Methodology and Computing in Applied Probability (2021) 23:1551–1579
https://doi.org/10.1007/s11009-020-09828-4

Analysis of a Queueing Model with Batch Markovian Arrival Process and General Distribution for Group Clearance

Srinivas R. Chakravarthy[1] · Shruti[2] · Alexander Rumyantsev[3,4]

Received: 16 January 2020 / Revised: 23 September 2020 / Accepted: 28 September 2020 /
Published online: 19 October 2020

Abstract

In this paper we consider a single server queueing model with under general bulk service rule with infinite upper bound on the batch size which we call *group clearance*. The arrivals occur according to a batch Markovian point process and the services are generally distributed. The customers arriving after the service initiation cannot enter the ongoing service. The service time is independent on the batch size. First, we employ the classical embedded Markov renewal process approach to study the model. Secondly, under the assumption that the services are of phase type, we study the model as a continuous-time Markov chain whose generator has a very special structure. Using matrix-analytic methods we study the model in steady-state and discuss some special cases of the model as well as representative numerical examples covering a wide range of service time distributions such as constant, uniform, Weibull, and phase type.

Keywords Queueing · Stochastic processes · Batch arrivals · Group clearance · Matrix-analytic method

Mathematics Subject Classification (2010) 60J27 · 60K25

1 Introduction

Batching (bulking) and group clearance are natural ways to improve throughput and utilization of a system used in various fields from public transportation (Grippa et al. 2019) and

✉ Srinivas R. Chakravarthy
schakrav@kettering.edu

[1] Departments of Industrial and Manufacturing Engineering, Mathematics, Kettering University, Flint, MI 48504, USA

[2] Department of Mathematics, Birla Institute of Technology and Science Pilani, Pilani Campus, Pilani, Rajasthan 333031, India

[3] Institute of Applied Mathematical Research, Karelian Research Centre of RAS, 11 Pushkinskaya Str., Petrozavodsk, Russia

[4] Petrozavodsk State University, 33 Lenina Pr., Petrozavodsk, Russia

cargo to blood screening, production systems, amusement parks etc. (a number of interesting applications with appropriate references can be found in Claeys et al. (2013)). In information technology batching is widely implemented in *telecommunication* and *computing* systems. For instance, packages in most common implementations of the TCP protocol are grouped into the so-called windows, which are sent to a receiver simultaneously. Another example is the packet encapsulation during transmission, when aggregates rather than standalone packages, equipped with a single header, are sent in a single transmission. The effect of increased throughput along with improved channel utilization by frame aggregation plays a crucial role in recent 802.11 standards and is further amplified in the case of IEEE 802.11n WLANs (Charfi et al. 2017).

By the same reason of improving the throughput and reducing the latency, data are being aggregated in a write buffer before the actual execution of write operation on a conventional hard disk drive is done. The data aggregation and group clearance feature on a NAND-flash based *solid state drive* (SSD) is related to the specifics of write operations, which are allowed to be performed only on a block level, as opposed to a more fine-grained page level (representing the granularity of read operations). Aggregation and write caching in modern SSDs is also related to the so-called data deduplication (replacing redundant data writes by pointers to an existing data blocks/pages), that may be performed in advance before the actual write operation is done; thus, improving the lifetime of an SSD by reducing the number of unnecessary writes (see Ivashko et al. 2018). Moreover, similar problems are studied at the hard disk drive (HDD) storage, see Xie et al. (2020) and Saxena et al. (2018) or database level, see O'Mullane et al. (2005). We stress that in the context of cloud computing, there is no natural upper bound on the batch size, and thus the batch service can be started for an arbitrary large batch (Saxena et al. 2018).

Aside from latency reduction, in *distributed computing* the workunit batching is used for improving the application turnaround, whereas the so-called workunit replication is used to improve reliability as well (see Mazalov et al. 2014). At that, the results of workunits completion are to be validated and assimilated into the overall result of an experiment. However, in a desktop grid (e.g. the BOINC-based desktop grid), being a highly parallel and volatile system, the coordination of computations is performed by a single project management server, which in many cases is a bottleneck (this effect is discussed in Ilya et al. (2017)). At that, it is important to improve efficiency of the assimilation phase, which is related to operations with workunits database. For the sake of efficiency, the assimilation operations are performed at group clearance basis, that is, all the completed and valid results are assimilated as a group in a single transaction, whereas the results arrived during the assimilation process (if any), are subject to the next assimilation (which, if required, starts immediately after completion of an ongoing one). The constant `one_pass_N_WU` limiting the number of results assimilated per transaction, can be made arbitrary large, thus making the *group clearance* nature of the assimilation phase evident.

Finally, with the current COVID-19 situation affecting the entire planet, we can find an application involving the pooling of the blood samples to test for this disease as well as for antibodies. If the test shows up a positive result then everyone involved in that group needs to be tested either individually or in subgroups. In the latter case, the strategy of subgroup selection is to be optimized (Cheng et al. 2019). The idea of pooling in queues originally appeared in Neuts and Chandramouli (1987), and very recently this model was put in use by a team from the Institute of Virology (Lohse et al. 2020).

The application examples in cloud storage (see, e.g., Saxena et al. (2018)), distributed computing (see, e.g., Ilya et al. (2017)) and blood screening (Lohse et al. 2020) motivate our interest to study a single-server batch arrival queue with a group clearance in which all those customers present at the time of the beginning of a service are taken into service. We are confident that there are other applications in practice where we can implement the *group clearance* rule which places no restriction on the size of the batch served. It should be pointed out that even though we place no restriction on the batch size, if the (finite) batch size goes beyond a certain point the assumption of unlimited size for the batch, for all practical purposes, will have little or no impact. This is like truncating an infinite sum in numerical computations. Often times, the assumption of infinite size leads to explicit or compact results as opposed to finite size problem.

Since the introduction of batch service queueing systems by Bailey (1954), the literature on such systems has grown significantly (see e.g., Abolnikov and Dukhovny (2003), Arumuganathan and Jeyakumar (2005), Baba (1996), Banik et al. (2008), Banik et al. (2009), Bar-Lev et al. (2007), Chakravarthy (1992), Chakravarthy et al. (2017), Chaudhry and Templeton (1983), Chaudhry and Gupta (1999), Claeys et al. (2010b), Germs and van Foreest (2013), Gold and Tran-Gia (1993), Hébuterne and Rosenberg (1999), Neuts (1967), Powell and Humblet (1986), Sikdar and Gupta (2005), Zee et al. (2001), Chakravarthy (1993), Chakravarthy and Dudin (2002a), Chakravarthy and Dudin (2002b), Dudin and Chakravarthy (2002a), Dudin and Chakravarthy (2002b), and Dudin and Chakravarthy (2003) and many others). Many variants of batch arrival/batch service systems are studied in the literature, and we summarize the recent developments in the next subsection.

1.1 Literature Survey

The batch service models have long history. Thus, in this subsection we focus only on the results of the recent decade, and refer the reader to literature surveys in Banerjee et al. (2015) and Sasikala and Indhira (2016) for an extensive discussion.

Two major classes of single-server batch service models studied in recent papers are the *discrete-time* (Claeys et al. 2010a; Claeys et al. 2010b; Claeys et al. 2013; Banerjee et al. 2014; Yu and Alfa 2015; Baetens et al. 2016; Baetens et al. 2017; Baetens et al. 2018; 2019; Panda and Goswami 2020) and *continuous-time* models (Saxena et al. 2018; D'Arienzo et al. 2019; Banerjee and Gupta 2012; Banerjee et al. 2015; Yu and Tang 2018; Pradhan and Gupta 2017; Pradhan et al. 2016; Pradhan and Gupta 2019; Gupta et al. 2020; Gupta and Banerjee 2019; Maity and Gupta 2015; Banik 2015; Vadivu and Arumuganathan 2015; Chaudhry et al. 2016; Jeyakumar and Senthilnathan 2017; Zeng and Xia 2017; Niranjan et al. 2018; Gupta and Banerjee 2018; Panda et al. 2018; Ayyappan and Karpagam 2018; Ayyappan and Nirmala 2018; Bank and Samanta 2020; Xie et al. 2020). The variety of techniques used for the analysis includes Kolmogorov equations, Supplementary variable techniques, Roots method, Matrix-Analytic Method, Embedded Markov chain analysis, Spectral methods, Asymptotic Quasi-Toeplitz Markov chain technique and Game theory, to name a few. Below we summarize a few features of the models studied, and give the corresponding classification of the recent papers with respect to these features. Finally, we outline the papers studying models close to the one studied in the present paper, and highlight the novelty of the present study.

Queue size Most of the papers study the system either with finite queue (FIN), or with infinite queue (INF). A notable exception is the paper (D'Arienzo et al. 2019) where a retrial system is studied.

Arrival process More analytically tractable are the memoryless arrival processes (Poisson in continuous time, and Geometric in discrete time case), but several works focus on general renewal processes, batch Poisson arrivals (Pradhan and Gupta 2017; Jeyakumar and Senthilnathan 2017; Ayyappan and Karpagam 2018; Ayyappan and Nirmala 2018), and, among the most general cases, the batch Markovian arrival process (BMAP) (Banik 2015; Bank and Samanta 2020), which though received less attention.

Service time distribution In the majority of cases, the service time distribution is assumed to be general (e.g. defined by its Laplace–Stieltjes transform), but several exceptions include memoryless (Gupta and Banerjee 2019; Maity and Gupta 2015; Panda et al. 2018; Panda and Goswami 2020; Baetens et al. 2017) and phase-type (PH) ((D'Arienzo et al. 2019)), which allow to obtain explicit results. Note that in discrete time models, single slot service is also used, see Claeys et al. (2010a), Baetens et al. (2016), and Baetens et al. (2018).

Batch service type The most widely studied is the classical General Bulk Service (GBS) rule introduced in Neuts (1967): the two finite constant threshold policy states that the server starts service of a batch of size larger or equal to $a \geq 1$, and can handle up to $b \geq a$ customers in a batch, with b finite. Some variations of this classical rule are: the possibility of serving a batch of size less than a with some probability (Saxena et al. 2018; Claeys et al. 2013), or a single customer service mode for batch smaller than a (Yu and Alfa 2015). A specific version of the GBS rule (when $a = b$) is the Fixed Bulk Service (FBS) policy (Claeys et al. 2010a; Yu and Tang 2018; Gupta and Banerjee 2019; Xie et al. 2020). A significant attention is also received by the systems with variable capacity or the so-called versatile Bulk Service rule, where the batch size taken for service is random (Pradhan et al. 2016; Maity and Gupta 2015; Jeyakumar and Senthilnathan 2017; Bank and Samanta 2020). Another interesting setup is considered in a series of papers (Baetens et al. 2016 2017, 2018, 2019), where the so-called two-class service policy was introduced. The model though is rather different from classical single-server batch service queue, since it is essentially a multiclass model. However, the policy is such that a sequence of customers of same class are taken into service, without an upper bound on the size of the batch, which is rather different from the GBS rule.

Size dependence Both variants are considered: the service time of a batch is thought to be dependent or independent of the batch size. The latter case is obvious for the FBS policy,

Model peculiarities Several papers include some specific model features, such as queue-length dependence of the service time distribution (Gupta and Banerjee 2018), vacations (Vadivu and Arumuganathan 2015; Jeyakumar and Senthilnathan 2017; Niranjan et al. 2018; Ayyappan and Nirmala 2018; Panda et al. 2018), breakdowns (Niranjan et al. 2018; Ayyappan and Karpagam 2018) and inter-customer correlation (Baetens et al. 2017).

Using this classification, we conclude that the present study considers a continuous time $BMAP/G^{(1,\infty)}/1$ model with size-independent service times and GBS service rule with $a = 1$ and $b = \infty$ which we call *group clearance*. The most similar models to the one considered in the present paper were studied in Banik (2015), Baetens et al. (2019), Saxena et al. (2018), and Zeng and Xia (2017). However, we stress the following differences:

- In the paper (Banik 2015) the $BMAP/G^{(a,b)}/1$ with a classical GBS discipline, given b finite, was treated by the matrix-analytic method. In contrast, in the present paper, the batch size taken into service is not bounded by finite b. It is not clear if the analysis in Banik (2015) can be extended to the case $b = \infty$ straightforwardly. However, likely that some asymptotic results can be obtained from the solution in Banik (2015) for b large, which could be an interesting point for further study.

- Though the batch service discipline in Baetens et al. (2019) allows one to take a batch of unbounded size for service, the discipline is rather specific, and the present model can only be obtained as a limiting (trivial) case of a single customer class. Moreover, the analysis in Baetens et al. (2019) is held in discrete time, and continuous time results may only be obtained in limiting case of the slot size approaching zero, which leads to complicated analysis.
- The paper (Saxena et al. 2018) analyzes a discrete-time model with general interarrival and service time distributions (the latter are dependent on the batch size), vacations and exhaustive service policy with a specific probabilistic batching and infinite upper bound on batch size. This discrete-time model resembles our model and generalizes it in some points. However, due to these generalizations, the analytical results are obtained in terms of generating functions, while in our model we obtain explicit results. However, it might be interesting to obtain our model as a limiting special case when the slot size goes to zero.
- The analysis in Zeng and Xia (2017) considers a finite queue system, and moreover, the arrival process is Poisson. However, the explicit results obtained in Zeng and Xia (2017) are interesting to be compared with the results of our analysis in case of a large queue capacity.

Summarizing the literature survey, to the best of our knowledge, the batch arrival and group clearance single server queueing model with general service times is studied for the first time in this paper.

1.2 Structure of the Paper

The structure of the paper is the following. The model under study is described in Section 2 and in Section 3 we study the model with general distribution of group clearance. We also provide the stationary performance measures of the system. In Section 3.1, we perform a busy period analysis. In Section 4, we perform a deeper analysis of the case when service time has a phase type distribution (we give the details on this type of distribution in the section), and provide explicit solutions for some special cases. A number of system performance measures are listed in Section 4.1. In Section 5, we provide some useful insights by means of illustrative numerical experiments, which include a wide range of service distributions including constant, uniform, Weibull among others, and some concluding remarks are given in Section 6.

2 Model Description

In this paper, we study a single-server queueing model in which the arrivals occur according to a versatile Markovian point process, namely, batch Markovian arrival process ($BMAP$), and services are offered *simultaneously* to all the customers present in the system at the service starting epoch. The service times are assumed to be generally distributed and is *independent* of the size of the group being served. That is, the system operates under the bulk service rule with infinite upper bound on the batch size. The server stays idle if at the completion of a service there are no customers waiting in the system. In Kendall notation, the system can be classified as $BMAP/G^{(1,\infty)}/1$.

The $BMAP$, which is dense in the class of arrival processes with domain on non-negative real axis, allows us to incorporate the correlation between inter-arrival times. The process goes through transitions among, say, m phases as follows. These transitions are governed by a sequence of square matrices, say, $\{D_k, k \geq 0\}$ of size m, such that D_0 governs transitions corresponding to no arrival and D_k governs those that correspond to an arrival of a batch (of size k, $k \geq 1$) of customers(for more details on this type of processes (see e.g., (Bladt and Nielsen 2017; Chakravarthy 2011; 2001; He 2014; Chakravarthy 2015)). The generator Q of the underlying Continuous-Time Markov Chain ($CTMC$, related to transitions of phase irrespective of the number of arrivals) is given by $Q = \sum_{k=0}^{\infty} D_k$. Let $\boldsymbol{\pi}$ be the stationary probability vector of the $CTMC$ uniquely satisfying

$$\boldsymbol{\pi} Q = \mathbf{0}, \quad \boldsymbol{\pi} \boldsymbol{e} = 1, \tag{1}$$

(where $\boldsymbol{e}$ is a column vector of ones, and $\mathbf{0}$ is the vector of zeros). Letting $D = Q - D_0$, the quantity $\lambda_g = \boldsymbol{\pi} D \boldsymbol{e}$ gives the arrival rate of batches and $\lambda = \boldsymbol{\pi} \sum_{k=1}^{\infty} k D_k \boldsymbol{e}$ is the expected number of customer arrivals per unit time. The service times of the group of customers present at the service starting epoch are iid., with distribution function $H(.)$ and mean μ^{-1}, and thus the traffic intensity is given by $\rho = \frac{\lambda_g}{\mu}$. Note the model studied in this paper is always stable for any $\rho > 0$.

3 Embedded Markov renewal process

In this section we study the model with a general service time distribution with CDF, $H(.)$, by means of the embedded Markov process method, introduced in Kendall (1953). Consider the embedded Markov renewal process $\{N_i, J_i, \tau_i, \;\; i \geq 0\}$ at departure epochs (with $N_i \geq 0$ being the number of customers, $J_i \in \{1, \ldots, m\}$ being the phase of the $BMAP$ right after departure, and $\tau_i \geq 0$ being the inter-departure times). The transition probability matrix $\tilde{P}(x)$ of the embedded process (transition kernel) consisting of elements $P_{nj}(x) = \mathrm{P}\{N_i = n, J_i = j, \tau_i \leq x\}, n \geq 0, 1 \leq j \leq m, x \geq 0$, is of the form

$$\tilde{P}(x) = \begin{bmatrix} B_0(x) & B_1(x) & B_2(x) & \cdots \\ A_0(x) & A_1(x) & A_2(x) & \cdots \\ A_0(x) & A_1(x) & A_2(x) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the (square) matrices of size m

$$B_i(x) = \int_0^x e^{D_0(x-t)} D A_i(t) dt, \;\; i \geq 0, \tag{2}$$

correspond to the probabilities that a departure at time 0 left the system empty and that the next departure (which occurs no later than time x) leaves behind i in the system and

$$A_i(x) = \int_0^x P(i, t) dH(t), \;\; i \geq 0, \tag{3}$$

are matrices (of dimension m) related to a service time (lasting no more than time x) of a group of customers that started right after the previous service completion epoch and during which time exactly i customers arrive. The elements $P_{ij}(n, t)$ of the matrix $P(n, t)$ are probabilities of n arrival events in $BMAP$ during time t, given the phase transition from i to j; $i, j = 1, \ldots, m$. The necessary details on the computation of $P(n, t)$ were first described in Neuts and Li (1996).

The transition probability matrix, $\hat{P} = \tilde{P}(\infty)$, of the corresponding embedded Markov chain (related to transitions of the process $\{N_i, J_i\}_{i\geq 0}$) is given by

$$\hat{P} = \begin{bmatrix} B_0 & B_1 & B_2 & \cdots \\ A_0 & A_1 & A_2 & \cdots \\ A_0 & A_1 & A_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where it follows from Eqs. 2 and 3, that

$$A_i = \int_0^\infty P(i,t)dH(t), \quad B_i = (-D_0)^{-1}DA_i, \quad i \geq 0.$$

Note that

$$A = \sum_{i=0}^{\infty} A_i = \int_0^\infty e^{Qt}dH(t), \quad B = \sum_{i=0}^{\infty} B_i = (-D_0)^{-1}DA. \tag{4}$$

It can easily be verified that A and B are stochastic matrices.

Let $\boldsymbol{x}$, partitioned as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots)$, be the steady-state probability vector of $\hat{P}$, where the j^{th} component of $\boldsymbol{x}_k$ gives the steady-state probability that soon after a departure, the system will have k customers waiting in the queue with the arrival process in phase j, $1 \leq j \leq m, k \geq 0$. That is,

$$\boldsymbol{x}\hat{P} = \boldsymbol{x}, \quad \boldsymbol{x}\boldsymbol{e} = 1. \tag{5}$$

Using the structure of $\boldsymbol{x}$, it is easy to transform Eq. 5 into the following steady-state equations:

$$\boldsymbol{x}_i = \boldsymbol{x}_0 B_i + \sum_{k=1}^{\infty} \boldsymbol{x}_k A_i, \quad i \geq 0, \tag{6}$$

subject to normalizing condition

$$\sum_{k=0}^{\infty} \boldsymbol{x}_k \boldsymbol{e} = 1.$$

Moreover, Eq. 6 allows to obtain explicit expressions for the steady-state probability vector $\boldsymbol{x}$.

Theorem 1 *The steady-state probability vector* $\boldsymbol{x}$ *is obtained as*

$$\begin{aligned} \boldsymbol{x}_0 &= \boldsymbol{u}A_0[I - (-D_0)^{-1}DA_0]^{-1}, \\ \boldsymbol{x}_i &= \boldsymbol{x}_0(-D_0)^{-1}DA_i + \boldsymbol{u}A_i, \ i \geq 1, \end{aligned} \tag{7}$$

where the vector $\boldsymbol{u}$ *is the unique solution to*

$$\boldsymbol{u}\Big[I - A + A_0(I - B_0)^{-1}(I - B)\Big] = \boldsymbol{0}, \tag{8}$$

subject to the normalizing condition

$$\boldsymbol{u}A_0[I - (-D_0)^{-1}DA_0]^{-1}\boldsymbol{e} + \boldsymbol{u}\boldsymbol{e} = 1.$$

Proof Denoting

$$\boldsymbol{u} = \sum_{k=1}^{\infty} \boldsymbol{x}_k,$$

the equations in Eq. 6 can be rewritten as

$$\begin{aligned} \boldsymbol{x}_0 &= \boldsymbol{x}_0 B_0 + \boldsymbol{u} A_0, \\ \boldsymbol{u} &= \boldsymbol{x}_0 (B - B_0) + \boldsymbol{u}(A - A_0), \end{aligned} \tag{9}$$

from which the stated result follows after routine simplifications. □

Interestingly, the vector $\boldsymbol{u}$ may be intuitively explained as the stationary probability vector of seeing at least one customer in the system at a departure epoch with the arrival process in various phases. By the group clearance property of the service discipline, the balance Eq. 9 in fact correspond to a reduced state Markov chain $\{\tilde{I}\{N_i > 0\}, J_i\}_{i \geq 0}$, where $\tilde{I}\{\cdot\}$ is the indicator function.

It is easy to verify that the inverses appearing in Eqs. 7 and 8 do indeed exist. Note also that the matrices $A_0 = \int_0^\infty e^{D_0 t} dH(t)$ and $A = \int_0^\infty e^{Qt} dH(t)$ can be obtained efficiently using uniformization method, which was first introduced by Jensen (1953). We will briefly outline some key steps and more details can be found in many books and papers (see e.g. He (2014)).

Suppose that $\tilde{\theta} = \max\limits_i |(D_0)_{i,i}|$ and

$$\gamma_n = \int_0^\infty e^{-\tilde{\theta} t} \frac{(\tilde{\theta} t)^n}{n!} dH(t), \ n \geq 0. \tag{10}$$

Defining $K_1 = I + \tilde{\theta}^{-1} D_0$ and $K_2 = I + \tilde{\theta}^{-1} Q$, we get

$$A_0 = \sum_{n=0}^\infty \gamma_n K_1^n, \quad A = \sum_{n=0}^\infty \gamma_n K_2^n. \tag{11}$$

The above infinite sums, for computational implementation, may be truncated at some n^* (based on the tail probabilities of γ_n), to guarantee that for a given $\epsilon > 0$, the condition, $\sum_{n=0}^{n^*} \gamma_n > 1 - \epsilon$, holds good.

3.1 Busy Period Analysis

In this section we consider the embedded Markov renewal process to obtain busy period distribution, as well as mean service completions (and customers served) during a busy period. Let G denote the matrix such that the $(j,k)^{th}$ element g_{jk}, $1 \leq j, k \leq m$, gives the probability that the underlying Markov renewal process eventually visits level **0** by visiting state $(0, k)$ for the first time starting from state $(1, j)$. Conditioning on the number of arrivals during a single transition (service time of a single group), and recalling the group clearance, it is easy to see that

$$G = A_0 + A_1 G + A_2 G + \cdots,$$

which implies that

$$G = (I - A + A_0)^{-1} A_0, \tag{12}$$

and it is easy to verify that $G\boldsymbol{e} = \boldsymbol{e}$.

Now let $\tilde{G}(n, x) = (\tilde{g}_{jk}(n, x), 1 \leq j, k \leq m), \ n \geq 1, x \geq 0$, be such that $\tilde{g}_{jk}(n, x)$ gives the conditional probability that the first passage time from $(1, j)$ to $(0, k)$, occurs no later than time x and exactly in n transitions. Let

$$G^*(z, s) = \sum_{n=1}^\infty z^n \int_0^\infty e^{-sx} d\tilde{G}(n, x).$$

Using the special structure induced by group clearance, it can be verified that

$$G^*(z,s) = z[A_0(s)+A_1(s)G^*(z,s)+A_2(s)G^*(z,s)+\cdots] = z[A_0(s)+(A(s)-A_0(s))G^*(z,s)], \quad (13)$$

where $A(s) = \sum_{i=0}^{\infty} A_i(s)$. Hence

$$G^*(z,s) = z[I - z(A(s) - A_0(s))]^{-1}A_0(s).$$

We note that

$$A_k(s) = \int_0^{\infty} e^{-st} dA_k(t),\ k \geq 0,$$

and, consistently with Eq. 12,

$$G = G^*(1,0) = [I - A + A_0]^{-1}A_0.$$

To obtain the sequence, $\hat{G}_n$, of the number of transitions (service completions) in a busy period, we look at $G^*(z,0)$. From Eq. 13, we get

$$G^*(z,0) = z[I - z(A - A_0)]^{-1}A_0 = \sum_{k=0}^{\infty} z^{k+1}(A - A_0)^k A_0,$$

and hence

$$\hat{G}_n = (A - A_0)^{n-1}A_0, \quad n \geq 1.$$

Note that the Laplace-Stieltjes transform of the distribution of the busy period is given by $G^*(1,s)$. It is easy to verify from Eq. 13 that

$$G^*(1,s) = [I - (A(s) - A_0(s))]^{-1}A_0(s).$$

Also, it can be verified that $\tilde{\boldsymbol{\mu}}$, the vector of order m, whose j^{th} component gives the mean number of service completions during a busy period given that the arrival process was in phase $j = 1, \ldots, m$ at the beginning of the busy period, can be obtained as

$$\tilde{\boldsymbol{\mu}} = \sum_{n=0}^{\infty} n\hat{G}_n \boldsymbol{e} = \sum_{n=1}^{\infty} n(A - A_0)^{n-1}A_0\boldsymbol{e},$$

which gives

$$\tilde{\boldsymbol{\mu}} = (I - A + A_0)^{-1}\boldsymbol{e}. \quad (14)$$

Alternatively, one can get an expression for $\tilde{\boldsymbol{\mu}}$ as follows.

$$\tilde{\boldsymbol{\mu}} = \frac{\partial G^*(z,s)}{\partial z}\boldsymbol{e}\bigg|_{z=1,s=0} = (I-A+A_0)^{-1}A_0\boldsymbol{e} + (A-A_0)(I-A+A_0)^{-2}A_0\boldsymbol{e} = (I-A+A_0)^{-1}\boldsymbol{e}.$$

Note that the vector of mean number of service completions during a busy period doesn't depend on the type of the batch size distribution as well as the parameter of the distribution. This is intuitively clear since we are looking at the number of service completions. However, the mean number of customers served during a busy period will depend on the average number of customers per batch arrival.

Suppose we denote by $\boldsymbol{\delta}$ the vector of size m such that its j^{th} component gives the mean duration of the busy period which started with the phase of the arrival process in state j. Then, we have,

$$\boldsymbol{\delta} = -\frac{\partial G^*(z,s)}{\partial s}\boldsymbol{e}\bigg|_{z=1,s=0} = [I-A+A_0]^{-1}(-A_0'(0))\boldsymbol{e} + [I-A+A_0]^{-1}(-A'(0)+A_0'(0))\boldsymbol{e}, \quad (15)$$

where

$$-A_0'(0) = \int_0^\infty t\, dA_0(t) = \int_0^\infty t\, exp(D_0 t)\, dH(t),$$

and

$$-A'(0) = \int_0^\infty t\, dA(t) = \int_0^\infty t\, exp(Qt)\, dH(t).$$

Thus, the following intuitively obvious result can be deduced from Eq. 15.

$$\boldsymbol{\delta} = \frac{1}{\mu}[I - A + A_0]^{-1}\boldsymbol{e} = \frac{1}{\mu}\tilde{\boldsymbol{\mu}}.$$

4 Generator approach

In this section we apply the celebrated matrix analytic method of Neuts (1981) to study the model in detail under the assumption of phase type (PH) distribution for the group service time, having representation (β, S) of order n. Recall, that the PH distribution is the time until absorption of a finite-state $CTMC$ with an absorbing state, where β is the initial state distribution (and a basic assumption $\beta\boldsymbol{e} = 1$ guarantees no atom at zero), S is square matrix of order n defining the transition rates between n non-absorbing states (phases), whereas $\boldsymbol{S}^0 := -S\boldsymbol{e}$ is the vector of absorption rates. Recall also that the mean service time is given by $\mu^{-1} = \beta(-S)^{-1}\boldsymbol{e}$.

We study the queueing model as a $CTMC$. Let $N(t)$ be the number of customers *waiting in the queue* at time t, $J_1(t)$ be the phase of the service process at time t should the server happens to be busy at that time, and $J_2(t)$ be the phase of the arrival process at time t. Then the triplet $\{N(t), J_1(t), J_2(t)\}_{t \geq 0}$ is a $CTMC$ with state space given by

$$\Omega = \{(0, k) : 1 \leq k \leq m\} \bigcup \{(i, j, k) : i \geq 0, 1 \leq j \leq n, 1 \leq k \leq m\}.$$

For convenience, we partition the set Ω into subsets $\mathbf{0}^* = \{(0, k) : 1 \leq k \leq m\}$, corresponding to an empty system, and $\mathbf{i} = \{(i, j, k) : i \geq 0, 1 \leq j \leq n, 1 \leq k \leq m\}$, for $i \geq 0$, being the number of customers in the *queue*. The generator, $\tilde{Q}$, of the $CTMC$ is given by

$$\tilde{Q} = \begin{bmatrix} D_0 & \beta \otimes D & 0 & 0 & 0 & 0 & \cdots \\ \boldsymbol{S}^0 \otimes I & S \oplus D_0 & I \otimes D_1 & I \otimes D_2 & I \otimes D_3 & I \otimes D_4 & \cdots \\ 0 & \boldsymbol{S}^0\beta \otimes I & S \oplus D_0 & I \otimes D_1 & I \otimes D_2 & I \otimes D_3 & \cdots \\ 0 & \boldsymbol{S}^0\beta \otimes I & 0 & S \oplus D_0 & I \otimes D_1 & I \otimes D_2 & \cdots \\ 0 & \boldsymbol{S}^0\beta \otimes I & 0 & 0 & S \oplus D_0 & I \otimes D_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

We note that $\otimes$ and $\oplus$ are Kronecker product and sum, respectively, and the special structure of the matrix $\tilde{Q}$ is related to the properties of the model. Namely, the main diagonal is related to transitions of the $CTMC$ without any arrivals or departures, and thus is related to phase changes of $J_1(t)$ or $J_2(t)$, except the element D_0 corresponding to subset $\mathbf{0}^*$ describing the phase change of $J_2(t)$ only (since the system is empty). The values $\boldsymbol{S}^0\beta \otimes I$ correspond to a service completion of a group of customers and an initiation of a service for a new group, whereas $\boldsymbol{S}^0 \otimes I$ is related to the clearance of a group that leaves the system empty. Finally, the values $I \otimes D_i$ correspond to an arrival of a batch of size i during a service completion, whereas $\beta \otimes D$ corresponds to an arrival of a group (with no regard to the batch size) to an empty system that initiates a service.

Let $\boldsymbol{y}$ partitioned as $\boldsymbol{y} = (\boldsymbol{y}^*, \boldsymbol{y}_0, \boldsymbol{y}_1, \cdots)$ be such that

$$\boldsymbol{y}\tilde{Q} = \mathbf{0}, \quad \boldsymbol{y}\boldsymbol{e} = 1. \tag{16}$$

The following theorem gives an expression for $\boldsymbol{y}$ which is explicit up to a vector $\boldsymbol{y}_0$.

Theorem 2 *The steady-state probability vector* $\boldsymbol{y}$ *is obtained as*

$$\begin{aligned} \boldsymbol{y}^* &= \boldsymbol{y}_0(\boldsymbol{S}^0 \otimes (-D_0)^{-1}), \\ \boldsymbol{y}_i &= \textstyle\sum_{j=0}^{i-1}\boldsymbol{y}_j(I \otimes D_{i-j})(-S \oplus D_0)^{-1}, \ i \geq 1, \end{aligned} \tag{17}$$

and $\boldsymbol{y}_0$ *is obtained by solving*

$$\begin{aligned} &\boldsymbol{y}_0\left[(\boldsymbol{S}^0\beta \otimes (-D_0)^{-1}D) + S \oplus D_0 + (I \otimes D)(-S \oplus Q)^{-1}(\boldsymbol{S}^0\beta \otimes I)\right] = \mathbf{0}, \\ &\boldsymbol{y}_0\left[\boldsymbol{e} + (\boldsymbol{S}^0 \otimes (-D_0)^{-1}\boldsymbol{e}) + (I \otimes D)(-S \oplus Q)^{-1}\boldsymbol{e}\right] = 1. \end{aligned}$$

Proof The equations corresponding to Eq. 16 are given by

$$\begin{gathered} \boldsymbol{y}^* D_0 + \boldsymbol{y}_0(\boldsymbol{S}^0 \otimes I) = \mathbf{0}, \\ \boldsymbol{y}^*(\beta \otimes D) + \boldsymbol{y}_0(S \oplus D_0) + \textstyle\sum_{k=1}^{\infty} \boldsymbol{y}_k(\boldsymbol{S}^0\beta \otimes I) = \mathbf{0}, \\ \textstyle\sum_{j=0}^{i-1}\boldsymbol{y}_j(I \otimes D_{i-j}) + \boldsymbol{y}_i(S \oplus D_0) = \mathbf{0}, \ i \geq 1, \\ \boldsymbol{y}^*\boldsymbol{e} + \textstyle\sum_{i=0}^{\infty}\boldsymbol{y}_i\boldsymbol{e} = 1. \end{gathered} \tag{18}$$

Suppose we denote

$$\boldsymbol{v} = \sum_{i=1}^{\infty} \boldsymbol{y}_i.$$

Then, the equations in Eq. 18 can be rewritten as

$$\begin{gathered} \boldsymbol{y}^* D_0 + \boldsymbol{y}_0(\boldsymbol{S}^0 \otimes I) = \mathbf{0}, \\ \boldsymbol{y}^*(\beta \otimes D) + \boldsymbol{y}_0(S \oplus D_0) + \boldsymbol{v}(\boldsymbol{S}^0\beta \otimes I) = \mathbf{0}, \\ \boldsymbol{y}_0(I \otimes D) + \boldsymbol{v}(S \oplus Q) = \mathbf{0}, \\ \boldsymbol{y}^*\boldsymbol{e} + \boldsymbol{y}_0\boldsymbol{e} + \boldsymbol{v}\boldsymbol{e} = 1. \end{gathered} \tag{19}$$

The stated result follows immediately after routine simplifications. □

Note that the special structure of the coefficient matrices involving Kronecker products and sums can be exploited in obtaining the steady-state vector, whereas Eq. 17 provides a recursive procedure to obtain $\boldsymbol{y}_i, \ i \geq 1$. Moreover, since the vector $\boldsymbol{v}$ can be obtained from the system Eq. 19 in advance, it is easy to find the truncation point, say, i^*, for the vectors $\boldsymbol{y}_i$ computation with given accuracy. The following intuitive results serve as accuracy checks in numerical computation.

Theorem 3 *We have*

$$\begin{gathered} \boldsymbol{y}^* + \textstyle\sum_{i=0}^{\infty}\boldsymbol{y}_i(\boldsymbol{e} \otimes I) = \boldsymbol{\pi}, \\ \dfrac{1}{(1 - \boldsymbol{y}^*\boldsymbol{e})}\textstyle\sum_{i=0}^{\infty}\boldsymbol{y}_i(I \otimes \boldsymbol{e}) = \mu\beta(-S)^{-1}, \end{gathered} \tag{20}$$

where $\boldsymbol{\pi}$ *is as defined in Eq.* 1.

Proof Note that the first equation in Eq. 18 can be transformed into

$$\boldsymbol{y}^*(\beta \otimes D_0) + \boldsymbol{y}_0(\boldsymbol{S}^0\beta \otimes I) = \mathbf{0}.$$

Now adding the above equation to the second and third (over all i) equations in Eq. 18, we get

$$(\beta \otimes \boldsymbol{y}^* Q) + \sum_{i=0}^{\infty} \boldsymbol{y}_i[(S + \boldsymbol{S}^0\beta) \oplus Q] = \mathbf{0}. \tag{21}$$

Post-multiplying Eq. 21 by $\boldsymbol{e} \otimes I$ and using the uniqueness of $\boldsymbol{\pi}$ we get the first statement in Eq. 20. Similarly, post-multiplying Eq. 21 by $I \otimes \boldsymbol{e}$ and using the uniqueness of the steady-state vector of $S + \boldsymbol{S}^0\beta$, we get the second stated result in Eq. 20. □

The steady-state probability vectors at arrival and departure epochs are given in the following theorem.

Theorem 4 *The steady-state probability vector $\boldsymbol{x}$ at departure epochs partitioned as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots)$ is given by*

$$\boldsymbol{x}_k = [\mu(1 - \boldsymbol{y}^*\boldsymbol{e})]^{-1}\boldsymbol{y}_k(\boldsymbol{S}^0 \otimes I), \ k \geq 0. \tag{22}$$

The steady-state probability vector $\boldsymbol{z}$ at arrival epochs, partitioned as $\boldsymbol{z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots)$, is given by

$$\boldsymbol{z}^* = \frac{1}{\lambda_g}(\boldsymbol{y}^* D), \quad \boldsymbol{z}_k = \frac{1}{\lambda_g}\boldsymbol{y}_k(I \otimes D), \ k \geq 0.$$

Proof Follows immediately from the definition of the steady-state vectors. □

Finally, let W_s denote the sojourn time in the system (STS) of a customer. The following theorem shows that STS has PH distribution.

Theorem 5 *W_s follows a PH distribution with representation given by (γ, L) of order $3n$ where*

$$\boldsymbol{\gamma} = \left(\boldsymbol{z}^*\boldsymbol{e}\beta, \sum_{k=0}^{\infty} \boldsymbol{z}_k(I \otimes \boldsymbol{e}), \mathbf{0}\right),$$

and

$$L = \begin{bmatrix} S & 0 & 0 \\ 0 & S & \boldsymbol{S}^0\beta \\ 0 & 0 & S \end{bmatrix}.$$

Proof First note that with probability $\boldsymbol{z}^*\boldsymbol{e}$ the arriving customer enters into the service immediately. So, in this case the sojourn time in the system is nothing but the service time. However, if the arrival sees the server busy then the customer has to wait for the current service to be over and then get into service. Thus, we can model the STS as a mixture of two phase type distributions and by a well-known result (see Neuts 1981) is again of phase type. Thus, the stated result follows. □

Note: It is clear from the representation of the $PH-$distribution that the sojourn time is independent of batch size distribution as well as the mean batch size.

4.1 The System Performance Measures

In this section we list a number of system performance measures of interest for $BMAP/PH/1$ along with their expressions.

1. *Probability that the server is idle.* The probability, P_{idle}, that the server is idle at an arbitrary time is given by

$$P_{idle} = \boldsymbol{y}^*\boldsymbol{e}.$$

It is important to point out that P_{idle} does not depend on the batch size distribution. This fact can be seen immediately from the steady-state equations given in Eq. 19. This observation can also be used as an internal accuracy check in numerical computation.

2. *Mean number of customers in the queue.* The mean, μ_{NQ}, number of customers in the queue is given by

$$\mu_{NQ} = \sum_{i=1}^{\infty} i\,\boldsymbol{y}_i\boldsymbol{e}.$$

Note that the measure, μ_{NQ}, depends only on the average batch size and not on the batch size distribution. To see this, we see from the third equation in Eq. 18,

$$\sum_{i=1}^{\infty} i\,\boldsymbol{y}_i = \sum_{j=0}^{\infty} \boldsymbol{y}_j\Big(I \otimes \sum_{k=1}^{\infty} kD_k\Big)(-S \oplus Q)^{-1},$$

which immediately implies that μ_{NQ} does not depend on the batch size distribution. Again, this observation can be used as another accuracy check in numerical computation.

3. *Mean number of customers in service.* A novel way to compute the mean number of customers in service is to look at the weighted average of the average batch size of an arrival and the (conditional) average number of customers left behind a departure conditioned on the fact that at least one customer is seen in the queue. The weights are given by $\boldsymbol{x}_0\boldsymbol{e}$ and $(1 - \boldsymbol{x}_0\boldsymbol{e})$. Thus, the mean, μ_{NS}, number of customers under service is given by

$$\mu_{NS} = \frac{\lambda}{\lambda_g}\boldsymbol{x}_0\boldsymbol{e} + \sum_{i=1}^{\infty} i\,\boldsymbol{x}_i\boldsymbol{e}.$$

The fact that μ_{NS} also depends only on the average batch size and not on the batch size distribution can be seen immediately from Eq. 22.

4. *Mean number of customers in the system.* The mean, μ_S, number of customers in the system is given by

$$\mu_S = \mu_{NQ} + \mu_{NS} = \frac{\lambda}{\lambda_g}\boldsymbol{x}_0\boldsymbol{e} + \sum_{i=1}^{\infty} i(\boldsymbol{y}_i\boldsymbol{e} + \boldsymbol{x}_i\boldsymbol{e}).$$

5. *Mean STS of a customer.* The mean, μ_{W_s}, STS of a customer is given by

$$\mu_{W_s} = \boldsymbol{\gamma}(-L)^{-1}\boldsymbol{e} = \tfrac{1}{\mu}z^*\boldsymbol{e} + \sum_{k=0}^{\infty} z_k(I \otimes \boldsymbol{e})(-S^{-1}\boldsymbol{e}) + \tfrac{1}{\mu}\sum_{k=0}^{\infty} z_k\boldsymbol{e} = \tfrac{1}{\mu} + \sum_{k=0}^{\infty} z_k(-S^{-1}\boldsymbol{e}) \otimes \boldsymbol{e}).$$

6. *Variance of STS of customers*. The variance, $\sigma^2_{W_s}$, of STS of customers is given by

$$\begin{aligned}\sigma^2_{W_s} &= 2\boldsymbol{\gamma}(-L)^{-2}\boldsymbol{e} - \mu^2_{WS} = \\ &= 2\beta(-S)^{-2}\boldsymbol{e} + 2\sum_{k=0}^{\infty} \boldsymbol{z}_k \left[(-S)^{-2}\boldsymbol{e}\otimes\boldsymbol{e} + \frac{1}{\mu}\left((-S)^{-1}\boldsymbol{e}\otimes\boldsymbol{e}\right) - \mu^2_{WS}\right] \\ &= \sigma^2 + 2\sum_{k=0}^{\infty} \boldsymbol{z}_k((-S)^{-2}\boldsymbol{e}\otimes\boldsymbol{e}) - \left[\sum_{k=0}^{\infty} \boldsymbol{z}_k((-S)^{-1}\boldsymbol{e}\otimes\boldsymbol{e})\right]^2.\end{aligned}$$

7. *Mean number of service completions during a busy period*. The mean, μ_{SC}, number of service completions during a busy period is obtained using the fact that the j^{th} component of the vector $\frac{1}{\boldsymbol{y}^*D\boldsymbol{e}}\boldsymbol{y}^*D$ gives the probability a busy period starts with the arrival process in phase $j, 1 \le j \le m$. Thus, we have

$$\mu_{SC} = \frac{1}{\boldsymbol{y}^*D\boldsymbol{e}}\boldsymbol{y}^*D\tilde{\boldsymbol{\mu}},$$

where $\tilde{\boldsymbol{\mu}}$ is as given in Eq. 14.

8. *Mean number of customers served during a busy period*. The mean, μ_{SR}, number of customers served during a busy period is obtained simply as

$$\mu_{SR} = \mu_{SC}\mu_{NS}.$$

4.2 Special Cases:

In this section we discuss special classes of $BMAP$ arrival processes and PH-type service time distributions.

Theorem 6 *In an $M^{[X]}/M/1$-type model define $\rho = \frac{\lambda_g}{\mu}$, and denote p_i the probability that an arriving batch is of size i. Then the steady-state probability vector $\boldsymbol{y}$ (and hence $\boldsymbol{x}$ and $\boldsymbol{z}$) is obtained explicitly as follows.*

$$y^* = \frac{1}{1+\rho+\rho^2}, \quad y_0 = \frac{\rho}{1+\rho+\rho^2}, \quad y_i = \frac{1}{1+\rho}\sum_{j=0}^{i-1} y_j p_{i-j}, \; i \ge 1.$$

Proof First note that in this case $D_0 = -\lambda_g$, $D_i = \lambda_g p_i$; $i \ge 1$, $D = \lambda$, $S = -\mu$, $\beta = 1$, and hence the steady-state equations given in Eq. 18 reduce to

$$\begin{aligned}&-\lambda_g y^* + \mu y_0 = 0,\\ &\lambda_g y^* - (\lambda_g + \mu)y_0 + \mu\textstyle\sum_{i=1}^{\infty} y_i = 0,\\ &\lambda_g\textstyle\sum_{j=0}^{i-1} y_j p_{i-j} - (\lambda_g + \mu)y_i = 0, \; i \ge 1,\\ &y^* + \textstyle\sum_{i=0}^{\infty} y_i = 1.\end{aligned}$$

The stated result follows immediately. □

Theorem 7 *In an $M^{[X]}/PH/1$-type model, the steady-state probability vector $\boldsymbol{y}$ (and hence $\boldsymbol{x}$ and $\boldsymbol{z}$) is obtained explicitly as follows.*

$$\begin{aligned}&\boldsymbol{y}_0 = \lambda_g y^*\beta(\lambda I - \lambda_g\boldsymbol{e}\beta - S)^{-1}\\ &\boldsymbol{y}_i = \lambda_g\textstyle\sum_{j=0}^{i-1} \boldsymbol{y}_j p_{i-j}(\lambda_g I - S)^{-1}, \; i \ge 1,\end{aligned}$$

where y^* is obtained using the normalizing condition.

Proof In this case, the steady-state equations in Eq. 18 are simplified to

$$-\lambda_g y^* + \boldsymbol{y}_0 \boldsymbol{S}^0 = 0, \tag{23}$$

$$\lambda_g y^* \beta + \boldsymbol{y}_0 (S - \lambda_g I) + \sum_{i=1}^{\infty} \boldsymbol{y}_i \boldsymbol{S}^0 \beta = \boldsymbol{0}, \tag{24}$$

$$\lambda_g \sum_{j=0}^{i-1} \boldsymbol{y}_j p_{i-j} + \boldsymbol{y}_i (S - \lambda_g I) = \boldsymbol{0},\ i \geq 1,$$

subject to the normalizing condition

$$y^* + \sum_{i=0}^{\infty} \boldsymbol{y}_i \boldsymbol{e} = 1.$$

It follows from Eq. 20 that

$$\sum_{i=0}^{\infty} \boldsymbol{y}_i = \mu(1 - y^*)\beta(-S)^{-1},$$

which implies

$$\sum_{i=0}^{\infty} \boldsymbol{y}_i \boldsymbol{S}^0 = \mu(1 - y^*).$$

Then, from Eq. 23 we get

$$\sum_{i=1}^{\infty} \boldsymbol{y}_i \boldsymbol{S}^0 = \mu(1 - y^*) - \lambda_g y^*. \tag{25}$$

Also, Eqs. 23 and 24 give

$$\lambda_g \boldsymbol{y}_0 \boldsymbol{e} = \sum_{i=1}^{\infty} \boldsymbol{y}_i \boldsymbol{S}^0.$$

Then, using Eqs. 24 and 25, we obtain

$$\lambda_g y^* \beta + \boldsymbol{y}_0 (S - \lambda_g I + \lambda_g \boldsymbol{e} \beta) = \boldsymbol{0},$$

from which the stated results follows immediately. □

Theorem 8 *In an $PH^{[X]}/M/1$-type model, with PH interarrival time distribution having representation ($\boldsymbol{\alpha}$, T), the steady-state probability vector y (and hence $\boldsymbol{x}$ and $\boldsymbol{z}$) is obtained explicitly as follows.*

$$\begin{aligned}
\boldsymbol{y}^* &= \mu \boldsymbol{y}_0 (-T)^{-1}, \\
\boldsymbol{y}_i &= \textstyle\sum_{j=0}^{i-1} \boldsymbol{y}_j \boldsymbol{T}_0 \boldsymbol{\alpha} p_{i-j} (\mu I - T)^{-1},\ i \geq 1, \\
\boldsymbol{y}_0 &= \mu \boldsymbol{\pi} [2\mu I - \mu \boldsymbol{e} \boldsymbol{\alpha} - T + \mu^2 (-T)^{-1}]^{-1}.
\end{aligned}$$

Proof In this case the model is of $PH^{[X]}/M/1$-type with $D_0 = T$, $D_i = p_i \boldsymbol{T}^0 \boldsymbol{\alpha}$; $i \geq 1$, $D = \boldsymbol{T}^0 \boldsymbol{\alpha}$, $S = -\mu$, $\beta = 1$. On substitution, the steady-state equations in Eq. 18 get reduced to

$$\boldsymbol{y}^{*}T+\mu\boldsymbol{y}_0=\boldsymbol{0}, \tag{26}$$

$$\boldsymbol{y}^{*}\boldsymbol{T}_0\boldsymbol{\alpha}+\boldsymbol{y}_0(T-\mu I)+\mu\sum_{i=1}^{\infty}\boldsymbol{y}_i=\boldsymbol{0}, \tag{27}$$

$$\sum_{j=0}^{i-1}\boldsymbol{y}_j\boldsymbol{T}^0\boldsymbol{\alpha}p_{i-j}+\boldsymbol{y}_i(T-\mu I)=\boldsymbol{0},\ i\geq 1,$$

$$\boldsymbol{y}^{*}\boldsymbol{e}+\sum_{i=0}^{\infty}\boldsymbol{y}_i\boldsymbol{e}=1.$$

Recalling that

$$\boldsymbol{y}^{*}+\sum_{i=0}^{\infty}\boldsymbol{y}_i=\boldsymbol{\pi}=\lambda\boldsymbol{\alpha}(-T)^{-1},$$

we have from Eqs. 26 and 27, that

$$\mu\boldsymbol{y}_0\boldsymbol{e}\boldsymbol{\alpha}+\boldsymbol{y}_0(T-\mu I)+\mu\sum_{i=1}^{\infty}\boldsymbol{y}_i=\boldsymbol{0},$$

which implies

$$\boldsymbol{y}_0[\mu I-\mu\boldsymbol{e}\boldsymbol{\alpha}-T]=\mu\sum_{i=1}^{\infty}\boldsymbol{y}_i=\mu[\boldsymbol{\pi}-\boldsymbol{y}_0-\mu\boldsymbol{y}_0(-T)^{-1}],$$

and hence

$$\boldsymbol{y}_0=\mu\boldsymbol{\pi}[2\mu I-\mu\boldsymbol{e}\boldsymbol{\alpha}-T+\mu^2(-T)^{-1}]^{-1}.$$

Thus, the steady-state probability vector $\boldsymbol{y}$ is obtained explicitly as stated. □

Finally we note that in the case of $MAP/PH/1$-queue (i.e., the arrivals occur singly), we have $D_K=0,\ K\geq 2$. The steady-state equations for $\boldsymbol{y}^{*}$, $\boldsymbol{y}_0$ and $\boldsymbol{v}$ are same as those of $BMAP/PH/1$. This is as is to be expected since these involve D_0 and D. The only simplification is in the computation of $\boldsymbol{y}_i,\ i\geq 1$, and the needed equations are:

$$\boldsymbol{y}_i=\boldsymbol{y}_{i-1}(I\otimes D)(-S\oplus D_0)^{-1},\ i\geq 1.$$

To conclude this section we note that specific results for steady-state probabilities obtained above allow us to proceed with explicit expressions for the matrix G defined in Eq. 12. In particular, for some special cases, we have

1. $BMAP/PH/1$:

$$G=\Big[I-(\beta\otimes I)(-S\oplus Q)^{-1}(\boldsymbol{S}^0\otimes I)+(\beta\otimes I)(-S\oplus D_0)^{-1}(\boldsymbol{S}^0\otimes I)\Big]^{-1}$$
$$\Big[(\beta\otimes I)(-S\oplus D_0)^{-1}(\boldsymbol{S}^0\otimes I)\Big]$$

2. $M^{[X]}/PH/1$: since $A_0=\beta(\lambda I-S)^{-1}\boldsymbol{S}^0$, and $A=1$, then $G=1$.
3. $PH^{[X]}/M/1$: $A_0=\mu(\mu I-T)^{-1}$, and $A=\mu(\mu I-Q)^{-1}$, where interarrival times are PH distributed with representation $(\boldsymbol{\alpha},T)$ and $Q=T+\boldsymbol{T}^0\boldsymbol{\alpha}$. Hence it is easy to obtain

$$G=\mu\Big[\mu I-T+\mu\boldsymbol{T}^0\boldsymbol{\alpha}(\mu I-Q)^{-1}\Big]^{-1}.$$

5 Numerical Examples

In this section we will discuss a few illustrative numerical examples that bring out the qualitative aspects of the model under study. During the development of the Fortran code to generate these numerical examples, we used a number of accuracy checks to validate the code. These accuracy checks include the results of Theorem 4.2 as well as the special cases outlined in Theorems 4.5 through 4.7.

In our examples below, we consider three service time distributions. These are:

$ToS-1$: **Erlang (*ErS*):** This is Erlang of order 5 with rate 5μ in each stage.
$ToS-2$: **Exponential (*ExS*):** This is an exponential distribution with rate μ.
$ToS-3$: **Hyperexponential (*HeS*)**: Here we look at mixture of two exponentials with rates 7.30μ, 0.730μ and 0.073μ, respectively, with mixing probabilities 0.8, 0.15 and 0.05.

Also, in our examples we consider three types of probability distributions for batch sizes in the arrival process. They are:

Poisson Batch Size (*PbS*) Here we assume that the arriving batch is of size k with probability given by
$e^{-\theta}\Big(\dfrac{\theta^{k-1}}{(k-1)!}\Big),\ k \geq 1$. Note that the mean batch size is given by $\theta + 1$.

Geometric Batch Size (*GbS*) Here the arriving batch is of size k with probability given by $(1-p)p^{k-1},\ k \geq 1$. Note that the mean batch size is given by $\dfrac{1}{1-p}$.

Uniform Batch Size (*UbS*) Here it is assumed that the batch size is uniformly distributed on $\{1, 2, \cdots, N\}$. Due to the finiteness of N, it is clear that we assume that $D_i = 0,\ i > N$. Note that the mean batch size is given by $0.5(N+1)$.

Note that, in order to compare various scenarios properly, we will fix $1+\theta = \dfrac{1}{1-p} = 0.5(N+1)$ so that the batch means in all these cases are the same.

Example 1 In this example, we look at the effect of arrival rate, arrival processes, service processes and probability distributions of batch sizes on some selected performance measures. Towards this end, we look at five different $BMAPs$ with representation $\{D_k\}$ such that $D_k = Dp_k,\ k \geq 1$, where the batch size probability mass function, $\{p_k\}$, is taken to be one of the three displayed above, and D_0 and D are as given below.

TaP 1 : **Erlang (*ErA*):** Here we consider an Erlang distribution of order 5 with rate 5λ. That is,

$$D_0 = \begin{pmatrix} -5\lambda_g & 5\lambda_g & & & \\ 0 & -5\lambda_g & 5\lambda_g & & \\ 0 & 0 & -5\lambda_g & 5\lambda_g & \\ 0 & 0 & 0 & -5\lambda_g & 5\lambda_g \\ 0 & 0 & 0 & 0 & -5\lambda_g \end{pmatrix},\ D = \begin{pmatrix} 0 & 0\,0\,0\,0 \\ 0 & 0\,0\,0\,0 \\ 0 & 0\,0\,0\,0 \\ 0 & 0\,0\,0\,0 \\ 5\lambda_g & 0\,0\,0\,0 \end{pmatrix}$$

TaP 2 : **Exponential (*ExA*):** This corresponds to the classical Poisson process with rate λ_g. That is,

$$D_0 = \left(-\lambda_g\right),\ D = \left(\lambda_g\right)$$

$TaP\ 3:$ **Hyperexponential (HeA):** We look at a mixture of two exponentials with rates 1.9λ and 0.19λ, respectively, with probabilities 0.9 and 0.1. That is,

$$D_0 = \begin{pmatrix} -1.9\lambda_g & 0 \\ 0 & -0.19\lambda_g \end{pmatrix}, D = \begin{pmatrix} 1.71\lambda_g & 0.19\lambda_g \\ 0.171\lambda_g & 0.019\lambda_g \end{pmatrix}$$

$TaP\ 4:$ MAP **with negative correlation (MnC):**

$$D_0 = \begin{pmatrix} -2.25\lambda_g & 2.25\lambda_g & 0 & 0 & 0 \\ 0 & -2.25\lambda_g & 2.25\lambda_g & 0 & 0 \\ 0 & 0 & -2.25\lambda_g & 2.25\lambda_g & 0 \\ 0 & 0 & 0 & -2.25\lambda_g & 0 \\ 0 & 0 & 0 & 0 & -4.5\lambda_g \end{pmatrix}, D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0225\lambda_g & 0 & 0 & 0 & 2.2275\lambda_g \\ 4.455\lambda_g & 0 & 0 & 0 & 0.045\lambda_g \end{pmatrix}$$

$TaP\ 5:$ MAP **with positive correlation (MpC):**

$$D_0 = \begin{pmatrix} -2.25\lambda_g & 2.25\lambda_g & 0 & 0 & 0 \\ 0 & -2.25\lambda_g & 2.25\lambda_g & 0 & 0 \\ 0 & 0 & -2.25\lambda_g & 2.25\lambda_g & 0 \\ 0 & 0 & 0 & -2.25\lambda_g & 0 \\ 0 & 0 & 0 & 0 & -4.5\lambda_g \end{pmatrix}, D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2.2275\lambda_g & 0 & 0 & 0 & 0.0225\lambda_g \\ 0.045\lambda_g & 0 & 0 & 0 & 4.455\lambda_g \end{pmatrix}.$$

All of the above $BMAP$ processes will be normalized so as to have a specified (batch) arrival rate, λ_g. Observe that these $BMAPs$ are qualitatively different with different variance and correlation structure. It is worth mentioning that (a) the first three arrival processes, namely ErA, ExA, and HeA, correspond to renewal processes and hence the correlation is 0; (b) the arrival process labeled MnC has correlated arrivals with correlation between two successive inter-arrival times given by -0.57855 and the arrivals corresponding to the processes labelled MpC has a positive correlation with values 0.57855; (c) the ratio of the standard deviations of the inter-arrival times of these five arrival processes with respect to ErA are, respectively, 1, 2.2361, 5.0194, 3.1758, and 3.1758.

We fix $\mu = 1$ and consider two values: $\lambda_g = 1, 2$ for the (batch) arrival rate, and two values for the mean (arrival) batch size: 3 and 5. That is, $\lambda = 3\lambda_g$ and $\lambda = 5\lambda_g$.

In Table 1, we display the three measures, P_{idle}, the coefficient of variation of the sojourn time, and μ_{NS}. As mentioned earlier these three measures do not depend on the distribution as well as the mean of the batch size.

A look at this table yields the following observations.

- As is to be expected, the probability of the server being idle decreases with increasing arrival rate λ_g for all combinations of arrival and service time distributions.
- With respect to the first three arrival processes, whose inter-arrival times form a renewal process, the probability, P_{idle} appears to increase as the variation in the inter-arrival times increases for all cases. However, for other two arrival processes, whose successive inter-arrival times are, respectively, negatively and positively correlated, we see a different behavior, indicating again the significant role played by the correlation. Note that these two arrival processes have the same variance also.
- As the variation in service times (or in the inter-arrival times) increases, the measure P_{idle} appears to increase. This seems to be true for all cases. However, for the HeA processes the rate of increase (as a function of variability in service times) is not that significant.
- Further, it is worthy to note that this measure, P_{idle} is insensitive with regard to the probability distributions of batch size for different parameters under all scenarios.

Table 1 Selected system performance measures

Measure	λ_g	ToS	ErA	ExA	HeA	MnC	MpC
P(idle)	1	ErS	0.145	0.287	0.447	0.244	0.396
		ExS	0.243	0.333	0.468	0.300	0.428
		HeS	0.412	0.435	0.487	0.422	0.520
	2	ErS	0.013	0.085	0.307	0.065	0.101
		ExS	0.074	0.143	0.317	0.122	0.187
		HeS	0.221	0.251	0.320	0.235	0.332
Coefficient of variation of the sojourn time	1	ErS	0.455	0.452	0.430	0.450	0.464
		ExS	0.846	0.825	0.787	0.818	0.833
		HeS	1.791	1.791	1.763	1.785	1.707
	2	ErS	0.403	0.414	0.410	0.414	0.422
		ExS	0.755	0.758	0.746	0.754	0.768
		HeS	1.510	1.528	1.550	1.516	1.523
Mean number of service completions during a BP	1	ErS	2.693	2.488	3.293	2.803	1.627
		ExS	1.913	2.000	2.345	2.224	1.595
		HeS	1.171	1.298	1.409	1.412	1.243
	2	ErS	14.630	5.378	4.792	5.722	3.786
		ExS	3.431	3.000	3.101	3.277	2.357
		HeS	1.309	1.493	1.669	1.651	1.341

- The measure, $CV = \frac{\sigma_{W_s}}{\mu_{W_s}}$, giving the coefficient of variation of the sojourn time exhibits an interesting phenomenon. For all scenarios in which the service times are ErS or ExS, we notice this measure appears to be less than one while for HeS services, the CV is greater than one; further MpC arrivals seem to have the largest CV.
- With regard to the mean number of service completions, μ_{SC}, during a busy period, we notice that an increase in the arrival rate causes an increase in this measure. This is as expected; however, the rate of increase is significant in the case of ErS (which has the least variability compared to the other two service distributions considered) and furthermore we noticed that the HeA arrivals appear to have the narrowest range with regard to the amount of change (going from $\lambda_g = 1$ to $\lambda_g = 2$) across all three services. That is, if one looks at the ratio $\frac{\mu_{SC}^{\lambda_g=2}}{\mu_{SC}^{\lambda_g=1}}$ by varying the type of services, we get the values 4.315, 1.011, 0.2707, 0.4418, and 0.7553, respectively, for ErA, ExA, HeA, MnC, and MpC.

Now we display the measures, μ_{NQ}, μ_{NS}, and μ_{SR}, in Table 2. A quick look at this table reveals the following observations.

- As is to be expected, as λ_g increases, the mean queue length μ_{NQ}, increases for all arrival and service processes.
- This measure appears to be the same across all the three batch sizes distributions for a fixed value of mean (arrival) batch size, λ. It is observed that for a particular distribution, as value of the parameter increases (increasing the mean batch size), the mean queue length increases under all combinations of service and arrival processes.

Table 2 System performance measures displayed as $(\mu_{NQ}, \mu_{NS}, \mu_{SR})$

TaP	λ_g	1		2	
	λ	3	5	3	5
ErA	ErS	(1.12, 4.63, 9.45)	(1.86, 7.71, 15.75)	(3.44, 9.51, 88.90)	(5.73, 15.86, 148.16)
	ExS	(1.81, 5.77, 7.58)	(3.02, 9.62, 12.63)	(5.19, 11.67, 22.23)	(8.65, 19.45, 37.06)
	HeS	(16.51, 21.61, 5.98)	(27.51, 36.02, 9.96)	(44.52, 52.23, 10.09)	(74.20, 87.04, 16.81)
ExA	ErS	(1.28, 5.49, 10.46)	(2.14, 9.15, 17.44)	(3.29, 9.85, 35.27)	(5.49, 16.42, 58.78)
	ExS	(2.00, 6.50, 9.00)	(3.33, 10.83, 15.00)	(5.14, 12.14, 21.00)	(8.57, 20.24, 35.00)
	HeS	(16.40, 21.71, 6.89)	(27.34, 36.19, 11.49)	(43.50, 51.51, 11.96)	(72.50, 85.85, 19.93)
HeA	ErS	(1.52, 6.94, 17.86)	(2.54, 11.57, 29.76)	(3.42, 12.07, 41.46)	(5.69, 20.11, 69.09)
	ExS	(2.32, 7.96, 13.23)	(3.87, 13.27, 22.05)	(5.34, 14.13, 27.24)	(8.90, 23.54, 45.41)
	HeS	(16.73, 22.58, 8.24)	(27.89, 37.64, 13.74)	(42.16, 50.98, 14.72)	(70.28, 84.97, 24.53)
MnC	ErS	(1.41, 5.37, 11.12)	(2.35, 8.96, 18.53)	(3.34, 9.75, 36.71)	(5.56, 16.25, 61.18)
	ExS	(2.06, 6.34, 9.53)	(3.43, 10.57, 15.89)	(5.20, 12.03, 22.40)	(8.66, 20.05, 37.33)
	HeS	(16.51, 21.70, 7.33)	(27.52, 36.17, 12.21)	(44.08, 51.93, 12.96)	(73.47, 86.55, 21.60)
MpC	ErS	(1.09, 6.06, 8.09)	(1.81, 10.10, 13.48)	(3.09, 9.76, 25.26)	(5.15, 16.27, 42.11)
	ExS	(1.92, 7.17, 8.37)	(3.21, 11.95, 13.94)	(4.96, 12.34, 17.39)	(8.27, 20.57, 28.99)
	HeS	(17.91, 24.15, 7.77)	(29.84, 40.25, 12.95)	(43.68, 52.67, 12.05)	(72.80, 87.78, 20.08)

Moreover, the sensitivity to the value of parameter for batch sizes distribution is more apparent for HeS processes.

- For all scenarios, we see that a higher variation in the service times yield a higher value for μ_{NQ}. Furthermore, we notice a significant change in μ_{NQ} when going from ExS to HeS for all the five arrival processes. This appears to be the case for both values of λ.
- With regard to this measure, it is interesting to observe that the degree of sensitivity to the variation in the successive inter-arrival times is more for HeS processes as compared to the other service processes, especially when $\lambda_g = 2$.
- The similar behavior is observed for the measure, μ_{NS} giving the mean number under service with respect to the arrival rate, λ_g and mean batch size.
- For all scenarios, we see that an increase in the variation in service times induces an increase in this measure. Moreover, the rate of increment is most significant for HeS processes for all the arrival processes.
- Further, it is interesting to observe that a higher variation in the successive inter-arrival times yield a higher value for μ_{NS}. However, when $\lambda_g = 2$, the reverse behavior is observed for HeS processes.
- With regard to the mean number of customers served, μ_{SR}, during a busy period, we notice that this measure increases with increase in value of λ_g for all arrival and service processes, as expected. Moreover, for a particular distribution, increasing the mean batch size causes an increase in μ_{SR} and the rate of increment is remarkable for ErS processes when $\lambda_g = 2$.
- We noticed an interesting observation that this measure appears to decrease as the variation in service time increases under all scenarios.
- Furthermore, as the variation in successive inter-arrival times increases, the measure μ_{SR} seems to increase for $\lambda_g = 1$ for all service processes. However, we see a different behavior for $\lambda_g = 2$.

Example 2 Here, we consider the inter-arrival times to be Erlang of various orders ranging from 1 to 10. We fix $\mu = 1$, vary the (batch) arrival rate, $\lambda_g = 1, 2$ and consider the three service time distributions like in Example 1. Note that a fixed batch arrival rate means, that the variance of inter-arrival times decreases inversely proportional to the order of Erlang.

In Figs. 1 and 2 we display some selected measures. These figures enable one to observe the following.

- As is to be expected $P(idle)$ decreases as λ_g is increased. Also, this measure appears to increase with increasing variability in the services or increasing variability in the inter-arrival times. While an increase in the order of Erlang causes a gradual decrease in this measure, the rate of decrease appears to be insignificant for the HeS process. Note that HeS has a higher variability compared to the other two services considered here.
- With respect to the coefficient of variation of the sojourn time, we notice that the order of Erlang appears to play no significant role for all the service processes considered; further HeS process seems to have the largest CV exceeding one. This appears to be the case for both values of λ_g.
- The mean number of service completions during a busy period exhibits an interesting phenomenon. For all the scenarios, we notice that this measure increases with the arrival rate; however, the rate of increase is remarkable for ErS processes. Further, as the variation in service time increases, this measure appears to decrease and the difference in values increases with λ_g. This phenomenon could be due to the fact that a smaller variation in the service times lead to longer busy period resulting in more service completions. We also notice that an increase in the order of Erlang results in a decrease in this measure in the case of ExS and HeS services; however, for ErS services this measure appears to increase.
- With regard to the mean number in the queue, we notice that an increase in the arrival rate or in the mean batch size causes an increase in this measure. This is as expected; however, the degree of sensitivity to the value of λ_g as well as of mean batch size is higher for HeS services. We also notice that this measure appears to be insensitive to the order of Erlang for all the service processes considered.
- As is to be expected the mean number in service increases as λ_g is increased. Also, this measure appears to increase with increasing variability in service times or increasing value of mean batch size. While an increase in the order of Erlang causes a decrease in this measure, the rate of decrease seems to be less significant for HeS process compared to the other two service time distributions considered.
- With respect to the mean number of customers served during a busy period, we notice that for a particular distribution, increasing the mean batch size causes an increase in μ_{SR}. Moreover, this measure decreases with an increase in variability in the service times and the difference in values increases with λ_g. Further, it is worthy to note that the order of Erlang appears to play no significant role for all scenarios; however when $\lambda_g = 2$, we see a remarkable increment for ErS process.

It is well-known that the class of continuous phase type distributions is dense in the class of all distributions with support on $[0, \infty)$. Thus, one can approximate a given general distribution on the non-negative real line with appropriate phase type distributions. However, in our next example we will look at four different general distributions for services. The four service time distributions considered are as follows.

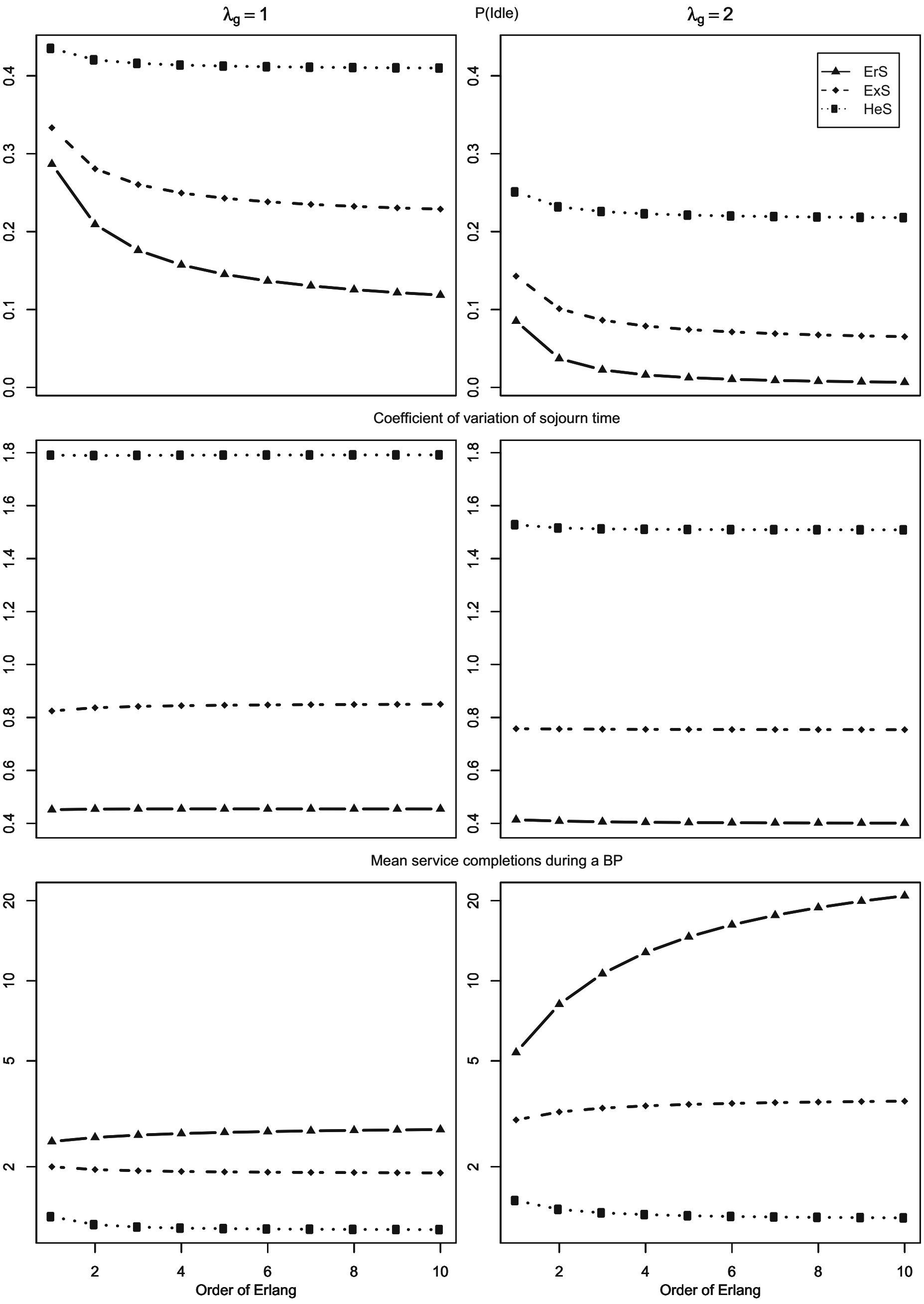

Fig. 1 Plot of selected measures under various scenarios

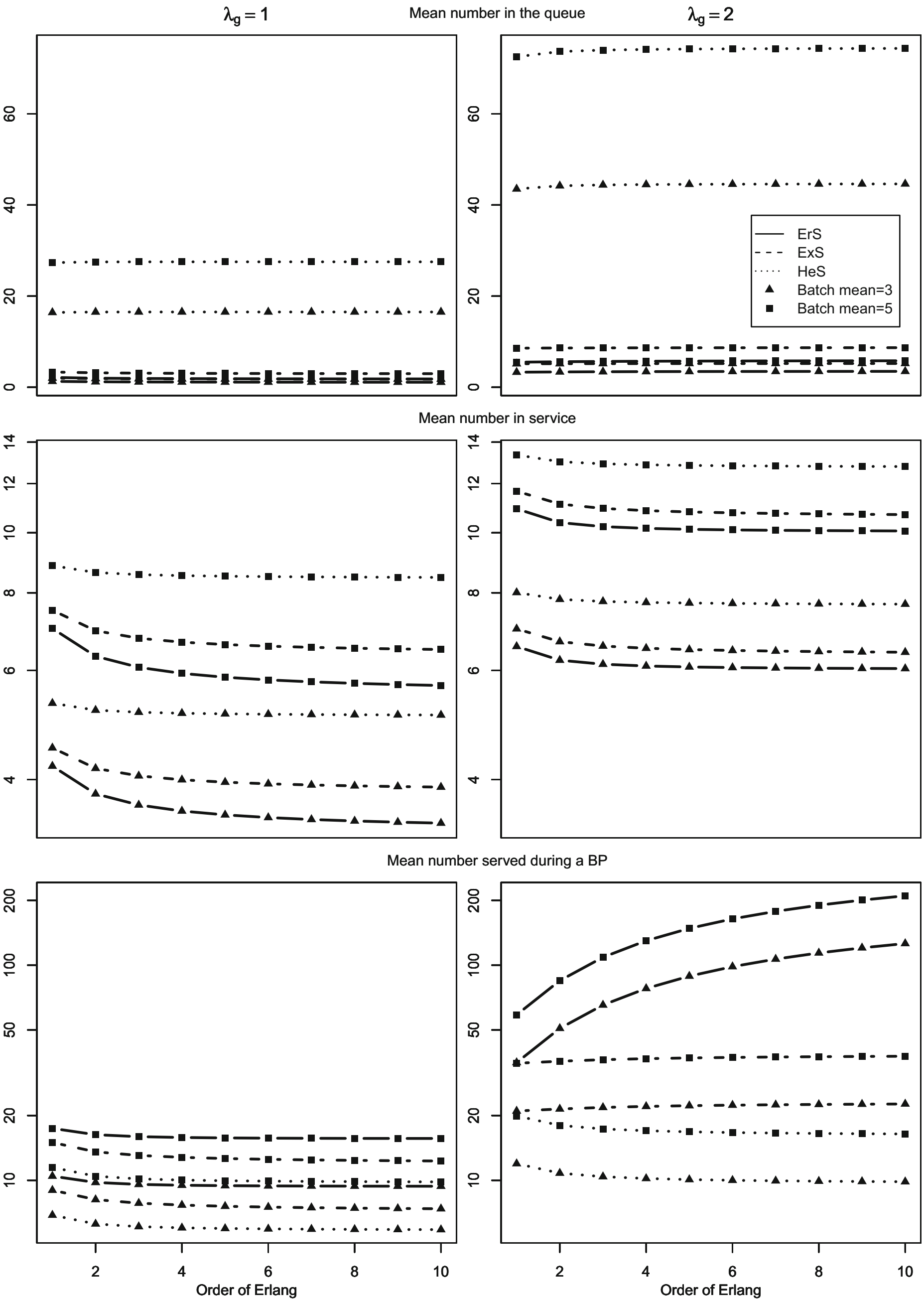

Fig. 2 Plot of selected measures under various scenarios using log scale for y-axis

1. **WBS:** Weibull with shape and scale parameters set at 0.5. The probability density function is given by:
$$f(t) = \frac{1}{\sqrt{2t}} e^{-\sqrt{2t}}, \; t \geq 0.$$
2. **UNS:** Uniform on the interval (0.5, 1.5).
3. **DPS:** Discrete probability mass function having masses at 0.5, 1, and 4,5, with probabilities, respectively, given by 0.7, 0.2, and 0.1.
4. **CNS:** Constant with a value of 1.

Note that the above distributions all have a mean of 1 but their standard deviations, respectively, are 2.2361, 0.2887, 1.1832, and 0.

Before we discuss the example, it is worth pointing out that the probabilities, $\{\gamma_n\}$ (see Eq. 10), discussed in Section 3.1, need to be calculated first. These probabilities depend on the type of arrival process as well as the service time distribution. Further, they need to be truncated and the cut-off or truncation point, n^*, is chosen by fixing $\epsilon = 10^{-4}$. for the tail probabilities.

For our next example, we will focus on the measure, μ_{SC}, the mean number of service completions during a busy period.

$$\mu_{SC} = \frac{1}{\boldsymbol{x}_0 \boldsymbol{e}} \boldsymbol{x}_0 (-D_0)^{-1} D \tilde{\boldsymbol{\mu}}, \tag{28}$$

where $\boldsymbol{x_0}$ is the steady-state probability vector that a service completion leaves the system idle.

Recall that that μ_{SC} is independent of the batch size distribution since the quantity on the right-hand side of Eq. 28 does not depend on the batch size distribution (see Eqs. 7 and 14).

Example 3 The purpose of this example to investigate the behavior of the performance measure, μ_{SC}, the mean number of service completions during a busy period, and (recalling) the cut-off point n^* such that

$$\sum_{i=n^*}^{\infty} \gamma_i \leqslant 10^{-4}, \tag{29}$$

in the case of general services. See Eq. 10 for the definition of γ_i. We use the same arrival processes that are used in Examples 1 and 2. We fix the parameters to have the same values. That is, we fix $\lambda_g = 1$ and $\mu = 1$.

In Figs. 3 and 4 below, we display the (log) mean number of service completions during a busy period, $ln(\mu_{SC})$ and (log) the cut-off points $ln(n^*)$ under different scenarios.

A quick note on the identifier in the Figs. 3 and 4. In Fig. 3, EW denotes the arrivals occur according to ERA process (defined in Example 1) and the service times are WBS (defined in Example 2). Similarly, ND corresponds to negatively correlated arrivals, namely, MnC and constant service times. That is, the first letter of the identifier corresponds to the arrival process used in the experiment, whereas the second defines the service time distribution. The following encoding is used for the first letter of the identifier: E – ErA process, X – ExA, H – HeA process, N – MnC process, P – MpC process. The second letter encoding is as follows: W – **WBS** distribution, U – **UNS** distribution, D – **DPS** distribution, C – **CNS** service time distribution. In Fig. 4, we display the two measures under consideration by looking at Erlangs of order 1 through 10, and for constant service times. This is to see the impact of low variability in the arrivals (and of course no variability in the services). Thus, the identifier $E1, \cdots E10$, correspond to Erlangs of order 1 through 10.

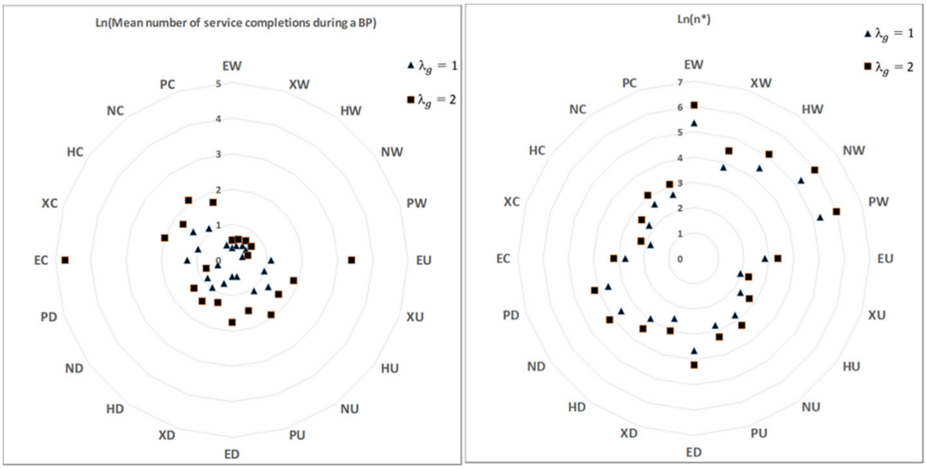

Fig. 3 Mean number of service completions during a busy period and cut-off points under various scenarios. The scenario is encoded as follows: first letter of the identifier E – ErA process, X – ExA process, H – HeA process, N – MnC process, P – MpC process; second letter encoding: W – **WBS** distribution, U – **UNS** distribution, D – **DPS** distribution, C – **CNS** service time distribution

A few key observations are as follows. (a) As is to be expected both μ_{SC} and n^* increase as λ_g is increased; (b) We notice for all scenarios that the mean number of service completions during a busy period decreases with an increase in the variability of the service times; whereas a reverse pattern is observed for the cut-off points. That is, as the variability in the services increases, one needs a large value of n, namely, n^*, such that Eq. 29 holds good. This appears to be the case for both values of λ_g; (c) It is worthy to note that, for ErA arrivals with either **UNS** or **CNS** services, a significant increase in the values of the measure μ_{SC} is seen as compared to the other scenarios; (d) With regard to the constant service times

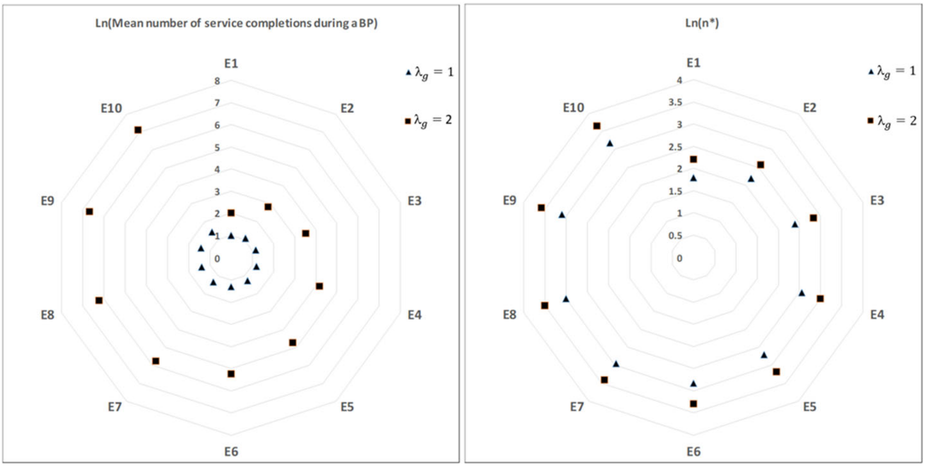

Fig. 4 Mean number of service completions during a busy period and cut-off or truncation points under constant service times, **CNS**, and Erlang arrival process of order $1, \ldots, 10$, encoded as $E1, \ldots, E10$, respectively

(i.e. **CNS**), we notice that the measure μ_{SC} and n^* increase as the order of Erlang increases. This can be explained intuitively as follows. As the order of Erlang increases the tail of the probability mass function, $\{\gamma_n\}$, of the number of Poisson (whose parameter depends on the order of Erlang) arrivals increases. This results in an increase in the probability of seeing at least one arrival during a service time (which is constant and equals 1). Thus, we see an increase in the average number of service completions during a busy period.

6 Concluding Remarks

In this paper we introduced bulk service rule with infinite upper bound in the context of a single server queue with batch Markovian arrivals and general service time distribution independent on the batch size being served. We analyzed this model using both embedded Markov renewal approach (in the case of general services) and using generator approach (in the case of phase type services). The steady-state probability vector is obtained almost explicitly for the general model and in many special cases we demonstrated how explicit solutions look like. Interesting numerical examples were presented. The model considered in this paper can be studied further in a number of ways. First, one can generalize this to include multiple servers. Secondly, we can model the system with server going on a vacation whenever the system becomes empty. In such a case, since the steady-state probabilities are likely to be obtained, it would be interesting to study the performance-energy tradeoff and optimize the vacation policy. Thirdly, one can model the server failures, repairs, and possibly backup server (but serving at a lower rate). These and other extensions are currently being investigated.

Acknowledgements AR thanks Dr. Natalia Nikitina for helpful discussions. The study of AR was partially supported by RFBR, projects 18-07-00147, 18-07-00156, 18-37-00094, 19-07-00303, 19-57-45022.

References

Abolnikov L, Dukhovny A (2003) Optimization in HIV screening problems. Int J Stoch Anal 16(4):361–374

Arumuganathan R, Jeyakumar S (2005) Steady state analysis of a bulk queue with multiple vacations, setup times with N-policy and closedown times. Appl Math Model 29(10):972–986

Ayyappan G, Karpagam S (2018) An $M^{[X]}/G(a,b)/1$ queueing system with breakdown and repair, stand-by server, multiple vacation and control policy on request for re-service. Mathematics 6(6):101. http://www.mdpi.com/2227-7390/6/6/101

Ayyappan G, Nirmala M (2018) An $M^{[X]}/G(a,b)/1$ queue with breakdown and delay time to two phase repair under multiple vacation. Applications and Applied Mathematics: An International Journal 13(2):639–663

Baba Y (1996) A bulk service $GI/M/1$ queue with service rates depending on service batch size. J Oper Res Soc Jpn 39(1):25–35

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. https://doi.org/10.1007/978-3-319-43904-4_3

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, ECMS 2017, vol 31, pp 649–655. http://www.scs-europe.net/dlib/2017/2017-0649.htm

Baetens J, Steyaert B, Claeys D, Bruneel H (2018) Delay analysis of a two-class batch-service queue with class-dependent variable server capacity. Math Methods Oper Res 88(1):37–57. https://doi.org/10.1007/s00186-017-0627-8

Baetens J, Steyaert B, Claeys D, Bruneel H (2019) Analysis of a batch-service queue with variable service capacity, correlated customer types and generally distributed class-dependent service times. Perform Eval 135:102012. https://linkinghub.elsevier.com/retrieve/pii/S0166531618302542

Bailey NTJ (1954) On queueing processes with bulk service. J R Stat Soc Series B Stat Methodol 16(1):80–87

Banerjee A, Gupta UC (2012) Reducing congestion in bulk-service finite-buffer queueing system using batch-size-dependent service. Perform Eval 69(1):53–70. https://linkinghub.elsevier.com/retrieve/pii/S0166531611001350

Banerjee A, Gupta UC, Goswami V (2014) Analysis of finite-buffer discrete-time batch-service queue with batch-size-dependent service. Comput Ind Eng 75:121–128. https://linkinghub.elsevier.com/retrieve/pii/S0360835214001892

Banerjee A, Gupta UC, Chakravarthy SR (2015) Analysis of a finite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service. Comput Oper Res 60:138–149. https://doi.org/10.1016/j.cor.2015.02.012. http://linkinghub.elsevier.com/retrieve/pii/S030505481500043X

Banik AD (2015) Single server queues with a batch Markovian arrival process and bulk renewal or non-renewal service. J Syst Sci Syst Eng 24(3):337–363. https://doi.org/10.1007/s11518-015-5268-y

Banik AD, Chaudhry ML, Gupta UC (2008) On the finite buffer queue with renewal input and batch Markovian service process: $GI/BMSP/1/N$. Methodol Comput Appl Probab 10(4):559–575

Banik AD, Gupta UC, Chaudhry ML (2009) Finite-buffer bulk service queue under Markovian service process: $GI/MSP^{(a,b)}/1/N$. Stoch Anal Appl 27(3):500–522

Bank B, Samanta SK (2020) Analytical and computational studies of the $BMAP/G^{(a,Y)}/1$ queue. Communications in Statistics - Theory and Methods, pp 1–29

Bar-Lev SK, Parlar M, Perry D, Stadje W, der Duyn Schouten FAV (2007) Applications of bulk queues to group testing models with incomplete identification. Eur J Oper Res 183(1):226–237

Bladt M, Nielsen BF (2017) Matrix-exponential distributions in applied probability, volume 81 of Probability Theory and Stochastic Modelling. Springer, Boston

Chakravarthy SR (1992) A finite capacity $GI/PH/1$ queue with group services. Nav Res Logist 39(3):345–357

Chakravarthy SR (1993) Analysis of a finite $MAP/G/1$ queue with group services. Queueing Syst 13(4):385–407. https://doi.org/10.1007/BF01149262

Chakravarthy SR (2001) The batch Markovian arrival process: A review and future work. In: Krishnamoorthy A et al.(eds) Advances in probability theory and stochastic processes. Notable Publications Inc., pp 21–39

Chakravarthy SR (2011) Markovian Arrival Processes, American Cancer Society. https://doi.org/10.1002/9780470400531.eorms0499

Chakravarthy SR (2015) Matrix-analytic queueing models, 2nd Edition, Birkhäuser, chap 8, pp 177–199

Chakravarthy SR, Dudin AN (2002a) A batch Markovian queue with a variable number of servers and group services. In: Latouche G, Taylor P (eds) Matrix-analytic methods - theory and applications. World Scientific Publishing Co., pp 63–88

Chakravarthy SR, Dudin AN (2002b) A multi-server retrial queue with $BMAP$ arrivals and group services. Queueing Syst 42(1):5–31. https://doi.org/10.1023/A:1019989127190

Chakravarthy SR, Maity A, Gupta UC (2017) An '(s, s)' inventory in a queueing system with batch service facility. Ann Oper Res 258(2):263–283

Charfi E, Gueguen C, Chaari L, Cousin B, Kamoun L (2017) Dynamic frame aggregation scheduler for multimedia applications in IEEE 802.11n networks. Trans Emerg Telecommun Technol 28(2): e2942

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

Chaudhry ML, Gupta UC (1999) Modelling and analysis of $M/G^{(a,b)}/1/N$ queue–a simple alternative approach. Queueing Syst 31(1-2):95–100

Chaudhry ML, Banik AD, Pacheco A, Ghosh S (2016) A simple analysis of system characteristics in the batch service queue with infinite-buffer and Markovian service process using the roots method: $GI/c-MSP^{(a,b)}/1/\infty$. RAIRO - Operations Research 50(3):519–551

Cheng Y, Yang Y, Du DZ (2019) A class of asymptotically optimal group screening strategies with limited item participation. Discret Appl Math 270:83–95. https://linkinghub.elsevier.com/retrieve/pii/S0166218X19302938

Ilya C, Natalia N, Evgeny I (2017) Task scheduling in desktop grids: Open problems. Open Eng 7(1):343. https://doi.org/10.1515/eng-2017-0038. https://www.degruyter.com/view/j/eng.2017.7.issue-1/eng-2017-0038/eng-2017-0038.xml

Claeys D, Laevens K, Walraevens J, Bruneel H (2010a) Complete characterisation of the customer delay in a queueing system with batch arrivals and batch service. Math Methods Oper Res 72(1):1–23. https://doi.org/10.1007/s00186-009-0297-2

Claeys D, Walraevens J, Laevens K, Bruneel H (2010b) A queueing model for general group screening policies and dynamic item arrivals. Eur J Oper Res 207(2):827–835. https://linkinghub.elsevier.com/retrieve/pii/S037722171000398X

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. Perform Eval 70(4):300–316. https://linkinghub.elsevier.com/retrieve/pii/S0166531612001344

D'Arienzo MP, Dudin AN, Dudin SA, Manzo R (2019) Analysis of a retrial queue with group service of impatient customers. J Ambient Intell Humaniz Comput 11(6):2591–2599. https://doi.org/10.1007/s12652-019-01318-x

Dudin A, Chakravarthy SR (2002a) Optimal hysteretic control for the $BMAP/G/1$ system with single and group service modes. Ann Oper Res 112(1):153–169. https://doi.org/10.1023/A:1020985106453

Dudin AN, Chakravarthy SR (2002b) A single server retrial queuing model with batch arrivals and group services. In: Artalejo JR, Krishnamoorthy A (eds) Advances in stochastic modelling. Notable Publications Inc., New Jersey

Dudin AN, Chakravarthy SR (2003) Multi-threshold control of the $BMAP/SM/1/K$ queue with group services. J Appl Math Stoch Anal 16(4):327–347. https://www.hindawi.com/archive/2003/276047/abs/

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

Gold H, Tran-Gia P (1993) Performance analysis of a batch service queue arising out of manufacturing system modelling. Queueing Syst 14(3-4):413–426

Grippa P, Schilcher U, Bettstetter C (2019) On access control in cabin-based transport systems. IEEE Trans Intell Transp Syst 20(6):2149–2156

Gupta GK, Banerjee A, Gupta UC (2020) On finite-buffer batch-size-dependent bulk service queue with queue-length dependent vacation. Qual Technol Quant Manag 17(5):501–527

Gupta G, Banerjee A (2018) On $M/G^{(a,b)}/1/N$ queue with batch size-and queue length-dependent service. In: International conference on mathematics and computing, Springer Nature, pp 249–262. https://doi.org/10.1007/978-981-13-2095-8_20

Gupta G, Banerjee A (2019) Steady state analysis of system size-based balking in $M/M^b/1$ queue. Int J Math Oper Res 14:319

He QM (2014) Fundamentals of matrix-analytic methods, Springer, New York

Hébuterne G, Rosenberg C (1999) Arrival and departure state distributions in the general bulk-service queue. Nav Res Logist (NRL) 46(1):107–118

Ivashko E, Rumyantsev A, Chernov I, Ponomarev V, Shabaev A (2018) Survey on deduplication techniques in flash-based storage. In: Proceedings of the 22nd conference of open innovations association FRUCT, vol 426, pp 25–33

Jensen A (1953) Markoff chains as an aid in the study of Markoff processes. Scand Actuar J 1953(sup1):87–91. https://doi.org/10.1080/03461238.1953.10419459

Jeyakumar S, Senthilnathan B (2017) Modelling and analysis of a bulk service queueing model with multiple working vacations and server breakdown. RAIRO - Operations Research 51(2):485–508

Kendall DG (1953) Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann of Math Stat 24(3):338–354

Lohse S, Pfuhl T, Berkó-Göttel B, Rissland J, Geißler T, Gärtner B, Becker SL, Schneitler S, Smola S (2020) Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. The Lancet Infectious Diseases https://doi.org/10.1016/S1473-3099(20)30362-5

Maity A, Gupta UC (2015) Analysis and optimal control of a queue with infinite buffer under batch-size dependent versatile bulk-service rule. OPSEARCH 52(3):472–489. https://doi.org/10.1007/s12597-015-0197-6

Mazalov VV, Nikitina NN, Ivashko EE (2014) Hierarchical two-level game model for tasks scheduling in a desktop grid. In: 2014 6th international congress on ultra modern telecommunications and control systems and workshops (ICUMT), pp 541–545

Neuts M, Li JM (1996) An algorithm for the P(n,t) matrices of a continuous BMAP. In: Chakravarthy SR, Alfa AS (eds) Matrix-analytic Methods in Stochastic Models. Marcel Dekker, pp 7–19

Neuts MF (1967) A general class of bulk queues with Poisson input. Ann of Math Stat 38(3):759–770

Neuts MF (1981) Matrix-Geometric solutions in stochastic models. Johns Hopkins University Press, Baltimore

Neuts MF, Chandramouli Y (1987) Statistical group testing with queueing involved. Queueing Syst 2(1):19–39

Niranjan SP, Chandrasekaran VM, Indhira K (2018) Performance characteristics of a batch service queueing system with functioning server failure and multiple vacations. J Phys Conf Ser 1000:012112. https://doi.org/10.1088/1742-6596/1000/1/012112

O'Mullane W, Li N, Nieto-Santisteban M, Szalay A, Thakar A, Gray J (2005) Batch is back: CasJobs, serving multi-TB data on the Web. In: IEEE international conference on Web services (ICWS'05), IEEE, pp 33–40

Panda G, Goswami V (2020) Effect of information on the strategic behavior of customers in a discrete-time bulk service queue. J Ind Manag Optim 16(3):1369–1388. https://doi.org/10.3934/jimo.2019007

Panda G, Banik AD, Guha D (2018) Stationary analysis and optimal control under multiple working vacation policy in a $GI/M^{(a,b)}/1$ queue. J Syst Sci Complexity 31(4):1003–1023. https://doi.org/10.1007/s11424-017-6172-y

Powell WB, Humblet P (1986) The bulk service queue with a general control strategy: Theoretical analysis and a new computational procedure. Oper Res 34(2):267–275

Pradhan S, Gupta UC (2017) Modeling and analysis of an infinite-buffer batch-arrival queue with batch-size-dependent service: $M^X/G_n^{(a,b)}/1$. Perform Eval 108:16–31. https://linkinghub.elsevier.com/retrieve/pii/S0166531616303078

Pradhan S, Gupta UC (2019) Analysis of an infinite-buffer batch-size-dependent service queue with Markovian arrival process. Ann Oper Res 277(2):161–196. https://doi.org/10.1007/s10479-017-2476-5

Pradhan S, Gupta UC, Samanta S (2016) Queue-length distribution of a batch service queue with random capacity and batch size dependent service: $M/G_r^Y/1$. OPSEARCH 53(2):329–343. https://doi.org/10.1007/s12597-015-0231-8

Sasikala S, Indhira K (2016) Bulk service queueing models – a survey. Int J Pure Appl Math 106(6):43–56. https://acadpubl.eu/jsi/2016-106-6-7-8/2016-106-6/6/6.pdf

Saxena A, Claeys D, Bruneel H, Zhang B, Walraevens J (2018) Modeling data backups as a batch-service queue with vacations and exhaustive policy. Comput Commun 128:46–59. https://linkinghub.elsevier.com/retrieve/pii/S0140366418302901

Sikdar K, Gupta UC (2005) Analytic and numerical aspects of batch service queues with single vacation. Comput Operat Res 32(4):943–966

Vadivu AS, Arumuganathan R (2015) Cost analysis of MAP/G(a, b)/1/N queue with multiple vacations and closedown times. Qual Technol Quant Manage 12(4):605–626. https://doi.org/10.1080/16843703.2015.11673438, publisher: Taylor & Francis

Xie M, Xia L, Xu J (2020) On $M/G^{[b]}/1/K$ queue with multiple state-dependent vacations: A real problem from media-based cache in hard disk drives. Perform Eval 139:102085. https://linkinghub.elsevier.com/retrieve/pii/S0166531620300055

Yu M, Alfa AS (2015) Algorithm for computing the queue length distribution at various time epochs in $DMAP/G^{(1,a,b)}/1/N$ queue with batch-size-dependent service time. Eur J Oper Res 244(1):227–239. https://linkinghub.elsevier.com/retrieve/pii/S0377221715000764

Yu M, Tang Y (2018) Analysis of the sojourn time distribution for $M/G^L/1$ queue with bulk-service of exactly size L. Methodol Comput Appl Probab 20(4):1503–1514. https://doi.org/10.1007/s11009-018-9635-2

Zee DJVD, Harten AV, Schuur P (2001) On-line scheduling of multi-server batch operations. IIE Trans 33(7):569–586

Zeng Y, Xia CH (2017) Optimal bulking threshold of batch service queues. J Appl Prob 54(2):409–423. https://www.cambridge.org/core/article/optimal-bulking-threshold-of-batch-service-queues/C1DD49670E8DFF9B2E55F82BF049BC29