Complete characterisation of the customer delay in a queueing system with batch arrivals and batch service

Dieter Claeys · Koenraad Laevens · Joris Walraevens · Herwig Bruneel

Received: date / Accepted: date

Abstract Whereas the buffer content of batch-service queueing systems has been studied extensively, the customer delay has only occasionally been studied. The few papers concerning the customer delay share the common feature that only the moments are calculated explicitly. In addition, none of these surveys consider models including the combination of batch arrivals and a server operating under the full-batch service policy (the server waits to initiate service until he can serve at full capacity). In this paper, we aim for a complete characterisation - i.e., moments and tail probabilities - of the customer delay in a discrete-time queueing system with batch arrivals and a batch server adopting the full-batch service policy. In addition, we demonstrate that the distribution of the number of customer arrivals in an arbitrary slot has a significant impact on the moments and the tail probabilities of the customer delay.

Keywords customer delay · moments · tail probabilities · batch arrivals · batch service · full-batch service policy

1 Introduction

Whereas servers in traditional queueing systems serve one customer at a time, batch servers process batches of customers. As batch servers appear in a spectrum of applications (for instance in transportation, production and manufacturing systems, telecommunications, et cetera), batch-service queueing models

D. Claeys · K. Laevens · J. Walraevens · H. Bruneel
Stochastic Modeling and Analysis of Communication Systems (SMACS) Research Group,
Department of Telecommunications and Information Processing (TELIN),
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
Tel.: +32 9 264 3414
Fax: +32 9 264 42 95
E-mail: Dieter.Claeys@telin.ugent.be

have been studied extensively, in discrete as well as in continuous time. In most of the papers (Arumuganathan and Jeyakumar 2005; Bailey 1954; Chang and Choi 2005; Chang and Takine 2005; Chaudhry and Templeton 1983; Claeys et al. 2007; Cohen 1969; Dümmler and Schömig 1999; Goswami et al. 2006; Gupta and Goswami 2002; Lee et al. 1996; Neuts 1967; Powell and Humblet 1986; Samanta et al. 2007; Sikdar and Gupta 2005; Yi et al. 2007), expressions are obtained for the distribution of the number of customers in the system at various time epochs and for several service policies (the service policy establishes when an available server is allowed to start processing).

The customer delay, however, has attracted less attention, for instance in Chaudhry and Templeton 1983; Cohen 1969; Dagsvik 1975; Downton 1955; Keilson 1962; Kim and Chaudhry 2006; Medhi 1975; Miller 1959. A common feature in Chaudhry and Templeton 1983; Downton 1955; Kim and Chaudhry 2006; Medhi 1975 is that single arrival models are studied. Cohen 1969; Dagsvik 1975; Keilson 1962; Miller 1959 consider queueing models with batch arrivals and batch service whereby the server commences processing if there is at least one customer present in the system. In this paper, we consider a batch-service queueing model with batch arrivals and the server adopts the full-batch service policy, meaning that he will initiate service only if the number of customers present in the system reaches or exceeds the batch server's capacity (the maximum number of customers that can be served in the same batch). Especially the combination of this service policy and batch arrivals complicates the analysis. Indeed, a randomly tagged customer does not only have to wait until batches of previously arrived customers are served, but experiences an additional delay when the server postpones the service of the batch containing the tagged customer until that batch is completely filled. This latter part of the delay is simply the sum of some geometrically distributed interarrival times in case of Bernoulli arrivals (either 0 or 1 customers arrive during a slot), while this is more complicated when several customers can arrive in a slot, i.e. when a batch arrival model is adopted.

In our conference paper Claeys et al. 2008, we have established the probability generating function (PGF) of the delay that a random customer experiences in this model. The resulting PGF is suitable for obtaining moments, but it is not useful to extract tail probabilities (probability that the delay of a random customer exceeds some large threshold) from. In addition, the tail probabilities have also not been studied in Chaudhry and Templeton 1983; Cohen 1969; Dagsvik 1975; Downton 1955; Keilson 1962; Kim and Chaudhry 2006; Medhi 1975; Miller 1959. In this paper, we aim for a complete characterisation of the customer delay: we deduce moments as well as accurate approximations of the tail probabilities.

This paper is organised as follows: in section 2, the model is specified in detail. The analysis is split up in two parts: first, the moments are studied in section 3. We therefore deduce the probability generating function (PGF) of the customer delay (section 3.1). This is done through a joint analysis of the two parts of the delay, as defined above. The customer delay is then of course simply the sum of both parts. This result is partly based on our conference

paper Claeys et al. 2008. Next, we extract the moments from this PGF in section 3.2, by applying the moment generating property of PGF's. After that, we show, through some examples, that the distribution of the number of customer arrivals in a random slot influences the moments significantly. If this would not have been the case, the (simpler) Bernoulli model would provide an accurate approximation.
Second, we analyze the tail probabilities in section 4. We therefore turn to another approach (section 4.1), because the obtained PGF in section 3.1 is not useful for the extraction of tail probabilities. We therefore characterize the customer delay as the maximum of two time periods. The first one is the time between arrival of a tagged customer and the departure of all batches containing only customers that arrived before this customer. The second one equals the time between arrival of the tagged customer and the time instant that enough future customers have arrived to fill the service batch of the tagged customer. The approach thus essentially boils down to a redefinition of the second part of the customer delay as defined earlier. The reason is that it is advantageous to deal with the maximum operator when calculating tail probabilities (section 4), while the sum operator is better suited for calculation of the moments (section 3). The resulting approximate formula for the tail probabilities turns out to be extremely accurate (section 4.2). We also demonstrate that the distribution of the number of customer arrivals per slot has a significant influence on the tail probabilities (section 4.3). The paper is finalised by drawing some conclusions in section 5.

2 Model description

The model has the following features:

- It is a discrete-time queueing model, i.e. the time axis is divided into fixed-length periods, referred to as slots.
- Several customers may arrive during a slot (which we call batch arrivals in this paper) and the number of arrivals during slot k is denoted by A_k. We assume that the sequence $\{A_k\}_{k\geq 1}$ consists of independent and identically distributed (IID) random variables (RV's), with common PGF $A(z)$. The number of customer arrivals during an arbitrary slot is denoted by A and has, in agreement with the previous notation, PGF $A(z)$.
- The queue is infinitely large. Therefore, all arriving customers can enter the queue and will eventually be served. The restriction of an infinite queue capacity is not stringent, as in most practical applications the queue is large in order to minimize the loss probability and the customer delay is not influenced considerably by this small fraction of customers that cannot enter the system.
- There is one batch server of capacity c (c is a constant) and this server operates under the full-batch service policy. That is, when the server becomes available and finds less than c customers in the queue, the server waits to

initiate service until at least c customers have accumulated in the system. At that moment, the server starts processing a batch of c customers.
- The server starts processing a batch at the beginning of a slot and finishes its service at the end of the same slot. Hence, the service of the customers in a batch together will take one slot. This yields that an arriving customer has to wait for service until at least the beginning of the next slot. The remaining time of a slot after the customer has arrived is not included in our definition of the customer delay, since we count the customer delay as an integral number of slots. The service times are also excluded from the customer delay. However, one can easily include them by multiplying the obtained PGF by its argument z.
- The queueing discipline is first-come-first-served (FCFS).

Summarized, our model is a discrete-time queueing model with batch arrivals and a batch server that operates under the full-batch service policy. This model can, in agreement with Kim and Chaudhry 2006, be denoted by $Geo^X/1^C/1$. The equilibrium condition of this queueing model requires that the load $\rho \triangleq \lambda/c < 1$, with $\lambda \triangleq \mathrm{E}[A] = A^{'}(1)$ (we use primes to indicate derivatives).

We close this section with the following convention: the queue content represents the number of customers in the queue (i.e., not being served), whereas the system content equals the number of customers in the total system. In other words, the system content is equal to the queue content plus the number of customers in service.

3 Moments of the customer delay

3.1 PGF of the customer delay

The delay W of a randomly tagged customer consists of two parts: the first part (W_1) is the time required to serve the 'older' batches and the second (W_2) is the time needed, starting at the end of the first part, until the batch containing the tagged customer is completely filled. Let us consider the example depicted in Fig. 1. The tagged customer is indicated by T, J represents its arrival slot, and the number of customers in the system in front of (behind) T at the beginning of slot $J+1$ is denoted by F (M) and is equal to 43 (2) in the example. Furthermore, we assume that the server capacity c equals 10 in this example. W_1 then equals $\lfloor \frac{F}{c} \rfloor = 4$ slots ($\lfloor . \rfloor$ represents the floor function, i.e. $\lfloor x \rfloor = \max\{n \in \mathbb{N} \mid n \leq x\}$). Whether W_2 is zero or not depends on the system content at the end of the first part, which we designate by P. At that time, the system contains the customers in front of the tagged customer T (F mod c with 'mod' the modulo operator), T itself and the customers in the queue behind T ($M + \sum_{k=1}^{W_1} A_{J+k}$), leading to $P = (F \bmod c) + 1 + M + \sum_{k=1}^{W_1} A_{J+k}$. Since $P = 9 < c = 10$ in this example, the server waits to initiate a new service. After two slots of waiting, 10 customers have accumulated in the system

and the batch containing T is processed ($W_2 = 2$ in the example).

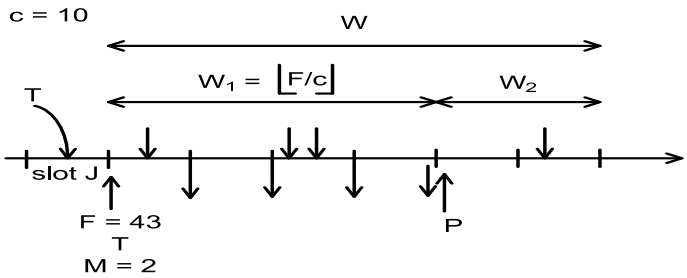


Fig. 1 Example of the two parts of the customer delay and illustration of some notations and relations between them

As demonstrated in the introductory example, $W = W_1 + W_2$, so that

$$W(z) \triangleq \mathrm{E}\left[z^W\right] = \mathrm{E}\left[z^{W_1} z^{W_2}\right] \quad .$$

The introductory example further illustrates that W_2 depends on W_1 through P and, as a result, they are independent if P is given. Exploiting this key observation and taking into account that $P \geq 1$ (because P contains at least the tagged customer), produces:

$$\begin{aligned} \mathrm{E}\left[z^{W_1} z^{W_2}\right] &= \sum_{p=1}^{\infty} \Pr\left[P = p\right] \mathrm{E}\left[z^{W_1} z^{W_2} | P = p\right] \\ &= \sum_{p=1}^{\infty} \Pr\left[P = p\right] \mathrm{E}\left[z^{W_1} | P = p\right] \mathrm{E}\left[z^{W_2} | P = p\right] \quad . \end{aligned} \tag{1}$$

In the subsequent lemmas, expressions are obtained for $\mathrm{E}\left[z^{W_2} | P = p\right]$ and $\mathrm{E}\left[z^{W_1} x^P\right]$. These will then be used in (1), which will yield the final formula for $W(z)$.

Lemma 1

$$\mathrm{E}\left[z^{W_2} | P = p\right] = 1 + (z-1) \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{x^p}{1 - zA(x)} \bigg|_{x=0} \quad . \tag{2}$$

Proof Let us start from the following relation:

$$\Pr[W_2 > m | P = p] = \Pr[p + \tilde{A}_1 + \cdots + \tilde{A}_m < c] \ , \qquad m \geq 0 \ , \tag{3}$$

with $\tilde{A}_k$ the number of customer arrivals during the kth slot after the first part of the tagged customer's delay. Multiplying both sides of (3) by z^m and

summing over all m produces

$$\begin{aligned}\frac{\mathrm{E}\left[z^{W_2}|P=p\right]-1}{z-1} &= \sum_{m=0}^{\infty} z^m \sum_{n=0}^{c-1} \Pr\left[p+\tilde{A}_1+\cdots+\tilde{A}_m=n\right] \\ &= \sum_{m=0}^{\infty} z^m \sum_{n=0}^{c-1} \frac{1}{n!}\frac{\partial^n}{\partial x^n} x^p A(x)^m \Big|_{x=0} \\ &= \sum_{n=0}^{c-1} \frac{1}{n!}\frac{\partial^n}{\partial x^n} \frac{x^p}{1-zA(x)}\Big|_{x=0} \quad ,\end{aligned}$$

which yields equation (2). Mark that the second step follows from the probability generating property of PGF's and the fact that $x^p A(x)^m$ is the PGF corresponding to $p+\tilde{A}_1+\cdots+\tilde{A}_m$ (due to the IID nature of the numbers of consecutive customer arrivals). Furthermore, the last step requires that $|zA(x)|<1$ in the neighbourhood of $x=0$, which is valid for at least every $z\leq 1$, because $|A(0)|<1$. □

As mentioned above, an expression for $\mathrm{E}\left[z^{W_1}x^P\right]$ is established in the following lemma:

Lemma 2

$$\mathrm{E}\left[z^{W_1}x^P\right] = \frac{x}{c}\sum_{m=0}^{c-1} D(u(z,x)\varepsilon_m,x)\frac{u(z,x)^c-x^c}{u(z,x)\varepsilon_m-x}\frac{u(z,x)\varepsilon_m}{u(z,x)^c} \quad , \tag{4}$$

with

- $D(z,x)\triangleq \mathrm{E}\left[z^F x^M\right]$,
- $z^{1/c}\triangleq |z|^{1/c}e^{\imath Arg(z)/c}$, *whereby* $\imath$ *characterises the imaginary unit,* $|z|$ *is the absolute value of* z *and* $Arg(z)$ *represents the principal value of the argument of* z *(i.e. it is a mapping in the interval* $]-\pi,\pi]$*),*
- $u(z,x)\triangleq (zA(x))^{1/c}$,
- $\varepsilon_m = e^{\imath 2\pi m/c}$ $,0\leq m\leq c-1$, *i.e. the consecutive* ε_m*'s constitute the* c *complex* c^{th} *roots of unity.*

Proof Let us denote the mass function of $D(z,x)$ by $d(n,m)\triangleq \Pr\left[F=n,M=m\right]$. Invoking the relations $W_1=\left\lfloor\frac{F}{c}\right\rfloor$ and $P=(F \bmod c)+1+M+\sum_{k=1}^{W_1}A_{J+k}$, we have that

$$\mathrm{E}\left[z^{W_1}x^P\right] = x\sum_{n=0}^{\infty}\sum_{l=0}^{c-1}\sum_{k=0}^{\infty} d(nc+l,k)z^n x^l x^k A(x)^n \quad .$$

In order to relate $\mathrm{E}\left[z^{W_1}x^P\right]$ with $D(z,x)$, we make use of the sifting property of the Kronecker delta function $\delta(.)$ ($\delta(.)$ is 1 when its argument is zero and 0 otherwise)

$$z^n = \sum_{j=0}^{c-1}\left(z^{1/c}\right)^{nc+l-j}\delta(l-j) \quad ,\forall(n,l)\in\mathbb{N}\times[0,c-1] \quad .$$

and the following relation between the Kronecker delta function and the c complex c^{th} roots of unity:

$$\delta(l-j) = \frac{1}{c}\sum_{m=0}^{c-1} \varepsilon_m^{nc+l-j} \quad ,\forall(n,l,j) \in \mathbb{N}^3 \quad .$$

We now obtain subsequently for $\mathrm{E}\left[z^{W_1}x^P\right]$:

$$\begin{aligned}
&\mathrm{E}\left[z^{W_1}x^P\right] \\
&= x\sum_{n=0}^{\infty}\sum_{l=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1} d(nc+l,k)u(z,x)^{nc+l-j}x^j x^k \delta(l-j) \\
&= x\sum_{n=0}^{\infty}\sum_{l=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1} d(nc+l,k)u(z,x)^{nc+l-j}x^j x^k \sum_{m=0}^{c-1}\frac{1}{c}\varepsilon_m^{nc+l-j} \\
&= \frac{x}{c}\sum_{m=0}^{c-1} D(u(z,x)\varepsilon_m, x)\sum_{j=0}^{c-1} u(z,x)^{-j}x^j\varepsilon_m^{-j} \quad .
\end{aligned}$$

Working out the second sum yields (4). □

Corollary 1

$$W_1(z) \triangleq \mathrm{E}\left[z^{W_1}\right] = \frac{1}{c}\sum_{m=0}^{c-1} D(z^{1/c}\varepsilon_m, 1)\frac{z-1}{z^{1/c}\varepsilon_m - 1}\frac{z^{1/c}\varepsilon_m}{z} \quad . \tag{5}$$

Proof Letting $x \to 1$ in (4) and making use of the definition of $u(z,x)$, yields formula (5). □

Now, it is obvious that an expression for $D(z,x)$ has to be established. This expression is provided in the following lemma:

Lemma 3

$$D(z,x) = Q(z)\frac{A(z)-A(x)}{\lambda(z-x)} \quad , \tag{6}$$

with $Q(z)$ the PGF corresponding to Q, the queue content at the beginning of a random slot in the steady state:

$$Q(z) = \frac{(1-\rho)(z^c-1)}{z^c - A(z)}\prod_{j=1}^{c-1}\frac{z-z_j}{1-z_j} \quad , \tag{7}$$

whereby the z_j's, $1 \le j \le c-1$, are the complex roots of $z^c - A(z)$ in the open complex unit disk $\{z : z \in \mathbb{C}, |z| < 1\}$.

Proof Let us denote the queue content at the beginning of slot J by Q_J and let B be the number of customer arrivals during slot J and before T. We then have that $F = Q_J + B$. Further, due to the IID character of the consecutive numbers of customer arrivals, and since M is also equal to the number of customer arrivals during slot J after the tagged customer, Q_J is independent of B and M and Q_J has the same distribution as Q. Hence,

$$D(z,x) = \mathrm{E}\left[z^F x^M\right] = \mathrm{E}\left[z^{Q_J}\right]\mathrm{E}\left[z^B x^M\right] = Q(z)\mathrm{E}\left[z^B x^M\right] \quad . \tag{8}$$

We can extract $Q(z)$ from the PGF $U(z)$ of the system content (i.e. including the customers in service) at the beginning of a random slot. Indeed, the system content at the beginning of an arbitrary slot is equal to the queue content at the previous slot plus the number of customer arrivals during the previous slot. The queue content of the previous slot is equally distributed as the queue content at the next slot due to the steady state. Hence, $Q(z) = U(z)/A(z)$. Making use of the expression for $U(z)$ found in Claeys et al. 2007, yields (7). Furthermore, $\mathrm{E}\left[z^B x^M\right]$ in expression (8) can be obtained by taking into account that an arbitrary customer is more likely to arrive in a slot with more customer arrivals (see e.g. Bruneel and Kim 1993):

$$\mathrm{E}\left[z^B x^M\right] = \frac{A(z) - A(x)}{\lambda(z-x)} \quad ,$$

finally leading to (6). □

Remark 1 The formula for the PGF of the system content from Claeys et al. 2008 is valid under the assumption that the highest common factor of the set of integers $\{\{c\} \cup \{n \in \mathbb{N} : \Pr[A = n] \neq 0\}\}$ equals 1 (this is usually the case).

Now, substitution of (2) in (1), using the probability generating property of PGF's, and taking into account (4) and (5) yields the final expression for $W(z)$:

Theorem 1

$$W(z) = W_1(z) + (z-1)\sum_{n=0}^{c-1}\frac{1}{n!}\frac{\partial^n}{\partial x^n}\frac{\mathrm{E}\left[z^{W_1}x^P\right]}{1 - zA(x)}\bigg|_{x=0} \quad , \tag{9}$$

with $W_1(z)$ given by (5) and $\mathrm{E}\left[z^{W_1}x^P\right]$ given by (4).

3.2 Extracting moments from the PGF

In this section, we compute the mean and the variance of the customer delay by applying the moment generating property of PGF's to (9).

Theorem 2 *The mean value of the customer delay equals*

$$\mathrm{E}[W] = \frac{1}{\lambda}\left[\sum_{j=1}^{c-1}\frac{1}{1-z_j} + \frac{A''(1) - \lambda(c-1)}{2(c-\lambda)}\right] \quad . \tag{10}$$

Proof Taking the first derivative of (9) at $z = 1$, yields:

$$\mathrm{E}\left[W\right] = \mathrm{E}\left[W_1\right] + \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{\mathrm{E}\left[x^P\right]}{1 - A(x)} \bigg|_{x=0} \quad . \tag{11}$$

$\mathrm{E}\left[W_1\right]$, in turn, is found by taking the first derivative of (5) at $z = 1$:

$$\mathrm{E}\left[W_1\right] = \frac{-\lambda c + 2\mathrm{E}\left[Q\right]\lambda + \lambda + A^{''}(1)}{2\lambda c} \quad . \tag{12}$$

In the latter expression, we have exploited that $Q(\varepsilon_m) = 0$ if $m \neq 0$. Further, applying the moment generating property to (7) yields:

$$\mathrm{E}\left[Q\right] = \sum_{j=1}^{c-1} \frac{1}{1 - z_j} - \frac{\lambda(c-1)}{2(c-\lambda)} + \frac{A^{''}(1)}{2(c-\lambda)} \quad . \tag{13}$$

Substituting (12) and (13) in (11) and taking into account that

$$\sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{\mathrm{E}\left[x^P\right]}{1 - A(x)} \bigg|_{x=0} = \frac{c-\lambda}{c\lambda} \sum_{j=1}^{c-1} \frac{1}{1 - z_j} \quad ,$$

(see appendix A for a proof of this identity), produces:

$$\begin{aligned} \mathrm{E}\left[W\right] = & \frac{-\lambda c}{2\lambda c} + \frac{1}{c} \sum_{j=1}^{c-1} \frac{1}{1 - z_j} - \frac{\lambda(c-1)}{2c(c-\lambda)} + \frac{A^{''}(1)}{2c(c-\lambda)} \\ & + \frac{\lambda}{2\lambda c} + \frac{A^{''}(1)}{2\lambda c} + \frac{c-\lambda}{c\lambda} \sum_{j=1}^{c-1} \frac{1}{1 - z_j} \quad . \end{aligned}$$

Rearrangement of the terms leads to (10). □

Remark 2 Comparison of formula (10) for $\mathrm{E}\left[W\right]$ with expression (13) for $\mathrm{E}\left[Q\right]$ learns us that $\mathrm{E}\left[W\right] = \mathrm{E}\left[Q\right]/\lambda$, which is in agreement with Little's law (see e.g. Fiems and Bruneel 2002).

The variance of the customer delay can also be obtained by applying the moment generating property of PGF's to (9):

$$\mathrm{Var}\left[W\right] = W^{''}(1) + \mathrm{E}\left[W\right] - \mathrm{E}\left[W\right]^2 \quad .$$

The second derivative of (9) at $z = 1$ reads:

$$\begin{aligned} & W^{''}(1) \\ & = W_1^{''}(1) + 2 \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{(1 - A(x)) \frac{\partial}{\partial z} \mathrm{E}\left[z^{W_1} x^P\right]\big|_{z=1} + A(x)\mathrm{E}\left[x^P\right]}{(1 - A(x))^2} \bigg|_{x=0} \quad . \end{aligned}$$

Computing the right-hand-side of this equation is difficult. In order to avoid this calculation, we follow an alternative route, leading to the following theorem:

Theorem 3 *The second derivative of $W(z)$ at $z=1$ can be written as*

$$W^{''}(1) = W_1^{''}(1) + 2\sum_{p=1}^{c-1}\sum_{w=0}^{\infty}\theta(p,w)\left[w\sum_{n=0}^{c-1-p}\frac{1}{n!}\frac{d^n}{dx^n}\frac{1}{1-A(x)}\bigg|_{x=0} + \sum_{n=0}^{c-1-p}\frac{1}{n!}\frac{d^n}{dx^n}\frac{A(x)}{(1-A(x))^2}\bigg|_{x=0}\right], \quad (14)$$

with

$$\theta(p,w) \triangleq \Pr\left[P=p, W_1=w\right] \quad , \quad (15)$$

$$\begin{aligned} W_1^{''}(1) = & \frac{12\mathrm{E}\left[Q\right]\lambda - 12\mathrm{E}\left[Q\right]\lambda c + 6A^{''}(1) + 2A^{'''}(1) + 6Q^{''}(1)\lambda}{6c^2\lambda} \\ & + \frac{6\mathrm{E}\left[Q\right](1)A^{''}(1) + 5\lambda c^2 - 6A^{''}(1)c + \lambda - 6c\lambda}{6c^2\lambda} \\ & + \frac{2}{c^2}\sum_{m=1}^{c-1}\frac{Q^{'}(\epsilon_m)\left[A(\epsilon_m)-1\right]\epsilon_m^2}{\lambda(\epsilon_m-1)^2} \quad , \end{aligned}$$

with $\mathrm{E}\left[Q\right]$ *given by (13),*

$$\begin{aligned} Q^{''}(1) = \frac{1}{6(c-\lambda)^2} & \left[6\left\{c\lambda + c\lambda^2 - \lambda^2 + A^{''}(1)c - c^2\lambda - A^{''}(1)\lambda\right\}\sum_{j=1}^{c-1}\frac{1}{1-z_j}\right. \\ & +6\left\{\lambda^2 - 2c\lambda + c^2\right\}\sum_{k=1}^{c-1}\sum_{l=1,l\neq k}^{c-1}\frac{1}{(1-z_k)(1-z_l)} \\ & +3A^{''}(1)\lambda - 3A^{''}(1)c\lambda - 2\lambda A^{'''}(1) + 4\lambda^2 \\ & -6c\lambda^2 + 2\lambda^2c^2 + c^3\lambda - c\lambda \\ & \left. + 3A^{''}(1)^2 - 3A^{''}(1)c^2 + 3A^{''}(1)c + 2cA^{'''}(1)\right] \quad , \end{aligned}$$

and

$$Q^{'}(\epsilon_m) = \frac{(c-\lambda)}{\epsilon_m(1-A(\epsilon_m))}\prod_{j=1}^{c-1}\frac{\epsilon_m - z_j}{1-z_j} \quad .$$

Proof Invoking (15) and the definition of PGF's transforms (9) into:

$$\begin{aligned} & W(z) \\ & = W_1(z) + (z-1)\sum_{p=1}^{\infty}\sum_{w=0}^{\infty}\theta(p,w)z^w\sum_{n=0}^{c-1}\frac{1}{n!}\frac{\partial^n}{\partial x^n}\frac{x^p}{1-zA(x)}\bigg|_{x=0} \quad . \quad (16) \end{aligned}$$

It is obvious that the nth derivative of $x^p/(1-zA(x))$ in $x=0$ equals zero if $n<p$. Further, by means of Leibniz's rule for the derivative of a product, we can write that

$$\frac{\partial^n}{\partial x^n}\frac{x^p}{1-zA(x)}\bigg|_{x=0} = \sum_{k=0}^{n}\binom{n}{k}\frac{\partial^k}{\partial x^k}x^p\bigg|_{x=0}\frac{\partial^{n-k}}{\partial x^{n-k}}\frac{1}{1-zA(x)}\bigg|_{x=0} \quad . \tag{17}$$

It is clear that the kth derivative in (17) is equal to zero if $k \neq p$. Hence, (17) reduces to

$$\frac{\partial^n}{\partial x^n}\frac{x^p}{1-zA(x)}\bigg|_{x=0} = \begin{cases} \binom{n}{p}p!\,\frac{\partial^{n-p}}{\partial x^{n-p}}\frac{1}{1-zA(x)}\Big|_{x=0} & \text{if } n \geq p \ , \\ 0 & \text{if } n < p \ . \end{cases} \tag{18}$$

Substituting (18) in (16) produces:

$$\begin{aligned} &W(z) \\ &= W_1(z) + (z-1)\sum_{p=1}^{\infty}\sum_{w=0}^{\infty}\theta(p,w)z^w\sum_{n=p}^{c-1}\frac{1}{(n-p)!}\,\frac{\partial^{n-p}}{\partial x^{n-p}}\frac{1}{1-zA(x)}\bigg|_{x=0} \\ &= W_1(z) + (z-1)\sum_{p=1}^{c-1}\sum_{w=0}^{\infty}\theta(p,w)z^w\sum_{n=0}^{c-1-p}\frac{1}{n!}\,\frac{\partial^{n}}{\partial x^{n}}\frac{1}{1-zA(x)}\bigg|_{x=0} \quad . \end{aligned} \tag{19}$$

Taking the second derivative of (19) at $z=1$ leads to (14). Finally, $W_1^{''}(1)$ is obtained by taking the second derivative of (5) at $z=1$ and $Q^{''}(1)$ and $Q^{'}(\varepsilon_m)$ can analogously be deduced by taking the appropriate derivative of (7). □

We can calculate $\theta(p,w)$ by inverting the PGF $\mathrm{E}\left[z^{W_1}x^P\right]$ (e.g. with the inverse discrete Fast Fourier Transform), given in (4).

Remark 3 The previous analysis simplifies considerably in the case of a Bernoulli distribution of the number of customer arrivals in an arbitrary slot:

- $Q(z) = \frac{z^c-1}{c(z-1)}$. This is because the denominator of $Q(z)$, $z^c - A(z)$, becomes a polynomial of degree c, of which we know the zeroes: $z_0 = 1, z_1, z_2, \ldots z_{c-1}$.
- $D(z,x) = Q(z) = \frac{z^c-1}{c(z-1)}$, because at most 1 customer can arrive in a slot and the tagged customer arrives during slot J.
- $W_1 = 0$. Indeed, due to the single-slot service times and Bernoulli arrivals, the queue cannot contain c or more customers upon an arrival of a customer. When the cth customer of a batch (i.e. the last customer of that batch) arrives during a slot, the c customers are served during the next slot, so that a possibly arriving customer during the latter slot finds an empty queue. This property can be verified by using $Q(z) = \frac{z^c-1}{c(z-1)}$, $A(z) = 1-\lambda+\lambda z$ and the property that $\frac{c}{(1-z)^2} = \frac{1}{cz}\sum_{m=0}^{c-1}\frac{z^{1/c}\varepsilon_m}{(1-z^{1/c}\varepsilon_m)^2}$ (see Appendix B) in (5).
- W_2 given P equals the sum of $c-p$ geometrically distributed interarrival times, if $p \leq c$ and $W_2 = 0$ else. This property can be verified by substituting $A(x)$ by $1-\lambda+\lambda x$ in (2).

Summarized, the PGF of the customer delay in case of a Bernoulli distribution reduces to

$$W(z) = \frac{1-(1-\lambda)z}{c(1-z)} \frac{[1-(1-\lambda)z]^c - (\lambda z)^c}{[1-(1-\lambda)z]} \quad ,$$

so that

$$\mathrm{E}\,[W] = \frac{c-1}{2\lambda} \quad ,$$

and

$$\mathrm{Var}\,[W] = \frac{-7+c^2-6c\lambda+6\lambda+6c}{12\lambda^2} \quad .$$

3.3 Comparison of the moments for several distributions for the number of customer arrivals in a slot

In this section, we compare the mean and the variance of the customer delay for the following distributions of the number of customer arrivals during an arbitrary slot:

- Bernoulli: $A(z) = 1-\lambda+\lambda z$
- Poisson: $A(z) = e^{\lambda(z-1)}$
- Geometric: $A(z) = 1/(1+\lambda-\lambda z)$
- The 'c-centered' distribution:

$$A(z) = \frac{c-\lambda}{c} + \frac{\lambda}{2c}(z^{c-1}+z^{c+1}) \quad .$$

  When customers arrive in this case, either $c-1$ or $c+1$ customers arrive and this occurs with an equal probability.

All cases lead to a mean number of per-slot arrivals of λ. The mean and variance of the customer delay are plotted versus the load ρ for these distributions in Fig. 2. The server capacity equals 10. We observe that the resulting $\mathrm{E}\,[W]$'s are nearly equal in the case of small load, while they differ considerably for larger values of the load. Further, the resulting $\mathrm{Var}\,[W]$'s differ even when the load is small. The moments of the customer delay in cases whereby several customers can arrive during a slot, can thus not be approximated well by their corresponding moments in case of Bernoulli arrivals (which are easier to calculate) or even Poisson arrivals. Finally, the curves corresponding to the Bernoulli distribution stop at $\rho = 0.1$. Indeed, $\lambda = 1$ in this case and λ also represents the probability that a customer arrives during a random slot in this case and probabilities cannot exceed 1. Hence, the moments of the customer delay in case of batch arrivals can thus not be approximated at all by their Bernoulli equivalents when $\lambda > 1$.

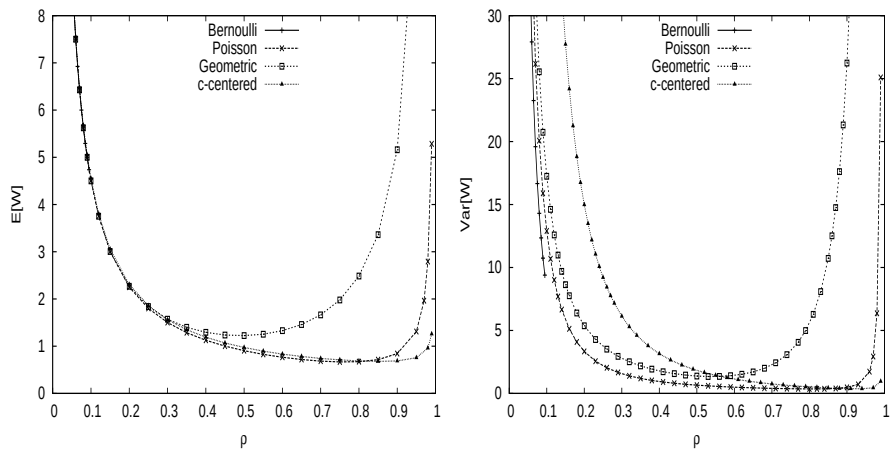


Fig. 2 $\mathrm{E}[W]$ and $\mathrm{Var}[W]$ versus ρ for several distributions of the number of customer arrivals during a slot

4 Tail probabilities of the customer delay

4.1 Approximation of the tail probabilities

The PGF of the customer delay (formula (9)) is not useful to obtain tail probabilities (only in the case of Bernoulli arrivals, it is quite straightforward to extract tail probabilities). Therefore, we resort to an alternative, approximate analysis. To this end, we redefine W_2 and call it $\tilde{W}_2$. As above, $\tilde{W}_2$ represents the time until the service batch with the tagged customer is completely filled, but now $\tilde{W}_2$ starts at the beginning of slot $J+1$, instead of after the first part of the customer delay. Note that $\tilde{W}_2$ can be smaller than W_1, when enough customers have arrived to fill the served batch of the tagged customer before the preceding batches are served (see e.g. the example depicted in Fig. 3). This implies that

$$W = \max(W_1, \tilde{W}_2) \ ,$$

which yields

$$\begin{aligned}\Pr[W > w] &= \Pr\left[W_1 > w \vee \tilde{W}_2 > w\right] \\ &= \Pr[W_1 > w] + \Pr\left[\tilde{W}_2 > w\right] - \Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right] \ .\end{aligned}$$

The major difficulty is the calculation of $\Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right]$. However, one intuitively expects that if w is large, this joint probability is small as compared to the marginal probabilities. Therefore, we equate this with zero, which leads to the following approximation formula:

$$\Pr[W > w] \simeq \Pr[W_1 > w] + \Pr\left[\tilde{W}_2 > w\right] \ . \tag{20}$$

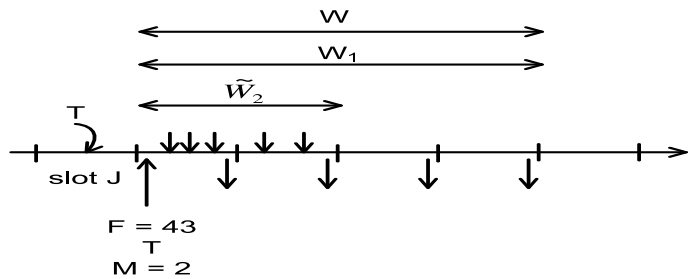


Fig. 3 Example of the two parts of the customer delay ($c = 10$) with the new definition of the second part of the customer delay

We evaluate this approximation in the next subsection. Let us first calculate both marginal tail probabilities separately in the subsequent theorems.

Theorem 4 *(i) $z^c - A(z)$ has one zero $\tilde{z}$ with smallest modulus outside the unit circle. This zero is a positive real number and it has multiplicity one.*

$$(ii)\ \Pr[W_1 > w] \simeq \frac{-\tilde{z}^{-(w+1)c}\,[A(\tilde{z}) - 1]\,(1-\rho)(\tilde{z}^c - 1)\prod_{i=1}^{c-1}\frac{\tilde{z}-z_i}{1-z_i}}{\lambda(\tilde{z}-1)^2\,[\tilde{z}^{c-1}c - A'(\tilde{z})]} \quad . \tag{21}$$

Proof (i) This is proved in Bruneel et al. 1994.
(ii) We obtain the following formula for $W_1(z^c)$ by substituting x by 1 and z by z^c in formula (5) and further substituting $Q(z)$ by its expression (7):

$$W_1(z^c) = \frac{(z^c-1)^2(1-\rho)z}{c\lambda z^c}\sum_{m=0}^{c-1}\frac{[A(z\varepsilon_m)-1]\,\varepsilon_m}{[z^c - A(z\varepsilon_m)]\,[z\varepsilon_m - 1]^2}\prod_{i=1}^{c-1}\frac{z\varepsilon_m - z_i}{1-z_i} \quad . \tag{22}$$

On account of (i), we have that $\tilde{z}\varepsilon_m^{-1}$ is a zero of $z^c - A(z\varepsilon_m)$ and that this zero has the same modulus as $\tilde{z}$, and this for all m, $0 \le m \le c-1$. Further, the equation $z^c - A(z\varepsilon_m)$ has no other zero $\hat{z}$ with modulus smaller than $\tilde{z}$, because $\hat{z}\varepsilon_m$ would then be a zero of $z^c - A(z)$ with a smaller modulus than $\tilde{z}$. As a result, $W_1(z^c)$ has c dominant poles $\tilde{z}\varepsilon_m^{-1}$ ($m = 0, \ldots, c-1$). In a similar manner as in Bruneel and Kim 1993, section 4.1.3.2, we obtain approximation formula (21) (for more details we refer to appendix C): □

An expression for $\Pr\left[\tilde{W}_2 > w\right]$ is deduced in the following theorem:

Theorem 5

$$\Pr\left[\tilde{W}_2 > w\right] = \sum_{m=0}^{c-2}\frac{1}{m!}\frac{\partial^m}{\partial z^m}A(z)^w\frac{z^c-1}{c(z-1)}\frac{A(z)-1}{\lambda(z-1)}\bigg|_{z=0} \quad . \tag{23}$$

Proof We start from the following relation (we refer to Fig. 3 for a reminder of the notations):

$$\Pr\left[\tilde{W}_2 > w\right] = \Pr\left[(F \bmod c) + 1 + M + \sum_{i=1}^{w} A_{J+i} < c\right] \quad . \tag{24}$$

In order to compute the right-hand-side of (24), we make use of the probability generating property of PGF's. Therefore, we first compute $\mathrm{E}\left[z^{(F \bmod c)} z^M\right]$. Along the same lines as we deduced $\mathrm{E}\left[z^{W_1} x^P\right]$, we find

$$\mathrm{E}\left[z^{(F \bmod c)} z^M\right] = \frac{1}{c} \sum_{j=0}^{c-1} \varepsilon_j \frac{z^c - 1}{z - \varepsilon_j} D(\varepsilon_j, z) \ . \tag{25}$$

Since

$$D(\varepsilon_m, z) = \begin{cases} \frac{A(z)-1}{\lambda(z-1)} & \text{if } m = 0 \ , \\ 0 & \text{if } m \neq 0 \ , \end{cases}$$

(see (6)), (25) transforms into:

$$\mathrm{E}\left[z^{(F \bmod c)} z^M\right] = \frac{z^c - 1}{c(z-1)} \frac{A(z) - 1}{\lambda(z-1)} \ . \tag{26}$$

Applying the probability generating property of PGF's to (24) and appealing to (26), produces

$$\Pr\left[\tilde{W}_2 > w\right] = \sum_{m=1}^{c-1} \frac{1}{m!} \frac{\partial^m}{\partial z^m} z A(z)^w \frac{z^c - 1}{c(z-1)} \frac{A(z) - 1}{\lambda(z-1)} \bigg|_{z=0} \ .$$

Applying Leibniz's rule for the derivative of a product finally yields (23). □

4.2 Evaluation of the approximation

In this section, we evaluate approximation formula (20). In fact, two types of possible inaccuracies appear in the approximation formula:

- The inaccuracy stemming from the approximation of $\Pr[W_1 > w]$, based on the dominant poles of $W_1(z^c)$.
- The error created by omitting the joint probability $\Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right]$.

Since it is generally known that approximations based on dominant poles are extremely accurate for $w \gg 1$ (see e.g. Bruneel and Kim 1993, section 4.1.3), we evaluate the second error type. Since the absolute value of the relative error

equals

$$\left|\frac{\Pr\left[W_1 > w\right] + \Pr\left[\tilde{W}_2 > w\right] - \Pr\left[W > w\right]}{\Pr\left[W > w\right]}\right|$$

$$= \left|\frac{\Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right]}{\Pr\left[W_1 > w\right] + \Pr\left[\tilde{W}_2 > w\right] - \Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right]}\right|$$

$$= \frac{1}{\frac{\Pr[W_1>w]+\Pr\left[\tilde{W}_2>w\right]}{\Pr\left[W_1>w\wedge\tilde{W}_2>w\right]} - 1}$$

$$\leq \frac{1}{\frac{\Pr[W_1>w]+\Pr\left[\tilde{W}_2>w\right]}{\min\left(\Pr[W_1>w],\Pr\left[\tilde{W}_2>w\right]\right)} - 1}$$

$$= \frac{\min\left(\Pr\left[W_1 > w\right], \Pr\left[\tilde{W}_2 > w\right]\right)}{\Pr\left[W_1 > w\right] + \Pr\left[\tilde{W}_2 > w\right] - \min\left(\Pr\left[W_1 > w\right], \Pr\left[\tilde{W}_2 > w\right]\right)} \ ,$$

we obtain the following upper bound β for the relative error:

$$\beta \triangleq \frac{\min\left(\Pr\left[W_1 > w\right], \Pr\left[\tilde{W}_2 > w\right]\right)}{\Pr\left[W_1 > w\right] + \Pr\left[\tilde{W}_2 > w\right] - \min\left(\Pr\left[W_1 > w\right], \Pr\left[\tilde{W}_2 > w\right]\right)} \ .$$

Note that $0 < \beta \leq 1$ and that $\beta = 1$ if and only if $\Pr\left[W_1 > w\right] = \Pr\left[W_2 > w\right]$. Also, $\Pr\left[W_1 > w\right] \ll \Pr\left[W_2 > w\right]$ or reversibly, leads to $\beta \ll 1$. In light-traffic situations (i.e. the load $\rho \to 0$), customers arrive rarely, so that the delay of a random customer is usually dominated by the second part, leading to $\Pr\left[W_1 > w\right] \ll \Pr\left[W_2 > w\right]$. On the other hand, customers arrive frequently in heavy-traffic circumstances (i.e. $\rho \to 1$), so that the delay is typically dominated by the first part, leading to $\Pr\left[W_1 > w\right] \gg \Pr\left[W_2 > w\right]$. In these cases, $\beta \ll 1$, so that the approximation is certainly accurate. As we expect that $\Pr\left[W_1 > w\right]$ increases as ρ increases, whereas $\Pr\left[W_2 > w\right]$ decreases, we expect that β first increases as a function of ρ, until β reaches one, and then β decreases when ρ increases. As a result, we also expect that there exists an interval wherein $\Pr\left[W_1 > w\right] \approx \Pr\left[W_2 > w\right]$, so that $\beta \approx 1$, implying that the approximation might be (but is not certain to be) inacurrate.
In order to exemplify these issues and to examine the magnitude of this interval (we assume that a relative error smaller than 10^{-3} is accurate), β is plotted versus the load in Fig. 4 for w-values of 10, 30 and 50 and for various distributions of the number of customer arrivals per slot. We observe that the approximation is indeed extremely accurate for $\rho \to 0$ and $\rho \to 1$ and there exists an interval wherein the approximation formula might be inaccurate. We see, however, that this interval is extremely small and that its length decreases as w increases. Finally, we observe that the position of the interval and the magnitude of β are highly dependent on the distribution of the number of

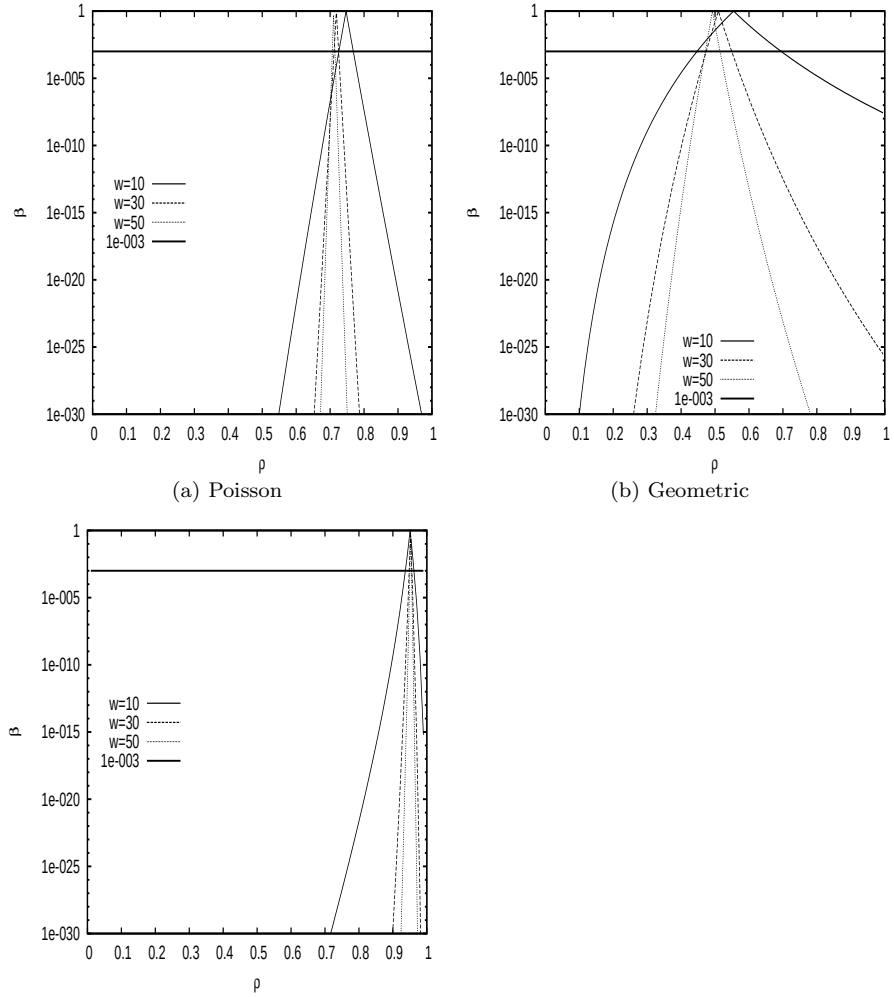
customer arrivals in a slot. Comparison with Fig. 2 learns that the position where $\beta = 1$ is close to the position where $\mathrm{E}[W]$ and $\mathrm{Var}[W]$ reach their minimum. The reason that $\mathrm{E}[W]$ and $\mathrm{Var}[W]$ increase after reaching their minimum is that W_1 becomes larger and starts to dominate over W_2. Before the minimum W_2 dominates. Thus the maximum of β is indeed expected to occur for the same ρ that $\mathrm{E}[W]$ and $\mathrm{Var}[W]$ reach their minimum. We also observe that the interval is larger when the load corresponding to $\beta = 1$ is smaller. The asymptote of $\mathrm{E}[W]$ and $\mathrm{Var}[W]$ for $\rho \to 1$ is then further away of their minimum, implying that W (and thus W_1) increases softly just after the minimum. Therefore, W_1 dominates only mildly for a wider region of ρ and the interval of possible inaccurateness is larger. Either way, the approximation is very accurate except for possibly a small interval in the load. The approximation is also better when w is larger.

4.3 Comparison of the tail probabilities for several distributions of the number of customer arrivals in a slot

In section 3.3, we have observed that the mean and the variance of the customer delay are significantly influenced by the distribution of the number of customer arrivals in an arbitrary slot. In this section, we study whether this is also the case for the probabilities $\Pr[W > w]$. In order to do so, we plot the approximation of $\Pr[W > w]$ versus w in Fig. 5 for several distributions of the number of customer arrivals per slot and for $\rho = 0.3$ (part a) and for $\rho = 0.8$ (part b). We see that the speed by which the probabilities $\Pr[W > w]$ decrease as a function of w differs significantly. In addition, we observe, in general, that when the mean and the variance of W are largest for some distribution, the probabilities decrease slower. For instance, we observe that $\Pr[W > w]$ decreases faster for the geometric distribution than for the c-centered distribution when $\rho = 0.3$, while it is the other way around for $\rho = 0.8$.

5 Conclusions

In this paper, we have deduced moments and accurate tail probabilities of the customer delay in a discrete-time queueing system with batch arrivals and a batch server that always processes at full capacity. In order to analyze the moments, we have conceived the customer delay as the sum of two non-overlapping parts, whereas for the tail probabilities, it has turned out to be more practical to interpret the delay as the maximum of two time periods. We have also demonstrated that the distribution of the number of customer arrivals in a random slot has a significant impact on the moments and the tail probabilities of the customer delay.



(a) Poisson (b) Geometric

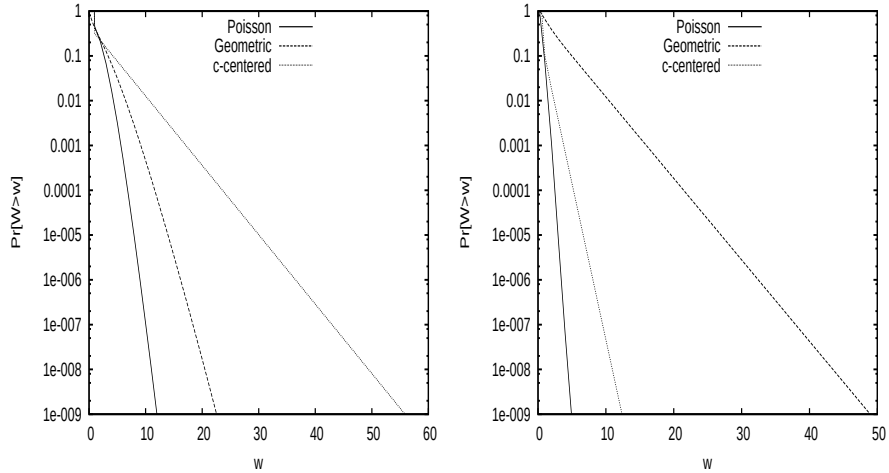
(c) c-centered

Fig. 4 β versus ρ for several values of w and various distributions of the number of customer arrivals during a slot

A Proof of $\sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{\mathrm{E}\left[x^P\right]}{1-A(x)}\Big|_{x=0} = \frac{c-\lambda}{c\lambda} \sum_{j=1}^{c-1} \frac{1}{1-z_j}$

First, we elaborate on $\mathrm{E}\left[x^P\right]$. Therefore, we substitute z by 1 in (4) and we use that $u(z,x) = (zA(x))^{1/c}$:

$$\begin{aligned}
\mathrm{E}\left[x^P\right] &= \frac{x}{c} \sum_{m=0}^{c-1} D\left(A(x)^{1/c}\epsilon_m, x\right) \frac{A(x)-x^c}{A(x)^{1/c}\epsilon_m - x} \frac{A(x)^{1/c}\epsilon_m}{A(x)} \\
&= \frac{x}{c} \sum_{m=0}^{c-1} Q\left(A(x)^{1/c}\epsilon_m\right) \frac{A\left(A(x)^{1/c}\epsilon_m\right) - A(x)}{\lambda\left[A(x)^{1/c}\epsilon_m - x\right]} \\
&\quad \cdot \frac{A(x)-x^c}{A(x)^{1/c}\epsilon_m - x} \frac{A(x)^{1/c}\epsilon_m}{A(x)} \\
&= \frac{x}{c} \sum_{m=0}^{c-1} \frac{(A(x)-1)(c-\lambda)}{c\left[A(x) - A\left(A(x)^{1/c}\epsilon_m\right)\right]} \prod_{j=1}^{c-1} \frac{A(x)^{1/c}\epsilon_m - z_j}{1-z_j} \\
&\quad \cdot \frac{A\left(A(x)^{1/c}\epsilon_m\right) - A(x)}{\lambda\left[A(x)^{1/c}\epsilon_m - x\right]^2} \frac{A(x)-x^c}{A(x)} A(x)^{1/c}\epsilon_m \ ,
\end{aligned}$$



(a) $\rho = 0.3$ (b) $\rho = 0.8$

Fig. 5 $\Pr[W > w]$ versus w for various distributions of the number of customer arrivals during a slot

whereby we have utilized expressions (6) and (7) of $D(z, x)$ and $Q(z)$ respectively. This leads us to the following expression for $\frac{\mathrm{E}[x^P]}{1-A(x)}$:

$$\frac{\mathrm{E}\left[x^P\right]}{1-A(x)} = \frac{c-\lambda}{c\lambda} x(x^c - A(x)) \cdot \left[\frac{-1}{cA(x)} \sum_{m=0}^{c-1} \frac{\epsilon_m A(x)^{1/c}}{(x-\epsilon_m A(x)^{1/c})^2} \prod_{j=1}^{c-1} \frac{\epsilon_m A(x)^{1/c} - z_j}{1-z_j}\right] . \tag{27}$$

It is obvious that the part between the square brackets is tedious. We will now elaborate on this part. Consider the following partial fraction expansion:

$$\frac{\prod_{j=1}^{c-1} \frac{x-z_j}{1-z_j}}{x^c - t} = \sum_{m=0}^{c-1} \frac{1}{x-\epsilon_m t^{1/c}} \prod_{j=1}^{c-1} \left(\frac{\epsilon_m t^{1/c} - z_j}{1-z_j}\right) \frac{1}{c\left[\epsilon_m t^{1/c}\right]^{c-1}} . \tag{28}$$

Taking the first derivative with respect to x of the right-hand side of this equation gives:

$$\frac{-1}{ct} \sum_{m=0}^{c-1} \frac{\epsilon_m t^{1/c}}{(x-\epsilon_m t^{1/c})^2} \prod_{j=1}^{c-1} \left(\frac{\epsilon_m t^{1/c} - z_j}{1-z_j}\right) .$$

Substituting t by $A(x)$ yields the tedious part between brackets in (27). Hence, this is equal to the first derivative with respect to x of the LHS of (28) evaluated at $t = A(x)$. This is equal to

$$\frac{\prod_{l=1}^{c-1} \frac{x-z_l}{1-z_l}}{x^c - A(x)} \left[\sum_{j=1}^{c-1} \frac{1}{x-z_j} - \frac{cx^{c-1}}{x^c - A(x)}\right] .$$

Hence,

$$\frac{\mathrm{E}\left[x^P\right]}{1-A(x)} = \frac{c-\lambda}{c\lambda} \prod_{l=1}^{c-1} \frac{x-z_l}{1-z_l} \left[\sum_{j=1}^{c-1} \frac{x}{x-z_j} - \frac{cx^c}{x^c - A(x)}\right] .$$

Because $x^c - A(x)$ is the part of the denominator of $U(x)$, the PGF of the system content, that produces the zeroes z_l, we can substitute $x^c - A(x)$ by $\prod_{l=1}^{c-1}(x - z_l)f(x)$, with $f(x)$ a function that has no zeroes in the open complex unit disk $\{z : z \in \mathbb{C}, |z| < 1\}$. Hence:

$$\frac{\mathrm{E}\left[x^P\right]}{1 - A(x)} = \frac{c-\lambda}{c\lambda}\prod_{l=1}^{c-1}\frac{1}{1-z_l}\left[\sum_{j=1}^{c-1} x \prod_{k=1,k\neq j}^{c-1}(x - z_k) - \frac{cx^c}{f(x)}\right] \quad .$$

Note that if we take the nth $(0 \leq n \leq c-1)$ derivative of the previous expression at $x = 0$, the second term vanishes. This implies consequently:

$$\begin{aligned}&\sum_{n=0}^{c-1}\frac{1}{n!}\frac{\partial^n}{\partial x^n}\frac{\mathrm{E}\left[x^P\right]}{1-A(x)}\bigg|_{x=0}\\ &= \frac{c-\lambda}{c\lambda}\prod_{l=1}^{c-1}\frac{1}{1-z_l}\sum_{n=0}^{c-1}\frac{1}{n!}\frac{\partial^n}{\partial x^n}\, x\sum_{j=1}^{c-1}\prod_{k=1,k\neq j}^{c-1}(x-z_k)\bigg|_{x=0} \quad . \end{aligned} \tag{29}$$

We can see that the rightmost side of (29) is the product of $\frac{c-\lambda}{c\lambda}\prod_{l=1}^{c-1}\frac{1}{1-z_l}$ and the Maclaurin expansion of a polynomial $H(z)$ of degree $c-1$, evaluated at $z = 1$, whereby $H(z) \triangleq z\sum_{j=1}^{c-1}\prod_{k=1,k\neq j}^{c-1}(z - z_k)$. This leads to:

$$\begin{aligned}&\sum_{n=0}^{c-1}\frac{1}{n!}\frac{\partial^n}{\partial x^n}\frac{\mathrm{E}\left[x^P\right]}{1-A(x)}\bigg|_{x=0}\\ &= \frac{c-\lambda}{c\lambda}\prod_{l=1}^{c-1}\frac{1}{1-z_l}\sum_{j=1}^{c-1}\prod_{k=1,k\neq j}^{c-1}(1-z_k)\\ &= \frac{c-\lambda}{c\lambda}\sum_{j=1}^{c-1}\frac{1}{1-z_j} \quad .\end{aligned}$$

□

B Proof of $\frac{c}{(1-z)^2} = \frac{1}{cz}\sum_{m=0}^{c-1}\frac{z^{1/c}\varepsilon_m}{(1-z^{1/c}\varepsilon_m)^2}$

By applying partial fraction expansion, we obtain:

$$\frac{1}{x^c - z} = \frac{1}{cz}\sum_{m=0}^{c-1}\frac{z^{1/c}\varepsilon_m}{x - z^{1/c}\varepsilon_m} \quad .$$

Taking the first derivative in z of both hand sides, we find

$$\frac{cx^{c-1}}{(x^c - z)^2} = \frac{1}{cz}\sum_{m=0}^{c-1}\frac{z^{1/c}\varepsilon_m}{\left(x - z^{1/c}\varepsilon_m\right)^2} \quad .$$

Letting $x \to 1$ finally produces:

$$\frac{c}{(1-z)^2} = \frac{1}{cz}\sum_{m=0}^{c-1}\frac{z^{1/c}\varepsilon_m}{\left(1 - z^{1/c}\varepsilon_m\right)^2} \quad .$$

□

C Details about the tail approximation

In Bruneel and Kim 1993, an approximation for $\Pr[X = n]$ (n large) is established for the situation where the PGF corresponding to the random variable X, $X(z)$, has one dominant singularity z^* and that this singularity is a pole:

$$\Pr[X = n] \approx -\mathrm{Res}\left[z^{-1-n}X(z), z^*\right] \ ,$$

with $\mathrm{Res}[f(z), z^*]$ the residue of $f(z)$ in z^*. The germ of their approach is to consider the right-hand-side of $X(z) = \sum_{n=0}^{\infty} \Pr[X = n]\, z^n$ as the Laurent series of the function $X(z)$ and to apply the residue theorem. In this paper, $W_1(z^c)$ has multiple dominant singularities, leading to a sum of residues. In addition, we have to take into account that the powers of the right-hand-side of $W_1(z^c) = \sum_{n=0}^{\infty} \Pr[W_1 = n]\, z^{nc}$ are multiples of c. Consequently, as a residue is the coefficient corresponding to z^{-1} in the Laurent series, we find

$$\Pr[W_1 = n] \approx -\sum_{j=0}^{c-1} \mathrm{Res}\left[z^{-1-nc}W_1(z^c), \tilde{z}\varepsilon_j^{-1}\right] \ . \tag{30}$$

As $\tilde{z}\varepsilon_j^{-1}$ is a pole of $W_1(z^c)$ with multiplicity one, for all $0 \le j \le c-1$, (30) transforms into

$$\Pr[W_1 = n] \approx -\sum_{j=0}^{c-1} \lim_{z \to \tilde{z}\varepsilon_j^{-1}} (z - \tilde{z}\varepsilon_j^{-1}) \frac{W_{1,N}(z^c)}{W_{1,D}(z^c)} z^{-nc-1} \ ,$$

with $W_{1,N}(z)$ the numerator of $W_1(z)$ and $W_{1,D}(z)$ the denominator of $W_1(z)$. Application of l'Hôpital's rule and calculation of the sum leads to

$$\Pr[W_1 = n] \approx -\tilde{z}^{-(n+1)c} \frac{W_{1,N}(\tilde{z}^c)}{\frac{\mathrm{d}}{\mathrm{d}z} W_{1,D}(z)\Big|_{z=\tilde{z}^c}} \ ,$$

and hence

$$\Pr[W_1 > w] \approx -\frac{\tilde{z}^{-(w+1)c}}{\tilde{z}^c - 1} \frac{W_{1,N}(\tilde{z}^c)}{\frac{\mathrm{d}}{\mathrm{d}z} W_{1,D}(z)\Big|_{z=\tilde{z}^c}} \ .$$

Making use of expression (22) yields formula (21).

Acknowledgements The third author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

References

1. Arumuganathan R, Jeyakumar S (2005) Steady state analysis of a bulk queue with multiple vacations, setup times with N-policy and closedown times. Appl Math Model 29:972–986
2. Bailey NTJ (1954) On queueing processes with bulk service. J R Stat Soc 16(1):80–87
3. Bruneel H, Kim BG (1993) Discrete-time models for communication systems including ATM. Kluwer Academic Publishers, Boston/Dordrecht/London
4. Bruneel H, Steyaert B, Desmet E, Petit GH (1994) Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. Eur J Oper Res 76:563–572
5. Chang SH, Choi DW (2005) Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations. Comp Oper Res 32:2213–2234
6. Chang SH, Takine T (2005) Factorization and stochastic decomposition properties in bulk queues with generalized vacations. Queueing Syst 50:165–183

7. Chaudhry ML, Templeton JGC (1983) A first course in bulk queues. John Wiley & Sons
8. Claeys D, Walraevens J, Laevens K, Bruneel H (2007) A discrete-time queueing model with a batch server operating under the minimum batch size rule. Proceedings of the 7th International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN 2007), St. Petersburg, September 10-14, pp.248–259
9. Claeys D, Laevens K, Walraevens J, Bruneel H (2008) Delay in a discrete-time queueing model with batch arrivals and batch services. Proceedings of the Fifth International Conference on Information Technology: New Generations (ITNG 2008), Las Vegas, Nevada, April 7-9, pp.1040–1045
10. Cohen JW (1969) The single server queue. North-Holland, Amsterdam; Wiley Interscience, New York
11. Dagsvik J (1975) The general bulk queue as a matrix factorisation problem of the Wiener-Hopf type. Part 1. Adv Appl Prob 7(3):636–646
12. Downton F (1955) Waiting time in bulk service queues. J R Stat Soc, Series B (Methodological) 17(2):256–261
13. Dümmler MA, Schömig AK (1999) Using discrete-time analysis in the performance evaluation of manufacturing systems. Proceedings of the 1999 International Conference on Semiconductor Manufacturing Operational Modeling and Simulation (SMOMS '99), San Francisco, California, January 18-20
14. Fiems D, Bruneel H (2002) A note on the discretization of Little's result. Oper Res Lett 30:17–18
15. Goswami V, Mohanty JR, Samanta SK (2006) Discrete-time bulk-service queues with accessible and non-accessible batches. Appl Math Comput 182:898–906
16. Gupta UC, Goswami V (2002) Performance analysis of finite buffer discrete-time queue with bulk service. Comp Oper Res 29:1331–1341
17. Keilson J (1962) The general bulk queue as a Hilbert problem. J R Stat Soc, Series B (Methodological) 24(2):344–358
18. Kim NK, Chaudhry ML (2006) Equivalences of batch-service queues and multi-server queues and their complete simple solutions in terms of roots. Stoch Anal Appl 24:753–766
19. Lee HW, Lee SS, Chae KC (1996) A fixed-size batch service queue with vacations. J Appl Math Stoch Anal 9:205–219
20. Medhi J (1975) Waiting time distributions in a Poisson queue with a general bulk service rule. Manag Sci 21(2):777–782
21. Miller RG (1959) A contribution to the theory of bulk queues. J R Stat Soc, Series B (Methodological) 21(2):320–337
22. Neuts MF (1967) A general class of bulk queues with Poisson input. Ann Math Stat 38:759–770
23. Powell WB, Humblet P (1986) The bulk service queue with a general control strategy: theoretical analysis and a new computational procedure. Oper Res 34(2):267–275
24. Samanta SK, Chaudhry ML, Gupta, UC (2007) Discrete-time $Geo^X|G^{(a,b)}|1|N$ queues with single and multiple vacations. Math Comp Model 45:93–108
25. Sikdar K, Gupta UC (2005) Analytic and numerical aspects of batch service queues with single vacation. Comp Oper Res 32:943–966
26. Yi XW, Kim NK, Yoon BK, Chae KC (2007) Analysis of the queue-length distribution for the discrete-time batch-service $Geo^X|G^{a,Y}|1|K$ queue. Eur J Oper Res 181:787–792