J. Appl. Prob. **54,** 409–423 (2017)
*doi:*10.1017/jpr.2017.8
© *Applied Probability Trust* 2017

OPTIMAL BULKING THRESHOLD OF BATCH SERVICE QUEUES

YUN ZENG [*] [**] AND
CATHY HONGHUI XIA,[*] [***] *The Ohio State University*

Abstract

Batch service has a wide application in manufacturing, communication networks, and cloud computing. In batch service queues with limited resources, one critical issue is to properly schedule the service so as to ensure the quality of service. In this paper we consider an $\mathrm{M}/\mathrm{G}^{[a,b]}/1/N$ batch service queue with bulking threshold a, max service capacity b, and buffer capacity N, where N can be finite or infinite. Through renewal theory, busy period analysis and decomposition techniques, we demonstrate *explicitly* how the bulking threshold influences the system performance such as the mean waiting time and time-averaged number of loss customers in batch service queues. We then establish a necessary and sufficient condition on the optimal bulking threshold that minimizes the expected waiting time. Enabled by this condition, we propose a simple algorithm which guarantees to find the optimal threshold in polynomial time. The performance of the algorithm is also demonstrated by numerical examples.

Keywords: Batch service; optimal bulking threshold; queueing; Poisson arrival

2010 Mathematics Subject Classification: Primary 60K25
Secondary 60J10; 60K05

1. Introduction

Batch service queues, also referred to as bulk service queues, have been analyzed for decades. The model can be applied to many applications, ranging from manufacturing to communication networks to transportation to service delivery to cloud computing. Neuts [14] stated the basic idea of bulking: at the time of a bulk service departure, if the number of customers in the queue is less than the *bulking threshold* a, the server will wait until there are a customers and then serve them in a bulk; if the number of customers in the queue is larger than the bulking threshold a, the server will serve as many customers as its *capacity* b allows. The queue has *buffer capacity* N, where N can be finite or infinite. The above queue-length-based threshold policy has been shown to be optimal in the sense of minimizing the mean delay for Poisson arrivals and infinite buffer capacity [9]. Such batch service queues have been studied mainly from two aspects: steady-state analysis and optimal control.

In this paper we focus on the optimal control aspect. We aim at finding the optimal bulking threshold a so as to minimize the average waiting time of customers in a batch service queue. Compared with buffer space or service capacity, the bulking threshold is the most easily controllable parameter. However, to the best of the authors' knowledge, no previous research

Received 8 March 2016; revision received 7 October 2016.

[*] Postal address: Department of Integrated Systems Engineering, The Ohio State University, 1971 Neil Avenue, Columbus, OH 43210, USA.
[**] Email address: zeng.153@osu.edu
[***] Email address: xia.52@osu.edu

has explicitly demonstrated how the system performance would change as a function of the bulking threshold. Moreover, an efficient way to obtain the optimal threshold is essential in practice. For example, in cloud computing it is critical to schedule the bulk services effectively so as to meet the service level agreement [11].

In the literature, much attention has been directed to the steady-state analysis of batch service queues. An analytical solution to the steady-state distribution of the $\mathrm{M}/\mathrm{G}^{[a,b]}/1$ queue was derived by Neuts [14] in 1967. Numerous extensions have then been made, focusing on various phenomenons with either finite buffers [7], [10], [12] or more complicated arrival [4] or server vacations [1], [15] or batch-size-dependent service [2]. See [8] for a background review. Research on the optimal control of bulk queueing systems is more limited. Deb and Serfozo [9] proposed a method to find the optimal threshold for $\mathrm{M}/\mathrm{M}^{[a,b]}/1$ queues with the objective of minimizing a linear holding cost. They believed that the most general case was intractable, as the format of the holding cost function made the trend of the objective function unpredictable under different thresholds. Tadj and Tadj [19] developed the optimal threshold to minimize a expected total cost for $\mathrm{M}/\mathrm{D}^{r}/1$ queues with constant service times, where r denotes both the bulking threshold and the service capacity. The result was later generalized to the case with general service times and N-policy in Tadj and Ke [17] and to the case with batch arrivals in Tadj and Ke [18]. Efforts on the optimal control of other parameters such as the buffer size include [3] and [5]. For a detailed survey on the topic of optimal control of batch service queues, we refer the reader to [16].

In this paper we study the optimal control of the bulking threshold a for $\mathrm{M}/\mathrm{G}^{[a,b]}/1/\mathrm{N}$ batch service queues under the most general setting. We assume that the arrival process is Poisson, the service times are independent and identically distributed (i.i.d.) of a general distribution, and $0 < a \leq b \leq N \leq \infty$. Our work is strongly motivated by Gold and Tran-Gia [10], who modelled the departure point queue length of an $\mathrm{M}/\mathrm{G}^{[a,b]}/1/\mathrm{N}$ queue as an embedded Markov chain and proposed a method to calculate the average waiting time in the queue through the arbitrary point queue length. We adopt the same approach and extend their steady-state analysis to the aspect of optimal control. We model the departure point queue length as a Markov chain, and set the regeneration point to be the service starting point with no customer in the queue. Through renewal theory and busy period analysis, we demonstrate *explicitly* how the bulking threshold influences the system performance such as the mean waiting time and time-averaged number of loss customers (loss rate) in bulking service queues. The loss rate analysis further indicates that the objective of minimizing the loss rate is a trivial one. We therefore focus on the objective to minimize the expected waiting time (in queue) of customers.

The main contributions of this work can be summarized as follow:

- we show that the expected waiting time $\mathbb{E}[W(a)]$ as a function of the bulking threshold a is monotonically decreasing before reaching the optimal threshold a_{opt} and monotonically increasing after reaching a_{opt}. This guarantees the existence and the uniqueness of the optimal threshold;
- we prove that $a_{\mathrm{opt}} = \min\{\lceil \lambda \mathbb{E}[W(a_{\mathrm{opt}})] \rceil, b\}$ is a necessary and sufficient condition on the optimal threshold;
- we propose a simple algorithm that guarantees to find the optimal threshold a_{opt} in polynomial time.

Specifically, the above results are achieved via the following roadmap. First, we show that the busy period ending point queue length is a censored Markov chain of the departure point

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

queue length Markov chain. More importantly, the limiting distributions of both Markov chains are independent of the bulk threshold a. We then show that the expected total waiting time experienced by all customers in the busy period is also independent of the bulk threshold a. Only the expected total waiting time experienced by customers in the idle period depends on a. This allows us to derive an important decomposition of the average waiting time in the queue and calculate the key component that depends on threshold a. The optimality condition is derived by comparing the difference between the average waiting time under threshold a and $a + 1$. This enables us to develop a search algorithm to identify the optimal bulking threshold. We further demonstrate the efficiency of our search algorithm via numerical examples.

Since our study is under the most general setting with $0 < a \le b \le N \le \infty$, most prior optimal control related studies become special cases. The corresponding results in special cases such as $\mathrm{M/G}^{[a,b]}/1/\infty$, $\mathrm{M/G}^{[a,\infty]}/1/\infty$, and $\mathrm{M/G}^{[a,\infty]}/1/\mathrm{N}$ (or, equivalently, $\mathrm{M/G}^{[a,N]}/1/\mathrm{N}$) queues can then be easily derived. Our main result is valid as long as the renewal theory works.

The rest of the paper is organized as follows. In Section 2 we establish the model and some analysing techniques. In Section 3 we decompose the average waiting time of customers in the queue. In Section 4 we show the optimal threshold condition. In Section 5 we present an algorithm to achieve the optimal threshold. Section 6 contains simulation results and Section 7 concludes our research.

2. The model and preliminary analysis

Consider an $\mathrm{M/G}^{[a,b]}/1/\mathrm{N}$ batch service queueing system. Customers arrive according to a Poisson process with arrival rate λ. A single server serves a bulk of customers simultaneously in a round of service. The server has bulking threshold a and service capacity b. The queue has buffer size N. The service times for all batches are i.i.d. and independent of the bulking threshold a. No priority rule is applied and the service is nonpreemptive.

The bulking mechanism works as follows: every time when the server is free, it will check whether the number of customers in the queue has reached the threshold. If not, it will wait until the number in queue reaches threshold and then start service; otherwise, it will immediately start a new service round to serve as many customers in the queue as the capacity permits. The customers who arrive after the start of this service round must wait for the next service round.

Let W denote the waiting time for an arbitrary customer in the queue (before the service begins). Our objective is to find the optimal threshold a_{opt} that minimizes the average waiting time $\mathbb{E}[W]$ for given parameters b, N, λ, and given service time distribution

$$a_{\mathrm{opt}} = \arg\min_{a} \mathbb{E}[W(a)].$$

In order to make the problem nontrivial, we assume that $0 < a \le b \le N \le \infty$, where a, b, and N must be integers. Under the above setting, the systems including $\mathrm{M/G}^{[a,b]}/1/\infty$, $\mathrm{M/G}^{[a,\infty]}/1/\infty$, and $\mathrm{M/G}^{[a,\infty]}/1/\mathrm{N}$ can all be considered as special cases.

2.1. Embedded Markov chain

To analyze an $\mathrm{M/G}^{[a,b]}/1/\mathrm{N}$ system, we use the following notation:

- $Q(t)$ denotes the number of customers in the queue (not including those in service) at time t;
- A_n denotes the number of new arrivals (including those who are rejected due to full buffer) during the nth service round;

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

- X_n denotes the number of customers in the queue upon the nth service completion (departure point queue length);
- S denotes the random variable for the service time.

According to the bulking service mechanism, the departure point queue length process $\{X_n\}$ must evolve as

$$X_{n+1} = \begin{cases} \min\{A_{n+1}, N\} & \text{if } X_n \le b, \\ \min\{A_{n+1} + X_n - b, N\} & \text{otherwise.} \end{cases} \tag{2.1}$$

Let p_i denote the probability that there are i arrivals in one round of service. Since the arrival process is Poisson, we have $p_i = \mathbb{E}[\mathrm{e}^{-\lambda S}(\lambda S)^i/i!]$. Clearly, the distribution of A_n is given by $\{p_i, i = 0, 1, \ldots\}$. Therefore, $\{X_n\}_{n=1}^{\infty}$ is a discrete-time Markov chain embedded in the stochastic process $\{Q(t),\ t \ge 0\}$. Denote $\mathbb{P}_{ij} = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}$ as the transition probability and let $\boldsymbol{P} = [\mathbb{P}_{ij}]$, then

$$\mathbb{P}_{ij} = \begin{cases} p_j, & i = 0, 1, \ldots, b,\ j = 0, 1, \ldots, N-1, \\ \displaystyle\sum_{k=N}^{\infty} p_k, & i = 0, 1, \ldots, b,\ j = N, \\ p_{j-i+b}, & i = b+1, \ldots, N,\ j = i-b, \ldots, N-1, \\ \displaystyle\sum_{k=N-i+b}^{\infty} p_k, & i = b+1, \ldots, N,\ j = N, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

It is easily checked that the Markov chain $\{X_n\}_{n=1}^{\infty}$ is irreducible and aperiodic. When $N < \infty$, the limiting distribution $\boldsymbol{\pi}$ always exists. When $N = \infty$ and $b < \infty$, the stationary condition $\lambda\mathbb{E}[S] < b$ is needed to guarantee the existence of $\boldsymbol{\pi}$.

When π exists, it can be computed by solving $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$, $\sum_{i=0}^{N} \pi_i = 1$. Chaudhry and Templeton [8] thoroughly discussed the case for $N = \infty$ and derived the probability generating function of π. Even in this case, explicit results of π are only available when the service times are exponentially distributed [8, p. 223].

2.2. Regenerative cycle

Consider the evolution of the stochastic process $\{Q(t),\ t \ge 0\}$. The system regenerates every time when the service starts with no customer left in the queue. These time epochs can be set as regeneration points. The time interval between two adjacent regeneration points forms a regenerative cycle. One regenerative cycle contains one busy period and one potential idle period. Denote by B the duration of the busy period; by I the duration of the idle period. One busy period consists of one or more service rounds.

Let $N(a)$ be the total number of admitted arrivals in one regenerative cycle, and $T(a)$ the total amount of waiting time in the queue experienced by all customers in a regenerative cycle, both under threshold a. By renewal reward theory, the average waiting time in the queue is given by

$$\mathbb{E}[W(a)] = \frac{\mathbb{E}[T(a)]}{\mathbb{E}[N(a)]}. \tag{2.3}$$

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

2.3. Limiting distribution

Let H_m denote the total number of service rounds in the mth busy period. Let t_j be the time epoch of the jth service completion in the busy period, $j = 1, \dots, H_m$. The departure point queue length must satisfy $Q(t_1^-) > b, \dots, Q(t_{H_m-1}^-) > b,\ Q(t_{H_m}^-) \le b$.

Let Y_m denote the queue length immediately before the end of the mth busy period. Then the busy period ending point queue length process $\{Y_m\}_{m=1}^{\infty}$ is a *censored* Markov chain by censoring $\{X_n\}_{n=1}^{\infty}$ from the state set $\{b+1, b+2, \dots\}$. The limiting distribution of $\{Y_m\}_{m=1}^{\infty}$, denoted by $\pi^\star$, is then given by normalizing π, the limiting distribution of $\{X_n\}_{n=1}^{\infty}$, over the censored state space $\{0, 1, \dots, b\}$; see [13].

$$\pi^\star = (\pi_i^\star), \qquad \pi_i^\star = \lim_{m\to\infty} \mathbb{P}\{Y_m = i\} = \frac{\pi_i}{\sum_{j=0}^{b} \pi_j}, \qquad i = 0, 1, \dots, b.$$

Since $\boldsymbol{P}$ is given by (2.2), and all entries of $\boldsymbol{P}$ are independent of threshold a, the following lemma is then immediate.

Lemma 2.1. *Both $\boldsymbol{\pi}$ and $\pi^\star$ are independent of the bulking threshold a.*

3. Dependency analysis

In this section we study how the bulking threshold would impact the system performance. We first present important decomposition results for the expected total/average waiting time experienced by customers in a regenerative cycle, which help identify the key components that depend on the threshold in calculating the expected waiting time. We then demonstrate the dependency relationship between time-average number of lost customers and the bulking threshold.

3.1. Decomposition of the total waiting time

To calculate $\mathbb{E}[W(a)]$, we need $\mathbb{E}[T(a)]$, the expected total waiting time experienced by all customers in one regenerative cycle under bulking threshold a.

Observe that $\mathbb{E}[T(a)]$ can be decomposed into two parts: $\mathbb{E}[T_{\mathrm{B}}(a)]$ and $\mathbb{E}[T_{\mathrm{I}}(a)]$, the expected total waiting time experienced by all customers in the busy period, and that in the idle period. Since the busy period and the idle period do not overlap, we have

$$\mathbb{E}[T(a)] = \mathbb{E}[T_{\mathrm{B}}(a)] + \mathbb{E}[T_{\mathrm{I}}(a)]. \tag{3.1}$$

Considering first $\mathbb{E}[T_{\mathrm{B}}(a)]$, we make the following claim.

Lemma 3.1. *It holds that $\mathbb{E}[T_{\mathrm{B}}(a)]$ is independent of the bulking threshold a, i.e.*

$$\mathbb{E}[T_{\mathrm{B}}(a)] = \mathbb{E}[T_{\mathrm{B}}] \perp a. \tag{3.2}$$

Proof. Recall that H_m is the number of services in the mth regenerative cycle. From the perspective of departure point queue length Markov chain $\{X_n\}$, H_m indicates the first time inside the mth regenerative cycle that $X_n \le b$. Since the evolution of $\{X_n\}_{n=1}^{\infty}$ (given by (2.1)), does not depend on a, the evolution of $\{H_m\}_{m=1}^{\infty}$ does not depend on a either. Duration B is the summation of H_m i.i.d. service times and, hence, is also independent of a. Waiting time $T_{\mathrm{B}}(a)$ is determined by H_m, B, and the arrival process, all of which are independent of a. Thus the expected value of $T_{\mathrm{B}}(a)$ is independent of a. □

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

We next compute $\mathbb{E}[T_{\mathrm{I}}(a)]$. Recall that the queue length at the end of a busy period is associated with the limiting distribution of $\{Y_m\}_{m=1}^{\infty}$ which is given by $\pi^\star$. An idle period occurs only when Y_m is below the bulking threshold a.

Suppose that $Y_m = i < a$. Then the duration of the idle period is $\mathbb{E}[I] = (a-i)/\lambda$. Each of the i customers (who are already in the queue) will have to wait for the entire idle period before entering service. The $a-i$ customers who arrive during the idle period, except the last one who triggers the batch service, will experience average waiting time $\mathbb{E}[I]/2$. This is because, for a Poisson arrival process, given that there are exactly $a-i-1$ arrivals in period I, the event times are uniformly distributed over I. Therefore,

$$\mathbb{E}[T_{\mathrm{I}}(a)] = \sum_{i=0}^{a-1} \pi_i^\star \biggl[i\frac{a-i}{\lambda} + (a-i-1)\frac{(a-i)}{2\lambda} \biggr] = \frac{1}{\lambda}\sum_{j=0}^{a-1} j \sum_{i=0}^{j} \pi_i^\star. \tag{3.3}$$

3.2. Decomposition of the average waiting time

We are now ready to compute the average waiting time using (2.3). First, let us compute $\mathbb{E}[N(a)]$, the expected total number of admitted arrivals in one regenerative cycle.

At service completion time epoch t_j, $j = 1, \ldots, H_m - 1$, the server will take b customers from the queue. At time epoch t_j, $j = H_m$, if $Y_m < a$, then an idle period starts until a customers fill a bulk; if $Y_m \ge a$ then the server takes Y_m customers immediately. Therefore,

$$\mathbb{E}[N(a)] = b(\mathbb{E}[H]-1) + \sum_{i=a}^{b} \pi_i^\star i + \sum_{i=0}^{a-1} \pi_i^\star a. \tag{3.4}$$

Define the following function:

$$\sigma(a) := \frac{\mathbb{E}[N(a)]}{\lambda} = \sum_{i=0}^{a-1} \pi_i^\star \frac{a}{\lambda} + \sum_{i=a}^{b} \pi_i^\star \frac{i}{\lambda} + (\mathbb{E}[H]-1)\frac{b}{\lambda}. \tag{3.5}$$

Combining (2.3)–(3.5) yields the following theorem.

Theorem 3.1. *In an* $\mathrm{M/G}^{[a,b]}/1/\mathrm{N}$ *queue, the average waiting time in queue is*

$$\mathbb{E}[W(a)] = \frac{\mathbb{E}[T_{\mathrm{B}}]}{\lambda\sigma(a)} + \frac{1}{\lambda^2\sigma(a)} \sum_{j=0}^{a-1} j \sum_{i=0}^{j} \pi_i^\star. \tag{3.6}$$

In the following special cases, we can derive $\mathbb{E}[W(a)]$ in closed form.

Theorem 3.2. *In an* $\mathrm{M/G}^{[a,\infty]}/1/\mathrm{N}$ *queue, the average waiting time in queue is*

$$\begin{aligned}\mathbb{E}[W(a)] = {} & \frac{1}{\lambda^2\sigma(a)} \sum_{j=0}^{a-1} j \sum_{i=0}^{j} p_i + \frac{1}{\lambda\sigma(a)} \sum_{i=0}^{N} \mathbb{E}\biggl[\mathrm{e}^{-\lambda S}\frac{(\lambda S)^i}{i!} S \biggr] \frac{i}{2} \\ & + \frac{1}{\lambda\sigma(a)} \sum_{i=N+1}^{\infty} \mathbb{E}\biggl[\mathrm{e}^{-\lambda S}\frac{(\lambda S)^i}{i!} S \biggr] \biggl[1 - \frac{N+1}{2(i+1)} \biggr] N,\end{aligned} \tag{3.7}$$

where

$$\sigma(a) = \sum_{i=0}^{a-1} p_i \frac{a}{\lambda} + \sum_{i=a}^{N-1} p_i \frac{i}{\lambda} + \sum_{i=N}^{\infty} p_i \frac{N}{\lambda}, \qquad p_i = \mathbb{E}\biggl[\mathrm{e}^{-\lambda S}\frac{(\lambda S)^i}{i!} \biggr]. \tag{3.8}$$

Proof. See Appendix A. □

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

Theorem 3.3. *In an* $\mathrm{M}/\mathrm{G}^{[a,\infty]}/1/\infty$ *queue, the average waiting time in queue is*

$$\mathbb{E}[W(a)] = \frac{\mathbb{E}[S^2]}{2\sigma(a)} + \frac{1}{\lambda^2\sigma(a)} \sum_{j=0}^{a-1} j \sum_{i=0}^{j} p_i,$$

where

$$\sigma(a) = \sum_{i=0}^{a-1} p_i \frac{a}{\lambda} + \sum_{i=a}^{\infty} p_i \frac{i}{\lambda}. \tag{3.9}$$

Proof. See Appendix B. □

In the most general case of an $\mathrm{M}/\mathrm{G}^{[a,b]}/1/\mathrm{N}$ queue with finite buffer and service capacity, Gold and Tran-Gia [10] proposed a procedure to compute $\mathbb{E}[W(a)]$. However, our Theorem 3.1 is different. We decompose $\mathbb{E}[W(a)]$ and give the part that depends on threshold a. Equation (3.6) has not been given before and it helps us to determine the optimal bulking threshold in Section 4.

In the special case where $b = N < \infty$, the Laplace–Stieltjes transform of $\mathbb{E}[W(a)]$ under a Markovian arrival process was given by [6]. However, the results are given in forms that are computationally less convenient than the above closed-form results.

3.3. Time-averaged number of losses

In this subsection we discuss how the bulking threshold a would impact the loss rate of a batch service queue when the buffer capacity N is finite. We make the following claim.

Lemma 3.2. *Consider an* $\mathrm{M}/\mathrm{G}^{[a,b]}/1/\mathrm{N}$ *queue with* $0 < a \le b \le N < \infty$. *Denote by* $l(a)$ *the time-averaged number of losses under bulking threshold* a. *Then* $l(a)$ *is a strictly decreasing function in* a.

Proof. Let $L(a)$ denote the total number of customers lost in one regenerative cycle under threshold a. Based on renewal theory, the time-average number of lost customers is given by $l(a) = \mathbb{E}[L(a)]/\mathbb{E}[B(a) + I(a)]$, where $B(a)$ (respectively, $I(a)$) represents the length of a busy period (respectively, an idle period) in one regenerative cycle under threshold a.

As we discussed in Section 3.1, the evolution of the queue length in each busy period is independent of the threshold a. Since customers can only be lost in the busy period, $\mathbb{E}[L(a)]$, the expected total number of customers lost in each regenerative cycle is independent of a.

Recall that H, the number of service rounds in one busy period, is independent of a. Since $B(a)$, the length of a busy period is simply the sum of H i.i.d. service times, $\mathbb{E}[B(a)]$ must also be independent of a. Observe that

$$\mathbb{E}[I(a)] = \sum_{i=0}^{a-1} \pi_i^\star \frac{a-i}{\lambda},$$

which is strictly increasing in a. Thus, $l(a)$, the time-average number of lost customers is strictly decreasing in a. □

The following theorem is then immediate.

Theorem 3.4. *When the objective is to minimize the time-averaged number of lost customers for an* $\mathrm{M}/\mathrm{G}^{[a,b]}/1/\mathrm{N}$ *queue with* $N < \infty$, *the optimal policy is a trivial one, namely* $a_{\mathrm{opt}} = b$.

In the remainder of the paper, we therefore focus on the more interesting objective to minimize the expected waiting time in bulking service queues.

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

4. Optimal threshold

We next investigate how to find the optimal bulking threshold a in M/G$^{[a,b]}$/1/N queue. Our objective is to minimize the expected waiting time given by (3.6).

Based on previous discussions, we know that $\mathbb{E}[T_{\mathrm{B}}(a)]$ and $\pi_i^\star$ are independent of threshold a. As a increases, both $\sigma(a)$ and the double summation term $\sum_{j=0}^{a-1} j \sum_{i=0}^{j} \pi_i^\star$ inside (3.6) increase. Note, from (3.5), that $\sigma(a)$ is proportional to the expected total number of customers arrived in a regenerative cycle. The double summation term is proportional to the expected total waiting time in the idle period. Thus, the idea behind the optimal threshold is to include as many customers in a regenerative cycle as possible but not to make them wait too long during the idle period.

In order to find the optimal threshold, we first analyze the difference between $\mathbb{E}[W(a)]$ and $\mathbb{E}[W(a+1)]$. First we make the following claim.

Lemma 4.1. *The difference between* $\mathbb{E}[W(a)]$ *and* $\mathbb{E}[W(a+1)]$ *is given by*

$$\Delta\mathbb{E}[W(a)] = \mathbb{E}[W(a+1)] - \mathbb{E}[W(a)] = \frac{g(a)}{\lambda^2 \sigma(a+1)} f(a), \tag{4.1}$$

in which $f(a)$ *and* $g(a)$ *are defined as*

$$g(a) = \sum_{i=0}^{a} \pi_i^\star, \qquad f(a) = a - \lambda\mathbb{E}[W(a)].$$

Proof. Move $\sigma(a)$ to the left-hand side of (3.6). We have

$$\sigma(a+1)\mathbb{E}[W(a+1)] - \sigma(a)\mathbb{E}[W(a)] = \frac{ag(a)}{\lambda^2}. \tag{4.2}$$

Note that $\sigma(a)$ can be rewritten as

$$\sigma(a) = \sum_{i=0}^{a-1} \pi_i^\star \frac{a-i}{\lambda} + \sum_{i=0}^{b} \pi_i^\star \frac{i}{\lambda} + (\mathbb{E}[H]-1)\frac{b}{\lambda}. \tag{4.3}$$

It has the following property:

$$\sigma(a+1) - \sigma(a) = \sum_{i=0}^{a} \pi_i^\star \frac{a+1-i}{\lambda} - \sum_{i=0}^{a-1} \pi_i^\star \frac{a-i}{\lambda} = \frac{g(a)}{\lambda}. \tag{4.4}$$

Then we can rewrite (4.2) as

$$\sigma(a+1)(\mathbb{E}[W(a+1)] - \mathbb{E}[W(a)]) + \frac{g(a)}{\lambda}\mathbb{E}[W(a)] = \frac{ag(a)}{\lambda^2}.$$

Thus,

$$\Delta\mathbb{E}[W(a)] = \frac{g(a)}{\lambda^2\sigma(a+1)}(a - \lambda\mathbb{E}[W(a)]) = \frac{g(a)}{\lambda^2\sigma(a+1)} f(a). \qquad \square$$

The next lemma reveals the increasing property of the function $f(a)$.

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

Lemma 4.2. *The function $f(a)$ is strictly increasing as threshold a increases from 1 to b.*

Proof. Since $f(a+1) - f(a) = 1 - \lambda\Delta\mathbb{E}[W(a)]$, it suffices to show that $\lambda\Delta\mathbb{E}[W(a)] < 1$ for all $1 \leq a \leq b$.

From (4.1) and (4.4), we have

$$\lambda\Delta\mathbb{E}[W(a)] = g(a)\frac{a/\lambda - \mathbb{E}[W(a)]}{\sigma(a) + g(a)/\lambda} < g(a)\frac{a/\lambda}{\sigma(a)}.$$

Based on (4.3), we have

$$\sigma(a) \geq \sum_{i=0}^{a-1} \pi_i^\star \frac{a}{\lambda} + \sum_{i=a}^{b} \pi_i^\star \frac{i}{\lambda} > \sum_{i=0}^{b} \pi_i^\star \frac{a}{\lambda} = \frac{a}{\lambda}. \tag{4.5}$$

It then follows that

$$\lambda\Delta\mathbb{E}[W(a)] < g(a) = \sum_{i=0}^{a} \pi_i^\star \leq 1.$$

Hence, $f(a+1) - f(a) > 0$ and $f(a)$ is a strictly increasing function in a. □

Observe that, from (4.1), since $g(a)$ and $\sigma(a+1)$ are both positive, the sign of $\Delta\mathbb{E}[W(a)]$ is the same as that of $f(a)$. We have the following three situations:

- if $f(a) > 0$ then $\mathbb{E}[W(a+1)] > \mathbb{E}[W(a)]$ and we prefer a to $a+1$;
- if $f(a) < 0$ then $\mathbb{E}[W(a+1)] < \mathbb{E}[W(a)]$ and we prefer $a+1$ to a;
- if $f(a) = 0$ then $\mathbb{E}[W(a+1)] = \mathbb{E}[W(a)]$.

Since $f(0) < 0$ and $f(a)$ is strictly increasing, the first point at which $f(a)$ turns nonnegative yields the optimal threshold. Before this point, we have $f(a) < 0$ and $\mathbb{E}[W(a)]$ decreases; on this point, we have $f(a) \geq 0$ and $\mathbb{E}[W(a)]$ is no smaller than $\mathbb{E}[W(a+1)]$; after this point, we have $f(a) > 0$ and $\mathbb{E}[W(a)]$ increases. In the special case when $f(a) = 0$, we have two adjacent points $a, a+1$ that minimize the average waiting time. We will take the smaller one, i.e. a as the optimal threshold. The next lemma summarizes these properties.

Lemma 4.3. *The optimal bulking threshold a_{opt} is the first point at which $f(a)$ turns nonnegative. The average waiting time $\mathbb{E}[W(a)]$, as a function of a, is strictly decreasing on $[0, a_{\text{opt}}]$, monotonically increasing on $[a_{\text{opt}}, a_{\text{opt}}+1]$ and strictly increasing on $[a_{\text{opt}}+1, \infty]$.*

We are now ready to present the main theorem on the optimal threshold.

Theorem 4.1. *In system $\mathrm{M/G}^{[a,b]}/1/\mathrm{N}$, a necessary and sufficient condition on the optimal threshold is*

$$a_{\text{opt}} = \min\{\lceil\lambda\mathbb{E}[W(a_{\text{opt}})]\rceil, b\}, \tag{4.6}$$

where $\lceil x \rceil$ denotes the ceiling function and returns the smallest integer that is larger than or equal to x.

Proof. See Appendix C. □

Theorem 4.1 establishes the property that the optimal bulking threshold should satisfy the condition $a = \lceil\lambda\mathbb{E}[W(a)]\rceil$ and it should not be larger than the server capacity b. The logic behind this result can be briefly interpreted as follows. Suppose we are tuning the bulking

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

threshold a at customer arriving epochs. Under PASTA (Poisson arrivals see time averages), a new (tagged) arrival is expected to see $\lambda\mathbb{E}[W(a)]$ customers waiting in the queue. Upon his/her arrival, the expected queue length is then $\lambda\mathbb{E}[W(a)] + 1$. Suppose that $\lambda\mathbb{E}[W(a)]$ is not an integer. We have $\lambda\mathbb{E}[W(a)] + 1 < \lceil\lambda\mathbb{E}[W(a)]\rceil + 1$. If the bulking threshold is set to be $\lceil\lambda\mathbb{E}[W(a)]\rceil + 1$, $\lceil\lambda\mathbb{E}[W(a)]\rceil + 2, \ldots$, then the threshold is not crossed and the tagged customer on average has to wait for more arrivals before being served. As we continue to increase the bulking threshold, by the discussion of Lemma 4.3, the expected waiting time also increases. On the other hand, if the threshold is set to be $\lceil\lambda\mathbb{E}[W(a)]\rceil - 1$, $\lceil\lambda\mathbb{E}[W(a)]\rceil - 2, \ldots$, then the tagged customer is likely to see a busy server and has to wait for service completions. Only when the threshold is set so that the tagged customer on average gets served immediately, which is exactly $\lceil\lambda\mathbb{E}[W(a)]\rceil$, the expected waiting time is minimized. Hence, the solution becomes optimal. A similar discussion applies to the case when $\lambda\mathbb{E}[W(a)]$ is an integer. In this case, $\lambda\mathbb{E}[W(a)]$ and $\lambda\mathbb{E}[W(a)] + 1$ lead to the same average waiting time and $\lceil\lambda\mathbb{E}[W(a)]\rceil$ is again optimal.

5. Algorithm for optimal threshold

We next present a simple algorithm to search for the optimal bulking threshold a that minimizes the expected waiting time in $\mathrm{M/G}^{[a,b]}/1/\mathrm{N}$ queues. We show that the algorithm guarantees to find the optimal solution in polynomial time.

Define $a_1 = \min\{\lceil\lambda\mathbb{E}[W(1)]\rceil, b\}$. We propose the following algorithm to search for the optimal threshold.

Algorithm 5.1. Thus,

```
(I) $a \leftarrow a_1$;
(II) Check if $a = \min\{\lceil\lambda\mathbb{E}[W(a)]\rceil, b\}$;
(III) If true, return $a_{\text{opt}} = a$;
(IV) If false, $a \leftarrow \lceil\lambda\mathbb{E}[W(a)]\rceil$, and go back to step (II).
```

In each iteration, $\mathbb{E}[W(a)]$ can be computed theoretically (by Theorem 3.2 and 3.3 for special cases; by Gold and Tran-Gia (1993) for general cases) or estimated by simulation.

The performance of the algorithm is established in the next theorem.

Theorem 5.1. *Algorithm 5.1 converges to the optimal bulking threshold a_{opt} within a_1 steps and takes polynomial time.*

Proof. Based on Theorem 4.1, (4.6) is the necessary and sufficient condition for optimality. Therefore, once the check result in step (II) is true, we have found the optimal solution.

To discuss the convergence of the algorithm, we assume that the check result of step (II) is always false so that the algorithm always takes step (IV) and keeps running.

For the initial threshold a_1, we have

$$a_1 = \min\{\lceil\lambda\mathbb{E}[W(1)]\rceil, b\} \geq \min\{\lceil\lambda\mathbb{E}[W(a_{\text{opt}})]\rceil, b\} = a_{\text{opt}}.$$

In order to enter step (IV), a_1 must be strictly larger than the optimal threshold. Based on Lemma 4.3 , we must have $f(a_1) = a_1 - \lambda\mathbb{E}[W(a_1)] > 0$. Since a_1 is not optimal, we also have $a_1 \neq \lceil\lambda\mathbb{E}[W(a_1)]\rceil$. Thus, $a_1 > \lceil\lambda\mathbb{E}[W(a_1)]\rceil = a_2$,in which a_2 is the threshold we obtain after the first iteration.

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

For threshold a_2, we also have

$$a_2 = \min\{\lceil \lambda \mathbb{E}[W(a_1)] \rceil, b\} \geq \min\{\lceil \lambda \mathbb{E}[W(a_{\text{opt}})] \rceil, b\} = a_{\text{opt}}.$$

Repeat the above argument, if the ith iteration enters step (IV), we will obtain a value a_{i+1} such that

$$a_{\text{opt}} \leq a_{i+1} < a_i \leq a_1.$$

Since a_1 is finite, the algorithm has to enter step (III) and return the optimal threshold a_{opt} in no more than a_1 steps.

We next show that the running time of the algorithm is polynomial.

Observe from (3.3), (3.5), and (3.6) that, as $a \to \infty$, we have

$$\sigma \sim \frac{a}{\lambda}, \qquad \mathbb{E}[W(a)] \sim \frac{a}{2\lambda}.$$

Thus, for large a_i, we have

$$\frac{a_{i+1}}{a_i} \approx \frac{\lambda \mathbb{E}[W(a_i)]}{a_i} \to \frac{1}{2}.$$

Thus, the algorithm takes at most $O(\log_2 a_1)$ iterations to find the optimal threshold, which is polynomial. □

6. Numerical examples

In this section we present two numerical examples to demonstrate the performance of Algorithm 5.1.

Example 6.1. Consider an $\mathrm{M/G}^{[a,\infty]}/1/\mathrm{N}$ queue with parameter settings $\lambda = 8$, $N = 500$, and service time following exponential distribution with mean 50. In this case, under a given threshold a, the corresponding expected waiting time $\mathbb{E}[W(a)]$ can be calculated in closed form using (3.7) in Theorem 3.2. In Table 1 we show the iterations of running Algorithm 5.1 in searching for the optimal bulking threshold that minimizes the expected waiting time. Each row presents the values of $\mathbb{E}[W(a)]$ and $\min\{\lceil \lambda \mathbb{E}[W(a)] \rceil, b\}$ under given threshold a. The optimality condition is checked according to Theorem 4.1, i.e. if $a = \min\{\lceil \lambda \mathbb{E}[W(a)] \rceil, b\}$ then threshold a is optimal. If the condition is not met, the value of $\min\{\lceil \lambda \mathbb{E}[W(a)] \rceil, b\}$ is then used as the threshold for the next iteration. Observe that at iteration two, the optimality condition is met, thus the algorithm terminates and the optimal bulking threshold is $a_{\text{opt}} = 279$.

Example 6.2. Consider an $\mathrm{M/G}^{[a,b]}/1/\mathrm{N}$ queue with parameter settings $\lambda = 3$, $N = 40$, $b = 35$, and service time distribution Γ(scale=20, shape=0.05). In this case, we do not have a closed-form formula to compute $\mathbb{E}[W(a)]$. In order to find the optimal threshold using Algorithm 5.1, in each iteration under a given threshold a, the expected waiting time $\mathbb{E}[W(a)]$ is estimated by discrete-event simulation. In order to estimate the steady-state performance of this batch service queue, for each simulation run, we set the replication length to be 100,000 time units out of which the first 10,000 time units were used as the warm-up period. We choose the number of replications so that the half width of the simulated average waiting time is smaller than 0.01. In Table 2 we show the iterations of running Algorithm 5.1, where column three presents the confidence interval for the estimated $\mathbb{E}[W(a)]$ under a given threshold a. Observe that the algorithm terminates at iteration three when the optimality condition is met. The optimal bulking threshold is therefore $a_{\text{opt}} = 20$.

Overall, we see that Algorithm 5.1 is highly efficient in finding the optimal threshold.

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

TABLE 1: Iterations of Algorithm 5.1.

Iteration	a	$\mathbb{E}[W(a)]$	$\min\{\lceil\lambda\mathbb{E}[W(a)]\rceil, b\}$	Optimal
0	1	37.4403	300	–
1	300	34.8525	279	No
2	279	34.8306	279	Yes

TABLE 2: Iterations of Algorithm 5.1.

Iteration	a	$\mathbb{E}[W(a)]$	$\min\{\lceil\lambda\mathbb{E}[W(a)]\rceil, b\}$	Optimal
0	1	9.726 ± 0.01	30	–
1	30	7.047 ± 0.01	22	No
2	22	6.646 ± 0.01	20	No
3	20	6.627 ± 0.01	20	Yes

7. Conclusion

In this paper we have studied the optimal control of batch service queues. In particular, we considered the $\mathrm{M/G}^{[a,b]}/1/\mathrm{N}$ batch service queueing system and focused on finding the optimal setting of the bulking threshold so as to minimize the average waiting time. Using busy period analysis and the renewal argument, we presented a useful decomposition of the average waiting time and derived the part that depends on the bulking threshold a. The optimal threshold that minimizes the average waiting time was proven to possess the following necessary and sufficient condition:

$$a_{\mathrm{opt}} = \min\{\lceil\lambda\mathbb{E}[W(a_{\mathrm{opt}})]\rceil, b\}.$$

Using this condition, we proposed an algorithm that can obtain the optimal threshold in polynomial time. Numerical results were presented to demonstrate the algorithm's efficiency.

Although these results were established with the restriction to Poisson arrivals and i.i.d. service times, they can be applied as good approximations to capture the high-level queueing dynamics of more general systems. Future research could focus on the optimal threshold analysis of batch service queueing systems with other types of arrival processes or with more servers. Another area of interest may be to study the bulking and resource allocation when there are multiple classes of jobs. These variations may pose additional difficulty in the analysis and are left for future investigations.

Appendix A.

Proof of Theorem 3.2. In an $\mathrm{M/G}^{[a,\infty]}/1/\mathrm{N}$ queue, the initiation of each batch service would result in an empty buffer thus a regeneration point. Hence,

$$\mathbb{E}[N(a)] = \sum_{i=0}^{a-1} p_i a + \sum_{i=a}^{N-1} p_i i + \sum_{i=N}^{\infty} p_i N,$$

where p_i is the probability of i arrivals in one service round. Apply $\sigma(a) = \mathbb{E}[N(a)]/\lambda$. We obtain (3.8).

Next, we compute the total waiting time of all customers in the queue conditioning on the service time $S = s$ and A_S the number of arrivals during the service time. Given that there

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

are i arrivals in s amount of time, the expected arriving epochs of these i customers evenly divide the service time s into $i+1$ intervals with equal mean length $\Delta = s/(i+1)$. Since at most N customers can join the queue, we only count the waiting time of the first $m = \min(i, N)$ customers. Thus,

$$\mathbb{E}[T_{\mathrm{B}} \mid A_{\mathrm{S}} = i, S = s] = (s - \Delta) + (s - 2\Delta) + \cdots + (s - m\Delta) = \biggl[1 - \frac{m+1}{2(i+1)}\biggr] ms.$$

Hence,

$$\mathbb{E}[T_{\mathrm{B}} \mid S = s] = \sum_{i=0}^{\infty} \mathrm{e}^{-\lambda s} \frac{(\lambda s)^i}{i!} \biggl[1 - \frac{(i \wedge N) + 1}{2(i+1)}\biggr] (i \wedge N) s. \tag{A.1}$$

For generally distributed S, we have

$$\begin{aligned}\mathbb{E}[T_{\mathrm{B}}] &= \mathbb{E}[\mathbb{E}[T_{\mathrm{B}} \mid S = s]] \\ &= \sum_{i=0}^{N} \mathbb{E}\biggl[\mathrm{e}^{-\lambda S} \frac{(\lambda S)^i}{i!} S\biggr] \frac{i}{2} + \sum_{i=N+1}^{\infty} \mathbb{E}\biggl[\mathrm{e}^{-\lambda S} \frac{(\lambda S)^i}{i!} S\biggr] \biggl[1 - \frac{N+1}{2(i+1)}\biggr] N,\end{aligned} \tag{A.2}$$

where the summation and expectation is interchanged by Fubini's theorem since the terms are nonnegative. Then $\mathbb{E}[W(a)]$ can be obtained by (3.6). □

Appendix B.

Proof of Theorem 3.3. In an $\mathrm{M}/\mathrm{G}^{[a,\infty]}/1/\infty$ queue, $\mathbb{E}[N(a)] = \sum_{i=0}^{a-1} p_i a + \sum_{i=a}^{\infty} p_i i$, which yields (3.9).

Equations (A.1) and (A.2) can be further simplified as

$$\mathbb{E}[T_{\mathrm{B}} \mid S = s] = \frac{s}{2} \sum_{i=0}^{\infty} \mathrm{e}^{-\lambda s} \frac{(\lambda s)^i}{i!} i = \frac{\lambda s^2}{2}, \qquad \mathbb{E}[T_{\mathrm{B}}] = \mathbb{E}[\mathbb{E}[T_{\mathrm{B}} \mid S = s]] = \frac{\lambda \mathbb{E}[S^2]}{2}.$$

Then $\mathbb{E}[W(a)]$ can be obtained by (3.6). □

Appendix C.

Proof of Theorem 4.1. When $\lceil \lambda \mathbb{E}[W(a_{\mathrm{opt}})] \rceil > b$, we have

$$f(b-1) = b - 1 - \lambda \mathbb{E}[W(b-1)] < b - \lceil \lambda \mathbb{E}[W(b-1)] \rceil \le b - \lceil \lambda \mathbb{E}[W(a_{\mathrm{opt}})] \rceil < 0.$$

Thus, $\mathbb{E}[W(1)] > \mathbb{E}[W(2)] > \cdots > \mathbb{E}[W(b)]$. The optimal threshold is $a_{\mathrm{opt}} = b$.

It remains to prove the necessity and sufficiency of (4.6) under the condition that

$$\lceil \lambda \mathbb{E}[W(a_{\mathrm{opt}})] \rceil \le b.$$

Necessity. If $a_{\mathrm{opt}} = 1$ then we have

$$\Delta \mathbb{E}[W(1)] = \mathbb{E}[W(2)] - \mathbb{E}[W(1)] \ge 0.$$

Then $f(1) = 1 - \lambda \mathbb{E}[W(1)] \ge 0$, and $\lceil \lambda \mathbb{E}[W(a_{\mathrm{opt}})] \rceil = \lceil \lambda \mathbb{E}[W(1)] \rceil = 1$.

If $a_{\mathrm{opt}} > 1$ then based on the definition of optimality, we have

$$\Delta \mathbb{E}[W(a_{\mathrm{opt}})] \ge 0 \quad \text{and} \quad \Delta \mathbb{E}[W(a_{\mathrm{opt}} - 1)] < 0.$$

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

Hence,

$$f(a_{\text{opt}}) \geq 0 \quad \text{and} \quad f(a_{\text{opt}} - 1) < 0,$$

which then yield

$$a_{\text{opt}} \geq \lambda \mathbb{E}[W(a_{\text{opt}})]. \tag{C.1}$$

On the other hand, from inequality (4.5), we have

$$\sigma(a) > \frac{a}{\lambda} \geq \frac{1}{\lambda} > \frac{g(a-1)}{\lambda}. \tag{C.2}$$

Since $f(a_{\text{opt}} - 1) < 0$, we have

$$\lambda \Delta \mathbb{E}[W(a_{\text{opt}} - 1)] = \frac{g(a_{\text{opt}} - 1)/\lambda}{\sigma(a_{\text{opt}})} f(a_{\text{opt}} - 1) > f(a_{\text{opt}} - 1).$$

Thus,

$$\lambda \mathbb{E}[W(a_{\text{opt}})] - \lambda \mathbb{E}[W(a_{\text{opt}} - 1)] > a_{\text{opt}} - 1 - \lambda \mathbb{E}[W(a_{\text{opt}} - 1)].$$

Eliminating $\lambda \mathbb{E}[W(a_{\text{opt}} - 1)]$ on both sides, we have

$$\lambda \mathbb{E}[W(a_{\text{opt}})] + 1 > a_{\text{opt}}. \tag{C.3}$$

Combining (C.1) and (C.3), we then have

$$\lambda \mathbb{E}[W(a_{\text{opt}})] \leq a_{\text{opt}} < \lambda \mathbb{E}[W(a_{\text{opt}})] + 1.$$

Therefore, $a_{\text{opt}} = \lceil \lambda \mathbb{E}[W(a_{\text{opt}})] \rceil$.

Sufficiency. If $\lceil \lambda \mathbb{E}[W(1)] \rceil = 1$ then $f(1) = 1 - \lambda \mathbb{E}[W(1)] \geq 0$. Thus, $a_{\text{opt}} = 1$.

Suppose that $a_{\text{opt}} > 1$ and we have $\lceil \lambda \mathbb{E}[W(a)] \rceil = a$ for a threshold a, then

$$\lambda \mathbb{E}[W(a)] \leq a < \lambda \mathbb{E}[W(a)] + 1.$$

The first part yields

$$f(a) = a - \lambda \mathbb{E}[W(a)] \geq 0.$$

The second part yields

$$a - 1 < \lambda \mathbb{E}[W(a)].$$

Subtracting $\lambda \mathbb{E}[W(a-1)]$ from both sides, we have

$$f(a-1) = a - 1 - \lambda \mathbb{E}[W(a-1)] < \lambda \Delta \mathbb{E}[W(a-1)] = \frac{g(a-1)/\lambda}{\sigma(a)} f(a-1).$$

Applying (C.2), we obtain $f(a-1) < 0$. Since $f(a)$ is strictly increasing, we know that a is optimal. □

References

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

[2] Banerjee, A., Chakravarthy, S. R. and Gupta, U. C. (2015). Analysis of a finite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service. *Comput. Operat. Res.* **60,** 138–149.

[3] Banik, A. D. (2009). Queueing analysis and optimal control of $BMAP/G^{(a,b)}/1/N$ and $BMAP/MSP^{(a,b)}/1/N$ systems. *Comput. Indust. Eng.* **57,** 748–761.

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press

[4] Banik, A. D. (2015). Single server queues with a batch Markovian arrival process and bulk renewal or non-renewal service. *J. Systems Sci. Systems Eng.* **24,** 337–363.

[5] Bar-Lev, S. K. *et al.* (2007). Applications of bulk queues to group testing models with incomplete identification. *Europ. J. Operat. Res.* **183,** 226–237.

[6] Chakravarthy, S. (1993). Analysis of a finite MAP/G/1 queue with group services. *Queueing Systems* **13,** 385–407.

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

[8] Chaudhry M. L. and Templeton J. G. C. (1983). *A First Course on Bulk Queues.* John Wiley, New York.

[9] Deb, R. K. and Serfozo, R. F. (1973). Optimal control of batch service queues. *Adv. Appl. Prob.* **5,** 340–361.

[10] Gold, H. and Tran-Gia, P.(1993). Performance analysis of a batch service queue arising out of manufacturing system modelling. *Queueing Systems* **14,** 413–426.

[11] Goswami, V., Patra, S. S. and Mund, G. B. (2012). Performance analysis of cloud computing centers for bulk services. *Internat. J. Cloud Appl. Comput.* **2,** 53–65.

[12] Gupta, U. C. and Banerjee, A. (2011). New results on bulk service queue with finite-buffer: $\mathrm{M/G}^{(a,b)}/1/N$. *Opsearch* **48,** 279–296.

[13] Meyer, C. D. (1989). Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Rev.* **31,** 240–272.

[14] Neuts, M. F. (1967). A general class of bulk queues with Poisson input. *Ann. Math. Statist.* **38,** 759–770.

[15] Sikdar, K. and Gupta, U. C. (2005) Analytic and numerical aspects of batch service queues with single vacation. *Comput. Operat. Res.* **32,** 943–966.

[16] Tadj, L. and Choudhury, G.(2005). Optimal design and control of queues. *Top* **13,** 359–412.

[17] Tadj, L. and Ke, J.-C. (2003). Control policy of a hysteretic queueing system. *Math. Meth. Operat. Res.* **57,** 367–376.

[18] Tadj, L. and Ke, J.-C. (2005). Control policy of a hysteretic bulk queueing system. *Math. Comput. Modelling* **41,** 571–579.

[19] Tadj, L. and Tadj, C.(2003). On an $\mathrm{M/D}^r/1$ queueing system. *J. Statist. Theory Appl.* **2,** 17–32.

https://doi.org/10.1017/jpr.2017.8 Published online by Cambridge University Press