Noname manuscript No.
(will be inserted by the editor)

Delay analysis of a two-class batch-service queue with class-dependent variable server capacity

Jens Baetens · Bart Steyaert · Dieter Claeys · Herwig Bruneel

Received: date / Accepted: date

Abstract In this paper, we analyse the delay of a random customer in a two-class batch-service queueing model with variable server capacity, where all customers are accommodated in a common single-server first-come-first-served queue. The server can only process customers that belong to the same class, so that the size of a batch is determined by the length of a sequence of same-class customers. This type of batch server can be found in telecommunications systems and production environments. We first determine the steady state partial probability generating function of the queue occupancy at customer arrival epochs. Using a spectral decomposition technique, we obtain the steady state probability generating function of the delay of a random customer. We also show that the distribution of the delay of a random customer corresponds to a phase-type distribution. Finally, some numerical examples are given that provide further insight in the impact of asymmetry and variance in the arrival process on the number of customers in the system and the delay of a random customer.

Keywords Delay · Discrete-Time · Batch Service · Two-Class · Variable Capacity

Dieter Claeys is a Postdoctoral Fellow with the Research Foundation Flanders (FWO-Vlaanderen), Belgium. Part of the research has been funded by the Interuniversity Attraction Poles Programma initiated by the Belgian Science Policy Office.

Jens Baetens
Ghent University, Dept. of Telecommunications and Information Processing, SMACS Research Group
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
E-mail: jens.baetens@ugent.be

Bart Steyaert · Dieter Claeys · Herwig Bruneel
Ghent University, Dept. of Telecommunications and Information Processing, SMACS Research Group
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

Dieter Claeys
Ghent University, Dept. of Industrial Systems Engineering and Product Design
Technologiepark 903
Zwijnaarde, Belgium

1 Introduction

In many applications, a single server can process several customers simultaneously in a group or batch. Such a batch-server queueing system differs from multi-server queueing systems in that a new customer cannot join a batch whose service has started, even if it is not a full batch, while in multi-server queueing systems a new customer is served as soon as there is at least one server available. Batch-service systems are common in practice, for example in production environments where a machine can heat or paint multiple components at the same time (Weng and Leachman (1993)) or transportation systems where a number of customers with the same destination are transported simultaneously (Lee and Kim (1994)). Due to the wide area of applications, batch-service queueing models have been studied thoroughly, for instance by Arumuganathan and Jeyakumar (2005), Banerjee and Gupta (2012), Banerjee et al (2015, 2014), Chang and Takine (2005), Chaudhry and Templeton (1983), Claeys et al (2012, 2013a,b, 2011), Goswami et al (2006), and Janssen and van Leeuwaarden (2005).

A common assumption in the above mentioned papers is that the service capacity is constant. However, in practice, capacity can be variable and stochastic, a feature that has been incorporated in only a few papers. The system content at various observation epochs in the $Geo/G^Y/1/N+B$ model, where Y denotes the stochastic capacity of the server, which is upper-bounded by B, and N is the maximum queue capacity, has been analysed by Chaudhry and Chang (2004). Server vacations have been included in this model by Chang and Choi (2005). In the context of vehicle dispatching strategies, Powell and Humblet studied various batch-service queues with stochastic capacity and several service (dispatching) policies (Powell and Humblet, 1986). They obtained the distribution of the queue length at departure epochs. Yi et al. further extended the work of Chang and Choi by including the general bulk service rule (Yi et al, 2007). More specifically, the distribution of the system content at various epochs has been established for the $Geo/G^{a,Y}/1/K$ queue. Germs and Van Foreest investigated the loss probabilities for the continuous-time $M^X/G^Y/1/K+B$ queue (Germs and Foreest, 2010). Furthermore, Pradhan et al (2015) obtained closed-form expressions for the queue-length distribution at departure epochs for the discrete-time $M/G_r^Y/1$ queue where the service process depends on the batch size. A similar feature in the models of Chaudhry and Chang, Chang and Choi, Germs and Van Foreest, Powell and Humblet, Pradhan et al., and Yi et al. is that the capacity of a batch is independent of the queue length and of the capacities of the previous batches. Germs and Foreest (2013) have recently developed an algorithmic method for the performance evaluation of the continuous-time $M(n)^{X(n)}/G(n)^{Y(n)}/1/K+B$ queue. In that model, both the arrival rate and service process (the service times as well as the capacities) depend on the queue size.

Another feature of the above models is that customers are indistinguishable, i.e., they all are of the same type. Although in many types of queueing systems several customer classes are included to account for customer differentiation, only a few papers on batch service consider multiple customer classes. Reddy et al (1993) study a multi-class batch-service queueing system with Poisson arrivals and a priority scheduling discipline, in the context of an industrial repair shop where the most critical machines are repaired first. Boxma et al (2008) study a polling system with Poisson arrivals and batch service. In this case, each customer class has a dedicated queue and the server visits the different queues in a cyclic manner. Boxma et al. focus on the influence of a number of different gating policies on the performance. Dorsman et al (2012) study a polling system with a renewal arrival process and batch service, where the batches are created by accumulation stations before they are added to a queue. Such a system can be used when a single server processes multiple product types with batching constraints. Dorsman et al. focus on optimizing the batch sizes of each class. A single-server station processing jobs in fixed-size batches belonging to multiple product classes is considered by Bitran and Tirupati (1989). The

interarrival times within each product class and the batch service times are assumed to have general, independent and identical distributions. A batch is constructed by aggregating the first N jobs in the queue. An approximation for the mean number of jobs, as well as some results on the output process, are derived. A refinement for some of these approximations, albeit for Poisson arrivals, can be found in Wu et al (2011). Fowler et al (2002) study a multiproduct $G/G/c$ model with batch-processing. For product class k, the interarrival times are generally distributed. Batches of size B_k are formed first, which are then added to a FCFS multiserver queue. As a result, steady-state approximations for the cycle time and work-in-progress are derived. In Huang et al (2001), class-dependent Poisson arrivals and exponential batch processing times are considered, and batches of different classes with a common maximum batch size are formed up front as well, and are then offered to a multiserver system. This results in an approximation for the average queue length. The previously mentioned papers on polling systems and priority queueing use a unique queue for each type of customer, while in this paper customers of both classes are accommodated in a common queue. We chose to use a single queue because it is not always feasible to implement multiple queues due to, for instance, lack of space in manufacturing environments or remaining memory.

In this paper, we analyse the delay of a random customer in a two-class discrete-time batch-service queueing model, with a batch size that is determined by the length of the sequence of same-class customers at the head of the queue. When the single server becomes available, it will simultaneously process the customer at the head of the queue, and all successive customers that are of the same class as the head customer. This, for instance, means that if the first customer is of class A, all of the following class A customers are also grouped in the batch that will be taken into service, until the next customer is of class B or no more customers are present, whichever occurs first. The system occupancy at random slot boundaries has been studied previously in our conference paper Baetens et al (2016). To the best of our knowledge, no other papers in the literature have studied the delay in a system that combines variable capacity batch service and multiple customer classes. Many papers on batch-service queueing models either do not study the delay, or only give the mean delay by using Little's law. Of the previously mentioned papers, only the papers of Claeys et al (2012, 2011) and Dorsman et al (2012) cover an extensive analysis of the delay of a random customer.

We would also like to highlight that the above-mentioned papers on batch service with multiple classes are continuous-time models, that assume a single customer arrival per arrival instant, while we consider the more general case of a generally distributed batch arrival process in a discrete-time setting, which adds an extra layer of complexity to the analysis. Also, in these contributions a distinct arrival process is defined for each class, whereas we define a single batch arrival process for the aggregated number of arrivals. The aggregated numbers of arrivals during consecutive slots are modelled as a sequence of independent and identically distributed random variables.

The queueing system studied in this paper can model, for example, a postal sorting center (Willems (2014)). Letters (the customers) heading for different destination areas (the classes) arrive in random order at the center and are put on a conveyor. At the end of the conveyor, a sorter (the single server) sorts the letters according to destination area. The sorter can simultaneously pick consecutive letters with the same destination area and put them in the corresponding box. The picking time of letters is only slightly sensitive to the number of letters picked: the basic motions involved, called therbligs (Freivalds and Niebel (2014)), are reaching for the letters, grasping them, searching for the corresponding box, and moving the letters to that box and releasing them. Only grasping is slightly dependent on the number of letters picked, which is only a small part of the total picking time.

The paper is structured as follows. In Section 2 we describe the discrete-time two-class queueing model with batch service in detail. This system consists of a single First-Come-First-Served (FCFS) queue of infinite size, and a single batch server with a variable capacity. In Section 3 we first derive a closed-form expression for the steady-state partial probability generating function (pgf) of the delay of a random customer when there are n customers in the queue at the arrival time of the random customer. Using the pgf of the system occupancy at customer arrival time instants, the pgf of the delay of a random customer can be calculated. Some numerical experiments are studied in Section 4 to illustrate the influence of the asymmetry and variance in the arrival process on the mean queue occupancy at customer arrival and mean delay of a random customer. Our conclusions are presented in Section 5.

2 Model description

2.1 Arrival and service process

Let us consider a discrete-time two-class batch-arrival queueing system with infinite queue size and a batch server whose capacity is stochastic. The classes of the customers are denominated as A and B. Arriving customers are inserted at the tail of the queue.

The aggregated numbers of customer arrivals in consecutive slots are modelled as a sequence of independent and identically distributed (i.i.d.) random variables, with common probability mass function (pmf) $e(n)$ and steady-state pgf $E(z)$. The mean aggregated number of customer arrivals per slot is denoted as $\lambda = \sum_{n=0}^{\infty} ne(n) = E'(1)$. A random customer is of class A with probability σ and of class B with probability $1-\sigma$.

When the server is or becomes available and finds a non-empty queue, a new service is initiated. The size of the batch is then determined by the number of consecutive customers at the front of the system that are of the same class. More specifically, the server starts serving a batch of n customers if and only if one of the following two cases occur:

- Exactly n customers are present and they are all of the same class.
- More than n customers are present, the n customers at the front of the queue are of the same class and the $(n+1)$-th customer is of the other class.

We define the class of a batch as the class of the customers within it. The service time of a batch is always a single slot.

2.2 Stability condition

The stability condition can easily be found by analysing the system under the condition that there are always many customers present in the queue. This means that the server will alternate between processing class A and B batches, which means that we can limit ourselves to studying 2 consecutive slots. We will only have a stable system when the average amount of work entering the system during two consecutive slots is less than the maximum expected amount of work processed during the same period. The expected amount of work processed is the sum of the expected length of a sequence of consecutive class A and B customers, which respectively follow a geometric distribution with parameter σ and $1-\sigma$. The stability condition is then given by

$$2\lambda < \frac{1}{1-\sigma} + \frac{1}{\sigma} \ . \tag{1}$$

If σ is either 0 or 1, then the stability condition is reduced to $\lambda < \infty$, i.e., the system is always stable. This is as expected, since in this case all customers are of the same class, which means that no matter how many customers arrive, the server will always take all waiting customers in a single batch. Also, if σ is equal to 0.5, than the maximum tolerable arrival rate reaches a minimum value.

3 Delay Analysis

In this section, we will calculate the steady-state pgf of the delay of a random customer. We define the delay of a random customer as the number of slots between the end of the arrival slot of the customer and the end of the slot in which the service of the batch, that the random customer belongs to, ends. The delay of a random customer is determined by the number of customers in the queue at arrival of the random customer which is the sum of the number of customers in the queue at the start of its arrival slot and the number of customers that arrive before the random customer during the same slot.

First, we calculate the partial pgf of the delay of a random customer if there are n customers in the queue before arrival of the customer by using a spectral decomposition or by modelling the delay as a phase-type distribution. In the second part, we first calculate the partial pgf's of the queue occupancy at customer arrival epochs if the first customer after arrival will be of class A or B. Finally, we use the previous results to obtain the steady-state probability generating function of the delay of a random customer.

3.1 Delay of a random customer with n customers in queue

We first define $D_{A,n}(z)$ and $D_{B,n}(z)$ as the partial steady-state pgf of the delay if there are n customers in the queue before the random customer and the first customer, that is the customer at the head of the queue, is respectively of class A and B. Let us consider the case where the first customer is of class A. If the second customer is also of class A (with probability σ), then that customer will be served in the same slot as the first. Hence the delay of the tagged customer is in this case in distribution equal to the delay of a customer who finds upon arrival $n-1$ customers in the queue with the first customer being of class A. If the second customer is of class B (with probability $1-\sigma$), then after one slot the first customer has been served and the tagged customer experiences a remaining delay which is in distribution equal to the delay of a customer who finds $n-1$ customers upon arrival with the first customer being of class B. Hence, we obtain the following expression

$$D_{A,n}(z) = \sigma D_{A,n-1}(z) + (1-\sigma) z D_{B,n-1}(z) \ .$$

Equivalently, we obtain for the case that the first customer is of class A

$$D_{B,n}(z) = (1-\sigma) D_{B,n-1}(z) + \sigma z D_{A,n-1}(z) \ .$$

If the random customer is the customer at the head of the queue, then it will be served first which means its delay will be a single slot. Summarising, we have

$$\begin{aligned} D_{A,0}(z) =& D_{B,0}(z) = z \\ \begin{bmatrix} D_{A,n}(z) \\ D_{B,n}(z) \end{bmatrix} =& \begin{bmatrix} \sigma & (1-\sigma)z \\ \sigma z & (1-\sigma) \end{bmatrix} \begin{bmatrix} D_{A,n-1}(z) \\ D_{B,n-1}(z) \end{bmatrix} \\ =& \begin{bmatrix} \sigma & (1-\sigma)z \\ \sigma z & (1-\sigma) \end{bmatrix}^n \begin{bmatrix} z \\ z \end{bmatrix} \ . \end{aligned} \tag{2}$$

We now decompose the matrix $\boldsymbol{M}(z)$, defined as

$$\boldsymbol{M}(z) := \begin{bmatrix} \sigma & (1-\sigma)z \\ \sigma z & (1-\sigma) \end{bmatrix} \quad .$$

The eigenvalues $\lambda_1(z)$ and $\lambda_2(z)$ of $\boldsymbol{M}(z)$ are given by

$$\begin{aligned} \lambda_1(z) =& \frac{1}{2} + \frac{1}{2}\sqrt{1+4\sigma(1-\sigma)(z^2-1)} \quad , \\ \lambda_2(z) =& \frac{1}{2} - \frac{1}{2}\sqrt{1+4\sigma(1-\sigma)(z^2-1)} \quad , \end{aligned} \tag{3}$$

and matrices of the right and left eigenvectors, denoted respectively by $\boldsymbol{R}(z)$ and $\boldsymbol{L}(z)$, corresponding with the matrix $\boldsymbol{M}(z)$ are

$$\begin{aligned} \boldsymbol{R}(z) =& \begin{bmatrix} \frac{(1-\sigma)z}{\lambda_1(z)-\sigma} & \frac{(1-\sigma)z}{\lambda_2(z)-\sigma} \\ 1 & 1 \end{bmatrix} =: \begin{bmatrix} r_1(z) & r_2(z) \\ 1 & 1 \end{bmatrix} \quad , \\ \boldsymbol{L}(z) =& \boldsymbol{R}^{-1}(z) = \begin{bmatrix} \frac{\sigma z}{2\lambda_1(z)-1} & \frac{\lambda_1(z)-\sigma}{2\lambda_1(z)-1} \\ \frac{\sigma z}{2\lambda_2(z)-1} & \frac{\lambda_2(z)-\sigma}{2\lambda_2(z)-1} \end{bmatrix} \quad . \end{aligned} \tag{4}$$

We also define $L_1(z)$ and $L_2(z)$ as the sums of the components of the first and second row of $\boldsymbol{L}(z)$ respectively which leads to the following equations

$$\begin{aligned} L_1(z) =& \frac{\lambda_1(z) - \sigma(1-z)}{2\lambda_1(z)-1} \quad , \\ L_2(z) =& \frac{\lambda_2(z) - \sigma(1-z)}{2\lambda_2(z)-1} \quad . \end{aligned} \tag{5}$$

Diagonalization of matrix $\boldsymbol{M}(z)$ in Eq. 2 leads to

$$\begin{bmatrix} D_{A,n}(z) \\ D_{B,n}(z) \end{bmatrix} = \boldsymbol{R}(z) \begin{bmatrix} \lambda_1(z)^n & 0 \\ 0 & \lambda_2(z)^n \end{bmatrix} \boldsymbol{L}(z) \begin{bmatrix} z \\ z \end{bmatrix} \quad . \tag{6}$$

In Appendix A, we show that the branching points, which are the points where $\lambda_1(z) = \lambda_2(z)$, can be removed so that there are no roots in the solution. With this we can also prove that $D_{A,n}$ and $D_{B,n}$ are polynomials of degree $n+1$.

A second approach to obtain the partial pgf's $D_{A,n}(z)$ and $D_{B,n}(z)$ is to analyse the delay of a random customer with n customers in the queue at arrival and the first customer is of class A or B using a discrete-time phase type distribution. More information on discrete-time phase type distributions can be found in the book of Latouche and Ramaswami (1999). The delay of a random customer corresponds to the absorption time of a discrete Markov Chain where the time spent in each state is equal to the service time of a batch. The transition probability from state i to state j is given by the probability that a batch of the corresponding class has a size of $i-j$ customers. The distribution of the delay of a random customer if the queue occupancy is equal to n at arrival of the random customer is the phase-type distribution $PH(\boldsymbol{\tau}_{A,n}, \boldsymbol{T}_n)$ or $PH(\boldsymbol{\tau}_{B,n}, \boldsymbol{T}_n)$ if the customer at the head of the queue is of class A or B, with $2(n+1)$ states. The first $n+1$ states correspond to the cases that there are $i = 0 \cdots n$ customers in the queue before the random customer and the customer at the head of the queue is of class A and the

next $n+1$ states with the cases that the customer at the head of the queue is of class B. The initial distributions of the phase-type distributions, denoted by $\boldsymbol{\tau}_{A,n}$ and $\boldsymbol{\tau}_{B,n}$, are given by

$$\begin{aligned}\boldsymbol{\tau}_{A,n} &= \big[\, \underbrace{0\cdots0}_{n}\ 1\ \underbrace{0\cdots0}_{n+1} \,\big] \ , \\ \boldsymbol{\tau}_{B,n} &= \big[\, \underbrace{0\cdots0}_{2n+1}\ 1 \,\big] \ ,\end{aligned}$$

and the transition matrix is characterized entirely by $\boldsymbol{T}_n$, a $(2n+2)\times(2n+2)$ matrix equal to

$$\boldsymbol{T}_n := \begin{bmatrix} \mathbf{0}_n & \boldsymbol{T}_{AB,n} \\ \boldsymbol{T}_{BA,n} & \mathbf{0}_n \end{bmatrix} \ ,$$

where $\mathbf{0}_n$ is a $(n+1)\times(n+1)$ zero matrix. The matrix $\boldsymbol{T}_{AB,n}$ is a strictly lower triangular matrix where the element in row i and column j is equal to the probability that the system processes a class A batch of size $i-j$, which leads to the matrix

$$\boldsymbol{T}_{AB,n} := \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ (1-\sigma) & 0 & \cdots & \cdots & \cdots & 0 \\ \sigma(1-\sigma) & (1-\sigma) & 0 & \cdots & \cdots & 0 \\ \sigma^2(1-\sigma) & \sigma(1-\sigma) & (1-\sigma) & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \sigma^{n-1}(1-\sigma) & \sigma^{n-2}(1-\sigma) & \cdots & \cdots & (1-\sigma) & 0 \end{bmatrix} \ ,$$

and the analogous matrix $\boldsymbol{T}_{BA,n}$ is given by

$$\boldsymbol{T}_{BA,n} := \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \sigma & 0 & \cdots & \cdots & \cdots & 0 \\ \sigma(1-\sigma) & \sigma & 0 & \cdots & \cdots & 0 \\ \sigma(1-\sigma)^2 & \sigma(1-\sigma) & \sigma & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \sigma(1-\sigma)^{n-1} & \sigma(1-\sigma)^{n-2} & \cdots & \cdots & \sigma & 0 \end{bmatrix} \ .$$

The partial pgf's $D_{A,n}(z)$ and $D_{B,n}(z)$ are given by

$$\begin{aligned} D_{A,n}(z) =& z\boldsymbol{\tau}_{A,n}(\boldsymbol{I} - z\boldsymbol{T}_n)^{-1}\boldsymbol{t}_n \\ D_{B,n}(z) =& z\boldsymbol{\tau}_{B,n}(\boldsymbol{I} - z\boldsymbol{T}_n)^{-1}\boldsymbol{t}_n \ , \end{aligned}$$

We obtain the column vector $\boldsymbol{t}_n$ by solving the equation

$$\boldsymbol{t}_n = (\boldsymbol{I} - \boldsymbol{T}_n)\mathbf{1} \ ,$$

where $\mathbf{1}$ is a column vector of ones.

3.2 Queue Occupancy

In order to find the pgf of the delay of a random customer, we must first find the partial pgfs of the number of customers in the queue at customer arrival and the first customer in the queue is of class A or B after arrival. For this, we will use the results from Baetens et al (2016) for the idle probability and the partial pgf's of the system occupancy at random slot boundaries if the server is processing a class A or B batch. The probability u_I that the server is idle in a random slot is given by

$$u_I = \frac{E(0)}{(1-E(0))}\left(\frac{U_A(\sigma)}{\sigma} + \frac{U_B(1-\sigma)}{1-\sigma}\right) \ , \tag{7}$$

and the partial pgf $U_A(z)$ of the system occupancy at the slot boundary of a random slot in which a class A batch is being processed, as

$$\begin{aligned} &U_A(z)\Big((z-\sigma)(z-(1-\sigma)) - \sigma(1-\sigma)E^2(z)\Big) \\ &\quad = \sigma(z-\sigma)\Big((z-(1-\sigma)) + (1-\sigma)E(z)\Big)\frac{E(z)-E(0)}{1-E(0)}\left(\frac{U_A(\sigma)}{\sigma} + \frac{U_B(1-\sigma)}{1-\sigma}\right) \\ &\qquad - \frac{\sigma(z-\sigma)zE(z)U_B(1-\sigma)}{1-\sigma} - (1-\sigma)zE^2(z)U_A(\sigma) \ , \end{aligned} \tag{8}$$

and the analogous partial pgf $U_B(z)$ if the server is processing a class B batch by

$$\begin{aligned} &U_B(z)\Big((z-\sigma)(z-(1-\sigma)) - \sigma(1-\sigma)E^2(z)\Big) \\ &\quad = (1-\sigma)(z-1+\sigma)\Big((z-\sigma) + \sigma E(z)\Big)\frac{E(z)-E(0)}{1-E(0)}\left(\frac{U_A(\sigma)}{\sigma} + \frac{U_B(1-\sigma)}{1-\sigma}\right) \\ &\qquad - \frac{(1-\sigma)(z-1+\sigma)zE(z)U_A(\sigma)}{\sigma} - \sigma zE^2(z)U_B(1-\sigma) \ . \end{aligned} \tag{9}$$

The combined pgf $U(z)$ of the system occupancy at random slot boundaries is defined as

$$U(z) := u_I + U_A(z) + U_B(z) \ .$$

The number of customers in the system at a random slot boundary is equal to the sum of the number of customers in the queue at the previous slot boundary and the number of customer arrivals during the previous slot. These numbers are respectively denoted by u_{k+1}, q_k and e_k. The resulting equation is

$$u_{k+1} = q_k + e_k \ ,$$

or in terms of the steady-state generating functions

$$U(z) = Q(z)E(z) \ , \tag{10}$$

where $Q(z)$ is the steady-state pgf of the queue occupancy at random slot boundaries, which has not been analysed in Baetens et al (2016). In order to analyse the delay of a random customer, we need the probability that the queue is empty and the partial pgf's of the queue occupancy and the customer at the head of the queue is a class A or B customer, which we will denote by q_0, $Q_A(z)$ and $Q_B(z)$. With these definitions, the steady-state pgf $Q(z)$ is given by

$$Q(z) = q_0 + Q_A(z) + Q_B(z) \ . \tag{11}$$

The queue is empty at a random slot boundary if the server is idle during the slot or if the server is processing a class A or B batch and all customers are of the same class. In case that there are i customers in the system at a random slot boundary, the probability that the length of a class A or B batch is equal to i is respectively given by σ^{i-1} and $(1-\sigma)^{i-1}$. The probabilities that the server is processing a class A or B batch and all customers are of the same class, are given by

$$\sum_{i=1}^{\infty} u_A(i)\sigma^{i-1} = \frac{U_A(\sigma)}{\sigma} \quad , \quad \sum_{i=1}^{\infty} u_B(i)(1-\sigma)^{i-1} = \frac{U_B(1-\sigma)}{1-\sigma} \quad ,$$

where $u_A(n)$ and $u_B(n)$ represent the pmf's of the system occupancy at random slot boundary if a class A or B batch is being processed. The probability q_0 that the queue is empty at random slot boundaries is then given by

$$q_0 := Q(0) = \frac{U(0)}{E(0)} = u_I + \frac{U_A(\sigma)}{\sigma} + \frac{U_B(1-\sigma)}{1-\sigma} = \frac{1}{1-E(0)}\left(\frac{U_A(\sigma)}{\sigma} + \frac{U_B(1-\sigma)}{1-\sigma}\right) . \tag{12}$$

On the other hand, if the queue is not empty at a random slot boundary then the first customer in the queue can only be of class A when the server is processing a class B batch in the same slot and the size of the batch is less than the number of customers in the system. The partial pgf $Q_A(z)$ of the queue occupancy at random slot boundaries and the first customer is of class A is given by

$$Q_A(z) = \sum_{i=2}^{\infty} \Pr[u_{B,k} = i] \sum_{j=1}^{i-1} \Pr[c_k = j] z^{i-j} \quad ,$$

where c_k is the size of the batch being processed during slot k. The size of a class B batch follows a geometric distribution with parameter $1-\sigma$ that is truncated by the number of customers in the system. This leads to

$$Q_A(z) = \sum_{i=2}^{\infty} \Pr[u_{B,k} = i] \frac{\sigma(z^i - z(1-\sigma)^{i-1})}{z-1+\sigma} \quad .$$

By invoking the definition of $U_B(z) := \sum_{i=1}^{\infty} \lim_{k\to\infty} \Pr[u_{B,k} = i] z^i$ from Baetens et al (2016), we obtain

$$Q_A(z) = \frac{\sigma}{z-1+\sigma} U_B(z) - \frac{\sigma}{1-\sigma} \frac{z}{z-1+\sigma} U_B(1-\sigma) \quad . \tag{13}$$

An analogous analysis for a class B customer at the head of the queue leads to

$$Q_B(z) := \frac{1-\sigma}{z-\sigma} U_A(z) - \frac{1-\sigma}{\sigma} \frac{z}{z-\sigma} U_A(\sigma) \quad . \tag{14}$$

The expected value of the system occupancy and queue occupancy at random slot boundaries, denoted by $E[U]$ and $E[Q]$, are given by

$$\begin{aligned} E[U]&2(1-2\sigma(1-\sigma)\lambda) \\ &=\left[\frac{U_A(\sigma)}{\sigma}+\frac{U_B(1-\sigma)}{1-\sigma}\right]\left[2-2\sigma(1-\sigma)\lambda+\frac{2\lambda}{1-E(0)}-4\sigma(1-\sigma)E''(1)\right. \\ &\quad\left.+\frac{2\sigma(1-\sigma)(2+\lambda)\lambda}{1-E(0)}+\frac{2\sigma(1-\sigma)E''(1)}{1-E(0)}\right]-\frac{U_A(\sigma)}{\sigma}2\sigma(1+\lambda) \\ &\quad-\frac{U_B(1-\sigma)}{1-\sigma}2(1-\sigma)(1+\lambda)-(1-u_I)\Big(2-2\sigma(1-\sigma)(\lambda^2+E''(1))\Big)\ , \end{aligned} \tag{15}$$

$$E[Q]=E[U]-\frac{U_B(1)-U_B(1-\sigma)}{\sigma}-\frac{U_A(1)-U_A(\sigma)}{1-\sigma}=E[U]-\lambda\ . \tag{16}$$

In order to convert the steady-state pgf of the queue occupancy at random slot boundaries to the queue occupancy at customer arrival epochs, we need the steady-state pgf of the number of arrivals in the same slot before arrival of the random customer, denoted by $B(z)$. The steady-state pgf of this random variable can be shown to be equal to, see e.g. Bruneel and Kim (1993),

$$B(z)=\frac{1-E(z)}{\lambda(1-z)}\ . \tag{17}$$

Since we observe the system at arrival of a random customer, we know that a new service will start in the next slot. This means we can define the pgf's of the queue occupancy at customer arrival epochs if the next service or the first customer in the queue is of class A or B as $N_A(z)$ and $N_B(z)$ and their respective pmf's by $n_A(i)$ and $n_B(i)$. The customer at the head of the queue will be of class A if the queue was empty at the start of the slot and the first arrival is of class A with probability σ or if the queue was not empty at the start of the slot and the customer at the head of the queue is of class A. The partial steady-state pgf $N_A(z)$ is then given by

$$N_A(z)=\sigma q_0 B(z)+B(z)Q_A(z)\ , \tag{18}$$

and the analogous expression in case of a class B customer at the head of the queue is

$$N_B(z)=(1-\sigma)q_0 B(z)+B(z)Q_B(z)\ . \tag{19}$$

We define the combined steady-state pgf of the queue occupancy at customer arrival epochs as $N(z)$. This function is found by taking the sum of the partial steady-state pgfs $N_A(z)$ and $N_B(z)$, given by Eqs. 18 and 19, leading to

$$N(z)=q_0 B(z)+B(z)Q_A(z)+B(z)Q_B(z)=Q(z)B(z)\ . \tag{20}$$

We also derive the mean queue occupancy at customer arrival epochs, denoted by $E[N]$, as

$$E[N]=E[Q]+B'(1)=E[Q]+\frac{E''(1)}{2\lambda}\ . \tag{21}$$

3.3 Delay of a random customer

The partial steady-state pgf of the delay of a random customer when the customer at the head of the queue is of class A, say $D_A(z)$, can be derived from Eq. 6 as

$$D_A(z) = \sum_{n=0}^{\infty} \Pr[n_A = n] \begin{bmatrix} 1 & 0 \end{bmatrix} \boldsymbol{R}(z) \begin{bmatrix} \lambda_1(z)^n & 0 \\ 0 & \lambda_2(z)^n \end{bmatrix} \boldsymbol{L}(z) \begin{bmatrix} z \\ z \end{bmatrix} .$$

By invoking Eqs. 4, 5 and 18, we obtain

$$\begin{aligned} D_A(z) =& \sum_{n=0}^{\infty} \Pr[n_A = n] \begin{bmatrix} r_1(z) & r_2(z) \end{bmatrix} \begin{bmatrix} \lambda_1(z)^n & 0 \\ 0 & \lambda_2(z)^n \end{bmatrix} \begin{bmatrix} L_1(z) \\ L_2(z) \end{bmatrix} z \\ =& \sum_{n=0}^{\infty} \Pr[n_A = n] \Big(r_1(z)(\lambda_1(z))^n L_1(z) z + r_2(z)(\lambda_2(z))^n L_2(z) z \Big) \\ =& r_1(z) N_A(\lambda_1(z)) L_1(z) z + r_2(z) N_A(\lambda_2(z)) L_2(z) z \ . \end{aligned} \tag{22}$$

A similar analysis for the partial pgf $D_B(z)$ of the delay of a random customer when the customer at the head of the queue is of class B leads to

$$D_B(z) := N_B(\lambda_1(z)) L_1(z) z + N_B(\lambda_2(z)) L_2(z) z \ . \tag{23}$$

By combining Eqs. 22 and 23, we obtain the steady-state pgf $D(z)$ of the delay of a random customer:

$$D(z) = \Big(r_1(z) N_A(\lambda_1(z)) + N_B(\lambda_1(z)) \Big) L_1(z) z + \Big(r_2(z) N_A(\lambda_2(z)) + N_B(\lambda_2(z)) \Big) L_2(z) z \ .$$

The mean delay of a random customer, denoted by $E[D]$, is given by

$$\begin{aligned} E[D] =& \Big(r_1'(1) N_A(1) + r_1(1) N_A'(1) \lambda_1'(1) + N_B'(1) \lambda_1'(1) \Big) L_1(1) + L'(1) \Big(N_A(1) + N_B(1) \Big) \\ &+ \Big(N_A(1) + N_B(1) \Big) L_1(1) \Big) + L_2'(1) \Big(r_2(1) N_A(0) + N_B(0) \Big) \ . \end{aligned}$$

Using the definitions in Eqs. 3 and 4, we obtain

$$E[D] = 1 + 2\sigma(1-\sigma) N'(1) + (1-2\sigma) \Big[(1-\sigma)((N_A(1) - N_A(0)) - \sigma(N_B(1) - N_B(0)) \Big] \ . \tag{24}$$

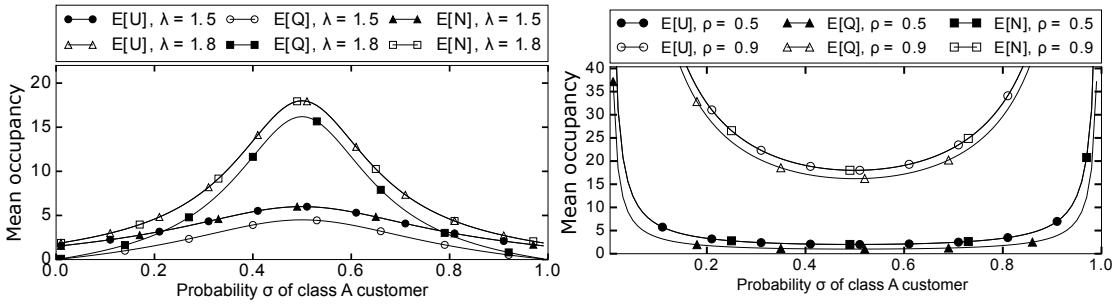
4 Numerical results

In this section, we will illustrate the results obtained in the previous section by using some numerical examples. To determine the influence of different parameters on the performance of the system, let us define the load of the system as

$$\rho := \frac{2\lambda}{\frac{1}{\sigma} + \frac{1}{1-\sigma}} \ .$$

In the first examples, we consider a geometric arrival process with mean arrival rate λ. The probability generating function or $E(z)$ is then equal to

$$E(z) = \frac{1}{1 + \lambda(1-z)} \ .$$



(a) For different arrival rates

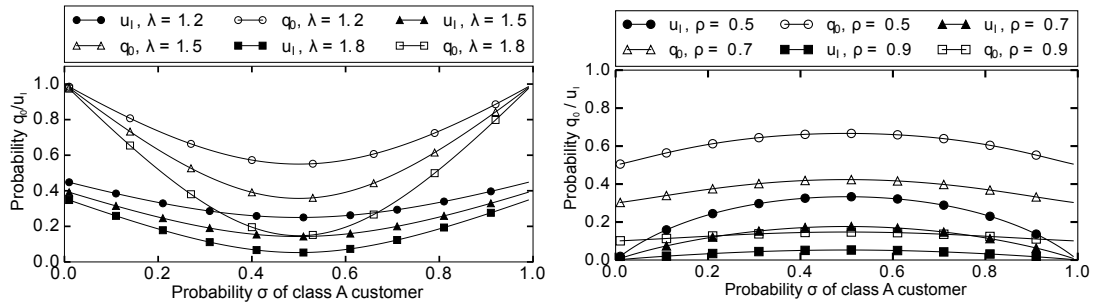
(b) For different loads

Fig. 1: Influence of σ on the mean system occupancy at slot boundaries $E[U]$, the queue occupancy at random slot boundaries $E[Q]$ and the queue occupancy at customer arrival epochs $E[N]$ for the system operating under different arrival rates (a) and loads (b)

In Fig. 1, $E[U]$ and $E[Q]$, respectively the mean system and queue occupancy at random slot boundaries, are depicted as well as $E[N]$, the mean queue occupancy at customer arrival epochs, see Eqs. 15, 16 and 21. We first note that $E[U]$ and $E[N]$ are identical, which is due to the memoryless property of the geometric arrival process. The difference between $E[U]$ and $E[Q]$ is equal to the average arrival rate λ which follows from Eq. 16. In Fig. 1a, we clearly see that, for fixed arrival rates, $\sigma = 0.5$ results in a maximum for the mean occupancies, and that values of σ closer to 0 or 1 significantly reduces the mean occupancies by allowing, on average, larger batches to be processed. Looking at the system under different loads results in the opposite behaviour, as can be seen in Fig. 1b. This is the result of the conflict between an increasing arrival rate in order to have the same load, and an increased mean number of customers being processed. It is clear that the effect of the increasing arrival rate is larger than the increase in average batch size. We also observe that the influence of σ, or the effect of the increased arrival rate, increases when the system is operating under higher loads.

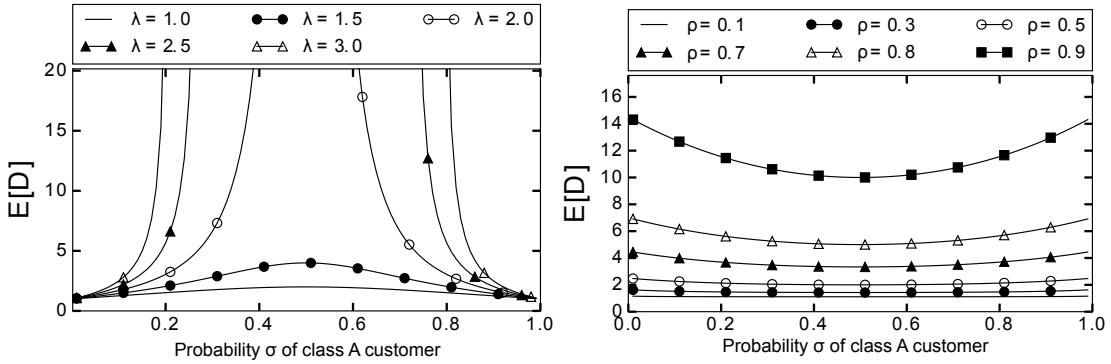
In Fig. 2, we study the influence of σ on u_I, the probability that the server is idle, and q_0, the probability that the queue is empty at random slot boundaries. Since q_0 is the sum of u_I and the probability that the server can process all waiting customers, see Eq. 12, it is always larger than u_I. We see in Fig. 2a that both q_0 and u_I reach a minimum for $\sigma = 0.5$, and values of σ closer to 0 or 1 leads to an increase of both q_0 and u_I, because the average number of customers being processed each service increases. We note that u_I converges to the probability that there are no arrivals during a single slot when σ goes to 0 or 1. In Fig. 2b, we look at the same system operating under different loads. We note that both u_I and q_0 reach a maximum for $\sigma = 0.5$ and that the probabilities decrease for σ closer to 0 or 1. For values of σ closer to 0 or 1, the effect of the decrease of u_I on q_0 is partially offset by an increase in the probability that the server can process all waiting customers simultaneously.

The expected delay of a random customer $E[D]$, calculated from Eq. 24 is shown in Fig. 3a for a number of arrival rates. We first note that only $\lambda < 2$ lead to a system that is stable for all values of σ. In the case that the system is stable for all values of σ, we see that there is a maximum if the arrival processes is symmetric or $\sigma = 0.5$. For $\lambda \geq 2$, there will be a σ value at which the system will become unstable and the delay will approach infinity if σ approaches this point of instability. We also observe that for very asymmetric systems or σ close to 0 or 1, the delay will go to a single slot or the service time of a single batch, even if the mean arrival rate is large. This is because almost all customers are of the same class which means that the mean



(a) For different arrival rates **(b)** For different loads

Fig. 2: Influence of σ on u_I and q_0 for the system operating under different arrival rates (a) and loads (b)



(a) For different arrival rates **(b)** For different loads

Fig. 3: Influence of σ on $E[D]$, the mean delay of a random customer, for different arrival rates (a) and loads (b)

service capacity is large so there is a high probability a newly arrived customer can be processed in the next slot.

In Fig. 3b, we show the mean delay for systems operating under different loads. We note that for small loads, the delay of a random customer is close to a single slot or to the service time of a single batch. This is because there is a high chance that the random customer will be the first customer in the queue and will be processed first. We also observe that there is a minimal expected delay for $\sigma = 0.5$ for all loads. An important remark is that, in contrast to the mean system and queue occupancies observed in Fig. 1, the mean delay stays limited, even for σ close to 0 or 1. This is because values of σ closer to 0 or 1 have two opposite consequences. The first consequence is an increased arrival rate which leads to a higher queue occupancy at customer arrival and indicates a longer expected delay. The opposite effect is that there will be larger sequences of same class customers which means that the server will process larger batches and the delay will decrease.

In the following examples we study the influence of the variance of the arrival process. To study the variance, we use an arrival process where the number of arrivals is determined by a weighted sum of a geometric distribution with mean $\frac{\lambda}{2\alpha}$ with probability α and a geometric

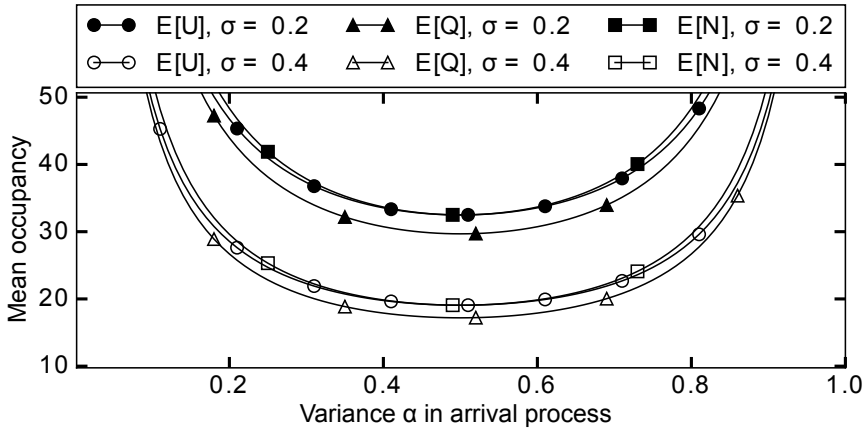


Fig. 4: Influence of variance in the arrival process on $E[U]$, $E[Q]$ and $E[N]$ for the system operating under a load $\rho = 0.9$ and σ equal to 0.2 or 0.4

distribution with mean $\frac{\lambda}{2(1-\alpha)}$ with probability $1-\alpha$. The probability generating function of this arrival process is given by

$$E(z) = \alpha \frac{1}{1 + \frac{\lambda}{2\alpha}(1-z)} + (1-\alpha)\frac{1}{1 + \frac{\lambda}{2(1-\alpha)}(1-z)} \ .$$

The mean arrival rate of this arrival process is still equal to λ, while the variance is given by

$$\text{Variance} = \frac{\lambda^2}{2\alpha(1-\alpha)} + \lambda - \lambda^2 \ .$$

This equation indicates that the variance is minimal for $\alpha = 0.5$, and approaches infinity for α equal to 0 or 1.

In Fig. 4, we show $E[U]$, $E[Q]$ and $E[N]$ as functions of α with a load $\rho = 0.9$ and $\sigma = 0.2$ or 0.4. We note that when α is 0.5 or when the variance is minimal, the three observed quantities are also minimal. For increasing variance in the arrival process or values of α closer to 0 or 1, we see that although the three quantities increase, the difference between $E[Q]$ and $E[U]$ remains constant and $E[N]$ increases faster than $E[Q]$. This is because $E[N]$ is the sum of $E[Q]$ and the average number of arrivals before a random customer. An increasing variance causes the number of arrivals before a random customer to increase, explaining the larger increase of $E[N]$ than that of $E[Q]$.

The influence of the arrival variance on u_I and q_0 is shown in Fig. 5. We clearly see that an increasing variance leads to an increasing idle probability. This is because the server cannot be idle if there was at least one arrival during the previous slot, and an increasing variance means that there is a higher probability that there are no arrivals, so the probability that the server will be idle increases. We also note that the influence of the variance in the arrival process is independent of the value of σ. In contrast to the effect on the idle probability, an increasing variance leads to a decreasing q_0. This is caused by a decreasing probability that the server can process all waiting customers simultaneously, which is clearly more significant than the increasing u_I. We note that the influence of the variance on q_0 decreases for σ closer to 0 or 1.

In Fig. 6 we show the influence of the arrival variance on $E[D]$ for the system operating under different loads and with $\sigma = 0.2$. We first observe that in contrast with the influence of σ on the

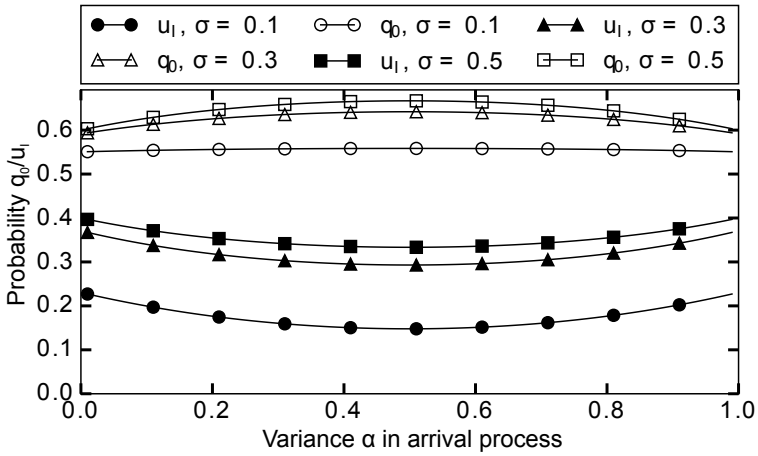


Fig. 5: Influence of variance in the arrival process on u_I and q_0 with a load of 0.5 and σ equal to 0.1, 0.3 or 0.5

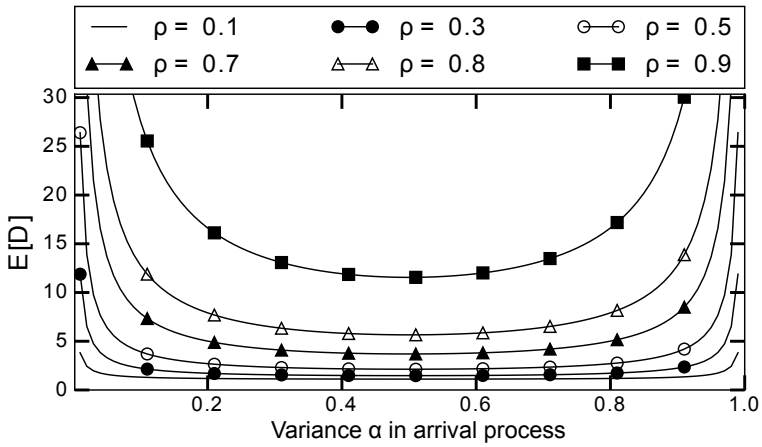


Fig. 6: Influence of variance in the arrival process on $E[D]$, the delay of a random customer, for $\sigma = 0.2$ and the system operating under different loads

mean delay, see Fig. 3b, the mean delay goes to infinity for α close to 0 or 1 or a high amount of variance in the arrival process. This is because a large variance means that there are many slots with no or few arrivals and few slots with many arrivals, which means that there is a higher probability that there will be many customers before the random customer, which in turn means that the delay will be longer. For smaller loads, we also note that the delay of a random customer will be close to the service time of a single batch, except for α close to 0 or 1.

5 Conclusions and future work

In this paper we have analysed a discrete-time two-class single-server queueing system with batch service. The sizes of the processed batches are determined by the number of customers in the

queue and their respective classes. From the queue occupancy at customer arrival epochs, we were able to determine the steady-state probability generating function of the delay of a random customer which was not done for batch service systems with variable capacity. Using these results, we have shown that the degree of asymmetry between the number of arrivals of the customer classes and an increased variance of the number of arrivals in each slot have a significant impact on the performance of the system.

There are a number of possible extensions that could be considered for this system. In a first extension we could introduce class-dependent general service-time distributions for class A and B batches. An analysis approach for this may be to analyse the system at service initiation opportunities, which are the boundaries of slots in which a new service is initiated or the server is idle. A second extension is that we could introduce clustering behaviour of same class customers by introducing correlation between the classes of two consecutive customers. This can for instance be done by using two different probabilities that the customer will remain of the same class if the previous customer was of class A or B. This allows us to tweak the lengths of class A or B customer sequences that arrive while maintaining a certain ratio of class A and B customers. A further possible extension is to do the delay analysis in case of more than 2 different customer classes. We also note that the model in this paper does not use a maximum service capacity. While this is not realistic, it is a good approximation in systems where the load is not too high or where the proportion of each customer type within the arrival stream is non negligible (that is σ not near to 0 and not near to 1). Under these assumptions, the length of a sequence of customers of the same class is limited which means that the service capacity will also be limited. In future work, we will incorporate a class-dependent maximum service capacity.

Appendix A Eliminating branching points

We analyse the steady-state pgf $D_{A,n}(z)$ for the delay of a random customer with n customers in the queue and the customer at the head of the queue is of class A. From Eq. 6 we obtain that the pgf $D_{A,n}(z)$ is equal to

$$D_{A,n}(z) = z\frac{(1-\sigma)z}{\lambda_1(z)-\sigma}\frac{\lambda_1(z)+\sigma(z-1)}{2\lambda_1(z)-1}\lambda_1(z)^n + z\frac{(1-\sigma)z}{\lambda_2(z)-\sigma}\frac{\lambda_2(z)+\sigma(z-1)}{2\lambda_2(z)-1}\lambda_2(z)^n \ .$$

We note that for any polynomial $f(z)$ with real coefficients then $f(z)+f(z*)$, with $z^\star$ the complex conjugate of z, gives a real number. This is also the case for the arguments $\lambda_1(z)$ and $\lambda_2(z)$ given by

$$\lambda_{1,2}(z) = \frac{1}{2}\Big(1 \pm \sqrt{1+4\sigma(1-\sigma)(z^2-1)}\Big) \ , \tag{25}$$

so that $f(\lambda_1(z)) + f(\lambda_2(z))$ is a function without roots and also without branching points. We first note that $(2\lambda_1(z)-1)$ and $(2\lambda_2(z)-1)$ can be written as

$$\begin{aligned} 2\lambda_1(z)-1 &= \sqrt{1+4\sigma(1-\sigma)(z^2-1)} \ , \\ 2\lambda_2(z)-1 &= -\sqrt{1+4\sigma(1-\sigma)(z^2-1)} \ . \end{aligned}$$

By using Newton's binomial expansion for the n-th powers, we obtain

$$D_{A,n}(z) = \left(\frac{1}{2}\right)^n \frac{(1-\sigma)z^2}{(2\lambda_1(z)-1)(\lambda_1(z)-\sigma)(\lambda_2(z)-\sigma)} \left[(\lambda_2(z)-\sigma)(\lambda_1(z)+\sigma(z-1))\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{1+4\sigma(1-\sigma)(z^2-1)}\right)^k - (\lambda_1(z)-\sigma)(\lambda_2(z)+\sigma(z-1))\sum_{k=0}^{n}\binom{n}{k}\left(-\sqrt{1+4\sigma(1-\sigma)(z^2-1)}\right)^k\right] .$$

Invoking the definitions of $\lambda_1(z)$ and $\lambda_2(z)$ in Eq. 25 results in

$$D_{A,n}(z) = \left(\frac{1}{2}\right)^n \frac{-z}{\sqrt{1+4\sigma(1-\sigma)(z^2-1)}} \left[\left(\frac{1-2\sigma}{2}-(1-\sigma)z-\frac{\sqrt{1+4\sigma(1-\sigma)(z^2-1)}}{2}\right) \sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{1+4\sigma(1-\sigma)(z^2-1)}\right)^k - \left(\frac{1-2\sigma}{2}-(1-\sigma)z+\frac{\sqrt{1+4\sigma(1-\sigma)(z^2-1)}}{2}\right) \sum_{k=0}^{n}\binom{n}{k}\left(-\sqrt{1+4\sigma(1-\sigma)(z^2-1)}\right)^k\right] .$$

Rearranging of the summations leads to

$$D_{A,n}(z) = \left(\frac{1}{2}\right)^n \frac{-z}{\sqrt{\cdots}} \left[\left(\frac{1-2\sigma}{2}-(1-\sigma)z\right)\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{\cdots}\right)^k - \left(-\sqrt{\cdots}\right)^k - \frac{\sqrt{\cdots}}{2}\sum_{k=0}^{n}\binom{n}{k}\left(\sqrt{\cdots}\right)^k + \left(-\sqrt{\cdots}\right)^k\right] ,$$

where we abbreviated $\sqrt{1+4\sigma(1-\sigma)(z^2-1)}$ as $\sqrt{\cdots}$. We clearly see that in the first summation only the terms when k is odd are non-zero and in the second summation only the even values of k remain. As a result we can write $D_{A,n}(z)$ as

$$D_{A,n}(z) = \left(\frac{1}{2}\right)^n z(1+2(1-\sigma)(z-1))\sum_{j=0}^{\lfloor\frac{n-1}{2}\rfloor}\binom{n}{2j+1}(1+4\sigma(1-\sigma)(z^2-1))^j + \left(\frac{1}{2}\right)^n z\sum_{j=0}^{\lfloor\frac{n}{2}\rfloor}\binom{n}{2j}(1+4\sigma(1-\sigma)(z^2-1))^j .$$

and the analogue equation for a class B customer at the head of the queue is

$$D_{B,n}(z) = \left(\frac{1}{2}\right)^n z(1+2\sigma(z-1))\sum_{j=0}^{\lfloor\frac{n-1}{2}\rfloor}\binom{n}{2j+1}(1+4\sigma(1-\sigma)(z^2-1))^j + \left(\frac{1}{2}\right)^n z\sum_{j=0}^{\lfloor\frac{n}{2}\rfloor}\binom{n}{2j}(1+4\sigma(1-\sigma)(z^2-1))^j .$$

We clearly see that both these functions are polynomials of degree $n+1$.

References

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 29:972–986

Baetens J, Steyaert B, Claeys D, Bruneel H (2016) System occupancy of a two-class batch-service queue with class-dependent variable server capacity. In: International Conference on Analytical and Stochastic Modeling Techniques and Applications, Springer, pp 32–44

Banerjee A, Gupta U (2012) Reducing congestion in bulk-service finite-buffer queueing system using batch-size-dependent service. Performance Evaluation 69(1):53–70

Banerjee A, Gupta U, Goswami V (2014) Analysis of finite-buffer discrete-time batch-service queue with batch-size-dependent service. Computers and Industrial Engineering 75:121–128

Banerjee A, Gupta U, Chakravarthy S (2015) Analysis of a finite-buffer bulk-service queue under markovian arrival process with batch-size-dependent service. Computers and Operations Research 60:138–149

Bitran GR, Tirupati D (1989) Approximations for product departures from a single-server station with batch processing in multi-product queues. Management Science 35(7):851–878, URL http://www.jstor.org/stable/2632313

Boxma O, van der Wal J, Yechiali U (2008) Polling with batch service. Stochastic Models 24(4):604–625

Bruneel H, Kim B (1993) Discrete-time models for communication systems including ATM. Kluwer Academic, Boston, USA

Chang S, Choi D (2005) Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations. Computers and Operations Research 32(9):2213–2234

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

Chaudhry M, Chang S (2004) Analysis of the discrete-time bulk-service queue $Geo/G^Y/1/N+B$. Operations Research Letters 32(4):355–363

Chaudhry M, Templeton J (1983) A first course in bulk queues. Wiley New York

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(6):528–549

Claeys D, Steyaert B, Walraevens J, Laevens K, Bruneel H (2012) Tail distribution of the delay in a general batch-service queueing model. Computers and Operations Research 39:2733–2741

Claeys D, Steyaert B, Walraevens J, Laevens K, Bruneel H (2013a) Analysis of a versatile batch-service queueing model with correlation in the arrival process. Performance Evaluation 70(4):300–316

Claeys D, Steyaert B, Walraevens J, Laevens K, Bruneel H (2013b) Tail probabilities of the delay in a batch-service queueing model with batch-size dependent service times and a timer mechanism. Computers and Operations Research 40(5):1497–1505

Dorsman J, der Mei RV, Winands E (2012) Polling with batch service. OR Spectrum 34:743–761

Fowler J, Phojanamongkolkij N, Cochran J, Montgomery D (2002) Optimal batching in a wafer fabrication facility using a multiproduct G/G/C model with batch processing. International Journal of Production Research 40(2):275–292

Freivalds A, Niebel B (2014) Niebel's methods, standards, and work design, vol 13. Mc Graw Hill, New York

Germs R, Foreest NV (2010) Loss probabilities for the $M^X/G^Y/1/K+B$ queue. Probability in the Engineering and Informational Sciences 24(4):457–471

Germs R, Foreest NV (2013) Analysis of finite-buffer state-dependent bulk queues. OR Spectrum 35(3):563–583

Goswami V, Mohanty J, Samanta S (2006) Discrete-time bulk-service queues with accessible and non-accessible batches. Applied Mathematics and Computation 182:898–906

Huang MG, Chang PL, Chou YC (2001) Analytic approximations for multiserver batch-service workstations with multiple process recipes in semiconductor wafer fabrication. IEEE Transactions on Semiconductor Manufacturing 14(4):395–405

Janssen A, van Leeuwaarden J (2005) Analytic computation schemes for the discrete-time bulk service queue. Queueing Systems 50:141–163

Latouche G, Ramaswami V (1999) Introduction to matrix analytic methods in stochastic modeling, vol 5. Siam

Lee H, Kim S (1994) Optimal dispatching of an infinite capacity shuttle with compound poisson arrivals: Control at a single terminal. Computers and Operations Research 21(1):67–78, DOI http://dx.doi.org/10.1016/0305-0548(94)90063-9, URL http://www.sciencedirect.com/science/article/pii/0305054894900639

Powell W, 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

Pradhan S, Gupta U, Samanta S (2015) Queue-length distribution of a batch service queue with random capacity and batch size dependent service: M/g r y/1. OPSEARCH pp 1–15

Reddy G, Nadarajan R, Kandasamy P (1993) A nonpreemptive priority multiserver queueing system with general bulk service and heterogeneous arrivals. Computers and operations research 20(4):447–453

Weng W, Leachman R (1993) An improved methodology for real-time production decisions at batch-process work stations. IEEE Transactions on Semiconductor Manufacturing 6(3):219–225

Willems D (2014) Modeling of a distribution center as a queuing system. Master's thesis, Ghent University

Wu K, McGinnis LF, Zwart B (2011) Approximating the performance of a batch service queue using the model. IEEE Transactions on Automation Science and Engineering 8(1):95–102

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