Gamma-Approximation for the Waiting Time Distribution Function of the $M/G/1-\infty$ Queue

Michael Menth, Robert Henjes, Christian Zepfel, and Phuoc Tran-Gia
Department of Distributed Systems, Institute of Computer Science, University of Würzburg, Germany
Email: {menth|henjes|zepfel|trangia}@informatik.uni-wuerzburg.de

***Abstract*— In this paper we propose to approximate the waiting time distribution function (DF) of the waiting customers in a $M/G/1-\infty$ queuing system by a Gamma-distribution whose parameters α and β are set by means of the first and second moment of the waiting time that are obtained from the Takacs recursion formula. Discrete-time analysis (DTA) is another approximation approach for the same objective. We show first for analytically feasible special cases that DTA is very accurate and then we use it to validate the accuracy of the new Gamma-approximation for wide parameter ranges regarding the service time distribution. We show that the Gamma-approximation respects well even the third moment of the service time distribution. As the new approach is very simple and fast, it may be used by engineers with only little background in queuing theory to calculate quantiles for real-time control loops in technical systems.**

Keywords: numerical methods, queuing theory

I. Introduction

Many problems in telecommunication networks can be modelled by queuing systems. If the customers are flows, their interarrival time follows usually a Poisson process [1]. Thus, the analysis of the waiting time of the $M/G/1-\infty$ queuing system is often required. Takacs recursion formula allows a simple calculation of the k-th moments (cf. 5.112 in [2]). The Pollaczek-Khintchine formula yields even the generating function of the entire waiting time distribution function (DF) [3]. This formula can be evaluated numerically [4]–[6], but this is not always an easy task without the appropriate tools. Explicit expressions in the time domain exist for special cases like the $M/M/1-\infty$, the $M/D/1-\infty$ queuing system [7], or some long-tail service distributions [8]. Approximations of the waiting time DF exist for general service time DFs in $M/G/1-\infty$ [9] and even for general interarrival time DFs, i.e. for $GI/G/1-\infty$ [10]. However, they are sufficiently complex, they must be adapted for specific service time distributions, or they provide good results only for specific parameter ranges.

The contribution of this paper is the presentation of a simple approximation of the waiting time DF for arbitrary service time DFs. It is based on the Gamma-distribution, therefore, we call it Gamma-approximation. It takes into account the first, second, and third moment of the service time DF.

Discrete time analysis (DTA) can be used to calculate the waiting time of any discrete time $GI/GI/1-\infty$ system with a finite interarrival and service time distribution [11]. We show first that it can be also used for the approximation of a continuous time $G/G/1-\infty$ queuing system. Then, we apply it to validate the accuracy of the waiting time DF obtained by the Gamma-approximation.

We tested the new approximation method for for a wide range of coefficients of variation $c_{var}[B]$ of the service time B and system utilization ρ, i.e. for $(c_{var}[B], \rho) \in [0.05; 4] \times [0; 0.95]$. The numerical results showed that the Gamma-approximation is very accurate. It also reflects well the impact of the third moment of the service time B on the waiting time DF and it is more accurate than another more general approximation method from [4]. As the Gamma-approximation runs within milliseconds, it is much faster than DTA, and it is also easier to apply than DTA or any other considered approximation method. Therefore, it is a useful tool for a quick analysis of $M/G/1-\infty$ queuing systems. As the new formula is very simple and fast, it may be used by engineers with only little background in queuing theory to calculate quantiles for technical systems in real-time.

The paper is structured as follows. Section II reviews basics about the $M/G/1-\infty$ queuing system, it presents the new Gamma-approximation. Section III explains the DTA-approximation and shows that it has a very good accuracy for a sufficiently strict convergence parameter. Section IV validates then the accuracy of the Gamma-approximation. Finally, we summarize this paper and draw conclusions in Section V.

II. Gamma-Approximation of the $M/G/1-\infty$ Waiting Time

In this section we review first the $M/G/1-\infty$ queuing system with existing approaches for the calculation or approximation of the waiting time distribution function (DF). Then we propose the Gamma-approximation for this task.

A. The $M/G/1-\infty$ Queuing System

The $M/G/1-\infty$ queuing system consists of a Poisson arrival process, i.e., the interarrival time A of the customers is exponentially distributed (M) with mean $E[A] = \frac{1}{\lambda}$ and their holding time B follows an identically and independently distributed (iid) general distribution (G) with mean $E[B]$. The k-th moments $E[W^k]$ of the waiting time W can be calculated by Takacs' recursion formula (cf. 5.112 in [2]). However, the DF of the waiting time cannot directly be computed in the time domain. The Pollaczeck-Khintchine

This work was funded by Siemens AG, Munich. The authors alone are responsible for the content of the paper.

solution provides a formula solely for its generating function [3]. The retransformation into the time domain is difficult and can be done analytically only for special cases.

1) The Takacs Recursion Formula: The k-th moments of the waiting time of all customers can be calculated by the Takacs recursion formula (cf. 5.112 in [2]):

$$E[W^k] = \frac{\lambda}{1-\rho} \cdot \sum_{i=1}^{k} \binom{k}{i} \cdot \frac{E[B^{i+1}]}{i+1} \cdot E[B^{k-i}] \quad (1)$$

with $\rho = \frac{E[B]}{E[A]}$ being the utilization of the system and $E[W^0] = 1$. Thus, the first and second moment of the waiting time are

$$E[W] = \frac{\lambda \cdot E[B^2]}{2 \cdot (1-\rho) \cdot E[B]} \quad (2)$$

$$E[W^2] = 2 \cdot E[W]^2 + \frac{\lambda \cdot E[B^3]}{3 \cdot (1-\rho)}. \quad (3)$$

In particular, we need the first and second moment of the waiting time regarding only waiting customers. As the waiting time probability is $p_w = \rho$, they are given by

$$E[W_{wc}] = \frac{E[W]}{p_w} \quad (4)$$

$$E[W_{wc}^2] = \frac{E[W^2]}{p_w}. \quad (5)$$

2) The Pollaczek-Khintchine Solution: The Laplace-Stieltjes transformation (LST) $X^*(s)$ of a DF $X(t)$ for a random variable X is defined by

$$X^*(s) = \int_0^\infty e^{-s \cdot t} dX(t). \quad (6)$$

The LST of the waiting time DF is given by the well known formula by Pollaczek and Khintchine

$$W^*(s) = \frac{s \cdot (1-\rho)}{s - \lambda + \lambda \cdot B^*(s)} \quad (7)$$

with $B^*(s)$ being the LST of the service time DF. There are means to get numerical results from this expression [4], [6], [12], but this is not an easy task without appropriate tools.

3) Explicit DFs for Special Cases: For some special cases it is possible to retransform the expression in Equation (7) back into the time domain such that an explicit DF is available. We present two of these special cases in the following.

The waiting time DF for the $M/M/1-\infty$ system can be calculated by

$$W(t) = 1 - \rho \cdot e^{-(1-\rho) \cdot t / E[B]}. \quad (8)$$

The solution for $M/D/1-\infty$ is somewhat more complex and numerically challenging (cf. (2.122) in [7]):

$$W(t) = 1 - (1-\rho) \cdot \sum_{n=m+1}^{\infty} e^{-\lambda \cdot (n \cdot E[B] - t)} \cdot \frac{\lambda^n}{n!} \cdot (n \cdot E[B] - t)^n \quad (9)$$

with m being $m = \lfloor \frac{t}{E[B]} \rfloor$. Other explicit solutions are given for a class of long-tail service time DFs in [8].

4) Approximative Solutions for the DF: There are also approximative solutions of the form

$$W(t) = 1 - (\alpha \cdot e^{-\beta \cdot t} + \gamma \cdot e^{-\delta \cdot t}) \quad (10)$$

if some preconditions regarding B are met (cf. (4.4.1) in [9]). The parameters α, β, γ, and δ are quite complex to calculate and there is not always a solution for them.

Another simpler approximation for the waiting time DF of the general $GI/G/1-\infty$ system is given in [10] of the form

$$W(t) = 1 - \alpha \cdot e^{-\eta \cdot t} \quad (11)$$

The rate parameter η is approximated based on the properties of the service time DF and there are various specialized formulae to adapt η to the exact type of the service time. We can calculate the waiting time DF for $M/Gamma/1-\infty$ with the following parameters:

$$\eta = \frac{2 \cdot (1-\rho)}{1 + c_{var}[B]^2} \cdot (1 - (1-\rho) \cdot \frac{1 - c_{var}[B]^2}{3 \cdot (1 + c_{var}[B]^2)}) \quad (12)$$

$$\alpha = \eta \cdot E[W]. \quad (13)$$

The latter equation assures the correct mean waiting time of the approximated DF. This approximation works good if ρ is large. We use it in the following to validate the Gamma-approximation where the DTA does not work. A similar approximation has been applied in [13] to calculate the quantiles of waiting times (cf. Section 1.3 in [13]).

B. The Gamma-Approximation

We introduce first the Gamma-distribution and some of its properties. Then we use the first and second moment of the waiting time of the waiting customers in an $M/G/1$–∞ system to determine the α- and β-parameter of a Gamma-distribution to get an estimate for its DF in a simple way.

1) The Gamma-Distribution: The base for the Gamma-*distribution* $\Gamma(\alpha, \beta)$ is the Gamma-*function* $\Gamma(z)$, which is defined by

$$\Gamma(z) = \begin{cases} 0 & \text{if } x < 0 \\ \int_0^\infty x^{z-1} \cdot e^{-t} dx & \text{if } 0 \le x \end{cases} \quad (14)$$

Most important properties of the Gamma-*function* $\Gamma(z)$ are

$$\Gamma(z+1) = z \cdot \Gamma(z) \text{ if } z > 0 \quad (15)$$

$$\Gamma(k+1) = k! \text{ if } k \in \mathbb{N}_0 \quad (16)$$

$$\Gamma(\frac{1}{2}) = \sqrt{\pi} \quad (17)$$

They are interesting to know, but they are not required in the following. The Gamma-*distribution* $\Gamma(\alpha, \beta)$ is given by its probability density function (pdf)

$$f_{\Gamma(\alpha,\beta)}(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{\beta^{-\alpha} \cdot t^{\alpha-1} \cdot e^{-t/\beta}}{\Gamma(\alpha)} & \text{if } t \ge 0 \end{cases} \quad (18)$$

The calculation of its DF $F_{\Gamma(\alpha,\beta)}$ and even of its inversion $F^{-1}_{\Gamma(\alpha,\beta)}$ is implemented by many tools for statistical analysis, e.g. in Matlab [14] by the commands “gampdf” and “gaminf”.

However, closed form expressions for $F_{\Gamma(\alpha,\beta)}$ exist only for integral values of α. Then, we have

$$F_{\Gamma(\alpha,\beta)}(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 - e^{-t/\beta} \cdot \sum_{0 \le i < \alpha} \frac{(t/\beta)^i}{i!} & \text{if } t \ge 0. \end{cases} \quad (19)$$

Thus, the Gamma-distribution $Gamma(k, \frac{1}{k \cdot \lambda})$ equals the Erlang-distribution $Erlang(k, \lambda)$. Therefore, the Gamma-distribution can be viewed as an extension of the Erlang-distribution towards $k \in \mathbb{R}^+$. The mean, the variance, and the coefficient of variation of the Gamma-distribution are

$$E[X] = \alpha \cdot \beta \quad (20)$$

$$VAR[X] = \alpha \cdot \beta^2 \quad (21)$$

$$c_{var}[X] = \frac{1}{\sqrt{\alpha}}. \quad (22)$$

Hence, the Gamma-distribution may be used to approximate distributions with a given mean and variance.

2) Estimation of the $M/G/1$–∞ Waiting Time Distribution by the Gamma-Distribution: The first and the second moment of the waiting time of the waiting customers of an $M/G/1$–∞ queueing system can be calculated by Equations (4) and (5). We use them to set the parameters α and β of the Gamma-distribution after a manipulation of Equations (20) and (21) by

$$\alpha = \frac{E[W]^2}{VAR[W]} = \frac{E[W]^2}{E[W^2] - E[W]^2} \quad (23)$$

$$\beta = \frac{E[W]}{\alpha}. \quad (24)$$

The resulting DF $F_{\Gamma(\alpha,\beta)}$ describes then the distribution of the waiting time of customers that are not immediately served upon arrival. The waiting time distribution of all customers is given by

$$F(t) = 1 - \rho + \rho \cdot F_{\Gamma(\alpha,\beta)}(t). \quad (25)$$

If the service time in $M/G/1-\infty$ is exponential, we get the $M/M/1-\infty$ system. As the waiting time DF of its waiting customers is exponential, we have $c_{var}[W] = 1$, and therefore, $\alpha = 1$ (cf. Equation (22)). In this case, the Gamma-approximation meets the exponential distribution exactly with the same mean $E[W]$. The question is now: how exact is the Gamma-approximation for other distributions of the service time? This is the issue in the remaining part of the paper. To that end, we compare its results with the ones of the DTA and the approximation given in Equation (13).

III. Discrete Time Analysis and its Accuracy

In this section, we explain the discrete time analysis (DTA) for the discrete time $GI/GI/1-D^{max}$ queuing system with bounded delay D^{max}. We use it to approximate the continuous time $GI/GI/1-\infty$ queue and identify potential sources of inaccuracies. Finally, we compare its results for the waiting time DF of an $M/D/1$–∞ and an $M/M/1$–∞ queuing system with analytical results and show that its accuracy depends on the parameters of the DTA-approximation.

A. Discrete-Time Analysis of the $GI/GI/1-D^{max}$ Queuing System with Bounded Delay D^{max}

The discrete time $GI/GI/1-D^{max}$ queuing system with bounded delay D^{max} is based on discrete time units, i.e., the interarrival time of its customers is distributed according to an iid general distribution and their holding time follows also an iid general distribution. The respective random variables are denoted by A and B. The value range of both distributions contains only multiples of a common basic time unit, but we omit this unit in the following.

1) State Transitions of the $GI/GI/1-D^{max}$ Queue: We analyze the discrete time $GI/GI/1-D^{max}$ queue by considering a discrete time Markov chain (DTMC) whose state represents the unfinished work in the buffer which is described by the random variable U. Upon arrival of a new customer, the unfinished work in the buffer is incremented by the new customer's service time B. If this exceeds the delay bound D^{max} of the buffer, the unfinished work is set to this delay bound. Afterwards, the unfinished work is decreased by the passing time units until the next customer arrives. Renewal points of the process exist shortly before (-) and after (+) the arrival instants. We number them t_n^- and t_n^+ and the states U_n^- and U_n^+, accordingly. The Markov chain evolves based on the following recursive stochastic equations.

$$U_{n+1}^- = max(U_n - A, 0) \quad (26)$$

$$U_{n+1}^+ = min(U_{n+1} + B, D^{max}).. \quad (27)$$

2) Discrete Time Analysis: An early use of discrete time analysis (DTA) can be found in [11], [15]–[18] with application to packet networks. The concept of DTA works as follows. An iteration algorithm starts with a distribution x_0 of the system state at the first renewal point. The distribution x_{n+1} of the system state at renewal point $n+1$ is calculated based on the distribution x_n of the system state at renewal point n and the distribution y of the factors. The calculation itself is described by a state transition function from one renewal point to the next one, i.e., it it denoted by the recursive stochastic equation

$$X_{n+1} = f(X_n, Y). \quad (28)$$

If the Markov chain is aperiodic, the series of the x_n converges to the stationary state distribution which characterizes the distribution of the system states at the renewal points after a long time. We recognize convergence in practice if the entries of two successive state distributions x_n and x_{n+1} differ not more than ε_c. If the Markov chain is periodic, there are modifications to the iteration algorithm such that the series x_n converges also to the stationary distribution [19]. The whole concept is extended to different types of renewal points, e.g. shortly before and after a customer arrival, and stationary distributions can be calculated for both types. Thus, we can calculate the distribution of the unfinished work in the buffer shortly before and after a customer arrival by DTA using Equations (26) and (27) as state transition functions. Note that the stationary state distribution shortly before a customer arrival yields the waiting time distribution for new customers.

B. Approximation of the Continuous Time $GI/GI/1-\infty$ by Discrete Time $GI/GI/1-D^{max}$ through DTA

The continuous time $GI/GI/1-\infty$ queue and the discrete time $GI/GI/1$–D^{max} queue with bounded delay D^{max} differ significantly regarding the nature of their interarrival time and service time distribution and regarding their buffer size. In addition, DTA is a numerical algorithm that terminates without having calculated the exact result. As these issues may lead to wrong approximation results, actions must be taken to keep the error small.

1) Continuous and Discrete Time Distributions: The DTA-approximation requires the transformation of the continuous DF F_c of the interarrival time and the service time into discrete DFs F_d which are step functions. We achieve this by increasing the step function F_d to F_c at the multiples of the basic time unit u. Thus, the approximation quality can be increased by decreasing the basic time unit u.

2) Infinite and Finite Distributions: The range of the interarrival time and the service time may be infinite for the continuous time $GI/GI/1-\infty$ queue, but it must be limited for the discrete time $GI/GI/1-D^{max}$ queue since this is a requirement of the DTA algorithm. Thus we choose a value $t_{max} : 1-F_c(t_{max}) < \varepsilon_f$ and set F_d to

$$F_d(t) = \begin{cases} 0 & \text{for } t = 0 \\ F_c(n \cdot u) & \text{for } t < t_{max} \wedge t \in ((n-1) \cdot u, n \cdot u] \\ 1 & \text{for } t \geq t_{max}. \end{cases} \tag{29}$$

Thus, the approximation quality can be increased by decreasing ε_f.

3) Infinite Buffer and Limited Delay: The continuous time $GI/GI/1$–∞ queue has an infinite buffer by definition, but the discrete time $GI/GI/1-D^{max}$ queue requires a limited delay which lies in the nature of DTA. This is obviously a source for approximation errors. To avoid them, the delay bound D^{max} must be set large enough, i.e., it must be set so large that the probability ε_d for customer to exceed this delay bound D^{max} is very small. This can be controlled by having a look at the stationary state distribution of the unfinished work shortly after the packet arrival. If the probability for the unfinished work to be D^{max} is smaller than ε_d, then the delay bound is sufficiently large. Thus, the approximation quality can be increased by decreasing ε_d.

4) Convergence Accuracy: The iteration algorithm of the DTA terminates if the difference of two consecutive distributions regarding the same renewal point differ less than ε_c in each component. Thus, the approximation quality can be increased by decreasing ε_c.

C. Validation of the DTA-Approximation of $GI/GI/1-\infty$

We explain first the generation of the factor distributions for the DTA and compare then the resulting DFs. Then, we show that the DTA-approximation of the discrete time $GI/GI/1-D^{max}$ system can be used to calculate the waiting time DF of the continuous time $GI/GI/1-\infty$ system. To that end, we validate its results by analytical values in the special cases of the $M/D/1-\infty$ and the $M/M/1-\infty$ system.

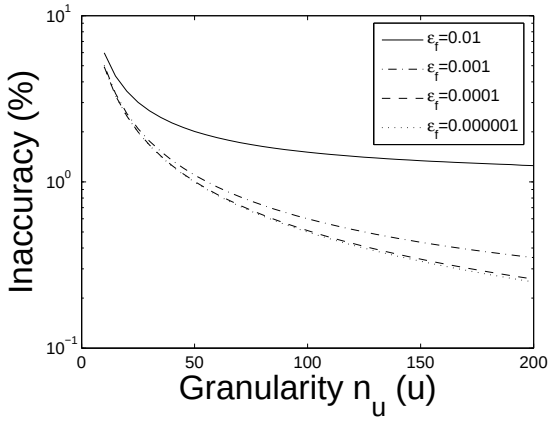
1) Generation of the Factor Distributions for DTA: We determine the time granularity by defining $E[A] = n_u \cdot u$ and choose $n_u = 100$ by default. As the DTA requires distributions instead of DFs to describe the factors A and B, we calculate them by

$$P(A' = k) = \begin{cases} 0 & \text{for } k = 0 \\ F_c^A(k) - F_c^A(k-1) & \text{for } k > 0. \end{cases} \tag{30}$$

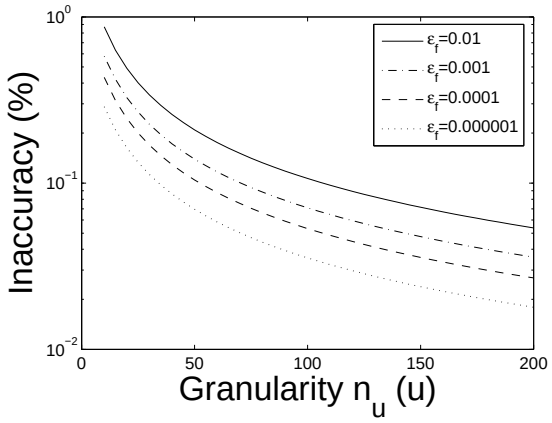
Note that the value k corresponds to a duration of $k{\cdot}u$. We limit the DFs for A and B according to Equation (29) with $\varepsilon_f = 10^{-4}$ by default and mark the discretized random variables by a "′". We test the system under a given utilization ρ. Thus, the mean of the service time is $E[B] = \rho \cdot E[A]$. Due to the discretization, the mean $E[X]$ and the coefficient of variation $c_{var}[X]$ of the discretized A' and B' differ slightly from the ones of the intended A and B. The discretization error of A' and B' leads also to a slightly different system utilization $\rho' = \frac{E[B']}{E[A']}$.

We use the exponential DF $F_c(t) = 1-e^{-t/E[A]}$ as the base for the interarrival time DF for A'. Figures 1(a) and 1(b) illustrate the inaccuracy of the discretized exponential distribution A' in relation to the continuous DF of A. Figure 1(a) shows the discretization error $R_E[A'] = \frac{E[A']-E[A]}{E[A]}$ regarding the mean $E[A]$. The discretized distribution has a slightly larger mean, but the difference decreases with increasing n_u. An decreasing discretization parameter ε_f can improve the result only up to $\varepsilon_f = 0.0001$. Figure 1(b) shows the discretization error $R_{c_{var}}[A'] = \frac{c_{var}[A']-c_{var}[A]}{c_{var}[A]}$ regarding the coefficient of variation $c_{var}[A]$. The discretized distribution has a slightly larger coefficient of variation, but the difference decreases with increasing n_u. In contrast to the mean, the accuracy of the coefficient of variation is further decreased by a decreasing ε_f.

2) Comparison of Analytical and Approximated Distribution Functions: We compare the waiting time DFs for an $M/D/1-\infty$ and an $M/M/1-\infty$ system calculated from the DTA-approximation of the $GI/GI/1$–D^{max} system and from the analytical formulae given in Equations (8) and (9). They are presented for a system utilization of $\rho = 0.9$ in Figures 2(a) and 2(b). We have plotted the complementary DF (CDF) because this makes the difference between the analytical and approximated values on a logarithmic y-scale more visible. The waiting time is given in multiples of the mean service time $E[B]$ since this is the invariant component in most systems. The approximations are shown for $n_u = 100$, $\varepsilon_f = 10^{-4}$, $\varepsilon_d = 10^{-10}$, and $\varepsilon_c = 10^{-\{6,7,8\}}$ for both considered queuing systems. The approximation with $\varepsilon_c = 10^{-6}$ yields significant deviations for large waiting times. However, $\varepsilon_c = 10^{-7}$ yields already a sufficiently good correspondence between the analytical and approximative CDF and the curve for $\varepsilon_c = 10^{-8}$ coincides with the analytical values. We observe within each of the figures that the inaccuracy increases for the same ε_c with increasing system utilization. When we compare the results

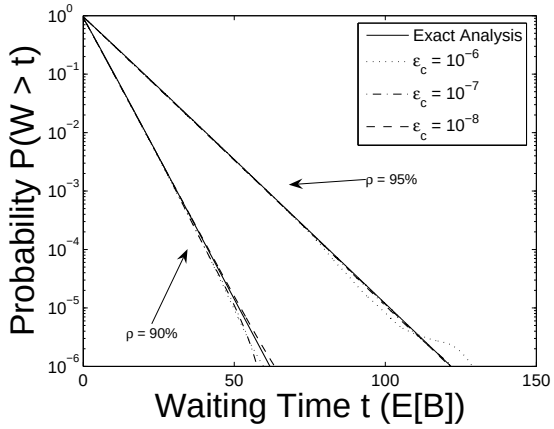


(a) Relative error of the mean rate $R_E[A'] = \frac{E[A']-E[A]}{E[A]}$.

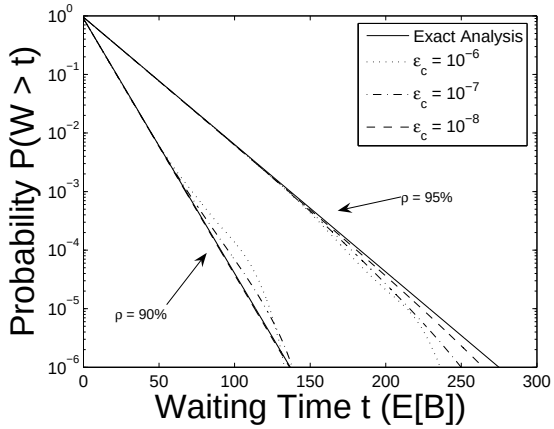


(b) Relative error of the coefficient of variation $R_{c_{var}}[A'] = \frac{c_{var}[A']-c_{var}[A]}{c_{var}[A]}$.

Fig. 1. Discretization error for an exponential DF depending on the discretization parameters n_u and ε_f.



(a) $M/D/1-\infty$ system.



(b) $M/M/1-\infty$ system.

Fig. 2. Analytical and approximated CDFs of the waiting time for a utilization of $\rho = 0.9$ and $\rho = 0.95$.

for $M/D/1-\infty$ and $M/M/1-\infty$, we also realize that the inaccuracy increases for increasing coefficients of variation $c_{var}[B]$ of the service time, too. Thus, the DTA leads to good and trustworthy results only for small or medium coefficients of variation $c_{var}[B]$ and moderate system utilization.

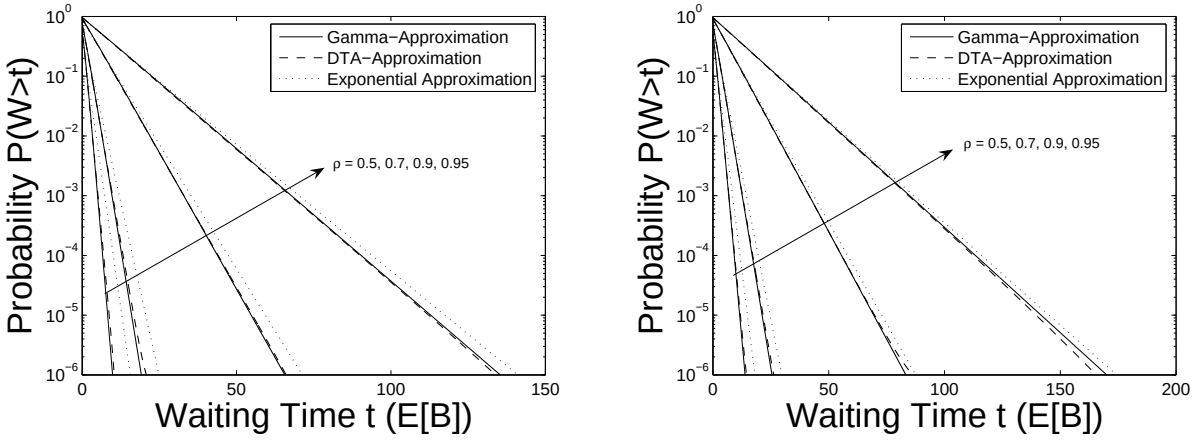
After all, we use the discretization and termination parameters set to $n_u = 100$, $\varepsilon_f = 10^{-4}$, $\varepsilon_d = 10^{-10}$, and $\varepsilon_c = 10^{-8}$ for the DTA analysis in the following.

IV. Validation of the Gamma-Approximation

In this section we illustrate the accuracy of the Gamma-approximation by a comparing its complementary DF (CDF) with the results obtained from the corresponding DTA-analysis in Section III and approximative results from the simple exponential $GI/G/1-\infty$ approximation in Equation (13). We consider first systems with a different coefficient of variation $c_{var}[B]$ of the service time and different utilization levels ρ. Then we study the impact of their third moment on the approximation accuracy for which we use a symmetric an one strongly asymmetric service time distribution.

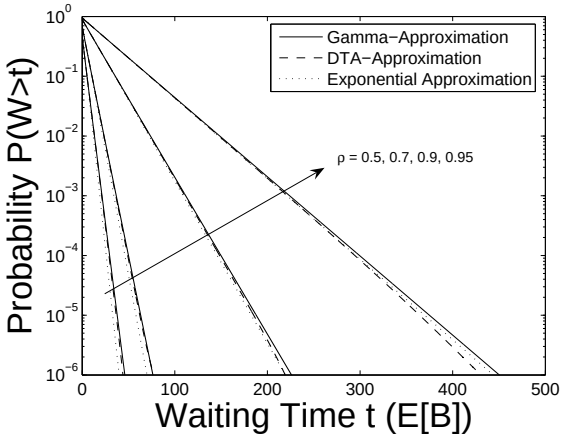
A. Impact of the First and Second Moment of the Service Time

The first moment of the service time determines the system utilization $\rho = \frac{E[B]}{E[A]}$ and the second moment determines the coefficient of variation of the service time by $c_{var}[B] = \frac{\sqrt{E[B^2]-E[B]^2}}{E[B]}$. Since ρ and $c_{var}[B]$ are more intuitive, we use them to control our parameter studies instead of $E[B]$ and $E[B^2]$. We use the Gamma-distribution as service time since it can be easily adapted to meet a given ρ and $c_{var}[B]$ (cf. Equations (23) and (24)). We discretize the continuous Gamma-DF according to Equation (29) to obtain approximated and finite DF A' and B' as input for the DTA-analysis. For the sake of a fair comparison, we use $E[B']$, $E[(B')^2]$, and $E[(B')^3]$ to calculate the first and the second moment of the waiting time of waiting customers in Equations (4) and (5) since they are required to fit the parameters α and β of the Gamma-distribution in Equations (23) and (24). Then, we use

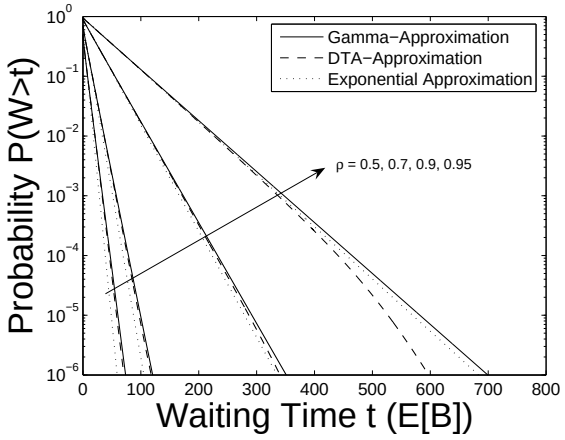


(a) $c_{var}[B] = 0.05$.

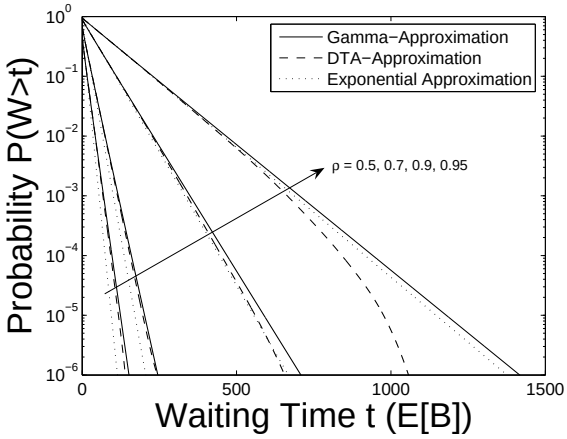
(b) $c_{var}[B] = 0.5$.



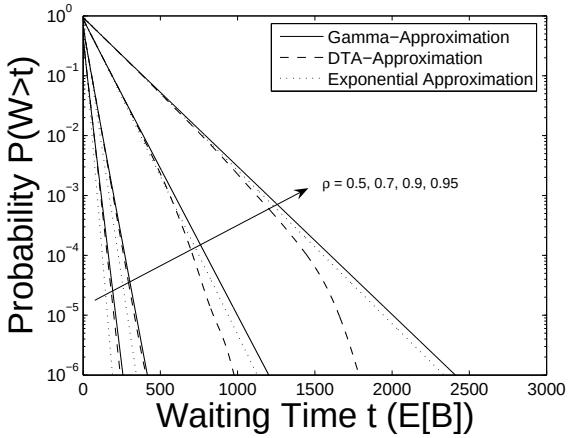
(c) $c_{var}[B] = 1.5$.



(d) $c_{var}[B] = 2.0$.



(e) $c_{var}[B] = 3.0$.



(f) $c_{var}[B] = 4.0$.

Fig. 3. The CDFs of the waiting time from the Gamma-, DTA-, and exponential approximation for an $M/Gamma/1-\infty$ system with various coefficients of variation $c_{var}[B]$ at various utilization levels ρ.

this distribution together with $\rho' = \frac{E[B']}{E[A']}$ to derive the Gamma-approximation in Equation (25).

Figures (3(a)) to (3(f)) compare the CDFs of the waiting time for the Gamma-approximation, the DTA-approximation, and the exponential approximation of Equation (13). We have chosen the coefficients of variation of the service time $c_{var} = \{0.05, 0.5, 1.5, 2.0, 3.0, 4.0\}$ in these figures to perform

a parameter study. Note that the Gamma-approximation cannot be parameterized to approximate the waiting time DF when the coefficient of the service time is $c_{var}[B]=0$. For $c_{var}[B]=1$ it yields exactly the analytical results, therefore, we omitted this figure. The different values for the system utilization $\rho = \frac{E[B]}{E[A]} = \{0.5, 0.7, 0.9, 0.95\}$ within a single figure are obtained by varying the mean service time $E[B]$.

The DTA-approximation is very good for low values of the system utilization ρ and the solid lines of the Gamma-approximation coincides with the dashed lines of the DTA-approximation in all figures for $\rho=0.5$ and $\rho=0.7$ whereas the exponential approximation shows significant deviations. It is known from [10] that the accuracy of the exponential approximation is only good for large values of ρ. Thus, the Gamma-approximation yields very good approximation results for small and medium system utilization. The DTA-approximation is also quite accurate for low coefficients of variation like $0 \leq c_{var}[B] \leq 1$ and the solid lines of the Gamma-approximation coincide with the dashed lines of the DTA-approximation for $c_{var}[B]=0.05$ and $c_{var}[B]=0.5$. For larger values of $c_{var}[B]$ and large values of ρ, the accuracy $\varepsilon_c = 10^{-8}$ of the DTA-approximation does not suffice to produce accurate results since the respective CDFs deviate significantly from a straight line in the logarithmic plot. In these cases, the correspondence of the solid lines for the Gamma-approximation and the dotted lines of the exponential approximation is relatively good. Hence, the Gamma-approximation leads to good approximation result for a very broad range of $c_{var}[B]$ and ρ. The high resolution of our numerical results proposes, that it can be used to calculate even large quantiles of the waiting time, e.g. the 99.999% percentile.

B. Impact of the Third Moment of the Service Time

The Takacs formula requires the third moment of the service time to calculate the second moment of the waiting time of waiting customers (cf. Equation (5)). Therefore, we are interested in the impact of this value on the CDF of the waiting time of a $M/G/1-\infty$ and in the approximation accuracy of the Gamma-approximation regarding this parameter. To that end, we consider two distributions with the same first and second moment. They are given in Table I. The symmetric distribution has a third moment of $E[B^3_{sym}] = 3.6 \cdot 10^6$ u^3 while the asymmetric distribution has a third moment of $E[B^3_{asym}]=9.72 \cdot 10^6$ u^3

TABLE I

SYMMETRIC AND ASYMMETRIC DISTRIBUTIONS, BOTH WITH A FIRST AND SECOND MOMENT OF $E[B]=100$ U AND $E[B^2]=18000$ U^2.

$B_{sym}=i$ (u)	$P(B_{sym}=i)$	$B_{asym}=i$ (u)	$P(B_{sym}=i)$
10	$\frac{40}{81}$	90	$\frac{80}{81}$
100	$\frac{1}{81}$		
190	$\frac{40}{81}$	900	$\frac{1}{81}$

We use both service time distributions to calculate the CDF of the waiting time for a system utilization of $\rho=0.9$, i.e., we set the mean of the interarrival to $E[A]=\frac{E[B]}{\rho}$. Figure 4 shows that the CDFs of the waiting times from the service times B_{sym} and B_{asym} differ notably, but it also shows that both the DTA and the Gamma-approximation account for this difference.

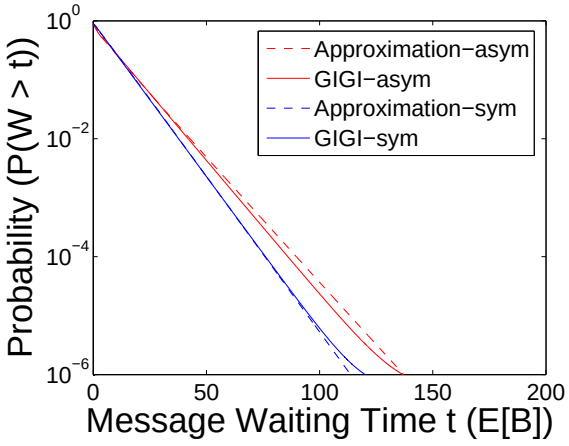


Fig. 4. The CDFs of the waiting time from the Gamma- and DTA-approximation for an $M/G/1-\infty$ system for $\rho=0.9$ with two different service times that differ only in their third but not in their first and second moment.

V. CONCLUSION

We have first reviewed the general $M/G/1-\infty$ queueing system and some approaches to characterize the waiting time of waiting customers. However, there is no explicit expression except for some special cases. The waiting time DF can only obtained by a numerical inversion of its Laplace-Stieltjes transform which is not possible with the appropriate software. Approximation methods are numerically also not simple, or they are specific to the used service time distribution. Thus, all methods require a substantial overview on queuing theory and appropriate tools. Therefore, we proposed the Gamma-approximation that estimates the waiting time distribution function (DF) by a Gamma-distribution based on the first three moments of the service time distribution.

To show the accuracy of the new Gamma-approximation, we used discrete time analysis (DTA) to get an estimate for the true waiting time DF. Therefore, we showed first the accuracy of the DTA approach by validating it with the known DF for the $M/D/1-\infty$ and the $M/M/1-\infty$ queueing system. Then, we showed by an analysis of the complementary DFs (CDFs) of the Gamma-, DTA- and an exponential approximation method that the accuracy of the waiting time CDF obtained from the Gamma-approximation is very accurate such that it can be even used for the calculation of 99.999% quantiles. We have studied a broad range of coefficients of variation $c_{var}[B]$ of the service time and the system utilization ρ, and the above results hold in all cases. The third moment of the service time has a minor impact on the CDF of the waiting time, but it is also well captured by the Gamma-approximation.

The computation time for the Gamma-approximation is negligible while the DTA-analysis takes minutes or hours to

produce sufficiently accurate results, and for large coefficients of variation $c_{var}[B]$ and a large utilization ρ it takes even days. This makes the advantage of the new calculation method obvious: it provides quite accurate estimates very quickly. Therefore, it is suitable for the implementation in real-time systems, e.g., where QoS measures like quantiles of the waiting time are needed to perform admission control.

After all, the Gamma-approximation is very simple to apply if the Gamma-distribution is available which is the case in most of today's mathematical toolboxes. Therefore, it is an attractive means for engineers with only little background in queuing theory. In addition, it calculates fast which makes it appropriate for application in technical control systems.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. John Shortle, Prof. Henk Tijms, and Prof. Ward Whitt for their valuable contributions.

REFERENCES

[1] J. W. Roberts, "Traffic Theory and the Internet," *IEEE Communications Magazine*, vol. 1, no. 3, pp. 94 – 99, Jan. 2001.

[2] L. Kleinrock, *Queueing Systems*, 1^{st} ed. New York: John Wiley & Sons, 1975, vol. 1: Theory.

[3] H. Takagi, *Queueing Analysis Volume 1: Vacation and Priority Systems*. North-Holland, 1991.

[4] J. Abate, G. L. Choudhury, and W. Whitt, "Calculation of the GI/G/1 Waiting Time Distribution an its Cumulants from Pollaczek's Formulas," *International Journal of Electronics and Communications*, vol. 47, no. 5/6, pp. 311–321, Sept. 1993.

[5] J. Abate and W. Whitt, "Numerical Inversion of Laplace Transforms of Probability Distributions," *INFORMS Journal on Computing*, vol. 7, no. 1, pp. 36–43, Winter 1995.

[6] J. F. Shortle, P. H. Brill, M. J. Fischer, D. Gross, and D. M. B. Masi, "An Algorithm to Compute the Waiting Time Distribution for the M/G/1 Queue," *INFORMS Journal on Computing*, vol. 16, no. 2, pp. 152–161, Spring 2004.

[7] N. U. Prabhu, *Queues and Inventories*. New York, London, Sydney: John Wiley & Sons, Inc., 1965.

[8] J. Abate and W. Whitt, "Explicit M/G/1 Waiting Time Distributions for a Class of Long-Tail Service Time Distributions," *Operations Research*, vol. 25, no. 1, pp. 25–31, Aug. 1999.

[9] H. C. Tijms, *Stochastic Modelling and Analysis: A Computational Approach*. John Wiley & Sons, 1986.

[10] J. Abate, G. L. Choudhury, and W. Whitt, "Exponential Approximations for Tail Probabilities in Queues I: Waiting Times," *Operations Research*, vol. 43, no. 5, pp. 885–901, Sept. 1995.

[11] M. H. Ackroyd, "Computing the Waiting Time Distribution for the G/G/1 Queue by Signal Processing Methods," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 52–58, January 1980.

[12] H. C. Tijms, *A First Course in Stochastic Models*. John Wiley & Sons, 2003.

[13] L. P. Seelen, H. C. Tijms, and M. H. Van Horn, *Tables for Multi-Server Queues*. North-Holland, 1985.

[14] "MATLAB – The Language of Technical Computing," http://www.mathworks.com/.

[15] M. H. Ackroyd, "Stationary and Cyclostationary Finite Buffer Behaviour Computation via Levinson's Method," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 10, pp. 2159 – 2170, Dec. 1984.

[16] P. Tran-Gia and H. Ahmadi, "Analysis of a Discrete-Time $G^X/D/1-S$ Queueing System With Applications in Packet-Switching Systems," in *IEEE Infocom*, New Orleans, 1988, pp. 0861–0870.

[17] P. Tran-Gia and R. Dittmann, "A Discrete-Time Analysis of the Cyclic Reservation Multiple Access Protocol," *Performance Evaluation*, vol. 16, pp. 185–200, 1992.

[18] P. Tran-Gia, "Discrete-Time Analysis Technique and Application to Usage Parameter Control Modeling in ATM Systems," in *8th Australian Teletraffic Research Seminar*, Melbourne / Australia, Dec. 1993.

[19] M. Menth, "A Framework for Modelling and Solving Discrete Finite Markov Chains," University of Würzburg, Institute of Computer Science, Technical Report, No. 326, Feb. 2004.