Peak congestion in multi-server service systems with slowly varying arrival rates

William A. Massey [a] and Ward Whitt [b]

[a] *Bell Laboratories, Lucent Technologies, Room 2C-120, Murray Hill, NJ 07974-0636, USA*
E-mail: will@research.bell-labs.com
[b] *AT&T Labs, Room 2C-178, Murray Hill, NJ 07974-0636, USA*
E-mail: wow@research.att.com

Received 15 January 1995; revised 7 November 1996

In this paper we consider the $M_t/G/\infty$ queueing model with infinitely many servers and a nonhomogeneous Poisson arrival process. Our goal is to obtain useful insights and formulas for nonstationary finite-server systems that commonly arise in practice. Here we are primarily concerned with the peak congestion. For the infinite-server model, we focus on the maximum value of the mean number of busy servers and the time lag between when this maximum occurs and the time that the maximum arrival rate occurs. We describe the asymptotic behavior of these quantities as the arrival changes more slowly, obtaining refinements of previous simple approximations. In addition to providing improved approximations, these refinements indicate when the simple approximations should perform well. We obtain an approximate time-dependent distribution for the number of customers in service in associated finite-server models by using the modified-offered-load (MOL) approximation, which is the finite-server steady-state distribution with the infinite-server mean serving as the offered load. We compare the value and lag in peak congestion predicted by the MOL approximation with exact values for $M_t/M/s$ delay models with sinusoidal arrival-rate functions obtained by numerically solving the Chapman–Kolmogorov forward equations. The MOL approximation is remarkably accurate when the delay probability is suitably small. To treat systems with slowly varying arrival rates, we suggest focusing on the form of the arrival-rate function near its peak, in particular, on its second and third derivatives at the peak. We suggest estimating these derivatives from data by fitting a quadratic or cubic polynomial in a suitable interval about the peak.

Keywords: time-dependent arrival rates, slowly varying arrival rates, nonstationary queues, multi-server queues, infinite-server queues, peak congestion, time lag, uniform acceleration expansions, modified-offered-load approximation

1. Introduction

This paper is a sequel to Eick, Massey and Whitt [4,5] in which we gave relatively simple formulas describing the mean number of busy servers as a function of time in an $M_t/G/\infty$ queue (having a nonhomogeneous Poisson arrival process). In addition to directly describing the behavior in this model, these formulas were intended to pro-

vide insight into the performance of corresponding nonstationary finite-server systems (delay or loss) commonly encountered in practice.

Our purpose here is to highlight some implications of our previous results for the commonly occurring case in which the arrival-rate function $\lambda(t)$ changes slowly relatively to the mean service time. In this case, steady-state analysis applied locally at each time t tends to be appropriate even though there may be significant changes in the arrival rate over a longer time scale. Our goal is to better understand when the direct steady-state analysis is appropriate and to determine what modifications are most important. For background, see Hall [10, p. 178], Newell [19, Chapter 4], Green, Kolesar and Svoronos [9], Green and Kolesar [6,7], Whitt [20] and references therein.

In particular, here we focus on the value and time of the maximum expected number of busy servers in the infinite-server model. We assume that the arrival-rate function can be expanded in a power series about its peak in a way that makes successive coefficients negligible compared to previous coefficients. Then we obtain a corresponding power-series expansion for the infinite-server mean. This enables us to identify the dominant terms in approximations for the value and time of peak congestion when the arrival-rate function changes slowly. Consistent with Eick et al. [4] and Green and Kolesar [7,8], we find that the most important modification to a direct steady-state approximation when the arrival rate changes slowly is a time lag in the peak mean behind the peak arrival rate.

We also want to see how the information about peak congestion in the infinite-server model enables us to predict peak congestion in associated finite-server models with a fixed number of servers and unlimited waiting space. The general idea is that the infinite-server model should provide a reasonable approximation when the *actual* number of servers in the finite-server model is greater than the *mean* number of busy servers in the infinite-server model. When this condition is violated for significant periods of time, then there should be a buildup of customers in queue not receiving service, which is not accurately accounted for by the infinite-server model.

To investigate the quality of infinite-server approximations for finite-server models, we consider Markovian $M_t/M/s$ delay models, for which we can calculate the exact time-dependent distribution of the number of customers in the system by numerically solving the Chapman–Kolmogorov forward equations, using a variant of the algorithm in Davis, Massey and Whitt [3], using a large finite waiting room to make the state space finite. The predicted location of the peak in the finite-server model is precisely the location of the peak in the infinite-server model. We compare the peak congestion and the location of the peak in the finite-server system to the exact and approximate peak congestion and location of the peak in the infinite-server system.

In order to obtain an approximation for the peak congestion in the finite-server model based on the peak value of the infinite-server mean, we use the *modified-offered-load* (MOL) *approximation*, as in Jagerman [11] and Massey and Whitt [16]. (Those papers focus on loss models, but the MOL approximation applies to delay models in the same way.) The MOL approximate distribution of the peak number of customers in the $M_t/M/s$ system is the steady-state distribution of the stationary $M/M/s$ model with

offered load equal to the peak infinite-server mean. Equivalently, the traffic intensity ρ in the steady-state distribution is the peak infinite-server mean divided by s. We show that the MOL approximation performs well when the traffic intensity is not too high.

Here is how this paper is organized. We state our main result in section 2. In section 3 we discuss the special case of sinusoidal arrival-rate functions considered in Eick et al. [5]. In section 4 we suggest that quadratic or cubic approximations fit in a neighborhood of the peak are likely to be more effective than sinusoidal approximations for realistic slowly-varying periodic arrival rates arising in practice. In section 5 we provide illustrative numerical comparisons. Finally, we prove our theorem in section 6.

Since we focus on peak congestion, our results here provide useful information about the number of servers needed to meet peak congestion. The more general problem of dynamic staffing to meet time-varying demand is considered in Jennings, Mandelbaum, Massey and Whitt [12].

2. The main result

We assume that the service times are independent and identically distributed, and independent of the arrival process. Let S denote a generic service-time random variable. Without loss of generality, we assume that a service time S has mean 1. Then the arrival-rate function $\lambda(t)$ is the *relative arrival rate*; the relative arrival rate is the time-dependent analog of the offered load. We assume that the system starts empty in the distant past. Then the number of busy servers at time t has a Poisson distribution for each t with a mean

$$m(t) = \int_{-\infty}^{t} G^c(u)\lambda(t-u)\,\mathrm{d}u = E\big[\lambda(t-S_e)\big], \tag{1}$$

where S_e is a random variable with the *service-time stationary-excess distribution*, i.e.,

$$P(S_e \leqslant t) = \int_0^t P(S > u)\,\mathrm{d}u, \tag{2}$$

see Eick et al. [4, eqs. (1) and (3)].

Formula (1) shows that the time-dependent mean coincides with the relative arrival rate $\lambda(t)$ except for a *random time lag* S_e. We thus say that there is a random time lag of S_e in $m(t)$ after $\lambda(t)$. In general, the actual time lag in the mean $m(t)$ behind the arrival rate $\lambda(t)$ differs from the mean $E[S_e]$ due to the nonlinearity of the arrival rate function $\lambda(t)$. However, the mean $E[S_e]$ is a natural initial approximation for the time lag; see Eick et al. [4, Remark 10 and section 3] and the discussion below. Fortunately, the moments of S_e are simply related to the moment of S, i.e.,

$$E\big[S_e^k\big] = \frac{E[S^{k+1}]}{kE[S]} = \frac{E[S^{k+1}]}{k}. \tag{3}$$

From (1) and (3), we can see the role of the service-time distribution.

Before stating our main result formally, we discuss the notion of peak congestion informally. For this informal discussion, let t_λ and t_m be the times of the peaks (maximum values) of $\lambda(t)$ and $m(t)$, respectively, which for simplicity we assume are unique. If $\lambda(t)$ is nondecreasing before t_λ, then $t_m > t_\lambda$; see Eick et al. [4, Theorem 5]. More generally, we typically have $t_m > t_\lambda$, but it is not difficult to construct counterexamples. Suppose that $\lambda(t)$ is unimodal in a relevant interval about t_λ and that t_m falls in this interval, so that $t_m > t_\lambda$. We are interested in the *time lag in the peaks*

$$L \equiv t_m - t_\lambda. \tag{4}$$

The initial approximation mentioned above is

$$L \approx ES_e = \frac{m_2}{2}, \tag{5}$$

where m_k is the kth moment of S. (Recall that $ES = 1$.) Approximation (5) was obtained from the linear and quadratic approximations in Eick et al. [4]; see Remark 10, Example 1 and section 3 there (especially Theorem 9).

We are also interested in the values of the peak $m(t_m)$ and $\lambda(t_\lambda)$. From (1) we see that $m(t_m) < \lambda(t_\lambda)$, because $m(t_m)$ is an average of $\lambda(t)$ for t to the left of t_m, where $\lambda(t) \leqslant \lambda(t_\lambda)$. We are interested in the *difference in the peaks*

$$D \equiv \lambda(t_\lambda) - m(t_m). \tag{6}$$

A natural initial approximation is $m(t_m) \approx \lambda(t_\lambda)$ or

$$D \approx 0. \tag{7}$$

Approximation (7) can be obtained from the *pointwise stationary approximation* (PSA), which approximates the distribution at time t by the steady-state distribution associated with the stationary model having arrival rate $\lambda(t)$. The steady-state mean number of busy servers in the infinite-server model associated with arrival rate $\lambda(t)$ is $\lambda(t)ES = \lambda(t)$.

Our primary purpose in this paper is to obtain refinements to approximations (5) and (7) in the case $\lambda(t)$ changes relatively slowly in a relevant interval about its peak t_λ. These refinements yield better approximations and indicate when the simple approximation in (5) and (7) should perform well. To obtain refinements to (5) and (7), we scale a fixed arrival-rate function in the neighborhood of its peak. In fact, our approach permits t_λ to be the location of any local maximum of the arrival rate function $\lambda(t)$. We then rescale $\lambda(t)$ so that only the neighborhood of t_λ is relevant.

Hence, let $\lambda(t)$ be any arrival-rate function with a local maximum at t_λ. We assume that $\lambda(t)$ has a Taylor-series expansion in the neighborhood of t_λ. Then we form a family of functions indexed by ε by letting

$$\lambda_\varepsilon(t) = \lambda\big(t_\lambda + \varepsilon(t - t_\lambda)\big) \tag{8}$$

and consider the behavior as $\varepsilon \to 0$. We thus think of the actual arrival rate function being $\lambda_\varepsilon(t)$ for some small ε. If we first move the peak t_λ to the origin, which we can do without loss of generality, then (8) is equivalent to the direct time scaling

$$\lambda_\varepsilon(t) = \lambda(\varepsilon t). \tag{9}$$

Since $\lambda(t)$ has a power-series expansion about t_λ, so does $\lambda_\varepsilon(t)$ in (8), and it takes the form

$$\lambda_\varepsilon(t) = \sum_{k=0}^{\infty} \frac{\lambda_\varepsilon^{(k)}(t_\lambda)}{k!}(t - t_\lambda)^k = \sum_{k=0}^{\infty} \frac{\lambda^{(k)}(t_\lambda)}{k!}\varepsilon^k(t - t_\lambda)^k\ , \tag{10}$$

where $\lambda_\varepsilon^{(k)}(t)$ and $\lambda^{(k)}(t)$ are the kth derivatives of $\lambda_\varepsilon(t)$ and $\lambda(t)$, respectively. The nature of a power series expansion is that its value in the neighborhood of a point can be approximated (to arbitrary precision) by using the derivatives of the function at that one point. This is the spirit of *perturbation theory* which is eloquently described by Bender and Orszag [2, p. 319]. When $\varepsilon = 1$, $\lambda_\varepsilon(t)$ becomes the original arrival-rate function $\lambda(t)$. When ε is close to 0, $\lambda_\varepsilon(t)$ corresponds to an arrival-rate function that is slowly varying and close to the constant rate of $\lambda(t_\lambda)$, a local maximum for $\lambda(t)$.

When ε is small, the successive coefficients of $(t - t_\lambda)^k$ in (10) are indeed negligible compared to all previous coefficients. Thus the representation (8) or (10) serves to justify the polynomial approximations previously considered in Eick et al. [4, section 3]. This approach also coincides with the uniform acceleration expansions in Massey [13], Eick et al. [4, Remark 15, p. 739] and Massey and Whitt [17].

Let $m_\varepsilon(t)$ be the infinite-server mean associated with $\lambda_\varepsilon(t)$ in (8), let $t_m(\varepsilon)$ be a local maximum time for m_ε, let $\overline{m}(\varepsilon) = m_\varepsilon(t_m(\varepsilon))$ be the local value attained and let $L(\varepsilon)$ and $D(\varepsilon)$ be the associated lag and difference. We will show that $t_m(\varepsilon)$ and $\overline{m}(\varepsilon)$ are well defined below. We justify (5) and (7) and identify the next most important terms by expanding $L(\varepsilon)$ and $D(\varepsilon)$ in powers of ε. Here is our main result.

Theorem 1. Suppose that λ is 6-times differentiable and $\lambda^{(k)}$ is bounded and Riemann integrable on the interval $(-\infty, t_\lambda]$ for $1 \leqslant k \leqslant 6$. Suppose that $ES = 1$ and $ES^6 < \infty$. If $\lambda_\varepsilon(t)$ is defined by (8), $\lambda^{(1)}(t_\lambda) = 0$ and $\lambda^{(2)}(t_\lambda) < 0$, then the associated mean function $m_\varepsilon(t)$ defined by (1) has a local maximum $\overline{m}(\varepsilon)$ at time $t_m(\varepsilon)$ for all suitably small ε, with a lag

$$\begin{aligned} L(\varepsilon) \equiv t_m(\varepsilon) - t_\lambda &= E[S_e] - \varepsilon\frac{\lambda^{(3)}(t_\lambda)}{2\lambda^{(2)}(t_\lambda)}\operatorname{Var}[S_e] \\ &\quad + \varepsilon^2\frac{\lambda^{(4)}(t_\lambda)}{6\lambda^{(2)}(t_\lambda)}E\big[\big(S_e - E[S_e]\big)^3\big] + \mathrm{O}\big(\varepsilon^3\big) \end{aligned} \tag{11}$$

and a difference

$$\begin{aligned} D(\varepsilon) &\equiv \lambda(t_\lambda) - m\big(t_m(\varepsilon)\big) \\ &= -\varepsilon^2\frac{\lambda^{(2)}(t_\lambda)}{2}\operatorname{Var}[S_e] + \frac{\varepsilon^3\lambda^{(3)}(t_\lambda)}{6}E\big[\big(S_e - E[S_e]\big)^3\big] + \mathrm{O}\big(\varepsilon^4\big) \end{aligned} \tag{12}$$

as $\varepsilon \to 0$.

In applications we typically have a single arrival-rate function $\lambda(t)$, not a parametric family $\lambda_\varepsilon(t)$ as in (8) or (9). For applications, it seems more meaningful to re-express (11) and (12) in terms of $\lambda_\varepsilon^{(k)}(t_\lambda)$, because our given arrival-rate function is $\lambda_\varepsilon(t)$ in (8) for some fixed small ε. When we do this, the ε factors prior to the final error terms disappear, i.e., (11) and (12) are equivalent to

$$L(\varepsilon) = E[S_e] - \frac{\lambda_\varepsilon^{(3)}(t_\lambda)}{2\lambda_\varepsilon^{(2)}(t_\lambda)}\,\mathrm{Var}[S_e] + \frac{\lambda_\varepsilon^{(4)}(t_\lambda)}{6\lambda_\varepsilon^{(2)}(t_\lambda)}E\big[\big(S_e - E[S_e]\big)^3\big] + \mathrm{O}\big(\varepsilon^3\big) \tag{13}$$

and

$$D(\varepsilon) = -\frac{\lambda_\varepsilon^{(2)}(t_\lambda)}{2}\,\mathrm{Var}[S_e] + \frac{\lambda_\varepsilon^{(3)}(t_\lambda)}{6}E\big[\big(S_e - E[S_e]\big)^3\big] + \mathrm{O}\big(\varepsilon^4\big). \tag{14}$$

Theorem 1 expressed this way could also be obtained by a direct Taylor series expansion, using the assumption that the kth derivative $\lambda^{(k)}(t_\lambda)$ is $\mathrm{O}(\varepsilon^k)$ for some small ε. Indeed, the first terms of $L(\varepsilon)$ and $D(\varepsilon)$ in (13) and (14) were already obtained this way via the quadratic approximation in Eick et al. [4, section 3]. The second-order approximation for the difference

$$D(\varepsilon) \approx -\varepsilon^2\frac{\lambda^{(2)}(t_\lambda)}{2}\mathrm{Var}[S_e] \tag{15}$$

coincides with the space shift in the quadratic approximation QUAD-D in Eick et al. [4, section 3]. Theorem 1 here adds new terms to what is directly deducible from Eick et al. [4].

From (8) and (9) and the proof of Theorem 1 we see that the successive terms in the expansions (including ones beyond the ones we display) depend on the service-time distribution through the central moments of S_e. The higher-order terms involving $E[(S_e - E[S_e])^3]$ will disappear when the distribution of S_e is symmetric about its mean, which happens if and only if the distribution of S is deterministic (because the density of S_e is $P(S > t)/E[S]$).

3. Sinusoidal arrival-rate functions

It is interesting to consider the special case of sinusoidal arrival-rate functions, because we can obtain convenient explicit formulas for them and because the asymptotic relation (8) is natural to consider for them. Their periodic form is also in the spirit of many real systems with daily cycles.

Hence, suppose that we have a family of arrival rate functions indexed by ε, defined by

$$\lambda_\varepsilon(t) = \overline{\lambda} + \beta\sin(\varepsilon t), \tag{16}$$

where as before the service time S has mean 1. Eick et al. [5] showed that

$$m_\varepsilon(t) = \overline{\lambda} + \beta\big(\sin(\varepsilon t)E\big[\cos(\varepsilon S_e)\big] - \cos(\varepsilon t)E\big[\sin(\varepsilon S_e)\big]\big), \tag{17}$$

where as before S_e has the stationary-excess distribution in (2). In this context, the peak for λ_ε is $t_\lambda(\varepsilon) = \pi/2\varepsilon$. Eick et al. [5] showed that the time lag $L(\varepsilon)$ and difference in the peaks $D(\varepsilon)$ are

$$L(\varepsilon) = \varepsilon^{-1}\arctan\big(E[\sin\varepsilon S_e]/E[\cos\varepsilon S_e]\big) \tag{18}$$

and

$$D(\varepsilon) = \beta - \beta\big(\big(E\cos(\varepsilon S_e)\big)^2 + \big(E\big[\sin(\varepsilon S_e)\big]\big)^2\big)^{1/2}. \tag{19}$$

What is nice about the sinusoidal arrival-rate function in (16) is that the notion of "slowly changing" is represented simply by the frequency ε. The arrival-rate function is slowly changing when ε is suitably small. Hence we can directly apply Theorem 1 in section 1.

However, there is a complication. As noted before Eick et al. [5, Theorem 4.4], the peak t_λ goes to infinity as $\varepsilon \to 0$. This is already accounted for in (18) and (19), but could be avoided at the outset if we moved the peak to the origin. The peak can be moved to the origin by replacing (16) with

$$\lambda_\varepsilon(t) = \overline{\lambda} + \beta\cos(\varepsilon t). \tag{20}$$

We can describe the asymptotic behavior as $\varepsilon \to 0$ in (16) or (20) either by applying Theorem 1 here or by directly letting $\varepsilon \to 0$ in (18) and (19). Indeed Eick et al. [5] already showed that limiting value $\overline{m}(0)$ is $\overline{\lambda} + \beta$ in their Theorem 4.4. They also showed that the limiting lag $L(0)$ is $1/2$ for a deterministic service-time distribution and 1 for an exponential service-time distribution; see (26), (16) and the remark below (17) in Eick et al. [5].

The asymptotic behavior is especially easy to see from Theorem 1, because $\lambda^{(k)}(t_\lambda) = 0$ when k is odd and $\lambda^k(t_\lambda) = 1$ when k is even. Note that quadratic and sine functions share the special property that $\lambda^{(3)}(t_\lambda) = 0$, which eliminates the second terms in (11) and (12). From Theorem 1, we obtain asymptotic formulas for the lag and difference, namely,

$$L(\varepsilon) = E[S_e] + \frac{\varepsilon^2}{6}E\big[\big(S_e - E[S_e]\big)^3\big] + \mathrm{O}\big(\varepsilon^4\big) \quad \text{as } \varepsilon \to 0 \tag{21}$$

and

$$D(\varepsilon) = \beta\frac{\varepsilon^2}{2}\,\mathrm{Var}[S_e] + \mathrm{O}\big(\varepsilon^4\big) \quad \text{as } \varepsilon \to 0. \tag{22}$$

We can also obtain formulas (21) and (22) from (18) and (19) with a little bit more work. In particular, we can apply familiar asymptotic expansions of trigonometric functions, 4.3.65, 4.3.66 and 4.4.42 of Abramowitz and Stegun [1],

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots, \tag{23}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots, \tag{24}$$

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \tag{25}$$

as $x \to 0$. From (18)–(19) and (23)–(25) plus 3.6.18, 3.6.21 and 3.6.22 of Abramowitz and Stegun [1], we obtain (21) and (22). Note that (21) and (22) are consistent with (5) and (7); i.e., $L(0) = m_2/2 = ES_e$, consistent with (5) and $D(0) = 0$, consistent with (7).

In the case of deterministic service times, (21) and (22) become

$$L(\varepsilon) = \frac{1}{2} + \mathrm{O}(\varepsilon^4) \quad \text{as } \varepsilon \to 0 \tag{26}$$

and

$$D(\varepsilon) = \frac{\beta\varepsilon^2}{24} + \mathrm{O}(\varepsilon^4) \quad \text{as } \varepsilon \to 0, \tag{27}$$

which is consistent with exact results in (26) and (27) of Eick et al. [5]. For the $M_t/D/\infty$ model with sinusoidal arrival rate, *all* error terms in the lag $L(\varepsilon)$ disappear because $\lambda^{(k)}(t_\lambda) = 0$ and $E[(S_e - E[S_e])^k] = 0$ for all *odd* k.

In the case of exponential service times

$$L(\varepsilon) = 1 - \frac{\varepsilon^2}{3} + \mathrm{O}(\varepsilon^4) \quad \text{as } \varepsilon \to 0 \tag{28}$$

and

$$D(\varepsilon) = \frac{\beta\varepsilon^2}{2} + \mathrm{O}(\varepsilon^4) \quad \text{as } \varepsilon \to 0. \tag{29}$$

Formulas (28) and (29) are consistent with exact results (16) and (18) of Eick et al. [5], namely,

$$L(\varepsilon) = \varepsilon^{-1} \arctan(\varepsilon), \tag{30}$$

$$D(\varepsilon) = \beta - \frac{\beta}{\sqrt{1+\varepsilon^2}}. \tag{31}$$

Formula (28) follows directly from (30) and (25) above. Formula (29) follows directly from (31) by taking a Taylor-series expansion.

The fact that the first error terms in (21) and (22) are of order $\mathrm{O}(\varepsilon^2)$ indicates that the approximations based on the limit $\varepsilon = 0$ should often perform well provided that ε is suitably small, as recently shown numerically for the cases of exponential and deterministic service-time distributions by Green and Kolesar [7]. The explicit constants given for the $\mathrm{O}(\varepsilon^2)$ terms help us understand departures from this limiting case.

To illustrate, we display both the approximations (28) and (29) and the exact values (30) and (31) for the lag $L(\varepsilon)$ and the difference $D(\varepsilon)$ for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) in table 1. Note that the lag and the relative the difference $D(\varepsilon)/\beta$ depend only on the frequency. For ε small, e.g., for $\varepsilon \leqslant 0.2$, the approximation (28) and (29) are quite accurate. More importantly,

Table 1
A comparison of approximations with exact values for the lag $L(\varepsilon)$ and the normalized difference $D(\varepsilon)/\beta$ for the $M_t/M/\infty$ model with the sinusoidal arrival rate in (16). The formulas are in (28)–(31).

frequency	lag $L(\varepsilon)$		relative difference $D(\varepsilon)/\beta$	
ε	exact	approx.	exact	approx.
10.0	0.1471	–	0.9005	–
5.0	0.2747	–	0.8039	–
2.0	0.5536	–	0.5528	–
1.0	0.7854	0.6667	0.2929	0.5000
0.5	0.9273	0.9167	0.1056	0.1250
0.2	0.9870	0.9867	0.0195	0.0200
0.1	0.9967	0.9967	0.0050	0.0050
0.0+	1.0000	1.0000	0.0000	0.0000

though, the simple approximations $L \approx ES_e = 1$ and $D \approx 0$ clearly perform well in this region.

4. Analyzing real periodic systems

In this section we make some suggestions about what seem to be appropriate ways to analyze real nonstationary multi-server service systems that are characterized by slowly changing arrival-rate functions. As before in this paper, "slowly-changing" means relative to the mean service time. Since many real systems clearly have periodic or nearly periodic arrival-rate functions, where the maximum is much greater than the minimum, it is natural to consider periodic arrival-rate functions.

The periodicity led many researchers, including Eick et al. [5], to consider the special case of sinusoidal arrival-rate functions. However, when the arrival-rate function changes slowly, the periodic nature tends to become less and less relevant. The periodic nature tends to be important only when the behavior at any time is influenced by the system behavior more than one cycle previously. However, when the arrival rate changes slowly, as when service times are in minutes with daily cycles, the relevant history to determine the system congestion at any time rarely goes back a full day. What really is relevant (when there are negligible queues of customers waiting to begin service) is a time interval extending back only several mean service times from the time of interest.

From section 1 we see that what is relevant to determine the congestion at times near the peak arrival-rate function is to know the arrival-rate function near the peak. In order to apply the first refined approximations (second terms) in (13) and (14), we need to know only the second and third derivatives of the arrival-rate function at its peak, $\lambda^{(2)}(t_\lambda)$ and $\lambda^{(3)}(t_\lambda)$. We suggest focusing on these quantities.

The problem with sinusoidal arrival-rate function models is that, if we take account of the full periodic structure, the frequency ε in (16) or (20) will be determined by the long-term behavior rather than the local behavior, because it is determined by

the cycle lengths. Hence, $\lambda^{(2)}(t_\lambda) = -\beta\varepsilon^2$ and $\lambda^{(3)}(t_\lambda) = 0$, where ε is determined by the cycle lengths. In contrast, what we should really do to obtain a good approximation is directly estimate $\lambda^{(2)}(t_\lambda)$ and $\lambda^{(3)}(t_\lambda)$ themselves by examining $\lambda(t)$ in the neighborhood of its peak t_λ. It may happen that $\lambda^{(2)}(t_\lambda) \approx -\beta\varepsilon^2$ and $\lambda^{(3)}(t_\lambda) \approx 0$, but it need not.

For example, an application may have a daily cycle with $\overline{\lambda} = \beta = 100$ and a mean service time of about 23 minutes. A direct sinusoidal model then dictates that $\varepsilon \approx 0.1$ (assuming $ES = 1$). This sinusoidal arrival rate function has second derivative $-\varepsilon^2\beta = -1$. However, the actual arrival-rate function could have a much bigger second derivative at its peak, say $\lambda^{(2)}(t_\lambda) \approx 25$. Since $\beta = 100$, this means that a sinusoidal arrival-rate function fit locally to $\lambda^{(2)}(t_\lambda)$ should have frequency $\varepsilon \approx 0.5$.

Having suggested estimating $\lambda^{(2)}(t_\lambda)$ and $\lambda^{(3)}(t_\lambda)$, it is appropriate to consider how. When considering possible estimation procedures, it is good to keep in mind that our real goal is to yield an approximation for the *integral*

$$m(t) = \int_{-\infty}^{t} G^c(t-u)\lambda(u)\,\mathrm{d}u \approx \int_{t-10}^{t} G^c(t-u)\lambda(u) \tag{32}$$

for $t_\lambda \leqslant t \leqslant t_m$. From (32), we clearly see that the behavior of $\lambda(t)$ in a very small immediate neighborhood of t_λ is less important than the *average* behavior over an interval centered at t_λ of length equal to a few mean service times. For example, a reasonable procedure is to fit a quadratic or cubic function to data over the interval $[t_\lambda - 4, t_\lambda + 4]$ using regression. A maximum likelihood estimator of the coefficients can be obtained from iterative weighted least squares, as was done for the linear case in Massey, Parker and Whitt [14]; see McCullagh and Nelder [18]. Massey, Parker and Whitt found that ordinary least squares was nearly as efficient.

5. Peak congestion with finitely many servers

In this section we investigate how well the exact and approximate formulas for the lag in the peak of the infinite-server mean $m(t)$ predict the lag in peak congestion for finite-server systems. We also investigate how well the MOL approximation using the exact or approximate peak $m(t_m)$ predicts the actual peak performance in finite-server delay systems. For this purpose, we consider the Markovian $M_t/M/s$ models with a nonhomogeneous Poisson arrival process and a sinusoidal arrival-rate function. We apply a variant of the algorithm described in Davis et al. [3] to compute the time-dependent probability distribution of the number $Q_s(t)$ of customers in the system at time t. We consider three specific performance measures: the probability of delay $P(Q_s(t) \geqslant s)$, the expected number in queue $E[(Q_s(t) - s)^+]$, where $(x)^+ = \max\{x, 0\}$, and the tail probability $P(Q_s(t) \geqslant s + 5)$.

We anticipate that, for a given arrival-rate function, the infinite-server approximation for the lag will perform better for larger s, because then the finite-server model is closer to an infinite-server model. To focus on this phenomenon, we describe the three performance measures as a function of s.

Table 2
The actual lags in peak congestion for three performance measures as a function of the number of servers, s, for the arrival-rate function $20 + 10\sin(0.2t)$.

number of servers s	peak delay probability	lag in delay probability $P(Q(t_m) \geqslant s)$	lag in tail probability $P(Q(t_m) \geqslant s+5)$	lag in mean number in queue
∞		0.99	0.99	0.99
55	0.000024	1.00	1.03	1.02
50	0.00048	1.01	1.08	1.04
45	0.0062	1.04	1.12	1.09
42	0.023	1.08	1.19	1.16
40	0.050	1.13	1.27	1.25
38	0.100	1.22	1.39	1.39
35	0.245	1.42	1.67	1.75
32	0.493	1.80	2.10	2.38
30	0.692	2.16	2.76	2.98
28	0.862	2.64	2.99	3.71
25	0.984	3.52	3.88	4.68
22	0.9997	4.52	4.87	4.78

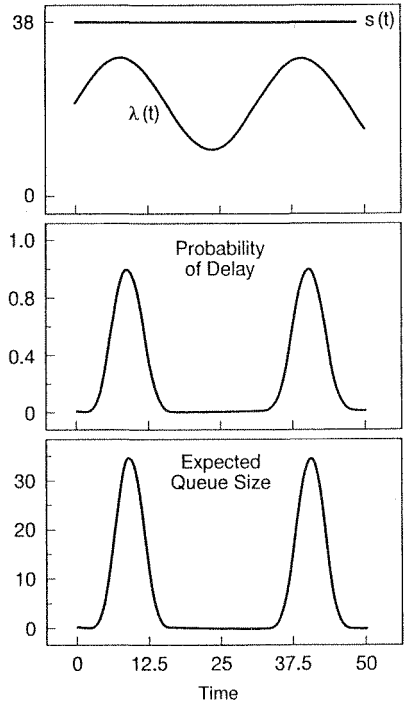

Figure 1. The arrival-rate function, probability of delay and the mean number of customers in queue in the $M_t/M/s$ model with $s = 38$ servers, $ES = 1$ and arrival-rate function $\lambda(t) = 20 + 10\sin(0.2t)$.

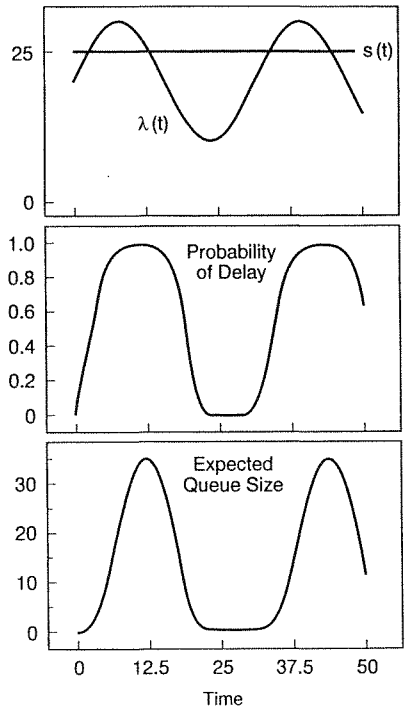

Figure 2. The arrival-rate function, probability of delay and the mean number of customers in queue in the $M_t/M/s$ model with $s = 25$ servers, $ES = 1$ and arrival-rate function $\lambda(t) = 20 + 10\sin(0.2t)$.

As a specific example, we consider the sinusoidal arrival-rate function (16) with $\overline{\lambda} = 20$, $\beta = 10$ and $\varepsilon = 0.2$. We have chosen the frequency ε small, so that the arrival-rate function is changing relatively slowly. From table 1, we see that the exact infinite-server lag of 0.987 is indeed close to the approximate infinite-server lag of $E[S_e] = E[S] = 1$. (For an exponential distribution, S_e is distributed the same as S.)

Table 2 displays the actual lags in the peak for the three performance measures as a function of s. When s is large, the actual lags are very close to the infinite-server lag. As long as the actual delay probability is relatively small, say less than 0.10, the infinite-server lag still yields a decent approximation. However, as the number of servers decreases, then the actual lag grows significantly. This behavior should be anticipated, especially when $s < 30$, because then the instantaneous traffic intensity exceeds 1 at the peak.

Figure 1 displays the three performance measures in the relatively good case in which $s = 38$. For $s \geqslant 38$, the infinite-server approximation performs pretty well. In contrast, figure 2 displays the same performance measures when $s = 25$. Since the infinite-server mean exceeds $s = 25$ for a substantial period, the infinite-server approximation no longer performs well.

Table 3
A comparison of the modified-offered-load (MOL) approximation with exact values of the peak delay probability and peak mean waiting time as a function of the number of servers for the model with arrival-rate function $\lambda(t) = 20 + 10\sin(0.2t)$ and mean service time $ES = 1$. The exact infinite-server peak mean 29.81 is used as the offered load for MOL.

number of servers s	delay probability		mean waiting time	
	exact	MOL	exact	MOL
50	0.00048	0.00048	0.0012	0.0012
45	0.0062	0.0062	0.0181	0.0185
42	0.0228	0.0232	0.077	0.080
40	0.050	0.051	0.190	0.200
38	0.100	0.104	0.442	0.483
35	0.245	0.268	1.44	1.81
32	0.493	0.601	4.28	8.78

Table 3 compares the MOL approximation with exact values for the peak delay probability and the peak mean waiting time before beginning service. (Recall that the mean waiting time equals the mean number in queue divided by the number of servers.) We used the peak infinite-server mean $m(t_m) = 29.81$. The PSA approximation $m(t_m) \approx 30.0$ obviously yields similar values, but the difference is noticeable with large numbers of servers; e.g., the 6% difference in offered load produces a 10% error in the delay probability when $s = 45$.

6. Proof of Theorem 1

In order to prove Theorem 1, we apply a previous result in Theorem 10 of Eick et al. [4] and Massey and Whitt [15]. The following weaker form of the previous result will suffice here. Let $S_e^{(n)}$ be a random variable with the n-fold iterated stationary-excess distribution, i.e., $S_e^{(n+1)} = (S_e^{(n)})_e$, where S_e is defined in (2).

Theorem 2. Suppose that λ is $(n+1)$-times differentiable and $\lambda^{(n+1)}$ is Riemann integrable on $[t_\lambda - x, t_\lambda]$ for each x. If $ES^{n+2} < \infty$ and $\lambda^{(k)}(t)$ is bounded on $(-\infty, t_\lambda]$, $0 \leqslant k \leqslant n+1$, then

$$\frac{m_\varepsilon(t)}{ES} \equiv E\big[\lambda_\varepsilon(t - S_e)\big] = \frac{m_n^\varepsilon(t)}{ES} + \frac{R_n^\varepsilon(t)}{ES}, \tag{33}$$

where λ_ε is defined in (8),

$$m_n^\varepsilon(t) = \sum_{k=0}^{n} (-1)^j \frac{\lambda_\varepsilon^{(j)}(t) E(S^{j+1})}{(j+1)!} \tag{34}$$

and

$$R_n^\varepsilon(t) = (-1)^{n+1} E\big[\lambda_\varepsilon^{(n+1)}\big(t - S_e^{(n+2)}\big)\big] \frac{E[S^{n+2}]}{(n+2)!} \tag{35}$$

with $|m_n^\varepsilon(t)| < \infty$ and $|R_n^\varepsilon(t)| < \infty$.

Let $\dot{x}(t)$ denote the derivative of a function x with respect to t.

Corollary. If, in addition to the conditions of Theorem 2, λ is $(n+2)$-times differentiable and $\lambda^{(n+2)}(t)$ is bounded on $(-\infty, t_\lambda)$, then

$$\dot{m}_\varepsilon(t) = \dot{m}_n^\varepsilon(t) + \dot{R}_n^\varepsilon(t) \tag{36}$$

with $|\dot{m}_n^\varepsilon(t)| < \infty$ and $|R_n^\varepsilon(t)| < \infty$ for $m_n^\varepsilon(t)$ in (34) and $R_n^\varepsilon(t)$ in (35).

Proof. By the bounded convergence theorem,

$$\dot{R}_n^\varepsilon(t) = (-1)^{n+1} E\big[\lambda_\varepsilon^{(n+2)}\big(t - S_e^{(n+2)}\big)\big]\frac{E[S^{n+2}]}{(n+2)!}.$$

The bound is obtained from

$$\frac{\lambda_\varepsilon^{(n+1)}(t+\varepsilon) - \lambda_\varepsilon^{(n+1)}(t)}{\varepsilon} = \lambda_\varepsilon^{(n+2)}(\theta_{t,\varepsilon}) \leqslant M,$$

where $t \leqslant \theta_{t,\varepsilon} < t + \varepsilon$ for all t and ε. □

In order to prove Theorem 1, we apply Theorem 2 and its corollary for the special case of $n = 4$. Recall that $ES = 1$. We first obtain

$$m_\varepsilon(t) = \mathsf{E}\big[\lambda_\varepsilon(t - S_e)\big] = \mathsf{E}\big[\lambda\big(t_\lambda + \varepsilon(t - t_\lambda - S_e)\big)\big] \tag{37}$$

$$= \lambda(t_\lambda) + \varepsilon^2\frac{\lambda^{(2)}(t_\lambda)}{2}\mathsf{E}\big[(t - t_\lambda - S_e)^2\big] + \varepsilon^3\frac{\lambda^{(3)}(t_\lambda)}{3!}\mathsf{E}\big[(t - t_\lambda - S_e)^3\big]$$

$$+ \varepsilon^4\frac{\lambda^{(4)}(t_\lambda)}{4!}\mathsf{E}\big[(t - t_\lambda - S_e)^4\big] + \mathrm{O}\big(\varepsilon^5\big). \tag{38}$$

The Corollary to Theorem 2 allows us to differentiate with respect to t in (38) in order to obtain

$$\dot{m}_\varepsilon(t) = \varepsilon^2\lambda^{(2)}(t_\lambda)\mathsf{E}\big[t - t_\lambda - S_e\big] + \varepsilon^3\frac{\lambda^{(3)}(t_\lambda)}{2}\mathsf{E}\big[(t - t_\lambda - S_e)^2\big]$$

$$+ \varepsilon^4\frac{\lambda^{(4)}(t_\lambda)}{6}\mathsf{E}\big[(t - t_\lambda - S_e)^3\big] + \mathrm{O}\big(\varepsilon^5\big). \tag{39}$$

(We use the fact that $\dot{R}_4^\varepsilon(t)$ is also $\mathrm{O}(\varepsilon^5)$.)

From (39), it follows that $m_\varepsilon(t)$ has a unique maximum $\overline{m}(\varepsilon)$ at time $t_m(\varepsilon)$ for all suitably small ε. Considering only the first term of (39), we see that $\dot{m}_\varepsilon(t) > 0$ for $t < t_\lambda + ES_e$ and all ε suitably small, while $\dot{m}_\varepsilon(t) < 0$ for $t > t_\lambda + ES_e$ and all ε suitably small. Now we want to construct an asymptotic expansion for $t_m(\varepsilon)$ of the form

$$t_m(\varepsilon) = \tau_m^{(0)} + \varepsilon\tau_m^{(1)} + \varepsilon^2\tau_m^{(2)} + \mathrm{O}\big(\varepsilon^3\big), \tag{40}$$

where

$$\dot{m}_\varepsilon\big(t_m(\varepsilon)\big) = 0. \tag{41}$$

Combining (39)–(41), we obtain

$$0 = \lambda^{(2)}(t_\lambda)\mathsf{E}\big[t_m(\varepsilon) - t_\lambda - S_e\big] + \varepsilon\frac{\lambda^{(3)}(t_\lambda)}{2}\mathsf{E}\big[\big(t_m(\varepsilon) - t_\lambda - S_e\big)^2\big]$$

$$+ \varepsilon^2\frac{\lambda^{(4)}(t_\lambda)}{6}\mathsf{E}\big[\big(t_m(\varepsilon) - t_\lambda - S_e\big)^3\big] + \mathrm{O}\big(\varepsilon^3\big). \tag{42}$$

If we equate the terms in (42) of order $\varepsilon^0 = 1$, then we obtain

$$\tau_m^{(0)} = t_\lambda + \mathsf{E}[S_e]. \tag{43}$$

If we set $t_m^*(\varepsilon) \equiv t_m(\varepsilon) - \tau_m^{(0)}$, then we get

$$t_m(\varepsilon) - t_\lambda - S_e = t_m^*(\varepsilon) + \mathsf{E}[S_e] - S_e. \tag{44}$$

Hence eq. (42) is equivalent to

$$0 = \lambda^{(2)}(t_\lambda)t_m^*(\varepsilon) + \varepsilon\frac{\lambda^{(3)}(t_\lambda)}{2}\big(t_m^*(\varepsilon)^2 + \mathsf{Var}[S_e]\big)$$

$$+ \varepsilon^2\frac{\lambda^{(4)}(t_\lambda)}{6}\big(t_m^*(\varepsilon)^3 + t_m^*(\varepsilon)\mathsf{Var}[S_e] - \mathsf{E}\big[\big(S_e - \mathsf{E}[S_e]\big)^3\big]\big) + \mathrm{O}\big(\varepsilon^3\big). \tag{45}$$

Now, if we equate like terms in (45) of order ε, we get

$$0 = \lambda^{(2)}(t_\lambda)\tau_m^{(1)} + \frac{\lambda^{(3)}(t_\lambda)}{2}\mathsf{Var}[S_e], \tag{46}$$

which gives us

$$\tau_m^{(1)} = -\frac{\lambda^{(3)}(t_\lambda)}{2\lambda^{(2)}(t_\lambda)}\,\mathsf{Var}[S_e]. \tag{47}$$

Finally, if we equate like terms in (45) of order ε^2, we get

$$0 = \lambda^{(2)}(t_\lambda)\tau_m^{(2)} - \frac{\lambda^{(4)}(t_\lambda)}{6}\mathsf{E}\big[\big(S_e - \mathsf{E}[S_e]\big)^3\big], \tag{48}$$

which yields

$$\tau_m^{(2)} = \frac{\lambda^{(4)}(t_\lambda)}{6\lambda^{(2)}(t_\lambda)}\mathsf{E}\big[\big(S_e - \mathsf{E}[S_e]\big)^3\big]. \tag{49}$$

We obtain the expansion for $m_\varepsilon(t_m(\varepsilon))$, and thus for $D(\varepsilon)$, by applying the asymptotic expansions for $m_\varepsilon(t)$ and $t_m(\varepsilon)$.

References

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, DC, 1972).

[2] C.M. Bender and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw-Hill, New York, 1978).

[3] 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.
[4] S.G. Eick, W.A. Massey and W. Whitt, The physics of the $M_t/G/\infty$ Queue, Oper. Res. 41 (1993) 731–742.
[5] S.G. Eick, W.A. Massey and W. Whitt, $M_t/G/\infty$ queues with sinusoidal arrival rates, Management Sci. 39 (1993) 241–252.
[6] L. Green and P. Kolesar, The pointwise stationary approximation for queues with nonstationary arrivals, Management Sci. 37 (1991) 84–97.
[7] L. Green and P. Kolesar, Simple approximations of peak congestion in $M_t/G/\infty$ queues with sinusoidal arrival rates, Columbia University (1995).
[8] L. Green and P. Kolesar, The lagged PSA for estimating peak congestion in multi-server Markovian queues with periodic arrival rates, Management Sci. 43 (1997) 80–87.
[9] L. Green, P. Kolesar and A. Svoronos, Some effects of nonstationarity on multiserver Markovian queueing systems, Oper. Res. 39 (1991) 502–511.
[10] R.W. Hall, *Queueing Methods for Services and Manufacturing* (Prentice-Hall, Englewood Cliffs, NJ, 1991).
[11] D.L. Jagerman, Nonstationary blocking in telephone traffic, Bell System Tech. J. 54 (1975) 625–661.
[12] O.B. Jennings, A. Mandelbaum, W.A. Massey and W. Whitt, Server staffing to meet time-varying demand, Management Sci. 42 (1996) 1383–1394.
[13] W.A. Massey, Asymptotic analysis of the time dependent $M/M/1$ queue, Math. Oper. Res. 10 (1985) 305–327.
[14] W.A. Massey, G.A. Parker and W. Whitt, Estimating the parameters of a nonhomogeneous Poisson process with linear rate, Telecommunication Systems 5 (1996) 361–388.
[15] W.A. Massey and W. Whitt, A probabilistic generalization of Taylor's theorem, Statist. Probab. Lett. 16 (1993) 51–54.
[16] W.A. Massey and W. Whitt, An analysis of the modified offered load approximation for the nonstationary Erlang loss model, Ann. Appl. Probab. 4 (1994) 1145–1160.
[17] W.A. Massey and W. Whitt, Uniform acceleration expansions for continuous-time Markov chains with time-varying rates, submitted (1996).
[18] P. McCullagh and J.A. Nelder, *Generalized Linear Models* (Chapman and Hall, London, 1983).
[19] G.F. Newell, *Applications of Queueing Theory* (Chapman and Hall, London, 1982).
[20] W. Whitt, The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase, Management Sci. 37 (1991) 307–314.

Queueing Systems 25 (1997) 173–233

Asymptotics for $M/G/1$ low-priority waiting-time tail probabilities

Joseph Abate [a] and Ward Whitt [b]

[a] *900 Hammond Road, Ridgewood, NJ 07450-2908, USA*
[b] *AT&T Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636, USA*
E-mail: wow@research.att.com

Received 7 June 1996; revised 18 October 1996

We consider the classical $M/G/1$ queue with two priority classes and the nonpreemptive and preemptive-resume disciplines. We show that the low-priority steady-state waiting-time can be expressed as a geometric random sum of i.i.d. random variables, just like the $M/G/1$ FIFO waiting-time distribution. We exploit this structures to determine the asymptotic behavior of the tail probabilities. Unlike the FIFO case, there is routinely a region of the parameters such that the tail probabilities have non-exponential asymptotics. This phenomenon even occurs when both service-time distributions are exponential. When non-exponential asymptotics holds, the asymptotic form tends to be determined by the non-exponential asymptotics for the high-priority busy-period distribution. We obtain asymptotic expansions for the low-priority waiting-time distribution by obtaining an asymptotic expansion for the busy-period transform from Kendall's functional equation. We identify the boundary between the exponential and non-exponential asymptotic regions. For the special cases of an exponential high-priority service-time distribution and of common general service-time distributions, we obtain convenient explicit forms for the low-priority waiting-time transform. We also establish asymptotic results for cases with long-tail service-time distributions. As with FIFO, the exponential asymptotics tend to provide excellent approximations, while the non-exponential asymptotics do not, but the asymptotic relations indicate the general form. In all cases, exact results can be obtained by numerically inverting the waiting-time transform.

Keywords: priority queues, $M/G/1$ queue, low-priority waiting time, tail probabilities, asymptotics, non-exponential asymptotics, asymptotic expansions, Laplace transforms, algebraic singularities

1. Introduction

In this paper we study the low-priority steady-state waiting-time distribution in the classical $M/G/1$ queue with two priority classes and the nonpreemptive and preemptive-resume disciplines. The priority structure tends to make the low-priority waiting-time distribution have a relatively long tail. We quantify this effect.

The Laplace transform of the low-priority waiting-time distribution and the first few moments are well known, e.g., see Cohen [30, section III.3.6], Heyman and So-