# SYSTEM PERFORMANCE OF A VARIABLE-CAPACITY BATCH-SERVICE QUEUE WITH GEOMETRIC SERVICE TIMES AND CUSTOMER-BASED CORRELATION

Jens Baetens[1]
Bart Steyaert[1]
Dieter Claeys[1,2,3]
Herwig Bruneel[1]
[1]SMACS Research Group
Department of Telecommunications and Information Processing
Ghent University
St-Pietersnieuwstraat 41, B 9000 Gent, Belgium
[2]Lean Enterprise Research Center
Department of Industrial Systems Engineering and Product Design
Ghent University
Technologiepark 903, 9052 Zwijnaarde (Gent), Belgium
[3]Department of Agile and Human Centered Production and Robotic Systems
Flanders Make
E-mail:{jens.baetens, bart.steyaert,dieter.claeys,herwig.bruneel}@ugent.be

## KEYWORDS

Queueing Performance; Batch Service; Variable Server Capacity; Two Classes; Geometric Service Times; Customer-based Correlation

## ABSTRACT

In many queueing systems the server processes several customers simultaneously. Although the capacity of a batch, that is the number of customers that can be processed simultaneously, is often variable in practice, nearly all batch-service queueing models in literature consider a constant capacity. In this paper, we extend previous work on a batch-service queueing model with variable server capacity, where customers of two classes are accommodated in a common first-come-first-served single-server queue. We include correlation between the classes of consecutive customers, and the service times are geometrically distributed. We establish the equations that govern the system behaviour, the stability condition, and an expression for the steady-state probability generating function of the system occupancy at random slot boundaries. In addition, some numerical results are shown to study the impact of the mean service times and of the customer-based correlation in the arrival process on the performance of the queueing system.

## INTRODUCTION

In manufacturing environments, a single machine often has the capability to process multiple products simultaneously in a single batch. The maximum number of products that can be processed at the same time, also called the service capacity, is often assumed to be a constant and bounded value, e.g. Banerjee and Gupta 2012; Claeys et al. 2012, 2013; Goswami et al. 2006; Weng and Leachman 1993. The difference with multi-server systems is that a newly arrived customer cannot join an ongoing service, even if the service is not completely full. In manufacturing, this can be found in for instance a furnace where a heating phase cannot be interrupted. In real systems, the maximum batch size or service capacity is often variable and stochastic, which has been studied in only a few papers. In many of these papers, the service capacity is generically distributed and does not depend on any parameter of the system, like the number of customers waiting in the queue, see the papers of Chaudhry and Chang 2004; Germs and Foreest 2010; Pradhan et al. 2015; Sikdar and Samanta 2016. A more detailed description of these papers can be found in our paper (Baetens et al. 2016). Also, Germs and Foreest (2013) have developed an algorithmic method to analyse the performance of continuous-time batch service queueing systems with arrival process, service time distribution and variable service capacity that depend on the number of customers in the queue.

None of the previous papers on batch service considered multiple customer classes. Differentiated service is common in priority queueing and polling systems, where the system can use different scheduling algorithms to optimize the performance of the system (e.g. Reddy et al. (1993); Boxma et al. (2008); Dorsman et al. (2012)).In contrast with polling systems and priority queueing, that use a unique queue for each possible class of customers, we use a common queue with a global First-Come-First-Served service discipline, because it is not always feasible to install a multi-queue system due to certain constraints like the increased cost of a more complicated system. An example of such a

Proceedings 31st European Conference on Modelling and
Simulation ©ECMS Zita Zoltay Paprika, Péter Horák, Kata Váradi,
Péter Tamás Zwierczyk, Ágnes Vidovics-Dancs, János Péter Rádics (Editors)
ISBN: 978-0-9932440-4-9/ ISBN: 978-0-9932440-5-6 (CD)

system is a furnace in a production line. A furnace can handle multiple customers simultaneously as long as the products must be heated to the same temperature and for the same duration.

In this paper, we analyse the performance of a two-class discrete-time batch-service queueing model, with a variable service capacity that depends on the queue size and on the specific classes of the successive customers. This kind of queueing systems can be found in many telecommunication technologies like optical burst switching networks (Chen et al. 2004) and wireless local area networks (Lu et al. 2005). To the best of our knowledge, the combination of batch service with variable capacity and multiple customer classes has only appeared in our previous paper Baetens et al. (2016). The difference with that paper is that we relax the assumption of fixed single-slot service times by considering geometric service times instead. We also introduce correlation between the classes of consecutive customers in order to model the tendency for same-class customers to arrive in clusters.

The paper is structured as follows. In the next section, we describe the discrete-time two-class queueing model with batch service in detail. In the third section we establish the system equations, from which we deduce the stability condition, and derive a closed-form expression for the steady-state probability generating function (pgf) of the system occupancy at random slot boundaries. Next, using the expressions obtained in this part, we evaluate the performance of the system through some numerical examples and study the impact of the mean service time and the degree of correlation between the classes of consecutive customers on the system performance. Finally, we draw some conclusions about the obtained results.

## MODEL DESCRIPTION

In this paper we study a discrete-time two-class queueing system with infinite queue size. This system uses a batch server with a stochastic service capacity based on the class of the customer at the head of the queue. We distinguish two different customer classes in the arrival process, which we will call class $A$ and $B$. When the server is idle or becomes available at the start of a new slot and finds a non-empty queue, a new service is initiated immediately. The capacity of this service is determined by the number of consecutive customers at the head of the queue that are of the same class which means it depends on the class of the batch being processed. More specifically, the server starts serving a batch of $n$ customers if and only if one of the following two cases occurs:

- 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 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 pgf $E(z)$. The mean of these i.i.d. random variables is denoted as $\lambda$. The class of a newly arrived customer depends on the class of the previous customer. If the previous customer was of class $A$ $(B)$, then the newly arrived customer will also be of class $A$ $(B)$ with probability (w.p.) $\alpha$ $(\beta)$. If $\alpha+\beta>1$, same-class customers will have a tendency to arrive in clusters. The service time of a random batch follows a geometric distribution with mean $\mu$, and does not depend on the class of the processed batch and its size. The pgf of the service time distribution is defined as

$$S(z) = \frac{z}{z+\mu(1-z)} \ .$$

## ANALYSIS

In this section, we first determine the system equations that capture the system behaviour. In the next part we analyse the conditions under which the system is stable. In the main part, we establish the steady-state pgf of the system occupancy, that is the number of customers in the system at the beginning of a random slot, including those in the ongoing service if the server is not idle.

### System Equations

We start by defining the variables we use in the system equations that capture the behaviour of the system at consecutive slot boundaries. The number of customers in the system, or the system occupancy at random slot boundaries, is denoted by $u_k$. The number of customers in the system while the server is idle and the previously processed batch is of class $A$ or $B$ is respectively denoted as $u_{I,A,k}$ and $u_{I,B,k}$. We also define the random variables $u_{A,k}$ and $u_{B,k}$ as the system occupancy at random slot boundaries if the server is processing a class $A$ or $B$ batch during slot $k$.

If the server is idle in slot $k$, then the server will remain idle if there are no new arrivals. On the other hand, when the server finds at least one newly arrived customer, then a new batch is started immediately. The probability that the class of the first arrival in slot $k$ is of class $A$ or $B$ depends on the class of the most recently processed batch which is why we distinguish between the two cases. The system equations if the previous batch was a class $A$ batch are given by

$$\begin{aligned} u_{I,A,k+1} &= 0 \ \text{ if } e_k = 0 \\ u_{A,k+1} &= e_k \ \text{ if } e_k > 0 \ \& \ \text{ first arrival of class } A \\ u_{B,k+1} &= e_k \ \text{ if } e_k > 0 \ \& \ \text{ first arrival of class } B, \quad (1) \end{aligned}$$

where $e_k$ represents the number of arrivals during slot $k$. The analogous equations if the previous batch was of class $B$ are

$$\begin{aligned}u_{I,B,k+1} &= 0 \text{ if } e_k = 0 \\ u_{A,k+1} &= e_k \text{ if } e_k > 0 \text{ \& first arrival of class } A \\ u_{B,k+1} &= e_k \text{ if } e_k > 0 \text{ \& first arrival of class } B. \quad (2)\end{aligned}$$

On the other hand, if the server is processing a class $A$ batch during slot $k$ and the service period does not end in slot $k+1$, then the service continues and the newly arrived customers are added to the tail of the queue. However, if the service ends, then the behaviour will be determined by the probability that all customers in the system at service initiation of the batch, processed during slot $k$, were of the same class. In the case that all waiting customers were of the same class, then the behaviour will be as if the server was idle in the previous slot. On the other hand, if at least one customer in the queue was a class $B$ customer, then the next batch will always be a class $B$ batch. The resulting system equations in the case of an ongoing class $A$ batch are

$$\begin{aligned}u_{A,k+1} =& u_{A,k} + e_k \text{ if Service not done} \\ u_{I,A,k+1} =& 0 \text{ if Service done} \\ & \text{\& } c_{A,k} = u_{A,\text{ini},k} \text{ \& } se_k = 0 \\ u_{A,k+1} =& se_k \text{ if Service done \& } c_{A,k} = u_{A,\text{ini},k} \\ & \text{\& } se_k > 0 \text{ \& first arrival of class } A \\ u_{B,k+1} =& se_k \text{ if Service done \& } c_{A,k} = u_{A,\text{ini},k} \\ & \text{\& } se_k > 0 \text{ \& first arrival of class } B \\ u_{B,k+1} =& u_{A,\text{ini},k} - c_{A,k} + se_k \\ & \text{if Service done \& } c_{A,k} < u_{A,\text{ini},k} \ , \quad (3)\end{aligned}$$

where $se_k$, $c_{A,k}$ ($c_{B,k}$) and $u_{A,ini,k}$ ($u_{B,ini,k}$) represent respectively the number of arrivals, the service capacity and the system occupancy at initiation of the ongoing service during slot $k$ which is of class $A$ ($B$). The analogous system equations for the case that the ongoing service is of class $B$ are given by

$$\begin{aligned}u_{B,k+1} =& u_{B,k} + e_k \text{ if Service not done} \\ u_{I,B,k+1} =& 0 \text{ if Service done} \\ & \text{\& } c_{B,k} = u_{B,\text{ini},k} \text{ \& } se_k = 0 \\ u_{B,k+1} =& se_k \text{ if Service done \& } c_{B,k} = u_{B,\text{ini},k} \\ & \text{\& } se_k > 0 \text{ \& first arrival of class } B \\ u_{A,k+1} =& se_k \text{ if Service done \& } c_{B,k} = u_{B,\text{ini},k} \\ & \text{\& } se_k > 0 \text{ \& first arrival of class } A \\ u_{A,k+1} =& u_{B,\text{ini},k} - c_{B,k} + se_k \\ & \text{if Service done \& } c_{B,k} < u_{B,\text{ini},k} \ . \quad (4)\end{aligned}$$

### Stability Condition

In this part we analyse a system in which the server is always busy and the variable server capacity is always smaller than the number of waiting customers, also called a saturated system. In such a system, the size of the processed batches is geometrically distributed and a class $A$ and $B$ batch are processed alternately. The system is stable when the mean number of customer arrivals during two consecutive service periods, which is equal to $2\mu\lambda$, is less than the mean number of customers that leave the system during the same two service periods. Since a class $A$ and $B$ batch leave the system during this time period, the mean number of customers that leave the system is the sum of the mean batch size of a class $A$ and $B$ batch. The batch size of a class $A$ and $B$ batch follow a geometric distribution with parameter $\alpha$ or $\beta$ respectively. The stability condition is then given by

$$2\mu\lambda < \frac{1}{1-\alpha} + \frac{1}{1-\beta} \ .$$

If $\alpha$ or $\beta$ is equal to 1, then the system will always be stable, since all customers that arrive are of the same class. The server can then group all waiting customers, which leaves an empty queue after service initiation. This also follows from the stability condition, which is reduced to $\lambda < \infty$ under the restriction that $\alpha$ or $\beta$ is equal to 1. Another element of interest in the stability condition, is the maximum allowed arrival rate, which reaches a minimum value for $\alpha = \beta = 0.5$. Finally, the load $\rho$ of the system is defined as the fraction of $\lambda$ versus the maximum allowed arrival rate, which leads to

$$\rho := \frac{2\mu\lambda}{\frac{1}{1-\alpha} + \frac{1}{1-\beta}} = 2\mu\lambda \frac{(1-\alpha)(1-\beta)}{2-\alpha-\beta} < 1 \ .$$

The stability condition implies that the load is smaller than 1.

### System Occupancy

Assuming the stability condition is met, we can define the steady-state pmf of the system occupancy at random slot boundaries, as

$$u(i) := \lim_{k\to\infty} \Pr[u_k = i] \ ,$$

with corresponding pgf

$$U(z) := \sum_{i=0}^{\infty} u(i) z^i \ .$$

We can split the generating function of the system occupancy $U(z)$ in two parts based on the class of the most recently initiated service. This leads to

$$U(z) = u_{I,A} + u_{I,B} + U_A(z) + U_B(z) \ ,$$

where we introduced the following definitions

$$\begin{aligned}u_I, A &:= \lim_{k\to\infty} \Pr[u_{I,A,k} = 0] \ , \\ u_I, B &:= \lim_{k\to\infty} \Pr[u_{I,B,k} = 0] \ , \\ U_A(z) &:= \sum_{i=1}^{\infty} \lim_{k\to\infty} \Pr[u_{A,k} = i] z^i \ , \\ U_B(z) &:= \sum_{i=1}^{\infty} \lim_{k\to\infty} \Pr[u_{B,k} = i] z^i \ .\end{aligned}$$

The probability that the server is idle and the previously initiated service contained class $A$ customers, denoted by $u_{I,A}$, is found by invoking system equations

Eq. (1) and Eq. (3).

$$u_{I,A} = \frac{S(E(0))}{1-E(0)}\frac{U_A(\alpha)}{\mu\alpha} \quad . \tag{5}$$

Using Eq. (2) and Eq. (4), we find the analogous equation if the previous batch is of class $B$

$$u_{I,B} = \frac{S(E(0))}{1-E(0)}\frac{U_B(\beta)}{\mu\beta} \quad . \tag{6}$$

The partial pgf $U_A(z)$ of the system occupancy when the server is processing a class $A$ batch can be split based on the state of the server in the previous slot. This leads to

$$\begin{aligned} U_A(z) &= E[z^{u_{A,k+1}}] \\ &= E[z^{u_{A,k+1}} I_{\{u_{I,A,k}=0\}}] + E[z^{u_{A,k+1}} I_{\{u_{I,B,k}=0\}}] \\ &\quad + E[z^{u_{A,k+1}} I_{\{u_{A,k}>0\}}] + E[z^{u_{A,k+1}} I_{\{u_{B,k}>0\}}] \ , \end{aligned} \tag{7}$$

where $I_{\{C\}}$ are indicator functions which are equal to 1 if event $C$ occurs and zero otherwise. Invoking the system equations in Eq. (1), we can write the first term of the right-hand side of Eq. (7) as

$$E[z^{u_{A,k+1}} I_{\{u_{I,A,k}=0\}}] = \alpha u_{I,A}(E(z)-E(0)) \quad . \tag{8}$$

Analogously, by using Eq. (2), we obtain

$$\begin{aligned} &E[z^{u_{A,k+1}} I_{\{u_{I,B,k}=0\}}] \\ &\qquad = (1-\beta)u_{I,B}(E(z)-E(0)) \quad . \end{aligned} \tag{9}$$

Using Eq. (3), we obtain the following equation for the third term of the right hand side of Eq. (7)

$$\begin{aligned} E[z^{u_{A,k+1}} I_{\{u_{A,k}>0\}}] &= (1-\frac{1}{\mu})U_A(z)E(z) \\ &\quad + \frac{\alpha}{\mu}\Big(\frac{U_A(\alpha)E(\alpha)}{\alpha S(E(\alpha))} - \frac{U_A(\alpha)}{\alpha}S(E(0))\Big) \\ &\quad \cdot \frac{S(E(z))-S(E(0))}{1-S(E(0))} \quad . \end{aligned} \tag{10}$$

The last term of $U_A(z)$ results in

$$\begin{aligned} &E[z^{u_{A,k+1}} I_{\{u_{B,k}>0\}}] \\ &= \frac{1-\beta}{\mu}\Big(\frac{U_B(\beta)E(\beta)}{\beta S(E(\beta))} - \frac{U_B(\beta)}{\beta}S(E(0))\Big) \\ &\quad \cdot \frac{S(E(z))-S(E(0))}{1-S(E(0))} + \frac{1-\beta}{\mu(z-\beta)} \\ &\quad \cdot \Big(\frac{U_B(z)E(z)}{S(E(z))} - \frac{z}{\beta}\frac{U_B(\beta)E(\beta)}{S(E(\beta))}\Big)S(E(z)) \quad . \end{aligned} \tag{11}$$

By combining Eqs. (8-11), we obtain

$$\begin{aligned} U_A(z)\Big(\mu-(\mu-1)E(z)\Big) &= \frac{1-\beta}{z-\beta}U_B(z)E(z) \\ &+ S(E(0))\Big(\frac{E(z)-E(0)}{1-E(0)} - \frac{S(E(z))-S(E(0))}{1-S(E(0))}\Big) \\ &\cdot \Big(U_A(\alpha)+\frac{1-\beta}{\beta}U_B(\beta)\Big) + \frac{S(E(z))-S(E(0))}{1-S(E(0))} \\ &\cdot \Big(\frac{U_A(\alpha)E(\alpha)}{S(E(\alpha))} + \frac{1-\beta}{\beta}\frac{U_B(\beta)E(\beta)}{S(E(\beta))}\Big) \\ &- \frac{(1-\beta)z}{\beta(z-\beta)}\frac{U_B(\beta)E(\beta)}{S(E(\beta))}S(E(z)) \quad . \end{aligned} \tag{12}$$

Multiplying both sides of Eq. (12) by $\frac{S(E(z))}{E(z)}$ results in

$$\begin{aligned} U_A(z) &= \frac{1-\beta}{z-\beta}U_B(z)S(E(z)) + S(E(0))\frac{S(E(z))}{E(z)} \\ &\cdot \Big(\frac{E(z)-E(0)}{1-E(0)} - \frac{S(E(z))-S(E(0))}{1-S(E(0))}\Big) \\ &\cdot \Big(U_A(\alpha)+\frac{1-\beta}{\beta}U_B(\beta)\Big) + \frac{S(E(z))-S(E(0))}{1-S(E(0))} \\ &\cdot \frac{S(E(z))}{E(z)}\Big(\frac{U_A(\alpha)E(\alpha)}{S(E(\alpha))} + \frac{1-\beta}{\beta}\frac{U_B(\beta)E(\beta)}{S(E(\beta))}\Big) \\ &- \frac{(1-\beta)z}{\beta(z-\beta)}\frac{U_B(\beta)E(\beta)}{S(E(\beta))}\frac{S(E(z))^2}{E(z)} \quad . \end{aligned} \tag{13}$$

A similar analysis leads to an equation for the partial pgf of the system occupancy if the customer is processing a class $B$ batch

$$\begin{aligned} U_B(z) &= \frac{1-\alpha}{z-\alpha}U_A(z)S(E(z)) + S(E(0))\frac{S(E(z))}{E(z)} \\ &\cdot \Big(\frac{E(z)-E(0)}{1-E(0)} - \frac{S(E(z))-S(E(0))}{1-S(E(0))}\Big) \\ &\cdot \Big(U_B(\beta)+\frac{1-\alpha}{\alpha}U_A(\alpha)\Big) + \frac{S(E(z))-S(E(0))}{1-S(E(0))} \\ &\cdot \frac{S(E(z))}{E(z)}\Big(\frac{U_B(\beta)E(\beta)}{S(E(\beta))} + \frac{1-\alpha}{\alpha}\frac{U_A(\alpha)E(\alpha)}{S(E(\alpha))}\Big) \\ &- \frac{(1-\alpha)z}{\alpha(z-\alpha)}\frac{U_A(\alpha)E(\alpha)}{S(E(\alpha))}\frac{S(E(z))^2}{E(z)} \quad . \end{aligned} \tag{14}$$

Using Eqs. (5), (6), (13) and (14), we obtain

$$\begin{aligned} &U(z)\big((z-\alpha)(z-\beta) - (1-\alpha)(1-\beta)S(E(z))^2\big) \\ &= \frac{S(E(0))}{1-E(0)}\frac{U_A(\alpha)}{\mu\alpha} + \frac{S(E(0))}{1-E(0)}\frac{U_B(\beta)}{\mu\beta} \\ &\quad + \Big(\frac{E(z)-E(0)}{1-E(0)} - \frac{S(E(z))-S(E(0))}{1-S(E(0))}\Big) \\ &\quad \cdot S(E(0))\frac{S(E(z))}{E(z)}\Bigg[(z-\alpha)(z-\beta)\Big(\frac{U_A(\alpha)}{\alpha}+\frac{U_B(\beta)}{\beta}\Big) \\ &\quad + (1-\alpha)S(E(z))(z-\alpha-\beta z+\alpha z)\frac{U_A(\alpha)}{\alpha} \\ &\quad + (1-\beta)S(E(z))(z-\beta-\alpha z+\beta z)\frac{U_B(\beta)}{\beta}\Bigg] \\ &\quad + \frac{S(E(z))-S(E(0))}{1-S(E(0))}\frac{S(E(z))}{E(z)}\Bigg[\Big((z-\alpha)(z-\beta) \\ &\quad + (1-\alpha)S(E(z))(z-\alpha-\beta z+\alpha z)\Big)\frac{U_A(\alpha)E(\alpha)}{\alpha S(E(\alpha))} \\ &\quad + \Big((z-\alpha)(z-\beta) + (1-\beta)S(E(z)) \\ &\quad \cdot (z-\beta-\alpha z+\beta z)\Big)\frac{U_B(\beta)E(\beta)}{\beta S(E(\beta))}\Bigg] \\ &\quad - (1-\beta)z\Big(z-\alpha+(1-\alpha)S(E(z))\Big) \\ &\quad \cdot \frac{U_B(\beta)E(\beta)}{\beta S(E(\beta))}\frac{S(E(z))^2}{E(z)} - \frac{U_A(\alpha)E(\alpha)}{\alpha S(E(\alpha))}\frac{S(E(z))^2}{E(z)} \\ &\quad \cdot (1-\alpha)z\Big(z-\beta+(1-\beta)S(E(z))\Big) \quad . \end{aligned} \tag{15}$$

In Eq. (15), the two remaining unknowns $U_A(\alpha)$ and $U_B(\beta)$ still have to be calculated. With the theorem of Rouché, we can easily prove that the common denominator of these partial pgf's, given by the left hand side of Eq. (15), has two zeros inside or on the unit circle. Each zero of the denominator must also be a zero of the numerator since generating functions are analytical functions inside the complex unit disk and bounded for $|z| \leq 1$. We can easily see that $z = 1$ is a zero of the denominator, which leads to the same condition as the normalisation condition. The other zero can be calculated numerically. The condition that the numerator of $U_A(z)$ is equal to zero for the second zero of the denominator, combined with the condition from the normalisation condition, constitutes a set of two linear equations that leads to a unique solution for the two remaining unknowns.

With these results we can also obtain the result of our previous paper, see Baetens et al. 2016, by using a mean service time $\mu = 1$, which corresponds to single-slot service times and $\beta = 1 - \alpha$. Substituting these assumptions in Eq. (13) and Eq. (14) results in

$$\begin{aligned}
&U_A(z)\big((z-\alpha)(z-1+\alpha) - \alpha(1-\alpha)E(z)^2\big)\\
&\quad= \alpha(z-\alpha)\frac{E(z)-E(0)}{1-E(0)}\Big(z-1+\alpha+(1-\alpha)E(z)\Big)\\
&\qquad\cdot\Big(\frac{U_A(\alpha)}{\alpha}+\frac{U_B(1-\alpha)}{1-\alpha}\Big) - (1-\alpha)zE(z)^2U_A(\alpha)\\
&\qquad - \alpha z(z-\alpha)E(z)\frac{U_B(1-\alpha)}{1-\alpha}\ .
\end{aligned}$$

## NUMERICAL RESULTS

In this section we will study the impact of different parameters on the probability $u_I$ that the server is idle, which is given by the sum of $u_{I,A}$ and $u_{I,B}$, and the mean system occupancy. The number of arrivals in each slot follows a geometric distribution with mean arrival rate $\lambda$. In Fig. 1, we show the impact of the mean service time $\mu$ on the probability that the server is idle. In this figure, the probabilities $\alpha$ and $\beta$ are both equal to 0.5 and results are obtained for a number of different mean arrival rates. We observe that for all arrival rates, the probability that the server is idle decreases when the mean service time increases, until it reaches the point that the probability is equal to 0 and the system becomes unstable. A higher value for $\mu$ results in, on average, longer service periods, which means that the probability that there are no arrivals during a service period decreases. The two requirements for the system to become idle after a service is finished are that all customers at service initiation must be of the same class and there are no arrivals during the service period. Because of its impact on the probability of this second requirement, it is clear that an increase in the mean arrival rate leads to a decrease of the probability that the server is idle.

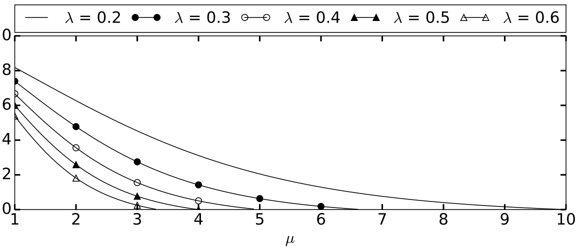

Fig. 1: Impact of Mean Service Time on the Idle Probability using $\alpha = \beta = 0.5$.

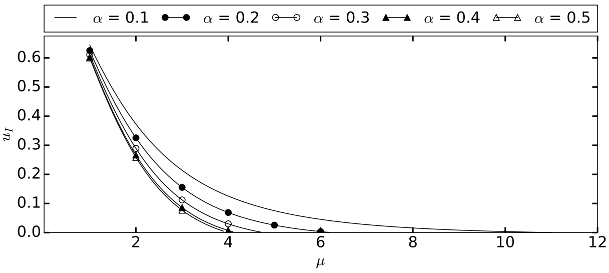

Fig. 2: Impact of Asymmetry in the Arrival Process on the Idle Probability

In Figure 1, we used a symmetric arrival process, that is the probability for a class $A$ and $B$ customer are equal. We can introduce asymmetry in the arrival process by using $\alpha < 0.5$ while keeping $\alpha + \beta = 1$. The impact of a bigger difference between $\alpha$ and $\beta$, or an increasing degree of asymmetry in the arrival process, on the idle probability, is shown in Fig. 2 using the mean arrival rate $\lambda = 0.5$. We note that an increasing degree of asymmetry in the arrival process leads to an increased idle probability for the same value of $\mu$. This is because the mean length of a sequence of class $A$ and $B$ customers increases for values of $\alpha$ closer to 0, which in turn increases the probability that all waiting customers are of the same class. This corresponds with the first requirement for the server to jump from a busy state to an idle state. We also note that more asymmetry in the arrival process allows using a slower server, that is a server with a higher mean service time.

In Fig. 3, we analyse the impact of asymmetry in the arrival process on the system occupancy, for an arrival process with mean arrival rate $\lambda = 0.5$ and $\alpha + \beta = 1$. We clearly see that increasing the degree of asymmetry significantly reduces the number of customers in the system, and allows the server to work more slowly while still being stable. The reason for this is that values of $\alpha$ closer to 0 lead to a higher mean length of a sequence of same-class customers, thus allowing the server to process higher service capacity. If the server can process larger batches, the system occupancy will be reduced and the service time must be higher to have the same mean number of customers in the system. We note that the point at which the system becomes unstable is inversely proportional to the parameter $\alpha$ and $\beta$ as can be seen in the deduction of the stability condition.

In the previous figures, we assumed there was no ten-

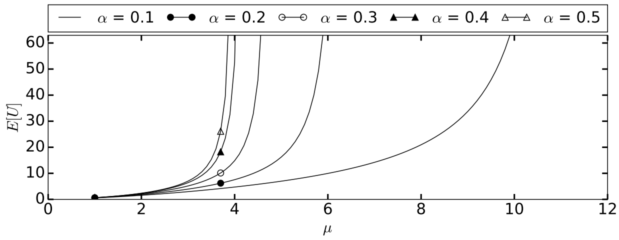

Fig. 3: Impact of Asymmetry in the Arrival Process on the System Occupancy

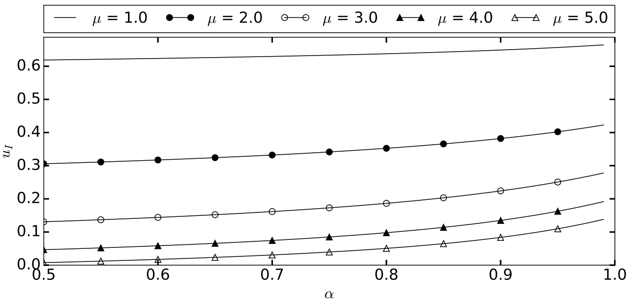

Fig. 4: Impact of Clustering on the Idle Probability

dency for same class customers to arrive in clusters or that $\beta = 1 - \alpha$. In the following results, we study the impact of this tendency for clustering by using values of $\alpha$ and $\beta$ so that $\alpha + \beta > 1$. In Fig. 4, we first analyse the influence of this tendency on the idle probability, by using $\beta = 0.7$, $\alpha$ varying between 0.5 and 1, and a mean arrival rate $\lambda = 0.5$. We clearly see that increasing $\alpha$, or the tendency for clustering, also leads to an increasing idle probability. This occurs because using more clustering in the arrival process means the expected length of sequences of same-class customers increases. This leads to a higher probability that all waiting customers are of the same class, which in turn results in a higher probability that the queue is empty after service initiation.

The influence of this tendency for clustering on the mean system occupancy, for the same system configuration as in the previous figure, is shown in Fig. 5. In case of a small mean service time, e.g. $\mu = 4$, we see that increasing the degree of clustering only has a very small influence on the average number of customers in the queue. On the other hand, for larger values of $\mu$, the system is stable only for a certain degree of clustering, and the more clustering in the system, the lower the mean system occupancy. An increased degree of clustering in the arrival process leads to a higher mean length of a sequence of same-class customers, which means that on average more customers can be processed. This increase in the mean service capacity means that the server processes larger batches, which leads to a lower mean system occupancy.

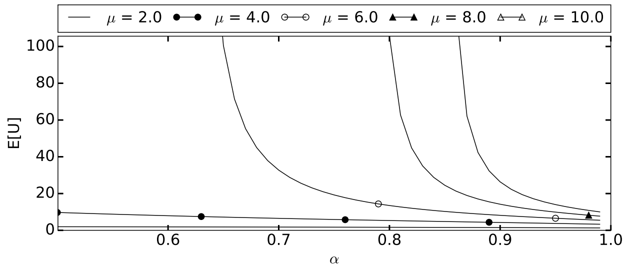

Fig. 5: Impact of Clustering on the System Occupancy

## CONCLUSIONS AND FUTURE RESEARCH

In this paper we have analysed the performance of a discrete-time two-class single-server queueing system with variable capacity batch service. The capacity of the batch server is determined by the length of the sequence of same-class customers at the head of the queue at service initiation. The service times of a batch of either class are geometrically distributed, and we also considered correlation between the classes of consecutive customers. During the analysis, we have derived the steady-state pgf of the number of customers in the system, also called the system occupancy, at random slot boundaries. Using the generating function technique, we have demonstrated the impact of the mean service time, asymmetry and clustering in the arrival process on two performance characteristics, more specifically on the idle probability and on the mean system occupancy.

There are a number of possible extensions that could be considered for this model. A first extension would be to find the steady-state pgf for the number of customers that are being processed by the batch server, and the customer delay. In a second extension we could extend the model to use a class-dependent general service time distribution for class A and B batches. Another possible extension is introducing an upper bound for the service capacity. We can also look at systems capable of processing more than 2 classes of customers. We expect that this will introduce an extra level of complexity because the class of the next batch, if not all customers in the queue at service initiation were of the same class, is not deterministic.

## REFERENCES

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.

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

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.

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

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., 2013. Analysis of a versatile batch-service queueing model with correlation in the arrival process. Performance Evaluation 70(4), 300–316.

Dorsman, J., der Mei, R. V., Winands, E., 2012. Polling with batch service. OR Spectrum 34, 743–761.

Germs, R., Foreest, N. V., 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, N. V., 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.

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. In: Communications, 2005. ICC 2005. 2005 IEEE International Conference on. Vol. 5. IEEE, pp. 2937–2941.

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, 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.

Sikdar, K., Samanta, S., 2016. Analysis of a finite buffer variable batch service queue with batch markovian arrival process and server's vacation. OPSEARCH 53 (3), 553–583.

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.

## AUTHOR BIOGRAPHIES

**JENS BAETENS** obtained his master in Computer Science Engineering at Ghent University in 2014. After graduation, he started his doctorate studies at the Stochastic Modelling and Analysis of Communication Systems (SMACS) research group within the department of Telecommunications and Information Processing (TELIN) at Ghent University. His main research interests include discrete-time batch service models and its applications.

**BART STEYAERT** received his Phd in Engineering Sciences in 2008 from Ghent University, Belgium. Since January 1990, he has been working as a researcher at the SMACS Research Group within the TELIN-Department at the same university. His main research interests include discrete-time queueing models, traffic control, and stochastic modelling of high-speed communications networks.

**DIETER CLAEYS** obtained his Ph.D. degree in Engineering in 2011 and has since worked at the SMACS research group at Ghent University. Since October 2015, he is an assistant professor at the department of Industrial Systems Engineering and Product Design at Ghent University. His main research interests include methods and time engineering, and the analysis of discrete-time queueing models and its applications to operations research and telecommunications systems.

**HERWIG BRUNEEL** is full Professor in the Faculty of Engineering and head of the TELIN-Department at Ghent University. He also leads the SMACS Research Group within this department. His main personal research interests include stochastic modeling and analysis of communication systems, and (discrete-time) queueing theory. He has published more than 500 papers on these subjects and is coauthor of the book H. Bruneel and B. G. Kim, "Discrete-Time Models for Communication Systems Including ATM" (Kluwer Academic Publishers, Boston, 1993). From October 2001 to September 2003, he served as the Academic Director for Research Affairs at Ghent University. Since 2009, he holds a career-long Methusalem grant from the Flemish Government at Ghent University, specifically on Stochastic Modeling and Analysis of Communication Systems.

## ACKNOWLEDGEMENT

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.