Control and recovery from rare congestion events in a large multi-server system

N.G. Duffield and W. Whitt

AT&T Laboratories, Room {A175, A117}, 180 Park Avenue, Florham Park, NJ 07932-0971, USA
E-mail: {duffield,wow}@research.att.com

Received 1 August 1996; revised 10 February 1997

We develop deterministic fluid approximations to describe the recovery from rare congestion events in a large multi-server system in which customer holding times have a general distribution. There are two cases, depending on whether or not we exploit the age distribution (the distribution of elapsed holding times of customers in service). If we do not exploit the age distribution, then the rare congestion event is a large number of customers present. If we do exploit the age distribution, then the rare event is an unusual age distribution, possibly accompanied by a large number of customers present. As an approximation, we represent the large multi-server system as an $M/G/\infty$ model. We prove that, under regularity conditions, the fluid approximations are asymptotically correct as the arrival rate increases. The fluid approximations show the impact upon the recovery time of the holding-time distribution beyond its mean. The recovery time may or not be affected by the holding-time distribution having a long tail, depending on the precise definition of recovery. The fluid approximations can be used to analyze various overload control schemes, such as reducing the arrival rate or interrupting services in progress. We also establish large deviations principles to show that the two kinds of rare events have the same exponentially small order. We give numerical examples showing the effect of the holding-time distribution and the age distribution, focusing especially on the consequences of long-tail distributions.

Keywords: multi-server systems, high congestion, recovery from congestion, overload control, long-tail distributions, transient behavior, fluid limits, fluid approximations, large deviations, Sanov's theorem, residual lifetimes, age distributions

1. Introduction

In this paper we study recovery from congestion in a large multi-server system. Our motivating application is a link in a high-bandwidth multi-service communication network, but there are many possible applications. We assume that service requests arrive according to a Poisson process with rate λ and require a certain bandwidth for a random holding (service) time with general cumulative distribution function (cdf) G having mean γ. As a simplifying assumption, we let the required bandwidth be 1 for each customer. (See section 6 for multi-class extensions, where different classes have different holding-time cdf's and required bandwidths.) We consider how the system recovers from congestion, with and without intervention, where the intervention may be to reduce the arrival rate or to interrupt some services in progress. Our goal is to obtain insights for system design and overload control.

We focus on large systems, where the offered load $\lambda\gamma$ is large and the capacity is even larger, so that demand exceeding capacity is a rare event. The size enables us to apply relatively simple deterministic fluid approximations in order to describe the transient behavior of the system. We show that these fluid approximations are asymptotically correct as the size increases.

Numerical results are not difficult to obtain when the holding-time distribution is exponential, e.g., see Abate and Whitt [2] and Davis, Massey and Whitt [13]. However, here we are primarily interested in non-exponential holding-time distributions. We focus on the impact of the holding-time cdf G beyond its mean. For instance, we investigate the consequence of G having a long tail; e.g., G might have a power tail, i.e., $G^c(t) \sim \alpha t^{-\beta}$ as $t \to \infty$, where G^c is the complementary cdf (ccdf), i.e., $G^c(t) = 1 - G(t)$, α and β are positive constants and $f(t) \sim g(t)$ means that $f(t)/g(t) \to 1$ as $t \to \infty$. We are interested in long-tail distributions because they are frequently reported in measurements of existing communications networks, e.g., see Cáceres, Danzig, Jamin and Mitzel [8], Leland, Taqqu, Willinger and Wilson [24], Paxson [28] and Crovella and Bestavros [12]. In models of stationary traffic, Willinger, Taqqu, Sherman and Wilson [31] have proved that long-tail on and off times in individual sources can cause self-similarity seen in aggregate traffic.

It is somewhat difficult to anticipate the impact of a long-tail holding-time cdf G upon recovery from congestion, because there is somewhat conflicting evidence. First, it is known that long-tail service times cause long-tail waiting times in single-server queues, e.g., see Abate, Choudhury and Whitt [1]. In that setting the impact can be great. However, in the fluid limit arising under heavy loads, the service-time cdf beyond its mean ceases to matter in single-server queues, e.g., see Chen and Mandelbaum [9] and Choudhury, Mandelbaum, Reiman and Whitt [11].

That background is not especially relevant, though, because the model here is *not* a single-server queue. If we consider the $M/G/s/0$ multi-server loss model or the $M/G/\infty$ infinite-server model, then the steady-state distribution of the number of busy servers has the *insensitivity* property, i.e., that distribution depends on the holding-time cdf G only through its mean. However, the *transient* behavior of these models does *not* have the insensitivity proper, e.g., see Eick, Massey and Whitt [16] and Davis, Massey and Whitt [13]. Moreover, when the system gets large, the holding-time cdf beyond its mean continues to matter. We will show that the deterministic fluid limit describing recovery from congestion in a multi-server system as the arrival rate increases depends strongly on the holding-time cdf G. On the other hand, the long-tail character of G might not have serious consequences. We will show that the impact of the cdf G beyond its mean upon recovery depends on how recovery is defined.

In order to simplify analysis, we will only consider the $M/G/\infty$ infinite-server model, but the conclusions are also applicable more generally to $M/G/s/r$ models with s servers and r extra waiting spaces, provided that the offered load (mean number of busy servers) $\lambda\gamma$ is indeed large and s is even larger; e.g., see example 2 in section 3. (It appears that the limit theorems in sections 7–10 extend to such finite-capacity systems, but that remains to be proven.) This paper continues to develop the idea that infinite-server models are very useful for understanding the behavior of multi-server systems; for related previous work, see Davis, Massey and Whitt [13], Eick, Massey and Whitt [16], Glynn and Whitt [19], Leung, Massey and Whitt [25] and Massey and Whitt [27]. This paper also continues to develop the idea that a large size, when viewed correctly, need not lead to excessive computational complexity, but instead can lead to statistical regularity and a simplification in performance analysis; for related previous work, see Chen and Mandelbaum [9], Glynn and Whitt [19] and Krichagina and Puhalskii [22].

It should be noted that the $M/G/\infty$ model is also of interest because it arises as the limit of the superposition of N possibly heterogeneous on-off (0–1 valued) processes as $N \to \infty$ with the mean off time increasing so that the overall mean remains fixed. Analysis similar to what we do here applies to superpositions of independent on-off processes (and more general component processes). We intend to discuss recovery from congestion in such alternative models elsewhere.

In addition to serving as an approximation for finite-capacity models, the infinite-capacity model is directly applicable to systems for which bandwidth allocations are somewhat elastic. We are thinking of applications which can function down to very low bandwidths, but for which a certain minimum allocation is desirable. If the allocation falls below this level, either too frequently or for too long, then the quality of service received by the application is deemed insufficient. A concrete example is packet telephony running over a link operating a rate-based feedback mechanism, with the application responding to rate-control messages by using a coarser encoding. Each call would use up to c units of bandwidth, but calls would never be blocked. If the total bandwidth is B and the number of calls present is n, then the allocated bandwidth would be $\min\{c, B/n\}$. As a first approximation, it seems reasonable to assume that the call holding time is independent of the allocated bandwidth, so that the $M/G/\infty$ model is directly appropriate.

In the $M/G/\infty$ model the obvious high-congestion event is having a large number of customers in the system. The likelihood of such an event is easy to determine because the steady-state number of customers in the system, denoted by N, has a Poisson distribution with mean $m = \lambda\gamma$, i.e.,

$$P(N \geqslant n) = \sum_{k=n}^{\infty} \frac{\mathrm{e}^{-m} m^k}{k!}. \tag{1.1}$$

From the point of view of (1.1), the cdf G beyond its mean plays no role, but (1.1) does not tell the whole story. The congestion is usually regarded as worse when the number of customers remains high for a longer period of time. This is certainly the case when there is inflow to a buffer with a rate proportional to the content of the $M/G/\infty$ system above a certain level, as in ATM switch models.

Let $N(t)$ be the number of customers in service in the $M/G/\infty$ system as a function of time t. We say that the system has recovered from a rare congestion event at time 0 when $N(t)$ first reaches a *recovery level* k with $m < k < n$. Thus, assuming that the rare congestion event is an unusually large number of customers present initially, i.e., the event $\{N(0) = n\}$, the *recovery time* is

$$T \equiv T_{n,k} = \inf\bigl\{t \geqslant 0\colon\ N(t) \leqslant k \mid N(0) = n\bigr\}. \tag{1.2}$$

We are interested in determining how the distribution of T depends on the parameters $m,\ n,\ k,\ \lambda$ and the holding-time cdf G. We show that, unlike in (1.1), the holding-time cdf G beyond its mean can play a big role in (1.2).

We are particularly interested in the distribution of T when $\lambda,\ m$ and n are large. In that case, we show that the recovery time T is approximately constant, being equal to an associated recovery time for the mean, which can be regarded as a fluid limit (shown in section 7). Paralleling (1.2), the *recovery time for the mean* is

$$\tau \equiv \tau_{n,k} = \inf\bigl\{t \geqslant 0\colon\ E\bigl(N(t) \mid N(0) = n\bigr) \leqslant k\bigr\} \tag{1.3}$$

for $m < k < n$. More generally, we suggest focusing on the time-dependent conditional mean $E(N(t) \mid N(0) = n)$. It turns out to be remarkably tractable.

A complication in our discussion of recovery is the precise meaning of the rare congestion event $\{N(0) = n\}$. There is no difficulty if the holding-time distribution is exponential, because then the stochastic process $\{N(t)\colon\ t \geqslant 0\}$ is Markov. However, we want to treat non-exponential holding-time distributions. A key to the remarkably simple analysis here is the assumption that the $M/G/\infty$ system is in steady state at time 0. In the steady-state setting, we condition on the event $\{N(0) = n\}$. However, in most applications we actually want to define the rare congestion event as a *first hitting time* of the level n by the process $N(t)$. To properly formalize the meaning for successive hitting times, we can let the hitting time be the first time to hit the high level n after first hitting a suitably low level, such as the mean m or 0. (When the arrival rate is not too large, 0 should suffice (and will be regenerative), but when the arrival rate is large, 0 will be visited too rarely.)

We propose our steady-state rare event, not only as an approximation for what we see at a random (steady-state) time, but also as an approximation for the associated hitting-time rare event. We conjecture that this approximation is asymptotically correct as $n \to \infty$, by which we mean that the age distribution conditional on $\{N(0) = n\}$ has the same distribution asymptotically as $n \to \infty$ for both definitions of the rare event $\{N(0) = n\}$, i.e., when it is a hitting time and when the system is in steady-state. Intuitively, this asymptotic equivalence is believable if we note that $P(N \geqslant n + 1 \mid N \geqslant n) \to 0$ as $n \to \infty$ when N has a Poisson distribution.

We also establish supporting theory for the hitting-time approximation in the case of a large arrival rate. Conditioning on $\{N(0) = n\}$ in steady state, we show that in the limit as $\lambda \to \infty$ and $n \to \infty$ with $n/\lambda \to \delta > \gamma$ the normalized path before time 0 is increasing. (See the corollary to theorem 6 in section 7.) We are able to describe the path before time 0 as well as the path after time 0, because the $M/G/\infty$ system is time reversible. We are able to characterize the entire path, not just the behavior at finitely many time points, by establishing a functional law of large numbers.

A second major theme in this paper is the idea that the notion of rare congestion in an $M/G/\infty$ system should include more than just large values for $N(t)$. In order to make the stochastic process $\{N(t)\colon\ t \geqslant 0\}$ a Markov process, we need to append the elapsed holding times (ages) of the $N(t)$ calls in progress at time t. (Alternatively, we could obtain a Markov process by appending the residual holding times of each customer in service at time t, but the residual times typically are not yet known at time t.) When the holding-time distribution is not nearly exponential, observed ages give important information about the residual holding times. Even as the arrival rate λ increases, the ages play an important role in the evolution of the system. Hence we also consider the empirical age distribution at time t, denoted by $A(t)$; $A(t,x)$ is the fraction of the $N(t)$ calls in progress with ages less than or equal to x.

Since the process $\{(N(t), A(t))\colon\ t \geqslant 0\}$ is Markov, it is natural to consider rare events for the pair $(N(t), A(t))$. We propose considering possible deviations of $A(t)$ from its mean as well as large values for $N(t)$. We relate the rarity of high values of $N(t)$ to the rarity of exceptional age distributions $A(t)$ by establishing a joint large deviations principle (LDP) for $N(t)$ and $A(t)$ as the arrival rate λ increases. The joint LDP for $N(t)$ and $A(t)$ implies LDPs for $N(t)$ and $A(t)$ separately. The LDP for $A(t)$ enables us to understand how likely are various possible age distributions. The two LDPs show that the two kinds of rare events have probabilities of the same exponentially small order. Thus, it is appropriate to consider unusual age distributions as well as large numbers of customers.

We also relate the age distribution $A(t)$ to the residual holding-time distribution $R(t)$; $R(t,x)$ is the fraction of the $N(t)$ calls in progress with residual holding times less than or equal to x. The same LDP applies to the pair $(N(t), R(t))$ as to $(N(t), A(t))$. This LDP helps reveal the likelihood of alternative recovery paths starting from $N(0) = n$. We mention here that the transient response of the LD rate-function describing buffer overflow after conditioning on a rare event has been investigated for single server queues by Duffield [15].

Here is how the rest of this paper is organized. In sections 2–7 we consider the case in which we do not use the age distribution, while in sections 8–12 we consider the case in which we do use the age distribution. In section 2 we present the basic supporting $M/G/\infty$ theory and the recovery equation for the mean, which approximates the recovery time of the process in (1.2). In section 3 we make comparisons with $M/M/s/0$ numerical results to show that the infinite-server results can serve as useful approximations for systems with finitely many servers. In section 4 we discuss applications to overload control. In section 5 we discuss approximations for losses and delays when the process $N(t)$ represents the random input rate into a service facility that processes fluid at a fixed rate c. In section 6 we discuss extensions to multiple classes. In section 7 we present limit theorems as $\lambda \to \infty$ and $n \to \infty$ showing that the approximation is asymptotically correct.

We begin considering the second case in which the age distribution is used in section 8. We show the importance of the ages in predicting the residual holding times when the holding-time distribution has a long tail. In section 9 we consider how the ages can be exploited in congestion control via call interruption. In section 10 we show that the new conditional mean in this setting is asymptotically correct for large systems. In section 11 we study the asymptotic behavior of the conditional mean as $t \to \infty$, focusing especially on the impact of long-tail distributions. Finally, in section 12 we present the supporting large deviation principles.

2. The conditional mean and the recovery equation

The $M/G/\infty$ model structure enables us to treat the customers in service independently of each other and of new arrivals. The following fundamental independence result characterizes the distribution of the recovery time T in (1.2) and serves as a basis for the limit theorems we prove later. The result can be proved by exploiting the fact that the arrival-time and service-time pairs constitute a Poisson random measure in the plane; e.g., see theorem 2.1 and (20) in Eick et al. [16] and references cited in [16]. A proof of part (a) is given in the appendix of [20]. Part (b) is contained in [16].

Theorem 1. (a) Conditional on $N(0) = n$, the n elapsed and n residual holding times at time 0 are each distributed as n i.i.d. random variables with the stationary-excess cdf associated with the holding-time cdf G, i.e.,

$$G_e(t) = \frac{1}{\gamma} \int_0^t G^c(u)\,\mathrm{d}u, \quad t \geqslant 0. \tag{2.1}$$

(b) The number of new arrivals after time 0 in the service at time t, denoted by $N_0(t)$, is independent of $N(0)$ and the residual holding times of the $N(0)$ initial customers and has a Poisson distribution with mean $mG_e(t)$ for each $t \geqslant 0$, where $m = \lambda\gamma$ and G_e is the stationary-excess cdf in (2.1).

Theorem 1 implies that the conditional mean and variance have remarkably simple expressions.

Corollary. The conditional number $(N(\pm t) \mid N(0) = n)$ has mean

$$M_n(t) \equiv E\big(N(t) \mid N(0) = n\big) = E\big(N(-t) \mid N(0) = n\big) = m + (n - m)G_e^c(t), \tag{2.2}$$

variance

$$\begin{aligned} V_n(t) &\equiv \mathrm{Var}\big(N(t) \mid N(0) = n\big) \\ &= \mathrm{Var}\big(N(-t) \mid N(0) = n\big) = nG_e^c(t)G_e(t) + mG_e(t) \end{aligned} \tag{2.3}$$

and is asymptotically normally distributed as $n \to \infty$ and $m \to \infty$.

Proof. Let $S_n(t)$ be the number of the original n customers remaining in the system at time t. By theorem 1,

$$\big(N(t) \mid N(0) = n\big) \overset{d}{=} S_n(t) + N_0(t), \tag{2.4}$$

where $\overset{d}{=}$ denotes equality in distribution (as stochastic processes with $t \geqslant 0$), $S_n(t)$ and $N_0(t)$ are independent stochastic processes and $S_n(t)$ has a binomial distribution for each t, i.e.,

$$P\big(S_n(t) = k\big) = \binom{n}{k} G_e^c(t)^k G_e(t)^{n-k}. \tag{2.5}$$

Formulas (2.2) and (2.3) are simple consequences. We get $(N(-t) \mid N(0) = n) \overset{d}{=} (N(t) \mid N(0) = n)$ because the $M/G/\infty$ system in steady-state is time reversible. Finally, as $n \to \infty$ and $m \to \infty$, after the usual normalization, the binomial and Poisson distributions converge to normal distributions, and the convolution of two normal distributions is a normal distribution. □

The corollary to theorem 1 allows us to characterize the recovery time for the mean.

Theorem 2. If there is no t such that $G^c(t-) > 0 = G^c(t)$, then $G_e^c(t)$ is continuous and strictly increasing. Then the recovery time for the mean, τ in (1.3), is the unique root t of the recovery equation

$$G_e^c(t) = \frac{k-m}{n-m}. \tag{2.6}$$

More generally,

$$\tau = \sup\left\{t\colon\ G_e^c(t) > \frac{k-m}{n-m}\right\}. \tag{2.7}$$

Proof. We apply the conditional mean in (2.2). Note that G_e^c has positive density $g_e(t) = G^c(t)/\gamma$ on the support of G, so that G_e^c is strictly decreasing and continuous on the support of G. The condition implies that $G_e^c(t) \to 0$ as t approaches the upper limit of support. Hence, there necessarily is a unique root of equation (2.6). If the condition does not hold, then there is a t^* such that $G_e^c(t^*-) > 0 = G_e^c(t^*)$ and G_e^c is continuous and strictly decreasing on $[0, t^*]$. In this case, if

$$G_e^c(t^*-) \geqslant \frac{k-m}{n-m} > G_e^c(t^*) = 0,$$

then clearly $\tau = t^*$. □

The recovery time for the mean is easy to calculate, e.g., by bisection search, because G_e^c is monotone. To illustrate how tractable the recovery equation in (2.6) is,

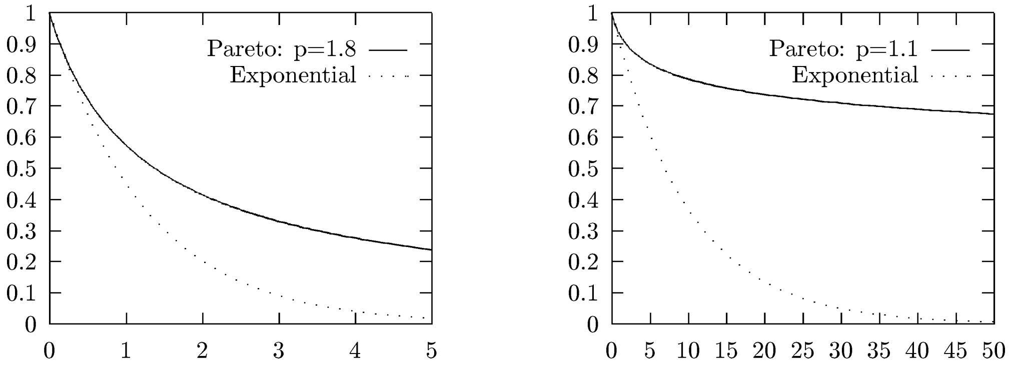

Figure 1. Stationary-excess ccdf for Pareto and Exponential distributions with common mean.

suppose that G is exponential with mean 1. Then $G_e^c(t) = G^c(t) = \mathrm{e}^{-t}$, $t \geqslant 0$, so that the recovery time for the mean is given explicitly by

$$\tau = \log\bigl((n-m)/(k-m)\bigr). \tag{2.8}$$

For a second example, let g be the Pareto density $p(1+t)^{-p-1}$, $t \geqslant 0$, for $p > 1$, which has mean $\gamma = (p-1)^{-1}$. Then the stationary-excess ccdf is $G_e^c(t) = (1+t)^{1-p}$ and

$$\tau = \left(\frac{n-m}{k-m}\right)^{1/(p-1)} - 1. \tag{2.9}$$

For a small system, the recovery time for the mean might not be very revealing, because the random recovery time T in (1.2) need not be close to τ. Indeed, even the expected recovery time ET need not be close to τ. However, τ is meaningful for large systems, because then T, ET and τ should all be close, as we show in section 7.

From (2.6) we can see that the impact of the cdf G on τ depends on whether the recovery level k is close to n or close to m. When the recovery level k is close to the m, the cdf G beyond its mean matters greatly. However, possibly counter to intuition, *when the recovery level k is not too far from the initial level n, the cdf G beyond its mean matters little.* If the recovery point k is closer to n than m, then the recovery time τ depends more on the initial portion of the cdf G_e than upon its tail. Note that the density of G_e is $g_e(t) = G^c(t)/\gamma$, which has the value $1/\gamma$ at the origin; this clearly depends upon the cdf G only through its mean. Since recovery in applications may actually not require returning to a level close to the mean, the holding-time cdf beyond its mean may indeed not matter much.

Consistent with this observation, we find that the impact of a long-tail cdf on τ differs little from the impact of a short-tail cdf with the same mean if k is suitably close to n, but it differs dramatically if k is suitably close to m. However, “suitably close” depends on the decay rate of the ccdf G_e^c.

Example 1. We illustrate by comparing the normalized conditional mean $[E(N(t) \mid N(0) = n) - m]/(n - m)$ for an exponential cdf and a Pareto cdf with a common mean in figure 1. As above, we consider the Pareto density $g(t) = p(1 + t)^{-p-1}$, $t \geqslant 0$, which has mean $\gamma = 1/(p - 1)$ and stationary-excess ccdf $G_e^c(t) = (1 + t)^{1-p}$. The cases of $p = 1.8$ and $p = 1.1$ are shown below in figure 1. For $p = 1.8$ and $(k - m)/(n - m) \geqslant 0.5$, recovery with the Pareto is not too different from recovery with an exponential. However, for $p = 1.1$, this is true only for $(k-m)/(n-m) \geqslant 0.9$, say.

We propose the recovery equation (2.6) as a good way to understand the way recovery occurs in a large system and the way it depends on the cdf G. To describe the relation between G and τ, we use stochastic order relations. For background, see Stoyan [29] and chapter 4 of Baccelli and Brémaud [3]. The relationship between the cdf's G and G_e is studied in Whitt [30]. For example, if G_1 is *stochastically less variable* than G_2, denoted by $G_1 \leqslant_{\mathrm{v}} G_2$, i.e., if $\int_0^\infty f\,\mathrm{d}G_1 \leqslant \int_0^\infty f\,\mathrm{d}G_2$ for all convex real-valued f for which the integrals are well defined (which requires equal means), then G_{1e} is *stochastically less* than G_{2e}, denoted by $G_{1e} \leqslant_{\mathrm{st}} G_{2e}$, i.e., $\int_0^\infty f\,\mathrm{d}G_{1e} \leqslant \int_0^\infty f\,\mathrm{d}G_{2e}$ for all nondecreasing real-valued f. ($G_{1e} \leqslant_{\mathrm{st}} G_{2e}$ if and only if $G_{1e}^c(t) \leqslant G_{2e}^c(t)$ for all t.) From (2.6), it is clear that if $G_{1e} \leqslant_{\mathrm{st}} G_{2e}$, then $\tau_1 \leqslant \tau_2$. Hence we have the following result.

Theorem 3. If two potential holding-time cdf's are ordered by $G_1 \leqslant_{\mathrm{v}} G_2$, then $\tau_1 \leqslant \tau_2$.

3. Systems with finitely many servers

To show that our infinite-server analysis also applies to systems with finitely many servers, we make numerical comparisons for the Erlang loss ($M/M/s/0$) model.

Example 2. We apply the algorithm in [2] for computing the time-dependent mean number of busy servers in the $M/M/s/0$ model given an arbitrary initial state. By (2.2), the infinite-server formula is simply

$$E\big(N(t) \mid N(0) = n\big) = n + (n - m)\,\mathrm{e}^{-t}, \quad t \geqslant 0. \tag{3.1}$$

We claim that the time-dependent mean $E(N(t) \mid N(0) = n)$ will be approximately independent of s for $s \geqslant n$ if m is suitably large and n is suitably large compared to m. We give two concrete examples: first, $\gamma = 1$, $m = \lambda\gamma = 400$ and $n = 500$ and, second, $\gamma = 1$, $m = \lambda\gamma = 100$ and $n = 130$. The first case should be more dramatic because m is larger and n exceeds m by $5\sqrt{m}$ as opposed to $3\sqrt{m}$. (The standard deviation of the steady-state number of busy servers in the infinite-server model is $\sqrt{m}$.) Table 1 provides supporting evidence by displaying the time-dependent mean for three values of s in each case, with $s = n$ in the first case and $s = \infty$ in the last case, with the last case coming from (3.1).

Table 1
The time-dependent mean $E(N(t) \mid N(0) = n)$ in the $M/M/s/0$ model as a function of s. The model parameters are $\gamma = 1$, $m = \lambda = 400$ and $n = 500$ in the first case and $\gamma = 1$, $m = \lambda = 100$ and $n = 130$ in the second case.

	$m = \lambda = 400$			$m = \lambda = 100$		
time t	$s = 500$	$s = 600$	$s = \infty$	$s = 130$	$s = 180$	$s = \infty$
0.0	500.00	500.00	500.00	130.00	130.00	130.00
1.0	435.15	436.79	436.79	109.42	111.04	111.04
2.0	412.93	413.53	413.53	103.34	104.06	104.06
3.0	404.76	404.98	404.98	101.17	101.49	101.49
4.0	401.75	401.83	401.83	100.39	100.55	100.55
5.0	400.64	400.67	400.67	100.10	100.20	100.20
∞	400.00	400.00	400.00	99.94	100.00	100.00

When s is significantly greater than n, the infinite-server approximation is essentially exact. When $s = n$, there is an error in the approximation, but it is not large.

4. Recovery with intervention

We now consider modifications in the conditional mean $E(N(t) \mid N(0) = n)$ and the recovery equation based on intervention.

Reduction of the arrival rate. One form of intervention is to reduce the arrival rate after time $t = 0$, where 0 is a point of high congestion, i.e., $N(0) = n$. In such control settings it is natural for the control epoch to be the first hitting time of the level n. Instead, our analysis is based on conditioning upon the event $\{N(0) = n\}$ in steady state. However, as indicated earlier, it seems reasonable to also apply our results to the first hitting time of level n.

Reducing the arrival rate to a new constant value is equivalent to reducing m in (2.2) and (2.6), which produces a new recovery equation of the same form. For $m < k < n$, clearly $(k - m)/(n - m)$ is increasing in m, so that the recovery time for the mean is reduced by decreasing m.

One possible strategy is to completely turn off arrivals until recovery is achieved. The resulting recovery time for the mean is the solution to (2.6) with $m = 0$. Further recovery afterwards with the arrivals turned on is again described by (2.6) with n replaced by k and k replaced by a new recovery level j with $m < j < k$.

More generally, we could decide to reduce the arrival rate to a time-dependent function $\lambda(t)$, $t \geqslant 0$, with $\lambda(t) \leqslant \lambda$. For example, this could be realized by admitting an arrival at time t with a time-dependent probability $\lambda(t)/\lambda$. Then $N_0(t)$ still has a Poisson distribution, but now with mean

$$EN_0(t) = \int_0^t \lambda(u) G^c(t - u) \, \mathrm{d}u, \tag{4.1}$$

which is still increasing in t. However, with (4.1), $E(N(t) \mid N(0) = n)$ need not be decreasing in t. Because of the lack of monotonicity of $E(N(t) \mid N(0) = n)$ when we use (4.1), it is natural to consider changing the definition of the recovery time for the mean to

$$\tau = \inf\Big\{t \geqslant 0\colon \sup_{u \geqslant t}\big(N(u) \mid N(0) = n\big) \leqslant k\Big\}. \tag{4.2}$$

In general, it is natural to pay attention to the whole path $E(N(t) \mid N(0) = n)$ for $t \geqslant 0$ as well as τ. The suprema $\sup_{0 \leqslant u \leqslant t} E(N(u) \mid N(0) = n)$ and $\sup_{u \geqslant t} E(N(u) \mid N(0) = n)$ for $t \geqslant 0$ both may be interesting.

We must be careful defining a corresponding modified recovery time T extending (1.2), because typically $\sup_{u \geqslant t} N(u) = \infty$ w.p.1. One approach is to consider the supremum up to a suitable finite time limit t_1, i.e.,

$$T = \inf\Big\{t\colon \sup_{0 \leqslant t \leqslant u \leqslant t_1} N(u) \leqslant k \mid N(0) = n\Big\} \tag{4.3}$$

with $T = t_1$ if the infimum is not attained. If $t_1 > \tau$ and $\tau < \infty$, where τ is defined by (4.2), then T will still be approximately equal to τ in large systems (see section 7).

It is important to note that our analysis so far relies on the arrival process being a Poisson process. Some methods of reducing the arrival rate would cause the Poisson property to be lost. For example, if we reject every other arrival after time 0, then the arrival process after time 0 becomes a renewal process with Erlang (E_2) interarrival times and rate $\lambda/2$. Obviously, changing the arrival process after time 0 leaves the number of customers in the system at time 0 and the ages of their holding times unchanged. It is significant that the critical mean formula $EN_0(t)$ in (4.1) actually does *not* depend on the Poisson property, but instead holds for any point process with arrival rate function $\lambda(u)$, $u > t$; see remark 2.3 of Massey and Whitt [27] and the analysis there. Thus, the analysis here is applicable to a large class of arrival rate controls.

5. Loss and delay

The framework developed so far can be used to approximately describe loss and delay in a service system with capacity c. In this setting, c is the fixed rate that fluid is processed, while the process $N(t)$ that we have been studying represents the random fluid input rate at time t. There may be no buffer, an infinite buffer, or a finite buffer of size b.

The natural simple approximation based on what we have done is to let the input rate at time 0 have the steady-state Poisson distribution as in (1.1) and let the conditional process $(N(t) \mid N(0) = n)$ be approximated by its conditional mean $E(N(t) \mid N(0) = n)$ for $t > 0$ and $t < 0$. We will use the $M/G/\infty$ model in section 2, for which $E(N(t) \mid N(0) = n) = m + (n - m)G_e^c(t)$. We assume that $c > m$.

No buffer. First consider the case of no buffer. Then the probability of loss (rate in $>$ rate out) at any time in steady-state is

$$P\big(N(0) > c\big) = \sum_{k=c+1}^{\infty} \frac{\mathrm{e}^{-m}m^k}{k!} \approx \Phi^{\mathrm{c}}\big((c + 0.5 - m)/\sqrt{m}\,\big), \tag{5.1}$$

where $\Phi^{\mathrm{c}}(x) = 1 - \Phi(x)$ and Φ the standard (mean 0, variance 1) normal cdf. From (5.1), we see that the likelihood of loss depends on the holding-time cdf G only through its mean.

We next describe the average fluid loss rate, denoted by r. Let ϕ be the density of the standard normal cdf Φ. Let $N(a, b)$ denote a normal random variable with mean a and variance b. We use basic properties of the conditional mean of a normal random variable; e.g., see lemma 1 on p. 593 of Choudhury, Leung and Whitt [10]. The average fluid loss rate is

$$\begin{aligned} r &= E\big(N(0) - c\big)^{+} = E\big(N(0) - c \mid N(0) > c\big)P\big(N(0) > c\big) \\ &\approx E\big(N(-(c - m), m\big) \mid N\big(-(c - m), m\big) > 0.5\big)P\big(N\big(-(c - m), m\big) > 0.5\big) \\ &= \left(-(c - m) + \sqrt{m}\frac{\phi((c - m + 0.5)/\sqrt{m})}{\Phi^{\mathrm{c}}((c - m + 0.5)/\sqrt{m})}\right)\Phi^{\mathrm{c}}\big((c - m + 0.5)/\sqrt{m}\big). \end{aligned} \tag{5.2}$$

However, when there is loss, the length of the period where it occurs depends on the cdf G beyond its mean, as we have shown. Moreover, conditional on the rate being $n > c$ at time 0, the total loss associated with this excursion of $N(t)$ above c depends on the cdf beyond the mean. Let

$$\tau_c(n) = \inf\big\{t \geqslant 0\colon\ E\big(N(t) \mid N(0) = n\big) \leqslant c\big\} \tag{5.3}$$

and let L be the total quantity of fluid lost while the rate experiences this excursion above c. Then our fluid approximation is

$$\begin{aligned} \big(L \mid N(0) = n\big) &\approx 2\int_0^{\tau_c(n)} \big[E\big(N(t) \mid N(0) = n\big) - c\big]\,\mathrm{d}t \\ &= 2(n - m)\int_0^{\tau_c(n)} G_e^c(t)\,\mathrm{d}t - 2(c - m)\tau_c(n), \end{aligned} \tag{5.4}$$

where

$$G_e^c\big(\tau_c(n)\big) = \frac{c - m}{n - m}. \tag{5.5}$$

The 2 appears in (5.4) to account for the mean before time 0 as well as the mean after time 0. From (5.4), we can see the influence of the cdf G.

Unlimited buffer capacity. Next suppose that there is an infinite-capacity buffer. The net rate into the buffer is positive while $N(t) > c$. The maximum buffer content associated with an excursion of $N(t)$ above c, $Q_{\max}$, is also approximated by (5.4). The maximum delay experienced by any particle of fluid during the excursion above c

is then $Q_{\max}/c$. However, the buffer is still emptying after $\tau_c \equiv \tau_c(n)$. The conditioned busy period (the length of the period that the buffer is non-empty) is then approximated by

$$\begin{aligned}&\bigl(B \mid N(0)=n\bigr)\\&\quad\approx 2\tau_c(n)+\inf\left\{u \geqslant 0\colon\ Q_{\max}=\int_{\tau_c}^{\tau_c+u}\bigl[c-E\bigl(N(t)\mid N(0)=n\bigr)\bigr]\,\mathrm{d}t\right\}\\&\quad=2\tau_c(n)+\inf\left\{u \geqslant 0\colon\ Q_{\max}=\int_{\tau_c}^{\tau_c+u}\bigl[c-m-(n-m)G_e^c(t)\bigr]\,\mathrm{d}t\right\},\end{aligned}\tag{5.6}$$

where $Q_{\max}$ is approximated by (5.4).

A finite buffer. Now suppose that there is a finite buffer of capacity b. Using the deterministic fluid approximation, let n_b be the level such that the buffer just fills at time 0, i.e., let n_b be such that

$$b=\int_0^{\tau_c(n_b)}\bigl[E\bigl(N(t)\mid N(0)=n_b\bigr)-c\bigr]\,\mathrm{d}t.\tag{5.7}$$

We actually consider the interval $[-\tau_c(n_b),0]$, but the means forward and backward from 0 are the same. Then we would approximate the probability of any loss at time 0 by

$$P\bigl(N(0)>n_b\bigr)=\sum_{k=n_b+1}^{\infty}\frac{\mathrm{e}^{-m}m^k}{k!}\approx\Phi^{\mathrm{c}}\bigl((n_b+0.5-m)/\sqrt{m}\bigr).\tag{5.8}$$

The expected loss associated with an excursion above c is

$$\sum_{n=c+1}^{\infty}P\bigl(N(0)=n\bigr)\bigl(\bigl(L\mid N(0)=n\bigr)-b\bigr)^{+},\tag{5.9}$$

where $(L \mid N(0)=n)$ is given by (5.4) and $N(0)$ has the Poisson distribution. To get an explicit value of (5.9), it seems necessary to calculate each term and then calculate the sum.

Suppose that we seek the probability that a quantity of fluid of at least size d is lost associated with an excursion above c. Then, let $n(b,d)$ be the largest level n such that

$$b+d\geqslant 2\int_0^{\tau_c(n)}\mathrm{d}t\bigl[E\bigl(N(t)\mid N(0)=n\bigr)-c\bigr],\tag{5.10}$$

where $\tau_c(n)$ is as defined in (5.3) Let L_b be the quantity of fluid lost during an excursion above c with buffer capacity b. Then

$$\begin{aligned}P\bigl(L_b>d\mid N(0)>c\bigr)&\approx P\bigl(N(0)>n(b,d)\mid N(0)>c\bigr)\\&\approx\frac{\Phi^{\mathrm{c}}((n(b,d)+0.5-m)/\sqrt{m})}{\Phi^{\mathrm{c}}((c+0.5-m)/\sqrt{m})}.\end{aligned}\tag{5.11}$$

In summary, the simple fluid approximation gives useful insight into the way loss and delays depend on the holding-time cdf G when the input rate is an $M/G/\infty$ process.

6. Multiclass extensions

The results so far extend easily to multiple classes of customers, which is very important for analyzing integrated-services networks. With the infinite-server model, multiple classes are easily treated separately as independent single classes.

To illustrate, suppose that there are c independent customer classes. For customer class i, service requests arrive according to a Poisson process with arrival rate λ_i and require a fixed bandwidth b_i for a random holding time having a general cdf G_i with mean γ_i. We now define the recovery time for the mean

$$\tau = \inf\left\{t \geqslant 0\colon\ \sup_{u \geqslant t} \sum_{i=1}^{c} b_i E\bigl[N_i(u) \mid N_1(0) = n_1, \ldots, N_c(0) = n_c\bigr] \leqslant k\right\}, \tag{6.1}$$

where it is understood that the recovery level k is in between the mean and the initial level, i.e.,

$$m \equiv \sum_{i=1}^{c} b_i m_i < k < \sum_{i=1}^{c} b_i n_i \equiv n, \tag{6.2}$$

with $m_i = \lambda_i \gamma_i$. Paralleling (4.2), we include the supremum in (6.1) because we need not have monotonicity. Paralleling (4.3), we can define a corresponding recover time T using an upper time limit t_1.

The initial congestion time 0 might be determined as the first hitting time of the total bandwidth process $\sum_{i=1}^{c} b_i N_i(t)$ to the level n. At that time, a congestion alarm goes off. We then observe the values $N_i(0) = n_i$ for $1 \leqslant i \leqslant c$ and consider how recovery occurs.

The analysis is a straightforward extension of the previous analysis, because the processes $\{N_i(t)\colon\ t \geqslant 0\}$ are conditionally mutually independent given the event $\{N_i(0) = n_i,\ 1 \leqslant i \leqslant c\}$. Paralleling (2.2),

$$E\bigl(N_i(t) \mid N_i(0) = n_i\bigr) = ES_{in}(t) + EN_{i0}(t) = m_i + (n_i - m_i)G_{ie}^{c}(t), \tag{6.3}$$

so that

$$E\left(\sum_{i=1}^{c} b_i N_i(t) \mid N_i(0) = n_i,\ 1 \leqslant i \leqslant c\right) = m + \sum_{i=1}^{c} b_i (n_i - m_i) G_{ie}^{c}(t). \tag{6.4}$$

Thus, we obtain the following generalization of theorem 2.

Theorem 4. Suppose that $G_i^c(t) > 0$ for all i and t. If (6.2) holds in the multiclass setting, then the recovery time for the mean defined in (6.1) is

$$\tau = \inf\left\{ t \geqslant 0\colon \ \sup_{u \geqslant t} \left\{ \sum_{i=1}^{c} b_i(n_i - m_i) G_{ie}^c(u) \right\} = k - m \right\}. \tag{6.5}$$

7. Large systems

We now return to the basic $M/G/\infty$ setting of section 2, and show that we are justified in focusing on the recovery time for the mean in (1.3) when the system is large. We show that the actual recovery time T in (1.2) will be close to the recovery time for the mean when the system is large. To do so, we consider the limit as the arrival rate increases and impose appropriate regularity conditions. Let T_λ and τ_λ be the recovery times in (1.2) and (1.3) as functions of λ. Let $\xrightarrow{p}$ denote convergence in probability.

Theorem 5. If $\lambda \to \infty$, $n \to \infty$ and $k \to \infty$ with $n/\lambda \to \delta$ and $k/\lambda \to \beta$, where $\gamma < \beta < \delta$, then

$$\tau_\lambda \to \tau \quad \text{and} \quad T_\lambda \xrightarrow{p} \tau,$$

where

$$\tau = \sup\bigl\{ t \geqslant 0\colon \ G_e^c(t) > (\beta - \gamma)/(\delta - \gamma) \bigr\}. \tag{7.1}$$

We now discuss the proof of theorem 5. First, the limit for τ_λ is an elementary consequence of (2.6) and (2.7). Hence the main problem is the limit for T_λ. We can establish the desired convergence by applying the law of large numbers, but it is important to apply the law of large numbers in function space, because we want the sample paths of the stochastic process $\{N(t)\colon\ t \geqslant 0 \mid N(0) = n\}$ to be suitably close to the conditional expectation $E(N(t) \mid N(0) = n)$ for all t in intervals $[0, t_0]$ for each t_0. (The conditional stochastic process $\{N(t)\colon\ t \geqslant 0 \mid N(0) = n\}$ is well defined by theorem 1 and (2.4).) It is not enough that the random variable $(N(t) \mid N(0) = n)$ be close to its mean with high probability for each t. We need the uniform convergence over bounded intervals in order for the first passage time to be continuous almost surely with respect to the deterministic limit process. Uniform convergence over bounded intervals also allows us to treat the supremum over bounded intervals in (4.3) by applying the continuous mapping theorem.

In fact we can obtain a functional weak law of large numbers (FWLLN) as a corollary of a functional central limit theorem (FCLT). For the FCLT, it is convenient to treat the two processes $S_n(t)$ and $N_0(t)$ in (2.4) separately. For $S_n(t)$, we can apply the FCLT for empirical distribution functions in sections 13 and 16 of Billingsley [5]. For $N_0(t)$, we can apply the FCLT for the $M/G/\infty$ model starting empty (and more general models) on p. 103 Borovkov [7]. (See Glynn and Whitt [19] and Krichagina and Puhalskii [22] for other work related to the FCLT for infinite-server queues.) In particular, theorem 5 is a consequence of the following result. Let $D[0,\infty)$ be the function space of all right-continuous real-valued functions with left limits, endowed with the usual Skorohod metric, which reduces to uniform convergence on bounded intervals at continuous limit functions, and let $\Rightarrow$ denote convergence in distribution. Let $N(\mu,\sigma^2)$ denote a normal distribution with mean μ and variance σ^2.

Theorem 6. Under the conditions of theorem 5,

$$\frac{(N_\lambda(\cdot) \mid N_\lambda(0)=n) - E(N_\lambda(\cdot) \mid N_\lambda(0)=n)}{\sqrt{\lambda}} \Rightarrow Z(\cdot) \quad \text{in } D[0,\infty) \text{ as } \lambda \to \infty,$$

where $\{Z(t)\colon\ t \geqslant 0\}$ is a Gaussian process with zero means and continuous sample paths. In particular,

$$Z(t) \overset{d}{=} N\bigl(0,\sigma^2(t)\bigr), \tag{7.2}$$

where

$$\sigma^2(t) = \delta G_e^c(t)G_e(t) + \gamma G_e(t). \tag{7.3}$$

The variance formula in (7.3) follows directly from (2.3).

As an immediate consequence of the FCLT in theorem 6, we obtain the associated functional weak law of large numbers (FWLLN).

Corollary. Under the conditions of theorem 5,

$$\lambda^{-1}\bigl[\bigl(N_\lambda(\cdot) \mid N_\lambda(0)=n\bigr) - E\bigl(N_\lambda(\cdot) \mid N_\lambda(0)=n\bigr)\bigr] \Rightarrow \theta(\cdot) \quad \text{in } D[0,\infty) \quad \text{as } \lambda \to \infty, \tag{7.4}$$

where $\theta(t)=0$ for $t \geqslant 0$, so that for all (deterministic) times T_*

$$\sup_{0\leqslant t\leqslant T_*}\left\{\left|\frac{(N_\lambda(t) \mid N_\lambda(0)=n) - m}{n-m} - G_e^c(t)\right|\right\} \Rightarrow 0 \quad \text{as } \lambda \to 0. \tag{7.5}$$

Proof. For (7.4) we apply theorem 6 and the continuous mapping theorem, theorem 5.5 of [5], using the mappings $h_\lambda(x) = x/\sqrt{\lambda}$ for $x \in D[0,\infty)$. The limit (7.5) is essentially just a restatement of (7.4) using the fact that

$$E\bigl(N_\lambda(t) \mid N_\lambda(0)=n\bigr) = m + (n-m)G_e^c(t). \tag{7.6}$$

By (7.6),

$$\begin{aligned} &\frac{(n-m)}{\lambda}\frac{[(N_\lambda(t) \mid N_\lambda(0)=n) - E(N_\lambda(t) \mid N_\lambda(0)=n)]}{n-m} \\ &\quad = \frac{(N_\lambda(t) \mid N_\lambda(0)=n) - m}{n-m} - G_e^c(t). \end{aligned}$$

Finally, use the conditions to obtain $(n-m)/\lambda \to \delta - \gamma$ as $\lambda \to \infty$. □

Proof of theorem 5. Theorem 5 is an elementary consequence of (7.5) and (7.6). First, by (7.6),

$$E\bigl(N_\lambda(t) \mid N_\lambda(0) = n\bigr) < k$$

if and only if

$$G_e^c(t) < \frac{k-m}{n-m},$$

where

$$\frac{k-m}{n-m} \to \frac{\beta-\gamma}{\delta-\gamma} \quad \text{as } \lambda \to \infty.$$

Hence $\tau_\lambda \to \tau$ as $\lambda \to \infty$ for τ in (7.1). (We exploit the fact that $G_e^c(t)$ is strictly decreasing in t.) We now turn to T_λ and apply (7.5). If we choose $T_* > \tau$ for T_* in (7.5), then by the closeness of sample paths $T_\lambda \to \tau$ as $\lambda \to \infty$ too. □

We point out that the FWLLN in the corollary to theorem 6 helps justify the use of the steady-state conditioning results to approximate hitting-time conditioning, as indicated in the introduction. By the time reversibility, the process before 0 is distributed the same as the process after 0. By (3.3), the limiting mean before time 0 is increasing.

Under slightly stronger conditions on the growth of k and n with λ, we can obtain a stronger distributional limit for T_λ.

Theorem 7. Let $\lambda \to \infty$ and $n \to \infty$ with $(k - \beta\lambda)/\sqrt{\lambda} \to 0$ and $(n - \delta\lambda)/\sqrt{\lambda} \to 0$. Assume that the support of G is greater than τ for τ in (6.1). Then $\sqrt{\lambda}\,(\tau_\lambda - \tau) \to 0$ as $\lambda \to \infty$ and

$$\sqrt{\lambda}(T_\lambda - \tau) \Rightarrow N\bigl(0, \sigma^2\bigr) \quad \text{as } \lambda \to \infty, \tag{7.7}$$

where

$$\sigma^2 = \frac{\gamma^2}{G^c(\tau)^2}\left[\frac{\delta(\beta-\gamma)(\delta-\beta) + \gamma(\delta-\beta)(\delta-\gamma)}{(\delta-\gamma)^2}\right]. \tag{7.8}$$

Proof. Apply (7.6). By the new conditions,

$$\left(\frac{k-m}{n-m}\right) - \left(\frac{\beta-\gamma}{\delta-\gamma}\right) = \mathrm{o}\bigl(1/\sqrt{\lambda}\bigr) \quad \text{as } \lambda \to \infty,$$

which implies that $\tau_\lambda - \tau = \mathrm{o}(1/\sqrt{\lambda})$ as $\lambda \to \infty$. Next, by theorem 6, recalling that $[(N_\lambda(T_\lambda) \mid N_\lambda(0) = n) - m]/(n-m) = (k-m)/(n-m)$ and $G_e^c(\tau_\lambda) = (k-m)/(n-m)$,

$$\begin{aligned}\sqrt{\lambda}\bigl(G_e^c(\tau_\lambda) - G_e^c(T_\lambda)\bigr) &= \sqrt{\lambda}\left(\frac{(N_\lambda(T_\lambda) \mid N_\lambda(0) = n) - m}{n-m} - G_e^c(T_\lambda)\right) \\ &\Rightarrow Z(\tau), \quad \text{as } \lambda \to \infty.\end{aligned} \tag{7.9}$$

(We use the fact that $Z(T_\lambda) \Rightarrow Z(\tau)$ as $\lambda \to \infty$, because Z has continuous sample paths and $T_\lambda \Rightarrow \tau$.) By theorem 5 and Taylor's theorem,

$$G_e^c(T_\lambda) - G_e^c(\tau_\lambda) = -(T_\lambda - \tau_\lambda)g_e(\tau_\lambda) + \mathrm{o}(T_\lambda - \tau_\lambda). \tag{7.10}$$

Combining (7.9) and (7.10), and using the differentiability of $G_e^c(t)$ in a neighborhood of τ (implied by the second condition), we obtain

$$\sqrt{\lambda}(T_\lambda - \tau_\lambda) \Rightarrow \frac{Z(\tau)}{-g_e(\tau)} \quad \text{as } \lambda \to \infty.$$

Next note that $g_e(t) = G^c(t)/\gamma$. Finally, by theorem 6,

$$Z(\tau) \stackrel{d}{=} N\bigl(0, \sigma^2(\tau)\bigr), \tag{7.11}$$

where

$$\sigma^2(\tau) = \delta G_e^c(\tau)G_e(\tau) + \gamma G_e(\tau). \tag{7.12}$$

Since $G_e^c(\tau) = (\beta - \gamma)/(\delta - \gamma)$, we obtain (7.8). □

As a practical consequence of theorem 7, we see that (in a stochastic sense)

$$T_\lambda \approx \tau + \mathrm{O}\bigl(1/\sqrt{\lambda}\bigr). \tag{7.13}$$

8. Exploiting the age distribution

The main reason for using the non-Markov $M/G/\infty$ model instead of a Markov model is to allow for non-exponential holding-time distributions. When the holding-time distribution is exponential, the elapsed holding times provide no information about the remaining holding times, by the lack-of-memory property of the exponential distribution. However, when the holding-time distribution is very different from exponential, the elapsed holding times can enable us to accurately predict the remaining holding times. Non-exponential holding-time distributions make stochastic models harder to analyze, but they help us predict remaining holding times by exploiting the elapsed holding times.

To see the impact of elapsed holding times on remaining holding times, it is useful to consider the failure rate or hazard rate associated with a cdf G with pdf g, i.e.,

$$r(x) \equiv \frac{g(x)}{G^c(x)}, \quad x \geqslant 0. \tag{8.1}$$

It is significant that the conditional remaining-holding-time ccdf can be expressed directly in terms of the failure rate by

$$H_x^c(t) \equiv \frac{G^c(x+t)}{G^c(x)} = \exp\biggl(-\int_x^{x+t} r(u)\,\mathrm{d}u\biggr), \quad t \geqslant 0. \tag{8.2}$$

To see (8.2), note that

$$G^c(x) = \exp\left(-\int_0^x r(u)\,\mathrm{d}u\right), \quad x \geqslant 0, \tag{8.3}$$

which in turn can be verified by taking the logarithm and then differentiating.

It is useful to employ stochastic comparison results related to aging that have been considered in reliability theory, e.g., see Barlow and Proschan [4]. For example, H_x^c is *stochastically increasing* in x, i.e.,

$$H_{x_1}^c(t) \leqslant H_{x_2}^c(t) \quad \text{for all } t \text{ when } x_1 < x_2 \tag{8.4}$$

if and only if the holding-time cdf G has *decreasing failure rate* (is DFR); see [4, p. 54]. Similarly, H_x is *stochastically decreasing* in x if and only if G has *increasing failure rate* (is IFR). Thus, if G is DFR (IFR), then recovery from congestion would be speeded up if we interrupt the longest (shortest) holding times.

The potential advantage of knowing the actual ages can be determined by comparing

$$H_\ell^c(t) \equiv \inf_x H_x^c(t) \quad \text{and} \quad H_u^c(t) \equiv \sup_x H_x^c(t). \tag{8.5}$$

When G is DFR, $H_\ell^c(t) = G^c(t) \leqslant G_e^c(t)$; when G is IFR, $H_u^c(t) = G^c(t) \geqslant G_e^c(t)$; when G is exponential, it is both DFR and IFR, so that $H_\ell^c(t) = G^c(t) = G_e^c(t) = H_u^c(t)$.

We now consider some concrete examples. First, the hyperexponential (H_k) ccdf is

$$G^c(t) = \sum_{i=1}^k p_i\,\mathrm{e}^{-\lambda_i t}, \quad t \geqslant 0, \tag{8.6}$$

where $\lambda_1 < \cdots < \lambda_k$. The H_k ccdf is known to be DFR. Moreover, it is easy to see that H_x is again hyperexponential, i.e.,

$$H_x^c(t) = \sum_{i=1}^k p_i(x)\,\mathrm{e}^{-\lambda_i t}, \quad t > 0, \tag{8.7}$$

where

$$p_i(x) = \frac{p_i\,\mathrm{e}^{-\lambda_i x}}{\sum_{j=1}^k p_j\,\mathrm{e}^{-\lambda_j x}}, \quad i = 1, \ldots, k. \tag{8.8}$$

It is easy to see that

$$H_x^c(t) \to \mathrm{e}^{-\lambda_1 t} \quad \text{as } x \to \infty. \tag{8.9}$$

Moreover, $H_x^c(t) \to 1$ as $x \to \infty$ for all t if and only if $r(t) \to 0$ as $t \to \infty$. This means that the remaining holding time gets large as the elapsed holding time gets large. Many long-tail distributions, such as the Weibull and Pareto distributions have the DFR property with $r(t) \to 0$ as $t \to \infty$. For the Weibull ccdf

$$G^c(t) = \mathrm{e}^{-(t/a)^c}, \quad t \geqslant 0, \tag{8.10}$$

it is easy to see directly that $H_x^c(t) \to 1$ as $x \to \infty$ for each t. An even more revealing property holds for the Pareto ccdf

$$G^c(t) = (1 + bt)^{-a}, \quad t \geqslant 0, \tag{8.11}$$

which we state as a theorem. Let $\stackrel{d}{=}$ denote equality in distribution.

Theorem 8. Let $Y(a,b)$ have the Pareto ccdf in (8.11) and let $Y_x(a,b)$ have the associated conditional remaining-holding-time ccdf in (8.2). Then

$$Y_x(a,b) \stackrel{d}{=} (1 + bx)Y(a,b). \tag{8.12}$$

Proof. Given (8.11),

$$H_x^c(t) = \frac{(1+bx)^a}{(1+b(x+t))^a} = \bigl(1 + \bigl(b/(1+bx)\bigr)t\bigr)^{-a}, \tag{8.13}$$

so that $H_x^c(t) = G^c(t/(1+bx))$, $t \geqslant 0$, which implies the stated result. □

Formula (8.12) dramatically shows how the distribution of the remaining holding time increases stochastically as the elapsed holding time increases, when the underlying holding-time distribution is Pareto.

It should thus prove to be more effective, but possibly more costly, to keep track of the age distribution and exploit it in predictions. Given that we know the age distribution, it does not matter how our time of interest is selected. It could be a hitting time or a random time (in steady-state). If we keep track of the n ages $x_1, \ldots, x_n$ of the customers in service, then we can use them to obtain a new estimate for the mean number remaining at time t, $ES_n(t)$.

The combinatorial possibilities for n calls with ages $x_1, \ldots, x_n$ makes the distribution of $S_n(t)$ somewhat complicated, but the mean has the simple form

$$ES_n(t) = \sum_{i=1}^{n} H_{x_i}^c(t) \tag{8.14}$$

for H_x in (8.2). As before, $ES_n(t)$ is decreasing in t. But if $ES_n(t)$ in (8.14) is used to find the overall conditional mean $E(N(t) \mid N(0) = n)$ by exploiting (2.4), then $E(N(t) \mid N(0) = n)$ need not be decreasing in t. Hence, we could use the modified recovery times in (4.2) and (4.3).

9. Congestion control via call interruption

Another form of intervention, which might be regarded as more drastic than reducing the input rate (section 4), is to interrupt and block some services in progress. If we elect to block j customers in service, then it is natural to ask which j should be selected. From the point of view of revenue, assuming that revenue is earned proportional to the holding time, then we might prefer to block the customers with shortest elapsed holding times. On the other hand, from the point of view of recovery from congestion, this might not be the best choice, because the services that have been in progress for a long time may be more likely to remain in progress a long time. Fortunately, we can evaluate the effect of any decision upon the recovery from congestion. Given n customers in service at time $t = 0$ with ages $x_1, \dots, x_n$, suppose that we decide to eliminate the last j (without necessarily assuming any ordering on the ages). Then, instead of (8.14), we obtain

$$ER'_n(t) = \sum_{i=1}^{n-j} H^c_{x_i}(t). \tag{9.1}$$

More generally, in order to determine which customers to interrupt, we can formulate and solve mathematical programs. To illustrate, suppose that there is a revenue $r_1 + r_2 x$ for completing service of a request with holding time x. Let the control variables be y_k, i.e., $y_k = 0$ if the kth customer is interrupted and $y_k = 1$ otherwise. One approach is to minimize the lost revenue, counting only the elapsed holding times, subject to the expected remaining number of customers in service at time t being less than some target value v. This is achieved by the *integer program*:

$$\min \sum_{k=1}^{n} (r_1 + r_2 x_k)(1 - y_k) \tag{9.2}$$

$$\text{subject to:} \quad \sum_{k=1}^{n} H^c_{x_k}(t) y_k \leqslant v, \quad y_k = 0 \text{ or } 1, \ 1 \leqslant k \leqslant n. \tag{9.3}$$

If we wanted to minimize the expected total lost revenue of interruptions, including the remaining holding times, then we would replace x_k in the objective function in (9.2) by $x_k + x'_k$, where x'_k is the conditional mean residual life, i.e.,

$$x'_k = \int_0^\infty H^c_{x_k}(t)\, \mathrm{d}t. \tag{9.4}$$

However, some of the lost future revenue associated with interrupted calls may be replaced by revenue from new arrivals.

Different interruption strategies suggest that it is interesting to compare the impact of different age distributions upon the empirical residual holding-time distribution. Note, however, that we typically would observe the value of $A(0)$, but the associated empirical residual holding-time distribution $R(0)$ is a random cdf. To compare random cdf's, we use notions of stochastic order on complete separable metric spaces with closed partial orders (using the natural orders on the underlying spaces, i.e., cdf's are ordered $F_1 \leqslant F_2$ if $F_1^c(t) \leqslant F_2^c(t)$ for all t); e.g., see Kamae, Krengel and O'Brien [21] and Lindvall [26, chapter IV]. Thus, $R_1(0) \leqslant_{\mathrm{st}} R_2(0)$ means that $Ef(R_1(0)) \leqslant Ef(R_2(0))$ for all nondecreasing real-valued functions defined on the space of cdf's, while, for possible realizations, $R_1(0) \leqslant R_2(0)$ means stochastic order for cdf's, i.e., $R_1(0,x) \geqslant R_2(0,x)$ for all x w.p.1. Since we typically know the ages, the natural condition is $A_1(0) \leqslant A_2(0)$ w.p.1, which means $A_1(0,x) \geqslant A_2(0,x)$ for all x w.p.1. However, this w.p.1 comparison implies the stochastic comparison $A_1(0) \leqslant_{\mathrm{st}} A_2(0)$.

Theorem 9. Assume that $N(0) = n$. Let $R_i(0)$ be the empirical residual holding-time distribution associated with the empirical age distribution $A_i(0)$ for $i = 1, 2$. If G is DFR (IFR) and $A_1(0) \leqslant_{\mathrm{st}} (\geqslant_{\mathrm{st}})\ A_2(0)$, then $R_1(0) \leqslant_{\mathrm{st}} R_2(0)$ and $\tau_1 \leqslant \tau_2$.

Proof. Condition on the event $\{N(0) = n\}$. If $A_1(0) \leqslant_{\mathrm{st}} A_2(0)$, then there are new versions $\widetilde{A}_1(0)$, and $\widetilde{A}_2(0)$, on a common probability space so that $P(\widetilde{A}_1(0) \leqslant \widetilde{A}_2(0)) = 1$ by Strassen's theorem, [26, p. 129]. Then, by the DFR property applied to each atom of $\widetilde{A}_i(0)$, $P(\widetilde{R}_1(0) \leqslant \widetilde{R}_2(0)) = 1$, which in turn implies that $\widetilde{R}_1(0) \leqslant_{\mathrm{st}} \widetilde{R}_2(0)$. Since $\widetilde{R}_i(0) \stackrel{d}{=} R_i(0)$, we have the desired conclusion: $R_1(0) \leqslant_{\mathrm{st}} R_2(0)$. Since $R_1(0) \leqslant_{\mathrm{st}} R_2(0)$, $ER_{1n}(t) \leqslant ER_{2n}(t)$ for all t, so that $\tau_1 \leqslant \tau_2$. □

So far, we have shown that we can obtain a suitable new recovery time for the mean if we do any or all of the following: (1) change the arrival rate, (2) keep track of the elapsed holding times and (3) interrupt some of the customers in service. It is worth remarking that the both the intervention and age-tracking strategies depend on the fact that the process $\{(N(t), A(t))\colon t \geqslant 0\}$ is Markovian. Changing the arrival rate amounts to introducing time-inhomogeneity into the process, while tracking ages and/or interrupting holding times in progress amount to conditioning with respect to a certain state at a given time.

10. Large systems with an unusual age distribution

We now consider an analog of the limiting behavior as $\lambda \to \infty$ given in theorem 5 when we condition upon the age distribution. We consider the conditional stochastic process $\{N(t)\colon t \geqslant 0 \mid N(0) = n,\ A(0) = F_n\}$, where F_n is a cdf compatible in the sense that it has has n jumps of size $1/n$. The recovery times τ and T in (4.2) and (4.3) have the conditional counterparts

$$\tau = \inf\Big\{t \geqslant 0\colon \sup_{u \geqslant t} E\big(N(u) \mid N(0) = n,\ A(0) = F_n\big) \leqslant k\Big\}, \tag{10.1}$$

$$T = \inf\Big\{t \geqslant 0\colon \sup_{t \leqslant u \leqslant t_1} \big(N(u) \mid N(0) = n,\ A(0) = F_n\big) \leqslant k\Big\}, \tag{10.2}$$

where we assume that t_1 is sufficiently large that $\tau < t_1$. The next theorem justifies the use of the mean recovery time when the system is large. As before, let τ_λ and T_λ be the recovery times in (10.1) and (10.2) as functions of λ. As a preliminary definition, following equations (8.2) and (8.14), we note that if the age is distributed according to the cdf F and the holding time is distributed according to the cdf G, then the remaining holding time has cdf ΘF given by

$$\Theta F(t) = \int_0^\infty \mathrm{d}F(x) H_x(t). \tag{10.3}$$

We call Θ the *residual time mapping*. It follows from (2.1) that G_e is a fixed point Θ, i.e., $\Theta G_e = G_e$. A sequence of cdf's $\{F_n\colon\ n \geqslant 1\}$ converges weakly to a cdf F if the associated probability measures converge weakly; see [5]. This means that $F_n(t) \to F(t)$ as $n \to \infty$ for all t that are continuity points of F.

Theorem 10. Assume that G is continuous. Under the conditions of theorem 5, with the addition that the cdf F_n converges weakly to some cdf F,

$$\begin{aligned}\lambda^{-1} M_n(t) &\equiv \lambda^{-1} E\bigl(N(t) \mid N(0) = n,\ A(0) = F_n\bigr) \to M(t)\\ &\equiv \delta(\Theta F)^c(t) + \gamma G_e(t) \quad \text{as } \lambda \to \infty\end{aligned} \tag{10.4}$$

and, for each $T_* > 0$,

$$\sup_{0\leqslant t\leqslant T_*} \bigl|\lambda^{-1}\bigl(N(t) \mid N(0) = n,\ A(0) = F_n\bigr) - M(t)\bigr| \Rightarrow 0 \quad \text{as } \lambda \to \infty, \tag{10.5}$$

so that

$$\tau_\lambda \to \tau \quad \text{and} \quad T_\lambda \xrightarrow{p} \tau \quad \text{as } \lambda \to \infty,$$

where

$$\tau = \inf\Bigl\{t \geqslant 0\colon\ \sup_{u\geqslant t} M(t) \leqslant \beta\Bigr\}. \tag{10.6}$$

Proof. Paralleling (3.1),

$$\bigl(N(t) \mid N(0) = n,\ A(0) = F_n\bigr) \stackrel{d}{=} \bigl(S_n(t) \mid N(0) = n,\ A(0) = F_n\bigr) + N_0(t),$$

where the two summands on the right are independent processes, so that

$$M_n(t) \equiv E\bigl(N(t) \mid N(0) = n,\ A(0) = F_n\bigr) = n(\Theta F_n)(t) + mG_e(t).$$

Using the continuity of G to obtain the continuity of Θ, we obtain $\lambda^{-1}M_n(t) \to M(t)$ as $\lambda \to \infty$ for $M(t)$ in (10.4), from which it follows that $\tau_\lambda \to \tau$ as $\lambda \to \infty$. As in section 6, we establish the FWLLN in (10.5) and desired limit for T_λ by establishing FCLTs for $(S_n(t) \mid N(0) = n,\ A_0 = F_n)$ and $N_0(t)$. The FCLT for $N_0(t)$ was established before. Hence, it remains to treat $S_n(t)$. Just as before (theorem 6), $\{S_n(t) \mid t \geqslant 0\}$ is the sum of n independent processes, but now these component processes are no longer identically distributed. Nevertheless, it is straightforward to modify the proof of theorem 16.4 of Billingsley [5] (the i.i.d. case) to obtain a FCLT in this case, in particular,

$$\frac{(S_n(\cdot) \mid N(0) = n,\ A_n(0) = F_n) - E(S_n(\cdot) \mid N(0) = n,\ A_n(0) = F_n)}{\sqrt{n}}$$
$$\Rightarrow Z(\cdot) \quad \text{in } D[0,\infty) \text{ as } n \to \infty,$$

where $\{Z(t)\colon\ t \geqslant 0\}$ is a Gaussian process. In particular, we can directly verify convergence of the finite-dimensional distributions and tightness. The finite-dimensional-distribution-limit is a consequence of the Lindeberg–Feller CLT for triangular arrays of non-identical summands, each of which is asymptotically negligible (here bounded), see Feller [18, theorem 3, p. 262, example e, p. 264, and pp. 518, 524]. As before, the FCLT implies the associated FWLLN (10.5), here using the maps $h_n(x) = x/\sqrt{n}$. As before, the FWLLN implies that $T_\lambda \xrightarrow{p} \tau$ as $\lambda \to \infty$. □

Theorem 10 motivates trying to understand how the mean occupation function $\{M(t)\colon\ t \geqslant 0\}$ in (10.4) depends on the holding-time cdf G and the limiting age cdf F. From (10.4), we see that it suffices to understand the stationary-excess map $G \to G_e$ defined in (2.1) and the residual-time mapping $(F, G) \to \Theta F$ defined in (10.3).

By essentially the same reason as in theorem 9, we obtain the following stochastic comparison result.

Theorem 11. If the holding-time cdf G is DFR and two prospective limiting age cdf's are ordered by $F_1 \leqslant_{\mathrm{st}} F_2$, then $\Theta F_1 \leqslant_{\mathrm{st}} \Theta F_2$, $M_1(t) \leqslant M_2(t)$ for all t and $\tau_1 \leqslant \tau_2$. If G is IFR, then the inequalities are reversed.

The following is another elementary stochastic comparison result, which follows easily from (10.3).

Theorem 12. If G_1 and G_2 are two holding-time cdf's for which $H_x^{(1)} \leqslant_{\mathrm{st}} H_x^{(2)}$ for all x, then $\Theta_1 F \leqslant_{\mathrm{st}} \Theta_2 F$.

Example 3. Consider the one-parameter family of Pareto ccdf's $G_p^c(x) = (p-1)(1+x)^{-p}$ for $p > 1$, all of which have mean 1. Since $G_{p_1}(x)$ crosses $G_{p_2}(x)$ for only one x, $G_{p_1} \leqslant_{\mathrm{v}} G_{p_2}$ when $p_1 > p_2$; see Stoyan [29, p. 12]. Hence $G_{p_1 e} \leqslant_{\mathrm{st}} G_{p_2 e}$, as noted before theorem 2. Clearly G_p is DFR for each p and $H_x^{(p)}$ is stochastically decreasing in p for all x. Hence, if $F_1 \leqslant_{\mathrm{st}} F_2$, then

$$\Theta_1 F_1 \leqslant_{\mathrm{st}} \Theta_2 F_1 \leqslant_{\mathrm{st}} \Theta_2 F_2$$

by theorems 11 and 12. Combining this with $G_{p_1 e} \leqslant_{\mathrm{st}} G_{p_2 e}$, we obtain $M_1(t) \leqslant M_2(t)$ for all t and $\tau_1 \leqslant \tau_2$, where $M_i(t)$ and τ_i are defined in terms of G_i and F_i, $i = 1, 2$. □

11. The asymptotic tail behavior of the mean

It is also interesting to see how the tail behavior of G_e and ΘF depends upon the tail behavior of F and G. Our motivation here is that, while $M(t)$ is generally not monotonic, if $(\beta - \gamma)/(\delta - \gamma)$ is small (for example, when we consider recovery from a level which is high to one which is close to the mean), then the recovery time should be determined by the asymptotic form of mean occupation $M(t)$ for large t.

The effect of G upon the tail behavior of G_e is easy to see because asymptotics for $G^c(t)$ is inherited by the integral (see, e.g., Erdelyi [17, p. 17]). For example, we thus have the following result.

Theorem 13. Let G be a cdf with mean γ. (a) If $G^c(x) \sim \alpha x^\beta \,\mathrm{e}^{-\eta x}$ as $x \to \infty$ for $\alpha > 0$ and $\eta > 0$, then $G^c_e(x) \sim G^c(x)/\eta\gamma$ as $x \to \infty$.

(b) If $G^c(x) \sim \alpha x^{-p}$ as $x \to \infty$ for $\alpha > 0$ and $p > 1$, then $G^c_e(x) \sim \alpha x^{-(p-1)}/(p-1)\gamma$.

More generally, it is convenient to capture the possible power law tails for F and G within the framework of regularly varying functions. Recall that a function f defined on $[0, \infty)$ is said to be *regularly varying* of *index* ρ if

$$\lim_{x\to\infty} f(\lambda x)/f(x) = \lambda^\rho, \quad \text{for all } \lambda > 0; \tag{11.1}$$

see Bingham, Goldie and Teugels [6].

We can extend theorem 13(b) by applying Karamata's theorem, theorem 1.5.11(ii) of [6, p. 28].

Theorem 14. If G is a cdf with mean γ such that $G^c(x)$ is regularly varying with index $-p$ for $p > 1$, then $G^c_e(x) \sim xG^c(x)/\gamma(p-1)$ as $x \to \infty$.

We now treat the residual time mapping Θ. First we obtain a general result showing that $(\Theta F)^c$ inherits an exponential tail from G^c.

Theorem 15. If $G^c(x) \sim \alpha\,\mathrm{e}^{-\eta x}$ as $x \to \infty$, then $(\Theta F)^c(x) \sim \alpha'\,\mathrm{e}^{-\eta x}$ as $x \to \infty$, where

$$\alpha' = \alpha \int_0^\infty \mathrm{d}F(y)\bigl[\mathrm{e}^{-\eta y}/G^c(y)\bigr] < \infty. \tag{11.2}$$

Proof. The condition implies that there are positive constants k and K such that

$$0 < k < \inf_x \mathrm{e}^{\eta x} G^c(x) \leqslant \sup_x \mathrm{e}^{\eta x} G^c(x) \leqslant K < \infty,$$

which in turn implies that

$$\mathrm{e}^{\eta x} H^c_y(x) = \frac{G^c(x+y)\,\mathrm{e}^{\eta(x+y)}}{G^c(y)\,\mathrm{e}^{\eta y}} \leqslant \frac{K}{k} < \infty.$$

Hence, we can apply the dominated convergence theorem. In particular,

$$\mathrm{e}^{\eta x} H^c_y(x) \to \frac{\alpha}{G^c(y)\,\mathrm{e}^{\eta y}} \quad \text{as } x \to \infty,$$

so that

$$\mathrm{e}^{\eta x}(\Theta F)^c(x) = \int_0^\infty \mathrm{d}F(y) H^c_y(x)\,\mathrm{e}^{\eta x} \to \int_0^\infty \mathrm{d}F(y)\frac{\alpha}{G^c(y)\,\mathrm{e}^{\eta y}} \quad \text{as } x \to \infty,$$

where the last integral is finite because

$$\frac{\alpha}{G^c(y)\,\mathrm{e}^{\eta y}} \leqslant \frac{\alpha}{k}. \qquad \square$$

Theorem 15 tells us that (within the class of probability distributions considered) if G has an exponentially decaying tail, so does ΘF even if F is long-tailed. Hence the tails of ΘF and G and, hence by theorem 13b, also G_e, are dominated by the same exponential rate, so that $M(t)$ will eventually relax exponentially at this rate to γ.

In the following theorem we examine the effect upon ΘF of power-tails in F and G.

Theorem 16. If $G^c(x) \sim \alpha x^{-q}\,\mathrm{e}^{-\eta x}$ and $F^c(x) \sim \beta x^{-p}$ for p and $q, \eta \geqslant 0$ (but not $q = \eta = 0$). Then

$$\mathrm{e}^{\eta x}(\Theta F)^c(x) \sim \begin{cases} \beta p \Gamma[p]\Gamma[q-p]\big/\Gamma[q]\; x^{-p}, & \text{if } p < q, \\ \beta p\; x^{-p}\log x, & \text{if } p = q, \\ \alpha'\; x^{-q}, & \text{if } p > q, \end{cases} \tag{11.3}$$

where α' is defined in (11.2), and is finite when $p > q$.

Proof. For any $r > 0$ we can divide up $(\Theta F)^c(x)$ as

$$(\Theta F)^c(x) = A_r(x) + B_r(x), \tag{11.4}$$

where

$$A_r(x) = \int_0^r \mathrm{d}F(y) H^c_y(x) \quad \text{and} \quad B_r(x) = \int_r^\infty \mathrm{d}F(y) H^c_y(x). \tag{11.5}$$

For each $r < \infty$, the asymptotics of $A_r(x)$ as $x \to \infty$ are as follows. By the assumption on G, $\mathrm{e}^{\eta x} x^q G^c(x+y)$ converges to α as $x \to \infty$, uniformly for $y \in [0, r]$. Hence

$$A_r(x) \sim x^{-q}\,\mathrm{e}^{-\eta x}\alpha_r \quad \text{as } x \to \infty, \tag{11.6}$$

where

$$\alpha_r = \alpha \int_0^r F(\mathrm{d}y)\big[\mathrm{e}^{-\eta y}\big/G^c(y)\big]. \tag{11.7}$$

Now we turn to the asymptotics of $B_r(x)$. By assumption on F and G, for all k', $k > 1$, we can choose r' such that for all $r > r'$ and $x \geqslant 0$

$$\begin{aligned}
\mathrm{e}^{\eta x} B_r(x) &\leqslant kk \int_r^\infty \mathrm{d}F(y)(1 + x/y)^{-q} \\
&\leqslant kF^c(r)(1 + x/r)^{-q} + k \int_r^\infty \mathrm{d}y \; F^c(y) \frac{\mathrm{d}}{\mathrm{d}y}(1 + x/y)^{-q} \quad \text{(by parts)} \\
&\leqslant kF^c(r)(1 + x/r)^{-q} + kk'\beta \int_r^\infty \mathrm{d}y \; y^{-p} \frac{\mathrm{d}}{\mathrm{d}y}(1 + x/y)^{-q} \\
&\quad \text{(since } y \mapsto (1 + x/y)^{-q} \text{ is non-decreasing)} \\
&= k\bigl(1 - k'\bigr)F^c(r)(1 + x/r)^{-q} + kk'\beta p \int_r^\infty \mathrm{d}y \; y^{-p-1}(1 + x/y)^{-q} \\
&\leqslant kk' C_r(x),
\end{aligned}$$

where

$$C_r(x) = \beta p \int_r^\infty \mathrm{d}y \, y^{-p-1}(1 + x/y)^{-q}. \tag{11.8}$$

The argument can be repeated to get a lower bound by replacing k, k' by their reciprocals. Thus for all $k > 1$ we can find y such that

$$k^{-1} C_r(x) \leqslant B_r(x) \leqslant kC_r(x). \tag{11.9}$$

The advantage of working with $C_r(x)$ is that its asymptotics are relatively easy to obtain. By making the change of variable $z = t/x$ in (11.8), $C_r(x)$ can be written as $\beta p x^{-p}$ times $\int_0^{x/r} \mathrm{d}z \, z^{p-1}(1 + z)^{-q}$. As $x \to \infty$, this integral is either convergent to a finite limit $\Gamma[p]\Gamma[q - p]/\Gamma[q]$ (if $q > p$), or is divergent and $\sim \log x$ (if $q = p$), or is divergent and $\sim (p - q)^{-1}(x/r)^{p-q}$ (if $q < p$).

Thus when $q \leqslant p$, $A_r(x)$ is $\mathrm{o}(C_r(x))$ as $x \to \infty$, and the stated result follows since $k > 1$ is arbitrary. When $q > p$ then both $A_r(x)$ and $C_r(x)$ are $\mathrm{O}(x^{-q})$ and so

$$\limsup_{x\to\infty} x^q \, \mathrm{e}^{\eta x} (\Theta F)^c(x) \leqslant \alpha_r + k\beta p(p - q)^{-1} r^{q-p}, \tag{11.10}$$

with a corresponding lower bound for the lim inf obtained by replacing k with k^{-1}. But for fixed k, we are free to take $r \to \infty$ and hence $r^{q-p} \to 0$. From $q < p$, is follows that $\alpha_r \to \alpha < \infty$ and the stated result obtains. □

We now examine the long-tailed case, $\eta = 0$, in more detail (using the notation of theorem 16). We assume that $q > 1$ in order that the stationary-excess distribution G_e exists. Theorem 16 says that the tail of ΘF is dominated by the longer of the tails of F and G. So the tail of (ΘF) behaves like $t^{-\min[p,q]}$ while that of G_e behaves like $t^{-(q-1)}$. Hence we can identify two regimes. If $p < q - 1$, customers in service at $t = 0$ leave more slowly that new customers arrive. Thus, provided $\delta - \gamma > 0$ is sufficiently small, $M(t)$ can be expected to increase initially before relaxing to γ

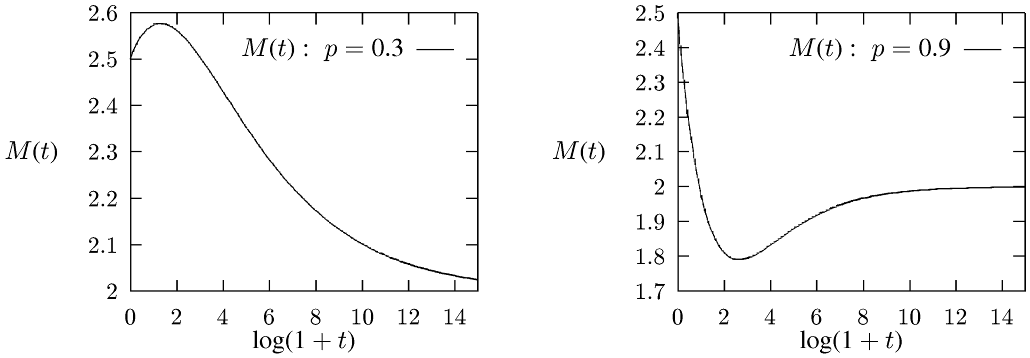

Figure 2. Evolution of $M(t)$. Left: $p = 0.3$, ages long. Right: $p = 0.9$, ages short.

according to a power law t^{-p}. But if $p > q - 1$, then new customers arrive more slowly, so we expect $M(t)$ to fall below γ according to a power law $t^{-\min[p,q]}$ before rising up to γ according to t^{1-q}.

Example 4. We illustrate this behavior for F and G Pareto with densities $p(1 + x)^{-1-p}$ and $q(1 + x)^{-1-q}$ respectively, so that $G^c(x) = (1 + x)^{-q}$, $\gamma = (q - 1)^{-1}$ and $G^c_e(x) = (1 + x)^{-q+1}$. In this case, one can express $(\Theta F)^c(x)$ directly in terms of the hypergeometric function ${}_2F_1$ without need of the approximation used during the proof of theorem 16, and

$$M(t) = \delta_2 F_1(p, q, 1 + p, -t) + \gamma\bigl(1 - (1 + x)^{1-q}\bigr).$$

In figure 2 we display $M(t)$ for $\delta = 2.5$, $q = 1.5$ and hence $\gamma = 2$. By (10.4), $M(t)$ goes from $\delta = 2.5$ to $\gamma = 2.0$, but figure 2 shows how. On the left $0.5 = q - 1 > p = 0.3$, while on the right $0.5 = q - 1 < p = 0.9$. Note that the horizontal time axis is logarithmic.

The case $p = 0.3$ corresponds to longer ages, while the case $p = 0.9$ corresponds to shorter ages. Longer (shorter) ages correspond to longer (shorter) residual times, since G is DFR. Longer residual times mean that the congestion will take longer to clear and, indeed, $M(t)$ goes up before it goes down when $p = 0.3$. On the other hand, shorter residual times mean that the congestion should clear relatively quickly and, indeed, $M(t)$ goes down quickly, even going below its eventual value $\gamma = 2.0$, when $p = 0.9$.

12. Large deviations

In this section we use large deviations theory to see how likely are various deviations from the mean in large systems. First, we can directly analyze what it means for the event $\{N \geqslant n\}$ to be unlikely. Since the event is a Poisson tail probability as indicated (1.1), we can directly show that

$$\frac{P(N \geqslant n)}{P(N = n)} \to 1 \quad \text{as } n \to \infty \tag{12.1}$$

and by Stirling's formula,

$$\begin{aligned} P(N = n) &\sim \frac{1}{\sqrt{2\pi n}} \left(\frac{m}{n}\right)^n \mathrm{e}^{-(m-n)} \\ &= \frac{1}{\sqrt{2\pi n}} \mathrm{e}^{-n(\log(n/m)-1+m/n)} \quad \text{as } n \to \infty. \end{aligned} \tag{12.2}$$

However, we are primarily interested in the case in which the mean m is large. To discuss this case, let N_m denote the Poisson distribution as a function of m. For moderately small tail probabilities, a normal approximation can be used, i.e., when m is not too small,

$$P\bigl(N_m \geqslant m + x\sqrt{m}\bigr) \approx \Phi^{\mathrm{c}}(x), \tag{12.3}$$

where $\Phi(x)$ is the standard (mean 0, variance 1) normal cdf and $\Phi^{\mathrm{c}}(x) = 1 - \Phi(x)$. However, for smaller tail probabilities it is better to use a large deviations approximation. If we let $m \to \infty$ and $n = cm$ (assumed to be integer) with $c > 1$, then

$$P(N_m = cm) \sim \frac{1}{\sqrt{2\pi cm}} \mathrm{e}^{-mI(c)} \quad \text{as } m \to \infty, \tag{12.4}$$

where

$$I(c) \equiv c(\log c - 1) + 1. \tag{12.5}$$

In fact, it is well known that a large deviation principle (LDP) holds for the distribution of $N(0)$.

Theorem 17. As $m \to \infty$, the distribution of $m^{-1}N_m$ satisfies an LDP with good rate function I in (12.5).

Proof. This follows from the Gärtner–Ellis theorem (see, e.g., Dembo and Zeitouni [14, section 2.3]) upon the observation that the cumulant generating function (cgf)

$$m^{-1} \log E\bigl[\mathrm{e}^{\theta N_m}\bigr] = \bigl(\mathrm{e}^{\theta} - 1\bigr) \tag{12.6}$$

exists for all θ, independent of m, is continuous and essentially smooth, and, as can be shown directly, has Legendre transform

$$I(c) = \sup_{\theta} \bigl\{c\theta - \bigl(\mathrm{e}^{\theta} - 1\bigr)\bigr\}. \tag{12.7}$$

□

We also note that the two asymptotic forms in (12.3) and (12.4) are different. If only $n \to \infty$ as in (12.1) and (12.2), then

$$\log P(N \geqslant n) \sim -n \log n \quad \text{as } n \to \infty, \tag{12.8}$$

whereas, by (12.7),

$$\log P(N_m \geqslant cm) \sim mI(c) \quad \text{as } m \to \infty. \tag{12.9}$$

We now turn to the empirical residual holding-time distribution at $t = 0$, conditional on $N(0) = n$. (Recall that the empirical residual holding-time distribution has the same distribution as the empirical age distribution.) The residual holding-time distribution can be expressed in terms of the random cdf S_n by

$$R_n(t) = 1 - n^{-1} S_n(t), \quad t \geqslant 0, \tag{12.10}$$

where $S_n(t)$ is the decreasing stochastic process whose marginals are distributed as in (2.5), so that

$$ER_n(t) = G_e(t), \quad t \geqslant 0. \tag{12.11}$$

By the FCLT supporting theorem 6,

$$\sqrt{n}\big(R_n(\cdot) - G_e(\cdot)\big) \Rightarrow W(\cdot) \quad \text{in } D[0,\infty) \text{ as } n \to \infty, \tag{12.12}$$

where W is a Gaussian process. From either (12.12) or (2.5),

$$\sqrt{n}\big(R_n(t) - G_e(t)\big) \Rightarrow \mathcal{N}\big(0, G_e(t) G_e^c(t)\big) \quad \text{as } n \to \infty, \tag{12.13}$$

where $\mathcal{N}(0, \sigma^2)$ is a random variable with mean 0 and variance σ^2. Hence, we already have some idea how $R_n(t)$ may deviate from $G_e(t)$

An alternative view is provided by an LDP. By virtue of theorem 1, an LDP for R_n follows directly from Sanov's theorem (see [14, section 6.2]). By reversibility of the underlying processes, an identical LDP holds for the distribution of the empirical age measures $A_n(0)$.

Theorem 18. Equipping the set of cdf's with the topology of weak convergence inherited from the space of measures, the empirical residual holding-time distributions R_n satisfy an LDP as $n \to \infty$ with good rate function J, where

$$J(F) = K(F, G_e), \tag{12.14}$$

the entropy of F relative to G_e (or their Kullback–Leibler "distance") defined as

$$K(F, G_e) = \int_0^\infty \mathrm{d}F(x) \log \frac{\mathrm{d}F}{\mathrm{d}G_e}(x), \tag{12.15}$$

for F absolutely continuous w.r.t. G_e, and ∞ otherwise.

Specifically, Sanov's theorem says that for Borel sets of cdf's $\mathcal{A}$,

$$\begin{aligned}-\inf_{F\in\mathcal{A}^0} J(F) &\leqslant \liminf_{n\to\infty} n^{-1}\log P\bigl(R_n\in\mathcal{A}\bigr)\\ &\leqslant \limsup_{n\to\infty} n^{-1}\log P\bigl(R_n\in\mathcal{A}\bigr) \leqslant -\inf_{F\in\bar{\mathcal{A}}} J(F),\end{aligned} \tag{12.16}$$

where $\mathcal{A}^0$ is the interior and $\bar{\mathcal{A}}$ the closure of $\mathcal{A}$. It is worth remarking that the stationary-excess cdf G_e in (2.1) always has a density g_e. Thus, a necessary (but not sufficient) condition for $J(F)$ to be finite is for F to have a density f, in which case (by an abuse of notation) we write $J(F)$ as

$$J(f) = \int_0^\infty \mathrm{d}t\ f(t)\log\frac{f(t)}{g_e(t)}. \tag{12.17}$$

The joint large deviation behavior of $N(0)$ and $R_n(0)$ can be seen informally by observing from (12.5) and (12.15)–(12.17) that for m large, the probability that $N(0)$ is near mc and $R_{N(0)}$ has density near f is roughly

$$\mathrm{e}^{-mI(c)}\,\mathrm{e}^{-mcJ(f)}. \tag{12.18}$$

We formulate this more precisely near the end of this section.

There are two important consequences of (12.15)–(12.17). First, the relative entropy in (12.17) enables us to quantify how rare are various age or residual life densities f that might appear instead of the stationary-excess density g_e.

Example 5. As in example 1, it is interesting to consider exponential and Pareto cdf's with a common mean. Given that $g_e(t) = \mathrm{e}^{-t}$, $t \geqslant 0$,

$$\begin{aligned}J(f) &= \int_0^\infty f(t)\log\left(\frac{f(t)}{\mathrm{e}^{-t}}\right)\mathrm{d}t\\ &= \int_0^\infty tf(t)\,\mathrm{d}t + \int_0^\infty f(t)\log f(t)\,\mathrm{d}t \approx \int_0^\infty tf(t)\,\mathrm{d}t.\end{aligned}$$

Hence, when $f(t) = f_p(t) \equiv p(1+t)^{-p-1}$, $t \geqslant 0$, $I(f) \approx 1/(p-1)$. Clearly $J(f_p)$ diverges as p approaches 1: longer residual holding-times become progressively more unlikely, and become impossible as the mean diverges, i.e., as $p \to 1$. We now make g_e longer tailed. If g_e is Pareto, in particular, if $g_e = f_q$, then

$$J(f_p) = \log(p/q) + (q-p)/p.$$

For fixed q, $J(f_p)$ is finite for all q. The curves of $J(f_p)$ for these two cases are shown in the left hand plot of figure 3. On the right hand plot we display $I(f)$ when $g_e = f_q$ as a function of q for f exponential. All these examples confirm that the likelihood of a given empirical residual holding-time distribution decreases with increasing disparity between its temporal characteristics and those of the stationary-excess distribution.

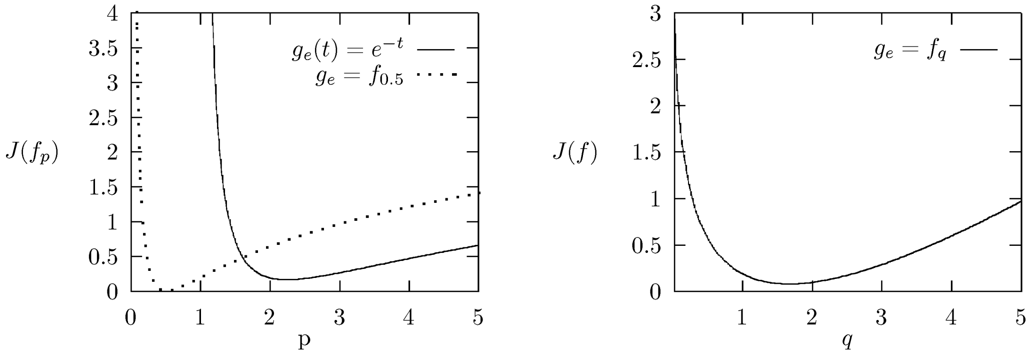

Figure 3. $J(f)$. Left: as a function of p for $f = f_p$. Right: as a function of q for $g_e = f_q$.

The second consequence of (12.15)–(12.17) is that the likelihood of R_n deviating from its expected value for large n is of the same exponential order as the likelihood of the event $\{N_m \geqslant cm\}$ for $c > 1$ in (12.9). This confirms that in considering rare congestion events we should consider the empirical age and residual holding-time distributions as well as large values for $N(0)$.

Example 6. In figure 4 we display contours of the joint rate function (12.18) when f is restricted to be in the family of Pareto densities $f_p = p(1 + x)^{-1-p}$ and $g_e = f_q$ with $q = 1$. Observe the narrowing of the contours upon moving to the right from $(1, 1)$. This illustrates that the prefactor c to J in (12.18) means that a given deviation in the age distribution become progressively less likely if it results from conditioning over an increasing number of customers.

We make some observations on the combination of example 5 and theorem 16. Recall that when F and G are Pareto with densities $f = f_p$ and $g = f_q$ and $p > q$, theorem 16 says that $(\Theta F)^c(x) \sim \alpha' x^{-q}$. Thus from the graphs in the left hand display of figure 3, one might expect that Θ has rendered the original age distribution F less unlikely by adjusting the tail exponent down from p to q. Indeed, if $R = \Theta A$ for a general age distribution A, then

$$J(R) = K(\Theta A, G_e) \overset{\text{(i)}}{=} K(\Theta A, \Theta G_e) \overset{\text{(ii)}}{\leqslant} K(A, G_e) = J(A). \tag{12.19}$$

Here, (i) follows since $\Theta G_e = G_e$, (ii) since Θ, being a Markov map, does not increase the relative entropy when applied to both its arguments; see [23, chapter 2]. However, more than one such age distribution A can be mapped into R by Θ. Indeed, when G is continuous we can approach equality in step (ii) above as closely as desired. This follows from the large deviation contraction principle. To see this, recall that by reversibility of the underlying processes, $A_n(0)$ and $R_n(0)$ satisfy an LDP with the same rate function as $n \to \infty$. For G continuous it follows that Θ is weakly continuous, i.e., $F_n \to F$ weakly implies $\Theta F_n \to \Theta F$ weakly. Thus by the large

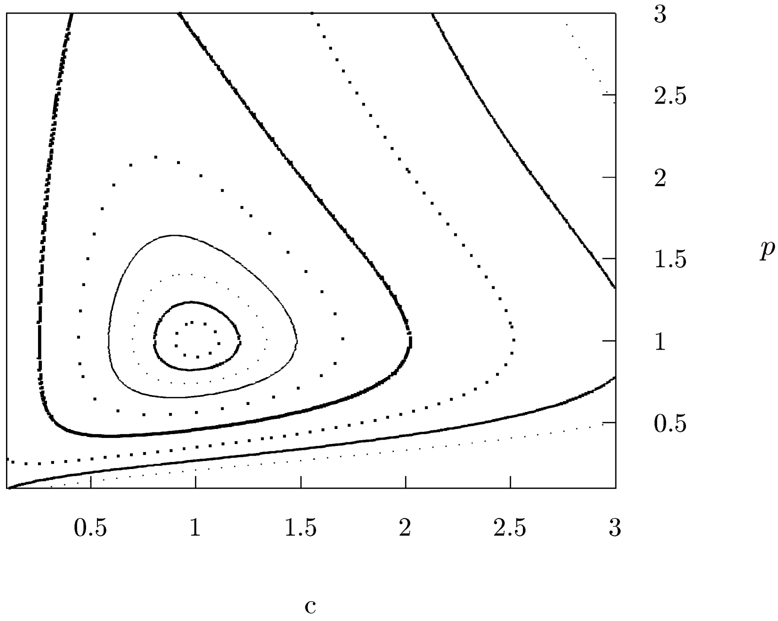

Figure 4. Contours of $I(c) + cJ(f_p)$ for $g_e = f_1$.

deviation contraction principle (see, e.g., [14, section 4.2])

$$J(R) = \inf_{A:R=\Theta A} J(A). \tag{12.20}$$

Instead of the pair $(N(t), A(t))$ we can focus on a single quantity containing both pieces of information, namely, the *empirical age measure*

$$\nu(t) = \sum_{i=1}^{N(t)} \delta_{y_{i,t}}, \tag{12.21}$$

where $y_{1,t}, \ldots, y_{N(t),t}$ are the ages of the holding times in process and δ_y is the Dirac measure (unit point mass) at y. The empirical age measure process $\{\nu(t)\colon\ t \geqslant 0\}$ is also a Markov process. We now justify the use of the joint rate function (12.18) by directly proving an LDP for the stationary empirical age measure $\nu(0)$ as the arrival rate $\lambda \to \infty$. By reversibility, the corresponding residual life measure $\Theta\nu(0)$ must satisfy the same LDP. We regard $\nu(0)$ is an element of $\mathcal{M}_+$, the set of positive measures (not necessarily probability measures) on $\mathbf{R}_+$. We topologize $\mathcal{M}_+$ with the topology of weak convergence (i.e., pointwise convergence on the set $\mathcal{C}_b(\mathbf{R}_+)$ of bounded continuous functions on $\mathbf{R}_+$) which is inherited the space $\mathcal{M}$ of signed measures on $\mathbf{R}_+$. The space $\mathcal{M}_+$ is metrizable as a complete separable space. For convenience, we denote $\int \mathrm{d}\nu(x) f(x)$ by $\langle \nu, f\rangle$ and $\langle \nu, 1\rangle$ by $\bar{\nu}$.

Theorem 19. As $m \to \infty$, the distribution of $m^{-1}\nu(0)$ satisfies an LDP with rate function

$$L(\nu) = \begin{cases} I(\bar\nu) + \bar\nu J(\nu/\bar\nu), & \bar\nu \neq 0, \\ I(0), & \bar\nu = 0. \end{cases} \tag{12.22}$$

Proof. Define for all $f \in \mathcal{C}_b(\mathbf{R}_+)$ and m the cgf

$$\begin{aligned}\Lambda_m(f) &= m^{-1}\log E\,\mathrm{e}^{\langle\nu(0),f\rangle} = m^{-1}\log E\langle G_e,\,\mathrm{e}^f\rangle^{\bar\nu(0)} = \langle G_e,\,\mathrm{e}^f\rangle - 1 \\ &\equiv \Lambda(f),\end{aligned} \tag{12.23}$$

independent of m. We now apply corollary 4.6.14 of [14]. For this purpose, it is straightforward to show that:

(i) $\Lambda(f)$ is finite for all $f \in \mathcal{C}_b(\mathbf{R}_+)$.

(ii) Λ is Gateaux differentiable; i.e., for all $f, g \in \mathcal{C}_b(\mathbf{R}_+)$, the function $\mathbf{R} \ni t \mapsto \Lambda(f + tg)$ is differentiable.

(iii) The set distributions of $m^{-1}\nu(0)$ is exponentially tight; i.e., for all $x < \infty$, there exists a compact set $\mathcal{K}_x \subset \mathcal{M}_+$ such that

$$\limsup_{\lambda\to\infty} m^{-1}\log P\left[m^{-1}\nu(0) \notin \mathcal{K}_x\right] < -x. \tag{12.24}$$

Item (iii) is due to the exponential tightness of the distributions of $m^{-1}\bar\nu(0)$ and $\nu(0)/\bar\nu(0)$, which follows from the goodness of the rate functions in the LDP's of theorems 17 and 18 (see, e.g., [14, exercise 1.2.9]). To see this, pick $0 < a < 1 < b$ and $\mathcal{K}$ a compact subset of probability measures in $\mathcal{M}_+$. Then (using $^{\mathrm{c}}$ to denote complements)

$$\begin{aligned}P\left[\nu(0) \in \left([a,b]\mathcal{K}\right)^{\mathrm{c}}\right] &= P\left[m^{-1}\bar\nu(0) \in [a,b]^{\mathrm{c}}, \nu(0)/\bar\nu(0) \in \mathcal{K}\right] \\ &\quad + P\left[m^{-1}\bar\nu(0) \in [a,b], \nu(0)/\bar\nu(0) \in \mathcal{K}^{\mathrm{c}}\right] \\ &\leqslant P\left[m^{-1}\bar\nu(0) \in [a,b]^{\mathrm{c}}\right] \\ &\quad + \sup_{n>am} P\left[\nu(0)/\bar\nu(0) \in \mathcal{K}^{\mathrm{c}} \mid \bar\nu(0) = n\right].\end{aligned} \tag{12.25}$$

Thus, $\limsup_{m\to\infty} m^{-1}\log P[m^{-1}\nu(0) \in ([a,b]\mathcal{K})^{\mathrm{c}}]$ is bounded above by the maximum of $-I(a)$, $-I(b)$ and

$$\limsup_{m\to\infty} m^{-1} \sup_{n>am} \log P\left[R_n \in \mathcal{K}^{\mathrm{c}}\right]$$

$$\leqslant \limsup_{m\to\infty} \sup_{n>am} (a/n)\log P\left[R_n \in \mathcal{K}^{\mathrm{c}}\right] \tag{12.26}$$

$$= a \limsup_{n\to\infty} n^{-1}\log P\left[R_n \in \mathcal{K}^{\mathrm{c}}\right] \tag{12.27}$$

$$\leqslant -a \inf_{\nu\in\overline{\mathcal{K}^{\mathrm{c}}}} J(\nu), \tag{12.28}$$

and so (12.24) can be fulfilled for each $x < \infty$ be choosing a, b, $\mathcal{K}$ appropriately.

Thus, [14, corollary 4.6.14], the distributions of $m^{-1}\nu(0)$ satisfy an LDP with a good convex rate function which is the Legendre transform of Λ:

$$\Lambda^*(\nu) = \sup_{f\in\mathcal{C}_b(\mathbf{R}_+)} \bigl\{\langle \nu, f\rangle - \Lambda(f)\bigr\}. \tag{12.29}$$

Then the given form of Λ^* follows from the duality [14, lemma 4.5.8] provided we can show, firstly, that Λ is the Legendre transform of L, and secondly that L is weak lower-semicontinuous. The second property follows from the weak lower-semicontinuity of K: see [14, lemma 6.2.12]. (For a sequence (ν_α) converging weakly to $\nu \neq 0$ then clearly $\liminf_\alpha L(\nu_\alpha) \geqslant L(\nu)$. If $\nu_\alpha \to 0$ then $\nu_\alpha/\bar{\nu}_\alpha$ need not converge. But in this case $L(\nu_\alpha) \geqslant I(\bar{\nu}_\alpha) \to I(0)$ as required.) We now establish the first property. For any $f \in \mathcal{C}_b(\mathbf{R}_+)$,

$$\sup_{\nu\in\mathcal{M}_+} \bigl\{\langle \nu, f\rangle - \Lambda(f)\bigr\} = \sup_{\omega\in\mathcal{M}_+^1;x\in\mathbf{R}_+} \bigl\{x\bigl(\langle \omega, f\rangle - K(\omega, G_e)\bigr) - I(x)\bigr\} \tag{12.30}$$

$$= \sup_{x\in\mathbf{R}_+} \bigl\{x \log\bigl\langle G_e,\, \mathrm{e}^f\bigr\rangle - I(x)\bigr\}. \tag{12.31}$$

The negative of the functional in the last display is strictly convex and steep on $\mathbf{R}_+$, so by differentiation the supremum is seen to occur for $x = \langle G_e, \mathrm{e}^f\rangle$, where the functional takes the value $\langle G_e, \mathrm{e}^f - 1\rangle = \Lambda(f)$, as required. □

References

[1] J. Abate, G.L. Choudhury and W. Whitt, Waiting-time tail probabilities in queues with long-tail service-time distributions, Queueing Systems 16 (1994) 311–338.

[2] J. Abate and W. Whitt, Calculating transient characteristics of the Erlang loss model by numerical transform inversion, Stochastic Models, to appear.

[3] F. Baccelli and P. Brémaud, *Elements of Queueing Theory* (Springer, New York, 1994).

[4] R.E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing* (Holt, Rinehart and Winston, New York, 1975).

[5] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).

[6] N.H. Bingham, C.M. Goldie and J.L. Teugels, *Regular Variation*, Encyclopedia of Mathematics and its Applications, Vol. 27 (Cambridge University Press, 1987).

[7] A.A. Borovkov, *Asymptotic Methods in Queueing Theory* (Wiley, New York, 1984).

[8] R. Cáceres, P.B. Danzig, S. Jamin and D.J. Mitzel, Characteristics of wide-area TCP/IP conversations, Computer Communication Review 21 (1991) 101–112.

[9] H. Chen and A. Mandelbaum, Discrete flow networks: bottleneck analysis and fluid approximation, Math. Oper. Res. 16 (1991) 408–446.

[10] G.L. Choudhury, K.K. Leung and W. Whitt, An inversion algorithm to compute blocking probabilities in loss networks with state-dependent rates, IEEE/ACM Trans. Networking 3 (1995) 585–601.

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

[12] M.E. Crovella and A. Bestavros, Self-similarity in World Wide Web traffic – evidence and possible causes, in: *Proc. Sigmetrics '96* (1996) pp. 160–169.

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

[14] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications* (Jones and Bartlett, Boston, 1993).

[15] N.G. Duffield, Conditioned asymptotics for tail probabilities in large multiplexers, Performance Evaluation, to appear.

[16] S.G. Eick, W.A. Massey and W. Whitt, The physics of the $M_t/G/\infty$ queue, Oper. Res. 41 (1993) 731–742.

[17] J.A. Erdelyi, *Asymptotic Expansions* (Dover, New York, 1956).

[18] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd ed. (Wiley, New York, 1971).

[19] P.G. Glynn and W. Whitt, A new view of the heavy-traffic limit for infinite-server queues, Adv. Appl. Probab. 23 (1991) 188–209.

[20] A.G. Greenberg, R. Srikant and W. Whitt, Resource sharing for book-ahead and instantaneous-request calls. AT&T Laboratories (1996), submitted.

[21] T. Kamae, U. Krengel and G.L. O'Brien, Stochastic inequalities on partially ordered spaces, Ann. Probab. 5 (1977) 899–912.

[22] E.V. Krichagina and A.A. Puhalskii, An asymptotic analysis of a closed queueing system with a GI/∞ service center, Institute for Problems in Information Transmission, Moscow, 1995.

[23] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).

[24] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Trans. Networking 2 (1994) 1–15.

[25] K.K. Leung, W.A. Massey and W. Whitt, Traffic models for wireless communication networks, IEEE J. Sel. Areas Commun. 12 (1994) 1353–1364.

[26] T. Lindvall, *Lectures on the Coupling Method* (Wiley, New York, 1992).

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

[28] V. Paxson, Empirically derived analytical models of wide-area TCP connections, IEEE/ACM Trans. Networking 2 (1994) 316–336.

[29] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models* (Wiley, New York, 1983).

[30] W. Whitt, The renewal-process stationary-excess operator, J. Appl. Probab. 22 (1985) 156–167.

[31] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson, Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level, in: *SIGCOMM Symp. on Commun. Arch. and Protocols* (1995) pp. 100–113.