Queueing Syst (2013) 75:169–188
DOI 10.1007/s11134-013-9366-6

Analysis of an $M/M/1+G$ queue operated under the FCFS policy with exact admission control

**Sudipta Das · Lawrence Jenkins ·
Debasis Sengupta**

Received: 16 September 2011 / Revised: 9 April 2013 / Published online: 14 June 2013
© Springer Science+Business Media New York 2013

Abstract In this article, we present an exact theoretical analysis of an $M/M/1$ system, with arbitrary distribution of relative deadline for the end of service, operated under the first come first served scheduling policy with exact admission control. We provide an explicit solution to the functional equation that must be satisfied by the workload distribution, when the system reaches steady state. We use this solution to derive explicit expressions for the loss ratio and the sojourn time distribution. Finally, we compare this loss ratio with that of a similar system operating without admission control, in the cases of some common distributions of the relative deadline.

Keywords Firm real-time system · Real-time queue · Exact admission control · Service time-dependent scheduling · Bounded sojourn time · Loss ratio

Mathematics Subject Classification (2000) Primary: 68M20 · 60K25 · Secondary: 90B22 · 90B36

1 Introduction

For some real-time systems such as web server, network router, or real-time database, the consequences of a missed deadline are not catastrophic, though timing require-

S. Das (✉) · L. Jenkins
Department of Electrical Engineering, Indian Institute of Science,
Bangalore 560012, India
e-mail: sdas@ee.iisc.ernet.in

L. Jenkins
e-mail: lawrn@ee.iisc.ernet.in

D. Sengupta
Applied Statistical Unit, Indian Statistical Institute, Kolkata 7000108, India
e-mail: sdebasis@isical.ac.in

ments are very critical. In these types of systems, named as firm real-time systems, one attempts to complete most of the jobs successfully. Such systems mainly consist of aperiodic jobs and it is typically not known when the job will arrive or what its service time and deadline will be [14]. If too many jobs arrive simultaneously, the system becomes overloaded and the jobs begin to miss their deadlines. A job may be executed partially before its deadline expires. This occurrence is undesirable from considerations of productive utilization of processor time as well as efficient running of applications. An application can be managed better if it is known at the arrival epoch of a job whether it can be completed. There have been attempts to screen jobs at the arrival stage through a scheduling test, so that the execution of all the admitted jobs (including the latest one) can be guaranteed. Ideally one would like to use an exact admission controller (EAC), which would admit a newly arrived job to the queue only if it is possible to complete this job as well as all the previously admitted jobs. The rejection of a job based upon the workload at arrival epoch would require that the service times of the jobs in the queue and the residual service time of the job in service are known.

A job is supposed to have a deadline either for its start of service or for its end of service. The systems where jobs include deadline for the beginning of the service and where the underlying scheduling policy is first come first served (FCFS) with admission control [6–8,13,19] or without it [2–4,12,16,18] are extensively studied in the literature. Other researchers [9,17] provided exact analyses of an $M/M/1+G$ queue operated under the FCFS scheduling policy without any admission control and job deadline for the end of service.

Cohen [8] investigated $M/M/1$ systems under the FCFS scheduling policy, where customers have constant deadline for the end of service, and an admission controller is used. Gavish et al. [11] partially generalized this work to the $M/G/1$ system and obtained explicit formulae for server utilization, loss ratio, mean sojourn time, mean server busy period, and distribution function of virtual waiting time in the $M/M/1$ case. A similar system was investigated by Van Dijk [20], who obtained an implicit relation satisfied by the workload distribution. Loris-Teghem [15] extended the analysis to queues with rational transforms for the interarrival and the service times, while keeping the deadline constant. Perry and Asmussen [19] studied a more general system in which jobs have stochastic deadline for the end of the service. They obtained a recursive solution of the workload distribution. This solution does not simplify further for general deadline distribution (except, for example, in the case of the exponential distribution), and does not appear to be suitable for computing the loss ratio. Also, Bekker [5] derives a formal solution for the steady-state workload density in $M/G/1$ systems in a generalized way for fixed deadline.

It transpires from the survey of the existing literature that explicit expressions for workload distribution, loss ratio, and other quantities of interest for the FCFS scheduler with EAC in the case of $M/M/1+G$ systems with deadline for the end of service has not been obtained till now. The $M/M/1+G$ system is the most basic and widely used model with stochastic deadline. It is the presence of admission control that makes the analyses difficult [21]. We show in this article that an explicit analysis for this model is possible.

In Sect. 2, we present the detailed model of the underlying system for subsequent analysis. In Sect. 3, we provide an explicit expression for the steady-state workload distribution of an $M/M/1+G$ queue for jobs with deadline for the end of service, operated under the FCFS scheduling policy with EAC. In Sect. 4, we present the virtual and the actual sojourn time distributions in the steady state of this system, as well as the loss ratio. Section 5 provides the waiting time distribution of a job under a different implementation of EAC. Finally, in Sect. 6, we show that the adoption of EAC indeed reduces the loss ratio of a $G/G/1+G$ system operating under the FCFS scheduling policy.

2 System model and notations

We model a firm real-time system with a single processor and an aperiodic workload as a single server queue with an infinite buffer. The infinite buffer ensures that there is no upper limit on the maximum number of jobs that can remain in the system. We assume that every job is ready as soon as it is released and it never suspends itself. The ith job, J_i, arrives at the instant A_i. It has a service time Y_i and a relative deadline D_i. The ith job requests that it is executed for Y_i time units during the time interval $[A_i, A_i + D_i]$, where $A_i + D_i$ is the absolute deadline (i.e., deadline for the end of service). We assume that job arrivals follow a Poisson process with rate λ and service times follow common distribution $G(\cdot)$ with mean $1/\mu$. The ratio $\rho = \lambda/\mu$ has been referred to as the normalized arrival rate. Also, the relative deadlines of the jobs follow common distribution $H(\cdot)$ having mean θ. The workload process is $\{V(t)\}_{t\geq 0}$. The workload seen by job J_i upon its arrival at time A_i is $V(A_i-)$. The job J_n is admitted to the system if and only if

$$V(A_n-) + Y_n \leq D_n.$$

Consequently,

$$V(A_n) = V(A_n-) + Y_n I_{\{V(A_n-)+Y_n\leq D_n\}},$$

where $I_{\{.\}}$ is the indicator variable. We define the loss ratio, α, as the limiting probability of rejection at the stage of admission control, i.e.,

$$\alpha = \lim_{n\to\infty} P(V(A_n-) + Y_n > D_n).$$

3 The workload process

In this section, as well as in Sects. 4 and 5, we analyze the performance of the FCFS scheduling policy with EAC, referred to as FCFS-EAC. We derive the workload distribution as a solution to a functional equation that must hold, when the system reaches steady state. In the next two propositions, we provide the equation for the $M/G/1+G$ queue and its solution for the $M/M/1+G$ queue.

Proposition 1 *The steady-state distribution of the workload process $V(t)$ of an $M/G/1+G$ queue, operating under the FCFS-EAC scheduling policy, consists of a density $f_V(\cdot)$ on $(0,\infty)$ satisfying*

$$f_V(u) = \lambda \int_0^u \int_u^\infty [G(z-v) - G(u-v)]\, \mathrm{d}H(z)\mathrm{d}F_V(v) \tag{1}$$

and possibly a point mass at 0 given by $\pi = 1 - \int_0^\infty f_V(u)\mathrm{d}u$.

Proof The workload seen by the job J_n on its arrival is $V(A_n-)$. For notational convenience, we use V_n for $V(A_n-)$. If X_{n+1} is the time between the arrivals of the job J_n and J_{n+1}, then

$$V_{n+1} = (V_n + Y_n I_{\{V_n+Y_n \le D_n\}} - X_{n+1}) \vee 0.$$

For $u \ge 0$,

$$\begin{aligned}\bar{F}_{V_{n+1}}(u) &= P\left((V_n + Y_n I_{\{V_n+Y_n \le D_n\}} - X_{n+1}) \vee 0 > u\right)\\ &= \int_0^\infty \lambda e^{-\lambda t} P(Y_n I_{\{Y_n \le D_n - V_n\}} > u + t - V_n)\mathrm{d}t.\end{aligned} \tag{2}$$

Now,

$$\begin{aligned}&P(Y_n I_{\{Y_n \le D_n - V_n\}} > u + t - V_n)\\ &= \int_0^\infty P(Y_n I_{\{Y_n \le D_n - v\}} > u + t - v)\mathrm{d}F_{V_n}(v)\\ &= \bar{F}_{V_n}(u+t) + \int_{u+t}^\infty \int_0^{u+t} P(Y_n I_{\{Y_n \le z - v\}} > u + t - v)\mathrm{d}F_{V_n}(v)\mathrm{d}H(z)\\ &= \bar{F}_{V_n}(u+t) + \int_{u+t}^\infty \int_0^{u+t} P(u + t - v < Y_n \le z - v)\mathrm{d}F_{V_n}(v)\mathrm{d}H(z).\end{aligned} \tag{3}$$

Assuming steady-state distribution as well as inserting the relation (3) in (2) yields, for $u > 0$,

$$\begin{aligned}\bar{F}_V(u) &= \int_0^\infty \lambda e^{-\lambda t} \left[\bar{F}_V(u+t) + \int_{u+t}^\infty \int_0^{u+t} \{G(z-v) - G(u+t-v)\}\, \mathrm{d}F_V(v)\mathrm{d}H(z)\right] \mathrm{d}t\\ &= \int_u^\infty \lambda e^{-\lambda(t-u)} \left[\bar{F}_V(t) + \int_0^t \int_t^\infty \{G(z-v) - G(t-v)\}\, \mathrm{d}H(z)\mathrm{d}F_V(v)\right] \mathrm{d}t.\end{aligned} \tag{4}$$

Since the last expression is differentiable, the workload distribution function $F_V(\cdot)$, has a density $f_V(\cdot)$ over $(0, \infty)$. The recursive relation (1), given in the proposition is obtained by differentiating (4). □

A different proof of the above proposition, by a level crossing argument, can be found in [19]. A solution to this implicit equation is difficult to obtain. Solution through a recursive relation was given in [19]. A recursive relation generally poses difficulty for further computations, such as that of the loss ratio. However, as the next proposition shows, a closed form expression can be obtained in the special case of the $M/M/1+G$ queue.

Proposition 2 *The steady-state distribution of the workload process $V(t)$ of an $M/M/1+G$ queue, operating under the FCFS-EAC policy, consists of a point mass π at 0 and a density $f_V(\cdot)$ on $(0,\infty)$ given by*

$$\frac{1}{\pi} = \mu e^{-B(0)} \int_0^\infty e^{B(u)-\mu u} \mathrm{d}u, \tag{5}$$

$$f_V(u) = \pi B'(u) e^{B(u)-\mu u - B(0)}, \tag{6}$$

where

$$B(u) = \lambda \left(e^{\mu u} \int_u^\infty e^{-\mu z} \bar{H}(z)\mathrm{d}z + \int_0^u \bar{H}(z)\mathrm{d}z \right), \quad \text{for } u \geq 0 \tag{7}$$

and $B'(u)$ is its first derivative, given by

$$B'(u) = \lambda\mu \int_u^\infty e^{-\mu(z-u)} \bar{H}(z)\mathrm{d}z, \quad \text{for } u \geq 0. \tag{8}$$

Proof For $G(x) = 1 - e^{-\mu x}$, we get from (1), for $u > 0$

$$\begin{aligned} f_V(u) &= \lambda \int_0^u e^{\mu v} \int_u^\infty \left(e^{-\mu u} - e^{-\mu z}\right) \mathrm{d}H(z)\mathrm{d}F_V(v) \\ &= \lambda\mu \left(\int_0^u e^{\mu v} \mathrm{d}F_V(v) \right) \left(\int_u^\infty e^{-\mu z} \bar{H}(z)\mathrm{d}z \right). \end{aligned} \tag{9}$$

Without loss of generality we assume that there exists a point mass π at point 0. Then

$$\frac{f_V(u)}{\int_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z} = \lambda\mu \left[\pi + \int_0^u e^{\mu v} f_V(v)\mathrm{d}v \right], \quad \text{for } u > 0. \tag{10}$$

Note that $f_V(u)$ has a derivative, since RHS of (9) is differentiable. Hence, we can differentiate both sides of (10). After simplifying, we get, for $u > 0$

$$\frac{1}{f_V(u)}\frac{\mathrm{d}f_V(u)}{\mathrm{d}u}+\frac{e^{-\mu u}\bar{H}(u)}{\int_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z}=\lambda\mu e^{\mu u}\int\limits_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z,$$

i.e.,

$$d\left[\ln\frac{f_V(u)}{\int_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z}\right]=\lambda\mu\left(e^{\mu u}\int\limits_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z\right)\mathrm{d}u.$$

It follows that

$$\begin{aligned}f_V(u)&=A_H e^{\lambda(e^{\mu u}\int_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z+\int_0^u \bar{H}(z)\mathrm{d}z)}\int\limits_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z\\&=\frac{A_H}{\lambda\mu}B'(u)e^{B(u)-\mu u}.\end{aligned}\tag{11}$$

Inserting (11) in (10) and then simplifying, we get

$$\frac{A_H}{\lambda\mu}=\pi e^{-\lambda\int_0^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z}.\tag{12}$$

Equation (6) is obtained by inserting (12) in (11). Now,

$$\pi+\int\limits_0^\infty f_V(u)\mathrm{d}u=1.\tag{13}$$

Hence, Eq. (6) is obtained by solving (13). □

Note that the emergence of an explicit solution to (1) hinges on the reduction of the double integral into two single integrals in (9). A similar technique was used by Bekker [5] in deriving the workload distribution for a finite-buffer $M^{(x)}/M^{(x)}/1$ queue.

We now consider models with three types of relative deadline distributions, namely degenerate, exponential, and uniform. For each type of distribution, we present the workload distribution functions.

Example 1 Degenerate distribution. The distribution function of the relative deadline is given by

$$H(z)=\begin{cases}0 & \text{if } z<\theta,\\ 1 & \text{if } z\geq\theta.\end{cases}$$

Hence,

$$\int_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z = \begin{cases} \frac{1}{\mu}(e^{-\mu u} - e^{-\mu\theta}) & \text{if } u \le \theta \\ 0 & \text{if } u > \theta \end{cases}$$

and

$$\int_0^u \bar{H}(z)\mathrm{d}z = \begin{cases} u & \text{if } u \le \theta, \\ \theta & \text{if } u > \theta. \end{cases}$$

Therefore,

$$B(u) = \begin{cases} \lambda\left(\frac{1}{\mu}(1 - e^{\mu(u-\theta)}) + u\right) & \text{if } u \le \theta, \\ \lambda\theta & \text{if } u > \theta. \end{cases}$$

Consequently, the workload distribution is determined by the expressions

$$\frac{1}{\pi} = e^{(\lambda-\mu)\theta-\rho(1-e^{-\mu\theta})}\left[1 + e^\rho \rho^{1-\rho}\int_{\rho e^{-\mu\theta}}^{\rho} x^{\rho-2}e^{-x}\mathrm{d}x\right], \tag{14}$$

$$f_V(u) = \begin{cases} \pi\lambda[1 - e^{-\mu(\theta-u)}]e^{(\lambda-\mu)u-\rho(e^{\mu u}-1)e^{-\mu\theta}} & \text{if } u \le \theta, \\ 0 & \text{if } u > \theta. \end{cases} \tag{15}$$

Equations (14) and (15) are equivalent to Eqs. (B-1) and (6) of [11].

Example 2 Exponential distribution. The distribution function of the relative deadline is given by

$$H(z) = 1 - e^{-\frac{z}{\theta}}.$$

Hence,

$$\int_u^\infty e^{-\mu z}\bar{H}(z)\mathrm{d}z = \frac{e^{-(\mu+\frac{1}{\theta})u}}{\mu + \frac{1}{\theta}}$$

and

$$\int_0^u \bar{H}(z)\mathrm{d}z = \theta(1 - e^{-\frac{u}{\theta}}).$$

Therefore,

$$B(u)=\lambda\theta\left(1-\frac{\mu}{\mu+\frac{1}{\theta}}e^{-\frac{u}{\theta}}\right).$$

Consequently, the workload distribution is determined by the expressions

$$\frac{1}{\pi}=1+\left(\frac{\mu+\frac{1}{\theta}}{\lambda\mu\theta}\right)^{\mu\theta}e^{\frac{\lambda\mu\theta}{(\mu+\frac{1}{\theta})}}\int_0^{\frac{\lambda\mu\theta}{(\mu+\frac{1}{\theta})}}x^{\mu\theta}e^{-x}\mathrm{d}x, \tag{16}$$

$$f_V(u)=\frac{\pi\lambda\mu}{\mu+\frac{1}{\theta}}e^{\frac{\lambda\mu\theta}{(\mu+\frac{1}{\theta})}-\frac{\lambda\mu\theta e^{-\frac{u}{\theta}}}{(\mu+\frac{1}{\theta})}-(\mu+\frac{1}{\theta})u}. \tag{17}$$

Equations (16) and (17) can also be found from the Corollary 2.8 of [19].

Example 3 Uniform distribution. The distribution function of the relative deadline is given by

$$H(z)=\begin{cases}\frac{z}{2\theta} & \text{if } z\le 2\theta,\\ 1 & \text{if } z>2\theta.\end{cases}$$

Hence,

$$\int_u^{\infty}e^{-\mu z}\bar{H}(z)\mathrm{d}z=\begin{cases}\frac{1}{\mu}\left[\left(1-\frac{1}{2\mu\theta}-\frac{u}{2\theta}\right)e^{-\mu u}+\frac{1}{2\mu\theta}e^{-2\mu\theta}\right] & \text{if } u\le 2\theta,\\ 0 & \text{if } u>2\theta.\end{cases}$$

and

$$\int_0^{u}\bar{H}(z)\mathrm{d}z=\begin{cases}u-\frac{u^2}{4\theta} & \text{if } u\le 2\theta,\\ \theta & \text{if } u>2\theta.\end{cases}$$

Therefore,

$$B(u)=\begin{cases}\rho\left[\left(1-\frac{1}{2\mu\theta}-\frac{u}{2\theta}\right)+\frac{1}{2\mu\theta}e^{\mu(u-2\theta)}\right]+\lambda u\left[1-\frac{u}{4\theta}\right] & \text{if } u\le 2\theta,\\ \lambda\theta & \text{if } u>2\theta.\end{cases}$$

Consequently, the workload distribution is determined by the expressions

$$f_V(u)=\begin{cases}\pi C(u) & \text{if } u\le 2\theta,\\ 0 & \text{if } u>2\theta.\end{cases}$$

and

$$\frac{1}{\pi} = 1 + \int_0^{2\theta} C(u)\mathrm{d}u$$

where,

$$C(u) = \frac{\rho}{2}\left[2\mu - \frac{(1 + \mu u e^{-\mu u} - e^{-2\mu\theta})}{\theta}\right] e^{\frac{\rho}{2\mu\theta}[-u\mu + (e^{\mu u}-1)e^{-2\mu\theta}] + \lambda u[1-\frac{u}{4\theta}]}.$$

□

4 Sojourn time and loss ratio

In order to compute the loss ratio, we need to find the probability of a job being denied admission by the EAC. Hence, we need to calculate the probability that the sum of the workload seen by a job on arrival and its service time is greater than its relative deadline. This sum is called the virtual sojourn time (i.e., the amount of time a job having infinite deadline will spend in the system). We provide the distribution of this quantity in the next proposition.

Proposition 3 *The steady-state distribution of the virtual sojourn time T of an incoming job in an $M/M/1+G$ queue, operating under the FCFS-EAC policy, has the density $f_T(\cdot)$ on $[0, \infty)$ given by*

$$f_T(s) = \mu\pi e^{B(s)-\mu s-B(0)}, \tag{18}$$

where π and $B(\cdot)$ are defined as in Proposition 2.

Proof Note that the workload seen by an incoming job is independent of its service time. Therefore, the virtual sojourn time distribution can be obtained by convolving the workload distribution with the service time distribution (assumed to be exponential). Hence, by using the result of Proposition 2, we have

$$\begin{aligned} f_T(s) &= \mu \int_0^s e^{-\mu(s-u)}\mathrm{d}F_V(u) \\ &= \mu\pi e^{-\mu s}\left[1 + e^{-B(0)}\int_0^s B'(u)e^{B(u)}\mathrm{d}u\right] \\ &= \mu\pi e^{B(s)-\mu s-B(0)}. \end{aligned}$$

□

We are now ready for the next proposition, which specifies the loss ratio.

Proposition 4 *The loss ratio of an $M/M/1+G$ queue, operating under the FCFS-EAC policy, is given by*

$$\alpha = 1 - \frac{\int_0^\infty \bar{H}(s)e^{B(s)-\mu s}\mathrm{d}s}{\int_0^\infty e^{B(s)-\mu s}\mathrm{d}s}, \tag{19}$$

where $H(\cdot)$ is the deadline distribution and $B(\cdot)$ is specified in Proposition 2.

Proof Let T and D be the virtual sojourn time and the relative deadline, respectively, of an incoming job. Hence, the loss ratio is

$$\begin{aligned}\alpha &= 1 - P(D > T)\\ &= 1 - \mu\pi e^{-B(0)} \int_0^\infty e^{B(s)-\mu s}\bar{H}(s)\mathrm{d}s.\end{aligned} \tag{20}$$

The expression given in Eq. (19) is obtained by substituting (6) in (20). □

Remark 1 An equivalent form of loss ratio given in (20) is

$$\frac{E_{T_2}\left[\exp\left\{\lambda\theta E_{T_1}(F(T_1)|T_1 > T_2)\right\}\bar{(H)}(T_2)\right]}{E_{T_2}\left[\exp\left\{\lambda\theta E_{T_1}(F(T_1)|T_1 > T_2)\right\}\right]},$$

where $F(\cdot)$ is the function defined by $F(t) = \frac{1}{\theta}\int_0^t \bar{H}(u)\mathrm{d}u$, and T_1 and T_2 are independent exponential random variates with mean $1/\mu$. The subscript to the expectations indicates the random variable, with respect to which the expectation is to be computed. The above expression is amenable to computation by Monte Carlo simulations.

Remark 2 It can be seen from Eq. (19) that

$$\lim_{\lambda\to 0} \alpha = P(D < Y),$$

where Y and D are the (mutually independent) service time and relative deadline of a job. Thus, when the arrival rate is very small, the loss ratio is the probability that the service time of a job exceeds its relative deadline.

Corollary 1 *The loss ratio of an $M/M/1+G$ system under the FCFS-EAC policy is less than that under the FCFS policy.*

Proof The loss ratio of an $M/M/1$ system operated under FCFS scheduling policy is (see Eq. (4.11) of [17])

$$\beta = 1 - \mu p_0 \int_0^\infty \bar{H}(s)e^{\lambda\int_0^s \bar{H}(x)\mathrm{d}x-\mu s}\mathrm{d}s, \tag{21}$$

where $p_0 = \left[\mu\int_0^\infty \exp\left\{\lambda\int_0^s \bar{H}(x)\mathrm{d}x - \mu s\right\}\mathrm{d}s\right]^{-1}$. One can write $1-\beta = E[\bar{H}(X)]$, where X has the probability density function $\mu p_0 e^{\lambda\int_0^s \bar{H}(x)\mathrm{d}x-\mu s}$. By the same token, (19) implies that one can write $1-\alpha = E[\bar{H}(X)\phi(X)]/E[\phi(X)]$, where X is the random variable described above, and ϕ is a real valued function over the positive real line, defined by $\phi(x) = \exp\left[\lambda \int_0^\infty \bar{H}(u+x)e^{-\mu u}\mathrm{d}u\right]$. Since $\phi(\cdot)$ and $\bar{H}(\cdot)$ are both decreasing functions, $\phi(X)$ and $\bar{H}(X)$ are positively correlated. The result follows. □

We have obtained, through Propositions 3 and 4, the distribution for the virtual sojourn time and the expression of the loss ratio, respectively. These allow us to specify the actual sojourn time distribution, through the next proposition. Here, while calculating the sojourn time distribution, we consider all the jobs, including those not admitted to the queue. The latter group is found to contribute a point mass at zero.

Proposition 5 *The steady-state distribution of the actual sojourn time T_a of an incoming job in an $M/M/1+G$ queue, operating under the FCFS-EAC policy, consists of a density $f_{T_\mathrm{a}}(\cdot)$ on $(0,\infty)$ given by*

$$f_{T_\mathrm{a}}(s) = \mu\pi \bar{H}(s)e^{B(s)-\mu s-B(0)}, \tag{22}$$

where π and $B(\cdot)$ are defined as in Proposition 2 and a point mass of α (given in Proposition 4) at 0.

Proof Let T and D be the virtual sojourn time and the relative deadline, respectively, of an incoming job, when the queue is in steady state. Then, the actual sojourn time of the job, $T_\mathrm{a} = TI_{\{T\le D\}}$, has the distribution function

$$\begin{aligned} P(T_\mathrm{a} \le s) &= 1 - \int_0^\infty f_T(u)P(uI_{\{u<D\}} > s)\mathrm{d}u \\ &= 1 - \int_s^\infty f_T(u)\bar{H}(u)\mathrm{d}u. \end{aligned} \tag{23}$$

We obtain Eq. (22) by differentiating both sides of (23). Since

$$\int_0^\infty f_{T_\mathrm{a}}(s)\mathrm{d}s < \int_0^\infty f_T(s)\mathrm{d}s = 1,$$

we conclude that T_a has a point mass at 0 given by (19). □

We now revert to the examples of relative deadline distributions considered in Sect. 3, and present the virtual sojourn time distribution, the actual sojourn time distribution and the loss ratio.

Example 1 (contd.) Degenerate distribution. The virtual sojourn time distribution has the density

$$f_T(s) = \begin{cases} \pi\mu e^{(\lambda-\mu)s-\rho(e^{\mu s}-1)e^{-\mu\theta}} & \text{if } s \le \theta, \\ \pi\mu e^{\lambda\theta-\mu s-\rho(1-e^{-\mu\theta})} & \text{if } s > \theta. \end{cases}$$

The actual sojourn time distribution consists of a density on $(0, \infty)$ satisfying

$$f_{T_{\mathrm{a}}}(s) = \pi\mu e^{(\lambda-\mu)s-\rho(e^{\mu s}-1)e^{-\mu\theta}} I_{\{s\leq\theta\}}$$

and a point mass at 0, which is the same as the loss ratio, given by

$$\alpha = \pi e^{\theta(\lambda-\mu)-\rho(1-e^{-\mu\theta})}. \tag{24}$$

Equation (24) is equivalent to Eq. (12) of [11]. □

Example 2 (contd.) Exponential distribution. The virtual sojourn time distribution has the density

$$f_T(s) = \pi\mu e^{\frac{\lambda\mu\theta}{(\mu+\frac{1}{\theta})} - \frac{\lambda\mu\theta e^{-\frac{s}{\theta}}}{(\mu+\frac{1}{\theta})} - \mu s},$$

over $[0, \infty)$. The actual sojourn time distribution consists of a density on $(0, \infty)$ satisfying

$$f_{T_{\mathrm{a}}}(s) = \pi\mu e^{\frac{\lambda\mu\theta}{(\mu+\frac{1}{\theta})} - \frac{\lambda\mu\theta e^{-\frac{s}{\theta}}}{(\mu+\frac{1}{\theta})} - (\mu+\frac{1}{\theta})s}$$

and a point mass at 0, namely the loss ratio, given by

$$\alpha = 1 - \pi\theta\left(\frac{(\mu+\frac{1}{\theta})}{\lambda\mu\theta}\right)^{1+\mu\theta} e^{\frac{\lambda\mu\theta}{(\mu+\frac{1}{\theta})}} \int_0^{\frac{\lambda\mu\theta}{(\mu+\frac{1}{\theta})}} x^{\mu\theta} e^{-x}\mathrm{d}x.$$

□

Example 3 (contd.) Uniform distribution. The virtual sojourn time T has the probability density function on $[0, \infty)$, defined by

$$f_T(s) = \begin{cases} \pi\mu e^{C_1(s)} & \text{if } s \leq 2\theta, \\ \pi\mu e^{C_2(s)} & \text{if } s > 2\theta; \end{cases}$$

and the actual sojourn time distribution consists of a density on $(0, \infty)$ of the form

$$f_{T_{\mathrm{a}}}(s) = \pi\mu\left(1 - \frac{z}{2\theta}\right) e^{C_1(s)} I_{\{s\leq 2\theta\}}$$

and a point mass at 0, namely the loss ratio, given by

$$\alpha = 1 - \pi\mu \int_0^{2\theta} \left(1 - \frac{s}{2\theta}\right) e^{C_1(s)}\mathrm{d}s,$$

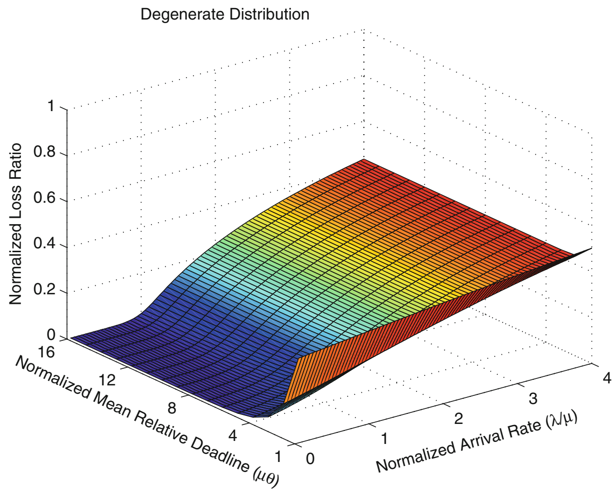

Fig. 1 Loss ratios for degenerate deadline distributions under the FCFC-EAC scheduling policy, for various values of normalized arrival rate (λ/μ) and normalized mean relative deadline ($\mu\theta$)

where

$$C_1(s) = \frac{\rho}{2\mu\theta}\left[(e^{\mu s} - 1)e^{-2\mu\theta} - s\mu\right] + (\lambda - \mu)s - \frac{s^2}{4\theta}$$

and

$$C_2(s) = \lambda\theta - \mu s - \rho\left[1 - \frac{1}{2\mu\theta}(1 - e^{-2\mu\theta})\right].$$

Figures 1, 2, and 3 show the loss ratios of the system for degenerate, exponential, and uniform deadline distributions, respectively, versus normalized arrival rate and normalized mean relative deadline. It is observed that loss ratio increases with the normalized arrival rate. For small normalized rates of arrival, the loss ratio is a decreasing function of the normalized mean relative deadline. This monotonicity prevails for all normalized arrival rates when the relative deadline is constant. This can also be shown analytically from (24) and (14). However, when the relative deadline has the exponential or uniform distribution and the normalized arrival rate is large, the loss ratio can increase with the normalized mean relative deadline, particularly for large values of the latter. A possible explanation of this apparently counter-intuitive phenomenon is that stochastic deadlines with large mean lead to admission of some jobs with very large service time, which in turn lead to non-admission of several relatively smaller jobs that could otherwise have been admitted. In order to investigate this matter, the mean service time of the admitted jobs was computed for the three deadline distributions considered here. The exact expressions for the mean service time are too messy to study the nature of the dependence on the mean deadline. However,

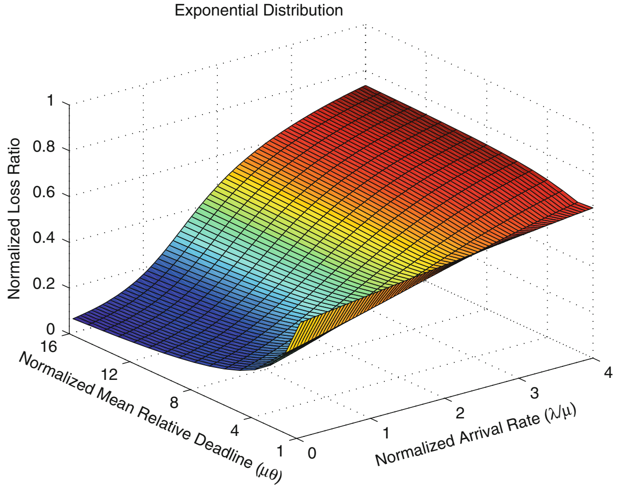

Fig. 2 Loss ratios for exponential deadline distributions under the FCFC-EAC scheduling policy, for various values of normalized arrival rate (λ/μ) and normalized mean relative deadline ($\mu\theta$)

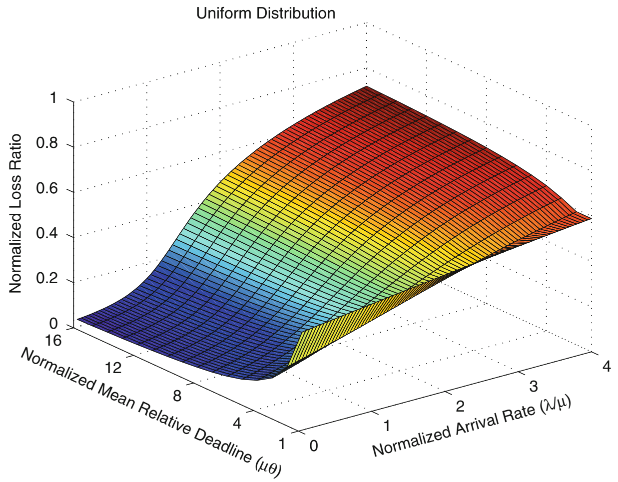

Fig. 3 Loss ratios for uniform deadline distributions under the FCFC-EAC scheduling policy, for various values of normalized arrival rate (λ/μ) and normalized mean relative deadline ($\mu\theta$)

numerical computations indicated that the mean service time of the admitted jobs increases with the mean deadline when the latter is large.

An interesting question that arises from Proposition 4 is, whether there is a distribution $H(\cdot)$ for which the loss ratio is minimized for a given normalized arrival rate and a normalized mean relative deadline. The expression given in Remark 1 can be useful in finding an answer to this question. Numerical studies indicated that the loss ratio is minimized when the relative deadline distribution is degenerate, even though this conjecture has not been proved analytically till now. It may be recalled that similar optimality of the degenerate distribution has been established in the case of FCFS systems with deadline for the end of service and FCFS-EAC systems with deadline for the beginning of service [17].

5 Deferred service decision

One generally needs an admission controller for implementation of the FCFS-EAC scheduling policy. The advantage of this controller is that a decision about the servicing of a job is taken immediately upon its arrival. In the case of FCFS systems, where generating an early decision is not crucial, one may dispense with the admission controller by deferring the decision till the epoch of a job reaching the server. The decision to accept or reject a job may be generated at that epoch by the server itself, depending on the service time of the job and its absolute deadline. Note that this arrangement may be viewed as merely a different implementation of EAC [6], as the set of serviced jobs remains the same. Even though the decision to service a job is generated at a later point of time, the advantage of this implementation is that an admission controller is not needed, as the decision can be taken by the server itself. In any case, sometimes it is only the server that knows the service time of a job [19]. Some researchers have also considered deferment of the service decision in the case of the earliest deadline first (EDF) scheduling policy [1,10].

As indicated above, for any realization of the queue operated under the FCFS-EAC scheduling policy, the actual service status of the jobs does not depend on the time of the service decision. Therefore, even if the service decision is delayed, the loss ratio would continue to be the same. Further, under the deferred decision implementation, the waiting time of a job is identical to the workload seen by it at arrival. However, the departure epochs of the rejected jobs under the two implementations are different, producing different sojourn time distributions. In the next proposition, we provide the sojourn time distribution of the jobs in an $M/M/1+G$ queue operated under the FCFS-EAC policy with deferred service decision.

Proposition 6 *The steady-state distribution of the sojourn time T_e of a job in an $M/M/1+G$ queue operating under the FCFS-EAC policy with deferred service decision, consists of a point mass at 0 and a density $f_{T_e}(\cdot)$ on $(0,\infty)$ given by*

$$P(T_e=0)=\pi\int_0^\infty \mu e^{-\mu y}H(y)\mathrm{d}y, \tag{25}$$

$$f_{T_e}(s)=f_T(s)\bar{H}(s)+\mu f_V(s)\int_0^\infty e^{-\mu y}H(s+y)\mathrm{d}y, \tag{26}$$

where π, $f_V(\cdot)$ and $f_T(\cdot)$ are defined as in Propositions 2 and 3, respectively.

Proof Let V, Y, and D be the waiting time, service time, and the relative deadline, respectively, of a job, when the queue is in steady state. Then, the sojourn time of a job, $T_e = V + YI_{\{V+Y\le D\}}$, has the distribution function

$$
\begin{aligned}
P(T_e \le s) &= 1 - P(s < V + Y \le D) - P(V > s, V + Y > D) \\
&= 1 - \int_s^\infty f_T(u)\bar{H}(u)\mathrm{d}u - \int_0^\infty \mu e^{-\mu y}\left(\int_s^\infty f_V(u)H(u+y)\mathrm{d}u\right)\mathrm{d}y.
\end{aligned} \tag{27}
$$

We obtain Eq. (26) by differentiating both sides of (27) and then simplifying. On the other hand,

$$P(T_e = 0) \;=\; P(V + YI_{\{V+Y\le D\}} = 0) \;=\; P(V + Y = 0 \le D) + \pi P(Y > D),$$

which simplifies to (25). □

6 Comparison between loss ratios of FCFS and FCFS-EAC

EAC is used as a means of ensuring guaranteed service of a job after its admission to the queue, however, one might expect that such an arrangement would also reduce the loss ratio of the system by eliminating the legacy of unproductive workload. Note that the FCFS-EAC policy entails full rejection (no server time allotted to a job that cannot be completed), as opposed to partial rejection in the case of FCFS alone. Thus, the loss ratio in the case of FCFS-EAC indicates the probability of a job not being admitted to the queue at all, whereas in the case of FCFS, the loss ratio is the probability of a job not being completed after possibly receiving some amount of service. The loss ratios of an $M/M/1$ system operated under these two policies are given in Eqs. (19) and (21).

In Figs. 4, 5, and 6, we provide the surface plots of the loss ratio of an $M/M/1$ system operated under FCFS-EAC, normalized by that of a similar system operated under FCFS, when the deadline distribution is degenerate, exponential, or uniform. We observe from the figures that, as expected from Corollary 1, the loss ratio under FCFS-EAC is indeed smaller than that under FCFS. In the next proposition, we show that such an order exist under a more general set-up. Here, we use the notations $\alpha^H_{\text{FCFS-EAC}}$ and α^H_{FCFS} to denote the loss ratios of the system, operated under the FCFS scheduling policy with and without EAC, respectively, and $H(\cdot)$ is the distribution function the relative deadline.

Proposition 7 *In a $G/G/1+G$ queue, the loss ratio under the FCFS scheduling policy can only be reduced when exact admission control is used, i.e.,*

$$\alpha^H_{\text{FCFS-EAC}} \le \alpha^H_{\text{FCFS}}.$$

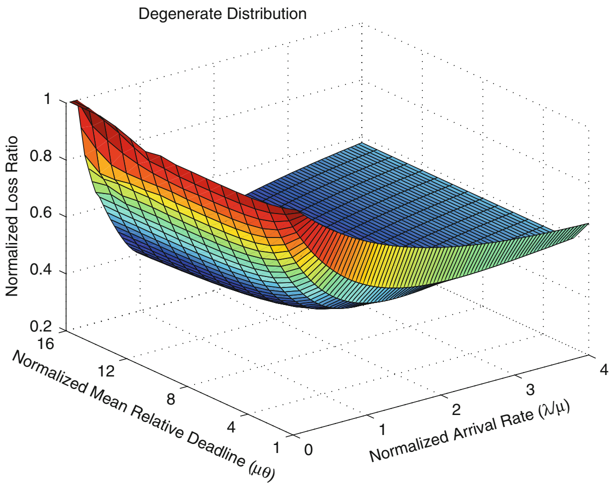

Fig. 4 Loss ratios for degenerate deadline distributions under the FCFS-EAC scheduling policy normalized by loss ratio under the FCFS scheduling policy, for various values of normalized arrival rate (λ/μ) and normalized mean relative deadline ($\mu\theta$)

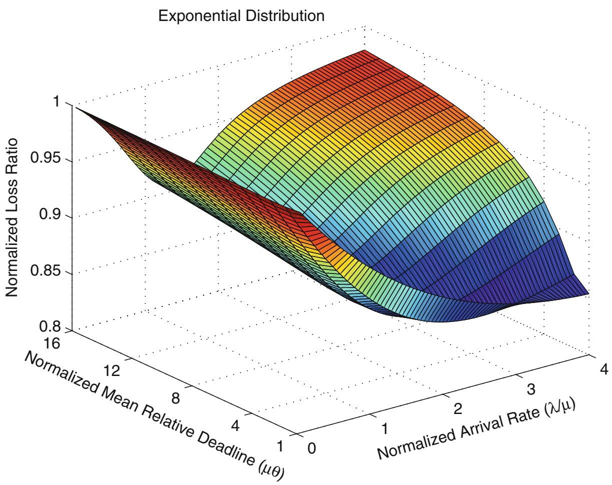

Fig. 5 Loss ratios for exponential deadline distributions under the FCFS-EAC scheduling policy normalized by loss ratio under the FCFS scheduling policy, for various values of normalized arrival rate (λ/μ) and normalized mean relative deadline ($\mu\theta$)

Proof As denoted in Sect. 2, let A_i, Y_i, and D_i be the arrival epoch, the service time, and the relative deadline of the ith job. If V_i is the workload upon arrival of the ith job under the FCFS scheduling policy, then

$$\alpha^H_{\text{FCFS}} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E[I_{\{V_i+S_i>D_i\}}]. \tag{28}$$

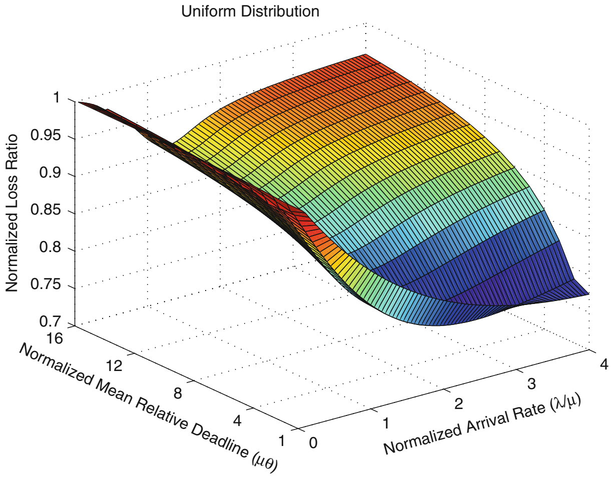

Fig. 6 Loss ratios for uniform deadline distributions under the FCFS-EAC scheduling policy normalized by loss ratio under the FCFS scheduling policy, for various values of normalized arrival rate (λ/μ) and normalized mean relative deadline ($\mu\theta$)

Likewise, if V_i^e is the workload upon arrival of the ith job under the FCFS-EAC scheduling policy, then

$$\alpha_{\text{FCFS-EAC}}^{H} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E[I_{\{(V_i^e+S_i>D_i\}}]. \tag{29}$$

Note that the workloads in the two cases follow the recursions

$$\begin{aligned} V_i = \big[&(V_{i-1}+Y_{i-1})I_{\{V_{i-1}+Y_{i-1}\le D_{i-1}\}} \\ &+(V_{i-1}\vee D_{i-1})I_{\{V_{i-1}+Y_{i-1}>D_{i-1}\}} - (A_i - A_{i-1})\big]\vee 0 \end{aligned} \tag{30}$$

$$\begin{aligned} V_i^e = \Big[&(V_{i-1}^e+Y_{i-1})I_{\{V_{i-1}^e+Y_{i-1}\le D_{i-1}\}} \\ &+V_{i-1}^e I_{\{V_{i-1}^e+Y_{i-1}>D_{i-1}\}} - (A_i - A_{i-1})\Big]\vee 0. \end{aligned} \tag{31}$$

We prove by induction that $V_i \geq V_i^e$ for all i. The inequality holds trivially for $i = 1$. Assuming that it holds for all indices up to $i-1$, we consider three cases: $V_{i-1} \leq D_{i-1} - Y_{i-1}$, $V_{i-1}^e \leq D_{i-1} - Y_{i-1} < V_{i-1}$, and $D_{i-1} - Y_{i-1} < V_{i-1}^e$.

In the first case,

$$\begin{aligned} V_i &= \big[(V_{i-1}+Y_{i-1}) - (A_i - A_{i-1})\big]\vee 0 \\ V_i^e &= \big[(V_{i-1}^e+Y_{i-1}) - (A_i - A_{i-1})\big]\vee 0. \end{aligned}$$

In the second case,

$$\begin{aligned} V_i &= \left[(V_{i-1} \vee D_{i-1}) - (A_i - A_{i-1})\right] \vee 0 \\ V_i^e &= \left[(V_{i-1}^e + Y_{i-1}) - (A_i - A_{i-1})\right] \vee 0. \end{aligned}$$

In the third case,

$$\begin{aligned} V_i &= \left[(V_{i-1} \vee D_{i-1}) - (A_i - A_{i-1})\right] \vee 0 \\ V_i^e &= \left[V_{i-1}^e - (A_i - A_{i-1})\right] \vee 0. \end{aligned}$$

In all the cases, we have $V_i \geq V_i^e$, which concludes the induction argument. The stated result follows from (28) and (29). □

Remark 3 When an $M/M/1+G$ system is operated under the FCFS scheduling policy, it follows from (21), by virtue of the dominated convergence theorem, that

$$\lim_{\lambda \to 0} \alpha_{\text{FCFS}}^H = 1 - \int_0^\infty \mu e^{-\mu s} \bar{H}(s) \mathrm{d}s.$$

Thus, when the arrival rate is very small, the loss ratio is the probability that the service time of a job exceeds its relative deadline. It has already been observed that the same conclusion holds in respect of an $M/M/1+G$ queue operated under the FCFS-EAC scheduling policy (see Remark 2). Therefore, Proposition 7 holds with equality when $\lambda \approx 0$. This fact explains why the ratio of $\alpha_{\text{FCFS-EAC}}^H/\alpha_{\text{FCFS}}^H$ plotted in Figs. 4, 5, and 6 approach the value 1 near the edge corresponding to $\lambda/\mu = 0$.

References

1. Abdelzaher, T., Lu, C.: Schedulability analysis and utilization bounds for highly scalable real-time services. In: IEEE Real-Time Technology and Applications Symp, pp. 11–25 (2001)
2. Baccelli, F., Boyer, P., Hebuterne, G.: Single-server queue with impatient customers. Adv. Appl. Probab. **16**(4), 887–905 (1984)
3. Barrer, D.Y.: Queueing with impatient customers and indifferent clerks. Oper. Res. **5**(5), 644–649 (1957)
4. Barrer, D.Y.: Queueing with impatient customers and ordered service. Oper. Res. **5**(5), 650–656 (1957)
5. Bekker, R.: Finite-buffer queues with workload-dependent service and arrival rates. Queueing Syst. **50**, 231–253 (2005)
6. Boxma, O.J., de Waal, P.R.: Multiserver queues with impatient customers. In: Proceedings of ITC, vol. 14, pp. 743–756. Elsevier, Amsterdam (1994)
7. Boxma, O., Perry, D., Stadje, W.: The M/G/1+G queue revisited. Queueing Syst. **67**, 207–220 (2011)
8. Cohen, J.W.: Single server queues with restricted accessibility. J. Eng. Math. **3**(4), 265–284 (1969)
9. Daley, D.J.: General customer impatience in queue GI/G/1. J. Appl. Probab. **2**(1), 186–205 (1965)
10. Das, S., Jenkins, L., Sengupta, D.: Loss ratios of EDF and FCFS scheduling policies under a performance enhancing modification. In: Proceedings of International Conference on Real-Time and Embedded Systems, pp. 128–131, RTES (2010)
11. Gavish, B., Schweitzer, P.: The Markovian queue with bounded waiting time. Manag. Sci. **23**(12), 1349–1357 (1977)

12. Haugen, R.B., Even, S.: Queueing systems with stochastic time out. IEEE Trans. Commun. **28**(12), 1984–1989 (1980)
13. Liu, L., Kulkarni, V.G.: Explicit solutions for the steady state distributions in M/PH/1 queues with workload dependent balking. Queueing Syst. **52**, 251–260 (2006)
14. Llamosi, A., Bernat, G., Burns, A.: Weakly hard real-time systems. IEEE Trans. Comput. **50**(4), 308–321 (2001)
15. Loris-Teghem, J.: On the waiting time distribution in a generalized queueing system with uniformly bounded sojourn times. J. Appl. Probab. **9**, 642–649 (1972)
16. Movaghar, A.: On queueing with customer impatience until the beginning of service. Queueing Syst. **29**, 337–350 (1998)
17. Movaghar, A.: On queueing with customer impatience until the end of service. Stoch. Models **22**, 149–173 (2006)
18. Palm, C.: Methods for judging the annoyance caused by congestion. Tele **2**, 1–20 (1953)
19. Perry, D., Asmussen, S.: Rejection rules in the M/G/1 queue. Queueing Syst. **19**, 105–130 (1995)
20. Van Dijk, N.M.: Queueing systems with restricted workload: an explicit solution. J. Appl. Probab. **27**, 393–400 (1990)
21. Zwart, A.P.: Loss rates in the M/G/1 queue with complete rejection. Technische Universiteit Eindhoven, Technical report (2003)