g_* is the *decreasing rearrangement* of q and $q(t) = \mathsf{E}[Q_\infty(t)]$. This is also the solution to the equation

$$a_r \frac{1}{T} \int_0^T \mathbb{1}_{\{q(t) > L_{\max}\}}\,\mathrm{d}t = a_c. \tag{60}$$

Moreover, the approximate maximum profit will be

$$\Pi_{\mathrm{SOL}}(L_{\max}) = a_r \int_{(a_c/a_r)T}^T q_*(t)\,\mathrm{d}t. \tag{61}$$

Table 1
Parameter values used in numerical example.

Arrival rate $\lambda(t)$	Holding time $1/\mu$	Initial load $Q_L(0)$	Revenue a_r	Cost a_c	Time interval T
$0.2t + 50$	10	24	10	5	20

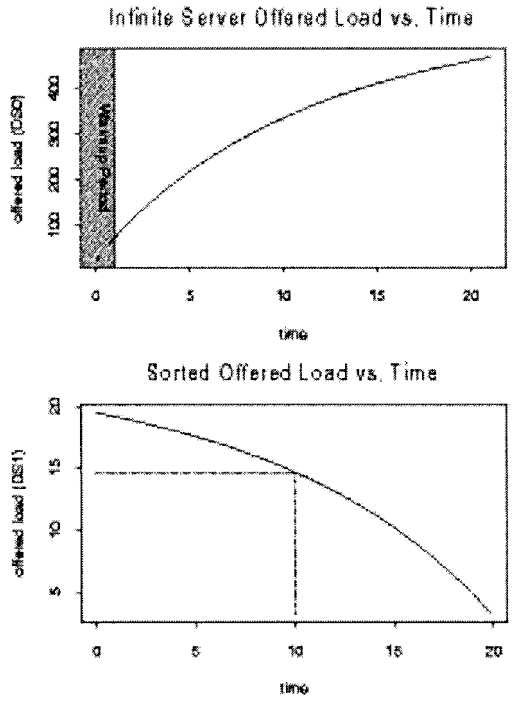

Figure 3. Graphs of the offered load and the sorted offered load as a function of time.

In table 1, we list the specific values that we use for the numerical example in this section. In figure 3 we plot both q and q_* side by side. Since q is an increasing function for this example, then q_* is merely the time reversal over the interval of length 20. Along with DS0 units we also use DS1 units where 1 DS1 channel equals 24 DS0 channels. In figure 3, we plot three quantities that estimate the average profit for a given number L of channels in inventory. The height of the bars corresponding to STAT (the black ones) estimates this profit using the steady state formulas for the carried load (STAT is for stationary). We do this by using the constant arrival rate that is a time average of the given rate function. The bars for SOL (the white ones) approximate the profit by using the sorted offered load estimation for the carried load. Finally, the bars for MOL (the gray ones) uses the modified offered load approximation of the carried load to estimate the profit.

In the case of an increasing arrival rate, we can prove that MOL always underestimates the carried load (see [Massey and Whitt, 35]). Since SOL always overestimates

Figure 4. Graphs of the STAT, SOL, and MOL profits as a function of the total number of DS1 lines provided.

the carried load, we can see by inspection that both of them are doing a good job of approximating the true profit function here. The STAT bars suggest that the maximal profit can be attained by using 22 DSl lines. This is doubly wrong. The true optimal number is 15 DSl lines which SOL estimates accurately and the MOL verifies. Not only is the true maximal profit half of what STAT estimates, STAT suggests a profit that is one fourth the true maximal profit. This example shows where nonstationary analysis makes a significant improvement over steady state analysis.

The above problem is an example of the interaction between queueing models and profit. In general, there is a natural relationship between pricing models for communication services and queues with time-varying rates and we have begun to explore this issue in [Lanning et al., 25]. The arrival rate for a queue can be viewed through the economic lens as "demand." Just as price shapes demand, we can model a queue with time varying rates as one where the arrival rate is a function of price. For loss models, blocking is a quality of service (QoS) constraint that the service provider may sell as a commodity. The queueing and economic issues then intertwine as a given price determines an arrival rate which determines a given QoS level. Queueing analysis can now be used to help service providers determine if they are selling precisely the QoS level that they are promising.

More recently, we have extended this type of economic queueing analysis to bandwidth exchange and provisioning. See [Hampshire et al., 14] as well as [Hampshire et al., 15] for more details.

3.5. Time reversible Markov chains

We now want to generalize the notion of a modified offered load approximation. The key to formulating this generalization lies in the theory of *time reversible Markov chains.* For the purposes of these applications, we do not need to define what a time reversible

Markov chain is (see [Kelly, 22] for details), we need only list some of the properties that these chains possess:

Tree condition. If the state space diagram for the Markov chain is bidirectional and, as an undirected graph, a tree (acyclic), then it is time reversible.

Independent joints. The joint process of two independent, time reversible Markov chains is also a time reversible Markov chain.

Restrictions to subsets. If Γ is a subset of the state space for a Markov chain, delete the transitions of the chain that leave Γ. The new chain inherits time reversibility from the first one. Moreover,

$$\mathsf{Pr}(X_\Gamma = i) = \mathsf{Pr}(X = i \mid X \in \Gamma) \tag{62}$$

for all $i \in \Gamma$, where X and X_Γ are the steady state limits for the original and new chains, respectively.

Using the first and last properties the $M/M/\infty$ queue is time reversible and the $M/M/L/L$ queue is the restriction of the $M/M/\infty$ infinite server model and so (50) is seen to be a special case of (62). Hence the MOL approximation method for the $Mt/G/L/L$ can be viewed as a "time reversibility" approximation.

This suggests a general recipe for constructing MOL approximations for the transient behavior of a Markov chain with time varying rates:

1. Show that this chain is the restriction of a Markov chain on a larger state space and this larger chain would be time reversible if it had constant rates.
2. Solve for the transient behavior of this larger chain.
3. Condition the transient probabilities of the larger chain to stay in the smaller state space of the original chain. This is the MOL approximation in general but also the *exact* answer for the steady state, constant rate case.

In the next section, we apply these generalized MOL approximation methods to a communication network.

3.6. Circuit switched networks

Recall that the classical circuit switch network model, as a Markov chain, is time reversible. This follows from the methods listed here for constructing time reversible Markov chains since

1. The $M/M/\infty$ queue is time reversible.
2. An independent collection of $M/M/\infty$ queues is time reversible.
3. The classical circuit switched network model is the restriction of independent $M/M/\infty$ queues to a smaller state space.

Therefore, the circuit switched network model is a time reversible Markov chain and we have an exact formula for its steady state distribution. Since the transient distribution for an independent collection of $M_t/M/\infty$ queues is known, we then have a MOL approximation for a circuit switched network with time varying rates.

Now we numerically investigate the following example from [Jennings and Massey, 19]: a two-link circuit switched network. We consider the three numerical examples given in our parameter table 2.

Table 2
Parameter values used in numerical example.

CSN case # 1	Arrival rate $\lambda_0(t)$	Arrival rate $\lambda_1(t)$	Arrival rate $\lambda_2(t)$	Channels c_1	Channels c_2
1	$10+5\sin(0.5\pi t)$	$10+5\sin(0.5\pi t)$	$10+5\sin(0.5\pi t)$	36	36
2	$3+\sin(0.4\pi t)$	$10+3\sin(0.4\pi t)$	$3+\sin(0.4\pi t)$	10	10
2	$3+\sin(0.2\pi t)$	$10+3\sin(0.2\pi t)$	$3+\sin(0.2\pi t)$	10	10

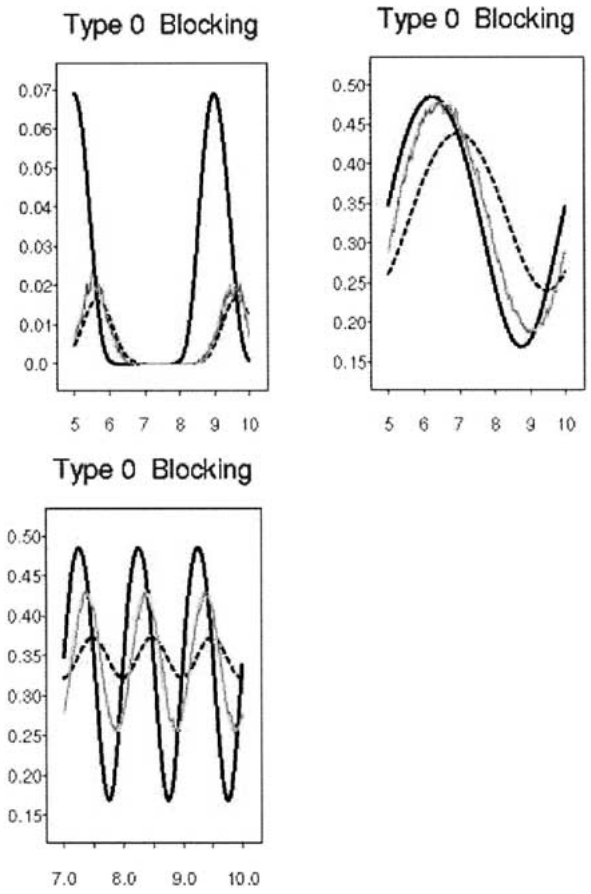

Figure 5. Graph of type 0 blocking probabilities for CSN cases 1, 2 and 3.

In figure 5, we plot the blocking probabilities (left to right, case #1 through case #3) for type 0 traffic (the calls that use channels from both links) where we compare the actual blocking probability (the light gray curves, where the numbers come from averaged numerical simulation), to the MOL approximation (the black dashed curves), and the PS approximation (the black solid curves).

For case #1, the actual blocking is small (under 2%) and MOL does significantly better than PS. For case #2, the actual blocking is high (25–50%) and PS does a better job than MOL. This is due to having arrival rates that are not varying too quickly in time. Finally, the parameters of case #3 differ from case #2 only in a fivefold scaling up of the frequency for the sinusoidal arrival rates. Here, neither the MOL nor the PS approximations work well. In practice, the MOL approximation is very useful since circuit switched network is typically designed for low blocking. Moreover, it can accommodate the changes in the blocking behavior due to different holding time distributions that give the same mean. This is something that the PS approximation cannot do.

This classical circuit switched network model has new applications with time varying rates. It can model the type of alternate routing that occurs when there is a link failure at a specific time. After that moment, all traffic on that link is rerouted.

4. Delay models

In contrast to loss models, another way to deal with limited communication resources is to have customers wait until the necessary resources are available. Now we have a *delay model*. This is a suitable model for communication services like file transfers that are not necessarily done in realtime. The $M_t/M_t/1$ and $M_t/M_t/L_t$ queues are the canonical nonstationary queueing models that we discuss in this section. We develop a theoretical asymptotic analysis framework that rigorizes the fluid and diffusion analysis of [Newell, 37]. The limit theorems obtain here give us approximation methods that complement to work of Ong and Taafe (see [40,41]), [Rolski, 46], as well as [Yin and Zhang, 51].

4.1. The $Mt/M_t/1$ queue

We discuss the first of two canonical examples for delay models. If Q is the steady state number for the $M/M/1$ queue, which has Poisson arrivals, exponentially distributed holding times, a single server, and an infinite buffer, then for all non-negative integers n:

$$\Pr(Q = n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \tag{63}$$

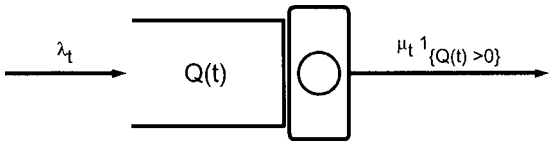

Figure 6. The $M_t/M_t/1$ queue.

for $\lambda < \mu$ (otherwise $\mathsf{Pr}(Q = n) = 0$ for all n) where λ is the mean Poisson arrival rate and the random service time S is exponentially distributed with $1/\mu = \mathsf{E}[S]$. Using Little's law to obtain the mean queueing delay $\mathsf{E}[Q]$ from the mean queue length $\mathsf{E}[Q]$ gives us for $\lambda < \mu$

$$\mathsf{E}[Q] = \frac{\lambda/\mu}{1 - \lambda/\mu} \quad \text{and} \quad \mathsf{E}[D] = \frac{\lambda}{\mu(\mu - \lambda)}. \tag{64}$$

There are two useful methods for constructing the $M_t/M_t/1$ queue length process. First we can construct the *transition probabilities* for the $M_t/M_t/1$ queue by defining the vector $\mathbf{p}(t)$, where

$$\mathbf{p}(t) = \big[\mathsf{Pr}\big(Q(t) = 0\big),\ \mathsf{Pr}\big(Q(t) = 1\big),\ \mathsf{Pr}\big(Q(t) = 2\big), \ldots\big], \tag{65}$$

to be the solution to the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = \mathbf{p}(t)\mathbf{A}(t), \tag{66}$$

where $\mathbf{A}(t)$ is the tridiagonal matrix

$$\mathbf{A}(t) = \begin{bmatrix} -\lambda_t & \lambda_t & & \\ \mu_t & -(\lambda_t + \mu_t) & \lambda_t & \\ & \mu_t & -(\lambda_t + \mu_t) & \ddots \\ & & \ddots & \ddots \end{bmatrix}. \tag{67}$$

The other constructive method is to build the *random sample paths* of the $M_t/M_t/1$ queue out of Poisson process using a reflection mapping. We can show that

$$Q(t) = X(t) - \inf_{0 \leqslant s \leqslant t} X(s) \tag{68}$$

with

$$X(t) \equiv \Pi_1\left(\int_0^t \lambda(s)\,\mathrm{d}s\right) - \Pi_2\left(\int_0^t \mu(s)\,\mathrm{d}s\right), \tag{69}$$

where Π_1 and Π_2 are independent, standard (rate 1) Poisson processes and $Q(0) = 0$.

We use the same method of asymptotic analysis for both constructions and refer to it as *uniform acceleration.* We achieve uniform acceleration by scaling both the arrival rate λ_t and the service rate μ_t with the same parameter η, and analyzing the asymptotic behavior of resulting process Q^η as $\eta \to \infty$. The uniformly accelerated transition probabilities are now given by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}^\eta(t) = \mathbf{p}^\eta(t)\eta\mathbf{A}(t), \tag{70}$$

and the uniformly accelerated random sample paths are given by:

$$Q^\eta(t) = X^\eta(t) - \inf_{0 \leqslant s \leqslant t} X^\eta(s), \tag{71}$$

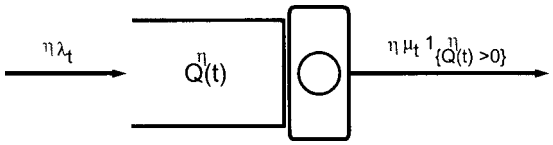

Figure 7. The uniformly accelerated $M_t/M_t/1$ queue.

where

$$X^{\eta}(t) \equiv \Pi_1\left(\int_0^t \eta\lambda(s)\,\mathrm{d}s\right) - \Pi_2\left(\int_0^t \eta\mu(s)\,\mathrm{d}s\right). \tag{72}$$

Uniform acceleration is the nonstationary analogue to steady state analysis. It also generates limit theorems that are the basis for time varying equilibrium, fluid and diffusion approximations of the original queueing system.

The transition probability analysis is given below and proved in both [Massey, 31,32]. Generalizations of this type of analysis can be found in [Yin and Zhang, 51].

Theorem 4.1 [Massey, 31]. If we let

$$\rho^*(t) \equiv \sup_{0\leqslant s<t} \frac{\int_s^t \lambda_r\,\mathrm{d}r}{\int_s^t \mu_r\,\mathrm{d}r}, \tag{73}$$

$\rho^*(t) < 1$, and $\rho(t) = \lambda(t)/\mu(t)$, then

$$\begin{aligned}\mathsf{Pr}\big(Q^{\eta}(t) = n\big) &= \big(1-\rho(t)\big)\rho(t)^n \\ &\quad + \frac{\rho'(t)}{\eta\mu(t)}\left(\frac{\rho(t)}{(1-\rho(t))^2} - \frac{n(n+1)}{2}\right)\rho(t)^{n-1} + \mathrm{O}\left(\frac{1}{\eta^2}\right),\end{aligned}$$

as $\eta \to \infty$. In particular,

$$\mathsf{Pr}\big(Q^{\eta}(t) > 0\big) = \rho(t) - \frac{\rho'(t)}{\eta\mu(t)(1-\rho(t))^2} + \mathrm{O}\left(\frac{1}{\eta^2}\right), \tag{74}$$

and

$$\mathsf{E}\big[Q^{\eta}(t)\big] = \frac{\rho(t)}{1-\rho(t)} - \frac{\rho'(t)(1+\rho(t))}{\eta\mu(t)(1-\rho(t))^4} + \mathrm{O}\left(\frac{1}{\eta^2}\right). \tag{75}$$

Moreover, if $\rho^*(t) \geqslant 1$, then

$$\mathsf{Pr}\big(Q^{\eta}(t) = n\big) \cong 0 \tag{76}$$

for all n as $\eta \to \infty$.

Now we state the limit theorems that follow from the sample path analysis. The proofs of these results can be found in [Mandelbaum and Massey, 27]. All the sample path results in this section draw heavily from the theory of *strong approximations* as discussed in [Ethier and Kurtz, 10].

Theorem 4.2 [Mandelbaum and Massey, 27]. If λ and μ are locally integrable functions, then $\lim_{\eta\to\infty}(1/\eta)Q^{\eta}(t) = Q^{(0)}(t)$ a.s. uniformly on compact sets, where

$$Q^{(0)}(t) = \int_0^t [\lambda_s - \mu_s]\,\mathrm{d}s - \inf_{0\leqslant s\leqslant t}[\lambda_r - \mu_r]\,\mathrm{d}r. \tag{77}$$

Moreover,

$$\lim_{\eta\to\infty}\frac{Q^{\eta}(t) - \eta Q^{(0)}(t)}{\sqrt{\eta}} \stackrel{\mathrm{d}}{=} Q^{(1)}(t),$$

where

$$Q^{(1)}(t) = B\left(\int_0^t [\lambda_s + \mu_s]\,\mathrm{d}s\right) - \inf_{s\in\Phi_t} B\left(\int_0^s [\lambda_r + \mu_r]\,\mathrm{d}r\right), \tag{78}$$

$\{B(t) \mid t \geqslant 0\}$ is standard (mean 0, variance t) Brownian motion, and finally

$$\Phi_t = \left\{0 \leqslant s \leqslant t \;\middle|\; \int_s^t [\lambda_r - \mu_r]\,\mathrm{d}r = Q^{(0)}(t)\right\}. \tag{79}$$

These random sample path limit theorems literally complement the transition probability limit theorems. The former results are zero when the latter results are nonzero and the reverse holds as well.

For more insight into the formulas, we now reduce these fluid and diffusion limits to the constant case. Assume that λ and μ are both constant over the interval $[0, t]$. We then have $Q^{(0)}(t) = (\lambda - \mu)^+ t$, assuming that $Q^{(0)}(0) = 0$, where $x^+ = \max(x, 0)$. We then have

$$\Phi_t = \begin{cases} \{t\}, & \text{if } \lambda < \mu, \\ [0, t], & \text{if } \lambda = \mu, \\ \{0\}, & \text{if } \lambda > \mu, \end{cases} \tag{80}$$

and so

$$Q^{(1)}(t) = \begin{cases} 0, & \text{if } \lambda < \mu, \\ B(2\lambda t) - \inf_{0\leqslant s\leqslant t} B(2\lambda s), & \text{if } \lambda = \mu, \\ B\big((\lambda + \mu)t\big), & \text{if } \lambda > \mu. \end{cases} \tag{81}$$

4.2. Virtual waiting time for the $M_t/G/1$ queue

The same analysis as above gives us a new result for the fluid and diffusion limit of the virtual waiting time (or workload) process for the $M_t/G/1$ queue. We construct a process $\{V(t) \mid t \geqslant 0\}$, where

$$V(t) = Y(t) - \inf_{0\leqslant s\leqslant t} Y(s), \tag{82}$$

with

$$Y(t) = \sum_{n=1}^{A(t)} S_n - t = \int_0^t S_{A(\tau)}\,\mathrm{d}A(\tau) - t. \tag{83}$$

The accelerated version of this process is formed by first representing $A(t)$ as $\Pi(\int_0^t \lambda(\tau)\,\mathrm{d}\tau)$. We then define $V^\eta(t) = Y^\eta(t) - \inf_{0\leqslant s\leqslant t} Y^\eta(s)$, where

$$Y^\eta(t) = \sum_{n=1}^{\Pi(\eta\int_0^t \lambda(\tau)\,\mathrm{d}\tau)} S_n - \eta t. \tag{84}$$

The physical interpretation is that we are simultaneously scaling up the arrival rate of the jobs and their processing rate. The sample path analysis for the $M_t/G/1$ queue is then

Theorem 4.3. If $\{V^\eta(t) \mid t \geqslant 0\}$ is the uniformly accelerated virtual waiting time process, then $\lim_{\eta\to\infty}(1/\eta)V^\eta(t) = V^{(0)}(t)$, where

$$V^{(0)}(t) = \sup_{0\leqslant s\leqslant t} \int_s^t \big(\lambda(\tau)\mathsf{E}[S] - 1\big)\,\mathrm{d}\tau, \tag{85}$$

and $\lim_{\eta\to\infty} \sqrt{\eta}((1/\eta)V^\eta(t) - V^{(0)}(t)) \overset{\mathrm{d}}{=} V^{(1)}(t)$, where

$$V^{(1)}(t) = B\left(\int_0^t \lambda(\tau)\,\mathrm{d}\tau \cdot \mathsf{E}\big[S^2\big]\right) - \inf_{s\in\Psi_t} B\left(\int_0^s \lambda(\tau)\,\mathrm{d}\tau\,\mathsf{E}\big[S^2\big]\right), \tag{86}$$

$\{B(t) \mid t \geqslant 0\}$ is standard Brownian motion, and

$$\Psi_t = \left\{0 \leqslant s \leqslant t \;\middle|\; \int_s^t \big(\lambda(\tau)\mathsf{E}[S] - 1\big)\,\mathrm{d}\tau = V^{(0)}(t)\right\}. \tag{87}$$

4.3. The $M_t/M_t/L_t$ queue

Our second canonical example of a delay model is the $M/M/L$ queue, which has Poisson arrivals, exponentially distributed holding times, L servers, and an infinite buffer.

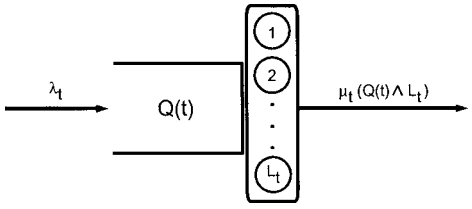

Figure 8. The $M_t/M_t/L_t$ queue.

If Q has the limiting distribution for an $M/M/L$ queue, then for all non-negative integers n,

$$G(\rho)\cdot \mathsf{Pr}(Q=n)=\begin{cases}\dfrac{\rho^n}{n!} & \text{if } n\leqslant L,\\[2ex] \dfrac{\rho^L(\rho/L)^{n-L}}{L!} & \text{if } n>L,\end{cases} \tag{88}$$

where $\rho=\mathsf{E}[Q_\infty]=\lambda\mathsf{E}[S]$ and

$$G(x)=\sum_{j=0}^{L}\frac{x^j}{j!}+\frac{x^L}{L!}\,\frac{x/L}{1-x/L}, \tag{89}$$

whenever $\rho<L$, otherwise $\mathsf{Pr}(Q=n)=0$ for all n. Moreover, the probability of queueing delay is

$$\mathsf{Pr}(Q\geqslant L)=\mathsf{Pr}(D>0)=\frac{\beta_L(\rho)}{1-\rho/L+\beta_L(\rho)\rho/L}. \tag{90}$$

The function β_L is the Erlang B formula given by (48) and (90) is the classical *Erlang C formula* for multi-server queues (see [Cooper, 3, p. 91]).

Now we define the $M_t/M_t/L_t$ queue length process $\{Q(t)\mid t>0\}$ using a sample path construction

$$Q(t)\equiv Q(0)+\Pi_1\biggl(\int_0^t\lambda_s\,\mathrm{d}s\biggr)-\Pi_2\biggl(\int_0^t\mu_s\bigl(Q(s)\wedge L_s\bigr)\,\mathrm{d}s\biggr), \tag{91}$$

where Π_1 and Π_2 are two independent, standard Poisson processes and $x\wedge y=\min(x,y)$.

This motivates us to define a new variation on uniform acceleration for the $M_t/M_t/L_t$ queue, namely

$$\begin{aligned}Q^\eta(t)&\equiv Q^\eta(0)+\Pi_1\biggl(\int_0^t\eta\lambda_s\,\mathrm{d}s\biggr)-\Pi_2\biggl(\int_0^t\eta\mu_s\biggl(\frac{1}{\eta}Q^\eta(s)\wedge L_s\biggr)\,\mathrm{d}s\biggr)\\&=Q^\eta(0)+\Pi_1\biggl(\int_0^t\eta\lambda_s\,\mathrm{d}s\biggr)-\Pi_2\biggl(\int_0^t\mu_s\bigl(Q^\eta(s)\wedge\eta L_s\bigr)\,\mathrm{d}s\biggr).\end{aligned}$$

Defining the service rate for the $M_t/M_t/1$ queue as $\mu_t\mathbb{1}_{\{Q_t>0\}}$, we see that the two uniform acceleration variants are the same for the single server case. These new asymptotics have the call center service provider interpretation of scaling up the *supply* (number of servers) in response to a similar scaling up of the *demand* (arrival rate).

Using the theory of strong approximations (see [Ethier and Kurtz, 10]), we can obtain a sample path analysis that yields the following fluid and diffusion limits.

Theorem 4.4 [Mandelbaum, Massey and Reiman, 28]. If λ and μ are locally integrable functions then $\lim_{\eta\to\infty}(1/\eta)Q^{\eta}(t) = Q^{(0)}(t)$ a.s. converging uniformly on compact sets, and

$$Q^{(0)}(t) = Q^{(0)}(0) + \int_0^t \lambda_s - \mu_s\big(Q^{(0)}(s) \wedge L_s\big)\,\mathrm{d}s. \tag{92}$$

Moreover,

$$\lim_{\eta\to} \frac{Q^{\eta}(t) - \eta Q^{(0)}(t)}{\sqrt{\eta}} \stackrel{\mathrm{d}}{=} Q^{(1)}(t),$$

and

$$\begin{aligned} Q^{(1)}(t) = Q^{(1)}(0) &- \int_0^t \mu_s \mathbb{1}_{\{Q^{(0)}(s)<L_s\}} Q^{(1)}(s)^+\,\mathrm{d}s \\ &+ \int_0^t \mu_s \mathbb{1}_{\{Q^{(0)}(s)\leqslant L_s\}} Q^{(1)}(s)^-\,\mathrm{d}s + B\bigg(\int_0^t \lambda_s + \mu_s\big(Q^{(0)}(s)\wedge L_s\big)\,\mathrm{d}s\bigg). \end{aligned} \tag{93}$$

The construction of the diffusion process involves a general notion of a nonsmooth derivative. If α is a real valued function that has right and left derivatives, then its *scalable Lipschitz derivative* is defined to be

$$\Lambda\alpha(x;\, y) = \alpha'(x+)y^+ - \alpha'(x-)y^-. \tag{94}$$

The function $\Lambda\alpha(x;\,\cdot)$ is *Lipschitz*, i.e. there exists a constant $M = \max(|\alpha'(x+)|, |\alpha'(x-)|)$ such that for all y_1 and y_2 we have

$$\big|\Lambda\alpha(x;\, y_1) - \Lambda\alpha(x;\, y_2)\big| \leqslant M|y_1 - y_2|. \tag{95}$$

The function $\Lambda\alpha(x;\,\cdot)$ is also *scalable*, i.e. we have

$$\Lambda\alpha(x;\, \gamma y) = \gamma\Lambda\alpha(x;\, y) \tag{96}$$

for all $\gamma \geqslant 0$. These derivatives have the following properties.

Theorem 4.5. Scalable Lipschitz differentiability is closed under operations that yield the following formulas:

- Addition formula

$$\Lambda(\alpha+\beta)(x;\, y) = \Lambda\alpha(x;\, y) + \Lambda\beta(x;\, y). \tag{97}$$

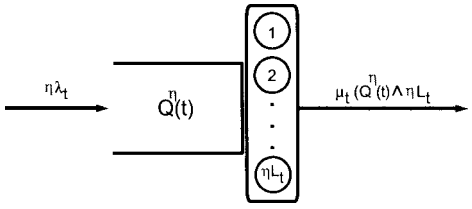

Figure 9. The uniformly accelerated $M_t/M_t/L_t$ queue.

- Multiplication formula

$$\Lambda(\alpha\beta)(x;\, y) = \alpha(x)\Lambda\beta(x;\, y) + \beta(x)\Lambda\alpha(x;\, y). \tag{98}$$

- Composition formula

$$\Lambda(\alpha \circ \beta)(x;\, y) = \Lambda\alpha\big(\beta(x);\, \Lambda\beta(x;\, y)\big). \tag{99}$$

- Maximum formula

$$\begin{aligned}\Lambda(\alpha \vee \beta)(x;\, y) = {} & \Lambda\alpha(x;\, y) \cdot \mathbb{1}_{\{\alpha(x)>\beta(x)\}} + \Lambda\beta(x;\, y) \cdot \mathbb{1}_{\{\alpha(x)<\beta(x)\}} \\ & + \Lambda\beta(x;\, y) \vee \Lambda\alpha(x;\, y) \cdot \mathbb{1}_{\{\alpha(x)=\beta(x)\}}.\end{aligned}$$

We can write the diffusion equation (93) more compactly as

$$Q^{(1)}(t) = Q^{(1)}(0) + \int_0^t \Lambda\alpha\big(Q^{(0)}(s);\, Q^{(1)}(s)\big)\,\mathrm{d}s + B\left(\int_0^t \alpha\big(\big(Q^{(0)}(s) \wedge L_s\big)\big)\,\mathrm{d}s\right)$$

where

$$\alpha(x) = \lambda_t - \mu_t \cdot (x \wedge L_t) \quad \text{and} \quad \alpha((x)) = \lambda_t + \mu_t \cdot (x \wedge L_t), \tag{100}$$

and get the following set of derivative formulas.

Corollary 4.6. The $M_t/M_t/L_t$ fluid limit, given $Q^{(0)}(0)$, is the unique solution to the autonomous differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} Q^{(0)}(t) = \alpha\big(Q^{(0)}(t)\big). \tag{101}$$

Moreover, the mean, variance, and covariance for the $M_t/M_t/L_t$ diffusion limit, solve the derivative formulas:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[Q^{(1)}(t)\big] = \mathsf{E}\big[\Lambda\alpha\big(Q^{(0)}(t);\, Q^{(1)}(t)\big)\big], \tag{102}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Var}\big[Q^{(1)}(t)\big] = 2\mathsf{Cov}\big[Q^{(1)}(t);\, \Lambda\alpha\big(Q^{(0)}(t);\, Q^{(1)}(t)\big)\big] + \alpha\big(\big(Q^{(0)}(t)\big)\big), \tag{103}$$

and for $s < t$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Cov}\big[Q^{(1)}(s),\, Q^{(1)}(t)\big] = \mathsf{Cov}\big[Q^{(1)}(s);\, \Lambda\alpha\big(Q^{(0)}(t);\, Q^{(1)}(t)\big)\big]. \tag{104}$$

Under the following conditions, the diffusion derivative formulas become autonomous differential equations.

Corollary 4.7. If the set of time points

$$\big\{t \mid Q^{(0)}(t) = L_t\big\} \tag{105}$$

has measure zero, then $\{Q^{(1)}(t) \mid t \geqslant 0\}$ is a Gaussian process whose mean, variance, and autocovariance solve the autonomous differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\left[Q^{(1)}(t)\right] = -\mu_t \mathbb{1}_{\{Q^{(0)}(t) \leqslant L_t\}} \mathsf{E}\left[Q^{(1)}(t)\right], \tag{106}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Var}\left[Q^{(1)}(t)\right] = -2\mu_t \mathbb{1}_{\{Q^{(0)}(t) \leqslant L_t\}} \mathsf{Var}\left[Q^{(1)}(t)\right] + \lambda_t + \mu_t\left(Q^{(0)}(t) \wedge L_t\right), \tag{107}$$

and for $s < t$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Cov}\left[Q^{(1)}(s), Q^{(1)}(t)\right] = -\mu_t \mathbb{1}_{\{Q^{(0)} \leqslant L_t\}} \mathsf{Cov}\left[Q^{(1)}(s), Q^{(1)}(t)\right]. \tag{108}$$

The times that $Q^{(0)}(t) = L_t$ are defined to be times of *critical loading.* Notice that a similar result can be seen in the constant rate case for the $M/M/1$ queue (see (81)). Here critical loading corresponds to the case of $\lambda = \mu$. When this occurs over the entire interval $[0, t]$, we have the only situation for the $M/M/1$ queue that $Q^{(1)}$ is reflecting Brownian motion and *not* a Gaussian process.

More recently, we have shown that this fluid and diffusion analysis extends to the virtual waiting time process for multiserver queues (see [Mandelbaum et al., 30]).

Finally, this analysis is a special case of a more general asymptotic analysis for Markovian service networks as discussed in [Mandelbaum et al., 28]. These networks incorporate many of the important features for call centers including: multiple servers, abandonment, priorities, bulk arrivals, network routing based on service completion, and network routing based on abandonment. We also show in [Mandelbaum et al., 28] that all Markovian service networks have fluid and diffusion limit theorems. Moreover, these fluid and diffusion approximations solve differential equations of a significantly smaller dimension than the that of the state space for the original queueing process.

Acknowledgements

The author's introduction to nonstationary queues starts with his Ph.D. thesis on the $M_t/M_t/1$ queue [Massey, 31], later published in [32]. I would like to thank my former thesis advisor Joseph Keller who first suggested to me this area of research back in 1979. I would also like to acknowledge all the co-authors that I have worked with in the area of queues with time-varying rates. They are Steve Eick, Jimmie Davis Jr., Nick Duffield, Nathaniel Grier, Robert Hampshire, Otis Jennings, Steve Lanning, Kin Leung, Avi Mandelbaum, Clement McCalla, Tyrone McKoy, Debasis Mitra, Geraldine Parker, Martin Reiman, Brian Rider, Alexander Stolyar, Qiong Wang, and Ward Whitt.

References

[1] P.J. Burke, The output of a queueing system, Operations Research 4(6) (1956) 699–704.
[2] F.H. Clarke, *Optimization and Nonsmooth Analysis* (SIAM Philadelphia, PA, 1990).

[3] R.B. Cooper, *Introduction to Queueing Theory* (Elsevier, North-Holland, Amsterdam, 1981).
[4] D.J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes* (Springer, New York, 1988).
[5] J.L. Davis, W.A. Massey and W. Whitt, Sensitivity to the service-time distribution in the nonstationary Erlang loss model, Management Science 41(6) (June 1995) 1107–1116.
[6] N.G. Duffield, W.A. Massey and W. Whitt, A nonstationary offered load model for packet networks, *Telecommunication Systems* 13(3/4) (March/April 2001) 271–296.
[7] S. Eick, W.A. Massey and W. Whitt, Infinite-server queues with sinusoidal arrival rates, Management Science 39 (January 1993) 241–252.
[8] S. Eick, W.A. Massey and W. Whitt, The physics of the $M(t)/G/\infty$ queue, Operations Research 41 (July/August 1993) 400–408.
[9] A.K. Erlang, Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges, The Post Office Electrical Engineers' Journal 10 (from the 1917 article in Danish in Elektroteknikeren, Vol. 13) (1918) 189–197.
[10] S.N. Ethier and T.G. Kurtz, *Markov Process: Characterization and Convergence* (Wiley, New York, 1986).
[11] L. Green and P.J. Kolesar, The pointwise stationary approximation for queues with nonstationary arrivals, Management Science (January 1991).
[12] N. Grier, W.A. Massey, T. McKoy and W. Whitt, The time-dependent Erlang loss model with retrials, Telecommunication Systems 7 (1997) 253–265.
[13] R. Hall, *Queueing Methods: For Services and Manufacturing* (Prentice-Hall, Englewood Cliffs, NJ, 1991).
[14] R.C. Hampshire, W.A. Massey, D. Mitra and Q. Wang, Provisioning for bandwidth trading, in: *Telecommunication Network Design and Management* (Kluwer Academic, Dordrecht, 2002) pp. 207–226.
[15] R.C. Hampshire, W.A. Massey and Q. Wang, Dynamic pricing for on-demand bandwidth services, Bell Labs Technical Report (2002).
[16] J.R. Jackson, Jobshop-like queueing systems, Managment Science 10(1) (October 1963) 131–142.
[17] D.L. Jagerman, Nonstationary blocking in telephone traffic, Bell System Technical Journal 54 (1975) 625–661.
[18] O.B. Jennings, A. Mandelbaum, W.A. Massey and W. Whitt, Server staffing to meet time-varying demand, Management Science 42(10) (October 1996) 1383–1394.
[19] O.B. Jennings and W.A. Massey, A modified offered load approximation for nonstationary circuit switched networks, in: *Selected Proceedings of the 3rd INFORMS Telecommunications Conf.*, Vol. 7, 1997, pp. 253–265.
[20] O.B. Jennings, W.A. Massey and C. McCalla, Optimal profit for leased lines services, in: *Proc. of the 15th Internat. Teletraffic Congress*, June 1997, pp. 803–814.
[21] J. Keilson and L.D. Servi, Networks of non-homogeneous $M/G/\infty$ systems, GTE Laboratories, Waltham, MA (1989).
[22] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, New York, 1979).
[23] F.P. Kelly, Notes on effective bandwidths, in: *Stochastic Networks: Theory and Applications*, eds. F.P. Kelly, S. Zachary and I.E. Ziedins, Royal Statistical Society Lecture Notes Series, Vol. 4 (Oxford Univ. Press, Oxford, 1996) pp. 141–168.
[24] T.G. Kurtz, Strong approximation theorems for density dependent Markov chains, Stochastic Processes and Their Applications 6 (1978) 223–240.
[25] S. Lanning, W.A. Massey, B. Rider and Q. Wang, Optimal pricing in queueing systems with quality of service constraints, in: *Teletraffic Engineering in a Competitive World, Proc. of the 16th Internat. Teletraffic Congress*, 1999, pp. 747–756.
[26] K. Leung, W.A. Massey and W. Whitt, Traffic models for wireless communications networks, IEEE Journal on Selected Areas in Communications 12 (October 1994) 1353–1364.

[27] A. Mandelbaum and W.A. Massey, Strong approximations for time dependent queues, MOR 20(1) (February 1995) 33–64.
[28] A. Mandelbaum, W.A. Massey and M.I. Reiman, Strong approximations for Markovian service networks, Queueing Systems 30 (1998) 149–201.
[29] A. Mandelbaum, W.A. Massey, M.I. Reiman and B. Rider, Time varying multiserver queues with abandonment and retrials, in: *Teletraffic Engineering in a Competitive World, Proc. of the 16th Internat. Teletraffic Congress*, 1999, pp. 355–364.
[30] A. Mandelbaum, W.A. Massey, M.I. Reiman, B. Rider and A. Stolyar, Queue lengths and waiting times for multiserver queues with abandonment and retrials, in: *Selected Proceedings of the 5th INFORMS Telecommunications Conf.*, to appear.
[31] W.A. Massey, Nonstationary queues, Stanford University (1981) Ph.D. thesis.
[32] W.A. Massey, Asymptotic analysis of the time dependent $M/M/1$ queue, MOR 10 (May 1985) 305–327.
[33] W.A. Massey, G.A. Parker and W. Whitt, Estimating the parameters of a nonhomogeneous Poisson processes with linear rate, Telecommunication Systems 5 (1996) 361–388.
[34] W.A. Massey and W. Whitt, Networks of infinite-server queues with nonstationary Poisson input, Queueing Systems 13(1) (May 1993) 183–250.
[35] W.A. Massey and W. Whitt, An analysis of the modified offered load approximation for the nonstationary Erlang loss model, Annals of Applied Probability 4(4) (November 1994) 1145–1160.
[36] W.A. Massey and W. Whitt, A stochastic model to capture space and time dynamics in wireless communication systems, Probability in the Engineering and Informational Sciences 8 (1994) 541—569.
[37] G.F. Newell, Queues with time-dependent arrival rates (parts I–III), Journal of Applied Probability 5 (1968) 436–451 (I), 579–590 (II), 591–606 (III).
[38] G.F. Newell, Approximate stochastic behavior of n-server service systems with large n, in: *Lecture Notes in Economics and Mathematical Systems*, Vol. 87 (Springer, Berlin, 1973).
[39] G.F. Newell, *Applications of Queueing Theory* (Chapman and Hall, London, 1982).
[40] K.L. Ong and M.R. Taaffe, Approximating nonstationary $Ph_t/M_t/S/C$ queueing systems, Annals of Operations Research 8 (1987) 103–116.
[41] K.L. Ong and M.R. Taaffe, Nonstationary queues with interrupted Poisson arrivals and unreliable/repairable servers, Queueing Systems 4 (1989) 27–46.
[42] C. Palm, Intensity variations in telephone traffic, Ericsson Technics 44 (1943) 1–189 (in German); English translation (North-Holland, Amsterdam, 1988).
[43] A. Prékopa, On Poisson and composed Poisson stochastic set functions, Studies in Mathematics 16 (1957) 142–155.
[44] A. Prékopa, On secondary processes generated by a random point distribution of Poisson type, Annales Univ. Sci. Budapest de Eötvös Nom. Sectio Math. 1 (1958) 153–170.
[45] R.T. Rockafellar, *Convex Analysis* (Princeton Univ. Press, Princeton, 1970) (paperback edition 1997).
[46] T. Rolski, Queues with nonstationary inputs, Queueing Systems 5(1–3) (1989) 113–129.
[47] M.H. Rothkopf and S.S. Oren, A closure approximation for the nonstationary $M/M/s$ queue, Management Science 25 (June 1979) 522–534.
[48] R. Serfozo, *Introduction to Stochastic Networks* (Springer, Berlin, 1999).
[49] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis, Queues, Communication, and Computing* (Chapman and Hall, New York, 1995).
[50] W. Willinger and V. Paxson, Where mathematics meets the Internet, Notices of the American Mathematical Society 5(8) (September 1998) 961–970; `http://www.ams.org/notices/199808/paxson.pdf`.
[51] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach* (Springer, New York, 1998).