Analysis of a versatile batch-service queueing model with correlation in the arrival process

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

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

Abstract

In the past, many researchers have analysed queueing models with batch service. In such models, the server typically postpones service until the number of present customers reaches a service threshold, whereupon service is initiated of a batch consisting of several customers. In addition, correlation in the customer arrival process has been studied for many different queueing models. However, correlated arrivals in batch-service models has attracted only modest attention. In this paper, we analyse a discrete-time D-BMAP/$G^{l,c}$/1 queue, whereby the service time of a batch is dependent on the number of customers within it. In addition, a timing mechanism is included, to avoid that customers suffer excessive waiting times because their service is postponed until the amount of customers reaches the service threshold. We deduce various useful performance measures related to the buffer content and we investigate the impact of the traffic parameters on the system performance through some numerical examples. We show that correlation merely has a small impact on the service threshold that minimizes the mean system content, and consequently, that the existing results of the corresponding independent system can be applied to determine a near-optimal service threshold policy, which is an important finding for practitioners. On the other hand, we demonstrate that for other purposes, such as performance evaluation and buffer management, correlation in the arrival process cannot be ignored, a conclusion that runs along the same lines as in queueing models without batch service.

Keywords: queueing, D-BMAP, batch service, service threshold

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

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

1. Introduction

Whereas traditional servers can serve only one customer at a time, batch servers process batches of customers. In fact, a traditional server can be perceived as a special type of batch server, namely whereby the capacity of the server (the maximum number of customers in a served batch) equals one. Batch service is ubiquitous in real life, for instance elevators in high buildings, transport vehicles, recreational devices in amusement parks, ovens in production processes, blood pooling (see e.g. [2], [12], [30]), ... In addition, in telecommunications, information packets are often grouped in larger entities (batches) and these batches are transmitted as a single entity, instead of all packets individually. This is mainly done for efficiency reasons, 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. Technologies using packet aggregation include Optical burst switched (OBS) networks [26], [61] and IEEE 802.11n WLANs [51]. More examples in telecommunications can be found in [13].
Although batch-service (1 server of capacity c) resembles multi-service (c servers of capacity one), it might be less performant: if a customer arrives when the batch server is processing less than c customers, this customer cannot join the ongoing service, whereas the customer would be served immediately by one of the available servers in the multiserver system. In view of this, one often enforces a service threshold for the minimum number of customers that have to be present before the available batch server is allowed to start processing. In practice, an operator typically has to select an efficient service threshold and this could have a huge impact on the performance of the system, as we will demonstrate later on. As batch-service queueing models have a wide area of applications, they have been studied extensively, as well in continuous ([3], [6], [7], [9], [18], [20], [22], [23], [57], [60], [65]) as in discrete time ([19], [28], [29], [31], [37], [39], [43], [46], [62], [70], [72]).

In many real-life circumstances, customer arrivals do not occur independently from each other. For instance, in modern telecommunication systems, a traffic source which is inactive in a given time slot is very likely to remain inactive for a long time (or during a large number of time slots) (see e.g. [36]). In order to cope with the correlated nature of arrivals, the Markovian arrival process (MAP) can be adopted. In case of a MAP, the probability of having an arrival depends on a background state which is governed by a Markov chain. Several variants of MAP exist: in case of a BMAP, customers arrive in batches instead of individually, whereas D-MAP and D-BMAP represent the discrete-time analogues of MAP and BMAP. Queueing models with MAP (or variants) have been studied extensively in the past, for instance the MAP is considered in [4], [5], [8], [14], [44], [48] and [52], the D-MAP is covered in [21], [25], [41], [68], [69], the BMAP is studied in [1], [10], [33], [53]–[55], [58], [59], [64] and [15], [21], [32], [34], [42], [47], [49], [63], [71] deal with D-BMAP.

Although batch-service queueing models and models with MAP (or variants) have been analyzed separately to a great extent, the combination has attracted much less attention. Exceptions are [11], [24], [38], [40], [66]. Gupta and Laxmu [38] studied the queue content at various epochs in the MAP/$G^{a,b}$/1/N queue. Chaudhry and Gupta [24] translated the analysis from [38] to discrete time, resulting in the analysis of the D-MAP/$G^{a,b}$/1/N queue. Gupta and Sikdar [40] extended [38] so that single vacations are included and Sikdar and Gupta [66] further extended this research to multiple vacations. Finally, Banik [11] analyzed the queue content at various epochs in the BMAP/$G^{(a,b)}$/1/N and BMAP/$MSP^{(a,b)}$/1/N systems. Our paper differs from these papers in several aspects. First, we consider the D-BMAP, which is more applicable in a telecommunications context due to the discrete nature of the information units that are typically used. Second, we include a dependency between the service time of a batch and the number of items within it. This is closer to reality, since the transmission time of a batch of information packets is typically longer when the batch contains more packets. Also in other application areas, this might be the case. Thirdly, we incorporate a timing mechanism, that avoids excessive delays due to postponing service until the service threshold is reached. This mechanism is of importance when the customers represent for instance real-time data packets. Further, we deduce an additional set of performance quantities compared to [11], [24], [38], [40], [66], where the queue content is established at service completion, pre-arrival and random times. We compute the system content (i.e. the amount of customers in the entire system, thus those in service included) at random slot boundaries, the queue content at random slot boundaries, the system content at the end of a service, the number of customers in a served batch, the queue content when the server is inactive, the queue content when the server processes at suboptimal capacity and the probability that the server processes a batch during a random slot. The number of customers in a served batch, for instance, is of major concern for practitioners, as it gives a clear indication of the efficiency of the server. Finally, we evaluate more thoroughly the influence of correlation on the behaviour of batch-service queueing systems and more specifically, we investigate the influence on the optimal service threshold.

Our conference paper [27] served as a starting point for this research. In [27], we have studied the system content in a batch-service queueing model with a service threshold, with geometrically distributed service times that are independent of the number of served customers, and with a customer arrival process modelled by a D-BMAP. This paper is an extension of [27], in the sense that we consider a more versatile model with service times that are generally distributed and dependent on the number of items in a served batch. In addition, a timing mechanism is included, to avoid excessive delays due to postponing service until the service threshold is reached. Furthermore, we establish various quantities related to the number of customers in the queue and server at specific time instants, instead of only the system content at random slot bounds as in [27]. We also elaborate upon the influence of correlation on the behaviour of the system.

The paper is organised as follows: the model is described in detail in section 2 and section 3 covers the analysis. The influence of correlation is studied through an example in section 4 and finally some conclusions are drawn in section 5.

2. Model

The queueing model under consideration has the following features:

- The time axis is divided into fixed-length contiguous slots.
- The queue is infinitely large.
- There is one batch server of capacity c (c a constant), which means that the server can process up to c customers simultaneously. When the server becomes available and finds at least as many customers as the service threshold l, it initiates a new service, whereas when the amount of available customers, say j, is smaller than l, the server initiates a service with probability β_j and with probability $1-\beta_j$ it postpones its service. This feature avoids that customers suffer excessive delays because the server waits to initiate service until enough customers have arrived.
- We assume that the already present customers remain in the queue when the server postpones 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 is the period between the start and end of the service of one batch of customers, and we assume that services are synchronised with respect to slot boundaries. The consecutive service times - a service time is the length of a service period, expressed in a number of slots - are dependent on the number of customers in the served batch. Given this number, the service time is independent of all previous service times. We denote the distribution of the service time T_j of a batch of j customers by $t_j(n) \triangleq \Pr[T_j = n]$ and its corresponding probability generating function (PGF) is represented by $T_j(z)$.
- Customers arrive during a slot (and not at slot boundaries). As a result, an arriving customer has to wait for service at least until the next slot mark. This is often referred to as late-arrival with delayed access (for instance in [37], [62]). Further, customers arrive in the buffer according to a homogeneous irreducible aperiodic D-BMAP, meaning that the distribution of the number of customer arrivals per slot depends on a background state which is determined by a homogeneous irreducible and aperiodic first-order Markov chain. The number of background states is finite and denoted by N. We designate the state during slot k by τ_k and during a random slot by τ. Next, let A_k be the amount of customers arriving in slot k. The arrival process is completely defined by the values $a(n, j|i)$:

$$a(n, j|i) \triangleq \lim_{k\to\infty} \Pr[A_k = n, \tau_{k+1} = j | \tau_k = i] \quad , \qquad n \geq 0; i, j \in \{1, \ldots, N\} \quad ,$$

denoting the probability that if the background state is i during a slot, there are n arrivals during this slot and the background state during the next slot is j. We put these probabilities in an $N \times N$ **matrix generating function** $\mathbf{A}(z)$, whose entries are defined as follows:

$$[\mathbf{A}(z)]_{ij} \triangleq \sum_{n=0}^{\infty} a(n, j|i) z^n \ .$$

The radius of convergence of $\mathbf{A}(z)$ is designated by $\Re$ and is equal to

$$\Re \triangleq \min_{i,j} \Re_{i,j} \ ,$$

with $\Re_{i,j}$ the radius of convergence of $[\mathbf{A}(z)]_{ij}$. Hence, each of the entries of $\mathbf{A}(z)$ is an analytic function in the open disk $\{z \in \mathbb{C} : |z| < \Re\}$.

It is worth noting that in traditional literature D-BMAP is described by standard notations $\mathbf{D}_0$, $\mathbf{D}_1$, $\mathbf{D}_2$, et cetera, whereby $[\mathbf{D}_n]_{ij} \triangleq a(n, j|i)$ (see e.g. [10], [15], [42], [63]). Hence, the connection with our notation is as follows:

$$\mathbf{A}(z) = \sum_{n=0}^{\infty} \mathbf{D}_n z^n \ .$$

We prefer working with probability generating matrix $\mathbf{A}(z)$ for a twofold reason: it completely describes the arrival process and it is convenient throughout the analysis. The following information can be extracted from $\mathbf{A}(z)$:

- Transition probabilities of the underlying Markov chain:

$$\Pr[\tau_{k+1} = j | \tau_k = i] = [\mathbf{A}(1)]_{ij} \ , \qquad \forall k \in \mathbb{N} \ .$$

- Stationary distribution $1 \times N$ vector $\boldsymbol{\pi}$ of the state of the underlying Markov chain:

$$[\boldsymbol{\pi}]_i \triangleq \lim_{k \to \infty} \Pr[\tau_k = i] \ , \qquad 1 \leq i \leq N \ ,$$

  is the solution of $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{A}(1)$ and the normalization condition $\boldsymbol{\pi}\mathbf{1} = 1$, with $\mathbf{1}$ the $N \times 1$ column vector whose N entries are equal to 1.

- Conditional PGF of the number of arrivals given that the background state during that slot equals i:

$$A_i(z) \triangleq \lim_{k \to \infty} \sum_{n=0}^{\infty} \Pr[A_k = n | \tau_k = i] \, z^n = [\mathbf{A}(z)\mathbf{1}]_i \ .$$

- Mean arrival rate λ, i.e., the average number of customer arrivals during a random slot:

$$\lambda \triangleq \sum_{i=1}^{N} \Pr[\tau = i] \, A_i^{'}(1) = \boldsymbol{\pi}\mathbf{A}^{'}(1)\mathbf{1} \ ,$$

  whereby

$$[\mathbf{A}^{'}(1)]_{ij} = \sum_{n=0}^{\infty} a(n, j|i) n = \frac{\mathrm{d}}{\mathrm{d}z}[\mathbf{A}(z)]_{ij}\Big|_{z=1} \ ,$$

  (we use primes to indicate derivatives). Note that with the above definitions, the stability condition of this system requires that the load $\rho \triangleq \frac{\lambda \mathrm{E}[T_c]}{c} < 1$ (we hereby have taken into account that the server nearly always processes c customers in case of heavy traffic).

The matrix generating function $\mathbf{A}(z)$ is convenient to deal with as the matrix generating function of the number of arrivals during n consecutive slots is equal to $\mathbf{A}(z)^n$. Similarly, the matrix generating function of the number of arrivals during the service of n customers equals $T_n(\mathbf{A}(z)) \triangleq \sum_{k=0}^{\infty} \Pr[T_n = k]\, \mathbf{A}(z)^k$.

During the analysis in the next section, we will make use of **spectral decomposition**[2]. We thereby assume that $\mathbf{A}(z)$ is diagonalizable, i.e., $\mathbf{A}(z)$ can be factorized as

$$\mathbf{A}(z) = \mathbf{R}(z)\mathbf{\Lambda}(z)\mathbf{R}^{-1}(z) \ , \tag{1}$$

with $\mathbf{\Lambda}(z)$ a diagonal matrix. The condition that $\mathbf{A}(z)$ is diagonalizable is standard in order to be able to apply the spectral decomposition approach ([35], [49], [50], [67], [71]) and is not a specific restriction on the generality of the model. Contribution [35] contains a detailed and extensive analysis on the conditions under which such a solution exists. It can be proved (see e.g. [56]) that $\mathbf{A}(z)$ is diagonalizable if and only if it possesses a complete set of eigenvectors, that the columns $\mathbf{r}_j(z)$ of $\mathbf{R}(z)$ then constitute a complete set of right eigenvectors and that the diagonal entries $\lambda_i(z)$ of $\mathbf{\Lambda}(z)$ are the eigenvalues of $\mathbf{A}(z)$, so that each $(\lambda_j(z), \mathbf{r}_j(z))$ is an eigenpair for $\mathbf{A}(z)$:

$$\mathbf{A}(z)\mathbf{r}_j(z) = \lambda_j(z)\mathbf{r}_j(z) \ , \qquad 1 \leq j \leq N \ .$$

Note that the eigenvectors are unique upon some factor, which we can, without loss of generality, fix by making the convention that the row sums of either $\mathbf{R}(z)$ or $\mathbf{R}^{-1}(z)$ are equal to one (the former implies the latter and vice versa):

$$\mathbf{R}(z)\mathbf{1} = \mathbf{1} \Leftrightarrow \mathbf{R}^{-1}(z)\mathbf{1} = \mathbf{1} \ .$$

This convention will turn out to be convenient for further calculations. Next, relation (1) implies that

$$\mathbf{A}(z)^n = \mathbf{R}(z)\mathbf{\Lambda}(z)^n\mathbf{R}^{-1}(z) \ ,$$

$$T_c(\mathbf{A}(z)) = \mathbf{R}(z)T_c(\mathbf{\Lambda}(z))\mathbf{R}^{-1}(z) \ ,$$

which means that

$$\mathbf{A}(z)^n\mathbf{r}_j(z) = \lambda_j(z)^n\mathbf{r}_j(z) \ , \qquad 1 \leq j \leq N \ ,$$

and

$$T_c(\mathbf{A}(z))\mathbf{r}_j(z) = T_c(\lambda_j(z))\mathbf{r}_j(z) \ , \qquad 1 \leq j \leq N \ .$$

In other words, each $\mathbf{r}_j(z)$ is a right eigenvector of $\mathbf{A}(z)^n$ and $T_c(\mathbf{A}(z))$ as well, with corresponding eigenvalues $\lambda_j(z)^n$ and $T_c(\lambda_j(z))$ respectively.

Next, since $\mathbf{A}(z)$ is a matrix with positive entries for all $z \in]0, \Re[$, it has one real and positive eigenvalue that exceeds the moduli of all other eigenvalues for these values of z ([56]). This eigenvalue is called the **Perron-Frobenius (PF) eigenvalue** and we let $\lambda_1(z)$ represent that eigenvalue. The PF eigenvalue and its corresponding right eigenvector satisfy $\lambda_1(1) = 1$, $\mathbf{r}_1(1) = \mathbf{1}$. In addition, it

[2]For a good introduction on matrix algebra and more specifically on spectral decomposition we recommend the book [56].

can be proved that $\lambda_1^{'}(1) = \lambda$. Indeed, it holds that $\mathbf{A}(z)\mathbf{r}_1(z) = \lambda_1(z)\mathbf{r}_1(z)$. Taking the first derivative at $z = 1$ and invoking $\lambda_1(1) = 1$ and $\mathbf{r}_1(1) = \mathbf{1}$ yields

$$\mathbf{A}^{'}(1)\mathbf{1} + \mathbf{A}(1)\mathbf{r}_1^{'}(1) = \lambda_1^{'}(1)\mathbf{1} + \mathbf{r}_1^{'}(1) \ . \tag{2}$$

Multiplying (2) to the left with $\boldsymbol{\pi}$, the steady-state vector of the state of the underlying Markov chain, relying on $\boldsymbol{\pi}\mathbf{A}(1) = \boldsymbol{\pi}$, $\boldsymbol{\pi}\mathbf{A}^{'}(1)\mathbf{1} = \lambda$ and $\boldsymbol{\pi}\mathbf{1} = 1$, produces

$$\lambda_1^{'}(1) = \lambda \ .$$

Before closing this section, we define the **vector generating function** $\mathbf{X}(z)$ of a random variable X that depends on the state τ_k (X_k represents its value at slot mark k) as the $1 \times N$ vector whose entries are defined as follows:

$$[\mathbf{X}(z)]_j \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{X_k}\mathbf{1}_{\{\tau_k=j\}}\right] \ ,$$

with $\mathbf{1}_{\{Y\}}$ the indicator function of Y.

3. Analysis

This section is organised as follows: first, we compute the joint vector generating function of the queue content, the server content and the remaining service time. From this formula and from some intermediate results, we then deduce various relevant quantities and finally we explain how performance measures can be extracted from these quantities.

3.1. Joint vector generating function $\mathbf{W}(z, x, y)$

In this subsection, we compute the joint vector generating function $\mathbf{W}(z, x, y)$ of the queue content, the server content and the remaining service time:

$$[\mathbf{W}(z, x, y)]_j \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k}x^{S_k}y^{R_k}\mathbf{1}_{\{\tau_k=j\}}\right] \ ,$$

with Q_k (S_k) the queue (server) content and R_k the remaining service time at slot boundary k. We commence by writing down the system equations, which express the relation between $(Q_{k+1}, S_{k+1}, R_{k+1})$ and (Q_k, S_k, R_k):

$$(Q_{k+1}, S_{k+1}, R_{k+1}) = \begin{cases} (Q_k + A_k, S_k, R_k - 1) & \text{if } R_k > 1 \\ (0, Q_k + A_k, T_{Q_k+A_k}) & \text{if } R_k \leq 1 \text{ and } l \leq Q_k + A_k < c \\ (Q_k + A_k - c, c, T_c) & \text{if } R_k \leq 1 \text{ and } Q_k + A_k \geq c \\ (0, Q_k + A_k, T_{Q_k+A_k}) & \text{if } R_k \leq 1,\ Q_k + A_k < l \text{ and service starts (with probability } \beta_{Q_k+A_k}) \\ (Q_k + A_k, 0, 0) & \text{if } R_k \leq 1,\ Q_k + A_k < l \text{ and service does not start (with probability } 1 - \beta_{Q_k+A_k}) \end{cases}$$

Indeed, in the first case, the service continues during slot $k+1$, so that customers that have arrived during slot k are stored in the queue. In the other cases, the server is available at slot mark $k + 1$. Whether a new service is initiated or not is described by the rules mentioned in section 2 and is thus dependent on the number of available customers.

Remark 1. *$S_k = 0$ does not necessarily imply $R_k = 0$. Indeed, since it is allowed that β_0 might differ from zero, it is possible to start a service with 0 customers. This can be interpreted as a server vacation, which duration is characterized by the PGF $T_0(z)$.*

The system equations can be translated into vector generating functions as follows:

$$\begin{aligned}[\mathbf{W}_{k+1}(z,x,y)]_j \triangleq & \mathrm{E}\left[z^{Q_{k+1}} x^{S_{k+1}} y^{R_{k+1}} \mathbf{1}_{\{\tau_{k+1}=j\}}\right] \\ = & \frac{1}{y}\mathrm{E}\left[z^{Q_k+A_k} x^{S_k} y^{R_k} \mathbf{1}_{\{R_k>1,\tau_{k+1}=j\}}\right] \\ & + \mathrm{E}\left[x^{Q_k+A_k} y^{T_{Q_k+A_k}} \mathbf{1}_{\{R_k\leq 1, l\leq Q_k+A_k<c, \tau_{k+1}=j\}}\right] \\ & + \left(\frac{x}{z}\right)^c T_c(y)\mathrm{E}\left[z^{Q_k+A_k} \mathbf{1}_{\{R_k\leq 1, Q_k+A_k\geq c, \tau_{k+1}=j\}}\right] \\ & + \mathrm{E}\left[x^{Q_k+A_k} y^{T_{Q_k+A_k}} \mathbf{1}_{\{R_k\leq 1, Q_k+A_k<l, \text{service starts}, \tau_{k+1}=j\}}\right] \\ & + \mathrm{E}\left[z^{Q_k+A_k} \mathbf{1}_{\{R_k\leq 1, Q_k+A_k<l, \text{no service start}, \tau_{k+1}=j\}}\right] \quad . \end{aligned} \tag{3}$$

We now calculate each term from the right-hand-side of (3) separately. We therefore introduce the $1 \times N$ row vectors $\mathbf{q}_{0k}(n)$, $\mathbf{d}_k(n)$ and $\mathbf{F}_k(z,x)$:

$$[\mathbf{q}_{0k}(n)]_j \triangleq \Pr\left[Q_k = n, R_k = 0, \tau_k = j\right] \quad , \tag{4}$$

$$[\mathbf{d}_k(n)]_j \triangleq \Pr\left[Q_k + A_k = n, R_k \leq 1, \tau_{k+1} = j\right] \quad , \tag{5}$$

$$[\mathbf{F}_k(z,x))]_j \triangleq \mathrm{E}\left[z^{Q_k} x^{S_k} \mathbf{1}_{\{R_k=1,\tau_k=j\}}\right] \quad . \tag{6}$$

Let us start with the first term from (3). We take the sum over all possible states τ_k during slot k:

$$\begin{aligned} & \mathrm{E}\Big[z^{Q_k+A_k} x^{S_k} y^{R_k} \mathbf{1}_{\{R_k>1,\tau_{k+1}=j\}}\Big] \\ & \qquad = \sum_{i=1}^{N} \mathrm{E}\left[z^{Q_k+A_k} x^{S_k} y^{R_k} \mathbf{1}_{\{R_k>1,\tau_{k+1}=j,\tau_k=i\}}\right] \quad . \end{aligned}$$

As A_k is independent of Q_k, S_k and R_k when τ_k and τ_{k+1} are given, this expresssion can be transformed into

$$\begin{aligned} & \mathrm{E}\Big[z^{Q_k+A_k} x^{S_k} y^{R_k} \mathbf{1}_{\{R_k>1,\tau_{k+1}=j\}}\Big] \\ & \qquad = \sum_{i=1}^{N} \mathrm{E}\left[z^{Q_k} x^{S_k} y^{R_k} \mathbf{1}_{\{R_k>1\}} | \tau_{k+1} = j, \tau_k = i\right] \Pr\left[\tau_k = i\right] \\ & \qquad\quad . \ \mathrm{E}\left[z^{A_k} | \tau_{k+1} = j, \tau_k = i\right] \Pr\left[\tau_{k+1} = j | \tau_k = i\right] \quad . \end{aligned}$$

Note that Q_k, S_k and R_k are independent of τ_{k+1} if τ_k is given. Indeed, Q_k, S_k and R_k are influenced by A_{k-1}, which is not dependent of τ_{k+1} if τ_k is known. As a result, we find,

$$\begin{aligned} & \mathrm{E}\Big[z^{Q_k+A_k} x^{S_k} y^{R_k} \mathbf{1}_{\{R_k>1,\tau_{k+1}=j\}}\Big] \\ & \qquad = \sum_{i=1}^{N} \mathrm{E}\left[z^{Q_k} x^{S_k} y^{R_k} \mathbf{1}_{\{R_k>1,\tau_k=i\}}\right] [\mathbf{A}(z)]_{ij} \quad . \end{aligned}$$

Applying the law of total probability yields

$$\begin{aligned}
&\mathrm{E}\Big[z^{Q_k+A_k}x^{S_k}y^{R_k}\mathbf{1}_{\{R_k>1,\tau_{k+1}=j\}}\Big] \\
&\qquad=\sum_{i=1}^{N}\Bigg[\mathrm{E}\left[z^{Q_k}x^{S_k}y^{R_k}\mathbf{1}_{\{\tau_k=i\}}\right]-\mathrm{E}\left[z^{Q_k}x^{S_k}y^{R_k}\mathbf{1}_{\{R_k=1,\tau_k=i\}}\right] \\
&\qquad\qquad-\mathrm{E}\left[z^{Q_k}x^{S_k}y^{R_k}\mathbf{1}_{\{R_k=0,\tau_k=i\}}\right]\Bigg][\mathbf{A}(z)]_{ij}\ .
\end{aligned}$$

Next, $R_k = 0$ (i.e., no service in slot k) implies that $S_k = 0$ and $Q_k < l$. As a consequence, we obtain, by taking into account the definitions of $\mathbf{W}_k(z,x,y)$ and those for $\mathbf{q}_{0k}(n)$ and $\mathbf{F}_k(z,x)$ ((4) and (6)):

$$\begin{aligned}
&\mathrm{E}\left[z^{Q_k+A_k}x^{S_k}y^{R_k}\mathbf{1}_{\{R_k>1,\tau_{k+1}=j\}}\right] \\
&\qquad=\sum_{i=1}^{N}\left[\mathbf{W}_k(z,x,y)-y\mathbf{F}_k(z,x)-\sum_{n=0}^{l-1}\mathbf{q}_{0k}(n)z^n\right]_i[\mathbf{A}(z)]_{ij}\ ,
\end{aligned}$$

which is nothing else than a matrix multiplication. Hence,

$$\begin{aligned}
&\mathrm{E}\left[z^{Q_k+A_k}x^{S_k}y^{R_k}\mathbf{1}_{\{R_k>1,\tau_{k+1}=j\}}\right] \\
&\qquad=\left[\left\{\mathbf{W}_k(z,x,y)-y\mathbf{F}_k(z,x)-\sum_{n=0}^{l-1}\mathbf{q}_{0k}(n)z^n\right\}\mathbf{A}(z)\right]_j\ . \qquad (7)
\end{aligned}$$

The third term from the right-hand-side of (3) can be established analogously as the first, which yields:

$$\begin{aligned}
&\mathbf{E}\Big[z^{Q_k+A_k}\mathbf{1}_{\{R_k\le 1,Q_k+A_k\ge c,\tau_{k+1}=j\}}\Big] \\
&\qquad=\left[\mathbf{F}_k(z,1)\mathbf{A}(z)+\sum_{n=0}^{l-1}\mathbf{q}_{0k}(n)z^n\mathbf{A}(z)-\sum_{n=0}^{c-1}\mathbf{d}_k(n)z^n\right]_j\ . \qquad (8)
\end{aligned}$$

The other terms are easier to calculate, because we just have to rely on definition (5) of $\mathbf{d}_k(n)$. As a result, we find for respectively the second, fourth and fifth term:

$$\mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\mathbf{1}_{\{R_k\le 1,l\le Q_k+A_k<c,\tau_{k+1}=j\}}\right]=\left[\sum_{n=l}^{c-1}\mathbf{d}_k(n)x^nT_n(y)\right]_j\ , \qquad (9)$$

$$\begin{aligned}
&\mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\mathbf{1}_{\{R_k\le 1,Q_k+A_k<l,\text{service starts},\tau_{k+1}=j\}}\right] \\
&\qquad=\left[\sum_{n=0}^{l-1}\mathbf{d}_k(n)\beta_nx^nT_n(y)\right]_j\ , \qquad (10)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}\left[z^{Q_k+A_k}\mathbf{1}_{\{R_k\le 1,Q_k+A_k<l,\text{no service start},\tau_{k+1}=j\}}\right] \\
&\qquad=\left[\sum_{n=0}^{l-1}\mathbf{d}_k(n)(1-\beta_n)z^n\right]_j\ . \qquad (11)
\end{aligned}$$

Owing to the definition of vector equality, the substitution of (7)-(11) in (3) produces in the steady state

$$\begin{aligned}\mathbf{W}(z,x,y) =& \frac{1}{y}\left\{\mathbf{W}(z,x,y) - y\mathbf{F}(z,x) - \sum_{n=0}^{l-1}\mathbf{q}_0(n)z^n\right\}\mathbf{A}(z) \\ &+ \sum_{n=l}^{c-1}\mathbf{d}(n)x^n T_n(y) \\ &+ \left(\frac{x}{z}\right)^c T_c(y)\left[\mathbf{F}(z,1)\mathbf{A}(z) + \sum_{n=0}^{l-1}\mathbf{q}_0(n)z^n\mathbf{A}(z) - \sum_{n=0}^{c-1}\mathbf{d}(n)z^n\right] \\ &+ \sum_{n=0}^{l-1}\mathbf{d}(n)\beta_n x^n T_n(y) + \sum_{n=0}^{l-1}\mathbf{d}(n)(1-\beta_n)z^n \ , \end{aligned} \tag{12}$$

whereby the $1\times N$ row vectors $\mathbf{q_0}(n)$, $\mathbf{d}(n)$ and $\mathbf{F}(z,x)$ represent the steady-state equivalents of $\mathbf{q}_{0k}(n)$, $\mathbf{d}_k(n)$ and $\mathbf{F}_k(z,x)$:

$$[\mathbf{q_0}(n)]_j \triangleq \lim_{k\to\infty}\Pr\left[Q_k = n, R_k = 0, \tau_k = j\right] \ , \tag{13}$$

$$[\mathbf{d}(n)]_j \triangleq \lim_{k\to\infty}\Pr\left[Q_k + A_k = n, R_k \leq 1, \tau_{k+1} = j\right] \ , \tag{14}$$

$$[\mathbf{F}(z,x))]_j \triangleq \lim_{k\to\infty}\mathrm{E}\left[z^{Q_k}x^{S_k}\mathbf{1}_{\{R_k=1,\tau_k=j\}}\right] \ .$$

Next, mark that definitions (13) and (14) imply that

$$\mathbf{q}_0(n) = \mathbf{d}(n)(1-\beta_n) \ , \qquad 0 \leq n \leq l-1 \ . \tag{15}$$

Indeed, "$Q_{k+1} = n, R_{k+1} = 0$" means that the server is not processing during slot $k+1$ and that n customers are present at the beginning of that slot. This can only be the case if the server is or becomes available at the end of slot k ($R_k \leq 1$) and if n customers are present at that moment (i.e. $Q_k + A_k = n$) and if the server does not start service anyway at slot mark $k+1$ (with probability $(1-\beta_n)$). Hence, $\Pr\left[Q_{k+1} = n, R_{k+1} = 0\right] = \Pr\left[Q_k + A_k = n, R_k \leq 1\right](1-\beta_n)$. Since $\rho < 1$, this becomes independent of the slot index k (or $k+1$), and thus leads to expression (15).

Substitution of (15) in (12) produces

$$\begin{aligned}\mathbf{W}(z,x,y)\left[\mathbf{I} - \frac{1}{y}\mathbf{A}(z)\right] =& \sum_{n=0}^{l-1}\mathbf{d}(n)(1-\beta_n)z^n\left[\mathbf{I} - \frac{\mathbf{A}(z)}{y}\right] \\ &+ \left(\frac{x}{z}\right)^c T_c(y)\sum_{n=0}^{l-1}\mathbf{d}(n)z^n[\mathbf{A}(z) - \mathbf{I}] \\ &+ \sum_{n=0}^{l-1}\mathbf{d}(n)\beta_n\left[x^n T_n(y)\mathbf{I} - z^n\left(\frac{x}{z}\right)^c T_c(y)\mathbf{A}(z)\right] \\ &+ \left(\frac{x}{z}\right)^c T_c(y)\mathbf{F}(z,1)\mathbf{A}(z) - \mathbf{F}(z,x)\mathbf{A}(z) \\ &+ \sum_{n=l}^{c-1}\mathbf{d}(n)\left[x^n T_n(y) - z^n\left(\frac{x}{z}\right)^c T_c(y)\right] \ , \end{aligned} \tag{16}$$

with $\mathbf{I}$ the $N \times N$ identity matrix. For the purpose of extracting performance measures in the next sections, it turns out to be more convenient to multiply

expression (16) to the right with $\mathbf{r}_i(z)$, the i-th right eigenvector of $\mathbf{A}(z)$. We obtain

$$\begin{aligned}\left[1-\frac{\lambda_i(z)}{y}\right]\mathbf{W}(z,x,y)\mathbf{r}_i(z) = &\left[1-\frac{\lambda_i(z)}{y}\right]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)(1-\beta_n)z^n \\ &+\left(\frac{x}{z}\right)^c T_c(y)[\lambda_i(z)-1]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n \\ &+\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\beta_n\left[x^nT_n(y)-z^n\left(\frac{x}{z}\right)^c T_c(y)\lambda_i(z)\right] \\ &+\left(\frac{x}{z}\right)^c T_c(y)\lambda_i(z)G_i(z,1)-\lambda_i(z)G_i(z,x) \\ &+\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[x^nT_n(y)-z^n\left(\frac{x}{z}\right)^c T_c(y)\right] \quad , \end{aligned} \tag{17}$$

with

$$G_i(z,x) \triangleq \mathbf{F}(z,x)\mathbf{r}_i(z) \ .$$

Substituting y by $\lambda_i(z)$, multiplying by z^c and letting $x \to 1$ leads to the following expression for $G_i(z,1)$:

$$\begin{aligned}\lambda_i(z)\left[z^c-T_c(\lambda_i(z))\right]G_i(z,1) = &T_c(\lambda_i(z))[\lambda_i(z)-1]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n \\ &+\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\beta_n\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\lambda_i(z)\right] \\ &+\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\right] \ . \end{aligned} \tag{18}$$

Finally, substituting y by $\lambda_i(z)$ in (17), multiplying by $z^c\left[z^c-T_c(\lambda_i(z))\right]$ and appealing to (18) yields

$$\begin{aligned}&z^c\lambda_i(z)\left[z^c-T_c(\lambda_i(z))\right]G_i(z,x) \\ &=z^cx^cT_c(\lambda_i(z))[\lambda_i(z)-1]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n \\ &\quad+x^cT_c(\lambda_i(z))\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\beta_n\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\lambda_i(z)\right] \\ &\quad+x^cT_c(\lambda_i(z))\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\right] \\ &\quad+\left[z^c-T_c(\lambda_i(z))\right]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\beta_n\left[z^cx^nT_n(\lambda_i(z))-x^cz^nT_c(\lambda_i(z))\lambda_i(z)\right] \\ &\quad+\left[z^c-T_c(\lambda_i(z))\right]\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cx^nT_n(\lambda_i(z))-x^cz^nT_c(\lambda_i(z))\right] \ . \end{aligned} \tag{19}$$

Expressions (17)-(19) provide enough information to deduce a spectrum of quantities related to the buffer content, which constitutes the subject of the next section.

3.2. Quantities related to the buffer content

3.2.1. System content at random slot boundaries

As the system content U equals the sum of the queue and the server content, its vector generating function $\tilde{\mathbf{U}}(z)$ is found by letting $y \to 1$ and $x \to z$ in (17) and applying (18) and (19), resulting in

$$\begin{aligned}[1-\lambda_i(z)]\left[z^c - T_c(\lambda_i(z))\right]\mathbf{U}(z)\mathbf{r}_i(z) =& (z^c-1)T_c(\lambda_i(z))[1-\lambda_i(z)]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n \\ &+\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{g}_{n,i}(z)\beta_n + \sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{h}_{n,i}(z)\ , \end{aligned} \tag{20}$$

with

$$\begin{aligned}\mathbf{g}_{n,i}(z) \triangleq& [(z^n - z^c)T_n(\lambda_i(z))T_c(\lambda_i(z)) + z^n(z^c-1)T_c(\lambda_i(z))\lambda_i(z) \\ &-z^c(z^n-1)T_n(\lambda_i(z))]\,\mathbf{r}_i(z)\ , \end{aligned}$$

$$\begin{aligned}\mathbf{h}_{n,i}(z) \triangleq& [T_n(\lambda_i(z))z^c\{(1-z^n) - T_c(\lambda_i(z))\} \\ &-T_c(\lambda_i(z))z^n\{(1-z^c) - T_n(\lambda_i(z))\}]\,\mathbf{r}_i(z)\ . \end{aligned}$$

The unknown vectors $\mathbf{d}(n)$ have to be determined by solving a set of linear equations. This is explained in section 3.3.1.

3.2.2. Queue content at random slot boundaries

The vector generating function $\mathbf{Q}(z)$ of the queue content at random slot boundaries is found by summing out both the server content and the remaining service time from $\mathbf{W}(z,x,y)$. Hence, letting $y \to 1$ and $x \to 1$ in (17) and applying (18), we find

$$\begin{aligned}&[1-\lambda_i(z)]\left[z^c - T_c(\lambda_i(z))\right]\mathbf{Q}(z)\mathbf{r}_i(z) \\ &\quad=(z^c-1)[1-\lambda_i(z)]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n \\ &\qquad+\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\beta_n\left[(1-z^n)\{z^c - T_c(\lambda_i(z))\} + (z^c-1)\{z^n\lambda_i(z) - T_n(\lambda_i(z))\}\right] \\ &\qquad+\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^c - z^n + (z^n-1)T_c(\lambda_i(z)) + (1-z^c)T_n(\lambda_i(z))\right]\ . \end{aligned} \tag{21}$$

Remark 2. *Note that in [45], where the service times are independent of the number of customers in a served batch, it holds that* $\mathbf{U}(z) = \mathbf{Q}(z)T(\mathbf{A}(z))$*, with* $T(z)$ *the PGF of the service times. Here, the service times are dependent on the number of customers in a served batch and it thus turns out that a similar relation as in [45] does not hold anymore.*

3.2.3. System content at the end of a service

The system content $\tilde{U}$ at the end of a service equals the sum of the queue content at the beginning of the last slot of the service and the customers that have arrived during that slot. Hence, by definition, we get

$$\tilde{\mathbf{U}}(z) = \frac{\mathbf{F}(z,1)\mathbf{A}(z)}{\mathbf{F}(1,1)\mathbf{1}}\ , \tag{22}$$

or, by multiplying both sides to the right with $\mathbf{1}$:

$$\tilde{U}(z) = \frac{\mathbf{F}(z,1)\mathbf{A}(z)\mathbf{1}}{\mathbf{F}(1,1)\mathbf{1}} \ .$$

Note however that we have deduced an expression for $G_i(z,x)$ and not for $\mathbf{F}(z,x)$. We therefore multiply both sides of (22) to the right with $\mathbf{r}_i(z)$, which yields

$$\tilde{\mathbf{U}}(z)\mathbf{r}_i(z) = \lambda_i(z)\frac{G_i(z,1)}{G_1(1,1)} \ . \tag{23}$$

We have hereby taken into account that $\mathbf{r}_1(1) = \mathbf{1}$.

3.2.4. Number of customers in a served batch

The number of customers in a random served batch, $\tilde{S}$, is equally distributed as the server content at the last slot of a random service period, which yields

$$\tilde{\mathbf{S}}(z) = \frac{\mathbf{F}(1,z)}{\mathbf{F}(1,1)\mathbf{1}} \ . \tag{24}$$

As we have deduced an expression for $G_i(z,x)$ instead of for $\mathbf{F}(z,x)$, we rewrite (24), by multiplying both sides to the right with $\mathbf{1}$, as

$$\tilde{S}(z) = \frac{G_1(1,z)}{G_1(1,1)} \ . \tag{25}$$

3.2.5. Queue content when the server is inactive

The queue content when the server is inactive (because none or not enough, i.e., less than l, customers are present), say $\tilde{Q}$, is found by taking into account that the server is not processing if and only if the remaining service time equals 0. Hence

$$\tilde{Q}(z) = \frac{\sum_{n=0}^{l-1} \mathbf{q}_0(n)\mathbf{1}z^n}{\sum_{n=0}^{l-1} \mathbf{q}_0(n)\mathbf{1}} \ . \tag{26}$$

3.2.6. Queue content when the server processes at suboptimal capacity

The vector generating function $\mathbf{Q}^*(z)$ of the queue content at random slot boundaries when the server processes at suboptimal capacity (i.e., the server is serving less than c customers: $R_k \neq 0$, $0 \leq S_k \leq c-1$) is found by applying the law of total probability and thereby taking into account the probability generating property of generating functions and $S_k = c \Rightarrow R_k \neq 0$,

$$\mathbf{Q}^*(z)\mathbf{r}_i(z) = \frac{\left[\mathbf{W}(z,1,1) - \mathbf{W}(z,0,0) - \frac{1}{c!}\frac{\partial^c}{\partial x^c}\mathbf{W}(z,x,1)\Big|_{x=0}\right]\mathbf{r}_i(z)}{1 - \left[\mathbf{W}(1,0,0) + \frac{1}{c!}\frac{\partial^c}{\partial x^c}\mathbf{W}(1,x,1)\Big|_{x=0}\right]\mathbf{r}_1(1)} \ . \tag{27}$$

Note that we have multiplied this equation with $\mathbf{r}_i(z)$, because we have established a formula for $\mathbf{W}(z,x,y)\mathbf{r}_i(z)$ and not for $\mathbf{W}(z,x,y)$ individually.

3.3. Performance measures

In the previous subsection, we have deduced various quantities related to the buffer content ((20), (21), (23), (25)-(27)). These formulas allow us to calculate performance measures such as moments and tail probabilities. As compared to the case of independent arrivals, this matter is more complicated now and

we therefore briefly explain how the mean value (subsection 3.3.2) and the tail probabilities (subsection 3.3.3) of the system content can be calculated (for a more extensive treatment we refer to more technical papers such as [35]). As formula (20) contains the unknown vectors $\mathbf{d}(n)$, we first compute these vectors in subsection 3.3.1.

3.3.1. Calculation of the vectors $\mathbf{d}(n)$

We start with rewriting (20) as follows:

$$[1-\lambda_i(z)]\,[z^c - T_c(\lambda_i(z))]\,\mathbf{U}(z)\mathbf{r}_i(z) = f_i(z)\ , \qquad 1 \le i \le N\ , \tag{28}$$

whereby $f_i(z)$ represents the right-hand-side of (20). Unlike the case of independent arrivals, it is impossible to construct an irrefutable mathematical proof, based on Rouché's theorem, to show that each of the equations $z^c - T_c(\lambda_i(z)) = 0\ , 1 \le i \le N$ necessarily has c solutions inside the closed complex unit disk. Nevertheless, an example where this is not the case has not been encountered up to now, and, to the best of our knowledge, such an example, if it exists, has yet to be constructed. Hence, for practical purposes, we can venture to state that the above equation has indeed c solutions inside the closed complex unit disk for each value of i, provided that the equilibrium condition $\rho < 1$ holds. Let us characterise the k-th solution of the i-th equation by $z_{i,k}$. As $\lambda_1(1) = 1$, one of the zeros of $z^c - T_c(\lambda_1(z))$ equals one. Without loss of generality, we let $z_{1,1}$ be that zero. As $\mathbf{U}(z)$ is analytic inside the closed complex unit disk, $f_i(z)$ must also vanish at $z_{i,k}$, $1 \le k \le c$ and this for all i, $1 \le i \le N$. This observation leads to $Nc-1$ linear equations in the $1 \times N$ vectors $\mathbf{d}(n)$, $0 \le n \le c-1$. The zero $z_{1,1}$ cannot be used as it produces the trivial equation $0 = 0$. Fortunately, we can resort to the normalisation condition to obtain another equation. Deriving (28) twice at $z = 1$ for $i = 1$ and taking into account that $\mathbf{r}_1(1) = \mathbf{1}$, $\mathbf{U}(1)\mathbf{1} = 1$ and $\lambda_1'(1) = \lambda$, produces the normalisation condition

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2} f_1(z)\bigg|_{z=1} = -2\lambda c(1-\rho)\ .$$

3.3.2. Mean value of the system content

The mean value of the system content is found by deriving (28) three times at $z = 1$ for $i = 1$ and taking into account that $\mathbf{r}_1(1) = \mathbf{1}$, $\lambda_1^{'}(1) = \lambda$ and $U^{'}(1)\mathbf{1} = \mathrm{E}\,[U]$, leading to

$$\mathrm{E}\,[U] = \frac{1}{6\lambda c(\rho-1)}\Bigg[\ \frac{\mathrm{d}^3}{\mathrm{d}z^3} f_1(z)\bigg|_{z=1} + 6\lambda c(1-\rho)\mathbf{U}(1)\mathbf{r}_1'(1) \\ - 3\lambda^3 T_c''(1) - 6\lambda_1''(1)T_c'(1)\lambda + 3\lambda_1''(1)c + 3\lambda c^2 - 3\lambda c\Bigg]\ .$$

Note hereby that $\mathbf{U}(1) = \boldsymbol{\pi}$, the steady-state probability vector of the background state.

3.3.3. Tail probabilities of the system content

In this paper, we assume an infinite buffer capacity. Nevertheless, buffers have a finite capacity, which causes customers to get rejected if they arrive

when the buffer is full. As a result, the loss ratio - defined as the fraction of customers that are rejected - is an important performance measure. In addition, the buffer capacity, say b, is typically designed such that loss is a rare event, and the tail probability $\Pr[U > b]$ in the corresponding infinite capacity model then provides a good approximation for the loss ratio [17]. Therefore, we calculate the tail probabilities and we achieve this by applying the dominant singularity approximation [17]. First, we divide (28) by $[1 - \lambda_i(z)]\,[z^c - T_c(\lambda_i(z))]$ and we sum both sides of this equation over i from 1 to N. On account of the distributive property of matrices, $\mathbf{R}(z)\mathbf{1} = \mathbf{1}$ and $\mathbf{U}(z)\mathbf{1} = U(z)$, we find

$$\begin{aligned} U(z) &= \sum_{i=1}^{N} \frac{f_i(z)}{[1 - \lambda_i(z)][z^c - T_c(\lambda_i(z))]} \\ &= \sum_{i=1}^{N} \frac{f_i(z) \prod_{k=1,k\neq i}^{N} [1 - \lambda_k(z)][z^c - T_c(\lambda_k(z))]}{\prod_{j=1}^{N} [1 - \lambda_j(z)][z^c - T_c(\lambda_j(z))]} \ . \end{aligned}$$

First, recall that $|\lambda_1(z)| > |\lambda_j(z)|$ for all $2 \leq j \leq N$ and for all $z \in]1, \Re[$. Second, note that when $\tilde{z} \in]1, \Re[$ is a zero of $[1 - \lambda_i(z)]$, $f_i(z)$ also vanishes at $z = \tilde{z}$ (observe equation (20)). Next, using similar arguments as in [1], where a continuous Markovian arrival process is considered, one can prove that $\lambda_1(z)$ is a strictly increasing and convex function for $z \in]1, \Re[$. Hence, $z^c - T_c(\lambda_1(z))$ will have a unique solution in this region if $\lim_{z\uparrow\Re} T_c(\lambda_1(z))/z^c > 1$, a requirement that we assume to be satisfied from now on. As a result, the dominant singularity of $U(z)$, z^*, is the zero from $z^c - T_c(\lambda_1(z))$ in $]1, \Re[$. Analogously as in the case of independent arrivals, we thus find the following approximate expression for the tail of U:

$$\Pr[U > n] \sim -\frac{(z^*)^{-(1+n)}}{z^* - 1} \frac{f_1(z^*)}{[1 - \lambda_1(z^*)][c(z^*)^{c-1} - T_c'(\lambda_1(z^*))\lambda_1'(z^*)]} \ , \tag{29}$$

where $f(n) \sim g(n)$ means that $\lim_{n\to\infty} f(n)/g(n) = 1$. Hence, expression (29) allows us to calculate the probability that the system content exceeds a threshold n, for large enough values of n. In practice, buffer dimensioning is an important assignment. For instance, one has to dimension the buffer so that the loss ratio is smaller than 10^{-6}. We can translate this problem to our setting: determine b such that $\Pr[U > b] \leq 10^{-6}$. Taking the Neperian logarithm of this equation and on account of (29), we obtain:

$$b \geq \frac{6 \ln 10 + \ln K}{\ln z^*} - 1 \ ,$$

with

$$K = -\frac{1}{z^* - 1} \frac{f_1(z^*)}{[1 - \lambda_1(z^*)][c(z^*)^{c-1} - T_c'(\lambda_1(z^*))\lambda_1'(z^*)]} \ .$$

Hence, the smallest buffer capacity $b \in \mathbb{N}$ that insures a loss ratio not larger than 10^{-6} is equal to

$$b = \left\lceil \frac{6 \ln 10 + \ln K}{\ln z^*} \right\rceil - 1 \ .$$

3.3.4. Probability that the server processes during a random slot

To conclude this section, we provide a performance measure that requires no manipulation of generating functions: the probability that the server processes

during a random slot. This probability ensues almost immediately from the definition of $\mathbf{q}_0(n)$:

$$\Pr[\text{server processes}] = 1 - \sum_{n=0}^{l-1} \mathbf{q}_0(n)\mathbf{1} \ .$$

4. Influence of correlation

In this section, we evaluate the influence of combining correlation in the arrival process and batch service on the behaviour of the system. To this end, we consider a numerical example whereby the number of background states N equals 2. We denote the probability that if the background state is i during a slot, the background state remains i during the next slot by $p_i \triangleq [\mathbf{A}(1)]_{ii}$, $i = 1, 2$ and we assume that $p_1 = p_2$. In view of the above assumptions, we define the coefficient of correlation γ between the states of two consecutive slots as

$$\gamma \triangleq \lim_{k\to\infty} \frac{\mathrm{E}[\tau_k\tau_{k+1}] - \mathrm{E}[\tau_k]\,\mathrm{E}[\tau_{k+1}]}{(\mathrm{Var}[\tau_k]\,\mathrm{Var}[\tau_{k+1}])^{1/2}} = 2p_1 - 1 \ .$$

We also assume that no customers arrive when the background state equals 1 and that the number of arrivals in the other case is geometrically distributed, i.e. $A_1(z) = 1$ and $A_2(z) = 1/(1 + 2\lambda - 2\lambda z)$. We further consider a server of capacity 10 ($c = 10$). The service times are geometrically distributed with the mean length being dependent on the number of customers in the served batch. More specifically, the average time to serve a batch of j customers is equal to $3 + j0.2$. Finally, the probability β_n that the server initiates a service when n customers are present ($n < l$) equals n/l.

In Fig.1, the mean system content $\mathrm{E}[U]$ is depicted versus the load ρ for several values of the correlation coefficient γ. It is assumed that the service threshold l equals 5. Fig.1 learns us that positive correlation ($\gamma > 0$) leads to a significant larger $\mathrm{E}[U]$ as compared to the independent case ($\gamma = 0$). Hence, disregarding positive correlation can lead to a severe underrating of the mean system content. Fig.1 also exhibits that ignoring negative correlation leads to some overestimation of $\mathrm{E}[U]$. We further perceive that these observations manifest themselves more as ρ increases. These conclusions are similar to those in multiserver systems with correlated arrivals (see e.g. [16], [36]).

Fig.2 shows the tail probabilities $\Pr[U > n]$ versus n in the case that the load ρ equals 0.6 and the service threshold l being 5. We perceive that positive correlation leads to much larger probabilities whereas negative correlation causes some smaller probabilities, which confirms the results of Fig.1.

When we take a look at the buffer capacity required to ensure that the loss ratio is smaller than 10^{-6} (Fig.3), we come to similar conclusions. Hence, we can state that correlation potentially has a huge impact on the system content.

Next, we investigate the server efficiency. Therefore, the filling degree - defined as $\mathrm{E}\left[\tilde{S}\right]/c$, the mean number of customers in a served batch divided by

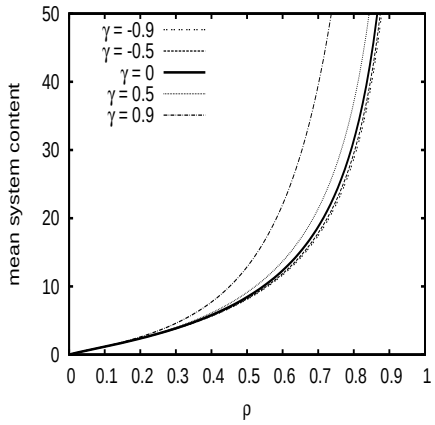


Figure 1: mean system content $\mathrm{E}[U]$ versus the load ρ for several values of the correlation coefficient γ; $l = 5$, $c = 10$, T_j geometrically distributed, $\mathrm{E}[T_j] = 3 + 0.2j$, $\beta_n = n/l$, $0 \leq n \leq l-1$

the server capacity, or, in other words, the ratio of the actual mean number of customers versus the maximum number of customers in a served batch - and the probability p that the server processes during a random slot are depicted versus the load in Fig.4. We observe that positive correlation leads to a larger filling degree and a smaller serving probability, whereas the opposite holds (in a lesser degree) for negative correlation. Hence, positive correlation leads to a more efficient usage of the server. Indeed, in case of positive correlation, long periods exist during which the server is idle because no customers arrive. On the other hand, when customers arrive, this is likely to happen during many contiguous slots, so that the server then serves more customers.

As determining the optimal service threshold (we define it as the one that minimizes $\mathrm{E}[U]$) is of the utmost importance in batch-service systems, we study whether correlation affects this optimum. For this purpose, the optimal threshold is shown versus ρ in Fig.5, for several values of γ. We perceive that the larger the correlation coefficient, the faster the optimum of l increases. Indeed, when, in the independent case, it becomes advantageous to postpone service until more customers have arrived, it can be beneficial in the correlated case to wait until even more customers have arrived, because when customers arrive it is very likely that other customers arrive in the subsequent slots.

We now investigate the impact of adopting the optimal threshold of the independent case in the correlated system. Therefore, we define the relative error as

$$\frac{\mathrm{E}[U]_{\tilde{l}_{\mathrm{opt}}} - \mathrm{E}[U]_{l_{\mathrm{opt}}}}{\left(\mathrm{E}[U]_{l_{\mathrm{opt}}} + \mathrm{E}[U]_{\tilde{l}_{\mathrm{opt}}}\right)/2},$$

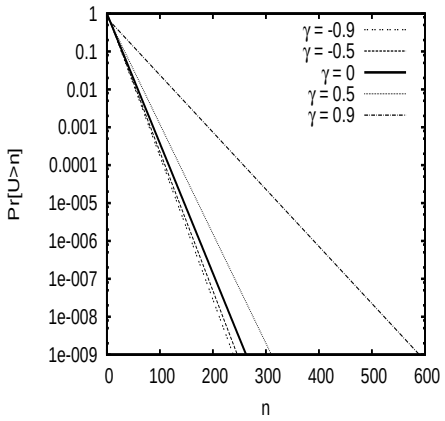


Figure 2: $\Pr[U > n]$ versus n for several values of the correlation coefficient γ; $\rho = 0.6$, $l = 5$, $c = 10$, T_j geometrically distributed, $\mathrm{E}[T_j] = 3 + 0.2j$, $\beta_n = n/l$, $0 \leq n \leq l-1$

with $\mathrm{E}[U]_{l_{\mathrm{opt}}}$ the mean system content in the correlated case when the optimal service threshold is adopted and $\mathrm{E}[U]_{\tilde{l}_{\mathrm{opt}}}$ the mean system content in the correlated system when the optimal threshold of the corresponding independent system is adopted. In Fig.6, the relative errors are depicted for the example in Fig.5. We observe that even when the optimal service threshold is different, the relative error is rather small. In view of this, the existing results of the corresponding independent system can be used to determine a near-optimal service threshold. Adopting this near-optimal threshold has only a marginal impact on the mean system content.

Before closing this section, we evaluate the impact of the probabilities β_n on the mean customer delay (equal to $\mathrm{E}[U]/\lambda$ due to Little's law). In Fig.7, the mean delay is depicted versus the load both for the case $\beta_n = n/l$ and $\beta_n = 0$ (the latter case was considered in our paper [27]). The left pane of the figure corresponds to $\gamma = 0.9$ and the right pane represents $\gamma = -0.9$. We observe that for small loads the mean delay tends to infinity when $\beta_n = 0$, whereas it is finite when $\beta_n = n/l$. When the load becomes larger, the difference between both policies diminishes. We can thus conclude that the inclusion of the β_n's in our model is a good mechanism to avoid excessive delays due to the service threshold, especially for small values of ρ.

Remark 3. *We have considered one example for the distribution of the number of arrivals: no arrivals occur during background state 1 and the arrivals are generated according to a geometric distribution during the second state. We however also have examined a large set of other examples (which we do not add to the paper due to page limitations). For instance, we have combined 0 arrivals in the first state ($A_1(z) = 1$) with other commonly adopted arrival distributions: Poisson ($A_2(z) = e^{2\lambda(z-1)}$), binomial ($A_2(z) = (1 - 2\lambda/m + 2\lambda z/m)^m$) for*

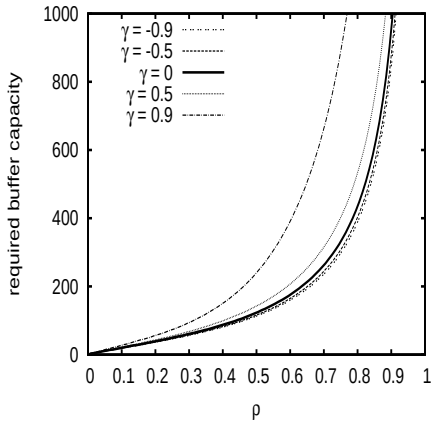
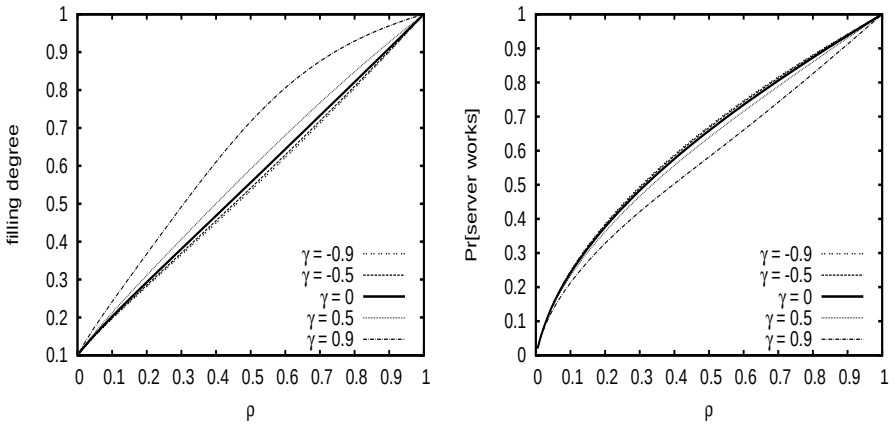


Figure 3: required buffer capacity versus the load ρ for several values of the correlation coefficient γ; $l = 5$, $c = 10$, T_j geometrically distributed, $\mathrm{E}[T_j] = 3 + 0.2j$, $\beta_n = n/l$, $0 \leq n \leq l-1$



(a) filling degree

(b) probability that the server processes during a random slot

Figure 4: server efficiency versus the load ρ for several values of the correlation coefficient γ; $l = 5$, $c = 10$, T_j geometrically distributed, $\mathrm{E}[T_j] = 3 + 0.2j$, $\beta_n = n/l$, $0 \leq n \leq l-1$

various values of m and batch Bernoulli ($A_2(z) = 1 - 2\lambda/m + 2\lambda z^m/m$) for various values of m. In addition, we have studied examples where arrivals can occur in both states, whereby the arrival rate in one of the states is m times larger than in the other state, for various values of m. We have considered all possible combinations of the above mentioned arrival distributions. In all these cases, the behaviour of the performance indices is quite similar as described in Figs 1-7, and the same conclusions could be drawn.

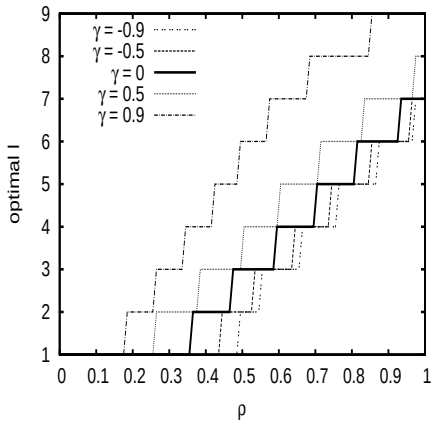


Figure 5: optimal service threshold l versus the load ρ for several values of the correlation coefficient γ; $c = 10$, T_j geometrically distributed, $\mathrm{E}[T_j] = 3 + 0.2j$, $\beta_n = n/l$, $0 \leq n \leq l-1$

Remark 4. *In the examples, we have noticed that positive correlation has a larger effect on the behaviour of the system than negative correlation. We can explain this intuitively. Therefore, let us call state 1 the "inactive" (no arrivals) and state 2 the "active" state. When being in the inactive state, the temporary load (the load during several consecutive slots of the same state) becomes very small. This effect causes a temporarily smaller system content. When being in the active state, the temporary load becomes very large, which causes a temporarily larger system content. Due to the queueing effect (i.e. the system content increases exponentially for large load), the effect of being in active state is larger than the effect of being in inactive state. Of course, the longer active periods last (i.e. the larger the correlation), the more this effect plays a role.*

5. Conclusions

In this paper, we have studied a discrete-time D-BMAP/$G^{l,c}$/1 queueing model, that includes service times that are dependent on the number of served customers and a mechanism to avoid that customers suffer excessive waiting times due to postponing service until more customers have arrived. We have deduced various useful performance measures related to the buffer content and we have investigated the impact of the traffic parameters on the system performance through some numerical examples. We have shown that correlation has only a small impact on the service threshold that minimizes the mean system content, and consequently, that the existing results of the corresponding independent arrivals system can be applied to determine a near-optimal service threshold policy, which is an important finding for practitioners. On the other hand, we have demonstrated that for other purposes, such as performance evaluation and buffer management, correlation in the arrival process cannot be

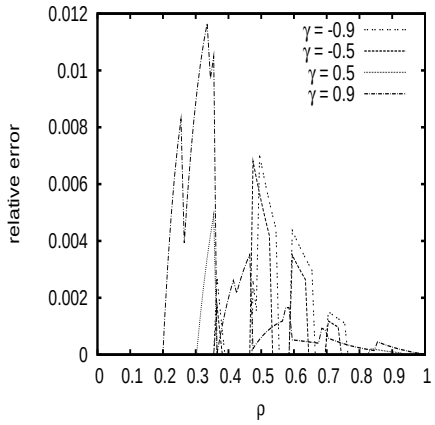
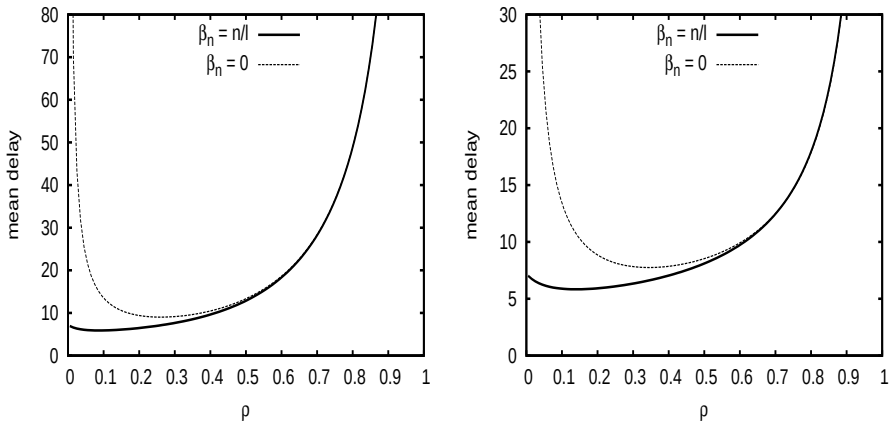


Figure 6: relative error versus the load ρ for several values of the correlation coefficient γ; $c = 10$, T_j geometrically distributed, $\mathrm{E}[T_j] = 3 + 0.2j$, $\beta_n = n/l$, $0 \leq n \leq l-1$



(a) $\gamma = 0.9$ (b) $\gamma = -0.9$

Figure 7: influence of β_n to the mean customer delay; $l = 5$, $c = 10$, T_j geometrically distributed, $\mathrm{E}[T_j] = 3 + 0.2j$

ignored, a conclusion that runs along the same lines as is found for queueing models without batch service.

References

[1] Abate, J., Choudhry, G.L., Whitt, W. (1994), Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models*, 10(1), 99–143.

[2] Abolnikov, L., Dukhovny, A. (2003), Optimization in HIV screening prob-

lems. *Journal of Applied Mathematics and Stochastic Analysis*, 16(4), 361–374.

[3] Adan, I.J.B.F., Resing, J.A.C. (2000), Multi-server batch-service systems. *Statistica Neerlandica*, 54(2), 202–220.

[4] Adan, I.J.B.F., Kulkarni, V.G. (2003), Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Systems*, 45, 113–134.

[5] Alfa, A.S. (1995), A discrete MAP/PH/1 queue with vacations and exhaustive time-limited service. *Operations Research Letters*, 18, 31-40.

[6] Arumuganathan, R., Jeyakumar, S. (2005), Steady state analysis of a bulk queue with multiple vacations, setup times with N-policy and closedown times. *Applied Mathematical Modelling*, 29(10), 972–986.

[7] Baba, Y. (1996), A bulk service GI/M/1 queue with service rates depending on service batch size. *Journal of the Operations Research Society of Japan*, 39(1), 25–34.

[8] Baiocchi, A. (1994), Analysis of the loss probability in MAP/G/1/K-queues Part I: Asymptotic theory. *Stochastic Models*, 10, 867-893.

[9] Banerjee, A., Gupta, U.C. (2012), Reducing congestion in bulk-service finite-buffer queueing system using batch-size-dependent service. *Performance Evaluation*, 69, 53-70.

[10] Banik, A.D., Gupta, U.C., Pathak, S.S. (2006), BMAP/G/1/N queue with vacations and limited service discipline. *Applied Mathematics and Computations*, 180, 707–721.

[11] Banik, A.D. (2009), Queueing analysis and optimal control of BMAP/$G^{(a,b)}$/1/N and BMAP/$MSP^{(a,b)}$/1/N systems. *Computers& Industrial Engineering*, 57, 748–761.

[12] Bar-Lev, S.K., Parlar, M., Perry, D., Stadje, W., Van der Duyn Schouten, F.A. (2007), Applications of bulk service queues to group testing models with incomplete identification. *European Journal of Operational Research*, 183, 226–237.

[13] Bellalta, B. (2009), A queueing model for the non-continuous frame assembly scheme in finite buffers. *Proceedings of the 16th international conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2009)*, Madrid, June 9-12, 219–233.

[14] Blondia, C. (1991), Finite capacity vacation models with non-renewal input. *Journal of Applied Probability*, 28, 174-197.

[15] Blondia, C. (1993), A discrete-time batch Markovian arrival process as B-ISDN traffic model. *Belgian Journal of Operations Research, Statistics and Computer Science*, 32, 3–23.

[16] Bruneel, H. (1988), Queueing behavior of statistical multiplexers with correlated inputs. *IEEE Transactions on Communications*, 36(12), 1339–1341.

[17] Bruneel, H., Steyaert, B., Desmet, E., Petit, G.H. (1992), Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. *European Journal of Operational Research*, 76, 563–572.

[18] Chakravarthy, S. (1992), A finite-capacity GI/PH/1 queue with group services. *Naval Research Logistics*, 39(3), 345–357.

[19] Chang, S.H., Choi, D.W. (2005), Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations. *Computers and Operations Research*, 32, 2213–2234.

[20] Chang, S.H., Takine, T. (2005), Factorization and stochastic decomposition properties in bulk queues with generalized vacations. *Queueing Systems*, 50, 165–183.

[21] Chaudhry, M.L. (1965), Correlated queueing. *CORS Journal*, 3, 142-151.

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

[23] Chaudhry, M.L., Gupta, U.C. (1999), Modelling and analysis of M/$G^{a,b}$/1/N queue A simple alternative approach. *Queueing Systems*, 31, 95-100.

[24] Chaudhry, M.L., Gupta, U.C. (2003), Analysis of a finite-buffer bulk-service queue with discrete-Markovian arrival process: D-MAP/$G^{a,b}$/1/N. *Naval Research Logistics*, 50(4), 345–363.

[25] Chaudhry, M.L., Gupta, U.C. (2003), Queue length distributions at various epochs in discrete-time D-MAP/G/1/N queue and their numerical evaluation. *International Journal of Information and Management Sciences*, 14(3), 67–83.

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

[27] Claeys, D., Walraevens, J., Laevens, K., Steyaert, B., Bruneel, H. (2010), A batch-service queueing model with a discrete batch Markovian arrival process. *Proceedings of the 17th international conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2010)*, Cardiff, June 14-16, 1–13.

[28] Claeys, D., Laevens, K., Walraevens, J., Bruneel, H. (2010), Complete characterisation of the customer delay in a queueing system with batch arrivals and batch service. *Mathematical Methods of Operations Research*, 72(1), 1–23.

[29] Claeys, D., Walraevens, J., Laevens, K., Bruneel, H. (2010), Delay analysis of two batch-service queueing models with batch arrivals: $Geo^X/Geo^c/1$. *4OR*, 8(3), 255–269.

[30] Claeys, D., Walraevens, J., Laevens, K., Bruneel, H. (2010), A queueing model for general group screening policies and dynamic item arrivals. *European Journal of Operational Research*, 207(2), 827–835.

[31] Claeys, D., Walraevens, J., Laevens, K., Bruneel, H. (2011), Analysis of threshold-based batch-service queueing systems with batch arrivals and general service times. *Performance Evaluation*, 68, 528–549.

[32] De Turck, K., De Vuyst, S., Fiems, D., Wittevrongel, S. (2008), Performance analysis of the IEEE 802.16e sleep mode for correlated downlink traffic. *Telecommunication Systems*, 39, 145–156.

[33] Ferrandiz, J.M. (1993), The BMAP/GI/1 queue with server set-up times and server vacations. *Advances in Applied Probability*, 25(1), 235-254.

[34] Frigui, I., Alfa, A.S., Xu, X. (1997), Algorithms for computing waiting time distributions under different queue disciplines for the D-BMAP/PH/1. *Naval Research Logistics*, 44, 559-576.

[35] Gail, H.R., Hantler, S.L., Taylor, B.A. (1996), Spectral analysis of M/G/1 and G/M/1 type Markov chains. *Advances in Applied Probability*, 28(1), 114–165.

[36] Gao, P., Wittevrongel, S., Bruneel, H. (2004), On the behavior of multi-server buffers with geometric service times and bursty input traffic. *IEICE Transactions on Communications*, 12, 3576–3583.

[37] Goswami, V., Mohanty, J.R., Samanta, S.K. (2006), Discrete-time bulk-service queues with accessible and non-accessible batches. *Applied Mathematics and Computation*, 182(1), 898–906.

[38] Gupta, U.C., Laxmi, P.V. (2001), Analysis of the MAP/$G^{a,b}$/1/N queue. *Queueing Systems*, 38, 109–124.

[39] Gupta, U.C., Goswami, V. (2002), Performance analysis of finite buffer discrete-time queue with bulk service. *Computers and Operations Research*, 29, 1331–1341.

[40] Gupta, U.C., Sikdar, K. (2004), A finite capacity bulk service queue with single vacation and Markovian arrival process. *Journal of Applied Mathematics and Stochastic Analysis*, 2004(4), 337-357.

[41] Gupta, U.C., Samanta, S.K., Sharma, R.K., Chaudhry, M.L. (2007), Discrete-time single-server finite buffer queues under discrete Markovian arrival process with vacations. *Performance Evaluation*, 64, 1–19.

[42] Herrmann, C. (2001), The complete analysis of the discrete time finite DBMAP/G/1/N queue. *Performance Evaluation*, 43, 95-121.

[43] Janssen, A.J.E.M., van Leeuwaarden, J.S.H. (2005), Analytic computation schemes for the discrete-time bulk service queue. *Queueing Systems*, 50, 141–163.

[44] Kasahara, S., Takine, T., Takahashi, Y., Hasegawa, T. (1996), MAP/G/1 queues under N-policy with and without vacations. *Journal of Operational Research Society Japan*, 39(2), 188-212.

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

[46] Kim, N.K., Chaudhry, M.L. (2006), Equivalences of batch-service queues and multi-server queues and their complete simple solutions in terms of roots. *Stochastic Analysis and Applications*, 24(4), 753–766.

[47] Kim, B., Kim, J. (2010), Queue size distribution in a discrete-time D-BMAP/G/1 retrial queue. *Computers and Operations Research*, 37(7), 1220–1227.

[48] Lee, H.W., Ahn, B.Y., Park, N.I. (2001), Decompositions of the queue length distributions in the MAP/G/1 queue under multiple and single vacations with N-policy. *Stochastic Models*, 17(2), 157-190.

[49] Lee, H.W., Moon, J.M., Kim, B.K., Park, J.G., Lee, S.W. (2005), A simple eigenvalue method for low-order D-BMAP/G/1 queues. *Applied Mathematical Modelling*, 29, 277-288.

[50] Li, S.Q. (1991), A general solution technique for discrete queueing analysis of multimedia traffic on ATM. *IEEE Transactions on Communications*, 39(7), 1115-1132.

[51] Lu, K., Wu, D., Fang, Y., Qiu, R.C. (2005), Performance analysis of a burst-frame-based MAC Protocol for ultra-wideband ad hoc networks. *Proceedings of the IEEE International Conference on Communications (ICC 2005)*, Seoul, May 16-20, Vol.5, 2937–2941.

[52] Lucantoni, D.M., Meier-Hellstern, K.S., Neuts, M.F. (1990), A single-server queue with server vacations and a class of non-renewal process. *Advances in Applied Probability*, 22, 676-705.

[53] Matendo, S.K. (1993), A single-server queue with server vacations and a batch Markovian arrival process. *Cahiers Centre Etudes Rech. Oper.*, 35(1-2), 87-114.

[54] Matendo, S.K. (1994), Some performance measures for vacation models with a batch Markovian arrival process. *Journal of Applied Mathematics and Stochastic Analysis*, 7(2), 111-124.

[55] Masuyama, H. (2003), *Studies on algorithmic analysis of queues with batch Markovian arrival streams*. Phd dissertation, Kyoto, Japan.

[56] Meyer, C. (2000), *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia.

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

[58] Niu, Z., Takahashi, Y. (1999), A finite-capacity queue with exhaustive vacation/close-down/setup times and Markovian arrival processes. *Queueing Systems Theory and Applications*, 31(1-2), 1-23.

[59] Niu, Z., Shu, T., Takahashi, Y. (2003), A vacation queue with setup and close-down times and batch Markovian arrival processes. *Performance Evaluation*, 54, 225–248.

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

[61] Qiao, C.M., Yoo, M.S. (1999), Optical burst switching (OBS) - a new paradigm for an optical Internet. *Journal of High Speed Networks*, 8(1), 69–84.

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

[63] Samanta, S.K., Gupta, U.C., Sharma, R.K. (2007), Analyzing discrete-time D-BMAP/G/1/N queue with single and multiple vacations. *European Journal of Operational Research*, 182(1), 321–339.

[64] Schellhaas, H. (1994), Single server queues with a batch Markovian arrival process and server vacations. *OR Spektrum*, 15, 189-196.

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

[66] Sikdar, K., Gupta, U.C. (2005), The queue length distributions in the finite buffer bulk-service MAP/G/1 queue with multiple vacations. *Sociedad de Estadística e Investigación Operativa Top*, 13(1), 75–103.

[67] Steyaert, B., Walraevens, J., Fiems, D., Bruneel, H. (2011), The NxD-BMAP/G/1 queueing model: queue contents and delay analysis. *Mathematical Problems in Engineering*, 2011, 1–31.

[68] Van Houdt, B., Lenin, R.B., Blondia, C. (2003), Delay distribution of (im)patient customers in a discrete time D-MAP/PH/1 queue with age-dependent service times. *Queueing Systems*, 45, 59–73.

[69] Van Velthoven, J., Van Houdt, B., Blondia, C. (2005), Response time distribution in a D-MAP/PH/1 queue with general customer impatience. *Stochastic Models*, 21, 745–765.

[70] Yi, X.W., Kim, N.K., Yoon, B.K., Chae, K.C. (2007), Analysis of the queue-length distribution for the discrete-time batch-service $Geo^X|G^{a,Y}|1|K$ queue. *European Journal of Operational Research*, 181, 787–792.

[71] Zhang, Z. (1991), Analysis of a discrete-time queue with integrated bursty inputs in ATM networks. *International Journal on Digital and Analog Communication Systems*, 4, 191–203.

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