ELSEVIER

Operations Research Letters 15 (1994) 223–230

operations research letters

A heavy-traffic expansion for asymptotic decay rates of tail probabilities in multichannel queues

Joseph Abate[a], Ward Whitt[b]

[a] *900 Hammond Road, Ridgewood, NJ 07450-2908, USA*
[b] *AT&T Bell Laboratories, Murray Hill, NJ 07974-0636, USA*

(Received 8 October 1992; revised 1 March 1994)

Abstract

We establish a heavy-traffic asymptotic expansion (in powers of one minus the traffic intensity) for the asymptotic decay rates of queue-length and workload tail probabilities in stable infinite-capacity multichannel queues. The specific model has multiple independent heterogeneous servers, each with i.i.d. service times, that are independent of the arrival process, which is the superposition of independent nonidentical renewal processes. Customers are assigned to the first available server in the order of arrival. The heavy-traffic expansion yields relatively simple approximations for the tails of steady-state distributions and higher percentiles, yielding insight into the impact of the first three moments of the defining distributions.

Key words: Queues; Multichannel queues; Waiting-time distribution; Tail probabilities; Asymptotics, Heavy traffic

1. Introduction

Let Q and W be the steady-state queue length (number in system) and workload (remaining service time of all customers in the system) at an arbitrary time in an infinite-capacity queueing model. In surprisingly great generality,

$$\sigma^{-k} P(Q > k) \to \beta \quad \text{as } k \to \infty \tag{1}$$

and

$$e^{\eta x} P(W > x) \to \alpha \quad \text{as } x \to \infty, \tag{2}$$

where σ and η are the *asymptotic decay rates*, and β and α are the *asymptotic constants*. For the GI/G/1 queue, such exponential small-tail asymptotics were established in 1953 by Smith [20]; these results require that the service-time distribution have a finite moment generating function in a neighborhood of the origin. For recent extensions and more references, see the works of Abate et al. [1]–[3], Asmussen and Perry [5], Asmussen and Rolski [6], Elwalid and Mitra [10], [11], Falkenberg [12], Neuts [18], Tijms [21] and Whitt

SSDI 0167-6377(94)00021-W

[23]. Corresponding small-tail asymptotics also typically hold for the steady-state random variables associated with the embedded processes, observing just before arrivals or just after departures, with the same asymptotic decay rates, but different asymptotic constants; see [3].

In even greater generality, we have the weaker limits

$$k^{-1} \log P(Q > k) \to \log \sigma \quad \text{as } k \to \infty \tag{3}$$

and

$$x^{-1} \log P(W > x) \to -\eta \quad \text{as } x \to \infty; \tag{4}$$

see the works of Chang [7] and Glynn and Whitt [13]. The limits (1) and (2) support the approximations

$$P(Q > k) \approx \beta \sigma^k \quad \text{and} \quad P(W > x) \approx \alpha e^{-\eta x}, \tag{5}$$

and the limits (1)–(4) support the cruder approximations

$$P(Q > k) \approx \sigma^k \quad \text{and} \quad P(W > x) \approx e^{-\eta x}. \tag{6}$$

Moreover, these approximations are often surprisingly good; see Tijms [21, Section 1.9, Ch. 4] and Ref. [2]. The cruder approximations in (6) are often good for high percentiles.

In [2] a heavy-traffic expansion was developed for the decay rate η in the GI/GI/1 model, which reveals how η depends on the first three moments of the interarrival-time and service-time distributions when the traffic intensity is not too low. The first term corresponds to the familiar heavy-traffic limit [16, 17]. The second term is especially revealing as a refinement of this heavy-traffic limit. Numerical examples in [1] clearly demonstrate the value of this approach. It is significant that the heavy-traffic refinements for η are much more tractable than the heavy-traffic refinements for the mean EW; i.e., we obtain useful exact formulas for η, whereas this is not possible for the mean [2, 22].

The purpose of this paper is to develop the corresponding heavy-traffic expansions for both η and σ in the general multichannel queue, with *m heterogeneous* servers and the superposition of *n* independent *nonidentical* renewal arrival processes. This multichannel model is of considerable interest to study the effect of statistical multiplexing in communication networks. The asymptotic decay rates play a key role in concepts of effective bandwidth for admission control; see the works of Chang [7], Elwalid and Mitra [10, 11] and Whitt [23]. The way to determine the decay rates for this model was indicated in [23]. Theoretical justification was provided first for the special case of phase-type distributions by Neuts [18] and then for the special case of a single server by Chang [7] and Glynn and Whitt [13]. However, in full generality, the formula for the decay rates remains to be justified.

Interestingly, the GI/GI/1 analysis extends directly to the $\sum_{i=1}^{n} \mathrm{GI}_i/\mathrm{GI}/m$ model with *m identical* servers in parallel and an arrival process that is the superposition of *n* independent and *identical* renewal arrival processes, because the decay rates σ and η are the *same* as for the GI/GI/1 model, after an appropriate adjustment of the time scale. However, this also remains to be fully proved.

Here is how this paper is organized. In Section 2 we review the equations determining the asymptotic decay rates; in Section 3 we establish the heavy-traffic expansion; and in Section 4 we make a few concluding remarks.

Since this paper was written, we have made significant extensions of the results given here and they are in the works of Choudhury and Whitt [9] and Glynn and Whitt [13, 14]. We have also developed numerical algorithms and evaluated the approximations based on the asymptotics. Choudhury et al. [8] show that the quality of the cruder asymptotic approximations in (6) based on the decay rates η and σ alone can deteriorate dramatically when the number of sources gets large.

2. The model and the determining equations

The model is specified by $n + m$ nonnegative mean-one random variables U_i, $1 \leqslant i \leqslant n$, and V_j, $1 \leqslant j \leqslant m$, and $n + m$ time-scaling factors (arrival and service rates) ρw_i, $1 \leqslant i \leqslant n$, and μ_j, $1 \leqslant j \leqslant m$, where it is understood that $\sum_{i=1}^{n} w_i = 1$ and $\sum_{j=1}^{m} \mu_j = 1$; i.e., the total service rate is 1 and the total arrival rate (which equals the traffic intensity) is ρ. The random variable $U_i/\rho w_i$ is a generic *interarrival time* in the ith arrival channel, while the variable V_j/μ_j is a generic *service-time* at the jth server. We assume that customers are assigned to the first available server in order of arrival, with some procedure to break ties. Interestingly, the tie-breaking mechanism does not affect the asymptotic decay rates.

Our asymptotic expansion as $\rho \to 1$ will involve the parameters w_i and μ_j, and the first three moments of U_i and V_j, which we assume are finite. Moreover, we assume that the moment generating functions Ee^{sV_j} are all finite for some positive s. This is a necessary (but not sufficient) condition for (1) and (2); see [2].

Let u_{ik} *and* v_{jk} *be the* kth moments of U_i and V_j, respectively. Let c_{ai}^2 and c_{sj}^2 be the squared coefficients of variation (SCV, variance divided by the square of the mean) of U_i and V_j. By the assumptions above,

$$u_{i1} = v_{j1} = 1, \qquad u_{i2} - 1 = c_{ai}^2 \quad \text{and} \quad v_{j2} - 1 = c_{sj}^2.$$

Note that u_{i3} and v_{j3} are the third moments of U_i and V_j instead of $U_i/\rho w_i$ and V_j/μ_j.

We now describe the equations that we conjecture determine the asymptotic decay rates. (As indicated in the introduction, the references explain where these equations come from and provide strong supporting evidence for their validity.) The geometric decay rate σ in (1) is the unique root z in (0, 1) of the *decay-rate equation*:

$$\sum_{i=1}^{n} a_i(z) = \sum_{j=1}^{m} s_j(z), \tag{7}$$

where $a_i(z) \equiv a_i(z, \rho)$ and $s_j(z)$ are the functions of z that are roots of the *individual congestion equations*:

$$Ee^{-a_i(z)U_i/\rho w_i} = z, \quad 1 \leqslant i \leqslant n, \tag{8}$$

and

$$Ee^{s_j(z)V_j/\mu_j} = \frac{1}{z}, \quad 1 \leqslant j \leqslant m. \tag{9}$$

Then the exponential decay rate $\eta \equiv \eta(\rho)$ in (2) is simply

$$\eta = \sum_{i=1}^{n} a_i(\sigma) = \sum_{j=1}^{m} s_j(\sigma). \tag{10}$$

3. The heavy-traffic expansion

In this section we show how to obtain heavy-traffic asymptotic expansions for the decay rates σ and η in (1) and (2) as $\rho \to 1$ from below, using a variant of the method used for the GI/GI/1 queue in [2]. For the GI/GI/1 queue, this idea goes back to Smith [20, p. 461], but he considers only the first term. Even the first term is useful here, because it provides additional support for claims in Section 2. The first term yields the exact decay rate of the limiting exponential in the heavy-traffic limit given in [16–17], as we will now demonstrate.

To do the asymptotics for the decay rates, it is convenient to work with the cumulant generating functions, which are the logarithms of the moment generating functions [15, p. 64]. The specifying equations (8)

and (9) become

$$\log Ee^{-sU_i/\rho w_i} = \log z, \quad 1 \leqslant i \leqslant n, \tag{11}$$

and

$$\log Ee^{sV_j/\mu_j} = -\log z, \quad 1 \leqslant j \leqslant m. \tag{12}$$

The coefficients of $s^n/n!$ in the power series expansion of these cumulant generating functions are the cumulants of $-U_i/w_i\rho$ and V_j/μ_j, respectively. The first three cumulants of U_i are $u_{i1} = 1$, $u_{i2} = u_{i1}^2 = u_{i2} - 1 \equiv c_{ai}^2$ and $u_{i3} - 3u_{i1}u_{i2} + 2u_{i1}^3 = u_{i3} - 3u_{i2} + 2$. We will express the cumulants via the moments below. The key parameters are

$$c_a^2 = \sum_{i=1}^{n} w_i c_{ai}^2 \quad \text{where } c_{ai}^2 = u_{i2} - 1, \tag{13}$$

$$c_s^2 = \sum_{j=1}^{m} \mu_j c_{sj}^2 \quad \text{where } c_{sj}^2 = v_{j2} - 1, \tag{14}$$

$$d_a = \sum_{i=1}^{n} w_i \frac{(u_{i3} - 3c_{ai}^2(c_{ai}^2 + 1) - 1)}{6}, \tag{15}$$

$$d_s = \sum_{j=1} \mu_j \frac{(v_{j3} - 3c_{sj}^2(c_{sj}^2 + 1) - 1)}{6}. \tag{16}$$

Theorem. *Assuming that the decay-rate equation* (7) *has a unique root in* (0, 1) *and the cumulant generating function* (11) *and* (12) *admit partial asymptotic expansions in powers of s, we obtain the following asymptotic expansions for the asymptotic decay rates* σ *and* η *in* (1) *and* (2):

$$\sigma = 1 - \frac{2(1-\rho)}{c_a^2 + c_s^2} + \left[\frac{8(d_s - d_a)}{(c_a^2 + c_s^2)^3} - \frac{2(c_a^2 - 1)}{(c_a^2 + c_s^2)^2}\right](1-\rho)^2 + O((1-\rho)^3) \quad \text{as } \rho \to 1 \tag{17}$$

and

$$\begin{aligned}\eta &= \frac{2(1-\rho)}{c_a^2 + c_s^2} - \left[\frac{8(d_s - d_a)}{(c_a^2 + c_s^2)^3} - \frac{2(c_a^2 - c_s^2)}{(c_a^2 + c_s^2)^2}\right](1-\rho)^2 + O((1-\rho)^3) \\ &= \frac{2(1-\rho)}{c_a^2 + c_s^2}(1 - X(1-\rho) + O((1-\rho)^2) \quad \text{as } \rho \to 1,\end{aligned} \tag{18}$$

where

$$X = \frac{4(d_s - d_a)}{(c_a^2 + c_s^2)^2} + \frac{(c_s^2 - c_a^2)}{c_a^2 + c_s^2}. \tag{19}$$

Proof. We start by expanding each cumulant generating function in powers of s. We keep only three terms, but it is easy to go further, provided higher moments are finite. We obtain the normalized cumulants (expressed via moments) as coefficients; i.e.,

$$\frac{-s}{\rho w_i} + \frac{s^2(u_{i2} - 1)}{2\rho^2 w_i^2} - \frac{s^3(u_{i3} - 3u_{i2} + 2)}{6\rho^3 w_i^3} + O(s^4) = \log z \tag{20}$$

and

$$\frac{s}{\mu_j} + \frac{s^2(v_{j2} - 1)}{2\mu_j^2} + \frac{s^3(v_{j3} - 3v_{j2} + 2)}{6\mu_j^3} + O(s^4) = -\log z \tag{21}$$

as $s \to 0$. We then use the inverse function theorem with (11) and (12), as with reversion of power series (matched power series) [4, p. 16], to write $s \equiv \tilde{a}_i(z)$ and $s \equiv \tilde{s}(z)$ as functions of $\varepsilon = -\log z$ where ($\tilde{a}_i$ and $\tilde{s}_j$ are the values of s as functions of $\log z$ and $-\log z$, respectively, yielding equality in (11) and (12)), i.e., $\tilde{a}_i(-\log z) = a_i(z)$ and $s_j(-\log z) = s_j(z)$, obtaining

$$\tilde{a}_i(\varepsilon) = \rho w_i \varepsilon + \frac{\rho w_i (u_{i2} - 1)\varepsilon^2}{2} + \rho w_i \left[\frac{(u_{i2} - 1)^2}{2} - \frac{(u_{i3} - 3u_{i2} + 2)}{6}\right]\varepsilon^3 + O(\varepsilon^4) \tag{22}$$

and

$$\tilde{s}_j(\varepsilon) = \mu_j \varepsilon - \frac{\mu_j (v_{j2} - 1)}{2}\varepsilon^2 + \mu_j \left[\frac{(v_{j2} - 1)^2}{2} - \frac{(v_{j3} - 3v_{j2} + 2)}{6}\right]\varepsilon^3 + O(\varepsilon^4) \tag{23}$$

as $\varepsilon \equiv \log z \to 0$.

Now we can apply (7) to obtain the asymptotic form of the characterizing equation

$$\delta(\varepsilon) \equiv \sum_{j=1}^{m} \tilde{s}_j(\varepsilon) - \sum_{i=1}^{n} \tilde{a}_i(\varepsilon) = 0. \tag{24}$$

Substituting (22), (23) and (13)–(16) into (24), we obtain

$$\frac{(c_s^2 + \rho c_a^2)\varepsilon}{2} + (d_s - \rho d_a)\varepsilon^2 + O(\varepsilon^3) = 1 - \rho. \tag{25}$$

We now intend to apply the inverse function theorem again to (25) to represent ε as a partial power series in $(1 - \rho)$. First, however, we eliminate ρ from the coefficients of ε^k in the left-hand side of (25). We write (25) as

$$A\varepsilon + B\varepsilon\omega + C\varepsilon^2 + D\varepsilon^2\omega + O(\varepsilon^3) = \omega \tag{26}$$

where $\omega = 1 - \rho$, $A = (c_s^2 + c_a^2)/2$, $B = -c_a^2/2$ and $C = (d_s - d_a)$. Then, substituting expression (26) for ω in each appearance of ω in the left-hand side of (26), we get

$$A\varepsilon + (C + AB)\varepsilon^2 + O(\varepsilon^3) = \omega. \tag{27}$$

Hence, by the inverse function theorem, we get

$$\varepsilon = \frac{2(1 - \rho)}{c_s^2 + c_a^2} - \left[\frac{8(d_s - d_a)}{(c_s^2 + c_a^2)^3} - \frac{2c_a^2}{(c_s^2 + c_a^2)^2}\right](1 - \rho)^2 + O((1 - \rho)^3). \tag{28}$$

Finally, since $\varepsilon = -\log z$, the geometric decay rate in (1) is

$$\sigma = e^{-\varepsilon} = 1 - \varepsilon + \frac{\varepsilon^2}{2} + o(\varepsilon^2) \quad \text{as } \varepsilon \to 0, \tag{29}$$

which is equivalent to (17).

Next, since $\eta = \psi(\sigma) = \tilde{\psi}(-\log\sigma)$, where $-\log\sigma = \varepsilon$, $\psi = \sum_{j=1}^{m} s_j$ and $\tilde{\psi} = \sum_{j=1}^{m} \tilde{s}_j$, by (10), we can apply (23) to obtain

$$\eta = \varepsilon - \frac{c_s^2}{2}\varepsilon^2 + O(\varepsilon^3) \quad \text{as } \varepsilon \to 0. \tag{30}$$

Then we can apply (25) to obtain (18). □

The first-order approximations $\sigma \approx 1 - \eta$ and $\eta \approx 2(1-\rho)/(c_a^2 + c_s^2)$ in (17) and (18) are the familiar heavy-traffic limits [16, 17]. Formulas (17)–(19) agree with previously derived formulas for the GI/GI/1 queue. For example, the correction factor X in (19) coincides with the previously derived GI/GI/1 result in the single-channel case [2]: For the GI/GI/1 model,

$$X = \frac{2(v_3 - 3c_s^2(c_s^2+1) - u_3 + 3c_s^2(c_s^2+1)) + 3((c_s^2)^2 - (c_a^2)^2)}{3(c_a^2 + c_s^2)^2}. \tag{31}$$

Note that $X = 0$ when $u_3 = v_3$ and $c_a^2 = c_s^2$, and thus in the M/M/1 special case.

Formulas (17)–(19) are especially important to develop insight into the way the decay rates depend on the underlying parameters. First, the arrival process beyond its rate affects approximations (17) and (18) solely via the parameters c_a^2 *and* d_a, while the service process beyond its rate affects these approximations solely via the parameters c_s^2 and d_s. From (18) (regarding the terms lexicographically), we see that η is decreasing (more congestion) in the arrival rate λ and the "second-moment" terms c_a^2 and c_s^2 in (13) and (14), as we expect. The decay rate η is also decreasing in the service "third-moment" parameter d_s in (16), but *increasing* in the arrival "third-moment" parameter d_a in (15). This last property can be understood by considering the fact that, for given first two moments, a high third moment usually means more mass near the origin, and thus more short interarrival times.

We can see the effect of the degree of heterogeneity in the way the four parameters c_a^2, d_a, c_s^2, d_s in expansions (17) and (18) are determined. From (13)–(16), we see that these parameters are all convex combinations of the corresponding single-source parameters, with the weights depending on the rate of the source.

4. Concluding remarks

The analysis in this paper extends to other single-channel processes. For example, if the ith single-channel arrival process is a Markov renewal process (MRP) as in Neuts [18] and in Section III.B of [23], then we replace Ee^{-sU_i} in (8) by $\hat{f}_i(s)$ where $\hat{f}_i(s)$ is the Perron–Frobenius eigenvalue of the matrix of Laplace transforms of the MRP. Instead of (16), we expand the function $\log \hat{f}_i(s/\rho w_i)$ about $s = 0$; see the works of Choudhury and Whitt [9] and Glynn and Whitt [13, 14]. Key structural properties of $\hat{f}_i(s)$, including its derivatives, are given in the Appendix of Neuts [19]. The coefficient $(u_{i2} - 1)$ in (20) should be replaced by the asymptotic variance constant of the associated rate-1 MRP, i.e., the limit of the index of dispersion for counts (IDC), i.e., $I_c(\infty)$ where

$$I_{ci}(t) = \frac{\operatorname{Var} A_i(t)}{EA_i(t)}, \quad t > 0, \tag{32}$$

and $A_i(t)$ counts the number of arrivals in the interval $[0; t]$; i.e., it is the heavy-traffic limit.

Similarly, if the jth single-channel service process is a Markov renewal process, then we replace Ee^{-sV_j} in (9) by $\hat{g}_j(-s)$ where $\hat{g}_j(s)$ is the Perron–Frobenius eigenvalue of the matrix of Laplace transforms of the Markov

renewal process. Instead of (21), we expand $\log \hat{g}_j(-s/\mu_j)$ about $s = 0$. The coefficient $(v_{j2} - 1)$ in (21) should be replaced by the asymptotic variance constant of the associated rate-1 MRP, i.e., $I_{cj}(\infty)$ where

$$I_{cj}(t) = \frac{\operatorname{Var} S_j(t)}{ES_j(t)}, \quad t > 0, \tag{33}$$

and $S_j(t)$ counts the number of service completions during the first t units of time that the server is busy.

In this paper we have considered only the asymptotic decay rates σ and η in (1) and (2). In [1–3] we discuss approximations and exact results for the asymptotic constants α and β in (1) and (2).

The heavy-traffic asymptotic expansion is a convenient approximation, which is useful to provide insight into the way the tail probabilities depend on the basic model data. It is known that asymptotic decay rates and asymptotic constants can actually be computed directly from transforms for a large class of models. For example, suppose that we have the Laplace–Stieltjes transform of the waiting time, Ee^{-sW}. Then the Laplace transform of $P(W > x)$ is

$$\hat{W}^c(s) \equiv \int_0^\infty e^{-sx} P(W > x)\,dx = \frac{1 - Ee^{-sW}}{s}. \tag{34}$$

Then $-\eta$ is the right-most singularity of $\hat{W}^c(s)$.

Having found η, we can find the asymptotic constant α in (2) by applying the final-value theorem for Laplace transforms, i.e.,

$$\alpha \equiv \lim_{x \to \infty} e^{\eta x} P(W > x) = \lim_{s \to -\eta} (s + \eta)\hat{W}^c(s + \eta). \tag{35}$$

To justify (35), we assume that (2) is valid, so that indeed the right-most singularity of $\hat{W}^c(s)$ is a simple pole. Furthermore, if we can represent $\hat{W}^c(s)$ as $N(s)/D(s)$ where $-\eta$ is a root of the equation $D(s) = 0$, then we can also express α as

$$\alpha = N(-\eta)/D'(-\eta). \tag{36}$$

where D' is the derivative of D.

Acknowledgment

This work is an outgrowth of joint work with Gagan Choudhury in [2]. We thank him for his help.

References

[1] J. Abate, G.L. Choudhury and W. Whitt, "Asymptotics for steady-state tail probabilities in structured Markov queueing models", *Stochastic Models* 10, 99–143 (1994).

[2] J. Abate, G.L. Choudhury and W. Whitt, "Exponential approximations for tail probabilities in queues, I: waiting times", *Oper. Res.*, to appear.

[3] J. Abate, G.L. Choudhury and W. Whitt, "Exponential approximations for tail probabilities in queues, II: sojourn time and workload", *Oper. Res.*, to appear.

[4] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C., 1972.

[5] S. Asmussen and D. Perry, "On cycle maxima, first passage problems and extreme value theory for queues", *Stochastic Models* 8, 421–459 (1992).

[6] S. Asmussen and T. Rolski, "Risk theory in a periodic environment: the Cramér–Lundberg approximation and Lundberg's inequality", *Math. Oper. Res.* 19, 410–433 (1994).

[7] C.S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks", IEEE Trans. Autom. Control 39, 913–931 (1994).
[8] G.L. Choudhury, D.M. Lucantoni and W. Whitt, "Squeezing the most out of ATM", submitted, 1993.
[9] G.L. Choudhury and W. Whitt, "Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 queue", *Stochastic Models* 10, 453–498 (1994).
[10] A.I. Elwalid and D. Mitra, "Effective bandwidths of general Markovian traffic sources and admission control of high speed networks", *ACM/IEEE Trans. Networks* 1, 329–343 (1993).
[11] A.I. Elwalid and D. Mitra, "Markovian arrival and service communication systems: Spectral expansions, separability and Kronecker-product forms", submitted, 1993.
[12] E. Falkenberg, "On the asymptotic behavior of the stationary distribution of Markov chains of M/G/1-type", *Stochastic Models* 10, 75–98 (1994).
[13] P.W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue", *Stud. Appl. Probab.*, 1994, to appear.
[14] P.W. Glynn and W. Whitt, "Large deviations behavior of counting processes and their inverses", *Queueing Systems*, to appear.
[15] B.V. Gnedenko and A.N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Reading, MA, 1968.
[16] D.L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, I", *Adv. in Appl. Probab.* 2, 150–177 (1970).
[17] D.L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, II: sequences, networks and batches", *Adv. in Appl. Probab.* 2, 355–369 (1970).
[18] M.F. Neuts, "The caudal characteristic curve of queues", *Adv. in Appl. Probab.* 18, 221–254 (1986).
[19] M.F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York, 1989.
[20] W. Smith, "On the distribution of queueing times", *Proc. Camb. Phil. Soc.* 49, 449–461 (1953).
[21] H.C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach*, Wiley, New York, 1986.
[22] W. Whitt, "An interpolation approximation for the mean workload in a GI/GI/1 queue", *Oper. Res.* 37, 936–952 (1989).
[23] W. Whitt, "Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues", *Telecommunication Systems* 2, 71–107 (1993).