Analysis of threshold-based batch-service queueing systems with batch arrivals and general service times

Dieter Claeys*, Joris Walraevens[1], Koenraad Laevens, and Herwig Bruneel

SMACS Research Group, Department TELIN, Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium

Abstract

Most research concerning batch-service queueing systems has focussed on some specific aspect of the buffer content. Further, the customer delay has only been examined in the case of single arrivals. In this paper, we examine three facets of a threshold-based batch-service system with batch arrivals and general service times. First, we compute a fundamental formula from which an entire gamut of known as well as new results regarding the buffer content of batch-service queues can be extracted. Secondly, we produce accurate light- and heavy-traffic approximations for the buffer content. Thirdly, we calculate various quantities with regard to the customer delay. This paper thus provides a whole spectrum of tools to evaluate the performance of batch-service systems.

Key words: batch arrivals, batch service, buffer content, light and heavy traffic, customer delay

1. Introduction

Whereas traditional servers in queueing systems can only serve one customer at a time, batch servers process batches of customers. In fact, a traditional server is a special type of batch server, namely whereby the capacity of the server (the maximum number of customers in a served batch) equals one. Examples of batch servers in real life include elevators in high buildings, transport vehicles, recreational devices in amusement parks, ovens in production processes, ... Furthermore, in telecommunications, it is often the case that information packets are grouped in larger entities (batches) and these batches are transmitted instead of all packets individually. This is mainly done for efficiency reasons,

*Corresponding author. Tel: +32 9 264 3411; fax: +32 9 264 42 95
Email addresses: Dieter.Claeys@telin.ugent.be (Dieter Claeys), Joris.Walraevens@telin.ugent.be (Joris Walraevens), Koenraad.Laevens@telin.ugent.be (Koenraad Laevens), Herwig.Bruneel@telin.ugent.be (and Herwig Bruneel)

[1] The second author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

since only one header per aggregated batch has to be constructed instead of one header per single information unit, thus leading to an increased goodput. Optical burst switched (OBS) networks, for instance, apply this method (see e.g. [10], [22]). At the edges of the optical burst network, IP packets with the same destination and Quality of Service (QoS) requirements are aggregated into optical bursts which are injected into the network.
As a last example of batch service we mention group-testing policies, which are useful in the testing of blood samples (see e.g. [1], [6]). For example, if blood samples have to be tested for the presence of some disease, great time savings might be obtained if a pool of samples is tested instead of all samples individually, especially when the prevalence of the disease is small.
An inherent aspect of batch-service systems is that a newly arriving customer cannot join the ongoing service, even if there is free capacity. Hence, capacity might be lost. In order to restrict this loss, a service threshold can be enforced for the minimum number of customers present before the available batch server is allowed to start processing (see e.g. [20]). The server is said to be dormant (see e.g. [24]) if it is waiting until enough customers have accumulated. In practice, an operator typically has to select an appropriate service threshold and this could have a huge impact on the performance of the system. Therefore, the operator needs tools to evaluate the influence of the service policy on the behaviour of the system.

Bailey [5] was presumably the first to consider a batch-service queueing system. He obtained the steady-state distribution of the buffer content at random time epochs in an $M/G^{(1,c)}/1$ queueing system - the superscript (l, c) indicates that the server capacity is equal to c and that the service threshold equals l $(1 \leq l \leq c)$. Since then, many papers, as well in continuous as in discrete time, have been published about batch-service. Downton [14] examined the customer delay for the same queueing model as in [5]. Neuts [20] studied the buffer content at random time epochs and at service completion times in a $M/G^{(l,c)}/1$ system. Further, the customer delay in an $M/M^{(l,c)}/1$ system was calculated by Medhi [19]. Chaudhry and Templeton [9] deduced the distributions of the buffer content and the customer delay in the systems $M/G^{(1,c)}/1/K$, $M/G^{(1,c)}/1$, $M/G^{(c,c)}/1$ and the discrete $Geo/G^{(1,c)}/1$. Powell and Humblet [21] calculated the buffer content at service completion times in the discrete $Geo^X/G^Y/1$ system for several possible service policies Y and Zhao and Campbell [25] studied the buffer content at random slots in the discrete $Geo^X/1^{(1,c)}/1$ system. Next, Arumuganathan and Jeyakumar [3] examined the buffer content at various time epochs in an $M^X/G^{(l,c)}/1$ model with N-policy, multiple vacations, setup and closedown times. The buffer content at several time epochs in an $M/G^{(l,c)}/1$ system with single vacations was computed by Sikdar and Gupta [24]. Kim and Chaudhry [18] studied equivalences between batch-service and multi-service systems and Samanta et al. [23] derived the buffer content at various time instants in a discrete $Geo^X/G^{(l,c)}/1/K$ system with vacations.

From the above literature overview, it follows that most research concerning batch-service queueing models has focused on some specific aspect of the buffer content and that the customer delay has only been studied in the case of single arrivals. In this paper, we study three novel aspects of the discrete-time $Geo^X/G^{(l,c)}/1$ system. First, we compute a fundamental formula from which an entire gamut of known as well as new results regarding the buffer content of such batch-service queues can be extracted. We also demonstrate that these expressions are useful tools to select a good service threshold (section 3). As in all batch-service systems, the resulting formulas contain unknown probabilities that have to be calculated numerically. This can become an unfeasible assignment especially for large c. Therefore, we deduce light- and heavy-traffic approximations of the buffer content, which require only a few (and sometimes no) numerical calculations (section 4). To the best of our knowledge, light- and heavy-traffic approximations have not been studied before for batch-service systems. Thirdly, we examine the customer delay in this model (section 5). Papers [14], [18] and [19] have studied the customer delay in case of single (Bernoulli) arrivals, which restricts the number of practical applications. We include batch arrivals, which, in our opinion, complicates the analysis considerably. We also demonstrate that the moments of the customer delay differ considerably as compared to the case of single (Bernoulli) arrivals. Hence, batch arrivals have to be taken into account. We would finally like to stress that the analysis, albeit for a discrete-time model, is in our opinion easily transferrable to a continuous-time model. The exact continous-time analysis of our model is to the best of our knowledge not studied in this much detail either (light- and heavy-traffic approximations, customer delay).

This paper is an extension of our conference papers [11] and [12]. In [11], we have computed the probability generating function (PGF) of the system content (the number of customers in the system, including those in service) for the same queueing model as in this paper. In this paper, we deduce, next to the PGF of the system content, several other significant PGF's related to the buffer content, among which of the queue content (i.e. the number of customers in the queue), the server content (i.e. the number of customers in service), the queue content when the server is dormant, et cetera. In addition, we deduce in this paper light- and heavy-traffic approximations of system content characteristics and we analyse the customer delay of this queueing system. In [12], we have studied the customer delay in a batch-service queueing system with batch arrivals for the first time. [12] serves as a starting point, in the sense that the model is very basic: the service threshold l is equal to the server capacity c and single-slot service times are considered. In our paper [13], we have focused on obtaining accurate approximations of the tail probabilities in the same model as in [12]. The main disadvantage of [12] and [13] is that single-slot service times restricts the applicability of the model, as service times are most likely variable in real-life applications. Therefore, the extension to generally distributed service times is crucial. The model in this paper also allows a general service threshold instead of the special case $l = c$ in [12] and [13], which is an important generalization as well. When c is large, it might take a long time to fill a complete batch when $l = c$. Therefore, a smaller l will give better performance in real-life applications.

2. Model Description

This section summarises the properties of the model studied in this paper.

- We consider a discrete-time queueing model, i.e. the time axis is divided into fixed-length contiguous periods, referred to as slots.
- During each slot, several customers can arrive. The number of customers that arrive 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 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, since in most practical applications the queue is large in order to minimize the loss probability.
- There is one batch server of capacity c (c fixed), which means that the server can process up to c customers simultaneously. The available server only starts service if the system contains at least as many customers as the service threshold l ($1 \leq l \leq c$). Hence, if the server finds less than l customers upon becoming available, it waits to start service (the server is then said to be dormant) until the beginning of the first slot that the system contains at least l customers. We assume that the already present customers remain in the queue when the server waits to start service. Hence, during each slot, the system content consists of the customers being served (the server content) and the customers waiting in the queue (the queue content).
- A service period (also called cycle) is the period between the start and end of the service of one batch of customers. The service periods are synchronized to slot marks, in the sense that the server always starts processing at the beginning of a slot and ends at the end of a slot (not necessarily the same one). This yields that an arriving customer has to wait for service at least until the beginning of the next slot. This part of the waiting time is not included in what we denote by the customer delay, since we count the customer delay as an integral number of slots. This kind of synchronisation is also known as LAS-DA (late arrival system with delayed access) (see e.g. [23]).
- A service time is the length of a service period, expressed in a number of slots. The consecutive service times are IID and have a general distribution. The length of a random service time is denoted by T and its PGF by $T(z)$. We assume that $\Pr[T = 0] = 0$.

- The queueing discipline is first-come-first-served (FCFS).

Summarized, this queueing model can thus be denoted by $Geo^X/G^{(l,c)}/1$.

Remark 1. *Most PGF's that occur in practice are analytical in a region that at least includes the closed unit disc $\{z \in \mathbb{C} : |z| \leq 1\}$. Throughout this paper, we assume that this is the case for $A(z)$ and $T(z)$. This implies that all order moments of A and T are finite and can be calculated from their PGF's, for instance* $\lambda \triangleq \mathrm{E}[A] = A^{'}(1)$ *(we use primes to indicate derivatives) and* $\mathrm{Var}[A] = A^{''}(1) - (A^{'}(1))^2 + A^{'}(1)$.

Since in heavy-traffic situations, on average, nearly always c customers leave the system every $\mathrm{E}[T]$ slots, the stability condition for this model reads $\lambda < \frac{c}{\mathrm{E}[T]}$.

3. Buffer Content

3.1. Joint PGF

In this section, we deduce the joint PGF $V(z,x,y)$ of the queue content, the server content and the remaining time of the current service period, i.e. we compute

$$V(z,x,y) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k} x^{S_k} y^{R_k}\right] \ ,$$

with Q_k (S_k) the queue (server) content at slot mark k and R_k the remaining number of slots of the service cycle at slot boundary k. In [11], we have obtained the following expression for $V(z,x,y)$:

$$\begin{aligned} \left[1 - \frac{A(z)}{y}\right] V(z,x,y) &= \left[1 - \frac{A(z)}{y}\right] \sum_{n=0}^{l-1} q_0(n) z^n \\ &+ T(y) x^c z^{-c} (A(z) - 1) \sum_{n=0}^{l-1} q_0(n) z^n \\ &+ T(y) x^c z^{-c} A(z) F(z,1) - A(z) F(z,x) \\ &+ T(y) \sum_{n=l}^{c-1} e(n) \left[x^n - x^c z^{n-c}\right] \ , \qquad (1) \end{aligned}$$

whereby

- $q_0(n) \triangleq \lim_{k\to\infty} \Pr[Q_k = n, R_k = 0] \ , \qquad n = 0, \ldots, l-1 \ ,$
- $F(z,x) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k} x^{S_k} \mathbf{1}_{R_k=1}\right]$, where $\mathbf{1}_X$ is the indicator function of X,
- $e(n) \triangleq \lim_{k\to\infty} \Pr[Q_k + A_k = n, R_k \leq 1] \ , \qquad n \geq 0 \ .$

We now completely characterize $V(z,x,y)$. We did not do this in [11] as we were only interested in the PGF of the system content $U(z) = V(z,z,1)$. The current approach has the advantage however that it enables to extract, next to the system content, various other significant performance measures.

In order to determine $F(z,x)$ and $F(z,1)$, we replace y by $A(z)$ in (1). Since $V(z,x,y)$ is analytic for $|x|<1$, $|y|<1$ and $|z|<1$, this leads to:

$$\begin{aligned} A(z)F(z,x) &= T(A(z))x^c z^{-c}(A(z)-1)\sum_{n=0}^{l-1} q_0(n)z^n \\ &+ T(A(z))x^c z^{-c}A(z)F(z,1) \\ &+ T(A(z))\sum_{n=l}^{c-1} e(n)\left[x^n - x^c z^{n-c}\right] \quad . \end{aligned} \tag{2}$$

We find an expression for $F(z,1)$ by letting $x \to 1$ in (2) and rearranging terms a bit:

$$A(z)F(z,1) = T(A(z))\frac{(A(z)-1)\sum_{n=0}^{l-1} q_0(n)z^n + \sum_{n=l}^{c-1} e(n)\left[z^c - z^n\right]}{z^c - T(A(z))} \quad . \tag{3}$$

$F(z,x)$ can be found by substituting (3) in (2) and these formulas for $F(z,x)$ and $F(z,1)$ result in the following expression for $V(z,x,y)$:

$$\begin{aligned} &\left[1-\frac{A(z)}{y}\right]V(z,x,y) = \frac{1}{z^c - T(A(z))} \\ &\quad \cdot \quad \left[\left(\left[1-\frac{A(z)}{y}\right]\left[z^c - T(A(z))\right]\right.\right. \\ &\qquad \left. + x^c\left[A(z)-1\right]\left[T(y)-T(A(z))\right]\right)\sum_{n=0}^{l-1} q_0(n)z^n \\ &\qquad + \left[T(y)-T(A(z))\right] \\ &\qquad\quad \left.\sum_{n=l}^{c-1} e(n)\Big[x^c\left\{T(A(z))-z^n\right\} - x^n\left\{T(A(z))-z^c\right\}\Big]\right] \quad . \end{aligned} \tag{4}$$

$V(z,x,y)$ is now completely computed except for the c unknown constants $q_0(n)$, $0 \le n \le l-1$ and $e(n)$, $l \le n \le c-1$. To calculate these constants, we first let $x \to z$ and $y \to 1$ in equation (4). This yields

$$\begin{aligned} V(z,z,1) &= \frac{T(A(z))}{z^c - T(A(z))} \\ &\quad \cdot \quad \left\{(z^c-1)\sum_{n=0}^{l-1} q_0(n)z^n + \frac{T^*(A(z))}{A(z)}\sum_{n=l}^{c-1} \tilde{u}_n(z^c - z^n)\right\} \quad , \end{aligned}$$

where we have introduced $\tilde{u}_n \triangleq \mathrm{E}\left[T\right]e(n)$ and where $T^*(z) \triangleq \frac{z[T(z)-1]}{\mathrm{E}[T](z-1)}$ represents the PGF of the position of an arbitrarily chosen slot in its current service cycle given that the server is processing in this slot. We can prove by means of Rouché's theorem that the denominator $z^c - T(A(z))$ has c zeroes $(z_0 = 1, z_1, \ldots, z_{c-1})$ in the closed complex unit disk $\{z \in \mathbb{C} : |z| \le 1\}$ (see e.g. [2]). Because $V(z,z,1)$ is a PGF and since PGF's are normalised ($V(1,1,1)=1$) and bounded inside and on this disk, the unknowns $q_0(n)$ and $\tilde{u}_n$ (and thus also $e(n)$) can be determined by solving a set of c linear equations (except for some special cases treated below in remark 2), consisting of the normalisation condition and $c-1$ equations expressing that the numerator of $V(z,z,1)$ vanishes at z_i, $1 \le i \le c-1$:

$$(z_i^c - 1)\sum_{n=0}^{l-1} q_0(n)z_i^n + \frac{T^*(A(z_i))}{A(z_i)}\sum_{n=l}^{c-1} \tilde{u}_n(z_i^c - z_i^n) = 0 \qquad , \ 1 \le \ i \le c-1 \ , \tag{5a}$$

$$c\sum_{n=0}^{l-1} q_0(n) + \sum_{n=l}^{c-1} \tilde{u}_n(c-n) = c - \mathrm{E}\,[T]\,\lambda \ . \tag{5b}$$

Remark 2. *If $l > 1$, the unknowns can only be determined as explained above if the period of $z^c - T(A(z))$ equals 1. The period p of a series $\sum_{j=-\infty}^{\infty} b_j z^j$ is defined as the largest integer for which $b_j = 0$ whenever j is not divisible by p. It can be proved (see e.g. [2]) that if the period is equal to p, then p zeroes of $z^c - T(A(z))$ are zeroes of $z^c - 1$. So, if $p > 1$, not all unknowns $q_0(n), 0 \leq n \leq l-1$ and $\tilde{u}_n, l \leq n \leq c-1$ can be solved as above. In these cases, one should then use intuitive arguments to reduce the problem into a solvable model. For instance, if $c = 4$, $l = 2$, and customers arrive by 2, the period of $z^c - T(A(z))$ equals 2, so that an equation of the set of equations is fulfilled regardless of the values of $q_0(n)$ and $\tilde{u}_n$. However, in this case the system always contains an even number of customers irrespective of the initial system content. Hence, $q_0(1) = \tilde{u}_3 = 0$. As a result, only $q_0(0)$ and $e(2)$ still have to be calculated, which can be done with the remaining equations.*

The resulting formula (4) for $V(z, x, y)$ turns out to be very useful in the determination of several other PGF's, such as those of the system content, the queue content, et cetera. This is illustrated in the next subsections.

3.2. Important quantities

3.2.1. Joint PGF of the queue and the server content

The joint PGF of the queue and the server content is extracted from (4) by summing out the remaining service time. Hence,

$$\tilde{V}(z,x) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k} x^{S_k}\right] = V(z,x,1) \ . \tag{6}$$

This leads to

$$\begin{aligned}\tilde{V}(z,x) = {} & \frac{1}{[z^c - T(A(z))]} \\ \cdot & \left\{ [z^c - x^c + T(A(z))(x^c - 1)] \sum_{n=0}^{l-1} q_0(n) z^n \right. \\ & \left. + \frac{T^*(A(z))}{A(z)} \sum_{n=l}^{c-1} \tilde{u}_n \left[x^n z^c - x^c z^n + T(A(z))(x^c - x^n)\right] \right\} \ .\end{aligned}$$

3.2.2. PGF of the system content

The system content at the beginning of a random slot, U, is defined as the sum of the queue content and the server content at the beginning of that slot. $U(z)$ is thus equal to $V(z, z, 1)$ (or equal to $\tilde{V}(z, z)$):

$$\begin{aligned}U(z) = {} & \frac{T(A(z))}{z^c - T(A(z))} \\ \cdot & \left\{ (z^c - 1) \sum_{n=0}^{l-1} q_0(n) z^n + \frac{T^*(A(z))}{A(z)} \sum_{n=l}^{c-1} \tilde{u}_n (z^c - z^n) \right\} \ .\end{aligned} \tag{7}$$

This expression has previously been found in our paper [11].

3.2.3. PGF of the queue content

The PGF of the queue content, $Q(z)$, is found by summing out both the server content and the remaining service time. Hence,

$$Q(z) = V(z,1,1) = \tilde{V}(z,1) = \frac{1}{z^c - T(A(z))} \cdot \left\{ (z^c - 1) \sum_{n=0}^{l-1} q_0(n) z^n + \frac{T^*(A(z))}{A(z)} \sum_{n=l}^{c-1} \tilde{u}_n (z^c - z^n) \right\} \quad . \tag{8}$$

This expression has previously been found in [17] (formula (16)). We observe that $U(z) = T(A(z))Q(z)$, which is also in agreement with the relation found in [17].

3.2.4. PGF of the server content

The PGF of the server content in the steady state, $S(z)$, is equal to $V(1,z,1)$ (or $\tilde{V}(1,z)$). l'Hôpital's rule and the moment generating property of PGF's yields

$$S(z) = \frac{1}{c - \mathrm{E}[T]\lambda} \cdot \left\{ [c + \mathrm{E}[T]\lambda(z^c - 1)] \sum_{n=0}^{l-1} q_0(n) + \sum_{n=l}^{c-1} \tilde{u}_n \left[z^n c - z^c n + \mathrm{E}[T]\lambda(z^c - z^n) \right] \right\} \quad . \tag{9}$$

$S(z)$ is a polynomial of degree c, as expected (the server content is between 0 and c). From (9), we can easily extract the corresponding probabilities:

$$\Pr[S = n] = \begin{cases} \sum_{m=0}^{l-1} q_0(m) & \text{if } n = 0 \ , \\ \tilde{u}_n & \text{if } l \le n \le c-1 \ , \\ 1 - \sum_{m=0}^{l-1} q_0(m) - \sum_{m=l}^{c-1} \tilde{u}_m & \text{if } n = c \ , \\ 0 & \text{else} \ . \end{cases}$$

In the remaining paragraphs, we characterise some random variables that indicate how efficiently the capacity of the server is used.

3.2.5. PGF of the queue content when the server is dormant

The number of customers waiting to be served if the server is not processing is denoted by $\tilde{Q}$. The PGF $\tilde{Q}(z)$ is given by:

$$\tilde{Q}(z) = \lim_{k \to \infty} \mathrm{E}\left[z^{Q_k} | S_k = 0 \right] \quad .$$

Hence,

$$\tilde{Q}(z) = \frac{\tilde{V}(z,0)}{\tilde{V}(1,0)} = \frac{\sum_{n=0}^{l-1} q_0(n) z^n}{\sum_{m=0}^{l-1} q_0(m)} \quad . \tag{10}$$

The corresponding probabilities are thus equal to

$$\Pr\left[\tilde{Q} = n\right] = \begin{cases} \frac{q_0(n)}{\sum_{m=0}^{l-1} q_0(m)} & \text{if } 0 \le n \le l-1 \ , \\ 0 & \text{else} \ . \end{cases}$$

3.2.6. PGF of the queue content when the server processes at suboptimal capacity

We denote the queue content when the server is processing but not at full capacity by Q^*. Its PGF is given by

$$\begin{aligned}
Q^*(z) &= \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k} | l \le S_k < c\right] \\
&= \lim_{k\to\infty} \frac{\mathrm{E}\left[z^{Q_k}\right] - \mathrm{E}\left[z^{Q_k}\mathbf{1}_{S_k<l}\right] - \mathrm{E}\left[z^{Q_k}\mathbf{1}_{S_k=c}\right]}{\Pr\left[l \le S_k < c\right]} \\
&= \frac{Q(z) - \lim_{k\to\infty}\mathrm{E}\left[z^{Q_k}\mathbf{1}_{S_k=0}\right] - \lim_{k\to\infty}\mathrm{E}\left[z^{Q_k}\mathbf{1}_{S_k=c}\right]}{\lim_{k\to\infty}\Pr\left[l \le S_k < c\right]} \\
&= \frac{Q(z) - \tilde{V}(z,0) - \frac{1}{c!}\frac{\partial^c}{\partial x^c}\tilde{V}(z,x)\Big|_{x=0}}{\sum_{n=l}^{c-1}\tilde{u}_n} \ . \qquad (11)
\end{aligned}$$

3.2.7. PGF of the number of customers in a served batch

We denote the PGF of the number of customers in a served batch by $\tilde{S}(z)$. Because the number of customers in service does not alter during a service cycle, $\tilde{S}(z)$ is equal to $\lim_{k\to\infty}\mathrm{E}\left[z^{S_k}|R_k = 1\right]$. Hence,

$$\tilde{S}(z) = \frac{F(1,z)}{F(1,1)} \ .$$

Making use of equation (2), yields

$$\tilde{S}(z) = z^c + \frac{1}{F(1,1)}\sum_{n=l}^{c-1} e(n)\left[z^n - z^c\right] \ , \qquad (12)$$

whereby $F(1,1)$ is found by replacing z by 1 in (3) and using l'Hôpital's rule:

$$F(1,1) = \frac{\lambda\sum_{n=0}^{l-1} q_0(n) + \sum_{n=l}^{c-1} e(n)(c-n)}{c - \mathrm{E}\left[T\right]\lambda} \ .$$

From (12), we easily extract the corresponding probabilities:

$$\Pr\left[\tilde{S} = n\right] = \begin{cases} \frac{e(n)}{F(1,1)} & \text{if } l \le n \le c-1 \ , \\ 1 - \frac{\sum_{n=l}^{c-1} e(n)}{F(1,1)} & \text{if } n = c \ , \\ 0 & \text{else} \ . \end{cases}$$

Numerous other PGF's of interest can be deduced from (4), such as the PGF of the number of customers left behind at service termination, the PGF of the system content as seen by new arrivals, et cetera. Furthermore, expression (4) of $V(z,x,y)$ is also used in section 5.1, where the customer delay is analysed.

PGF's (6)-(12) now enable us to extract several performance measures:

- Since we assume that $A(z)$ and $T(z)$ are analytic at $z = 1$, it is easy to prove that the obtained PGF's (7)-(12) are analytic at $z = 1$. Hence, all order moments of (7)-(12) can be calculated. In section 3.3, we demonstrate that these moments provide an efficient tool to evaluate the impact of the service threshold l.

- From formula (6), it is possible to calculate the covariance between the server and the queue content.
- From formulas (7), (8) and (11), it is possible to calculate tail probabilities from respectively the system content, the queue content and the queue content when the server processes at suboptimal capacity, by means of the dominant pole approximation (see e.g. [8]).

3.3. Some examples

In this section, we demonstrate that moments of the above PGF's can be used to evaluate the impact of the service threshold l. We compare thresholds 1, 5 and 10 via the mean and the variance of the system content and via the mean server utilization (which is defined as $\mathrm{E}\left[\tilde{S}\right]$ - the mean number of customers in a served batch - divided by the server capacity c). We initially consider a batch-service queueing system with capacity c equal to 10, with Poisson arrivals (i.e. $A(z) = e^{\lambda(z-1)}$) and geometrically distributed service times with a mean length of 10 slots.

In Fig. 1, the mean system content is depicted versus λ ($0 < \lambda < 1(= c/\mathrm{E}[T])$ - the stability condition is fulfilled for these values). The figure exhibits that a small threshold should be adopted in light-traffic situations. Moreover, $\mathrm{E}[U] \rightarrow K_l$, with $K_1 = 0$ and $K_l > 0$ if $l > 1$. In section 4.1, we examine the light-traffic behaviour of the system content in detail and we find that $K_1 = 0$ and $K_c = (c-1)/2$ for a broad set of distributions of A and T.
Fig. 1 further shows that larger thresholds produce the smallest mean system content as λ increases. Finally, it seems that the distinct thresholds lead to an equal performance when the arrival rate becomes large ($\lambda \rightarrow 1$). This can be explained intuitively: when the mean arrival rate λ tends to 1, the system contains many customers. This implies on the one hand that the server nearly always serves at full capacity, even for small thresholds. On the other hand, the server is seldom in a dormant state, even for large thresholds. Hence, the mean system content is more or less equal for every threshold. In section 4.2, we study the heavy-traffic behaviour of the system content and we find that it is indeed independent of the service threshold.

An interesting question is the following: what is the magnitude of the performance gain (or drop) when adopting $l = c$ instead of $l = 1$? In view of this, we define the relative difference in the mean system content as $2(\mathrm{E}[U_1] - \mathrm{E}[U_c])/(\mathrm{E}[U_1] + \mathrm{E}[U_c])$, whereby U_1 (U_c) represents the system content when $l = 1$ ($l = c$). Analogously, the relative difference in the variance of the system content reads $2(\mathrm{Var}[U_1] - \mathrm{Var}[U_c])/(\mathrm{Var}[U_1] + \mathrm{Var}[U_c])$. In Fig. 2 the relative differences in the mean (part a) and the variance (part b) are plotted versus λ. We observe that if λ equals 0.75, the mean system content is approximately 10% smaller for $l = c$ than for $l = 1$. The relative difference in the variance reaches its peak earlier and the gain amounts to approximately 30%. Furthermore, the 'transition point' (i.e. where the relative difference becomes positive,

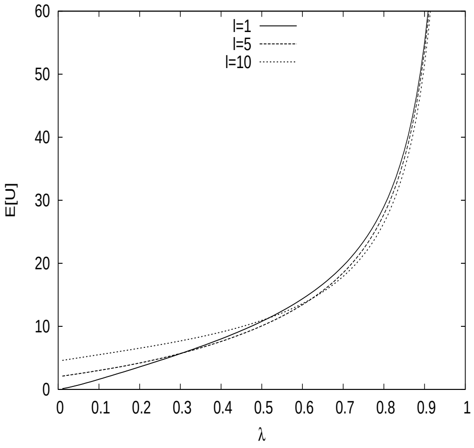
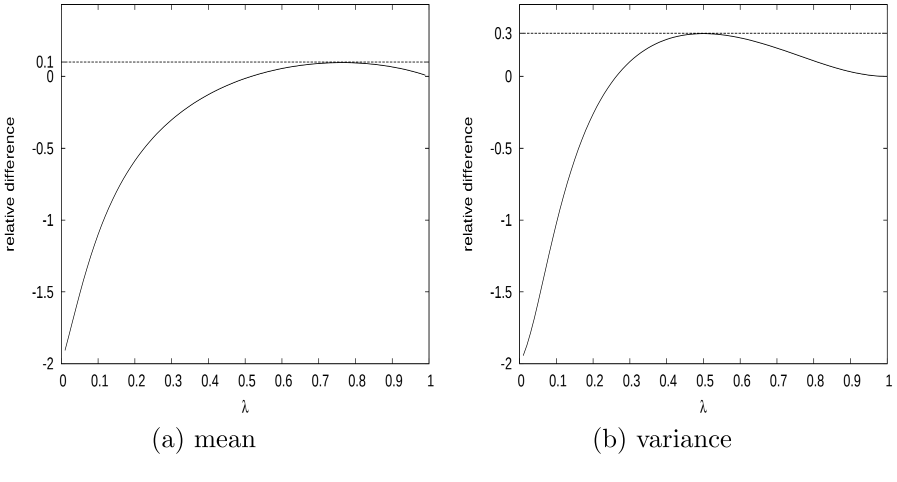


Figure 1: Mean system content versus λ

and hence where $l = c$ becomes better than $l = 1$) appears for a smaller λ. The figure also exhibits that although $l = 1$ is better in terms of $\mathrm{E}[U]$ if λ equals 0.4, the opposite holds in terms of $\mathrm{Var}[U]$. We further notice that the relative differences tend to 0 for $\lambda \to 1$.

In practice, the service of a batch can be expensive. In this case, it is of the utmost importance that the dissipation of capacity is limited. In view of this, the mean server utilization becomes a valuable performance measure. Fig. 3 shows the mean server utilization versus λ. The figure exhibits that the mean server utilization increases gradually from l/c to 1 and that the higher the threshold,



(a) mean

(b) variance

Figure 2: Relative difference versus λ

the better the server utilization. Hence, if the best utilization is desired in some situation, $l = c$ should be adopted. However, this restriction is sometimes relaxed: one might for example restrict the set of possible thresholds so that the mean server utilization is at least 50% and pick then from this set the threshold that causes the smallest mean system content.

To close this section, we examine the influence of the distribution of the service times on the mean system content. Fig. 4 shows, both for $l = 1$ (part a) and for $l = c$ (part b), the mean system content versus λ for the following distributions of the service times with a mean length of 10 slots:

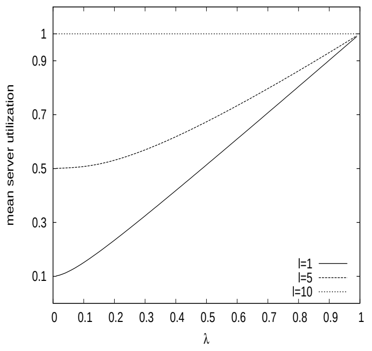


Figure 3: Mean server utilization versus λ

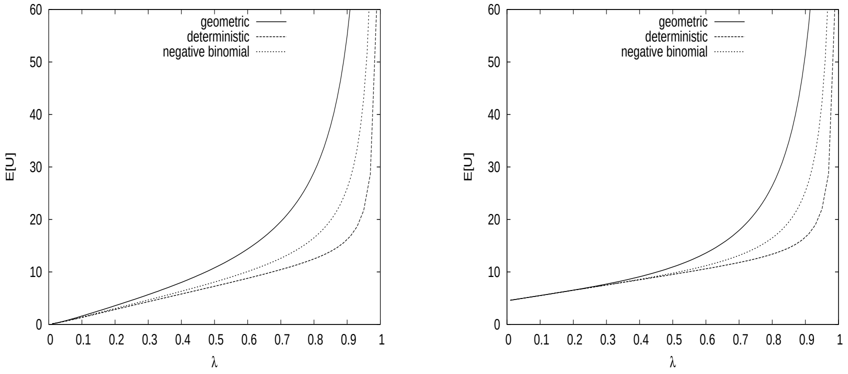
- Geometric: $T(z) = \frac{(1-0.9)z}{1-0.9z}$
- Deterministic: $T(z) = z^{10}$
- Negative binomial: $T(z) = \left(\frac{(1-0.7)z}{1-0.7z}\right)^3$

On the one hand, we observe that $\mathrm{E}[U]$ is hardly influenced by the distribution of the service times when λ is small. The light-traffic approximation of the system content obtained in the next session will indeed point out that this is in general the case. When, on the other hand, λ is larger, $T(z)$ has a huge impact on $\mathrm{E}[U]$. We observe that $\mathrm{E}[U]$ is largest in the geometric case while it is smallest in the deterministic case. These cases correspond respectively with the largest and the smallest variance of T. From the obtained heavy-traffic approximations of $\mathrm{E}[U]$ in section 4.2, we can indeed draw the conclusion that, in general, $\mathrm{E}[U]$ increases as $\mathrm{Var}[T]$ increases.

4. Approximations for the system content

4.1. Light-traffic approximation for $U(z)$

In order to calculate the performance measures from section 3.2, quite some numerical work is required, namely the computation of the c zeroes of $z^c -$



(a) $l = 1$ (b) $l = c = 10$

Figure 4: $\mathrm{E}[U]$ versus λ for several distributions of the service times with an equal mean value of 10 slots

$T(A(z))$ inside the closed complex unit disk and the solution of a set of equations. Especially the calculation of the zeroes can be a severe and even unfeasible assignment when c is large. Therefore, we establish light- and heavy-traffic approximations of the system content in respectively this section and section 4.2. These approximations require no numeric calculation of zeroes anymore. In this first subsection, we examine the light-traffic behaviour of the system content by expanding formula (7) for $U(\lambda, z)$ in a Taylor series about $\lambda = 0$ and only retaining the constant and the linear terms since the others are negligible when $\lambda \to 0$ (note that we substitute every function $f(z)$ that is dependent on λ by $f(\lambda, z)$ to underline this dependency). This leads to a formula whereby it is required to solve a set of equations but no zeroes have to be computed anymore, which eliminates this bottleneck (section 4.1.1). When $l = c$ and $l = 1$, the set of equations can be solved explicitly, so that we obtain fully-analytic formulas in these cases (sections 4.1.2 and 4.1.3 respectively). We close this section by evaluating the approximation formula through an example (section 4.1.4).

4.1.1. General l

Let us denote the Taylor series expansion of $q_0(n)$ about $\lambda = 0$ by $\sum_{k=0}^{\infty} \alpha_k(n)\lambda^k$. Along the same lines, $\sum_{k=0}^{\infty} \beta_k(n)\lambda^k$ represents the analogous expansion for $\tilde{u}_n$. Taking into account that

$$A(\lambda, z) = 1 + \lambda A^{(1)}(0, z) + \frac{\lambda^2}{2} A^{(1,1)}(0, z) + O(\lambda^3) \ ,$$

$$T(A(\lambda, z)) = 1 + \lambda \mathrm{E}[T]\, A^{(1)}(0, z) + \frac{\lambda^2}{2}\left[T^{''}(1) A^{(1)}(0, z)^2 + \mathrm{E}[T]\, A^{(1,1)}(0, z)\right] + O(\lambda^3) \ ,$$

$$\frac{T^*(A(\lambda, z))}{A(\lambda, z)} = 1 + \lambda \frac{T^{''}(1) A^{(1)}(0, z)}{2\mathrm{E}[T]} + O(\lambda^2) \ ,$$

with

$$A^{(\mathbf{1}_n, \mathbf{2}_m)}(x, y) \triangleq \frac{\partial^n}{\partial \lambda^n} \frac{\partial^m}{\partial z^m} A(\lambda, z)\Big|_{\lambda = x, z = y} \ ,$$

whereby $\mathbf{k}_n$ represents a series consisting of n consecutive k's, the series expansion of $U(\lambda, z)$ about $\lambda = 0$ reads:

$$\begin{aligned}
U(\lambda, z) =& \frac{(z^c - 1)\sum_{n=0}^{l-1} \alpha_0(n) z^n + \sum_{n=l}^{c-1} \beta_0(n)(z^c - z^n)}{z^c - 1} + \lambda \frac{1}{(z^c - 1)^2} \\
&\cdot \Bigg[\mathrm{E}\left[T\right] A^{(1)}(0, z) z^c \left\{ (z^c - 1) \sum_{n=0}^{l-1} \alpha_0(n) z^n + \sum_{n=l}^{c-1} \beta_0(n)(z^c - z^n) \right\} \\
&+ (z^c - 1) \left\{ (z^c - 1) \sum_{n=0}^{l-1} \alpha_1(n) z^n + \sum_{n=l}^{c-1} \beta_1(n)(z^c - z^n) \right\} \\
&+ (z^c - 1) \frac{T^{''}(1) A^{(1)}(0, z)}{2\mathrm{E}\left[T\right]} \sum_{n=l}^{c-1} \beta_0(n)(z^c - z^n) \Bigg] + O(\lambda^2) \ . \qquad (13)
\end{aligned}$$

Hence, in order to characterize the constant and linear term of $U(\lambda, z)$ fully, $\alpha_k(n)$ and $\beta_k(n)$, $k = 0, 1$, the constant and linear terms of the unknowns $q_0(n)$ and $\tilde{u}_n$, have to be calculated. Remember that $q_0(n)$ and $\tilde{u}_n$ can be found by solving a set of equations expressing that $U(\lambda, 1) = 1$ and that the numerator of $U(\lambda, z)$ must vanish for the zeroes $z_i(\lambda)$ of the denominator inside the complex unit disk (equations (5a) and (5b)). Therefore, we first deduce the constant $(f_{i,0})$, linear $(f_{i,1})$ and quadratic $(f_{i,2})$ terms of the Taylor series expansion of $z_i(\lambda) \triangleq \sum_{k=0}^{\infty} f_{i,k} \lambda^k$ (it will become clear later why we also need the quadratic term of the zeroes). In view of this, we expand $T(A(\lambda, z))$ and $z_i(\lambda)$ in a Taylor series about $\lambda = 0$, we apply Newton's binomium and we take into account that $A^{(2)}(0, f_{i,0}) = 0$ (since $A(0, z) = 1$), to transform $z_i(\lambda)^c = T(A(\lambda, z_i(\lambda)))$ into

$$\begin{aligned}
&f_{i,0}^c + \lambda c f_{i,0}^{c-1} f_{i,1} + \lambda^2 \left[\frac{c(c-1)}{2} f_{i,0}^{c-2} f_{i,1}^2 + c f_{i,0}^{c-1} f_{i,2} \right] + O(\lambda^3) \\
=&1 + \lambda \mathrm{E}\left[T\right] A^{(1)}\left(0, f_{i,0}\right) \\
&+ \frac{\lambda^2}{2} \left[T^{''}(1) A^{(1)}(0, f_{i,0})^2 + \mathrm{E}\left[T\right] \{ A^{(1,1)}(0, f_{i,0}) + 2A^{(1,2)}(0, f_{i,0}) f_{i,1} \} \right] + O(\lambda^3) \ .
\end{aligned}$$

We now equate the constant term at the left-hand-side with the constant term at the right-hand-side and repeat this for the linear and quadratic terms. The following set of equations is produced:

$$\begin{cases}
f_{i,0}^c = 1 \ , \\
c f_{i,0}^{c-1} f_{i,1} = \mathrm{E}\left[T\right] A^{(1)}\left(0, f_{i,0}\right) \ , \\
\frac{c(c-1)}{2} f_{i,0}^{c-2} f_{i,1}^2 + c f_{i,0}^{c-1} f_{i,2} \\
= \frac{1}{2} \left[T^{''}(1) A^{(1)}(0, f_{i,0})^2 + \mathrm{E}\left[T\right] \{ A^{(1,1)}(0, f_{i,0}) + 2A^{(1,2)}(0, f_{i,0}) f_{i,1} \} \right] \ .
\end{cases}$$

It is directly clear that the first equation has c solutions: the c complex c-th roots of one, $\varepsilon_i \triangleq e^{(\imath 2\pi i)/c}$, with $\imath$ the imaginary unit and $i = 0, \ldots, c-1$. Hence,

$$f_{i,0} = \varepsilon_i \ , \qquad 0 \leq i \leq c-1 \ . \qquad (14)$$

The corresponding $f_{i,1}$'s can be found by replacing $f_{i,0}$ by ε_i in the second equation, leading to

$$f_{i,1} = \frac{\varepsilon_i \mathrm{E}\left[T\right]}{c} A^{(1)}\left(0, \varepsilon_i\right) \ , \qquad 0 \leq i \leq c-1 \ . \qquad (15)$$

Finally, the use of (14) and (15) in the third equation produces

$$
\begin{aligned}
f_{i,2} = \frac{\varepsilon_i}{2c} \Bigg[& \left\{ T^{''}(1) - \frac{c-1}{c} \mathrm{E}\left[T\right]^2 \right\} A^{(1)}(0, \varepsilon_i)^2 \\
& + \mathrm{E}\left[T\right] A^{(1,1)}(0, \varepsilon_i) + 2 \frac{\varepsilon_i \mathrm{E}\left[T\right]^2}{c} A^{(1)}(0, \varepsilon_i) A^{(1,2)}(0, \varepsilon_i) \Bigg] \ . \qquad (16)
\end{aligned}
$$

This concludes the calculation of $f_{i,0}$, $f_{i,1}$ and $f_{i,2}$ ($0 \le i \le c-1$). As a next step, we expand equations (5a) and (5b) in a series expansion and we thereby make use of formulas (14)-(16) for $f_{i,0}$, $f_{i,1}$ and $f_{i,2}$. Owing to Newton's binomium and taking (14) into account (and thus that $f_{i,0}^c = 1$), we get

$$
\begin{cases}
\lambda c \varepsilon_i^{-1} f_{i,1} \sum_{n=0}^{l-1} [\alpha_0(n) + \lambda \alpha_1(n)][\varepsilon_i^n + \lambda n \varepsilon_i^{n-1} f_{i,1}] \\
+ \left[1 + \lambda \frac{T^{''}(1)}{2\mathrm{E}[T]} A^{(1)}(0, \varepsilon_i) \right] \sum_{n=l}^{c-1} \left[\beta_0(n) + \lambda \beta_1(n)\right] \left[1 - \varepsilon_i^n + \lambda f_{i,1} \{ c \varepsilon_i^{-1} - n \varepsilon_i^{n-1} \}\right] \\
+ O(\lambda^2) = 0 \ , \qquad 1 \le i \le c-1 \ , \\
\\
c \sum_{n=0}^{l-1} [\alpha_0(n) + \lambda \alpha_1(n)] + \sum_{n=l}^{c-1} [\beta_0(n) + \lambda \beta_1(n)](c-n) + O(\lambda^2) = c - \mathrm{E}\left[T\right] \lambda \ .
\end{cases}
\qquad (17)
$$

Equating the constant term at the left-hand-side with the constant term at the right-hand-side yields:

$$
\sum_{n=l}^{c-1} \beta_0(n)[1 - \varepsilon_i^n] = 0 \ , \qquad 1 \le i \le c-1 \ , \qquad (18a)
$$

$$
c \sum_{n=0}^{l-1} \alpha_0(n) + \sum_{n=l}^{c-1} \beta_0(n)(c-n) = c \ , \qquad (18b)
$$

(18a) thus consists of $c-1$ equations for $c-l$ unknowns $\beta_0(n)$ ($l \le n \le c-1$). The only solution is the obvious solution $\beta_0(n) = 0$ (we give an intuitive explanation of this fact in appendix B). Next, we equate the linear terms of (17) and on account of $\beta_0(n) = 0$, we obtain:

$$
c \varepsilon_i^{-1} f_{i,1} \sum_{n=0}^{l-1} \alpha_0(n) \varepsilon_i^n + \sum_{n=l}^{c-1} \beta_1(n)(1 - \varepsilon_i^n) = 0 \ , \qquad 1 \le i \le c-1 \ , \qquad (19a)
$$

$$
c \sum_{n=0}^{l-1} \alpha_1(n) + \sum_{n=l}^{c-1} \beta_1(n)(c-n) = -\mathrm{E}\left[T\right] \ . \qquad (19b)
$$

Hence, (19a) produces $c-1$ equations for c unknowns $\alpha_0(n)$ ($0 \le n \le l-1$) and $\beta_1(n)$ ($l \le n \le c-1$). Together with equation (18b) and $\beta_0(n) = 0$, these unknowns can thus be calculated. The $\alpha_1(n)$ however cannot be calculated from (18a)-(19b). Therefore, we finally also establish the quadratic term in the series expansion of the set of equations (5a) and (5b). Taking expression (14) for $f_{i,0}$ into account, we find the following $c-1$ extra equations:

$$
\begin{aligned}
& c \varepsilon_i^{-1} f_{i,1} \sum_{n=0}^{l-1} \alpha_1(n) \varepsilon_i^n + c \varepsilon_i^{-1} f_{i,1}^2 \sum_{n=0}^{l-1} \alpha_0(n) n \varepsilon_i^{n-1} \\
& + \left[\frac{c(c-1)}{2} \varepsilon_i^{-2} f_{i,1}^2 + c \varepsilon_i^{-1} f_{i,2} \right] \sum_{n=0}^{l-1} \alpha_0(n) \varepsilon_i^n + f_{i,1} \sum_{n=l}^{c-1} \beta_1(n) [c \varepsilon_i^{-1} - n \varepsilon_i^{n-1}] \\
& + \sum_{n=l}^{c-1} \beta_2(n)(1 - \varepsilon_i^n) + \frac{T^{''}(1)}{2\mathrm{E}\left[T\right]} A^{(1)}(0, \varepsilon_i) \sum_{n=l}^{c-1} \beta_1(n)(1 - \varepsilon_i^n) = 0 \ , \qquad 1 \le i \le c-1 \ . \qquad (20)
\end{aligned}
$$

The appearance of $f_{i,2}$ in these equations is the reason why we above deduced an expression for it. Formula (20) produces $c-1$ equations in the l unknowns $\alpha_1(n)$ together with $c-l$ extra unknowns $\beta_2(n)$. Hence, together with (19b), this makes it possible to obtain $\alpha_1(n)$. Summarized, the light-traffic formula for $U(\lambda,z)$ is found by replacing $\beta_0(n)$ by 0 in (13), leading to:

$$\begin{aligned} U(\lambda,z) = &\sum_{n=0}^{l-1} \alpha_0(n) z^n + \lambda \frac{1}{z^c-1} \\ &\cdot \left[\mathrm{E}\left[T\right] A^{(1)}(0,z) z^c \sum_{n=0}^{l-1} \alpha_0(n) z^n + (z^c-1) \sum_{n=0}^{l-1} \alpha_1(n) z^n + \sum_{n=l}^{c-1} \beta_1(n)(z^c - z^n) \right] \\ &+ O(\lambda^2) \ . \end{aligned} \tag{21}$$

Solving (18b)-(19a) produces $\alpha_0(n)$ $(0 \le n \le l-1)$ and $\beta_1(n)$ $(l \le n \le c-1)$ completely. Using these results in the equations (19b) and (20) finally yields $\alpha_1(n)$ $(0 \le n \le l-1)$ completely. Note that $\beta_2(n)$ $(l \le n \le c-1)$ are also determined, but we do not need these.

Remark 3. *Note that $U(z)$, $q_0(n)$, $\tilde{u}_n$ and z_i can only be expanded in a Taylor series if they are analytic at $\lambda = 0$. In appendix A, we prove that if $A(z)$ is analytic at $\lambda = 0$ for all z in the closed complex unit disk (this assumption is not stringent, since this is usually the case), then these functions are also analytic at $\lambda = 0$.*

Remark 4. *The light-traffic formulas are valid under the assumption that $z^c - T(A(z))$ is aperiodic. Indeed, the approximation is based on the series expansion of equations (5), which only make sense in case of aperiodicity of $z^c - T(A(z))$ (see Remark 2).*

4.1.2. Special case: $l = c$

In this case, formula (21) for $U(\lambda,z)$ transforms into

$$\begin{aligned} U(\lambda,z) = &\sum_{n=0}^{c-1} \alpha_0(n) z^n + \lambda \frac{1}{z^c-1} \\ &\cdot \left[\mathrm{E}\left[T\right] A^{(1)}(0,z) z^c \sum_{n=0}^{c-1} \alpha_0(n) z^n + (z^c-1) \sum_{n=0}^{c-1} \alpha_1(n) z^n \right] + O(\lambda^2) \ , \end{aligned} \tag{22}$$

and equations (19a) and (18b) into

$$\begin{cases} \sum_{n=0}^{c-1} \alpha_0(n) \varepsilon_i^n = 0 \ , & 1 \le i \le c-1 \ , \\ \sum_{n=0}^{c-1} \alpha_0(n) = 1 \ . \end{cases} \tag{23}$$

These c equations in c unknowns have a unique solution

$$\alpha_0(n) = \frac{1}{c} \ , \qquad 0 \le n \le c-1 \ , \tag{24}$$

because

$$\sum_{n=0}^{c-1} \varepsilon_i^n = \frac{\varepsilon_i^c - 1}{\varepsilon_i - 1} = 0 \ , \qquad 1 \le i \le c-1 \ .$$

Next, we calculate $\sum_{n=0}^{c-1} \alpha_1(n) z^n$. Note that this is in fact a polynomial, say $p(z)$, of degree $c-1$. On account of $\sum_{n=0}^{c-1} \alpha_0(n)\varepsilon_i^n = 0$ (see equation (23)), expression (15) for $f_{i,1}$, and

$$\sum_{n=0}^{c-1} n\varepsilon_i^{n-1} = \frac{c}{\varepsilon_i(\varepsilon_i - 1)} \quad , \qquad 1 \le i \le c-1 \ ,$$

equations (20) and (19b) become

$$\begin{aligned} p(\varepsilon_i) &= \frac{\mathrm{E}\,[T]}{c(1-\varepsilon_i)} A^{(1)}(0,\varepsilon_i) \ , \qquad 1 \le i \le c-1 \ , \\ p(1) &= -\frac{\mathrm{E}\,[T]}{c} \ , \end{aligned}$$

We thus obtain a set of c data points of a polynomial of degree $c-1$. By virtue of Lagrange's interpolation formula (see e.g. [4]), we find

$$\begin{aligned} p(z) =& -\frac{\mathrm{E}\,[T]}{c} \prod_{k=1}^{c-1} \frac{z-\varepsilon_k}{1-\varepsilon_k} + \sum_{i=1}^{c-1} \frac{\mathrm{E}\,[T]}{c(1-\varepsilon_i)} A^{(1)}(0,\varepsilon_i) \frac{z-1}{\varepsilon_i - 1} \prod_{j=1, j\neq i}^{c-1} \frac{z-\varepsilon_j}{\varepsilon_i - \varepsilon_j} \\ =& -\frac{\mathrm{E}\,[T]\,(z^c-1)}{c^2(z-1)} + (z^c - 1)\frac{\mathrm{E}\,[T]}{c^2} \sum_{i=1}^{c-1} \frac{A^{(1)}(0,\varepsilon_i)\varepsilon_i}{(z-\varepsilon_i)(1-\varepsilon_i)} \ . \qquad (25) \end{aligned}$$

The combination of (22), (24) and (25) yields the final expression for $U(\lambda, z)$:

$$\begin{aligned} U(\lambda,z) =& \frac{z^c-1}{c(z-1)} \\ &+ \lambda\frac{\mathrm{E}\,[T]}{c^2} \left\{ \frac{1}{z-1}\left[1 - z^c + cz^c \frac{\partial}{\partial x} A(x,z)\bigg|_{x=0}\right] \right. \\ &\left. + \ (z^c-1)\sum_{i=1}^{c-1} \frac{\partial}{\partial x} A(x,y)\bigg|_{x=0,y=\varepsilon_i} \frac{\varepsilon_i}{(1-\varepsilon_i)(z-\varepsilon_i)} \right\} + O(\lambda^2) \ . \qquad (26) \end{aligned}$$

4.1.3. Special case: $l = 1$

In this case, formula (21) for $U(\lambda, z)$ transforms into

$$\begin{aligned} U(\lambda,z) =& \alpha_0(0) + \lambda \frac{1}{z^c-1} \\ & \cdot \left[\mathrm{E}\,[T]\, A^{(1)}(0,z) z^c \alpha_0(0) + (z^c-1)\alpha_1(0) + (z-1)\sum_{n=1}^{c-1} \beta_1(n) \frac{z^c - z^n}{z-1} \right] \\ & + O(\lambda^2) \ . \qquad (27) \end{aligned}$$

Below, we subsequently establish expressions for $\alpha_0(0)$, $\alpha_1(0)$ and $\sum_{n=1}^{c-1}\beta_1(n)(z^c - z^n)/(z-1)$ and then substitute them in (27) to obtain the light-traffic formula for $U(\lambda, z)$. First, equation (18b) implies that

$$\alpha_0(0) = 1 \ . \qquad (28)$$

Secondly, $\sum_{n=1}^{c-1}\beta_1(n)(z^c - z^n)/(z-1)$ is a polynomial, say $\tilde{p}(z)$, of degree $c-1$, of which a set of $c-1$ data points is extracted from (19a):

$$\tilde{p}(\varepsilon_i) = \frac{\mathrm{E}\,[T]\, A^{(1)}(0,\varepsilon_i)}{1-\varepsilon_i} \quad , \qquad 1 \le i \le c-1 \ . \qquad (29)$$

Application of Lagrange's interpolation formula on (29) and $\tilde{p}(0) = 0$ produces:

$$\tilde{p}(z) = \sum_{i=1}^{c-1} \frac{\mathrm{E}[T]\, A^{(1)}(0, \varepsilon_i)}{1 - \varepsilon_i} \frac{z}{\varepsilon_i} \prod_{j=1, j \neq i}^{c-1} \frac{z - \varepsilon_j}{\varepsilon_i - \varepsilon_j} \ . \tag{30}$$

Since $\varepsilon_0 = 1, \varepsilon_1, \ldots, \varepsilon_{c-1}$ are the c complex c-th roots of one, we have that $(z^c - 1) = (z - 1) \prod_{j=1}^{c-1} (z - \varepsilon_j)$ and thus that

$$\prod_{j=1, j \neq i}^{c-1} (z - \varepsilon_j) = \frac{z^c - 1}{(z - \varepsilon_i)(z - 1)} \ ,$$

and consequently

$$\prod_{j=1, j \neq i}^{c-1} (\varepsilon_i - \varepsilon_j) = \frac{c\varepsilon_i^{c-1}}{\varepsilon_i - 1} = \frac{c}{\varepsilon_i(\varepsilon_i - 1)} \ .$$

As a result, (30) transforms into

$$\tilde{p}(z) = \frac{z^c - 1}{c(z - 1)} z \mathrm{E}[T] \sum_{i=1}^{c-1} \frac{A^{(1)}(0, \varepsilon_i)}{\varepsilon_i - z} \ . \tag{31}$$

Finally, $\alpha_1(0)$ is obtained through the normalisation condition $U(\lambda, 1) = 1$. The linear term in (27) has to be equal to zero at $z = 1$. This condition, after application of l'Hôpital's rule ($A^{(1)}(0, 1) = 0$, since $A(\lambda, 1) = 1$), yields

$$\alpha_1(0) = -\frac{\mathrm{E}[T]}{c} - \frac{\mathrm{E}[T]}{c} \sum_{i=1}^{c-1} \frac{A^{(1)}(0, \varepsilon_i)}{\varepsilon_i - 1} \ , \tag{32}$$

where we have used that $A^{(1,2)}(0, 1) = 1$ (because owing to the moment generating property of PGFs $A^{(2)}(\lambda, 1) = \mathrm{E}[A] = \lambda$) and $\sum_{n=0}^{c-1} \beta_1(n)(c - n) = \tilde{p}(1)$. Finally, making use of (28), (31) and (32) in (27) results in the following light-traffic formula for $U(\lambda, z)$:

$$U(\lambda, z) = 1 + \lambda \mathrm{E}[T] \left\{ \frac{z^c}{z^c - 1} A^{(1)}(0, z) - \frac{1}{c} + \frac{1 - z}{c} \sum_{i=1}^{c-1} A^{(1)}(0, \varepsilon_i) \frac{\varepsilon_i}{(\varepsilon_i - z)(1 - \varepsilon_i)} \right\} \ . \tag{33}$$

Remark 5. *Formulas (26) and (33) are closed-form expressions: no numerical work is required.*

Remark 6. *Formula (26) expresses that in the case $l = c$, the system content tends to a uniform distribution for $\lambda \to 0$, while (33) shows that when $l = 1$, $\Pr[U = 0]$ tends to 1 (for an intuitive explanation of these facts, we refer to appendix B). This is in agreement with the findings in section 3.3 that the constant term of the mean system content in case of light-traffic equals 0 when $l = 1$, while it equals $(c - 1)/2$ if $l = c$.*

Remark 7. *Equations (18b) and (19a) together with formula (15) for $f_{i,1}$ reveal that the constant term in the light-traffic approximation is only influenced by the service times through their mean value. When $l = 1$ or $l = c$, this also holds for the linear term (see formulas (26) and (33)). The variance of the service times thus only has a small influence on $\mathrm{E}[U]$ in case of light traffic. This is in agreement with the findings in section 3.3.*

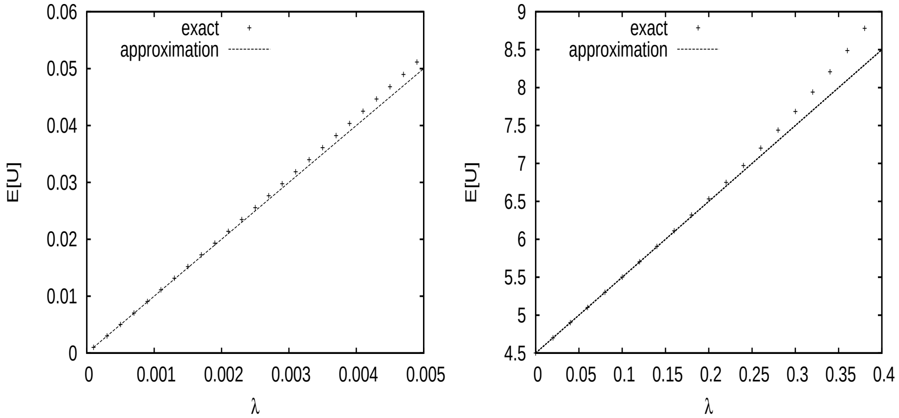
Remark 8. *Light-traffic approximations of the moments of the system content can be obtained by taking derivatives of (21) at $z = 1$. Indeed, since $U(\lambda, z)$ is analytic at ($\lambda = 0, z = 1$), the order of taking derivatives can be changed (first to λ and then to z or vice versa).*

4.1.4. Evaluation of the approximation

In this section, we demonstrate the accurateness of light-traffic approximation formula (21) through a small example. Fig. 5 shows the mean system content and its approximation for $l = 1$ (part a) as well as for $l = 10$ (part b), in case of Poisson arrivals, geometric service times and a server capacity c equal to 10. We perceive that the approximation is indeed accurate for small values of λ. Fig. 5 also leads one to suspect that the larger l, the longer the range of values of λ where the approximation is accurate. In order to verify this, we define the relative error of the approximation as

$$\frac{2\left(\mathrm{E}\left[U\right] - \mathrm{E}\left[U\right]_a\right)}{\mathrm{E}\left[U\right] + \mathrm{E}\left[U\right]_a} ,$$

with $\mathrm{E}\left[U\right]_a$ the approximated value of $\mathrm{E}\left[U\right]$, found by taking the first derivative at $z = 1$ of (21). In Fig. 6, this relative error is depicted versus λ for each value of l. Fig. 6 indeed leads to the conclusion that the larger the value of l, the more accurate the light-traffic formula (21).



(a) $l = 1$ (b) $l = 10$

Figure 5: $\mathrm{E}\left[U\right]$ and its approximation; Poisson arrivals, geometric service times with $\mathrm{E}\left[T\right] = 10$, $c = 10$

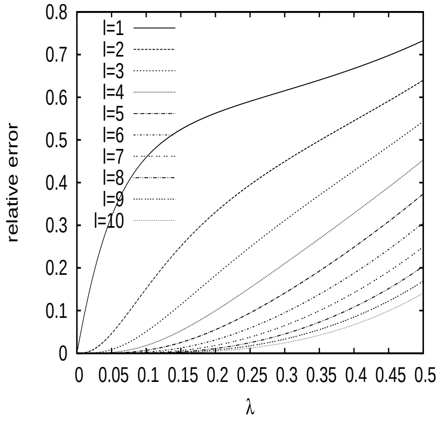


Figure 6: Relative error of the light-traffic approximation of $\mathrm{E}\left[U\right]$, for all values of l; Poisson arrivals, geometric service times with $\mathrm{E}\left[T\right] = 10$, $c = 10$

4.2. Heavy-traffic approximation

We here deduce heavy-traffic approximations for the most important performance measures ($\mathrm{E}[U]$, $\mathrm{Var}[U]$ and $\Pr[U=n]$). The approach essentialy boils down to demonstrate that $q_0(n)$ and $\tilde{u}_n$ tend to zero as λ goes to $c/\mathrm{E}[T]$ and then exploiting these findings to deduce heavy-traffic formulas. So, let us start by proving that $q_0(n)$ and $\tilde{u}_n$ tend to zero. The rate-in-rate-out principle leads to the following relation:

$$\Pr[R=0] = 1 - \frac{\lambda \mathrm{E}[T]}{L(\lambda)} \ , \tag{34}$$

whereby $L(\lambda)$ represents the mean number of customers in a served batch. Indeed, the mean number of customers leaving the system at the end of an arbitrary slot equals the product of the fraction of slots where a service is completed - this equals the probability that the server is not dormant ($1-\Pr[R=0]$) divided by the mean service length $\mathrm{E}[T]$ - and the mean number of customers in a served batch ($L(\lambda)$). Relation (34) together with $l \leq L(\lambda) \leq c$ and $\Pr[R=0] \geq 0$, yields

$$\lim_{\lambda \uparrow \frac{c}{\mathrm{E}[T]}} L(\lambda) = c \ , \tag{35}$$

and

$$\lim_{\lambda \uparrow \frac{c}{\mathrm{E}[T]}} \Pr[R=0] = 0 \ , \tag{36}$$

which is in accordance with Fig. 3, where the mean server utilization (which equals $L(\lambda)/c$) tends to one. Combining (36) with the definitions of $q_0(n)$ and $\tilde{u}_n$ implies that

$$\lim_{\lambda \uparrow \frac{c}{\mathrm{E}[T]}} q_0(n) = 0 \ , \tag{37}$$

and that

$$\lim_{\lambda \uparrow \frac{c}{\mathrm{E}[T]}} \tilde{u}_n = \mathrm{E}[T] \lim_{\lambda \uparrow \frac{c}{\mathrm{E}[T]}} \lim_{k \to \infty} \Pr[Q_k + A_k = n, R_k = 1] \ .$$

Further, if $Q_k + A_k = n, R_k = 1$, the server starts the service of n customers at the beginning of slot $k+1$. However, (35) states that the mean number of served customers tends to c, so that

$$\lim_{\lambda \uparrow \frac{c}{\mathrm{E}[T]}} \tilde{u}_n = 0 \ , \qquad n < c \ . \tag{38}$$

Next, we take advantage of (37) and (38) to formulate an approximation for $\mathrm{E}[U]$, $\mathrm{Var}[U]$ and $\Pr[U=n]$. To this end, we introduce $N(z)$ and $D(z)$ as respectively the numerator and denominator of $U(z)$ (formula (7)). Observe that $N(1) = D(1) = 0$ and that, owing to the normalisation condition, $N^{'}(1) = D^{'}(1)$. As a result, we have

$$\mathrm{E}[U] = U^{'}(1) = \frac{N^{''}(1) - D^{''}(1)}{2D^{'}(1)} \ ,$$

and

$$U^{''}(1) = \frac{2N^{'''}(1)D^{'}(1) - 2D^{'}(1)D^{'''}(1) - 3D^{''}(1)N^{''}(1) + 3D^{''}(1)^2}{6D^{'}(1)^2} \ .$$

In addition, when λ goes to $c/\mathrm{E}[T]$, $D^{'}(1) = c - \mathrm{E}[T]\lambda$ tends to zero, as well as the numerator and all its derivatives become zero at $z=1$ (since $q_0(n) \to 0$ and

$\tilde{u}_n \to 0$). On account of these findings, $U^{'}(1)$ and $U^{''}(1)$ go to infinity according to the following expressions:

$$U^{'}(1) \sim \frac{-D^{''}(1)}{2D^{'}(1)} \ , \tag{39}$$

$$U^{''}(1) \sim \frac{1}{2}\left(\frac{D^{''}(1)}{D^{'}(1)}\right)^2 \ . \tag{40}$$

Taking the appropriate derivatives of $D(z)$ and relying on (39) and (40) yields the intended heavy-traffic approximations for $\mathrm{E}\left[U\right]$ and $\mathrm{Var}\left[U\right]$:

$$\mathrm{E}\left[U\right] \sim \frac{T^{''}(1)\lambda^2 + \mathrm{E}\left[T\right]A^{''}(1) - c(c-1)}{2(c - \mathrm{E}\left[T\right]\lambda)} \ , \tag{41}$$

$$\mathrm{Var}\left[U\right] \sim \frac{1}{2}\left(\frac{T^{''}(1)\lambda^2 + \mathrm{E}\left[T\right]A^{''}(1) - c(c-1)}{c - \mathrm{E}\left[T\right]\lambda}\right)^2 \ . \tag{42}$$

Finally, on account of (37) and (38), it follows that $U(z) \to 0$ for $|z| < 1$, which, in turn, implies that $\mathrm{Pr}\left[U = n\right] \to 0$ for finite n.

Before finishing this section, we evaluate formula (41) through a small example with Poisson arrivals, geometrically distributed service times with mean value 10 and a server capacity equal to 10. Fig. 7 shows the relative differences between the exact values and the approximations, for $l = 1$, $l = 5$ and $l = c$. We observe that the approximation is accurate for a large arrival intensity and that it fits better when l is larger. This is logical, since the approximation exploits the fact that the server nearly always processes at full capacity in case of heavy traffic.

Remark 9. *Note that (41) and (42) are independent of l. The heavy-traffic behaviour is thus independent of the threshold l.*

Remark 10. *The appearance of $T^{''}(1)$ in the numerator of (41) and (42) shows that the system content increases if the variance of the service times increases, which is a typical result in queueing theory.*

Remark 11. *We could in fact have adhered to the notations from section 3.2.7, so that $\mathrm{E}\left[\tilde{S}\right]$ would denote the mean number of customers in a served batch. However, we have introduced $L(\lambda)$ to emphasize the dependency of the mean number of customers in a served batch on λ.*

5. Customer delay

The time (counted in number of slots) that a randomly tagged customer remains in the queue (W) consists of two components; the first (W_1) is the time required to serve batches of ‘older’ customers. The second part (W_2) is the time needed, starting from the end of the first waiting time, to fill the batch

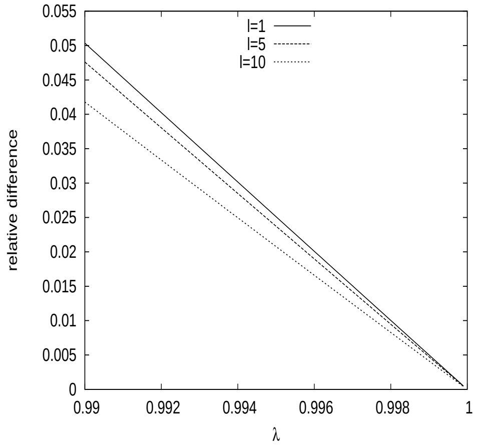


Figure 7: Relative difference between $\mathrm{E}\left[U\right]$ and heavy-traffic approximation formula (41), for $l = 1$, $l = 5$ and $l = c$

containing the tagged customer with at least l customers. Dependence exists between W_1 and W_2. Indeed, the first waiting time influences the number of customers left at the end of this waiting time (denoted by P), while the second waiting time depends on P. In Fig. 8, W_1, W_2 and P are indicated together with several other notations that are introduced further. We first compute the joint PGF of W_1 and W_2 (section 5.1), from which several characteristics related to the customer delay will be derived (section 5.2).

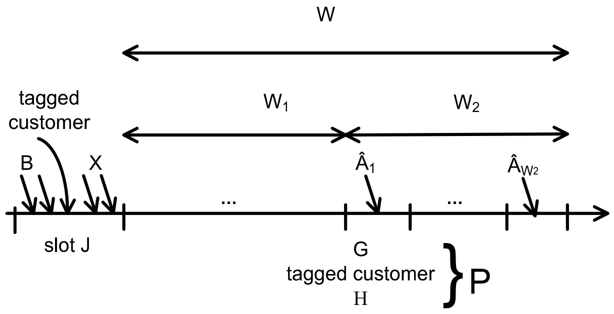


Figure 8: Illustration of the notations

5.1. Joint PGF of W_1 and W_2

We denote the joint PGF of W_1 and W_2 by $\tilde{W}(z, x)$, i.e.

$$\tilde{W}(z,x) \triangleq \mathrm{E}\left[z^{W_1} x^{W_2}\right] \quad ,$$

and we condition on P:

$$\begin{aligned}\tilde{W}(z,x) &= \sum_{p=1}^{\infty} \Pr\left[P=p\right] \mathrm{E}\left[z^{W_1}x^{W_2}|P=p\right] \\ &= \sum_{p=1}^{\infty} \Pr\left[P=p\right] \mathrm{E}\left[z^{W_1}|P=p\right] \mathrm{E}\left[x^{W_2}|P=p\right] \ . \qquad (43)\end{aligned}$$

In the last step of (43) we exploit that W_1 and W_2 are independent if P is given (W_1 only influences W_2 through the value of P). Next, we compute $\mathrm{E}\left[x^{W_2}|P=p\right]$. Therefore, we make use of the following relation between W_2 and P:

$$\Pr[W_2 > m|P=p] = \Pr[p + \hat{A}_1 + \cdots + \hat{A}_m < l] \ , \qquad m \geq 0, \qquad (44)$$

with $\hat{A}_j$ the number of arrivals during the j^{th} slot after the end of the first waiting time. Multiplication of both sides of (44) by x^m and summing over all m yields

$$\begin{aligned}\frac{\mathrm{E}\left[x^{W_2}|P=p\right]-1}{x-1} &= \sum_{m=0}^{\infty} x^m \Pr\left[p + \hat{A}_1 + \cdots + \hat{A}_m < l\right] \\ &= \sum_{m=0}^{\infty} x^m \sum_{n=0}^{l-1} \Pr\left[p + \hat{A}_1 + \cdots + \hat{A}_m = n\right] \\ &= \sum_{m=0}^{\infty} x^m \sum_{n=0}^{l-1} \frac{1}{n!} \frac{\partial^n}{\partial y^n} y^p A(y)^m \bigg|_{y=0} \\ &= \sum_{n=0}^{l-1} \frac{1}{n!} \frac{\partial^n}{\partial y^n} \frac{y^p}{1-xA(y)} \bigg|_{y=0} \ ,\end{aligned}$$

whereby step 3 makes use of the probability generating property of PGF's and the IID character of the arrival process. The last equation requires that $|xA(y)| < 1$ in the neighbourhood of $y = 0$. We thus have that:

$$\mathrm{E}\left[x^{W_2}|P=p\right] = 1 + (x-1) \sum_{n=0}^{l-1} \frac{1}{n!} \frac{\partial^n}{\partial y^n} \frac{y^p}{1-xA(y)} \bigg|_{y=0} \ . \qquad (45)$$

Note that the second term of (45) vanishes if $p \geq l$. Indeed, when $p \geq l$, the second waiting time is equal to zero. Substituting (45) into (43) yields:

$$\tilde{W}(z,x) = P(z,1) + (x-1) \sum_{n=0}^{l-1} \frac{1}{n!} \frac{\partial^n}{\partial y^n} \frac{P(z,y)}{1-xA(y)} \bigg|_{y=0} \ , \qquad (46)$$

with

$$P(z,y) \triangleq \mathrm{E}\left[z^{W_1}y^P\right] \ .$$

In order to compute $P(z,y)$, we first calculate the joint PGF $W(z,x,y)$ of W_1, the number of customers ahead (G) and the number of customers behind (H) the tagged customer at the end of the first waiting time (see Fig. 8 for an illustration of these variables), i.e.

$$W(z,x,y) \triangleq \mathrm{E}\left[z^{W_1}x^G y^H\right] \ .$$

Since P is equal to $G + H + 1$, $P(z,y)$ is then equal to $yW(z,y,y)$. Call the slot wherein the tagged customer arrives slot J and denote the queue content at the beginning of this slot by Q_J. Furthermore, B (X resp.) is the number of customer arrivals during slot J and before (after resp.) the tagged customer. We consider two situations depending on whether the remaining service time at the beginning of slot J, denoted by R_J, equals 0 or not:

- $R_J = 0$. In this case the server is not processing during slot J. As a consequence, the server can start a new service at slot $J+1$ if there are enough customers. Also, the number of 'older' customers equals $Q_J + B$. The first waiting time W_1 is thus equal to $\left\lfloor \frac{Q_J+B}{c} \right\rfloor$ service periods, with $\lfloor . \rfloor$ the floor function, i.e. $\lfloor x \rfloor = \max\{n \in \mathbb{Z} | n \leq x\}$. $(Q_J + B) \bmod c$ 'older' customers are served in the same batch as the tagged customer, with 'mod' the modulo operator. Hence, $G = (Q_J + B) \bmod c$. The number of customers behind the tagged customer at the end of the first waiting time is the sum of the number of customers that arrive during slot J but after the tagged one and those that arrive during the first waiting time. Hence, $H = X + \sum_{i=1}^{W_1} A_{J+i}$.
- $R_J \geq 1$. In this case, the server first continues $R_J - 1$ slots with the current service period. After that, $Q_J + B$ customers are ahead of the tagged one and another $\left\lfloor \frac{Q_J+B}{c} \right\rfloor$ service periods are part of W_1. Hence, $W_1 = \left\lfloor \frac{Q_J+B}{c} \right\rfloor$ service periods $+R_J - 1$. Analogously as in the first case, $G = (Q_J + B)\text{mod } c$ and $H = X + \sum_{i=1}^{W_1} A_{J+i}$.

We split the computation of the joint PGF $W(z, x, y)$ of W_1, G and H in two parts corresponding to these two situations:

$$W(z,x,y) \quad = \quad \mathrm{E}\left[z^{W_1} x^G y^H \mathbf{1}_{R_J=0}\right] + \mathrm{E}\left[z^{W_1} x^G y^H \mathbf{1}_{R_J\geq 1}\right] \quad . \tag{47}$$

For the first component we have:

$$\mathrm{E}\left[z^{W_1} x^G y^H \mathbf{1}_{R_J=0}\right] = \sum_{n=0}^{\infty} \sum_{m=0}^{c-1} \sum_{k=0}^{\infty} d(nc+m,k,0) T(zA(y))^n x^m y^k \quad , \tag{48}$$

with

$$d(n,m,k) \triangleq \Pr\left[Q_J + B = n, X = m, R_J = k\right] \quad .$$

Due to the IID number of per-slot customer arrivals, we can write the corresponding PGF $D(z, x, y)$ as the following product:

$$D(z,x,y) = \mathrm{E}\left[z^B x^X\right] V(z,1,y) \quad , \tag{49}$$

with $V(z, x, y)$ as defined in section 3.1. Taking into account that an arbitrary customer is more likely to arrive in a slot with more customer arrivals (see e.g. [7]), $\mathrm{E}\left[z^B x^X\right]$ is equal to

$$\mathrm{E}\left[z^B x^X\right] = \frac{A(z) - A(x)}{\lambda(z-x)} \quad . \tag{50}$$

In order to relate $\mathrm{E}\left[z^{W_1} x^G y^H \mathbf{1}_{R_J=0}\right]$ with $D(z, x, y)$, we first introduce some notations. The function $u(z, y)$ is defined as the 'principal c^{th} root' of $T(zA(y))$, i.e.

$$u(z,y) \triangleq |T(zA(y))|^{1/c} e^{\imath \operatorname{Arg}\left(T(zA(y))\right)/c} \quad , \tag{51}$$

whereby $|z|$ is the absolute value of z, $\mathrm{Arg}(z)$ is the principal value of the argument of z and $\imath$ is the imaginary unit. Next, $\delta\langle l = j\rangle$ is the Kronecker-Delta function (i.e. $\delta\langle l = j\rangle = 1$ if $l = j$ and $\delta\langle l = j\rangle = 0$ if $l \neq j$). We now obtain subsequently for $\mathrm{E}\left[z^{W_1}x^G y^H \mathbf{1}_{R_J=0}\right]$, starting from (48):

$$\begin{aligned}
&\mathrm{E}\left[z^{W_1}x^G y^H \mathbf{1}_{R_J=0}\right] \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1} d(nc+m,k,0)u(z,y)^{nc+m-j}x^j y^k \delta\langle m = j\rangle \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1} d(nc+m,k,0)u(z,y)^{nc+m-j}x^j y^k \sum_{i=0}^{c-1}\frac{1}{c}\varepsilon_i^{nc+m-j} \\
&= \frac{1}{c}\sum_{i=0}^{c-1} D(u(z,y)\varepsilon_i,y,0)\sum_{j=0}^{c-1} u(z,y)^{-j}x^j\varepsilon_i^{-j} \\
&= \frac{u(z,y)^c - x^c}{cu(z,y)^c}\sum_{i=0}^{c-1} D(u(z,y)\varepsilon_i,y,0)\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i - x} \quad , \qquad (52)
\end{aligned}$$

whereby we used the standard property $\delta\langle m = j\rangle = \sum_{i=0}^{c-1}\frac{1}{c}\varepsilon_i^{nc+m-j}$ in step 2 and $\varepsilon_i = e^{\imath 2\pi i/c}$ as in the previous section. We continue with the second part of (47). In a similar way as formulas (48) and (52), we find

$$\begin{aligned}
&\mathrm{E}\left[z^{W_1}x^G y^H \mathbf{1}_{R_J\geq 1}\right] \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{r=1}^{\infty} d(nc+m,k,r)T(zA(y))^n(zA(y))^{r-1}x^m y^k \quad ,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[z^{W_1}x^G y^H \mathbf{1}_{R_J\geq 1}\right] &= \frac{1}{czA(y)}\frac{u(z,y)^c - x^c}{u(z,y)^c} \\
&\cdot \sum_{i=0}^{c-1}\left[D(u(z,y)\varepsilon_i,y,zA(y)) - D(u(z,y)\varepsilon_i,y,0)\right]\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i - x} \quad . \qquad (53)
\end{aligned}$$

Substitution of (52) and (53) in (47) produces:

$$\begin{aligned}
W(z,x,y) =& \frac{u(z,y)^c - x^c}{cu(z,y)^c zA(y)} \\
&\cdot\left[[zA(y)-1]\sum_{i=0}^{c-1} D(u(z,y)\varepsilon_i,y,0)\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i - x}\right. \\
&\left. + \sum_{i=0}^{c-1} D(u(z,y)\varepsilon_i,y,zA(y))\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i - x}\right] \quad .
\end{aligned}$$

Making use of formulas (49), (50), (51), (4) and $V(z,1,0) = \sum_{n=0}^{l-1} q_0(n)z^n$ yields:

$$\begin{aligned}W(z,x,y) = \frac{T(zA(y)) - x^c}{c\lambda T(zA(y))} \sum_{i=0}^{c-1} \frac{A(u(z,y)\varepsilon_i) - A(y)}{u(z,y)\varepsilon_i - y} \frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i - x} \\ \cdot \left\{ [zA(y) - 1] \sum_{n=0}^{l-1} q_0(n)(u(z,y)\varepsilon_i)^n + \sum_{n=l}^{c-1} e(n) \left[T(zA(y)) - (u(z,y)\varepsilon_i)^n\right] \right\} \\ \cdot \frac{1}{zA(y) - A(u(z,y)\varepsilon_i)} \ .\end{aligned}$$

Hence,

$$\begin{aligned} P(z,y) &= yW(z,y,y) \\ &= y\frac{T(zA(y)) - y^c}{c\lambda T(zA(y))} \sum_{i=0}^{c-1} \frac{A(u(z,y)\varepsilon_i) - A(y)}{[u(z,y)\varepsilon_i - y]^2} \frac{u(z,y)\varepsilon_i}{zA(y) - A(u(z,y)\varepsilon_i)} \\ &\quad \cdot \left\{ [zA(y) - 1] \sum_{n=0}^{l-1} q_0(n)(u(z,y)\varepsilon_i)^n \right. \\ &\quad \left. + \sum_{n=l}^{c-1} e(n) \left[T(zA(y)) - (u(z,y)\varepsilon_i)^n\right] \right\} \ . \end{aligned} \tag{54}$$

Substitution of (51) and (54) in (46) produces the joint PGF $\tilde{W}(z,x)$. As was the case in section 3, the obtained joint PGF enables us to extract several characteristics. In the next subsections, we derive the PGF of the total customer delay and the marginal PGF's of the first and the second delay.

Remark 12. *The analysis of W_2 given $P = p$ is substantially facilitated in the event of single arrivals, since this then equals the sum of max($l - p$,0) shifted geometrically distributed service times. One can verify that if $A(z)$ is substituted by $1 - \lambda + \lambda z$, (45) reduces to this sum.*

5.2. Important quantities

5.2.1. PGF of the customer delay

Since the customer delay W is the sum of the first and the second waiting time, $W(z)$ is found by substituting x by z in (46):

$$W(z) = \tilde{W}(z,z) = P(z,1) + (z-1) \sum_{n=0}^{l-1} \frac{1}{n!} \frac{\partial^n}{\partial y^n} \left. \frac{P(z,y)}{1 - zA(y)} \right|_{y=0} \ . \tag{55}$$

5.2.2. PGF of the first waiting time

The PGF of the first waiting time is found by summing out the second waiting time in $\tilde{W}(z,x)$. Hence, substituting x by 1 in (46) gives:

$$W_1(z) \triangleq \mathrm{E}\left[z^{W_1}\right] = \tilde{W}(z,1) = P(z,1) \ . \tag{56}$$

5.2.3. PGF of the second waiting time

Substituting z by 1 and x by z in (46) produces:

$$W_2(z) \triangleq \mathrm{E}\left[z^{W_2}\right] = \tilde{W}(1,z) = 1 + (z-1)\sum_{n=0}^{l-1} \frac{1}{n!}\frac{\partial^n}{\partial y^n}\left.\frac{P(1,y)}{1-zA(y)}\right|_{y=0} . \tag{57}$$

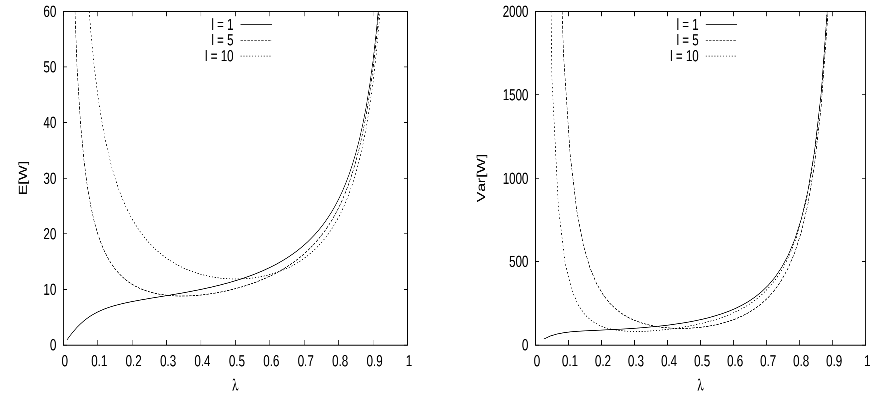
Expressions (55) - (57) enable us to calculate moments of the first, second and total customer delay (note that the mean delay can also be obtained by applying Little's law to the mean queue content - see e.g. [15]). These moments serve as performance measures. Furthermore, formula (55) can be used to calculate the correlation between the first and the second waiting time.

5.3. Examples

The purpose of this section is twofold: (i) we demonstrate that the above quantities can be applied to study the behaviour of batch-service systems and (ii) we investigate the influence of the distribution of the number of customer arrivals on the customer delay.

Fig. 9 shows the mean (part a) and the variance (part b) of the customer delay, in the case of Poisson arrivals, geometrically distributed service times with mean value equal to 10 and a server capacity equal to 10. We can draw the same conclusions as was the case for the system content, except that the curves go to infinity when the service threshold $l > 1$ and $\lambda \to 0$.

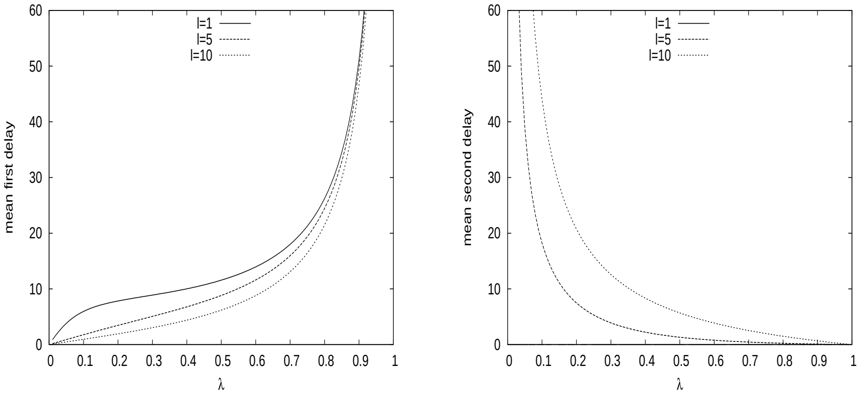
Taking a closer look at $\mathrm{E}[W_1]$ and $\mathrm{E}[W_2]$ (respectively parts a and b of Fig. 10) learns us that the second waiting time is large in light-traffic scenarios, whereas the first is large in heavy-traffic circumstances. Note also that the mean second waiting time equals zero when l equals 1, since the server starts serving when at least one customer (i.e. the tagged customer) is present in this case.



(a) $\mathrm{E}[W]$ (b) $\mathrm{Var}[W]$

Figure 9: $\mathrm{E}[W]$ and $\mathrm{Var}[W]$ versus λ for several values of l

As the inclusion of batch arrivals constitutes one of the major contributions



(a) $\mathrm{E}\left[W_1\right]$

(b) $\mathrm{E}\left[W_2\right]$

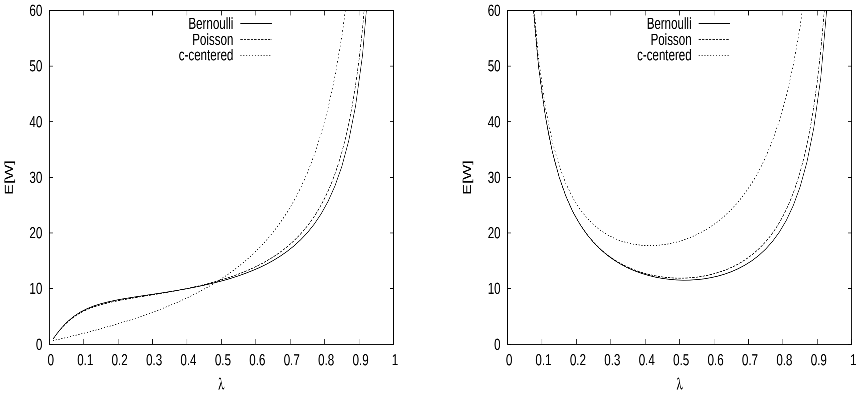
Figure 10: $\mathrm{E}\left[W_1\right]$ and $\mathrm{E}\left[W_2\right]$ vs λ for several values of l

of this paper, we compare the mean (Fig. 11) and the variance (Fig. 12) of the waiting time for three cases of the distribution of the number of customer arrivals in a slot: Bernoulli (i.e. the single-arrival case), Poisson, and the 'c-centered' distribution. The latter has a PGF

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

i.e., either 0, $c-1$ or $c+1$ customers arrive. We observe on the one hand that the performance in case of Poisson arrivals approximates the performance in the event of Bernoulli arrivals. On the other hand, the 'c-centered' arrival distribution clearly leads to rather different results. We can thus conclude that the distribution of the number of customer arrivals in an arbitrary slot plays a significant role in the performance of the system.

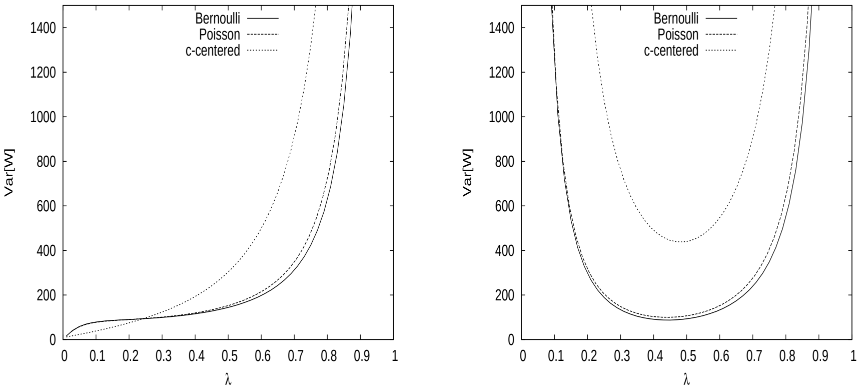
As a closing, we investigate whether the relative difference in the mean delays between $l = 1$ and $l = c$ (defined as $2(D_1 - D_c)/(D_1 + D_c)$, with D_1 (D_c) the customer delay when $l = 1$ ($l = c$)) is affected by the distribution of the arriving batch sizes. Therefore, we consider Fig. 13, where the relative differences in the mean (part a) and the variance (part b) of the customer delay are plotted for the three above mentioned distributions. We observe that the position of the transition points as well as the extent of the performance gain are highly affected by the distribution of the sizes of the arriving batches. Also, the difference between the Poisson and the Bernoulli distribution is larger here. Hence, we can conclude that the inclusion of batch arrivals in the model is a necessity.



(a) $l = 1$

(b) $l = 10$

Figure 11: $\mathrm{E}\left[W\right]$ versus λ for several arrival processes



(a) $l = 1$

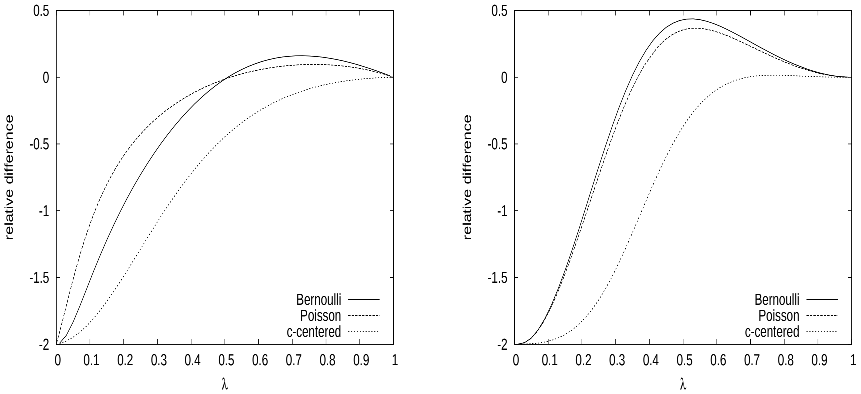
(b) $l = 10$

Figure 12: $\mathrm{Var}\left[W\right]$ versus λ for several arrival processes

6. Conclusions

In this paper, we have studied a general (batch arrivals, general service times, threshold-based service policy) batch-service queueing model. We have first computed the joint probability generating function (PGF) of the number of customers in the queue, the number of customers in service and the remaining service time. From this joint PGF, we have extracted various (known as well as new) important quantities related to the buffer content.

These formulas are semi-analytic, in the sense that they contain some unknown probabilities that have to be calculated numerically. This can become an unfeasible assignment when the batch size is large. Therefore, in a second part of the paper, we have established accurate light- and heavy-traffic approximations of



(a) $\mathrm{E}\,[W]$ (b) $\mathrm{Var}\,[W]$

Figure 13: Relative difference versus λ for several arrival processes

the system content, which require fewer numerical calculations or none at all. In the last part of the paper, we have studied the customer delay. We have first computed the joint PGF of the first (time to serve older batches) and the second (time until the number of present customers reaches or exceeds the service threshold) waiting time. From that joint PGF, we have extracted the marginal PGF's and the PGF of the customer delay. We have illustrated that these PGF's again enable to extract performance measures, by which batch-service systems can be examined. We have also evaluated the influence of the distribution of the sizes of the arriving batches on the mean and the variance of the customer delay. This influence turned out to be significant, implying that the inclusion of batch arrivals in a batch-service queueing model is a necessity.
In conclusion, we feel this paper provides various useful tools to evaluate a wide range of practical batch-service queueing systems.

Acknowledgement: We would like to thank the referees for their valuable comments and suggestions which led to improvements in the paper.

A. Analyticity of the zeroes z_i and $U(z)$ in $\lambda = 0$

In this section, we show that, if $A(\lambda, z)$ is analytic in

$$\mathcal{D} = \{(\lambda, z) : |\lambda| < \delta, |z| < 1 + \gamma\} \ , \qquad \delta > 0, \gamma > 0 \ ,$$

then (i) the zeroes $z_j, 0 \leq j \leq c-1$ of $z^c - T(A(\lambda, z))$ are analytic in $\lambda = 0$, (ii) $q_0(n)$ and $\tilde{u}_n$ are analytic in $\lambda = 0$ and (iii) $U(\lambda, z)$ is analytic in $\lambda = 0$ for $|z| \leq 1$.

From $A(\lambda, z)$ being analytic in $\mathcal{D}$, it follows that $f(\lambda, z) \triangleq z^c - T(A(\lambda, z))$

is analytic in $\mathcal{D}$ (mark that we have previously assumed that the radius of convergence of $T(z)$ is larger than 1). Hence, $f(\lambda, z)$ is analytic in a neighbourhood of the points $(0, \varepsilon_j)$, $0 \leq j \leq c-1$ (a). Further, $f(0, \varepsilon_j) = 0$ and

$$\frac{\partial}{\partial z} f(\lambda, z) \bigg|_{\lambda=0, z=\varepsilon_j} = \frac{c}{\varepsilon_j} \neq 0 \qquad \text{(b)} \ .$$

From (a) and (b) and the implicit function theorem, it follows that there exists a unique function $z_j(\lambda)$, that satisfies

$$f(\lambda, z_j(\lambda)) = 0 \ ,$$

and

$$z_j(0) = \varepsilon_j \ ,$$

and that is analytic in $\lambda = 0$ and this for all j, $0 \leq j \leq c-1$. Next, it is possible to prove that (i) implies (ii) by virtue of the implicit function theorem (see e.g. [16]). Finally, from the calculus of analytic functions, it follows that (iii) also holds.

B. Intuitive explanation of the constant terms in the light-traffic approximation

In case of light traffic, the time between slots during which one or more customers arrive is long (and tends to infinity). Hence, if the system content is below the service threshold l, it will take a long time until at least l customers have accumulated. When, finally, l or more customers are present, these are served immediately as the service time is negligible in this case. As a result, the fraction of slots during which l or more customers are present in the system tends to zero. In fact, we can claim that if the system content equals i, $0 \leq i < l$, and customers arrive so that the system content goes to $nc+j$,$(n,j) \in \ \{\mathbb{N}^2 : j < c\}$, the system content practically immediately evolves to 0 if $j \geq l$ or to j if $j < l$. In addition, due to the BASTA property, the system content at a random slot boundary is equally distributed as the system content at the beginning of a slot during which customers arrive. On account of the above observations, the system content is practically always below l and is governed by a Markov chain with the following transition matrix:

$$\mathbf{P} = [p_{ij}]_{0 \leq i,j \leq l-1} \ ,$$

with

$$p_{ij} = \begin{cases} \sum_{m=l-i}^{c-i} \phi(m) & , \ j = 0, \\ \phi(c-i+j) & , \ j > 0 \text{ and } j \leq i, \\ \phi(j-i) & , \ j > 0 \text{ and } j > i, \end{cases}$$

whereby

$$\phi(m) \triangleq \sum_{n=0}^{\infty} \lim_{k \to \infty} \Pr\left[A_k = nc + m | A_k > 0\right] \ , \qquad 0 \leq m \leq c \ .$$

Of course, when $l = 1$, the system content nearly always equals zero, implying that $\Pr[U = 0] \to 1$. When, on the other hand, $l = c$, $\mathbf{P}$ is a circulant matrix. In addition, as $z^c - T(A(z))$ is aperiodic, the Markov chain governed by the transition matrix $\mathbf{P}$ is irreducible. These two facts imply that the system content is uniformly distributed between 0 and $c - 1$.

References

[1] L. Abolnikov, A. Dukhovny, Optimization in HIV screening problems, Journal of Applied Mathematics and Stochastic Analysis 16(4) (2003) 361–374.

[2] I.J.B.F. Adan, J.S.H. van Leeuwaarden, E.M.M. Winands, On the application of Rouché's theorem in queueing theory, Operations Research Letters 34 (2006) 355–360.

[3] R. Arumuganathan, S. Jeyakumar, Steady state analysis of a bulk queue with multiple vacations, setup times with N-policy and closedown times, Applied Mathematical Modelling 29 (2005) 972–986.

[4] K.E. Atkinson, An introduction to numerical analysis, John Wiley & Sons, New York, 1978.

[5] N.T.J. Bailey, On queueing processes with bulk service, Journal of the Royal Statistical Society, Series B (Methodological) 16(1) (1954) 80–87.

[6] S.K. Bar-Lev, M. Parlar, D. Perry, W. Stadje, F.A. Van der Duyn Schouten, Applications of bulk service queues to group testing models with incomplete identification, European Journal of Operational Research 183 (2007) 226–2374.

[7] H. Bruneel, Buffers with stochastic output interruptions, Electronics Letters 19 (1983) 461–463.

[8] H. Bruneel, B. Steyaert, E. Desmet, G.H. Petit, Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues, European Journal of Operational Research 76 (1994) 563–572.

[9] M.L. Chaudhry, J.G.C. Templeton, A first course in bulk queues, John Wiley & Sons, 1983.

[10] Y. Chen, C. Qiao, X. Yu, Optical burst switching (OBS): a new area in optical networking research, IEEE Network 18(3) (2004) 16–23.

[11] D. Claeys, J. Walraevens, K. Laevens, H. Bruneel, A discrete-time queueing model with a batch server operating under the minimum batch size rule, Lecture Notes in Computer Science 4712 (2007) 248–259.

[12] D. Claeys, K. Laevens, J. Walraevens, H. Bruneel, Delay in a discrete-time queueing model with batch arrivals and batch services, Proceedings of ITNG 2008, International conference on information technology : new generations, 7-9 April 2008, Las Vegas, Nevada, USA, pp. 1040–1045.

[13] D. Claeys, K. Laevens, J. Walraevens, H. Bruneel, Complete characterisation of the customer delay in a queueing system with batch arrivals and batch service, Accepted in Mathematical Methods of Operations Research.

[14] F. Downton, Waiting time in bulk service queues, Journal of the Royal Statistical Society, Series B (Methodological) 17(2) (1955) 256–261.

[15] D. Fiems, H. Bruneel, A note on the discretization of Little's result, Operations Research Letters 30 (2002) 17–18.

[16] G. Hooghiemstra, M. Keane, S. Van De Ree, Power series for stationary distributions of coupled processor models, Siam Journal of Applied Mathematics 48(5) (1988) 1159–1166.

[17] N.K. Kim, K.C. Chae, M.L. Chaudhry, An invariance relation and a unified method to derive stationary queue lengths, Operations Research 52(5) (2004) 756–764.

[18] N.K. Kim, M.L. Chaudhry, Equivalences of batch-service queues and multiserver queues and their complete simple solutions in terms of roots, Stochastic Analysis and Applications 24 (2006) 753–766.

[19] J. Medhi, Waiting time distributions in a Poisson queue with a general bulk service rule, Management Science, Theory Series 21(2) (1975) 777–782.

[20] M.F. Neuts, A general class of bulk queues with Poisson input, Annals of Mathematical Statistics 38 (1967) 759–770.

[21] W.B. Powell, P. Humblet, The bulk service queue with a general control strategy: theoretical analysis and a new computational procedure, Operations Research 34(2) (1986) 267–275.

[22] C.M. Qiao, M.S. Yoo, Optical burst switching (OBS) - a new paradigm for an optical Internet, Journal of high speed networks 8(1) (1999) 69–84.

[23] S.K. Samanta, M.L. Chaudhry, U.C. Gupta, Discrete-time $Geo^X|G^{(a,b)}|1|N$ queues with single and multiple vacations, Mathematical and Computer Modelling 45 (2007) 93–108.

[24] K. Sikdar, U.C. Gupta, Analytic and numerical aspects of batch service queues with single vacation, Computers and Operations Research 32 (2005) 943–966.

[25] Y.Q. Zhao, L.L. Campbell, Equilibrium probability calculations for a discrete-time bulk queue model, Queueing Systems 22 (1996) 189–198.