System occupancy in a multiclass batch-service queueing system with limited variable service capacity

Jens Baetens, Bart Steyaert, Dieter Claeys & Herwig Bruneel

Annals of Operations Research

ISSN 0254-5330

Ann Oper Res
DOI 10.1007/s10479-019-03470-1

Springer

Annals of Operations Research
https://doi.org/10.1007/s10479-019-03470-1

S.I.: QUEUEING THEORY AND NETWORK APPLICATIONS II

System occupancy in a multiclass batch-service queueing system with limited variable service capacity

Jens Baetens[1] · Bart Steyaert[1] · Dieter Claeys[2,3] · Herwig Bruneel[1]

Abstract
In multi-class telecommunications or manufacturing systems, customers belonging to the same class can often be processed together. This results in a service capacity that depends on the classes of the customers in the queue. In this paper, we analyse a discrete-time batch-service queue with two customer classes. The single batch server can group all same-class customers at the head of the queue up to a constant class-dependent maximum service capacity. We focus on the analysis of the system occupancy at service initiation opportunities, and also compute both a light- and heavy traffic approximation in order to reduce the numerical complexity introduced by the maximum service capacities. Additionally, we propose a method for interpolating between these approximations in order to study the behaviour in the intermediate region. We also deduce the system occupancy and its approximations at random slot boundaries. In the numerical experiments, we examine the conditions under which these proposed approximations are accurate.

Keywords Batch service · Two-class · Variable service capacity · Generally distributed service times · Correlated customer types

1 Introduction

Batch-service queueing systems are often found in manufacturing (Niranjan et al. 2017), transportation systems (Bountali and Economou 2017), and telecommunication systems (Bellalta and Oliver 2009) where packets are grouped together based on similarities in the production process or destination. Due to their wide range of applications, this type of queueing system has been studied extensively, for instance, by Chaudhry and Templeton (1983), Arumuganathan and Jeyakumar (2005), Banerjee and Gupta (2012), Banerjee et al. (2015),

✉ Jens Baetens
jens.baetens@ugent.be

1 Department of Telecommunications and Information Processing, SMACS Research Group, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

2 Department of Industrial Systems Engineering, Ghent University, Technologiepark 903, 9052 Zwijnaarde, Belgium

3 Industrial Systems Engineering (ISyE), Flanders Make, Lommel, Belgium

Published online: 19 November 2019

Banerjee et al. (2014), Chang and Takine (2005), Claeys et al. (2013a), Claeys et al. (2013b), Claeys et al. (2012), Goswami et al. (2006), Janssen and van Leeuwaarden (2005), Pradhan and Gupta (2017). While the capacity of the batch server is assumed to be constant in these contributions, this service capacity often depends on the environment and the content of the queue. An example of a model with variable service capacity has been studied by Chaudhry and Chang (2004) where they analysed the system content in the $Geo/G^Y/1/N+B$ model . Server vacations were also incorporated in the previous model by Chang and Choi (2005). Yi et al. (2007) further extended this model by using the general bulk service rule. In the previously mentioned papers with a stochastic service capacity, the service capacity was independent of the state of the system or the contents of the queue. Germs and Foreest (2013) evaluated the $M(n)^{X(n)}/G(n)^{Y(n)}/1/K+B$ batch-service queueing system, where the variable service capacity as well as the arrival process and service time distributions depend on the number of customers in the queue.

In the previously mentioned papers, the batch server did not distinguish between different customer classes and processed all customers identically. However, in many applications the same server is often capable of processing multiple classes, such as a machine that can handle multiple types of products with slightly different parameters or a transportation systems that can transmit customers to different destinations. The combination of batch-service and customer differentiation has been mostly studied in the context of priority queueing systems and polling systems. Vardakas and Logothetis (2009) study the delay of a packet in priority-based optical networks using k customer classes. The optimal scheduling policies in a polling system with switch-in and switch-out times and unbounded service capacities have been examined by Van Der Wal and Yechiali (2003). Boxma et al. (2008) analysed a similar model with a Poisson service process and looked at the impact of different service disciplines. Dorsman et al. (2012) studied a polling system with an inner and outer layer. The batches are formed in accumulation stations inside the outer layer before being passed to the inner part. Priority queueing and polling systems both rely on customer reordering or different queues for different customer classes in order to be able to construct groups of customers. This is however not always feasible due to certain requirements of the system, such as strict fairness rules that do not allow packets to skip ahead of other packets. In our model, we guarantee these rules by using a shared queue for packets of different classes that does not allow reordering of customers. The use of such a queue results in a global First-Come-First-Served(FCFS) service discipline, which is also described in Bruneel et al. (2012).

In this paper, we study the system occupancy of a two-class discrete-time batch-service queueing system where the batch server can group all consecutive same-class customers at the head of the queue until a certain class-dependent maximum capacity is reached. This means that if the customer at the head of the queue is of class A, all of the following class A customers are also processed in the same batch until either the next customer is of class B, or the maximum capacity for a class A batch is reached, or the queue is empty, whichever occurs first. This means that the size of a batch, also called the service capacity, is variable and depends on the classes of the waiting customers. Since the length of a sequence of same-class customers will have a significant impact on the performance of the system, we also include a tendency for clustering of same-class customers. This tendency for clustering, also called customer-based correlation, has been described in more detail by Bruneel et al. (2012) as well. A simplified model without maximum service capacities has been studied by Baetens et al. (2016, 2017, 2018a, b). The main contributions of this paper are the inclusion of maximum service capacities in order to model a much more realistic model and the approximations at low, high or intermediate loads (which is obtained by interpolating the first two). We propose these approximations in order to reduce the numerical complexity,

which can be very high since it depends on the maximum service capacities. We restrict the number of distinct customer classes to 2 in order to clearly observe the impact of the maximum service capacities on the complexity of the analysis. Note that an extension to a general number of classes will significantly further increase the complexity of the analysis, and therefore falls beyond the scope of the current paper. Applications of this type of batch-service queueing system can be found in telecommunication systems that require the use of strict fairness rules which can be guaranteed by using a single shared queue, or Lean manufacturing systems where goods flow in a FIFO fashion through downstream process.

In Sect. 2, we start by giving a more detailed description of the discrete-time two-class batch-service queueing system. The analysis in Sect. 3 focuses on the steady-state pgf of the system occupancy at service initiation opportunities and random slot boundaries. An important part of the analysis is the proof of the number of zeroes of the denominator. This number is equal to the sum of the maximum service capacities and can therefore become computationally expensive. In order to remedy this, we also present light-traffic and heavy-traffic approximations. Finally, we will also propose an interpolation of these two in order to study the behaviour of the system under a moderate load. Some numerical experiments are presented in Sect. 4 to evaluate the impact of different parameters on the behaviour of the system and evaluate the accuracy of the proposed approximations. We end by drawing some conclusions in Sect. 5.

2 Model description

The queueing system considered in this paper consists of a single batch server that can process two classes of customers (called class A and B), and a single queue of infinite size. Newly arrived customers are added to the tail of the common queue with a global FCFS service discipline, meaning that no overtaking is allowed. When the server is or becomes available and finds at least one waiting customer in the queue, a new service is initiated immediately. The batch server can then group all consecutive customers at the head of the queue that belong to the same class, until either the end of the queue is reached, or the next customer is of a different type. The size of a batch is furthermore bound by a constant class-dependent value, denoted by C_A or C_B for a batch of class A or B customers respectively. This results in a stochastic service capacity that depends both on the number of customers in the queue and their respective classes. The service time of a batch follows a generic class-dependent distribution, with pgfs $S_A(z)$ and $S_B(z)$ respectively for a class A or B batch. We note that the service time only depends on the class of the batch and is independent of the number of customers in the batch and of the class and size of previous batches. Such a service process corresponds, for instance, with a furnace where the heating period is not influenced by the occupancy of the furnace. Also the time necessary for forming the groups is assumed to be negligible compared to the time spent in service.

The total number of arrivals in the system during a single slot is independent from slot-to-slot, and is characterized by the probability mass function (pmf) $e(n)$ and pgf $E(z)$. The mean number of customer arrivals in a single slot is denoted by the parameter $\lambda = E'(1)$. Customers arriving in the system can either be of class A or B, and the probability that a random customer belongs to either class depends on the class of its predecessor. Since the batch server can group consecutive customers of the same class, the probabilities that two consecutive customers are of the same class will have a significant impact on the performance of the system. These probabilities can be used to incorporate a tendency for clustering which

in practice often occurs, for instance due to presorting of customer orders. The probabilities that a random customer is a class A or B customer if its predecessor was respectively also a class A or B customer are denoted by α and β. Hence, a random customer is of class A with probability (w.p.) $\frac{1-\beta}{2-\alpha-\beta}$.

This type of model can, for instance, be used to study the performance of postal distribution centres where a sorter can pick items from the front of a conveyor and sort them according to their destination area. Consecutive letters with the same destination can be sorted simultaneously, and the sorting time of such a group is only slightly sensitive to the number of letters that are grouped together, since the most significant part of the processing time stems from moving the items to their corresponding box which is independent of the size of the group. Other examples in manufacturing can be found in a paint booth where multiple products can be painted together, or a furnace where items can be heated simultaneously.

3 Analysis

In this section, we aim to calculate the pgf of the system occupancy at service initiation opportunities. These time instants correspond with the slot boundaries of idle slots and of slots in which a service is initiated. We chose these time instances in order to simplify the analysis, because examining the system directly at random slot boundaries introduces a number of complications arising from the generic service times. It will be clear that finding a unique solution for the remaining unknowns in the pgf of the system occupancy at service initiation opportunities can become computationally expensive. For this reason, we also propose three methods of approximating the system.

3.1 System equations

In this part of the analysis, we will capture the behaviour of the system at consecutive time instants in a set of equations. In this system, a batch is formed by grouping all consecutive same-class customers at the head of the queue, until the maximum, denoted by C_A or C_B respectively for class A or B batches, or a customer of the other class is reached. The number of customers in the queue at the boundary of the k-th service initiation opportunity is denoted by the random variable u_k. We can differentiate between four different server states, based on the class of the ongoing batch, or the previous batch in case the server is idle. The random variables for the system occupancy at the k-th initiation opportunity when the server was busy with a class A or B batch, are denoted by $u_{A,k}$ or $u_{B,k}$ respectively. On the other hand, the random variables $u_{I,A,k}$ and $u_{I,B,k}$ represent the system occupancy in case that the server is idle during the k-th initiation opportunity and the previous processed batch was of class A or B. It is important to keep track of the previous batch in case the server becomes idle, because the probability that the next arrival is of a certain class depends on the class of the last customer in the previous batch, which is equal to the class of the batch.

This leads to the following system equations in case that the system was idle in the slot following the k-th service initiation opportunity and the previous batch contained class A customers.

$$\begin{aligned} u_{I,A,k+1} &= 0, \text{ if } e_k = 0 \\ u_{A,k+1} &= e_k, \text{ if } e_k > 0 \,\&\, \text{next customer of class } A (\text{w.p. } \alpha) \\ u_{B,k+1} &= e_k, \text{ if } e_k > 0 \,\&\, \text{next customer of class } B (\text{w.p. } 1-\alpha), \end{aligned} \tag{1}$$

where e_k is the random variable of the number of arrivals during the slot following the k-th service period. Analogously, in case the previous batch was of class B

$$\begin{aligned} u_{I,B,k+1} &= 0, \text{ if } e_k = 0 \\ u_{A,k+1} &= e_k, \text{ if } e_k > 0 \,\&\, \text{next customer of class } A (\text{w.p. } 1-\beta) \\ u_{B,k+1} &= e_k, \text{ if } e_k > 0 \,\&\, \text{next customer of class } B (\text{w.p. } \beta). \end{aligned} \tag{2}$$

However, if the server initiated a class A service at the k-th service initiation opportunity, then we can differentiate between three behaviours. First, if the system occupancy $u_{A,k}$ is smaller than or equal to C_A and all customers in the system belong to the same class, then the system will become idle after service completion if there were no arrivals during the service of class A, or, in case that there was at least one arrival, start a new class A or B batch. The class of this batch is determined by the class of the first arrival. The condition that all customers in the system belong to the same class corresponds with the condition that the service capacity of the service initiated in the k-th service initiation instant, denoted by c_k, must be equal to $u_{A,k}$. This behaviour can be seen in the first three lines of Eq. (3). On the other hand, if the system occupancy $u_{A,k}$ at the k-th service initiation opportunity is larger than C_A and the first C_A customer all are class A customers, then the capacity of the initiated batch c_k is equal to C_A, which is the maximum batch size. When this occurs, in the next service initiation opportunity, a new service will be initiated since there are still waiting customers, and the class of the service depends on the class of the (C_A+1)-th customer in the system at the k-th service initiation opportunity. The system equations that correspond to this behaviour can be seen in the fourth and fifth line of Eq. (3). If neither of the above conditions are true, then it is certain that not all customers in the system belong to the same class nor is the maximum service capacity C_A reached of the service initiated in the k-th service initiation opportunity. In this case, a new service will be initiated in the $k+1$-th initiation opportunity and its class must be of the other type (otherwise the first customer in the system would be taken in the previous batch). Using $e_{A,k}$ as the number of arrivals during the service period of a class A batch, this results in

$$\begin{aligned} u_{I,A,k+1} &= 0, \text{ if } c_k = u_{A,k} \,\&\, e_{A,k} = 0 \\ u_{A,k+1} &= e_{A,k}, \text{ if } c_k = u_{A,k} \,\&\, e_{A,k} > 0 \,\&\, \text{next customer of class } A (\text{w.p. } \alpha) \\ u_{B,k+1} &= e_{A,k}, \text{ if } c_k = u_{A,k} \,\&\, e_{A,k} > 0 \,\&\, \text{next customer of class } B (\text{w.p. } 1-\alpha) \\ u_{A,k+1} &= u_{A,k} - C_A + e_{A,k}, \text{ if } c_k = C_A \,\&\, u_{A,k} > C_A \,\&\, \text{next customer of class } A (\text{w.p. } \alpha) \\ u_{B,k+1} &= u_{A,k} - C_A + e_{A,k}, \text{ if } c_k = C_A \,\&\, u_{A,k} > C_A \,\&\, \text{next customer of class } B (\text{w.p. } 1-\alpha) \\ u_{B,k+1} &= u_{A,k} - c_k + e_{A,k}, \text{ if } c_k < \min(C_A, u_{A,k}), \end{aligned} \tag{3}$$

and analogously if a class B service was initiated at the k-th service initiation opportunity

$$\begin{aligned} u_{I,B,k+1} &= 0, \text{ if } c_k = u_{B,k} \,\&\, e_{B,k} = 0 \\ u_{B,k+1} &= e_{B,k}, \text{ if } c_k = u_{B,k} \,\&\, e_{B,k} > 0 \,\&\, \text{next customer of class } B (\text{w.p. } \beta) \\ u_{A,k+1} &= e_{B,k}, \text{ if } c_k = u_{B,k} \,\&\, e_{B,k} > 0 \,\&\, \text{next customer of class } A (\text{w.p. } 1-\beta) \\ u_{B,k+1} &= u_{B,k} - C_B + e_{B,k}, \text{ if } c_k = C_B \,\&\, u_{B,k} > C_B \,\&\, \text{next customer of class } B (\text{w.p. } \beta) \\ u_{A,k+1} &= u_{B,k} - C_B + e_{B,k}, \text{ if } c_k = C_B \,\&\, u_{B,k} > C_B \,\&\, \text{next customer of class } A (\text{w.p. } 1-\beta) \\ u_{A,k+1} &= u_{B,k} - C_B + e_{B,k}, \text{ if } c_k < \min(C_B, u_{B,k}), \end{aligned} \tag{4}$$

where $e_{B,k}$ is the random variable of the number of arrivals during service period of the class B batch initiated at the k-th service period.

3.2 Stability condition

The steady-state pgfs that we will obtain during the analysis in the next section, will only be valid as long as the system is stable. In order to find the condition under which this is the case, we study a saturated system, that is a system where the service capacity is never limited by the number of waiting customers. In such a system, the server is never idle and the mean number of customers that arrive during the service time of a random batch must be smaller than the mean number of customers that are processed during the same batch. The first step is to find the probabilities that a random batch is either of class A or B, denoted by $Pr[\tau = A]$ and $Pr[\tau = B]$ respectively. The probability that the k-th batch, denoted by $Pr[\tau_k = A]$, contains class A customers is given by

$$Pr[\tau_k = A] := \alpha^{C_A} Pr[\tau_{k-1} = A] + (1 - \beta^{C_B}) Pr[\tau_{k-1} = B],$$

and an analogous equation can be found for the probability $Pr[\tau_k = B]$ that the k-th batch is of class B. By letting k approach infinity and using the fact that $Pr[\tau = A] + Pr[\tau = B] = 1$, we obtain the following probabilities

$$Pr[\tau = A] := \frac{1 - \beta^{C_B}}{2 - \alpha^{C_A} - \beta^{C_B}},\ Pr[\tau = B] := \frac{1 - \alpha^{C_A}}{2 - \alpha^{C_A} - \beta^{C_B}}. \tag{5}$$

We note that if both C_A and C_B approach infinity, then these probabilities approach the value of 0.5 which means that the class of successive batches alternates.

The mean number of customers that can be processed during a single batch is given by the mean length of a sequence of same-class customers which resembles a modified geometric distribution. For a batch of class A customers, the pgf and expected value of the service capacity are equal to

$$\hat{C}_A(z) := \frac{(1 - \alpha)z + (1 - z)(\alpha z)^{C_A}}{1 - \alpha z},\ \hat{C}'_A(1) = \frac{1 - \alpha^{C_A}}{1 - \alpha}.$$

This results in the following stability condition

$$\begin{aligned} \lambda &< \frac{Pr[\tau = A]\frac{1-\alpha^{C_A}}{1-\alpha} + Pr[\tau = B]\frac{1-\beta^{C_B}}{1-\beta}}{Pr[\tau = A]S'_A(1) + Pr[\tau = B]S'_B(1)} \\ &< \frac{(1 - \beta^{C_B})(1 - \alpha^{C_A})}{(1 - \beta^{C_B})S'_A(1) + (1 - \alpha^{C_A})S'_B(1)} \left(\frac{1}{1 - \alpha} + \frac{1}{1 - \beta} \right). \end{aligned} \tag{6}$$

The load ρ of the system is then equal to

$$\rho = \lambda \frac{(1 - \beta^{C_B})S'_A(1) + (1 - \alpha^{C_A})S'_B(1)}{(1 - \beta^{C_B})(1 - \alpha^{C_A})} \frac{(1 - \alpha)(1 - \beta)}{2 - \alpha - \beta}.$$

3.3 System occupancy at service initiation opportunities

In this part, we will calculate the pgf of the system occupancy at service initiation opportunities, denoted by $U(z)$. The probability mass function (pmf) corresponding to the pgf $U(z)$ is denoted by $u(n)$. As mentioned earlier in Sect. 3.1, we distinguish 4 different server states. We define $U_{I,A}$ and $U_{I,B}$ as the probabilities that the system is idle and the previous batch was a class A or B batch. The partial pgfs of the system occupancy at a random service initiation opportunity in which a class A or B batch service is initiated, are denoted by $U_A(z)$ and $U_B(z)$ (with the pmfs $u_A(n)$ and $u_B(n)$). We can then write $U_A(z)$, $U_B(z)$, $U(z)$ as

$$U_A(z) = \sum_{n=1}^{\infty} u_A(n) z^n$$
$$U_B(z) = \sum_{n=1}^{\infty} u_B(n) z^n$$
$$U(z) = \sum_{n=0}^{\infty} u(n) z^n = U_{I,A} + U_{I,B} + U_A(z) + U_B(z). \tag{7}$$

From Eqs. (1)–(4), we derive that the system remains idle if there are no arrivals and can only become idle if all waiting customers could be grouped together at service initiation and no customers have arrived during the service period. This leads to the following probabilities

$$U_{I,A} = \frac{S_A(E(0))}{1 - E(0)} \sum_{i=1}^{C_A} u_A(i) \alpha^{i-1}, \; U_{I,B} = \frac{S_B(E(0))}{1 - E(0)} \sum_{i=1}^{C_B} u_B(i) \beta^{i-1}, \tag{8}$$

where $u_A(i)$ and $u_B(i)$ are the probabilities that there are i customers in the system at a random service initiation opportunity in which, respectively, a class A or B batch is initiated.

On the other hand, if the server can initiate a class A batch, then by using the system equations from Sect. 3.1, we can obtain the following expression for $U_A(z)$

$$\begin{aligned} U_A(z) = {} & \alpha U_{I,A} \sum_{i=1}^{\infty} e(i) z^i + (1-\beta) U_{I,B} \sum_{i=1}^{\infty} e(i) z^i + \sum_{i=1}^{C_A} \sum_{j=1}^{\infty} \alpha^i u_A(i) e_A(j) z^j \\ & + \sum_{i=C_A+1}^{\infty} \sum_{j=0}^{\infty} \alpha^{C_A} u_A(i) e_A(j) z^{i-C_A+j} + \sum_{i=1}^{C_B} \sum_{j=1}^{\infty} (1-\beta) \beta^{i-1} u_B(i) e_B(j) z^j \\ & + \sum_{i=C_B+1}^{\infty} \sum_{j=0}^{\infty} (1-\beta) \beta^{C_B-1} u_B(i) e_B(j) z^{i-C_B+j} \\ & + \sum_{i=2}^{C_B} \sum_{j=0}^{\infty} \sum_{n=1}^{i-1} (1-\beta) \beta^{n-1} u_B(i) e_B(j) z^{i-n+j} \\ & + \sum_{i=C_B+1}^{\infty} \sum_{j=0}^{\infty} \sum_{n=1}^{C_B-1} (1-\beta) \beta^{n-1} u_B(i) e_B(j) z^{i-n+j}. \end{aligned}$$

Working out the sums for the number of arrivals and the sizes of the processed batches results in

$$\begin{aligned} U_A(z) = {} & \left(S_A(E(z)) - S_A(E(0)) \frac{1 - E(z)}{1 - E(0)} \right) \sum_{i=1}^{C_A} u_A(i) \alpha^i \\ & + S_A(E(z)) \left(\frac{\alpha}{z} \right)^{C_A} \left(U_A(z) - \sum_{i=1}^{C_A} u_A(i) z^i \right) + \frac{1-\beta}{\beta - z} S_B(E(z)) \sum_{i=1}^{C_B} u_B(i) \beta^i \\ & - \frac{1-\beta}{\beta} S_B(E(0)) \frac{1 - E(z)}{1 - E(0)} \sum_{i=1}^{C_B} u_B(i) \beta^i - \frac{1-\beta}{\beta - z} S_B(E(z)) \sum_{i=1}^{C_B} u_B(i) z^i \\ & - \frac{1-\beta}{\beta - z} S_B(E(z)) \left(1 - \left(\frac{\beta}{z} \right)^{C_B} \right) \left(U_B(z) - \sum_{i=1}^{C_B} u_B(i) z^i \right). \end{aligned} \tag{9}$$

Analogously, we obtain that the similar expression for the partial pgf $U_B(z)$ is given by

$$\begin{aligned}U_B(z) &= \left(S_B(E(z)) - S_B(E(0))\frac{1-E(z)}{1-E(0)}\right)\sum_{i=1}^{C_B} u_B(i)\beta^i \\ &\quad + S_B(E(z))\left(\frac{\beta}{z}\right)^{C_B}\left(U_B(z) - \sum_{i=1}^{C_B} u_B(i)z^i\right) + \frac{1-\alpha}{\alpha-z}S_A(E(z))\sum_{i=1}^{C_A} u_A(i)\alpha^i \\ &\quad - \frac{1-\alpha}{\alpha}S_A(E(0))\frac{1-E(z)}{1-E(0)}\sum_{i=1}^{C_A} u_A(i)\alpha^i - \frac{1-\alpha}{\alpha-z}S_A(E(z))\sum_{i=1}^{C_A} u_A(i)z^i \\ &\quad - \frac{1-\alpha}{\alpha-z}S_A(E(z))\left(1-\left(\frac{\alpha}{z}\right)^{C_A}\right)\left(U_A(z) - \sum_{i=1}^{C_A} u_A(i)z^i\right). \end{aligned} \tag{10}$$

We can now introduce the auxiliary functions $R_A(z)$ and $R_B(z)$, which are respectively the right-hand side of the Eqs. (9) and (10), without the terms of the partial pgfs $U_A(z)$ and $U_B(z)$, multiplied respectively by $z^{C_A}(\beta - z)$ and $z^{C_B}(\alpha - z)$. These auxiliary functions are then given by

$$\begin{aligned}R_A(z) &= z^{C_A}(\beta-z)\left(S_A(E(z)) - S_A(E(0))\frac{1-E(z)}{1-E(0)}\right)\sum_{i=1}^{C_A} u_A(i)\alpha^i \\ &\quad - (\beta-z)S_A(E(z))\alpha^{C_A}\sum_{i=1}^{C_A} u_A(i)z^i + (1-\beta)z^{C_A}S_B(E(z))\sum_{i=1}^{C_B} u_B(i)\beta^i \\ &\quad - \frac{1-\beta}{\beta}(\beta-z)z^{C_A}S_B(E(0))\frac{1-E(z)}{1-E(0)}\sum_{i=1}^{C_B} u_B(i)\beta^i \\ &\quad - (1-\beta)\beta^{C_B}z^{C_A-C_B}S_B(E(z))\sum_{i=1}^{C_B} u_B(i)z^i, \end{aligned} \tag{11}$$

and

$$\begin{aligned}R_B(z) &= z^{C_B}(\alpha-z)\left(S_B(E(z)) - S_B(E(0))\frac{1-E(z)}{1-E(0)}\right)\sum_{i=1}^{C_B} u_B(i)\beta^i \\ &\quad - (\alpha-z)S_B(E(z))\beta^{C_B}\sum_{i=1}^{C_B} u_B(i)z^i + (1-\alpha)z^{C_B}S_A(E(z))\sum_{i=1}^{C_A} u_A(i)\alpha^i \\ &\quad - \frac{1-\alpha}{\alpha}(\alpha-z)z^{C_B}S_A(E(0))\frac{1-E(z)}{1-E(0)}\sum_{i=1}^{C_A} u_A(i)\alpha^i \\ &\quad - (1-\alpha)\alpha^{C_A}z^{C_B-C_A}S_A(E(z))\sum_{i=1}^{C_A} u_A(i)z^i. \end{aligned}$$

The next step is to substitute the partial pgf $U_B(z)$ in $U_A(z)$ and to multiply the right- and left-hand side of the expression by $z^{C_A+C_B}(\alpha - z)(\beta - z)$. With the previous definitions of $R_A(z)$ and $R_B(z)$, and moving all terms of $U_A(z)$ to the left-hand side, we obtain the following equation for $U_A(z)$

$$U_A(z)\Big[(\alpha-z)(\beta-z)\left[z^{C_A}-\alpha^{C_A}S_A(E(z))\right]\left[z^{C_B}-\beta^{C_B}S_B(E(z))\right]$$
$$-(1-\alpha)(1-\beta)S_A(E(z))S_B(E(z))\left(z^{C_A}-\alpha^{C_A}\right)\left(z^{C_B}-\beta^{C_B}\right)\Big]$$
$$=(\alpha-z)\left[z^{C_B}-\beta^{C_B}S_B(E(z))\right]R_A(z)-(1-\beta)S_B(E(z))\left(z^{C_B}-\beta^{C_B}\right)z^{C_A-C_B}R_B(z). \tag{12}$$

For the partial pgf of the system occupancy in a random service initiation opportunity in which the server has started a class B service, we obtain the analogous equation

$$U_B(z)\Big[(\alpha-z)(\beta-z)\left[z^{C_A}-\alpha^{C_A}S_A(E(z))\right]\left[z^{C_B}-\beta^{C_B}S_B(E(z))\right]$$
$$-(1-\alpha)(1-\beta)S_A(E(z))S_B(E(z))\left(z^{C_A}-\alpha^{C_A}\right)\left(z^{C_B}-\beta^{C_B}\right)\Big]$$
$$=(\beta-z)\left[z^{C_A}-\alpha^{C_A}S_A(E(z))\right]R_B(z)-(1-\alpha)S_A(E(z))\left(z^{C_A}-\alpha^{C_A}\right)z^{C_B-C_A}R_A(z). \tag{13}$$

By combining Eqs. (8), (12) and (13), we obtain the pgf $U(z)$ of the system occupancy at random service initiation opportunities. This leads to

$$\begin{aligned}
U(z)=&\frac{S_A(E(0))}{1-E(0)}\sum_{i=1}^{C_A}u_A(i)\alpha^{i-1}+\frac{S_B(E(0))}{1-E(0)}\sum_{i=1}^{C_B}u_B(i)\beta^{i-1}\\
&+\Big[(\alpha-z)(\beta-z)\left[z^{C_A}-\alpha^{C_A}S_A(E(z))\right]\left[z^{C_B}-\beta^{C_B}S_B(E(z))\right]\\
&-(1-\alpha)(1-\beta)S_A(E(z))S_B(E(z))\left(z^{C_A}-\alpha^{C_A}\right)\left(z^{C_B}-\beta^{C_B}\right)\Big]^{-1}\\
&\cdot\Big[\Big((\alpha-z)\left[z^{C_B}-\beta^{C_B}S_B(E(z))\right]-(1-\alpha)S_A(E(z))\left(z^{C_A}-\alpha^{C_A}\right)z^{C_B-C_A}\Big)\\
&\cdot\Big[z^{C_A}(\beta-z)\left(S_A(E(z))-S_A(E(0))\frac{1-E(z)}{1-E(0)}\right)\sum_{i=1}^{C_A}u_A(i)\alpha^i\\
&-(\beta-z)S_A(E(z))\alpha^{C_A}\sum_{i=1}^{C_A}u_A(i)z^i+(1-\beta)z^{C_A}S_B(E(z))\sum_{i=1}^{C_B}u_B(i)\beta^i\\
&-\frac{1-\beta}{\beta}(\beta-z)z^{C_A}S_B(E(0))\frac{1-E(z)}{1-E(0)}\sum_{i=1}^{C_B}u_B(i)\beta^i\\
&-(1-\beta)\beta^{C_B}z^{C_A-C_B}S_B(E(z))\sum_{i=1}^{C_B}u_B(i)z^i\Big]\\
&+\Big((\beta-z)\left[z^{C_A}-\alpha^{C_A}S_A(E(z))\right]-(1-\beta)S_B(E(z))\left(z^{C_B}-\beta^{C_B}\right)z^{C_A-C_B}\Big)\\
&\cdot\Big[z^{C_B}(\alpha-z)\left(S_B(E(z))-S_B(E(0))\frac{1-E(z)}{1-E(0)}\right)\sum_{i=1}^{C_B}u_B(i)\beta^i
\end{aligned}$$

$$- (\alpha - z)S_B(E(z))\beta^{C_B}\sum_{i=1}^{C_B} u_B(i)z^i + (1-\alpha)z^{C_B}S_A(E(z))\sum_{i=1}^{C_A} u_A(i)\alpha^i$$
$$- \frac{1-\alpha}{\alpha}(\alpha - z)z^{C_B}S_A(E(0))\frac{1-E(z)}{1-E(0)}\sum_{i=1}^{C_A} u_A(i)\alpha^i$$
$$- (1-\alpha)\alpha^{C_A}z^{C_B-C_A}S_A(E(z))\sum_{i=1}^{C_A} u_A(i)z^i\Bigg]\Bigg],$$

or by using the auxiliary functions $R_A(z)$ and $R_B(z)$

$$U(z) = \frac{S_A(E(0))}{1-E(0)}\sum_{i=1}^{C_A} u_A(i)\alpha^{i-1} + \frac{S_B(E(0))}{1-E(0)}\sum_{i=1}^{C_B} u_B(i)\beta^{i-1}$$
$$+ \Bigg[(\alpha - z)(\beta - z)\left[z^{C_A} - \alpha^{C_A}S_A(E(z))\right]\left[z^{C_B} - \beta^{C_B}S_B(E(z))\right]$$
$$- (1-\alpha)(1-\beta)S_A(E(z))S_B(E(z))\left(z^{C_A} - \alpha^{C_A}\right)\left(z^{C_B} - \beta^{C_B}\right)\Bigg]^{-1}$$
$$\cdot \Bigg[\left((\alpha - z)\left[z^{C_B} - \beta^{C_B}S_B(E(z))\right] - (1-\alpha)S_A(E(z))\left(z^{C_A} - \alpha^{C_A}\right)z^{C_B-C_A}\right)R_A(z)$$
$$+ \left((\beta - z)\left[z^{C_A} - \alpha^{C_A}S_A(E(z))\right] - (1-\beta)S_B(E(z))\left(z^{C_B} - \beta^{C_B}\right)z^{C_A-C_B}\right)R_B(z)\Bigg].$$

In this expression, there still are $C_A + C_B$ unknowns, namely $u_A(i)$, $1 \le i \le C_A$, and $u_B(j)$, $1 \le i \le C_A$. These parameters correspond with the probabilities that there are i or j customers in the system when respectively a class A or B service has been initiated. By using the theorem of Rouché, see Adan et al. (2006), we can prove that the denominator of $U(z)$ has $C_A + C_B + 2$ zeroes inside or on the unit circle. We first define the following functions

$$f(z) = (z-\alpha)(z-\beta)\left(z^{C_A} - \alpha^{C_A}S_A(E(z))\right)\left(z^{C_B} - \beta^{C_B}S_B(E(z))\right),$$
$$g(z) = (1-\alpha)(1-\beta)S_A(E(z))S_B(E(z))\left(z^{C_B} - \beta^{C_B}\right)\left(z^{C_A} - \alpha^{C_A}\right)$$
$$h(z) = \frac{g(z)}{f(z)} = G_A\left(\frac{S_A(E(z))}{z^{C_A}}\right)T_A(z)G_B\left(\frac{S_B(E(z))}{z^{C_B}}\right)T_B(z),$$

where

$$G_A(z) := \frac{(1-\alpha^{C_A})z}{1-\alpha^{C_A}z}, T_A(z) := \frac{1-\alpha}{1-\alpha^{C_A}}\frac{z^{C_A} - \alpha^{C_A}}{z-\alpha},$$

and analogously expressions for $G_B(z)$ and $T_B(z)$. The function $G_A(z)$ corresponds with a shifted geometric pgf with mean $(1-\alpha^{C_A})^{-1}$ and $T_A(z)$ also corresponds with a pgf (deterministic for $\alpha = 0$, uniform for $\alpha = 1$ and otherwise a truncated geometric with mean $C_A(1-\alpha^{C_A})^{-1} - (1-\alpha)^{-1}$). We know that a pgf $X(z)$ satisfies the inequality $|X(z)| \le 1 + \epsilon X'(1) + O(\epsilon^2)$ on the contour $|z| = 1 + \epsilon, \epsilon > 0$. Although $G_A(z)\left(\frac{S_A(E(z))}{z^{C_A}}\right)$ is clearly not a pgf, by writing it as a series with the probabilities $g_A(n)$ and considering that $\left(\frac{S_A(E(z))}{z^{C_A}}\right)^n$ is upper bounded by $1 + n\epsilon(S_A'(1)E'(1) - C_A)$ for $|z| = 1 + \epsilon$ we can easily prove that

$$\left|G_A\left(\frac{S_A(E(z))}{z^{C_A}}\right)\right| \leq \sum_{n=0}^{\infty} g_A(n)\left|\left(\frac{S_A(E(z))}{z^{C_A}}\right)^n\right|$$
$$\leq \sum_{n=0}^{\infty} g_A(n)\Big(1+n\epsilon(S'_A(1)E'(1)-C_A)\Big)$$
$$\leq 1+\epsilon G'_A(1)\big(S'_A(1)E'(1)-C_A\big),$$

which results in the inequality

$$|h(z)| \leq 1+\epsilon\left(G'_A(1)(S'_A(1)\lambda-C_A)+T'_A(1)+G'_B(1)(S'_B(1)\lambda-C_B)+T'_B(1)\right)$$
$$\leq 1+\epsilon\left(\frac{\lambda S'_A(1)-C_A}{1-\alpha^{C_A}}+\frac{C_A}{1-\alpha^{C_A}}-\frac{1}{1-\alpha}\frac{\lambda S'_B(1)-C_B}{1-\beta^{C_B}}+\frac{C_B}{1-\beta^{C_B}}-\frac{1}{1-\beta}\right)$$
$$\leq 1+\epsilon\left(\frac{\lambda S'_A(1)}{1-\alpha^{C_A}}+\frac{\lambda S'_B(1)}{1-\beta^{C_B}}-\frac{1}{1-\alpha}-\frac{1}{1-\beta}\right).$$

By using the stability condition in Eq. (6), we can prove that $|h(z)| = |g(z)/f(z)| < 1$ for $|z| = 1+\epsilon$. Using the results obtained in Bruneel and Kim (1993), it is clear that the function $f(z)$ has $C_A + C_B + 2$ zeroes inside or on the unit circle which means that since $|f(z)| > |g(z)|$ the denominator of $U(z)$ also has $C_A + C_B + 2$ zeroes inside or on the unit circle. We note that there are three zeroes that require a more detailed look. The zeroes $z = \alpha$ or $z = \beta$, are clearly also zeroes of the numerator of $U(z)$ and result in redundant equations. The other noteworthy zero is $z = 1$, which corresponds with the normalisation condition of the pgf $U(z)$ and leads to the equation $U(1) = 1$. The resulting set of $C_A + C_B$ equations allows us to find a unique solution for all the remaining unknowns.

An important note is that using a higher maximum service capacity C_A or C_B significantly increases the computational complexity of the system while the change in the performance might be negligible when C_A and C_B are much larger than the expected batch size (denoted respectively by $\frac{1}{1-\alpha}$ and $\frac{1}{1-\beta}$). For this reason, we will study both light- and heavy-traffic approximations in order to observe the behaviour of the batch server in these important edge cases. We will present an interpolation method to obtain an accurate approximation for all loads.

3.3.1 Light-traffic approximation

We start by deducing a light-traffic approximation of the system occupancy at service initiation, by expanding $U(z)$ in a Taylor series around $\lambda = 0$ and retaining the constant and linear terms. Aside from an intrinsic interest in the behaviour of the system under the light-traffic condition, the light-traffic approximation is also useful in analysing queueing systems in moderate traffic by interpolating between the light- and heavy-traffic approximations, see Reiman and Simon (1988), Whitt (1989), which is also demonstrated in Sect. 3.3.3. The approach that is adopted in this section is similar to the one followed in Claeys et al. (2011). Also, in the appendix of Claeys et al. (2011), they have shown a proof for the analyticity of $U(z)$ in $\lambda = 0$ and, therefore, for the idle probabilities $U_{I,A}$ and $U_{I,B}$, and the partial pgfs $U_A(z)$ and $U_B(z)$. To start, we rewrite the pgf of the system occupancy at service initiation opportunities as

$$U(\lambda, z) = U_{I,A}(\lambda) + U_{I,B}(\lambda) + \frac{N_A(\lambda, z)}{Den(\lambda, z)} + \frac{N_B(\lambda, z)}{Den(\lambda, z)},$$

where $N_A(\lambda, z)$ and $N_B(\lambda, z)$ are the numerator of $U_A(z)$ and $U_B(z)$ [or the right-hand side of Eqs. (12) and (13)], $Den(\lambda, z)$ is the common denominator of $U_A(z)$ and $U_B(z)$, and $U_{I,A}(\lambda)$ and $U_{I,B}(\lambda)$ correspond with the idle probabilities $U_{I,A}$ and $U_{I,B}$. We rewrote this expression in order to show the dependency on λ.

We denote the Taylor series expansion of the unknowns $u_A(i)$ $(1 \le i \le C_A)$ and $u_B(i)$ $(1 \le i \le C_B)$ obtained in the previous section respectively by $\sum_{k=0}^{\infty} u_{A,k}(i)\lambda^k$ and $\sum_{k=0}^{\infty} u_{B,k}(i)\lambda^k$. When $\lambda = 0$, a service will never be initiated. As a result the constant terms $u_{A,0}(i)$ and $u_{B,0}(i)$ are equal to zero.

The first quantities we will analyse are the idle probabilities $U_{I,A}$ and $U_{I,B}$. From Eq. (8), we clearly see that these probabilities depend on the arrival rate λ since they contain $E(0)$, the probability that there are no arrivals, and the unknowns $u_A(i)$ $(1 \le i \le C_A)$ and $u_B(i)$ $(1 \le i \le C_B)$. Expanding $U_{I,A}$ and $U_{I,B}$ around $\lambda = 0$ yields

$$\begin{aligned}
U_{I,A}(\lambda) &= \sum_{k=0}^{\infty} U_{I,A,k}\lambda^k = \frac{1 + S'_A(1)E_1(0)\lambda + O(\lambda^2)}{-E_1(0) - \lambda E_2(0) - O(\lambda^2)} \sum_{i=1}^{C_A} (u_{A,1}(i) + u_{A,2}(i)\lambda + O(\lambda^2))\alpha^{i-1} \\
&= -\frac{\sum_{i=1}^{C_A} u_{A,1}(i)\alpha^{i-1}}{E_1(0)} - \lambda\Bigg(S'_A(1) \sum_{i=1}^{C_A} u_{A,1}(i)\alpha^{i-1} + \frac{\sum_{i=1}^{C_A} u_{A,2}(i)\alpha^{i-1}}{E_1(0)} \\
&\quad - \frac{E_2(0)\sum_{i=1}^{C_A} u_{A,1}(i)\alpha^{i-1}}{E_1(0)^2} \Bigg) + O(\lambda^2), \\
U_{I,B}(\lambda) &= \sum_{k=0}^{\infty} U_{I,B,k}\lambda^k = -\frac{\sum_{i=1}^{C_B} u_{B,1}(i)\beta^{i-1}}{E_1(0)} - \lambda\Bigg(S'_B(1) \sum_{i=1}^{C_B} u_{B,1}(i)\beta^{i-1} + \frac{\sum_{i=1}^{C_B} u_{B,2}(i)\beta^{i-1}}{E_1(0)} \\
&\quad - \frac{E_2(0)\sum_{i=1}^{C_B} u_{B,1}(i)\beta^{i-1}}{E_1(0)^2} \Bigg) + O(\lambda^2),
\end{aligned}$$

where we introduced the following Taylor expansion at $\lambda = 0$ of the pgf $E(z)$ of the arrival process

$$E(z) = 1 + \lambda E_1(z) + \lambda^2 E_2(z) + O(\lambda^3).$$

Next, the Taylor expansion of $U_A(z)$ and $U_B(z)$ has to be calculated. The constant and linear terms of the series expansion of $R_A(z) = \sum_{k=0}^{\infty} R_{A,k}(z)\lambda^k$, see Eq. (11), are given by

$$\begin{aligned}
R_{A,0}(z) &= 0 \\
R_{A,1}(z) &= z^{C_A}(\beta - z)\left[1 - \frac{E_1(z)}{E_1(0)}\right] \sum_{i=1}^{C_A} u_{A,1}(i)\alpha^i - (\beta - z)\alpha^{C_A} \sum_{i=1}^{C_A} u_{A,1}(i) z^i \\
&\quad + (1 - \beta)z^{C_A}\left[1 - \frac{\beta - z}{\beta}\frac{E_1(z)}{E_1(0)}\right] \sum_{i=1}^{C_B} u_{B,1}(i)\beta^i - (1 - \beta)\beta^{C_B} z^{C_A - C_B} \sum_{i=1}^{C_B} u_{B,1}(i) z^i,
\end{aligned}$$

and analogous expressions can be found for $R_B(z)$. Using these definitions, we obtain for $N_{A,0}(z)$ and $N_{A,1}(z)$, the constant and linear terms of the numerator $N_A(\lambda, z)$,

$$\begin{aligned}
N_{A,0}(z) &= (\alpha - z)[z^{C_B} - \beta^{C_B}]R_{A,0}(z) - (1 - \beta)[z^{C_B} - \beta^{C_B}]z^{C_A - C_B} R_{B,0}(z) = 0 \\
N_{A,1}(z) &= [z^{C_B} - \beta^{C_B}]\Big[(\alpha - z)R_{A,1}(z) - (1 - \beta)z^{C_A - C_B} R_{B,1}(z)\Big].
\end{aligned}$$

Analogous terms can be found for the numerator of $U_B(z)$, and the constant and linear terms of the Taylor series expansion of the denominator at $\lambda = 0$ are given by

$$\mathrm{Den}_0(z) = [z^{C_A} - \alpha^{C_A}][z^{C_B} - \beta^{C_B}][(z-\alpha)(z-\beta) - (1-\alpha)(1-\beta)]$$

$$\mathrm{Den}_1(z) = -(z-\alpha)(z-\beta)E_1(z)\Big(\alpha^{C_A}[z^{C_B} - \beta^{C_B}]S'_A(1) + \beta^{C_B}[z^{C_A} - \alpha^{C_A}]S'_B(1)\Big)$$
$$- (1-\alpha)(1-\beta)[z^{C_A} - \alpha^{C_A}][z^{C_B} - \beta^{C_B}]E_1(z)[S'_A(1) + S'_B(1)]. \quad (14)$$

With these definitions, we obtain that the Taylor expansions of $U_A(z)$ and $U(z)$ around $\lambda = 0$ are equal to

$$U_A(z) = 0 \cdot \lambda^0 + \frac{N_{A,1}(z)}{Den_0(z)}\lambda^1 + O(\lambda^2),$$

$$U(z) = -\frac{\sum_{i=1}^{C_A} u_{A,1}\alpha^{i-1} + \sum_{i=1}^{C_B} u_{B,1}\beta^{i-1}}{E_1(0)}\lambda^0 + \lambda^1\Bigg[\frac{N_{A,1}(z) + N_{B,1}(z)}{Den_0(z)} - S'_A(1)\sum_{i=1}^{C_A} u_{A,1}\alpha^{i-1}$$
$$- S'_B(1)E_1(0)\sum_{i=1}^{C_B} u_{B,1}\beta^{i-1} - \frac{\sum_{i=1}^{C_A} u_{A,2}\alpha^{i-1} + \sum_{i=1}^{C_B} u_{B,2}\beta^{i-1}}{E_1(0)}$$
$$+ \frac{E_2(0)}{E_1(0)^2}\Bigg[\sum_{i=1}^{C_A} u_{A,1}\alpha^{i-1} + \sum_{i=1}^{C_B} u_{B,1}\beta^{i-1}\Bigg]\Bigg] + O(\lambda^2).$$

The mean system occupancy under the light-traffic condition is given by

$$E[U]_L = 0 \cdot \lambda^0 + \left[\frac{N''_{A,1}(1) + N''_{B,1}(1)}{2Den'_0(1)} - \frac{Den''_0(1)}{2Den'_0(1)^2}(N'_{A,1}(1) + N'_{B,1}(1))\right]\lambda^1 + O(\lambda^2).$$

In order to be able to calculate the constant and linear terms of the approximated mean system occupancy at service initiation opportunities, only the linear terms of the unknowns $u_A(i)$ and $u_B(i)$ have to be deduced. We denote the Taylor series expansion of the i-th zero $z_i(\lambda)$ of the denominator, given in Eq. (14), by $\sum_{k=0}^{\infty} z_{i,k}\lambda^k$. In order to fully characterize the light-traffic approximation of the mean system occupancy at service initiation opportunities, we need to find the constant terms of the zeroes in order to have a unique solution for the linear terms of the unknowns $u_A(i)$ and $u_B(i)$. The constant terms of the zeroes are the solutions of $Den_0(z) = 0$, resulting in

$$z_{i,0} = \begin{cases} 1 & i = 0 \\ \alpha\epsilon_{C_A,i} & 1 \le i \le C_A - 1 \\ \beta\epsilon_{C_B,i-C_A+1} & C_A \le i \le C_A + C_B - 2 \\ \alpha + \beta - 1 & i = C_A + C_B - 1 \\ \alpha & i = C_A + C_B \\ \beta & i = C_A + C_B + 1 \end{cases},$$

where $\epsilon_{c,i} = e^{(\iota 2\pi i)/c}$ is the i-th complex c-th root of one, with ι as the imaginary unit and $i = 0, \ldots, c-1$. The cases $z_{i,0} = \alpha$ and $z_{i,0} = \beta$ result in redundant equations.

The equations corresponding with these zeroes are either the normalisation condition $U(1) = 1$, or that the numerator vanishes for the zero resulting in the following equations

$$\begin{cases} U_{I,A,0} + U_{I,B,0} = 1 & , i = 0 \\ N_{A,1}(z_{i,0}) = 0 & , 1 \le i \le C_A - 1 \\ N_{B,1}(z_{i,0}) = 0 & , C_A \le i \le C_A + C_B - 2 \\ N_{A,1}(z_{i,0}) + N_{B,1}(z_{i,0}) = 0 & , i = C_A + C_B - 1 \end{cases}$$

After solving this set of equations resulting in a unique solution for the remaining unknowns, the constant and linear terms of the light traffic approximation of the mean system occupancy at service initiation opportunities are fully characterized.

We note that while we must still solve a system with $C_A + C_B$ equations and $C_A + C_B$ unknowns to obtain the light-traffic approximation, the constant terms of the zeroes of the denominator are known exactly which results in a significant reduction of the computational complexity.

3.3.2 Heavy-traffic approximations

The previous approximation can be used to analyse the behaviour of the system when the load is very small. In this section, we will study the other side of the spectrum and introduce a heavy-traffic approximation. In this case, the server will rarely be idle and the service capacity will almost always be equal to the corresponding maximum service capacity.

First, we define $N_U(z)$ and $D_U(z)$ as respectively the numerator and denominator of $U(z)$. From Eqs. (12) and (13), it is clear that $N_U(1) = D_U(1) = 0$ and the normalisation condition dictates that $U(1) = 1$. As a result, we obtain, after applying l'Hôpital's rule,

$$E[U] = U'(1) = \frac{N''_U(1) - D''_U(1)}{2D'_U(1)}.$$

We now let λ converge to the arrival rate at which the system becomes unstable, see Eq. (6). Under this condition, the system is nearly unstable meaning that the number of customers in the system is typically very high and the variables $u_A(i)$ $(1 \leq i \leq C_A)$ and $u_B(j)$ $(1 \leq j \leq C_B)$ go to zero. This means that both $R_A(z)$ and $R_B(z)$ go to zero, which leads to the numerator $N_U(z)$ and all of its derivatives also going to zero. The mean system occupancy at service initiation opportunities can then be approximated by

$$U'(1) \sim -\frac{D''_U(1)}{2 \cdot D'_U(1)}.$$

The first and second derivative of the denominator of $U(z)$, evaluated at $z = 1$, are equal to

$$\begin{aligned} D'_U(1) &= (2-\alpha-\beta)(1-\alpha^{C_A})(1-\beta^{C_B}) \\ &\quad - (1-\alpha)(1-\beta)\lambda[(1-\beta^{C_B})S'_A(1) + (1-\alpha^{C_A})S'_B(1)] \\ &= (2-\alpha-\beta)(1-\alpha^{C_A})(1-\beta^{C_B})(1-\rho), \end{aligned}$$

and

$$\begin{aligned} D''_U(1) &= 2(1-\alpha^{C_A})(1-\beta^{C_B}) + 2(2-\alpha-\beta)[C_A(1-\beta^{C_B}) + C_B(1-\alpha^{C_A})] \\ &\quad - (1-\alpha)(1-\beta)\Big[(S''_A(1)\lambda^2 + S'_A(1)E''(1))(1-\beta^{C_B}) + (S''_B(1)\lambda^2 \\ &\quad + S'_B(1)E''(1))(1-\alpha^{C_A})\Big] - 2(1-\alpha)(1-\beta)S'_A(1)S'_B(1)\lambda^2(1-\alpha^{C_A}-\beta^{C_B}) \\ &\quad - 2S'_A(1)\lambda[(2-\alpha-\beta)\alpha^{C_A}(1-\beta^{C_B}) + (1-\alpha)(1-\beta)(C_B + (1-\beta^{C_B})C_A)] \\ &\quad - 2S'_B(1)\lambda[(2-\alpha-\beta)\beta^{C_B}(1-\alpha^{C_A}) + (1-\alpha)(1-\beta)(C_A + (1-\alpha^{C_A})C_B)]. \end{aligned}$$

We note that there are no unknowns in the equation for the mean system occupancy under heavy-traffic approximations, which means that the computational complexity is no longer dependent on the maximum service capacities C_A and C_B.

Using the previous expressions, we can also write the expression for the heavy-traffic approximation as a function of the load ρ, which leads to

$$U'(1) \sim \frac{f_H(\rho)}{1-\rho}, \tag{15}$$

where

$$f_H(\rho) = \frac{-D''_U(1)}{2(2-\alpha-\beta)(1-\alpha^{C_A})(1-\beta^{C_B})}.$$

When looking at the behaviour of this function, we noticed that the limit for ρ going to 0 is negative. Therefore we modify the Eq. (15) by adding $f_H(0)$. This leads to the heavy-traffic approximation $E[U]_H$, which is

$$E[U]_H = \frac{f_H(\rho)}{1-\rho} + |f_H(0)|,$$

where $f_H(0)$ is the limit of the function $f_H(\rho)$ for ρ going to 0. We observed experimentally that adding the term $|f_H(0)|$ significantly increased the accuracy of the heavy-traffic approximation. This does not mean however that the heavy-traffic approximation is always positive since the minimum of the function $f_H(\rho)$ is not used.

Intuitively, we expect that the heavy-traffic approximation is a lower-bound for the mean system occupancy, since it and the mean occupancy are both monotonically increasing functions, the heavy-traffic approximation only becomes zero at significant loads (this occurs at $\rho = 0$ for the mean system occupancy) and both functions approach the same value at $\rho = 1$. This expectation is also observed experimentally. Now, we also know that the impact of the term by which we increased the heavy-traffic approximation so that it is 0 at $\rho = 0$, diminishes when the load increases because the mean system occupancy goes to infinite and therefore the term becomes negligible when ρ approaches 1. Due to these considerations, we expect that adding the additional term significantly improves the approximation at intermediate loads and also slightly increases the accuracy at heavy loads.

3.3.3 Interpolation of light- and heavy-traffic approximation

While the previous two approximations are accurate when the load is either low or high enough, at medium loads the accuracy of these methods deteriorates. To remedy this, we propose a method of approximating the system occupancy based on interpolating between the light and heavy-traffic approximations. The function for calculating this interpolated approximation $E[U]_I$ that we propose is as follows

$$E[U]_I = E[U]_L \frac{1-\rho^{a_L}}{1-\rho} + \rho^{a_H} E[U]_H, a_L \geq 2, a_H \geq 1,$$

where we note that both the light- and heavy-traffic approximations are functions of the load. Increasing a_L increases the impact of the light-traffic approximation $E[U_L]$ at medium loads, while increasing a_H decreases the impact of the heavy-traffic approximation $E[U_H]$. Note that we set $a_H \geq 2$, since for $a_H = 1$, this term would also contribute to the light-load approximation (due to its proportionality to λ). In this interpolation, we use $a_H = 2$ in order to minimize the impact of the heavy-traffic approximation at low loads. With this interpolation, the parameter a_L can be freely chosen but has a significant impact on the accuracy of the interpolation. Based on our experiments, we can derive that $a_L = 4$ results in a very good interpolated approximation. Using $a_L = 4$, we note that the interpolated approximation is a lower bound but becomes an upper bound for ρ, α and β sufficiently high.

3.4 System occupancy at random slot boundaries

Previously, we studied the system at service initiation opportunities in order to avoid the difficulties introduced by using generic and class-dependent service times. In this part of the analysis, we will use the previous results in order to calculate the pgf of the system occupancy at random slot boundaries, denoted by $U_R(z)$. The first step is to find the probabilities that the type τ_R of a random slot is either an idle slot and the previous service was a class A or B batch, or busy processing class A or B customers, denoted correspondingly by $Pr[\tau_R = I, A]$, $Pr[\tau_R = I, B]$, $Pr[\tau_R = A]$ or $Pr[\tau_R = B]$. By using the mean service times $S'_A(1)$ and $S'_B(1)$, we obtain

$$
\begin{aligned}
Pr[\tau_R = I, A] &= \frac{U_{I,A}}{U_{I,A} + U_{I,B} + U_A(1)S'_A(1) + U_B(1)S'_B(1)},\\
Pr[\tau_R = I, B] &= \frac{U_{I,B}}{U_{I,A} + U_{I,B} + U_A(1)S'_A(1) + U_B(1)S'_B(1)},\\
Pr[\tau_R = A] &= \frac{U_A(1)S'_A(1)}{U_{I,A} + U_{I,B} + U_A(1)S'_A(1) + U_B(1)S'_B(1)},\\
Pr[\tau_R = B] &= \frac{U_B(1)S'_B(1)}{U_{I,A} + U_{I,B} + U_A(1)S'_A(1) + U_B(1)S'_B(1)}.
\end{aligned}
$$

Based on Bruneel and Kim (1993), we find that the number of arrivals during the elapsed service period of a class A service results in $(S_A(E(z)) - 1)/[S'_A(1)(E(z) - 1)]$, and an analogous expression for a class B service holds as well. The service time of a batch of type A (or B) is independent of the number of customers in the queue at a type A (or B) service initiation, which means we obtain that the pgf $U_R(z)$ of the system occupancy at random slot boundaries is equal to

$$
\begin{aligned}
U_R(z) := {} & Pr[\tau_R = I_A] + Pr[\tau_R = I_B] + Pr[\tau_R = A]\frac{U_A(z)}{U_A(1)}\frac{S_A(E(z)) - 1}{S'_A(1)(E(z) - 1)}\\
& + Pr[\tau_R = B]\frac{U_B(z)}{U_B(1)}\frac{S_B(E(z)) - 1}{S'_B(1)(E(z) - 1)},
\end{aligned}
$$

resulting in the following mean system occupancy

$$
U'_R(1) = E[U_R] = Pr[\tau_R = A]\Big[\frac{U'_A(1)}{U_A(1)} + \frac{S''_A(1)\lambda}{2S'_A(1)}\Big] + Pr[\tau_R = B]\Big[\frac{U'_B(1)}{U_B(1)} + \frac{S''_B(1)\lambda}{2S'_B(1)}\Big].
$$

The methods previously developed for obtaining light- and heavy-traffic approximations of the system occupancy at service initiation opportunities can also be used to approximate the pgf $U_R(z)$ of the system occupancy at random slot boundaries.

3.4.1 Light-traffic approximation

Now, we will also give a light-traffic approximation of the system occupancy at random slot boundaries. We start by looking at the probability $Pr[\tau_R = I, A]$ that a random slot is an idle slot and the most recently initiated batch was a class A batch which is equal to

$$Pr[\tau_R = I, A] = \frac{U_{I,A,0} + U_{I,A,1}\lambda}{U_{I,A,0} + U_{I,A,1}\lambda + U_{I,B,0} + U_{I,B,1}\lambda + \frac{N'_{A,1}(1)}{Den'_0(1)}S'_A(1)\lambda + \frac{N'_{B,1}(1)}{Den'_0(1)}S'_B(1)\lambda}$$

$$= \frac{U_{I,A,0}}{U_{I,A,0} + U_{I,B,0}} + \frac{U_{I,A,1}U_{I,B,0} - U_{I,A,0}[U_{I,B,1} + \frac{N'_{A,1}(1)}{Den'_0(1)}S'_A(1) + \frac{N'_{B,1}(1)}{Den'_0(1)}S'_B(1)]}{(U_{I,A,0} + U_{I,B,0})^2}\lambda + O(\lambda^2),$$

and a similar approach for $Pr[\tau_R = A]$ results in

$$Pr[\tau_R = A] = \frac{\frac{N'_{A,1}(1)}{Den'_0(1)}\lambda S'_A(1)}{U_{I,A,0} + U_{I,A,1}\lambda + U_{I,B,0} + U_{I,B,1}\lambda + \frac{N'_{A,1}(1)}{Den'_0(1)}\lambda S'_A(1) + \frac{N'_{B,1}(1)}{Den'_0(1)}\lambda S'_B(1)}$$

$$= 0 \cdot \lambda^0 + \frac{\frac{N'_{A,1}(1)}{Den'_0(1)}S'_A(1)}{U_{I,A,0} + U_{I,B,0}}\lambda + O(\lambda^2).$$

The probabilities corresponding with class-B services are computed analogously. We also obtain the following Taylor-series expansion around $\lambda = 0$ for the pgf of the number of arrivals during a service period that occur before the random slot:

$$\frac{S_A(E(\lambda, z)) - 1}{S'_A(1)(E(\lambda, z) - 1)} = 1 \cdot \lambda^0 + \frac{S''_A(1)E_1(z)}{2S'_A(1)}\lambda + O(\lambda^2).$$

With these Taylor-series for the probabilities of the different server states and the pgf of the number of arrivals, we obtain that the pgf of the system occupancy at random slot boundaries can be written as

$$U_R(z) = 1 \cdot \lambda^0 + \frac{[\frac{N_{A,1}(z)}{Den_0(z)} - \frac{N'_{A,1}(1)}{Den'_0(1)}]S'_A(1) + [\frac{N_{B,1}(z)}{Den_0(z)} - \frac{N'_{B,1}(1)}{Den'_0(1)}]S'_B(1)}{U_{I,A,0} + U_{I,B,0}}\lambda + O(\lambda^2).$$

From this expression, we derive the following light-traffic approximation for the mean system occupancy $E[U_R]_L$ at random slot boundaries

$$E[U_R]_L := \frac{N''_{A,1}(1)S'_A(1) + N''_{B,1}(1)S'_B(1) - \frac{Den''_0(1)}{Den'_0(1)}(N'_{A,1}(1)S'_A(1) + N'_{B,1}(1)S'_B(1))}{2Den'_0(1)(U_{I,A,0} + U_{I,B,0})}\lambda + O(\lambda^2).$$

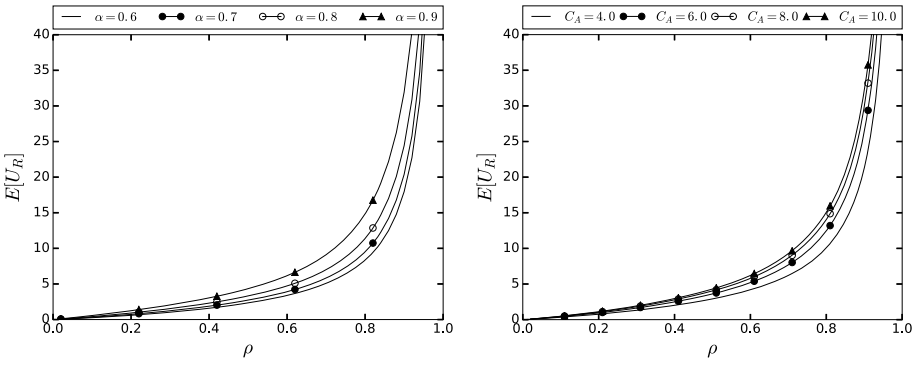
3.4.2 Heavy-traffic approximation

The heavy-traffic approximation of the system occupancy at random slot boundaries can be calculated analogously as the heavy-traffic approximation at service initiation opportunities. Since the idle probabilities approach zero when the system is operating under a high load, the pgf of the system occupancy at random slot boundaries can be approximated by

$$U_R(z) \sim \frac{U_A(z)(S_A(E(z)) - 1) + U_B(z)(S_B(E(z)) - 1)}{[U_A(1)S'_A(1) + U_B(1)S'_B(1)][E(z) - 1]}.$$

We note that the probabilities $U_A(1)$ and $U_B(1)$ are, under the heavy-traffic assumption, equal to the probabilities $Pr[\tau = A]$ and $Pr[\tau = B]$, see Eq. (5). Using this approximated pgf, we obtain the following approximation for the mean system occupancy at random slot boundaries, denoted by $E[U_R]_H$,

$$E[U_R]_H = E[U]_H + \lambda\frac{S''_A(1)U_A(1) + S''_B(1)U_B(1)}{S'_A(1)U_A(1) + S'_B(1)U_B(1)},$$



(a) $C_A = C_B = 4, \alpha = \beta$ **(b)** $\alpha = \beta = 0.7, C_A = C_B$

Fig. 1 Impact of clustering (**a**) and maximum service capacities (**b**) on the mean system occupancy at random slot boundaries as a function of the load ρ

where we used $E[U]_H$, the heavy-traffic approximation of the mean system occupancy at service initiation opportunities, which we derived previously.

3.4.3 Interpolation of light- and heavy-traffic approximation

Finally, we propose an analogous interpolation function for obtaining an approximation for the mean system occupancy at random slot boundaries, denoted by $E[U_R]_I$. This approximation is analogous and given by

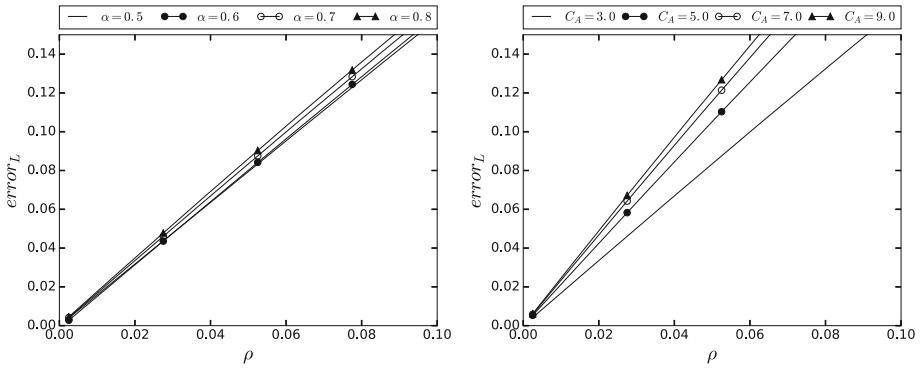
$$E[U_R]_I = E[U_R]_L \frac{1-\rho^{a_L}}{1-\rho} + \rho^{a,H} E[U_R]_H, a_L \geq 2, a_H \geq 1.$$

We can use the same values for the parameters a_L and a_H that we used in Sect. 3.3.3.

4 Discussion of results and numerical examples

In this section, we will evaluate the impact of different parameters on the performance of the system using a number of numerical examples. In the following examples, we will use geometrically distributed service times $S_A(E(z))$ and $S_B(E(z))$ with a mean of 3 slots, and the maximum service capacities C_A and C_B of class A and B batches are assumed to be equal. The arrival process of the total number of arrivals in each slot is also geometrically distributed, and in the following examples the same-class probability α is always equal to β.

In Fig. 1, we show the impact of the degree of clustering in the arrival process (a) and of the maximum service capacities (b) on the mean system occupancy at random slot boundaries, given by $E[U_R]$. We see that increasing both the degree of clustering or the maximum service capacities result in an increase of the mean system occupancy. While this might be counter-intuitive because increasing either parameter can lead to larger batches and in turn a better system performance, we note that the figures are plotted as a function of the load of the system. Since increasing either parameter can improve the system performance, the mean arrival rate λ is also changed in order to maintain a constant load. We also observe that changes in the probability α are most significant when the degree of clustering is already high, while the opposite holds for the service capacity. We also observe that, as long as the



(a) $C_A = C_B = 3$ **(b)** $\alpha = \beta = 0.7$

Fig. 2 Accuracy of the light-traffic approximation using geometric arrival and service processes

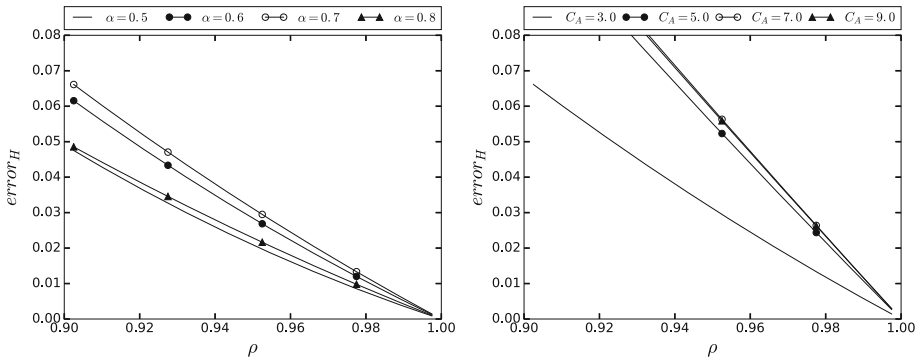
maximum service capacities are significantly larger than the expected length of a sequence of same class customers, changing the maximum service capacities or same-class probabilities only lead to negligible changes to the mean system occupancy.

Next, we will evaluate the accuracy of the proposed approximations by studying the relative error, defined as

$$error_x := \frac{|E[U] - E[U]_x|}{E[U]},$$

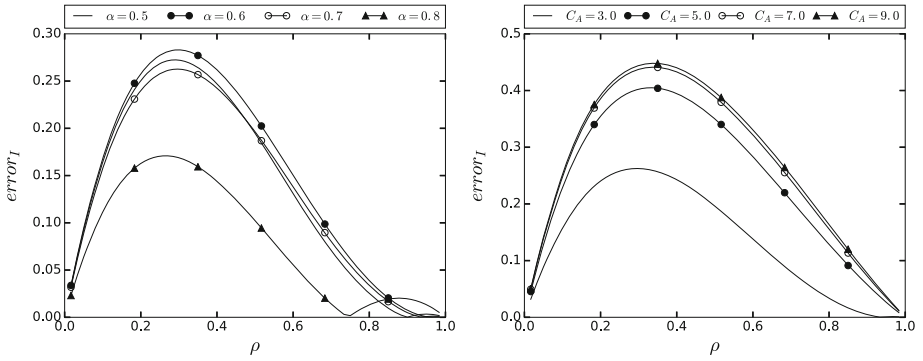
where x is either L, H or I for respectively the light-traffic, heavy-traffic or interpolated approximation. In Fig. 2, we show the relative error of the light-traffic approximation as a function of the load, and study the impact of clustering (by varying the parameter α in Fig. 2a) and of the maximum service capacities (Fig. 2b). It is clear that increasing both parameters also results in a higher relative error. The reason for this is that increasing either parameter results in a system that can process a higher mean arrival rate before becoming unstable, which means that λ also increases in order to maintain a stable load. While this results in a slight increase of the error of the light traffic approximation for changes to α or β, we see in Fig. 2b that higher values for the maximum service capacities result in a significantly higher relative error. However, when the maximum service capacity becomes larger than the expected length of a sequence of same-class customers, a point is quickly reached at which higher service capacities only result in a negligible increase of the error of the light-traffic approximation.

We can also investigate the accuracy of the heavy-traffic approximation by looking at $error_H$, which is the relative error of the heavy traffic approximation. We note that this is the modified heavy-traffic approximation where we made sure that it goes to 0 when the load approaches 0. This error is shown in Fig. 3 for identical arrival and service time distributions as in previous figures. We observe that the impact of the maximum service capacities on the relative error, see Fig. 3b, is similar to the effect on the relative error of the light-traffic approximation. This means that higher maximum service capacities result in a higher relative error. However, once C_A and C_B are much larger than the expected length of a sequence of same-class customers, further increasing the maximum capacities results in a insignificant increase of the relative error. The impact of increasing same-class probabilities is different than the one we observed for the light-traffic approximation. For the heavy-traffic



(a) Varying $\alpha = \beta$, $C_A = C_B = 3$ **(b)** Varying $C_A = C_B$, $\alpha = \beta = 0.7$

Fig. 3 Accuracy of the heavy-traffic approximation using geometric arrival and service processes



(a) Varying $\alpha = \beta$, $C_A = C_B = 3$ **(b)** Varying $C_A = C_B$, $\alpha = \beta = 0.7$

Fig. 4 Accuracy of the interpolated approximation using geometric arrival and service processes

approximation, the relative error first increases when the same-class probabilities increase but at higher values of α and β, the error starts decreasing again.

While these approximations allow us to study the behaviour of the system under light- or heavy-traffic conditions, their accuracy deteriorates at moderate loads. For this reason, we proposed an interpolation between the light- and heavy-traffic approximation which we called the interpolated approximation introduced in Sect. 3.3.3. The accuracy of this interpolation, with the parameter $a_L = 4$, is shown in Fig. 4. It is clear that the approximation by interpolation yields quite accurate results but reaches its worst point around a load ρ of 0.3. The impact of increasing the maximum service capacity, see Fig. 4b, is analogue to the observed impact on the accuracy of the light- and heavy-traffic approximations in that a higher value for C_A and C_B leads to a higher error but a point is quickly reached at which this increase is negligible. However, the impact of increasing same-class probabilities is more complicated. The relative errors for $\alpha = 0.5$, 0.6 and 0.7 are only slightly different from each other, but the relative error of $\alpha = 0.8$ is much smaller and also has a point at which it becomes 0. This is the point at which the interpolated approximation goes from being a lower bound, to being an upper bound. We note that the most accurate results are obtained when the expected length of a sequence of same-class customers is sufficiently larger than the maximum service capacities.

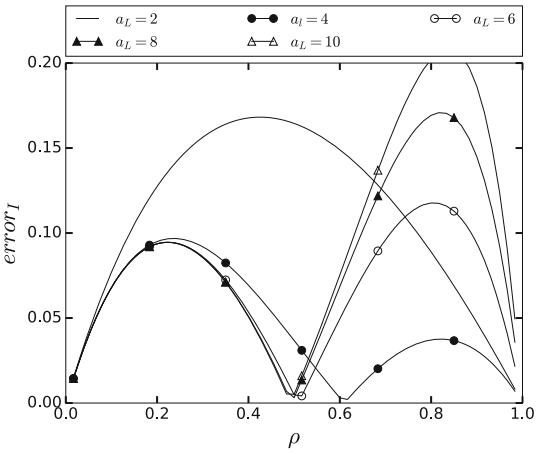


Fig. 5 Effect of different options for a_L (with the parameters $\alpha = \beta = 0.9, C_A = C_B = 3$)

In Sect. 3.3.3, we proposed that $a_L = 4$ results in an accurate approximation. In order to justify this choice, we show the relative error of the interpolated approximation for a number of different values of a_L in Fig. 5. It is clear that $a_L = 2$ results, for all loads, in a worse approximation than $a_L = 4$. While there is a region where using $a_L > 4$ results in a lower relative error, the accuracy of these options decreases significantly after the interpolation approximation goes from a lower bound to an upper bound. Therefore, these values for a_L result on average in a worse approximation. However, in Fig. 4, we see that the interpolation approximation does not always become an upper bound. In these cases, higher values of a_L result in a slight increase of the accuracy of the interpolation approximation but only when the load ρ is larger than 0.5. At low loads, the change in the accuracy of the interpolation approximation for $a_L \geq 4$ is negligible.

We will end this section on the numerical results by showing the mean system occupancy at service initiation opportunities and the obtained approximations for three different combinations of same-class probabilities and maximum service capacities for a range of values for the load ρ, see Table 1. In this table, we also compare the results obtained by the analysis of this paper with the expected system occupancy $E[U]_\infty$ of the simplified model without maximum service capacities, which is analysed in Baetens et al. (2016, 2017, 2018a, b). We note that in this simplified model, only 2 parameters must be solved and is therefore much less complex to compute. We observe that the light-traffic approximation is rather inaccurate for the considered loads and can therefore, on its own, only be used in a system under strict light-traffic conditions. However, the situation is different for the heavy-traffic approximation which is still rather accurate at loads of 0.8 and even at 0.6 for the case that $\alpha = \beta = 0.9$. We also observe that, when $\alpha = \beta = 0.5$ or 0.7, the heavy-traffic approximation becomes negative at low loads which is expected (see Sect. 3.3.3). When comparing the interpolation and the heavy-traffic approximation, we notice that the interpolated approximation performs better for $\rho < 0.8$. For higher loads, which approximation is the most accurate will depend on the mean length of a sequence of same-class customers and the corresponding maximum service capacities. Based on some numerical experiments, we see that the heavy-traffic approximation will be more accurate at loads $\rho > 0.8$ when the expected length of a sequence of each class of customers is less than the maximum service capacity of that class, or $Pr[\tau = A]/(1-\alpha) + Pr[\tau = B]/(1-\beta) < Pr[\tau = A]C_A + Pr[\tau = B]C_B$. Finally, we will compare the interpolation approximation with the approximation obtained by removing

Table 1 The computed mean system occupancy at service initiation opportunities compared to the three different types of approximations and the results from the simplified model for a number of different same-class probabilities, maximum service capacities and loads

$\alpha = \beta$	$C_A = C_B$	ρ	$E[U]$	$E[U]_L$	$E[U]_I$	$E[U]_H$	$E[U]_\infty$
0.5	3	0.2	0.16569	0.11667	0.14148	−0.1071	0.20028
		0.4	0.52858	0.23333	0.38697	0.04762	0.67833
		0.6	1.46818	0.35000	1.07118	0.85714	1.94999
		0.8	4.92900	0.46667	4.12373	4.28571	6.52955
		0.9	12.5072	0.52500	11.4470	11.8929	16.3135
0.9	3	0.2	0.26676	0.18067	0.24141	0.39847	1.76941
		0.4	0.82279	0.36133	0.76578	1.11859	5.75216
		0.6	2.13734	0.54200	2.13081	2.64283	13.9999
		0.8	6.61326	0.72267	6.85879	7.38356	38.9374
		0.9	16.0648	0.81300	16.5559	16.9911	88.9091
0.7	9	0.2	0.36912	0.21325	0.25784	−0.2075	0.39388
		0.4	1.29140	0.42651	0.70506	0.07753	1.40710
		0.6	3.50974	0.63976	1.96587	1.59373	3.88889
		0.8	10.6939	0.85302	7.66034	8.03473	11.9886
		0.9	25.3757	0.95695	21.3924	22.3360	28.5556

the maximum service capacities. We note that the approximation of the simplified model is most accurate when the expected length of a sequence of same-class customers (determined by the parameters α and β) is much smaller than the maximum service capacities, which for instance occurs when $\alpha = \beta = 0.7$ and $C_A = C_B = 9$. However, when the length of a sequence of same-class customers is close to the maximum service capacities, then the approximation of the simplification will only be accurate when the load is sufficiently low (see the case $\alpha = \beta = 0.5$ and $C_A = C_B = 3$). In the last case ($\alpha = \beta = 0.9$ and $C_A = C_B = 3$), which occurs when the maximum service capacities are much smaller than the expected length, the approximation by simplification is useless even at small loads. This behaviour corresponds to our predictions in our previous papers, see Baetens et al. (2016, 2017, 2018a, b). Combining these observations, we obtain the following conditions to determine which method for approximating is preferred:

- $\frac{Pr[\tau=A]}{1-\alpha} + \frac{Pr[\tau=B]}{1-\beta} << Pr[\tau = A]C_A + Pr[\tau = B]C_B \rightarrow$ Use the simplified model
- $\frac{Pr[\tau=A]}{1-\alpha} + \frac{Pr[\tau=B]}{1-\beta} \sim Pr[\tau = A]C_A + Pr[\tau = B]C_B$ and $\rho > 0.8 \rightarrow$ Use the modified heavy-traffic approximation
- Otherwise $\rightarrow$ Use the interpolation approximation.

5 Conclusions

In this paper, we examined a batch-service queueing system with a variable and class-dependent service capacity. The main contribution of this paper is the inclusion of maximum class-dependent service capacities which is instrumental in modelling a realistic system. We started the analysis by establishing the system equations of the variable-capacity batch-service queueing system. We used these system equations to calculate the probability generating

function of the system occupancy at service initiation opportunities and gave the proof for the number of zeroes in the denominator, which is equal to the sum of the maximum service capacities. The numerical complexity of finding these unknowns depends on the sum of the maximum service capacities and can therefore numerically be very complicated. Therefore, we present both a light- and heavy-traffic approximation in order to reduce this numerical complexity. We also propose an interpolation between the light- and heavy-traffic approximation in order to study the behaviour in the intermediate region. We completed the analysis by extending these results to random slot boundaries. In the numerical experiments, we focused on the impact of clustering between same-class customers and of the maximum service capacities on the mean system occupancy. We also discussed the impact of these parameters on the relative error of the light-traffic, heavy-traffic and interpolated approximation. Finally, we consider the conditions under which each of these approximations are accurate and compare the results of this paper with the results obtained by using a simplified model without maximum service capacities. Some guidelines have also been included to predict which method of approximating would be most accurate or useful.

References

Adan, I., Leeuwaarden, J. V., & Winands, E. (2006). On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, *34*(3), 355–360.

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., Claeys, D., Steyaert, B., & Bruneel, H. (2017). System performance of a variable-capacity batch-service queue with geometric service times and customer-based correlation. In *31st European conference on Modelling and Simulation*.

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 S. Wittevrongel & T. Phung-Duc (Eds.), *Analytical and stochastic modelling techniques and applications* (pp. 32–44). Switzerland: Springer

Baetens, J., Steyaert, B., Claeys, D., & Bruneel, H. (2018a). Delay analysis of a two-class batch-service queue with class-dependent variable server capacity. *Mathematical Methods of Operations Research*, *88*, 1–21.

Baetens, J., Steyaert, B., Claeys, D., & Bruneel, H. (2018b). Delay analysis of a variable-capacity batch-server queue with general class-dependent service times. In *AIP conference proceedings* (Vol. 1978, p. 4). American Institute of Physics (AIP).

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., & 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.

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.

Bellalta, B., & Oliver, M. (2009). A space-time batch-service queueing model for multi-user MIMO communication systems. In *Proceedings of the 12th ACM international conference on modeling, analysis and simulation of wireless and mobile systems* (pp. 357–364). ACM.

Bountali, O., & Economou, A. (2017). Equilibrium threshold joining strategies in partially observable batch service queueing systems. *Annals of Operations Research*, *277*, 1–23.

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*. Boston: Kluwer Academic.

Bruneel, H., Mélange, W., Steyaert, B., Claeys, D., & Walraevens, J. (2012). A two-class discrete-time queueing model with two dedicated servers and global FCFS service discipline. *European Journal of Operational Research*, *223*(1), 123–132.

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. G. (1983). *A first course in bulk queues*. New York: Wiley.

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.

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.

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

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.

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

Niranjan, S., Chandrasekaran, V., & Indhira, K. (2017). State dependent arrival in bulk retrial queueing system with immediate Bernoulli feedback, multiple vacations and threshold. In *IOP conference series: materials science and engineering* (Vol. 263). IOP Publishing.

Pradhan, S., & Gupta, U. (2017). Analysis of an infinite-buffer batch-size-dependent service queue with Markovian arrival process. *Annals of Operations Research*, *277*, 1–36.

Reiman, M. I., & Simon, B. (1988). An interpolation approximation for queueing systems with Poisson input. *Operations Research*, *36*(3), 454–469.

Van Der Wal, J., & Yechiali, U. (2003). Dynamic visit-order rules for batch-service polling. *Probability in the Engineering and Informational Sciences*, *17*(3), 351–367.

Vardakas, J. S., & Logothetis, M. D. (2009). Packet delay analysis for priority-based passive optical networks. In *First international conference on emerging network intelligence* (pp. 103–107).

Whitt, W. (1989). An interpolation approximation for the mean workload in a GI/G/1 queue. *Operations Research*, *37*(6), 936–952.

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.