View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Ghent University Academic Bibliography

Performance analysis of a discrete-time queueing system with customer deadlines

Herwig Bruneel and Tom Maertens
Stochastic Modelling and Analysis of Communication Systems Research Group (SMACS)
Department of Telecommunications and Information Processing (TELIN)
Ghent University (UGent)
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Email: {hb,tmaerten}@telin.UGent.be

***Abstract*—This paper studies a discrete-time queueing system where each customer has a maximum allowed sojourn time in the system, referred to as the "deadline" of the customer. Deadlines of consecutive customers are modelled as independent and geometrically distributed random variables. The arrival process of new customers, furthermore, is assumed to be general and independent, while service times of the customers are deterministically equal to one slot each. For this queueing model, we are able to obtain exact formulas for quantities as the mean system content, the mean customer delay, and the deadline-expiration ratio. These formulas, however, contain infinite sums and infinite products, which implies that truncations are required to actually compute numerical values. Therefore, we also derive some easy-to-evaluate approximate results for the main performance measures. These approximate results are quite accurate, as we show in some numerical examples. Possible applications of this type of queueing model are numerous: the (variable) deadlines could model, for instance, the fact that customers may become impatient and leave the queue unserved if they have to wait too long in line, but they could also reflect the fact that the service of a customer is not useful anymore if it cannot be delivered soon enough, etc.**

Keywords*—*queueing; discrete-time; deadlines; closed-form results; power-series approximation

I. Introduction

In a typical queueing model, customers present themselves near some service facility to receive some kind of service, and – if they cannot be served immediately upon arrival – wait patiently in a queue until the server is available for them. In some cases, however, customers may leave the queue unserved if their time in the queue becomes too big. There may be various reasons for such behaviour. One of them is customer impatience [9], [10], [17], [19]–[21], [24] (for instance, when trying to reach a call center [1]). In this case, it is the customer that takes the decision to abandon prematurely, e.g., because the customer (usually, a human being in this case) does not like to wait any longer or because the customer has other tasks to attend. On the other hand, also the system itself may decide to remove customers from the queue if servicing those customers is deemed not to be useful anymore after some time in the queue, e.g., because those customers (audio or video streaming packets in a telecommunications network [5], [13], [15], for instance) would not arrive soon enough at their next destination (a playout buffer [7], [12], [14], [16], [18], [23], for instance) if they had to wait any longer. In either case, the queueing system is "special" in the sense that customers may disappear from the system without ever reaching the service facility.

We start with the analysis of the queueing performance of the system. This results in an exact, yet complicated expression for the probability generating function (pgf) of the number of customers in the system. From this pgf, we derive both exact and approximate expressions for the mean system content, the mean customer delay and the deadline-expiration ratio. Although this approach demonstrates that a queueing analysis based on pgfs also seems very suitable for studying this type of queueing models (i.e., models with customer deadlines), the main drawback of the obtained results is that they are expressed in terms of infinite sums and products which have to be truncated in order to obtain numerical results.

Therefore, we also propose an alternative approach. In particular, we express all relevant performance measures in the form of power series, not, as usual, in terms of the load or the traffic intensity of the system (see, e.g., [2], [8]) but as functions of the deadline parameter. In this way, we arrive at approximate, yet much simpler expressions than with the "exact" method. We show with some numerical examples that the approximate expressions are quite accurate. Moreover, they are much more suitable to study the influence of the deadline parameter. In our opinion, this alternative approach is the main contribution of this paper and distinguishes the paper of the other literature on customer deadlines. The approach leads to accurate approximations, so it may be useful in the solution of more advanced queueing models with customer deadlines. It should be noted that the approach with power series in a parameter other than the load has also proven to be useful in other types of queueing models, e.g., a Generalized Processor Sharing model [22] and a model with train arrivals [4].

II. Mathematical model

We consider a *discrete-time* queueing system with one server and an infinite waiting room. As in all discrete-time models, the time axis is divided into fixed-length intervals referred to as *slots*. New customers may enter the system at any given (continuous) point on the time axis, but services are synchronized to (i.e., can only start and end at) slot boundaries. We assume that the service of each customer requires exactly one slot. The arrival process of new customers in the system is characterized by means of a sequence of i.i.d. non-negative discrete random variables with common pgf $E(z)$, i.e.,

$$E(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{e_k}\right], \tag{1}$$

with e_k denoting the number of arrivals during slot k. The mean number of customer arrivals per slot, in the sequel referred to as the *(mean) arrival rate*, is defined as λ: $\lambda \triangleq E'(1)$.

As mentioned above, the special feature of the queueing model at hand is the fact that customers may leave the system before they have actually received service. Here we make a distinction between the *queue*, which collects the customers that are actually waiting for service, and the *system*, which encompasses all the customers, also the customer in service. Customers that have entered the server, possibly after having spent some time in the queue, stay in the system until their service ends. However, no customer waits in the queue longer than a prescribed maximum time duration, referred to as the *waiting deadline* of the customer. We assume that the waiting deadlines of the customers may be different from one customer to another, but that they are statistically independent and geometrically distributed with parameter σ. So their pgf is given by

$$S(z) = \frac{(1-\sigma)z}{1-\sigma z}, \tag{2}$$

while their mean value equals $S'(1) = 1/(1-\sigma)$. In the remainder, the mean waiting deadline is represented by D. The geometric nature of the deadlines implies that the probability that the deadline of a waiting customer expires at the end of a slot, does not depend on the amount of time the customer already spent in the queue and is simply given by σ. This property is crucial in the analysis of the system.

It should be noted that the literature on queueing systems with customer deadlines makes a distinction between waiting deadlines and *residence deadlines*. Waiting deadlines are defined as deadlines until the beginning of service (i.e., once a customer enters the server, it stays in the system until its service ends), whereas residence deadlines are interpreted as deadlines until the end of service (i.e., a customer may also leave the system during service due to deadline expiration). Since we assume one-slot service times, however, we do not have to make a distinction here. Therefore, we can just talk about "deadline" instead of "waiting deadline" in the sequel.

The structure of the rest of this paper is as follows. In section III, we analyze the queueing performance of the system. Section IV is devoted to an alternative approach in which we express all relevant performance measures of the system in the form of power series in the parameter σ. In section V, we discuss the theoretical results and we compare the different approximations by means of some numerical examples. Section VI finally states some conclusions and indicates some possible future work.

III. Performance analysis

Let u_k denote the system content, i.e., the number of customers present in the system, at the beginning of the slot k. Then the following recursive system equation can be established:

$$u_{k+1} = \sum_{i=1}^{(u_k-1)^+} a_{i,k} + e_k, \tag{3}$$

with $(\ldots)^+$ indicating the quantity $\max(0,\ldots)$. In equation (3), the $a_{i,k}$s are a sequence of i.i.d. Bernoulli random variables with parameter σ, i.e., with common pgf $A(z) = 1-\sigma+\sigma z$. $a_{i,k}$ can be interpreted as the indicator function (taking values 1 or 0) of the event that the i-th customer in the queue at the beginning of slot k does not leave the queue unserved at the end of slot k.

For all k, let $U_k(z)$ denote the pgf of u_k. Then, from equation (3), we can derive

$$U_{k+1}(z) = E(z) \cdot \mathrm{E}\left[z^{\sum_{i=1}^{(u_k-1)^+} a_{i,k}}\right], \tag{4}$$

with $\mathrm{E}[\cdot]$ the expectation operator. The second factor in the right hand side of (4) can be expanded further by means of the law of total probability, yielding

$$U_{k+1}(z) = E(z)\left[\mathrm{Prob}[u_k=0] + \sum_{j=1}^{\infty} \mathrm{Prob}[u_k=j]\, A^{j-1}(z)\right], \tag{5}$$

or, equivalently,

$$U_{k+1}(z) = E(z)\left[U_k(0) + \frac{U_k(A(z)) - U_k(0)}{A(z)}\right]. \tag{6}$$

Now, let us assume that the queueing system at hand is stable. In fact, it is not difficult to see that the system is always stable if the parameter σ is strictly less than 1, because in that case the deadlines are finite (with probability 1) and, hence, also the sojourn times of the customers in the system are necessarily finite. On the other hand, if $\sigma = 1$, the system reduces to a simple discrete-time buffer without deadlines, which is stable if and only if the mean number of customers entering the system per slot, given by λ, is strictly less than 1. We now let the time parameter k go to infinity in equation (6). Assuming the system reaches a steady state, then both functions $U_k(\cdot)$ and $U_{k+1}(\cdot)$ converge to a common limit function $U(\cdot)$, which denotes the pgf of the system content at the beginning of an arbitrary slot in steady state. As a result, equation (6) translates into

$$U(z) = F(z)\left[U(A(z)) + (A(z)-1)U(0)\right], \tag{7}$$

where $F(z) = E(z)/A(z)$.

We are now faced with the problem of solving the (non-classical) functional equation (7). If $\sigma = 1$, this is very simple, because in that case $A(z) = z$ and (7) is, in fact, a simple linear equation for $U(z)$ with the well-known [3] solution

$$U(z) = \frac{(1-\lambda)(z-1)E(z)}{z - E(z)}. \tag{8}$$

If $\sigma < 1$, however, the problem is less trivial. One way to

proceed is to use equation (7) recursively, as follows:

$$\begin{aligned} U(z) =& \; F(z)\left[U(1-\sigma+\sigma z)+\sigma(z-1)U(0)\right] \\ =& \; F(z)F(1-\sigma+\sigma z)U(1-\sigma^2+\sigma^2 z) \\ & + F(z)F(1-\sigma+\sigma z)\sigma^2(z-1)U(0) \\ & + F(z)\sigma(z-1)U(0) \\ =& \; \cdots \\ =& \; U(1)\prod_{i=0}^{\infty} F(1-\sigma^i+\sigma^i z) \\ & + U(0)\sum_{i=0}^{\infty}\sigma^{i+1}(z-1)\prod_{j=0}^{i} F(1-\sigma^j+\sigma^j z), \quad (9) \end{aligned}$$

where we have used the fact that $\lim_{n\to\infty}\sigma^n = 0$ for $\sigma < 1$. In the above result, $U(1) = 1$ (normalization) and $U(0)$ can be obtained by choosing $z = 0$ and solving the resulting equation for U(0). This leads to

$$U(0) = \frac{\prod_{i=0}^{\infty} F(1-\sigma^i)}{1+\sum_{i=1}^{\infty}\sigma^i\prod_{j=0}^{i-1}F(1-\sigma^j)}. \quad (10)$$

Insertion of (10) in (9) finally yields an explicit expression for the pgf $U(z)$.

In principle, various moments of the system-content distribution can be obtained by computing derivatives of this expression at $z = 1$. This, however, results in expressions containing both infinite sums and infinite products; expression (10) for the quantity $U(0)$ suffers from the same inconvenience. We have found that, in practice, the easiest way to circumvent these difficulties is to truncate the infinite sum and product in the expression of $U(0)$ as follows:

$$U(0) \approx U_K(0) \triangleq \frac{\prod_{i=0}^{K-1} F(1-\sigma^i)}{1+\sum_{i=1}^{K-1}\sigma^i\prod_{j=0}^{i-1}F(1-\sigma^j)}, \quad (11)$$

where the integer K is such that $1-\sigma^K$ is "close enough" to 1. The mean system content $\mathrm{E}[u]$ can be computed from $U(0)$, departing from the functional equation (7) rather than from the explicit expression of $U(z)$. Indeed, by taking first derivatives at $z = 1$ of both sides of (7), we find that

$$U'(1) = F'(1) + U'(1)A'(1) + A'(1)U(0), \quad (12)$$

which easily leads to

$$\mathrm{E}[u] = \frac{\lambda - \sigma(1-U(0))}{1-\sigma}. \quad (13)$$

By applying (the discrete-time version of) Little's theorem (see, e.g., [3], [6], [11]), the mean delay (system time) $\mathrm{E}[d]$ of a customer can be obtained as $\mathrm{E}[u]/\lambda$. We get that

$$\mathrm{E}[d] = \frac{\lambda - \sigma(1-U(0))}{\lambda(1-\sigma)}. \quad (14)$$

In terms of the mean deadline $D = 1/(1-\sigma)$, this gives

$$\mathrm{E}[d] = D - \frac{(D-1)(1-U(0))}{\lambda}. \quad (15)$$

Equation (15) clearly illustrates that the mean delay of a customer cannot be higher than the mean deadline D, as expected intuitively.

Another quantity of interest in the context of a queueing system with deadlines is the fraction of customers that leave the queue unserved due to the expiration of their deadline. We call this fraction the *deadline-expiration ratio* in the sequel. It can be computed as

$$r_{\mathrm{ex}} = \frac{\lambda - (1-U(0))}{\lambda}, \quad (16)$$

where the numerator corresponds to the mean number of customers leaving the queue unserved per slot, i.e., the difference between the mean number of arrivals in a slot (λ) and the mean number of customers receiving service per slot ($1-U(0)$).

As soon as the quantity $U(0)$ has been computed, the other performance measures, $\mathrm{E}[u]$, $\mathrm{E}[d]$, and r_{ex}, can be easily obtained from (13), (15), and (16), respectively. In particular, if $U(0)$ is approximated by $U_K(0)$ (see equation (11)), we refer to the corresponding approximations of $\mathrm{E}[u]$, $\mathrm{E}[d]$ and r_{ex} as $\mathrm{E}[u]_K$, $\mathrm{E}[d]_K$ and $r_{\mathrm{ex},K}$.

IV. Power-series approximation

The main drawback of the results obtained so far is that they are expressed in terms of infinite sums and products which have to be truncated in order to obtain numerical results. In this section, we take an alternative approach to arrive at easily computable formulas. Specifically, we aim for a representation of the pgf $U(z)$ (and the performance measures derived from it) in the form of a *power series* in the parameter σ:

$$U(z) = \sum_{i=0}^{\infty}\sigma^i V_i(z), \quad (17)$$

where the functions $V_i(z)$ are independent of σ. This so-called 'power series approximation' (PSA) technique was introduced in [8] and initially expressed system characteristics as functions of the load. Our approach differs from that conventional approach in that we construct a power series in σ rather than in λ.

Instead of trying to solve the functional equation (7) for the pgf $U(z)$, we now focus on the derivation of the functions $V_i(z)$, for $i = 0, 1, 2, 3, \cdots$. In order to do so, we first determine series expansions for all the quantities appearing in (7). We know that $A(z) = 1+\sigma(z-1)$. Next, the quantity $U(A(z))$ can be written as

$$\begin{aligned} U(A(z)) =& U(1+\sigma(z-1)) \\ =& U(1) + U'(1)\sigma(z-1) + U''(1)\frac{(\sigma(z-1))^2}{2} \\ & + U'''(1)\frac{(\sigma(z-1))^3}{6} + \cdots, \quad (18) \end{aligned}$$

Then by introducing the expansion (17), we get that

$$\begin{aligned} U(A(z)) \approx& V_0(1) + \sigma V_1(1) + \sigma^2 V_2(1) + \sigma^3 V_3(1) \\ & + \sigma(z-1)\left[V_0'(1) + \sigma V_1'(1) + \sigma^2 V_2'(1)\right] \\ & + \frac{\sigma^2(z-1)^2}{2}\left[V_0''(1) + \sigma V_1''(1)\right] \\ & + \frac{\sigma^3(z-1)^3}{6}V_0'''(1). \quad (19) \end{aligned}$$

Now, expanding both sides of the equation (7) in powers of σ leads to

$$\begin{aligned}[1+\sigma(z-1)][V_0(z)+\sigma V_1(z)+\sigma^2V_2(z)+\sigma^3V_3(z)]\\ \approx\ E(z)[V_0(1)+\sigma V_1(1)+\sigma^2V_2(1)+\sigma^3V_3(1)]\\ +\sigma(z-1)E(z)[V_0'(1)+\sigma V_1'(1)+\sigma^2V_2'(1)]\\ +\frac{\sigma^2(z-1)^2}{2}E(z)[V_0''(1)+\sigma V_1''(1)]\\ +\frac{\sigma^3(z-1)^3}{6}E(z)V_0'''(1)\\ +\sigma(z-1)E(z)\\ \times[V_0(0)+\sigma V_1(0)+\sigma^2V_2(0)+\sigma^3V_3(0)].\end{aligned} \tag{20}$$

It is clear that, for normalization,

$$U(1)=V_0(1)+\sigma V_1(1)+\sigma^2V_2(1)+\sigma^3V_3(1)+\cdots=1, \tag{21}$$

for all σ. Consequently, $V_0(1)=1$ and $V_i(1)=0$ for $i>0$.

We now identify the coefficients of equal powers of σ on both sides of equation (20) to determine explicit expressions for the functions $V_i(z)$, for $i=0,1,2,3$. With the coefficients of σ^0, we easily find that

$$V_0(z)=E(z). \tag{22}$$

Next, for σ^1, we initially obtain

$$\begin{aligned}V_1(z)+(z-1)V_0(z)\\ =E(z)[V_1(1)+\lambda(z-1)+E(0)(z-1)].\end{aligned} \tag{23}$$

Since $V_0(z)=E(z)$ and $V_1(1)=0$, this results in

$$V_1(z)=(z-1)E(z)[\lambda-1+E(0)]. \tag{24}$$

Identifying coefficients of σ^2 leads to

$$\begin{aligned}V_2(z)+(z-1)V_1(z)\\ =E(z)\Bigg[V_2(1)+(z-1)V_1'(1)\\ +\frac{(z-1)^2}{2}V_0''(1)+(z-1)V_1(0)\Bigg].\end{aligned} \tag{25}$$

From (22)-(24), we find that

$$\begin{aligned}V_1'(1)&=\lambda-1+E(0),\\ V_0''(1)&=E''(1),\qquad\text{and}\\ V_1(0)&=-E(0)[\lambda-1+E(0)],\end{aligned} \tag{26}$$

respectively. $V_2(1)=0$, so this yields

$$\begin{aligned}V_2(z)=&(z-1)E(z)\Bigg\{[\lambda-1+E(0)][2-E(0)-z]\\ &+\frac{E''(1)}{2}(z-1)\Bigg\}.\end{aligned} \tag{27}$$

In a similar way, by identifying the coefficients of σ^3, we finally produce that

$$\begin{aligned}V_3(z)=&(z-1)E(z)\\ &\times\Bigg\{[\lambda-1+E(0)][1-3E(0)+E^2(0)]\\ &+(z-1)[\lambda-1+E(0)][\lambda-2+E(0)+z]\\ &+\frac{(z-1)^2}{6}[E'''(1)-3E''(1)]+\frac{E(0)E''(1)}{2}\Bigg\}.\end{aligned} \tag{28}$$

Combining the results in equations (22), (24), (27), and (28), we now dispose of the following explicit approximate expression ($\hat{U}(z)$) for the pgf $U(z)$:

$$\hat{U}(z)\triangleq V_0(z)+\sigma V_1(z)+\sigma^2V_2(z)+\sigma^3V_3(z). \tag{29}$$

Equation (29) can now be used to derive explicit closed-form expressions for various performance measures of the queueing system at hand, in terms of the basic system parameters, (i.e., the pgf $E(z)$ of the arrival process and the probability σ which characterizes the deadline distribution). First, we derive from (29) an approximation $\hat{U}(0)$ for the probability of an empty system:

$$\begin{aligned}\hat{U}(0)=&\ V_0(0)+\sigma V_1(0)+\sigma^2V_2(0)+\sigma^3V_3(0)\\ =&\ E(0)+\sigma E(0)[1-\lambda-E(0)]\\ &+\sigma^2E(0)\Bigg\{[1-\lambda-E(0)][2-E(0)]+\frac{E''(1)}{2}\Bigg\}\\ &+\sigma^3E(0)\Bigg\{[1-\lambda-E(0)][3-\lambda-4E(0)+E^2(0)]\\ &-\frac{E(0)E''(1)}{2}+\frac{3E''(1)-E'''(1)}{6}\Bigg\}.\end{aligned} \tag{30}$$

Next, we derive three different approximate expressions for the mean system content $\mathrm{E}[u]$, referred to as $\mathrm{E}[\hat{u}_1]$, $\mathrm{E}[\hat{u}_2]$, and $\mathrm{E}[\hat{u}_3]$, respectively. The first approximation is obtained directly from the approximate expression (29) of $U(z)$:

$$\begin{aligned}\mathrm{E}[\hat{u}_1]\triangleq&\hat{U}'(1)\\ =&\lambda-\sigma[1-\lambda-E(0)]-\sigma^2[1-\lambda-E(0)][1-E(0)]\\ &-\sigma^3\Bigg\{[1-\lambda-E(0)][1-3E(0)+E^2(0)]\\ &-\frac{E(0)E''(1)}{2}\Bigg\}.\end{aligned} \tag{31}$$

The second approximation is obtained by replacing $U(0)$ by $\hat{U}(0)$ in the exact equation (13):

$$E[\hat{u}_2]\triangleq\frac{\lambda-\sigma(1-\hat{U}(0))}{1-\sigma}. \tag{32}$$

Finally, inspired by the results of some numerical examples (see next section), we define a third (heuristic) approximation as the arithmetic mean of the first two approximations:

$$E[\hat{u}_3]\triangleq\frac{E[\hat{u}_1]+E[\hat{u}_2]}{2}. \tag{33}$$

Corresponding approximations for the mean customer delay are given by $\mathrm{E}\left[\hat{d}_i\right]\triangleq\mathrm{E}[\hat{u}_i]/\lambda$ $(i=1,2,3)$. In view of

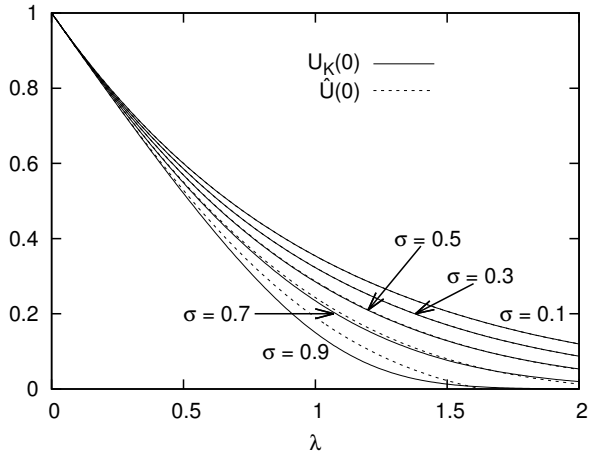

Fig. 1. Probability of empty system $U_K(0)$ and $\hat{U}(0)$ versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

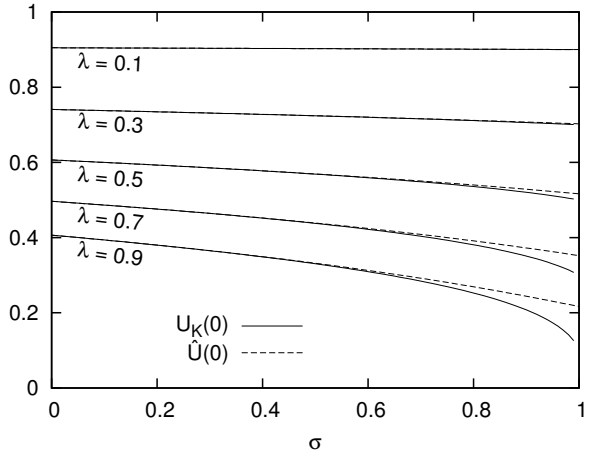

Fig. 2. Probability of empty system $U_K(0)$ and $\hat{U}(0)$ versus deadline-distribution parameter σ, for Poisson arrivals and various values of the mean arrival rate λ

equation (16), an approximation for the deadline-expiration ratio can be computed as

$$\hat{r}_{ex} \triangleq \frac{\lambda - (1 - \hat{U}(0))}{\lambda} \quad . \tag{34}$$

V. Discussion of results and numerical examples

In this section, we discuss the results obtained in the previous sections, both from a qualitative perspective and by means of some numerical examples. In particular, we also validate the approximate power-series results against more accurate results, obtained by truncation of infinite sums and products.

Let us consider the (common) case of Poisson arrivals, i.e., $E(z) = e^{\lambda(z-1)}$. Other choices of the arrival distribution, such as a geometric distribution or a binomial distribution are also possible and even lead to simpler formulas than the Poisson distribution, because the corresponding pgf $E(z)$ in these cases is a rational function of z rather than a transcendental function in the Poisson case. We do not further discuss such

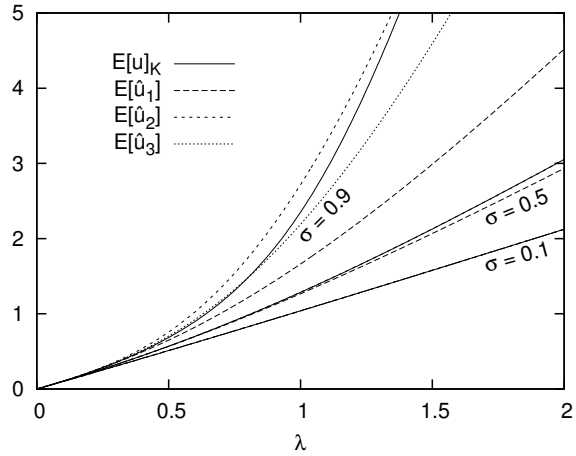

Fig. 3. Mean system content $E[u]_K$, $E[\hat{u}_1]$, $E[\hat{u}_2]$ and $E[\hat{u}_3]$, versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

choices here because they basically lead to the same qualitative conclusions on the system behavior as the Poisson assumption.

In Fig. 1, we have plotted the "exact" result $U_K(0)$ (with $K = 2000$) and the approximation $\hat{U}(0)$ for the probability of an empty system, according to formulas (11) and (30) respectively, versus the arrival rate λ, for various values of the deadline-distribution parameter σ. The figure shows that the probability of an empty system decreases when the arrival rate increases, as expected. It also shows that this probability decreases more slowly when the deadlines get smaller, i.e., when σ decreases, which can be attributed to the fact that more customers leave the queue prematurely. Note that, owing to the finite deadlines of the customers, the queue remains empty for $\lambda > 1$ with a positive probability, even though in this case more customers arrive per slot than the server can handle. The figure finally also illustrates the accuracy of the power-series approximation $\hat{U}(0)$, which is very good for all values of λ, as long as σ is not too high (say, $\sigma < 0.75$).

Very similar conclusions can be drawn from Fig. 2, where we have plotted $U_K(0)$ (with $K = 2000$) and $\hat{U}(0)$ versus the deadline-distribution parameter σ, for various values of the arrival rate λ. This figure illustrates very clearly that the probability of an empty system decreases when the deadlines get longer (i.e., for higher values of σ), because the number of customers that leave the queue prematurely goes down in these circumstances. The accuracy of the power-series approximation for low values of λ is also very striking.

Fig. 3 shows the "exact" results $E[u]_K$ (again, for $K = 2000$), and our three power-series approximations $E[\hat{u}_1]$, $E[\hat{u}_2]$ and $E[\hat{u}_3]$, for the mean system content, versus the mean arrival rate λ, for various values of the deadline-distribution parameter σ. As expected, all the curves increase with λ. The figure also makes clear that, for a given arrival rate λ, the mean system content increases when the deadlines become longer, i.e., when the parameter σ takes higher values. Again, we note that the system remains stable for $\lambda > 1$, due to the finite length of the deadlines (if $\sigma < 1$). In the most interesting region of the graph, i.e., where the mean arrival rate λ is smaller than 1, we observe that the power-series

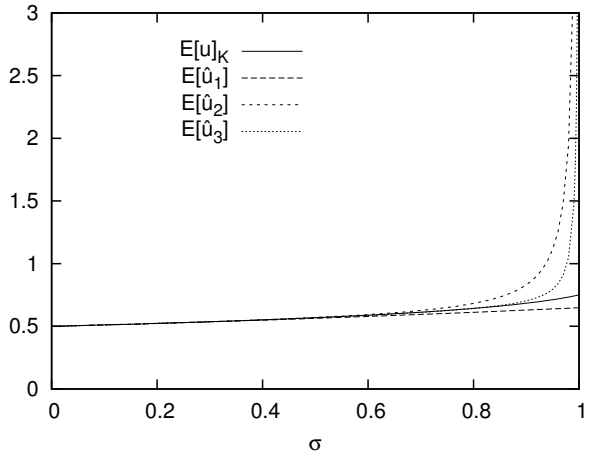

Fig. 4. Mean system content $E[u]_K$, $E[\hat{u}_1]$, $E[\hat{u}_2]$ and $E[\hat{u}_3]$, versus the deadline-distribution parameter σ, for Poisson arrivals with $\lambda = 0.5$

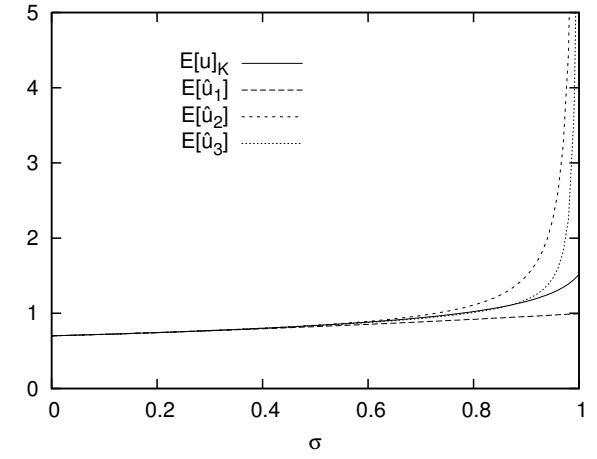

Fig. 5. Mean system content $E[u]_K$, $E[\hat{u}_1]$, $E[\hat{u}_2]$ and $E[\hat{u}_3]$, versus the deadline-distribution parameter σ, for Poisson arrivals with $\lambda = 0.7$

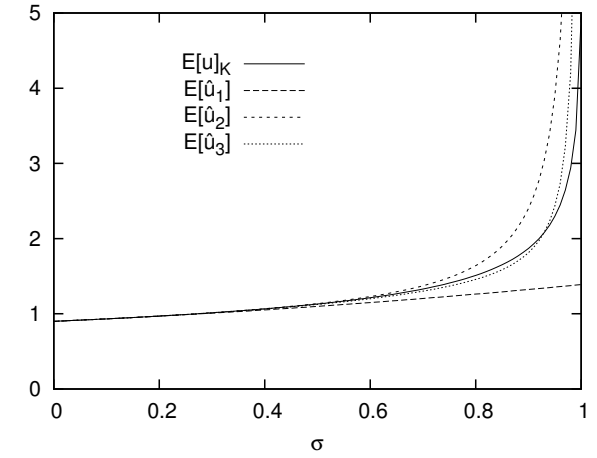

Fig. 6. Mean system content $E[u]_K$, $E[\hat{u}_1]$, $E[\hat{u}_2]$ and $E[\hat{u}_3]$, versus the deadline-distribution parameter σ, for Poisson arrivals with $\lambda = 0.9$

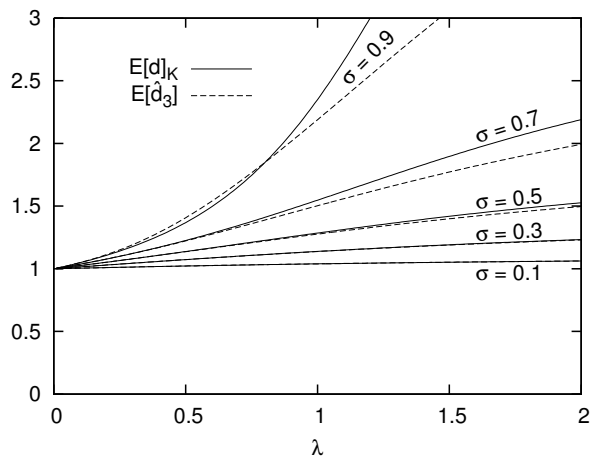

Fig. 7. Mean customer delay $E[d]_K$ and $E\left[\hat{d}_3\right]$, versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

approximations $E[\hat{u}_1]$ and $E[\hat{u}_2]$ are quite close as long as $\sigma < 0.5$, but the third approximation $E[\hat{u}_3]$ is even good for such "high" values of σ as 0.9. This is also very clearly illustrated in Figs. 4, 5 and 6, where we have plotted the "exact" and approximate values of the mean system content as functions of the deadline parameter σ, for three different values of λ (less than 1). In all cases, $E[\hat{u}_3]$ turns out to be the best approximation for all $\sigma < 0.9$. For this reason, we only use the third power-series approximation in the sequel.

The approximate mean customer delay $E\left[\hat{d}_3\right]$ is compared with the "exact" values $E[d]_K$ in Fig. 7. Again, the accuracy of the approximation turns out to be very good in the region $\lambda < 1$ for all displayed values of σ. As expected, the mean delay increases with the arrival rate λ. Also, the mean delay is kept smaller (for all relevant values of λ) as the mean deadline of the customers decreases. Specifically, for the five curves in Fig. 7, we have that the mean deadline $D = 1/(1-\sigma)$ takes values 10, 3.33, 2, 1.42 and 1.11 for $\sigma = 0.9$, $\sigma = 0.7$, $\sigma = 0.5$, $\sigma = 0.3$, and $\sigma = 0.1$, respectively. It is very clear that the curves for the mean customer delay stay well below these mean deadlines, for all values of the arrival rate λ.

Finally, some results for the deadline-expiration ratio are shown in Fig. 8. As the deadline-expiration ratio is computed directly from the probability of an empty system (see equations (16) and (34)), we expect the power-series approximation $\hat{r}_{ex}$ to be accurate as long as σ is not too high (say, $\sigma < 0.75$). This is indeed confirmed by Fig. 8. Furthermore, the figure reveals that, for a given deadline-distribution parameter $\sigma < 1$, the fraction of customers that leave the queue unserved grows steadily with the arrival rate λ. An intuitive explanation of this observation is not so obvious, in view of the fact that in a system without deadlines (i.e., $\sigma = 1$) the deadline-expiration ratio is constant and equal to zero for all values of λ, either smaller than 1 (stable system) or larger than 1 (instable system). In a system with deadlines ($\sigma < 1$), we expect the deadline-expiration ratio to increase with λ when $\lambda > 1$, because in this case the server cannot handle more than 1 customer per slot, which implies that at least $\lambda - 1$ customers per slot leave the system prematurely. Perhaps more surprisingly, according to Fig. 8, the deadline-expiration ratio also grows with λ in the region $\lambda < 1$, which means that the fraction of customers that do get served before they leave the

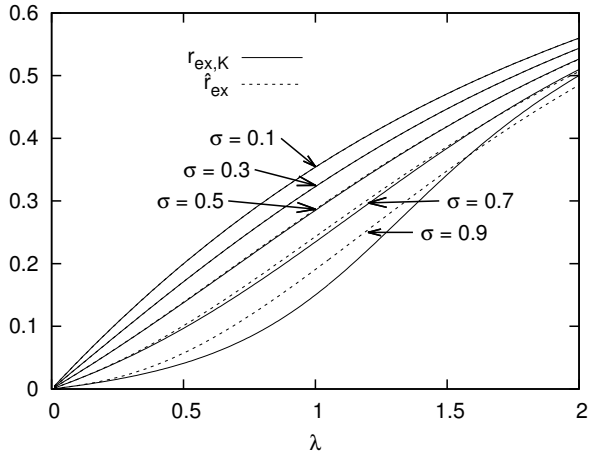

Fig. 8. Deadline-expiration ratio $r_{ex,K}$ and $\hat{r}_{ex}$, versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

system decreases when λ increases. A possible explanation for this behavior lies in the fact that for increasing λ the length of the queue grows and customers are more likely to reach their deadline while waiting.

VI. Conclusions and future work

This paper has examined a relatively simple model for a discrete-time single-server queueing system in which customers are subjected to deadlines. We have been able to derive nearly exact but complicated formulas, as well as simpler approximate formulas for the main performance measures of the system. From the methodological point of view, we believe that one of the main contributions of our paper lies in the power-series approximation method that we have developed in section 3, a technique that may be useful in the solution of other queueing models that lead to hard-to-solve functional equations such as equation (7). In terms of numerical results, we have been able to explain most of the observed dependencies between performance measures and system parameters intuitively.

The main restriction of this work seems to be the assumption that the service times of the customers are *deterministically* equal to one slot each and that deadlines of the customers are *i.i.d.* and *geometrically distributed*. Future work will focus on generalizations of these assumptions.

Acknowledgment

This research has been co-funded by the Interuniversity Attraction Poles (IAP) Programme initiated by the Belgian Science Policy Office.

References

[1] T. Aktekin and R. Soyer. Bayesian analysis of queues with impatient customers: Applications to call centers. *Naval Research Logistics*, 59(6):441–456, 2012.

[2] J.P.C. Blanc. On a numerical method for calculating state probabilities for queueing systems with more than one waiting line. *Journal of Computational and Applied Mathematics*, 20:119–125, 1987.

[3] H. Bruneel and B.G. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic, Boston, USA, 1993.

[4] K. De Turck, D. Fiems, S. Wittevrongel, and H. Bruneel. A Taylor series expansions approach to queues with train arrivals. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, pp. 447-455, 2011.

[5] M. Feldman and J. Naor. Non-Preemptive Buffer Management for Latency Sensitive Packets. In *2010 Proceedings IEEE INFOCOM*, 2010.

[6] D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30:17–18, 2002.

[7] T. Hofkens, K. Spaey, and C. Blondia. Transient analysis of the D-BMAP/G/1 queue with an application to the dimensioning of a playout buffer for vbr video. *Lecture Notes in Computer Science*, 3042:1338–1343, 2004.

[8] G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM Journal of Applied Mathematics*, 48(5):1159-1166, 1988.

[9] E. Hyon and A. Jean-Marie. Scheduling Services in a Queuing System with Impatience and Setup Costs. *Computer Journal*, 55(5):553–563, 2012.

[10] S. Kapodistria. The M/M/1 queue with synchronized abandonments. *Queueing Systems*, 68(1):79–109, 2011.

[11] L. Kleinrock. *Queueing systems, part I*. Wiley, New York, USA, 1975.

[12] N. Laoutaris and I. Stavrakakis. Intrastream synchronization for continuous media streams: A survey of playout schedulers. *IEEE Network Magazine*, 16(3):30–40, 2002.

[13] F. Li. Competitive Scheduling of Packets with Hard Deadlines in a Finite Capacity Queue. In *IEEE INFOCOM 2009*, pages 1062–1070, 2009.

[14] M. Li, T. Lin, and S. Cheng. Arrival process-controlled adaptive media playout with multiple thresholds for video streaming. *Multimedia Systems*, 18(5):391–407, 2012.

[15] R. Li and A. Eryilmaz. Scheduling for End-to-End Deadline-Constrained Traffic With Reliability Requirements in Multihop Networks. *IEEE-ACM Transactions on Networking*, 20(5):1649–1662, 2012.

[16] S. Park and J. Kim. An adaptive media playout for intra-media synchronization of networked-video applications. *Journal of Visual Communication and Image Representation*, 19(2):106–120, 2008.

[17] N. Perel and U. Yechiali. Queues with slow servers and impatient customers. *European Journal of Operational Research*, 201(1):247–258, 2010.

[18] B. Steyaert, K. Laevens, D. De Vleeschauwer, and H. Bruneel. Analysis and design of a playout buffer for vbr streaming video. *Annals of Operations Research*, 162(1):159–169, 2008.

[19] R. Sudhesh. Transient analysis of a queue with system disasters and customer impatience. *Queueing Systems*, 66(1):95–105, 2010.

[20] J. Van Velthoven, B. Van Houdt, and C. Blondia. On the probability of abandonment in queues with limited sojourn and waiting times. *Operations Research Letters*, 34(3):333–338, May 2006.

[21] A. Ward and S. Kumar. Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research*, 33(1):167–202, 2008.

[22] J. Walraevens, J.S.H. van Leeuwaarden, and O.J. Boxma. Power series approximations for two-class generalized processor sharing systems. *Queueing Systems: Theory and Applications*, 66(2):107–130, 2010.

[23] J. Yang, H. Hu, H. Xi, and L. Hanzo. Online buffer fullness estimation aided adaptive media playout for video streaming. *IEEE Transactions on Multimedia*, 13(5):1141–1153, 2011.

[24] D. Yue, W. Yue, and G. Xu. Analysis of customers' impatience in an M/M/1 queue with working vacations. *Journal of Industrial and Management Optimization*, 8(4):895–908, 2012.