Queueing Systems 39, 55–78, 2001

The Remaining Service Time upon Reaching a High Level in $M/G/1$ Queues

PIETER-TJERK DE BOER * ptdeboer@cs.utwente.nl
Telematics Systems and Services, Department of Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

VICTOR F. NICOLA nicola@cs.utwente.nl
Telematics Systems and Services, Department of Electrical Engineering, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

JAN-KEES C.W. VAN OMMEREN J.C.W.vanOmmeren@math.utwente.nl
Faculty of Mathematical Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Received 15 July 1998; Revised 26 April 2001

Abstract. The distribution of the remaining service time upon reaching some target level in an $M/G/1$ queue is of theoretical as well as practical interest. In general, this distribution depends on the initial level as well as on the target level, say, B. Two initial levels are of particular interest, namely, level "1" (i.e., upon arrival to an empty system) and level "$B-1$" (i.e., upon departure at the target level).

In this paper, we consider a busy cycle and show that the remaining service time distribution, upon reaching a high level B due to an arrival, converges to a limiting distribution for $B \to \infty$. We determine this asymptotic distribution upon the "first hit" (i.e., starting with an arrival to an empty system) and upon "subsequent hits" (i.e., starting with a departure at the target) into a high target level B. The form of the limiting (asymptotic) distribution of the remaining service time depends on whether the system is stable or not. The asymptotic analysis in this paper also enables us to obtain good analytical approximations of interesting quantities associated with rare events, such as overflow probabilities.

Keywords: remaining service time, asymptotics, $M/G/1$ queue, level crossing

AMS subject classification: primary 60K25, secondary 68M20, 90B22

1. Introduction

In this paper, we study the distribution of the remaining service time upon reaching a high level (typically corresponding to full buffer) due to a customer arrival in an $M/G/1$ queueing system. This problem was originally motivated by research on the efficient simulation of consecutive loss in such queueing systems [3]; other applications include an approximate calculation of consecutive-loss probabilities (see section 8.3), and improving RESTART simulation (see [6]).

* Corresponding author.

Consider an $M/G/1/B$ queue without service interruptions. Initially, assume that it is empty, i.e., there are neither customers waiting nor in service. After some time, the queue may become full, i.e., there are a total of B customers in it, one of which is being served. We are interested in the distribution of the remaining service time of the customer being served at the moment full buffer is reached. After the full-buffer state is left, the queue will sooner or later either become empty (marking the end of the busy cycle), or reach full buffer again during the same busy cycle; more full-buffer periods may follow in the same busy cycle. Because of the memoryless arrival process, the second and later full-buffer hits are stochastically equivalent, so we will refer to them as *subsequent hits* in this paper. The first full-buffer hit in a busy cycle is in general different from subsequent full-buffer hits, and will be referred to as the *first hit*.

A huge amount of literature exists on the study of the single-server queue with all its variants; however, little is related to this problem. The closest we found was a discussion of the distribution of idle periods in a stable $GI/M/1$ queue in [4, chapter II.5.10]. The stable $GI/M/1$ queue is the dual of the unstable $M/G/1$ queue, so those idle periods correspond to the remaining service times for "subsequent" hits to full buffer in an unstable $M/G/1$ queue. Our analysis is more comprehensive, as it treats the stable as well as the unstable $M/G/1$ queue, and also the "first" as well as "subsequent" hits. In [4, chapter III.6.3], there is a discussion of a related subject: the stationary joint distribution of the number of customers and the past service time in an $M/G/1$ queue. In [1] the equilibrium distributions of the past and remaining service times upon arrival to a given level in an $M/G/1$ queue are calculated; equilibrium here implies that no distinction between first and subsequent hits is made: they are "mixed" according to the frequency with which they occur. Finally, in [5] the expected value of the remaining service time upon arrival to a given level in $G/G/1$ queues is studied.

We start by introducing some notation in section 2. Next, we derive some results for a hypothetical "doubly-unbounded" $M/G/1$ queue in section 3. These results are used in section 4 to find approximate results (accurate for large B) for the real bounded $M/G/1/B$ queue. Those results allow us to calculate the distributions of past and remaining service times in section 5. However, this analysis does not hold for systems where the average service time equals the average inter-arrival time; to derive results for this case, we use a limit procedure in section 6. As a by-product of the analysis in this paper, we can also obtain an (asymptotically tight) approximation for the probability of reaching full buffer in a busy cycle, as demonstrated in section 7. Section 8 illustrates the accuracy of our results by comparing them with results from exact numerical analysis and simulation. We present a summary of the results together with conclusions in section 9. Note: the results in this paper are only valid if a technical condition is satisfied (see (4)); this excludes cases where the service time distribution has a heavy tail.

2. Notation

Throughout this paper, we will use some notational conventions which are introduced here. First, three generic random variables are defined:

- X is the (total) service time,
- Y is past service time upon hitting full buffer,
- Z is the remaining service time upon hitting full buffer.

Note that the distributions of Y and Z can be defective (in cases where there is a nonzero probability that full buffer is not reached in a given busy cycle); the defect will be represented by a probability mass at $+\infty$. We will also consider the distributions of Y and Z conditional on reaching full buffer, and denote these by $Y|\text{fb}$ and $Z|\text{fb}$, respectively; these are nondefective, of course.

For probability distributions the following notation is used, using the random variable W as an example:

- $f_W(\cdot)$ is the probability density function,
- $F_W(\cdot)$ is the distribution function,
- $\overline{F}_W(\cdot)$ is the complementary distribution function: $\overline{F}_W(w) = 1 - F_W(w)$,
- $\widetilde{F}_W(\cdot)$ is the Laplace–Stieltjes transform of $F_W(\cdot)$: $\widetilde{F}_W(s) = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}F_W(t)$.

We use the symbol $\mathbb{P}(E)$ to denote the probability of the event E, and $\mathbb{E}W$ to denote the expected value of a random variable W.

The arrival process is Poisson; its arrival rate is denoted by λ, and the system load is denoted by ρ, with $\rho = \lambda\mathbb{E}(X)$.

Finally, the symbol for approximate equality ($\approx$) in this paper is understood to imply equality in the limit of infinite buffer size B; i.e., the limit for $B \to \infty$ (sometimes $i \to \infty$) of the quotient of the left-hand side and the right-hand side is 1.

3. The doubly-unbounded $M/G/1$ queue

In this section, we study the "doubly-unbounded" $M/G/1$ queue. This hypothetical system is identical to the usual $M/G/1$ queue with infinite buffer, except for one detail: if the buffer becomes empty, the service process continues, so the buffer content (number of customers in the system) can become *negative*; in fact, we allow it to become infinitely negative. Of course, this has no physical interpretation, but it is useful as a step towards studying the bounded $M/G/1/B$ system in the next section.

For this doubly-unbounded queue, we consider the state of the system at the beginning of service epochs. Because of the memoryless arrival process, these instants together form an *(embedded) Markov chain*. Let N_n (with $-\infty < N_n < \infty$) denote the buffer content at the nth embedded point, i.e., at the beginning of the nth service period ($n \geqslant 1$).

We define q_j as the probability of exactly j arrivals during one service interval. Therefore,

$$q_j = \int_0^\infty \frac{(\lambda s)^j}{j!} \, \mathrm{e}^{-\lambda s} \, \mathrm{d}F_X(s).$$

Next, we define $r_i^{(n)}$ as the probability that $N_n = i$, assuming that the first service starts in state 0. Furthermore, we define r_i as the expected number of times the Markov chain visits state i. Clearly,

$$r_i = \sum_{n=1}^{\infty} r_i^{(n)}.$$

Note that although r_i is the expected number of visits during an infinite interval, it is (in general) a finite number, because the system will eventually drift to either $-\infty$ (if $\rho < 1$) or $+\infty$ (if $\rho > 1$).

It is easily seen that $r_i^{(n)}$ must satisfy the following recursion for $n > 1$:

$$r_i^{(n)} = \sum_{j=0}^{\infty} q_j r_{i-j+1}^{(n-1)}, \tag{1}$$

with boundary condition at $n = 1$:

$$r_i^{(1)} = I_{i=0} \stackrel{\text{def.}}{=} \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

because we start at level 0. Suppose one were to start in state m instead of 0, which corresponds to replacing $I_{i=0}$ by $I_{i=m}$ in the above equation. Since the system is doubly-unbounded, the resulting solution r_i' would just be a translated copy of the original solution, i.e., $r_i' = r_{i-m}$.

We now define $V(z)$ as the z-transform of q_j; it can be expressed in terms of the Laplace–Stieltjes transform $\widetilde{F}_X(\cdot)$ as follows:

$$V(z) = \sum_{j=0}^{\infty} z^j q_j = \widetilde{F}_X(\lambda - \lambda z). \tag{3}$$

Below, we will also need the solutions of the equation

$$V(K) = K. \tag{4}$$

It is easily seen that $V(z)$ is a convex function, that $V(1) = 1$ (so 1 is a solution to (4)), and that $V'(1) = \rho$. Because of these facts, (4) can have at most one other solution, which must be greater than 1 if $\rho < 1$, and less than 1 if $\rho > 1$. For our analysis, we

Table 1
Properties of K_1 and K_2, solutions of $V(K) = K$.

Case	K_1	$V'(K_1)$	K_2	$V'(K_2)$
$\rho < 1$	1	ρ	$K > 1$	> 1
$\rho > 1$	$K < 1$	< 1	1	ρ

assume that this second solution of (4) does indeed exist,[1] and we denote it by K ($\neq 1$). Again because of convexity, $V'(K)$ must be greater than 1 if $\rho < 1$, and less than 1 if $\rho > 1$. We denote the two solutions of (4) by K_1 and K_2, where $0 < K_1 < K_2$; table 1 summarizes their properties.

In appendix A we prove the following theorem:

Theorem 1. Given the recursion (1) with initial condition (2), the sum $r_i = \sum_{n=1}^{\infty} r_i^{(n)}$ (which can be interpreted as the expected number of visits to state i of the embedded Markov chain of the doubly-unbounded system) has the following properties:

$$r_i = \frac{K_1^{-i}}{1 - V'(K_1)} \quad \text{for } i \leqslant 0, \tag{5}$$

and

$$\lim_{i \to \infty} K_2^i r_i = \frac{1}{V'(K_2) - 1}, \tag{6}$$

where $0 < K_1 < K_2$ are the two solutions of $V(K) = K$.

Note that (6) can also be written as

$$r_i \approx \frac{K_2^{-i}}{V'(K_2) - 1} \quad \text{for } i \gg 0. \tag{7}$$

4. The bounded $M/G/1$ queue

Let us now turn to the "real" system, the bounded $M/G/1/B$ queue. Because of the Poisson arrival process, we can again define an embedded Markov chain with embedding points at the beginning of service epochs. At those embedded points, the state variable of interest is the number of customers in the system. Starting in state A (i.e., with A customers in the system, and at the beginning of a service period), we study the evolution of the embedded Markov chain until absorption, which happens in either of two ways:

[1] For $\rho < 1$, such a solution $K > 1$ obviously only exists if the Laplace transform $\widetilde{F}_X(\cdot)$ exists for negative values of its argument. If the tail of the probability distribution of the service time X decays less than exponentially fast, this is a problem. In particular, K does not exist for distributions with a heavy tail, so the results in this paper are not applicable in such cases.

- Full buffer: if during one service, so many arrivals occur that there would be B or more customers in the system just before the completion of this service, full buffer is reached.
- Empty system: if there is only one customer left in the system, and during his service no others arrive, the system would be empty at the completion of this service.

We will now proceed to determine the expected number of times E_i the embedded Markov chain visits state i, starting from level A and ending in one of the above two absorbing states.

In the previous section, we have determined r_i for the doubly-unbounded queue starting in state 0. Those results can be used to obtain E_i as follows: if one would just use the r_i results (shifted by A, to accommodate the fact that we start in state A instead of 0), one would overestimate E_i. In order to compensate for this error, we compare the expected number of times state i is visited in the bounded system (E_i) and in the unbounded system ($E'_i = r_{i-A}$):

- First, we have a contribution which is the same for both E_i and E'_i, corresponding to the evolution up to absorption (i.e., full buffer or empty system).
- Second, if the systems reach level 0 before level B, the bounded system stops (empty system), whereas the unbounded system continues, giving some additional contribution to E'_i. This is exactly as large as the contribution that would be produced by starting from state 0, which we know is given by r_i. In order to cancel this contribution, we need to determine the probability α that the unbounded system indeed reaches level 0 before level B. Then the correction term for E_i is clearly given by $-\alpha r_i$.
- Third, if the systems reach level B before level 0, the bounded system stops (full buffer), whereas the unbounded system continues, giving an additional contribution to E'_i (for $i < B$) only if it down-crosses into state $B-1$ later on. This contribution is exactly as large as the contribution that would be produced by starting from state $B-1$, which is given by r_{i-B+1}. In order to cancel this contribution, we need to determine the probability β that the unbounded system down-crosses into level $B-1$ before reaching level 0. Then the correction term is given by $-\beta r_{i-B+1}$.
  Note that if $\rho > 1$, the system may not return to level $B-1$ after having passed level B; in this case β is not equal to $1-\alpha$. On the other hand, if $\rho < 1$, then $\beta = 1-\alpha$.

Figure 1 shows four typical sample paths of the number of customers in the buffer as a function of time in the unbounded system. The filled circles represent the embedding points of the embedded Markov chain. The lines at levels 0 and B represent the absorption of the bounded system at empty system and full buffer, respectively. The dotted parts of the paths are the parts that must be compensated for by the above procedure. Note the difference between what happens to paths that reach level B and to paths that reach level 0. In the former case, compensation is necessary only if the buffer content returns to level $B-1$ (which may never happen if the arrival rate is higher than the service rate, i.e., $\rho > 1$).

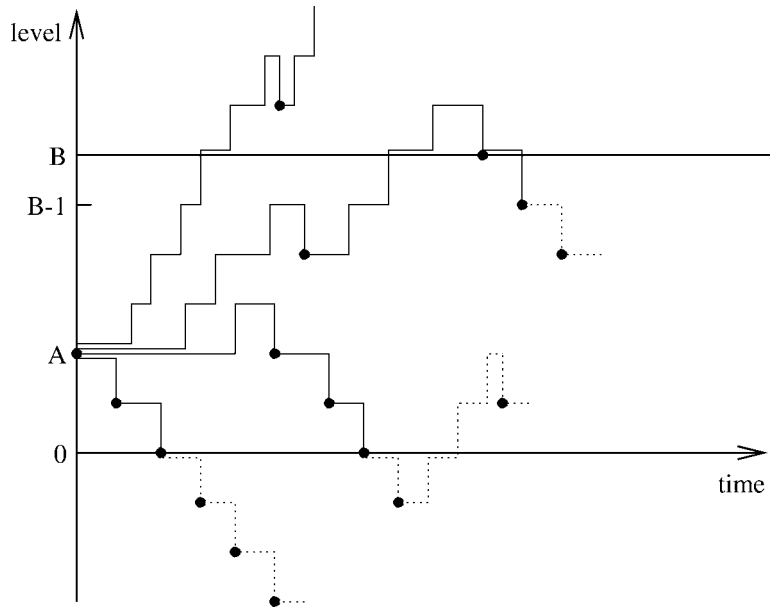

Figure 1. Illustration of typical sample paths in the doubly-unbounded queue. In the bounded queue, the dotted parts of the sample paths must be cancelled.

From the above, it follows that E_i is given by r_{i-A} (that is, the expected number of visits to level i starting from level A in the doubly-unbounded system), minus the contribution due to sample paths beyond absorption at level 0 (i.e., αr_i) and level B (i.e., βr_{i-B+1}):

$$E_i = r_{i-A} - \alpha r_i - \beta r_{i-B+1}. \tag{8}$$

The value of the starting level A is determined by whether first or subsequent hits are being considered. The values of α and β can be determined by applying the appropriate boundary conditions, as will be shown in the sequel.

4.1. First hit: $A = 1$

In the case of first hit, we start in state 1, thus $A = 1$. In the bounded system, the embedded Markov chain cannot reach state 0 or $B - 1$ because of absorption (which would occur before or upon entering either state). As we also do not start in either of these states, we know that $E_0 = 0$ and $E_{B-1} = 0$. By inserting this into (8), we find the conditions for α and β:

$$0 = E_0 = r_{-1} - \alpha r_0 - \beta r_{-B+1} \tag{9}$$

and

$$0 = E_{B-1} = r_{B-2} - \alpha r_{B-1} - \beta r_0. \tag{10}$$

Substituting for r_i from (5) and (7) we get the following two equations:

$$K_1 - \alpha - \beta K_1^{B-1} = 0$$

and

$$\frac{K_2^{-(B-1)}(K_2 - \alpha)}{V'(K_2) - 1} \approx \frac{\beta}{1 - V'(K_1)}.$$

By substituting from (5) and (7) into (8), and then using the above equation to eliminate β, we can write E_i for $1 \ll i \leqslant B$ as follows:

$$\begin{aligned} E_i &\approx \frac{K_2^{-i+1}}{V'(K_2) - 1} - \alpha \frac{K_2^{-i}}{V'(K_2) - 1} - \beta \frac{K_1^{-i+B-1}}{1 - V'(K_1)} \\ &\approx \frac{K_2^{-i}(K_2 - \alpha)}{V'(K_2) - 1} - \frac{K_1^{-i+B-1} K_2^{-(B-1)}(K_2 - \alpha)}{V'(K_2) - 1} \\ &= \frac{K_2^{-(B-1)}(K_2 - \alpha)}{V'(K_2) - 1}\big(K_2^{B-i-1} - K_1^{B-i-1}\big). \end{aligned}$$

Writing this properly as a limit gives us:

Theorem 2. The expected number of visits E_{B-j} to state $B - j$ of the $M/G/1/B$ embedded Markov chain, starting from state 1, has the following asymptotic behaviour for large B:

$$\lim_{B\to\infty} \frac{E_{B-j}}{K_2^{-(B-1)}(K_2^{j-1} - K_1^{j-1})} = \frac{K_2 - \alpha}{V'(K_2) - 1}. \tag{11}$$

For the moment, we do not need the value of α and defer its calculation to section 7.

4.2. Subsequent hits: $A = B - 1$

In the case of subsequent hits, we start in state $B - 1$, thus $A = B - 1$. In the bounded system, the embedded Markov chain cannot reach this state again (because of absorption), so the total number of visits to this state must be 1, i.e., $E_{B-1} = 1$. Furthermore, since state 0 is an absorbing state, it is considered unreachable, so $E_0 = 0$. By inserting this into (8), we find the equations for determining α and β:

$$1 = E_{B-1} = r_0 - \alpha r_{B-1} - \beta r_0$$

and

$$0 = E_0 = r_{-B+1} - \alpha r_0 - \beta r_{-B+1}.$$

Substituting for r_i from (5) and (7) in the above, we get

$$-\alpha \frac{K_2^{-(B-1)}}{V'(K_2) - 1} + (1 - \beta) \frac{1}{1 - V'(K_1)} \approx 1$$

and

$$-\alpha + (1-\beta)K_1^{B-1} = 0.$$

Solving for α and β yields

$$\alpha = (1-\beta)K_1^{B-1} \tag{12}$$

and

$$\frac{1}{1-\beta} \approx -\frac{K_1^{B-1}K_2^{-(B-1)}}{V'(K_2)-1} + \frac{1}{1-V'(K_1)}. \tag{13}$$

By substitution into (8), we find for $1 \ll i \leqslant B$

$$E_i \approx \frac{\frac{K_1^{-i+B-1}}{1-V'(K_1)} - \frac{K_1^{B-1}K_2^{-i}}{V'(K_2)-1}}{\frac{1}{1-V'(K_1)} - \frac{K_1^{B-1}K_2^{-(B-1)}}{V'(K_2)-1}}.$$

Note that because $K_1/K_2 < 1$ for any $\rho \neq 1$, the second terms of both the numerator and the denominator vanish for large B and large i, which yields:

Theorem 3. The expected number of visits E_{B-j} to state $B-j$ of the $M/G/1/B$ embedded Markov chain, starting from state $B-1$, has the following asymptotic behaviour for large B:

$$\lim_{B\to\infty} \frac{E_{B-j}}{K_1^{j-1}} = 1, \tag{14}$$

which for $\rho < 1$ reduces to

$$\lim_{B\to\infty} E_{B-j} = 1. \tag{15}$$

Remark. Here we give a less rigorous, but more intuitive explanation of (15).

Consider the embedded Markov chain of the bounded system at service beginning epochs, in the limit for $B \to \infty$. The probability of going from state m to state $m-1$ is simply equal to the probability of no arrivals during a service period, which we henceforth denote by γ. Starting in state m, define $\overline{E}_m$ to be the expected number of visits to state m before reaching any state above m. Clearly, $\overline{E}_m \geqslant 1$, since the starting in state m is also counted. Furthermore, for infinitely high levels $\overline{E}_m$ is independent of m, so $\overline{E}_m$ is equal to some constant E. Since we are considering subsequent hits, E_m (as defined earlier) is the expected number of visits to state m, starting from state $B-1$ until absorption due to a full buffer or an empty system. One can easily see that it must satisfy the recursion $E_m = \gamma E_{m+1}\overline{E}_m$, which for sufficiently large m reduces to

$$E_m \approx C E_{m+1}$$

with $C = \gamma E$. Since the starting state $B-1$ is never visited again until absorption, $E_{B-1} = 1$ is a boundary condition for the above recursion. It follows that

$$E_m \approx C^{B-m-1} \quad \text{for } 1 \ll m \leqslant B-1.$$

To determine C, we use the following two arguments. Since the embedded Markov chain eventually reaches an absorbing state at full buffer or empty system, E_m must be bounded for all m and B (with $m < B$), which is possible only if $C \leqslant 1$. On the other hand, since $\rho < 1$, there is a nonzero probability that the embedded Markov chain eventually reaches state 0. Therefore, E_m must not vanish for low values of m even at large B, which is possible only if $C \geqslant 1$. Obviously, the only value of C which satisfies both conditions is $C = 1$. Consequently, $E_m \approx 1$, for $1 \ll m \leqslant B-1$.

5. Past and remaining service time distributions

Denote by X_n the duration of the service that starts at the nth embedded point ($n \geqslant 1$). As for the doubly-unbounded system, N_n is the state (number of customers) of the system at the nth embedded point. Without loss of generality, we assume that after absorption (due to either full buffer or empty system) the embedded Markov chain enters state 0 and stays there, i.e., N_n becomes 0. Define S_n to be the time, starting from the nth embedded point, until full buffer would be reached in the absence of any further service completions. Clearly, S_n has an Erlang–$(B-N_n)$ distribution, whose density for a given $N_n = i$ we denote by $g_i(s)$; thus

$$g_i(s) = f_{S_n}(s \mid N_n = i) = \lambda \frac{(\lambda s)^{B-i-1}}{(B-i-1)!}\, \mathrm{e}^{-\lambda s}.$$

Write the past service time distribution as a sum over a set of disjoint events, which together cover all ways the event $Y \leqslant y$ can happen:

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \sum_{n=1}^{\infty} \sum_{i=1}^{B-1} \mathbb{P}(S_n \leqslant y \wedge S_n \leqslant X_n \wedge N_n = i).$$

Note that if $S_n \leqslant X_n$, then n is the last embedded point before reaching full buffer, in which case $Y = S_n$. Furthermore, the second summation is over the nonabsorbing states $1 \leqslant i \leqslant B-1$, thus restricting the first summation to embedded points until absorption. Next, conditioning on $N_n = i$ gives

$$F_Y(y) = \sum_{n=1}^{\infty} \sum_{i=1}^{B-1} \mathbb{P}(S_n \leqslant y \wedge S_n \leqslant X_n \mid N_n = i)\, \mathbb{P}(N_n = i).$$

Using the independence of S_n and X_n, and $\sum_{n=1}^{\infty} \mathbb{P}(N_n = i) = E_i$, we find:

$$F_Y(y) = \sum_{i=1}^{B-1} E_i \int_0^{\infty} \int_0^{\infty} I_{s \leqslant y}\, I_{s \leqslant x}\, \mathrm{d}F_X(x) g_i(s)\, \mathrm{d}s = \int_0^{y} \overline{F}_X(s) H(s)\, \mathrm{d}s,$$

where $H(s)$ is defined as

$$H(s) = \sum_{i=1}^{B-1} g_i(s) E_i = \lambda \sum_{i=0}^{B-2} \frac{(\lambda s)^i}{i!} \mathrm{e}^{-\lambda s} E_{B-i-1}. \tag{16}$$

By differentiation, one finds the probability density of Y:

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = H(y)\overline{F}_X(y), \tag{17}$$

which holds only if $\overline{F}_X(x)$ is continuous at $x = y$. At a discontinuity of $\overline{F}_X(\cdot)$, f_Y does not exist.

Similarly we can write for the remaining service time distribution upon reaching full buffer

$$\begin{aligned}
\mathbb{P}(z < Z < \infty) &= \sum_{n=1}^{\infty} \sum_{i=1}^{B-1} \mathbb{P}(X_n - S_n > z \mid N_n = i)\, \mathbb{P}(N_n = i) \\
&= \sum_{i=1}^{B-1} E_i \int_0^{\infty} \int_0^{\infty} I_{x-s>z}\, \mathrm{d}F_X(x) g_i(s)\, \mathrm{d}s \\
&= \int_0^{\infty} \overline{F}_X(z+s) H(s)\, \mathrm{d}s.
\end{aligned}$$

Differentiation yields the probability density:

$$\begin{aligned}
f_Z(z) &= -\frac{\mathrm{d}\mathbb{P}(Z > z)}{\mathrm{d}z} = -\frac{\mathrm{d}}{\mathrm{d}z} \int_z^{\infty} \overline{F}_X(t) H(t-z)\, \mathrm{d}t \\
&= \overline{F}_X(z) H(0) + \int_{t=z}^{t=\infty} \overline{F}_X(t)\, \mathrm{d}H(t-z) \\
&= \overline{F}_X(z) H(0) - \overline{F}_X(z) H(0) - \int_z^{\infty} H(t-z)\, \mathrm{d}\overline{F}_X(t) \\
&= \int_z^{\infty} H(t-z)\, \mathrm{d}F_X(t).
\end{aligned} \tag{18}$$

Just like f_Y, also f_Z does not exist at discontinuities of $\overline{F}_X(\cdot)$.

5.1. First hit

For the first hit and $\rho \neq 1$, E_i is given by (11), and using (16) we find the asymptotic expression for $H(s)$:

$$H(s) \approx \frac{K_2^{-(B-1)} \lambda (K_2 - \alpha)}{V'(K_2) - 1} \left(\sum_{i=0}^{B-2} \frac{(\lambda s K_2)^i}{i!} \mathrm{e}^{-\lambda s} - \sum_{i=0}^{B-2} \frac{(\lambda s K_1)^i}{i!} \mathrm{e}^{-\lambda s} \right)$$

$$\approx \frac{K_2^{-(B-1)}\lambda(K_2-\alpha)}{V'(K_2)-1}\left(\sum_{i=0}^{\infty}\frac{(\lambda s K_2)^i}{i!}\,\mathrm{e}^{-\lambda s}-\sum_{i=0}^{\infty}\frac{(\lambda s K_1)^i}{i!}\,\mathrm{e}^{-\lambda s}\right)$$
$$= \frac{K_2^{-(B-1)}\lambda(K_2-\alpha)}{V'(K_2)-1}\big(\mathrm{e}^{\lambda(K_2-1)s}-\mathrm{e}^{\lambda(K_1-1)s}\big).$$

According to (17), we find the probability density of the past service time upon reaching full buffer by multiplying $H(y)$ by $\overline{F}_X(y)$. For $\rho < 1$ we have $K_1 = 1$, so this can be simplified to

$$f_Y(y) \approx \frac{K_2^{-(B-1)}\lambda(K_2-\alpha)}{V'(K_2)-1}\big(\mathrm{e}^{\lambda(K_2-1)y}-1\big)\overline{F}_X(y),$$

and for $\rho > 1$, we have $K_2 = 1$, yielding

$$f_Y(y) \approx \frac{\lambda(1-\alpha)}{\rho-1}\big(1-\mathrm{e}^{\lambda(K_1-1)y}\big)\overline{F}_X(y).$$

The above distributions are, in general, defective. If only the *conditional* distribution of the remaining service times is of interest (i.e., conditional on reaching full buffer), the above expressions and (18) can easily be normalized, leading to:

Theorem 4. The probability densities of the past and remaining service times in an $M/G/1/B$ queue, conditional on reaching full buffer, and starting from empty system, have the following asymptotic form:

$$\lim_{B\to\infty} f_{Y|\mathrm{fb}}(y) = \frac{\lambda}{1-\rho}\big(\mathrm{e}^{\lambda(K-1)y}-1\big)\overline{F}_X(y) \tag{19}$$

and

$$\lim_{B\to\infty} f_{Z|\mathrm{fb}}(z) = \frac{\lambda}{1-\rho}\int_z^{\infty}\big(\mathrm{e}^{\lambda(K-1)(t-z)}-1\big)\,\mathrm{d}F_X(t), \tag{20}$$

with $K = K_2$ if $\rho < 1$ and $K = K_1$ if $\rho > 1$.

5.2. *Subsequent hits*

For subsequent hits and $\rho \neq 1$, E_i is given by (14). As in the previous section, $H(s)$ can be found using (16), yielding

$$H(s) \approx \lambda \mathrm{e}^{\lambda(K_1-1)s},$$

so according to (17) the past service time density is

$$f_Y(y) \approx \lambda \mathrm{e}^{\lambda(K_1-1)y}\overline{F}_X(y), \tag{21}$$

and the remaining service time density is

$$f_Z(z) \approx \lambda\int_z^{\infty}\mathrm{e}^{\lambda(K_1-1)(t-z)}\,\mathrm{d}F_X(t). \tag{22}$$

Note that for $\rho > 1$, these distributions are not defective. For $\rho < 1$ we have $K_1 = 1$, which reduces the above expressions to

$$f_Y(y) \approx \lambda \overline{F}_X(y), \qquad f_Z(z) \approx \lambda \overline{F}_X(z).$$

These are defective distributions. Their total probability is easily shown to be ρ, allowing us to calculate the densities conditional on reaching full buffer.

Theorem 5. The probability densities of the past and remaining service times in an $M/G/1/B$ queue, conditional on reaching full buffer, and starting from full buffer, have the following asymptotic form:

$$\lim_{B\to\infty} f_{Y|\mathrm{fb}}(y) = \begin{cases} \dfrac{\lambda}{\rho}\overline{F}_X(y) & \text{for } \rho < 1, \\ \lambda \mathrm{e}^{\lambda(K_1-1)y}\overline{F}_X(y) & \text{for } \rho > 1 \end{cases}$$

and

$$\lim_{B\to\infty} f_{Z|\mathrm{fb}}(z) = \begin{cases} \dfrac{\lambda}{\rho}\overline{F}_X(z) & \text{for } \rho < 1, \\ \lambda \displaystyle\int_z^\infty \mathrm{e}^{\lambda(K_1-1)(t-z)}\,\mathrm{d}F_X(t) & \text{for } \rho > 1. \end{cases}$$

6. The limit $\rho \to 1$

For $\rho = 1$, equation (4) has only one solution (K_1 and K_2 approach 1 as ρ approaches 1). Since the analysis so far assumes two distinct solutions of (4), the obtained results may not hold for $\rho = 1$. However, we show that the limits of these results as $\rho \uparrow 1$ and as $\rho \downarrow 1$ exist and are identical, so we can assume them to be the result for $\rho = 1$.

6.1. First hit

In order to calculate the limit for the first hit, we need to examine the behaviour of K for ρ near 1. Consider the function

$$g(\rho, z) = \begin{cases} \dfrac{V(z) - z}{1 - z} & \text{for } z \neq 1, \\ 1 - \rho & \text{for } z = 1, \end{cases}$$

where $V(z)$ is given in (3). The function $g(\rho, z)$ as defined above is continuous at $z = 1$, because $\lim_{z\to 1} g(\rho, z) = 1 - \rho$ (using L'Hospital's rule).

Clearly, for $\rho \neq 1$, the solution K of the equation $g(\rho, K) = 0$, is the same $K \neq 1$ which is defined in section 3 as a solution of (4). Note that for $\rho \to 1$, also $K \to 1$. Calculation shows that for all ρ

$$\left.\frac{\partial g(\rho, z)}{\partial z}\right|_{z=1} = -\frac{\lambda^2 \mathbb{E}(X^2)}{2} \quad \text{and} \quad \left.\frac{\partial g(\rho, z)}{\partial \rho}\right|_{z=1} = -1.$$

Then the implicit function theorem applied to $g(\rho, z)$ at $z = 1$ and $\rho = 1$ implies that

$$\lim_{\rho\to 1} \frac{\mathrm{d}K}{\mathrm{d}\rho} = -\frac{\partial g(\rho, z)/\partial\rho}{\partial g(\rho, z)/\partial z}\bigg|_{z=1} = -\frac{2}{\lambda^2\mathbb{E}(X^2)}.$$

Using L'Hospital's rule, we get the following limit:

$$\begin{aligned}\lim_{\rho\to 1} \frac{\lambda}{1-\rho}\big(\mathrm{e}^{\lambda(K-1)y} - 1\big) &= \lim_{\rho\to 1} \lambda \frac{(\mathrm{d}/\mathrm{d}\rho)(\mathrm{e}^{\lambda(K-1)y} - 1)}{(\mathrm{d}/\mathrm{d}\rho)(1-\rho)} \\ &= -\lambda \frac{\mathrm{d}}{\mathrm{d}\rho}\big(\lambda(K-1)y\big)\bigg|_{\rho=1} = \frac{2y}{\mathbb{E}(X^2)},\end{aligned}$$

which we substitute into (19) to find the conditional probability density of the past service time for the first hit:

$$\lim_{\rho\to 1} f_{Y|\mathrm{fb}}(y) \approx \frac{2y}{\mathbb{E}(X^2)} \overline{F}_X(y).$$

Similarly, the limit of the conditional remaining service time distribution for the first hit (as given by (20)) is

$$\lim_{\rho\to 1} f_{Z|\mathrm{fb}}(z) \approx \frac{2}{\mathbb{E}(X^2)} \int_z^\infty (t - z)\, \mathrm{d}F_X(t).$$

6.2. Subsequent hits

To get the past service time distribution for subsequent hits when $\rho = 1$, we need to calculate the limit of (21) and (22) for $\rho \to 1$. These limits are trivial, since these functions turn out to be continuous at $\rho = 1$. So all results derived in section 5.2 are also valid for $\rho = 1$.

7. Approximation for full-buffer probability

In section 5, we found expressions for the asymptotic distributions of the past and remaining service times upon reaching a high level (e.g., full buffer in the bounded system). It was noted that these distributions are defective; i.e., the total probability of these distributions is less than 1. This defect of course represents the fact that the system does not always reach full buffer.

In section 4 we defined α to be the probability that the bounded system hits level 0 before reaching full buffer. As any path must either be absorbed at 0 or at B, we can conclude that the probability of reaching full buffer (i.e., absorption at B) is given by $1 - \alpha$, which we calculate in the following.

7.1. *First hit*

Eliminating β from (9) and (10) in section 4.1 gives:

$$\frac{K_2^{-(B-1)}(K_2-\alpha)}{V'(K_2)-1} \approx \frac{K_1^{-(B-1)}(K_1-\alpha)}{1-V'(K_1)}.$$

For both $\rho < 1$ ($K = K_2$) and $\rho > 1$ ($K = K_1$), this is

$$\frac{K^{-(B-1)}(K-\alpha)}{V'(K)-1} \approx \frac{1-\alpha}{1-\rho},$$

so

$$1-\alpha \approx \frac{K^{-(B-1)}(K-1)}{(V'(K)-1)/(1-\rho) - K^{-(B-1)}},$$

which for large B approaches

$$1-\alpha \approx \begin{cases} \dfrac{K(K-1)(1-\rho)}{V'(K)-1}K^{-B} & \text{for } \rho < 1, \\ 1-K & \text{for } \rho > 1. \end{cases} \tag{23}$$

Remark. Relationship with large deviation results:

For $\rho < 1$, the decay rate of the full-buffer probability in (23) is given by $\log K$, where K is determined from (4). For verification, we will now show that this is equal to the decay rate obtained using large deviation theory.

According to [7], the large-deviations calculation of the decay rate of the full-buffer probability starts by finding the nontrivial solution θ^* of the equation $\widetilde{F}_X(-\theta^*) = (\lambda+\theta^*)/\lambda$. Then the decay rate is given by $\log K'$, with $K' = (\lambda+\theta^*)/\lambda$. Using the latter equality to rewrite the former in terms of K', we get $\widetilde{F}_X(\lambda - \lambda K') = K'$. This is equivalent to (4), so $K' = K$.

Finally, it is interesting to note that $V'(K)$ can easily be shown to be the traffic intensity (i.e., average arrival rate divided by average service rate) in the so-called "θ^*-conjugate" system, in which the inter-arrival and service time distributions are exponentially twisted with the parameters θ^* and $-\theta^*$, respectively, [7]. This has applications in simulation speed-up techniques based on importance sampling.

7.2. *Subsequent hits*

For calculating the full-buffer probability for subsequent hits, we start from equations (12) and (13). By substituting the latter into the former, we obtain:

$$\alpha \approx \frac{K_1^{B-1}}{-\frac{K_1^{B-1}K_2^{-(B-1)}}{V'(K_2)-1} + \frac{1}{1-V'(K_1)}}.$$

This can be simplified by considering the cases $\rho < 1$ and $\rho > 1$ separately. First the case $\rho < 1$, which implies that $K_1 = 1$ and $K = K_2$:

$$1 - \alpha \approx 1 - \frac{1}{-\frac{K^{-(B-1)}}{V'(K)-1} + \frac{1}{1-\rho}} \approx \rho, \tag{24}$$

where the second step uses the fact that B is large and $K > 1$. For the case $\rho > 1$, which implies that $K = K_1$ and $K_2 = 1$, we find:

$$1 - \alpha \approx 1 - \frac{1}{-\frac{1}{\rho-1} + \frac{K^{-(B-1)}}{1-V'(K)}} \approx 1 - \big(1 - V'(K)\big)K^{B-1}, \tag{25}$$

where the second step uses the fact that B is large and $K < 1$. From this, one sees that the full-buffer probability $1 - \alpha \approx 1$, which is not surprising, since we start from just below full buffer in a system with a higher arrival rate than service rate ($\rho > 1$).

8. Numerical validations and an application

In the previous sections, we have derived several asymptotic results which are valid for infinitely high levels in $M/G/1$ queues. For sufficiently high levels, the results may still be used as approximations. In general, it is difficult to calculate error bounds for these approximations. However, we note that many of the approximations involve neglecting terms of the form K^B (if $K < 1$) or K^{-B} (if $K > 1$). For such a term to be very small, B must be very large and/or K must be far from 1. The value of K depends both on the form of the service time distribution $F_X(\cdot)$ and on ρ. If ρ approaches 1, K also approaches 1. Consequently, we can expect the approximation to be good if B is large and ρ is not close to 1.

8.1. Example: $M/D/1/B$

In order to illustrate the validity of the approximations, we first consider a simple $M/D/1/B$ queue. We assume the deterministic service time to be 1, thus $\mathbb{E}(X) = 1$, and $\rho = \lambda$. This leaves two parameters to vary, namely, ρ (traffic intensity) and B (buffer size).

First, we will test our approximations for the full-buffer probability, presented in section 7. For this queue, the true value of the full-buffer probability can also be computed numerically. To validate our approximations, table 2 shows both the approximate and the true values of the probability of reaching full buffer, starting from an empty system (i.e., first hit) and starting from level $B - 1$ (i.e., subsequent hits). As expected, the approximation is good for large B and for ρ not close to 1.

The accuracy of the approximations for the remaining service time distribution upon first and subsequent full-buffer hits is illustrated in figure 2. (Note in this example that because of the deterministic service time of 1, the remaining service time cannot exceed 1.) As a reference, simulation results are shown (with solid lines); these of course

Table 2
Comparison of approximation and true values for the probability of reaching full buffer in an $M/D/1/B$ queue.

Full buffer	ρ	B	Approximation	Exact	Difference (%)
First hit	0.8	5	$8.327 \cdot 10^{-2}$	$9.851 \cdot 10^{-2}$	15
		10	$9.659 \cdot 10^{-3}$	$9.835 \cdot 10^{-3}$	1.8
		20	$1.2996 \cdot 10^{-4}$	$1.2999 \cdot 10^{-4}$	0.024
		40	$2.3526 \cdot 10^{-8}$	$2.3526 \cdot 10^{-8}$	0.000004
	0.95	5	0.06891	0.1933	64
		10	0.04144	0.06759	39
		20	0.01498	0.01742	14
		40	0.001959	0.001995	1.8
	1.05	5	0.09370	0.2700	65
		10	0.09370	0.1560	40
		20	0.09370	0.1101	15
		40	0.09370	0.09570	2.1
	1.2	5	0.3137	0.3900	20
		10	0.3137	0.3233	3.0
		20	0.3137	0.3139	0.069
		40	0.3137	0.3137	0.000035
Subsequent hit	0.95	5	0.95	0.8597	11
		10	0.95	0.9184	3.4
		20	0.95	0.9419	0.9
		40	0.95	0.9491	0.09
	1.05	5	0.9674	0.9060	6.8
		10	0.9800	0.9668	1.4
		20	0.9925	0.9912	0.13
		40	0.9990	0.9989	0.002

have some small statistical errors, at most 0.004 with 95% confidence. The analytical approximations (plotted with dashed lines) are given by the expressions shown in the figure, which follow directly from the analysis in section 5. Clearly, the approximate analytical distributions agree quite well with the simulation results, especially considering the fact that we have a relatively small B and ρ close to 1.

8.2. Example: $M/H_2/1/B$

As an example with nondeterministic service time, we consider an $M/H_2/1/B$ queue. We choose the hyperexponential service time distribution such that the service rate is either 2 (with probability $1/2$), or $2/3$ (with probability $1/2$); thus $\mathbb{E}(X) = 1$ and $\rho = \lambda$.

Let us first test the approximation for the full-buffer probability. Table 3 shows the results from our approximation, as well as results from a numerical computation for comparison. Clearly, for ρ not close to 1, our approximation is quite good. A comparison with table 2 suggests that for the same ρ, the approximations are better for the $M/D/1/B$

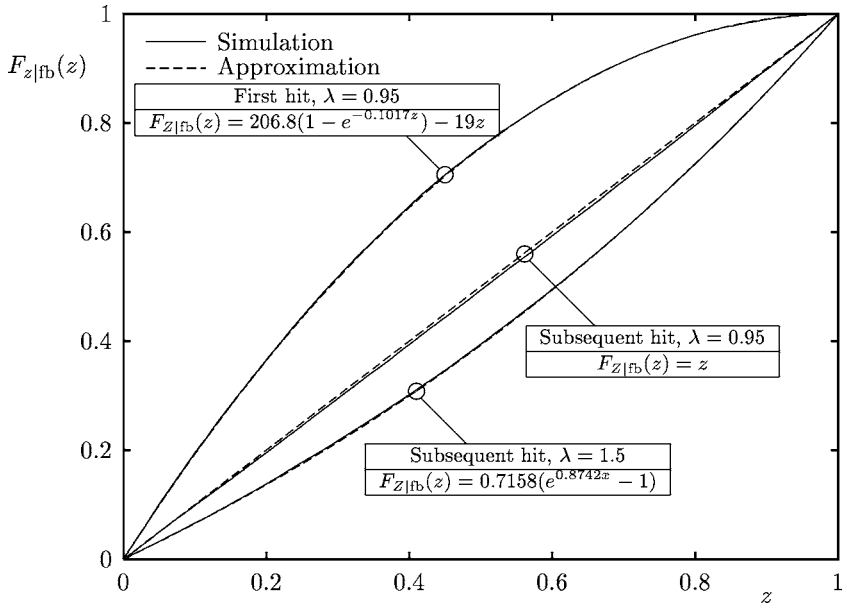

Figure 2. Remaining service time distributions in an $M/D/1/10$ queue.

system than for the $M/H_2/1/B$ system. Presumably, this is due to the larger variance of the service time in the latter system.

For the $M/H_2/1/B$ queue, figure 3 shows the remaining service time distribution obtained from our analytical approximations and from simulations. Again, the agreement between our approximations and the simulation results is evident.

8.3. Application: estimation of consecutive cell loss probabilities in an $M/G/1/B$ queue

As an application example, we consider the calculation of consecutive-cell-loss (CCL) probabilities in $M/G/1/B$ queues. This has applications in the estimation of the Quality-of-Service (QoS) provided by ATM (Asynchronous Transfer Mode) telecommunication systems. The word "cell" in CCL refers to the packets of data with which an ATM system transports information; for the purpose of the present paper, they are just customers arriving to the queue. The n-CCL probability, denoted by γ_n, is defined as the probability that during one busy cycle, at least once a group of n consecutive arrivals (cells) are all lost (due to buffer overflow). The estimation of γ_n has been studied before in [3], using simulation.

In order to simplify the calculation of γ_n, start by noting that all n cells that are lost consecutively in the n-CCL event, must arrive during a single full-buffer period, since between two full-buffer periods at least one arrival is accepted. This allows us to write

Table 3
Comparison of approximation and true values for the probability of reaching full buffer in an $M/H_2/1/B$ queue.

Case	ρ	B	Approximation	Exact	Difference (%)
First hit	0.4	10	0.0005005	0.0005007	0.054
		20	$3.098 \cdot 10^{-7}$	$3.098 \cdot 10^{-7}$	0.00
	0.6	10	0.00717	0.00728	1.5
		20	$1.213 \cdot 10^{-4}$	$1.213 \cdot 10^{-4}$	0.024
	0.8	10	0.03051	0.03618	16
		20	$5.145 \cdot 10^{-3}$	$5.284 \cdot 10^{-3}$	2.6
	1.2	10	0.1363	0.1744	22
		20	0.1363	0.1436	5.1
	1.6	10	0.3175	0.3238	1.9
		20	0.3175	0.3177	0.05
	2.0	10	0.4342	0.4355	0.3
		20	0.4342	0.4343	0.00
Subsequent hit	0.4	10	0.4000	0.3997	0.07
		20	0.4000	0.4000	0.00
	0.6	10	0.6000	0.5942	0.97
		20	0.6000	0.5999	0.016
	0.8	10	0.8000	0.7629	4.9
		20	0.8000	0.7946	0.68
	1.2	10	0.9563	0.9441	1.3
		20	0.9899	0.9894	0.053
	1.6	10	0.9885	0.9882	0.03
		20	0.9997	0.9997	0.00
	2.0	10	0.9972	0.9972	0.00
		20	1.0000	1.0000	0.00

γ_n as follows:

$$\gamma_n = \gamma p_{1n} + \frac{\gamma(1 - p_{1n})\phi p_n}{1 - \phi(1 - p_n)}, \tag{26}$$

where the following definitions have been used:

- γ = the probability of reaching full-buffer in a busy cycle (i.e., reaching level B, after starting from level 0 and before reaching 0 again).
- ϕ = the probability of reaching yet another full-buffer period in the same busy cycle (i.e., reaching level B, after starting from level $B - 1$ and before hitting level 0).
- p_{1n} = the probability of n or more arrivals during the first full-buffer period in a busy cycle.
- p_n = the probability of n or more arrivals during a subsequent full-buffer period.

Note that these probabilities are well-defined due to the fact that the arrival process is memoryless, and the fact that the second and later full-buffer periods are stochastically equivalent.

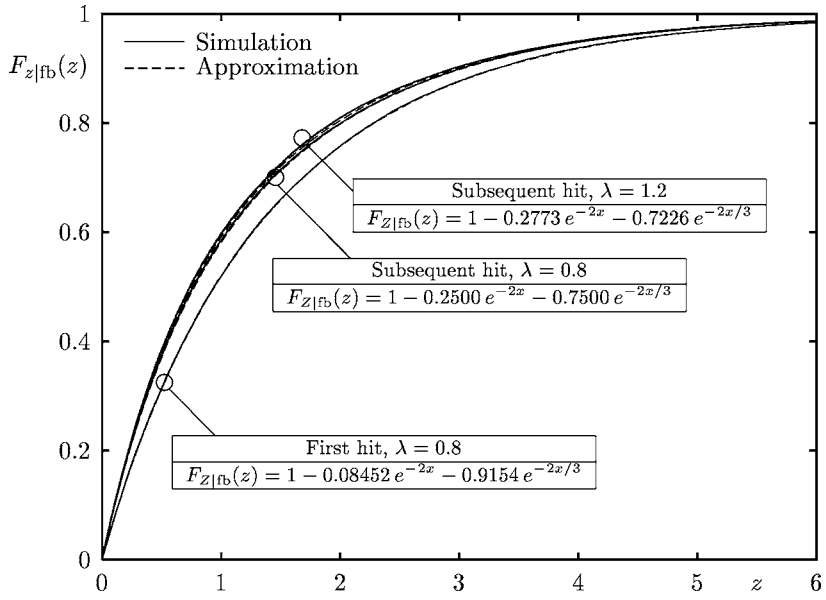

Figure 3. Remaining service time distributions in an $M/H_2/1/10$ queue.

Two of the four probabilities involved, namely, γ and ϕ, are simply the probabilities of reaching first and subsequent full-buffer periods, respectively, which can (for large B) be approximated by $1 - \alpha$ as calculated in section 7 (equations (23) and (24)). The other two, p_{1n} and p_n, are the probabilities that at least n (Poisson) arrivals occur during a first and a subsequent full-buffer period, respectively. If the distributions of those durations are known, these two probabilities can be estimated using a straightforward integration:

$$p_{1n} = \int_0^\infty \mathrm{e}^{-\lambda x} \sum_{i=n}^{\infty} \frac{(\lambda x)^i}{i!}\, \mathrm{d}G_1(x), \tag{27}$$

$$p_n = \int_0^\infty \mathrm{e}^{-\lambda x} \sum_{i=n}^{\infty} \frac{(\lambda x)^i}{i!}\, \mathrm{d}G(x), \tag{28}$$

where $G_1(\cdot)$ and $G(\cdot)$ are the distribution functions of the duration of first and subsequent full-buffer periods, respectively. If the overflow level B is high enough, $G_1(\cdot)$ and $G(\cdot)$ can be approximated by the asymptotic distributions that we have derived in the present paper (theorems 4 and 5).

As an example, consider an $M/D/1/B$ queue, with arrival rate 0.8 and deterministic service time $d = 1$. The approximate values for γ and ϕ can be read from table 2. The duration of the first full-buffer period asymptotically has a density $\mathrm{d}G_1(x)/\mathrm{d}x$ of the form $(\mathrm{e}^{0.43084(1-x)} - 1)$, while the distribution $G(x)$ of the duration of subsequent full-buffer periods asymptotically is uniform on [0, 1]. By numerical evaluation of the integrals in (27) and (28), approximate values of p_{1n} and p_n can be calculated. Finally,

Table 4
Analytic approximation of n-CCL probability for $M/D/1/B$ queues, with $\lambda = 0.8$ and $d = 1$.

B	n	γ_n (anal. appr.)	γ_n (exact)
5	1	$5.412 \cdot 10^{-2}$	$6.018 \cdot 10^{-2}$
	4	$7.239 \cdot 10^{-4}$	$7.246 \cdot 10^{-4}$
	16	$1.330 \cdot 10^{-17}$	$1.329 \cdot 10^{-17}$
	64	$1.174 \cdot 10^{-98}$	$1.175 \cdot 10^{-98}$
10	1	$6.278 \cdot 10^{-3}$	$6.352 \cdot 10^{-3}$
	4	$8.388 \cdot 10^{-5}$	$8.398 \cdot 10^{-5}$
	16	$1.542 \cdot 10^{-18}$	$1.542 \cdot 10^{-18}$
	64	$1.362 \cdot 10^{-99}$	$1.363 \cdot 10^{-99}$
20	1	$8.447 \cdot 10^{-5}$	$8.448 \cdot 10^{-5}$
	4	$1.130 \cdot 10^{-6}$	$1.130 \cdot 10^{-6}$
	16	$2.075 \cdot 10^{-20}$	$2.075 \cdot 10^{-20}$
	64	$1.832 \cdot 10^{-101}$	$1.834 \cdot 10^{-101}$

the four probabilities are substituted into (26) to obtain the n-CCL probability. The results are shown in table 4, for several values of B and n.

For this relatively simple problem, it is also possible to calculate the n-CCL probability exactly, by carefully considering an embedded Markov chain; see [2]. The resulting values are also listed in table 4.

Clearly, the agreement between the analytical approximation and the exact results is very good; in fact, it is much better than should be expected. For example, consider $B = 5$: at such a low buffer size the approximations from the present paper are generally rather bad, and indeed table 2 lists an error of about 15% for the approximation of γ used here. Since γ_n is directly proportional to γ (according to (26)), a 15% error in γ should also contribute a 15% error to γ_n. At $n = 1$, γ_n indeed has an error of this order (11%), but at $n = 4$, 16 and 64, the error in γ_n is just 0.1%. It seems as if the large error in the approximation of γ is compensated for by errors in the approximations of ϕ, $\mathrm{d}G_1(\cdot)$ and $\mathrm{d}G(\cdot)$. Correlations between these four errors are of course to be expected, since they all come from one approximation method. However, it is surprising that they cancel so well; further analysis of this may be of interest.

9. Summary and concluding remarks

In this paper, we have derived analytical approximations for the probability densities of the past and remaining service time upon reaching a high level, e.g., full buffer, in $M/G/1$ queues. Table 5 summarizes the main results for these distributions, conditional on reaching full buffer. As a by-product, we also obtained approximations for the probability of reaching full buffer (for the first and subsequent hits) in a busy cycle; those are given in (23)–(25). However, the results in this paper are only valid if a solution,

Table 5
Probability densities of past and remaining service times upon reaching a high level in $M/G/1$ queues.

Case		Past service time density $f_{Y\|\mathrm{fb}}(y)$	Remaining service time density $f_{Z\|\mathrm{fb}}(z)$
First hit	$\rho \neq 1$	$\frac{\lambda}{1-\rho}(\mathrm{e}^{\lambda(K-1)y}-1)\overline{F}_X(y)$	$\frac{\lambda}{1-\rho}\int_z^\infty (\mathrm{e}^{\lambda(K-1)(t-z)}-1)\,\mathrm{d}F_X(t)$
	$\rho = 1$	$\frac{2}{\mathbb{E}(X^2)}y\overline{F}_X(y)$	$\frac{2}{\mathbb{E}(X^2)}\int_z^\infty (t-z)\,\mathrm{d}F_X(t)$
Subs. hit	$\rho \leqslant 1$	$\frac{1}{\mathbb{E}(X)}\overline{F}_X(y)$	$\frac{1}{\mathbb{E}(X)}\overline{F}_X(z)$
	$\rho > 1$	$\lambda \mathrm{e}^{\lambda(K-1)y}\overline{F}_X(y)$	$\lambda\int_z^\infty \mathrm{e}^{\lambda(K_1-1)(t-z)}\,\mathrm{d}F_X(t)$

Note: K is defined as the solution not equal to 1 of equation (4). For some distributions such a solution does not exist, and the above results are not valid.

unequal to 1, of (4) exists; in particular, this excludes cases where the distribution of the service time has a heavy tail.

Validations of the approximations are carried out by means of comparisons with true values obtained from exact numerical results and simulations. The approximations are shown to be most accurate for high levels and ρ not too close to 1, although the approximate distributions in table 5 remain surprisingly accurate for ρ near 1.

An extension of our results to queueing systems other than $M/G/1$ would be of much interest. Our present analysis is based on an embedded Markov chain formulation, which cannot be applied to most $GI/G/1$ systems; for such systems, a different approach needs to be devised. The only other category of queueing systems which does lend itself to an embedded Markov chain analysis, is $GI/M/1$, for which a study of the remaining inter-arrival time at service completion epochs would be of interest.

Appendix A. Solving the doubly-unbounded system

In this appendix, we present a proof of theorem 1. For definitions and properties of N_n, q_j, $V(\cdot)$, K_1 and K_2, see section 3.

Define the following double-sided z-transform of the distribution of N_n:

$$R^{(n)}(z) = \sum_{i=-\infty}^{\infty} r_i^{(n)} z^i.$$

From (2) we get

$$R^{(1)}(z) = 1,$$

and from (1) we have (for $0 < |z| \leqslant K_2$)

$$R^{(n)}(z) = \frac{V(z)}{z} R^{(n-1)}(z),$$

as can be seen either by expanding the summations, or by noting that N_n is just the sum of N_{n-1} and another independent random number with the generating function $V(z)/z$. Next, we easily find:

$$R^{(n)}(z) = R^{(1)}(z)\left(\frac{V(z)}{z}\right)^{n-1} = \left(\frac{V(z)}{z}\right)^{n-1}.$$

Now define $R(z)$ to be the double-sided z-transform of r_i, to find:

$$R(z) = \sum_{i=-\infty}^{\infty} r_i z^i = \sum_{i=-\infty}^{\infty} \sum_{n=1}^{\infty} r_i^{(n)} z^i = \sum_{n=1}^{\infty} R^{(n)}(z) = \sum_{n=1}^{\infty} \left(\frac{V(z)}{z}\right)^{n-1} = \frac{z}{z - V(z)}. \tag{A.1}$$

The change of the order of summation above is possible on the ring[2] $\{z\colon K_1 < |z| < K_2\}$ in the complex plane. So, (A.1) is valid on that ring.

Note that $R(z)$ itself is only defined on the ring; however, the right-hand side of (A.1) is also analytical outside the ring (except of course at K_1 and K_2), so it is the (unique) analytical continuation of $R(z)$.

Behaviour of r_i for $i \leqslant 0$. We have already shown that $|V(z)| < |z|$ on the ring $\{z\colon K_1 < |z| < K_2\}$. By Rouché's theorem (see, e.g., [8]), this means that $V(z) - z$ has exactly as many zeros on the disc $\{|z| < K_2\}$ as z does, while the latter of course has exactly one zero (at 0). We already know that $V(z) - z$ has a zero at K_1, which is on the disc. So that must be its only zero. Consequently, $z = K_1$ is the only pole of $R(z)/z = 1/(z - V(z))$ on the disc. Now calculate the residue of this pole:

$$\begin{aligned}\operatorname{Res}\left(\frac{R(z)}{z}, K_1\right) &= \lim_{z \to K_1} (z - K_1)\frac{R(z)}{z} = \lim_{y \to 1} (y-1)K_1 \frac{R(K_1 y)}{K_1 y} \\ &= \lim_{y \to 1} \frac{(y-1)K_1}{K_1 y - V(K_1 y)} = \lim_{y \to 1} \frac{K_1}{K_1 - K_1 V'(K_1 y)} = \frac{1}{1 - V'(K_1)},\end{aligned}$$

[2] Note that the last summation (a simple geometric sum) converges on the ring because there $|V(z)| < |z|$. This follows from observing that:

- Because of the convexity of $V(z)$ and the fact that $V(K_1) = K_1$ and $V(K_2) = K_2$, we have $V(z) < z$ for any (real and positive) z for which $K_1 < z < K_2$.
- Write z, which is in general a complex number, as $x + \mathrm{i}y$. With the definition (3) of $V(z)$ in terms of a Laplace transform of a positive function, one easily verifies that $|V(z)| = |V(x + \mathrm{i}y)| \leqslant |V(|x|)| \leqslant |V(|x + \mathrm{i}y|)|$.

where L'Hospital's rule was used in the fourth step. Knowing the only pole's residue, we can split $R(z)/z$ as follows (for $K_1 < |z| < K_2$):

$$\frac{R(z)}{z} = \sum_{i=0}^{\infty} r'_{i+1} z^i + \frac{1}{1 - V'(K_1)} \cdot \frac{1}{z - K_1} = \sum_{i=0}^{\infty} r'_{i+1} z^i + \frac{1/K_1}{1 - V'(K_1)} \cdot \frac{K_1/z}{1 - K_1/z}$$
$$= \sum_{i=0}^{\infty} r'_{i+1} z^i + \frac{1/K_1}{1 - V'(K_1)} \sum_{i=-\infty}^{-1} \left(\frac{z}{K_1}\right)^i.$$

The reason the first term $\sum_{i=0}^{\infty} r'_{i+1} z^i$ contains no negative powers of z, is that this term results from removing the only pole $R(z)$ has on $|z| < K_2$, so this term must be analytic on this disc, and therefore can be written as a Taylor series. Multiply the above by z to find

$$R(z) = \sum_{i=1}^{\infty} r'_i z^i + z \frac{1/K_1}{1 - V'(K_1)} \sum_{i=-\infty}^{-1} \left(\frac{z}{K_1}\right)^i = \sum_{i=1}^{\infty} r'_i z^i + \frac{1}{1 - V'(K_1)} \sum_{i=-\infty}^{0} \left(\frac{z}{K_1}\right)^i,$$

from which (5) directly follows.

Limit behaviour of r_i for large i. Using the final value theorem (Abelian theorem, see [8]) for z-transforms and applying L'Hospital's rule yield the following limit:

$$\lim_{i\to\infty} K_2^i r_i = \lim_{z\to 1}(1 - z)R(K_2 z) = \lim_{z\to 1} \frac{(1 - z)K_2 z}{K_2 z - V(K_2 z)}$$
$$= \lim_{z\to 1} \frac{(1 - z)K_2 - K_2 z}{K_2 - K_2 V'(K_2 z)} = \frac{1}{V'(K_2) - 1},$$

thus proving (6).

References

[1] S. Asmussen, Equilibrium properties of the $M/G/1$ queue, Z. Wahrscheinlichkeitstheorie Verwandte Gebiete 58 (1981) 267–281.

[2] P.T. de Boer, Analysis and efficient simulation of queueing models of telecommunication systems, Ph.D. thesis, University of Twente (2000) chapter 3.

[3] P.T. de Boer and V.F. Nicola, Hybrid importance sampling estimation of consecutive cell loss probability, AEÜ Internat. J. Electronics Commun. 52 (1998) 133–140.

[4] J.W. Cohen, *The Single Server Queue*, 2nd ed. (North-Holland, Amsterdam, 1982).

[5] D. Fakinos, The expected remaining service time in a single server queue, Oper. Res. 30 (1982) 1014–1018.

[6] M.J.J. Garvels, The splitting method in rare event simulation, Ph.D. thesis, University of Twente (2000) chapter 4.

[7] J.S. Sadowsky, Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue, IEEE Trans. Automat. Control 36 (1991) 1383–1394.

[8] E.C. Titchmarsh, *Theory of Functions* (Oxford Univ. Press, London, 1952).