Queueing Systems (2020) 95:251–269
https://doi.org/10.1007/s11134-020-09654-y

Batch service systems with heterogeneous servers

Jan-Kees van Ommeren[1] · Niek Baer[1] · Nishant Mishra[2] · Debjit Roy[3]

Received: 4 August 2019 / Revised: 30 March 2020 / Published online: 11 June 2020

Abstract
Bulk-service multi-server queues with heterogeneous server capacity and thresholds are commonly seen in several situations such as passenger transport or package delivery services. In this paper, we develop a novel decomposition-based solution approach for such queues using arguments from renewal theory. We then obtain the distribution of the waiting time measure for multi-type server systems. We also obtain other useful performance measures such as utilization, expected throughput time, and expected queue lengths.

Keywords Heterogeneous capacity · Threshold service · Decomposition

Mathematics Subject Classification 60K25 · 90B22

1 Introduction

Heterogeneous servers have been a useful modeling construct for performance analysis of telecommunication networks [4]. During recent years, other applications of hetero-

✉ Jan-Kees van Ommeren
j.c.w.vanommeren@utwente.nl

Niek Baer
niekbaer@hotmail.com

Nishant Mishra
nishant.mishra@uclouvain.be

Debjit Roy
debjit@iima.ac.in

[1] Stochastic Operations Research, University of Twente, Postbox 217, 7500 AE Enschede, The Netherlands

[2] Center for Operations Research and Econometrics, UCLouvain, Louvain-la-Neuve 1348, Belgium

[3] Production and Quantitative Methods Area, Indian Institute of Management Ahmedabad, Ahmedabad, India

geneous servers with bulk service have been found in container terminal systems [11], passenger transport systems, and parcel distribution networks. In a passenger transport system, the servers can vary in their carrying capacity (for example, a 20-seater bus vs. a 40-seater bus) and serves (transports) the passengers in a batch. Hence, the servers may have heterogeneous capacity. Likewise, we can model a parcel distribution network using heterogeneous capacity servers. The freight movers own or rent trucks from the market of different capacities. The trucks vary in their body length and carrying capacity. For instance, Light Duty Box Trucks typically have a carrying capacity of 8600–14,050 lbs, whereas Heavy-duty Flatbed Trucks typically have a carrying capacity of 26,000–52,000 lbs. The trucks have a threshold load limit before they leave the origin dispatch unit. The performance analysis problem (such as the number and quantity of loads waiting at a depot before being dispatched) can be modeled using heterogeneous capacity resources (multiple types of trucks). However, the literature on performance analysis of bulk-service queues with heterogeneous servers and threshold service is scarce.

In Arora [3], a heterogeneous two-server queueing process fed by Poisson arrivals and exponential service time distributions has been considered under the bulk-service discipline. Time-dependent probabilities for the queue length have been obtained in terms of Laplace transforms, from which different measures associated with the queueing process, like the mean queue-length, could be determined. Goswami and Samanta [8] analyzed a discrete-time bulk-service queueing system with two heterogeneous servers, i.e., two batch servers working with different service rates. They assumed the interarrival times of customers and service times of batches to be independent and geometrically distributed. They obtain closed-form expressions for the steady-state probabilities at an arbitrary epoch with the help of the displacement operator method and derive the outside observer's observation epoch probabilities and waiting time distribution measured in slots. Chakka and Van Do [4] proposed a new HetSigma queue for performance analysis of wireless communication systems. They use negative customers to model server failures, packet losses, and load balancing in networks. They analyze joint Markov modulation of the arrival and service processes, superposition of K Compound Poisson Process (CPP) streams of positive customer arrivals and a CPP of negative customer arrivals in each modulating phase for a multi-server queue with c non-identical servers, and generalized exponential service times.

Using a matrix geometric method, Kumar and Madheswari [10] obtained the stationary queue length distribution and mean system size for a Markovian queue with two heterogeneous servers and multiple vacations. Using a generating function technique, Ammar [2] analyzed the transient behavior (exact time dependent solutions) of a two-processor heterogeneous system with catastrophes, server failures and repairs. The tasks arrive according to a Poisson process and service times are exponentially distributed. Each task requires exactly one processor for its execution and the scheduling policy is FCFS. Using the embedded method, Keaogile et al. [9] presented an exact analysis for finding the probability generating function of the steady state number of customers in a discrete time queue with two heterogeneous servers.

While there are several studies that analyze batch service queues with single or multiple homogeneous servers, studies on analysis of batch service queues with heterogeneous server systems are limited (for example, see Chang and Harn [5], Chen et al. [6], Gold and Tran-Gia [7], Aalto [1]). We contribute to the literature in the following ways: (1) we analyze batch service systems with heterogeneous servers and threshold service capacity. In many practical settings, the server may have a large service capacity; however, the service is initiated only when a threshold capacity of the server is utilized. Such settings are commonly observed in amusement parks, bus services, multi-trailer systems due to revenue or system functionality considerations. (2) We perform an exact analysis of batch service systems using a combination of Markov chains and system state decomposition. We decompose the system using a free and busy periods state of the system, perform separate analysis, and then combine the results from the free and the busy periods to obtain the joint probability of the number of free servers and the number of customers waiting in the queue. This joint probability leads both to the waiting time distribution of a customer and to performance measures such as used server capacity, expected throughput time, expected waiting time, and expected queue lengths. Such measures are useful for batch service system design. Due to the complexity of the analysis, we show the results for exponential service times only. However, our work can be extended to batch service queues with heterogeneous servers and service times which have a phase-type distribution, albeit with significant numerical costs associated with a large system state space.

The rest of the paper is organized as follows: In Sect. 2, we describe the queueing model. In Sect. 3, we perform the analysis of the queueing system with two types of servers and batch service. We extend our analysis to more than two server types with batch services in Sect. 4. Results from numerical experiments are included in Sect. 5. Finally, we draw our conclusions in Sect. 6.

2 Model

We consider a system with S types of batch servers characterized by their rates and capacity, denoted as the $M_\lambda / \sum_{\sigma=1}^{S} M_{\mu_\sigma}^{T_\sigma, B_\sigma} / \sum_{\sigma=1}^{S} N_\sigma$ queue. Here λ denotes the Poisson customer arrival rate; for $\sigma = 1, \ldots, S$, let N_σ denote the number of servers of type σ, each with a bulk exponential service rate of μ_σ, with a maximum batch size B_σ and a minimum batch size T_σ. The service times are assumed to be independent of the batch sizes in process. Furthermore, we assume that if a type σ server is free, there are fewer than T_σ customers in the queue. We assume for $1 \le \tau < \sigma \le S$ that $T_\tau \le T_\sigma$; if $T_\tau = T_\sigma$ then we assume that type τ servers have priority.

In the following section, we analyze the system with $S = 2$ types of servers. In Sect. 4, we give some results for the general case $S \ge 2$.

3 Analysis of the $M_\lambda / M_{\mu_1}^{T_1, B_1} + M_{\mu_2}^{T_2, B_2} / N_1 + N_2$ queue

We analyze the $M_\lambda / M_{\mu_1}^{T_1, B_1} + M_{\mu_2}^{T_2, B_2} / N_1 + N_2$ queue to find the waiting time distribution. We first concentrate on the joint probability function of the number of busy servers of type $\sigma = 1, 2$ and the queue length. We then find the residual waiting time distribution of a tagged customer, conditional on the free servers and the number of

Table 1 Transition from state (n_1, n_2, n_Q)

To state	Rates conditional on		
	$n_1 < N_1$	$n_1 = N_1, n_2 < N_2$	$n_1 = N_1, n_2 = N_2$
$(n_1, n_2, n_Q + 1)$	$\lambda I_{\{n_Q+1<T_1\}}$	$\lambda I_{\{n_Q+1<T_2\}}$	λ
$(n_1 + 1, n_2, 0)$	$\lambda I_{\{n_Q+1=T_1\}}$		
$(n_1, n_2 + 1, 0)$		$\lambda I_{\{n_Q+1=T_2\}}$	
$(n_1 - 1, n_2, n_Q)$	$n_1\mu_1$	$N_1\mu_1 I_{\{n_Q<T_1\}}$	$N_1\mu_1 I_{\{n_Q<T_1\}}$
$(n_1, n_2, (n_Q - B_1)^+)$		$N_1\mu_1 I_{\{n_Q\geq T_1\}}$	$N_1\mu_1 I_{\{n_Q\geq T_1\}}$
$(n_1, n_2 - 1, n_Q)$	$n_2\mu_2$	$n_2\mu_2$	$N_2\mu_2 I_{\{n_Q<T_2\}}$
$(n_1, n_2, (n_Q - B_2)^+)$			$N_2\mu_2 I_{\{n_Q\geq T_2\}}$

waiting customers who arrived either before or after the tagged customer. The waiting time of a customer equals the residual waiting time given the number of busy servers and the queue length at arrival and that there are no waiting customers who arrived after this customer. By the fact that ‘Poisson arrivals see time averages’, the joint probability function of the number of busy servers and the queue length at arrival was already determined, so we can find the unconditional waiting time distribution.

3.1 The joint probability function of the number of busy servers of each type and the queue length

The state space for this queue can be expressed using a three-tuple (n_1, n_2, n_Q), with $n_\sigma = 0, \ldots, N_\sigma, \sigma = 1, 2$ and $n_Q = 0, 1, \ldots$, where n_1 and n_2 denote the number of busy servers of type 1 and type 2 respectively and n_Q denotes the number of waiting customers. Note that if $n_\sigma < N_\sigma$ then $n_Q < T_\sigma$. Due to our assumptions, we can find the transition rates from (n_1, n_2, n_Q); these transition rates are given in Table 1. The rate of service completion is proportional to the number of type 1 and type 2 busy servers, i.e., $n_1\mu_1 + n_2\mu_2$. If there are fewer than T_σ customers in the queue, a type σ server will not begin its service.

To analyze the continuous time Markov chain (CTMC), we split the state space into a free period and a busy period (see Fig. 1). During the free period (FP) at least one server is free. The states (n_1, n_2, n_Q) with $n_1 + n_2 < N_1 + N_2$, and $n_Q < T_\sigma$ and $n_\sigma < N_\sigma, \sigma = 1, 2$ correspond to the free period. During the busy period (BP), all servers are busy, i.e., the states (N_1, N_2, n_Q) with $n_Q = 0, 1, \ldots$. When a BP starts, we know that $n_Q = 0$. When a FP starts, we know that either $n_1 = N_1 - 1$ and $n_2 = N_2$ or $n_1 = N_1$ and $n_2 = N_2 - 1$. However, when we enter a FP, the distribution of n_Q is unknown. We now discuss the free and the busy period analysis in the following sections.

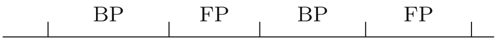

Fig. 1 Timeline of free and busy periods

Busy period analysis

During the BP, all the servers are busy. Therefore, we only have to focus on the queue length, which, together with the state '-1' representing the FP, can be described by a CTMC process.

The transition rate from '-1' (the state representing the FP) to '0' is arbitrarily chosen since the sojourn time in '-1' in combination with the probability that the process is in state '-1' is only used to compute the expected length of the BP. A BP starts always with $N_Q = 0$ customers in the queue. We are interested in $\pi_{\text{BP}}(n)$, the conditional probability of being in state 'n' given that all servers are busy, and $E(T_B)$, the expected length of a BP. The rate up from any state is λ and the rate down to or below state $n = 0, 1, \ldots$ equals $N_1\mu_1 \sum_{k=1}^{B_1} \pi(n+k) + N_2\mu_2 \sum_{k=1}^{B_2} \pi(n+k)$, where we used that during a BP all servers are busy (see Fig. 2). The steady state probabilities for the queue length N_Q are defined by the following balance equations:

$$\lambda\pi(-1) = N_1\mu_1 \sum_{k=0}^{T_1-1} \pi(k) + N_2\mu_2 \sum_{k=0}^{T_2-1} \pi(k),$$

$$\lambda\pi(n) = N_1\mu_1 \sum_{k=1}^{B_1} \pi(n+k) + N_2\mu_2 \sum_{k=1}^{B_2} \pi(n+k).$$

It is easily checked that the steady state probabilities for the states '-1' and 'n' (i.e., $\pi(-1)$ and $\pi(n)$) are provided by

$$\pi(-1) = \frac{N_1\mu_1(1-\alpha^{T_1}) + N_2\mu_2(1-\alpha^{T_2})}{N_1\mu_1(1-\alpha^{T_1}) + N_2\mu_2(1-\alpha^{T_2}) + \lambda}, \tag{1}$$

$$\pi(n) = \alpha^n(1-\alpha)(1-\pi(-1)) \text{ for } n = 0, 1, \ldots, \tag{2}$$

with α equal to the unique solution in (0, 1) of the equation

$$\lambda(1-\alpha) = N_1\mu_1\alpha(1-\alpha^{B_1}) + N_2\mu_2\alpha(1-\alpha^{B_2}), \tag{3}$$

for $\lambda < N_1\mu_1 B_1 + N_2\mu_2 B_2$.

The length of the BP and its expectation can be found by considering the above described Markov chain as a regenerative process, which regenerates when it enters state '-1'. The number of consecutive times that the states are positive equals the

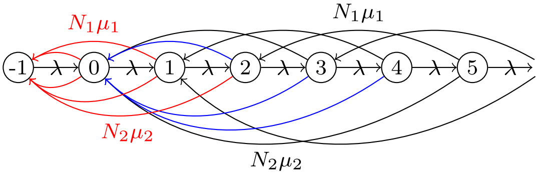

Fig. 2 CTMC for the busy period where $N_1 = 4$, $T_1 = 2$, $B_1 = 3$, $N_2 = 2$, $T_2 = 3$, $B_2 = 5$

length of a BP. The theory of regenerative processes now gives

$$\pi(-1) = \frac{1/\lambda}{1/\lambda + E(T_{\mathrm{BP}})},$$

which, combined with (1), leads to

$$E(T_{\mathrm{BP}}) = \frac{1-\pi(-1)}{\lambda\pi(-1)} = \frac{1}{N_1\mu_1(1-\alpha^{T_1}) + N_2\mu_2(1-\alpha^{T_2})}. \tag{4}$$

The conditional distribution of the queue length during a BP is found by (2). This gives

$$\pi_{\mathrm{BP}}(n) = \alpha^n(1-\alpha), n = 0, 1, \ldots. \tag{5}$$

In the analysis of the FP, we need to know in which states this FP starts, or, equivalently, how the BP ends. The probability that the BP is ended by a type σ server becoming free which sees $n_Q < T_\sigma$ customers at the end of the BP is denoted by $P(\text{type } \sigma \text{ server}, n_Q)$. Note that the end of a BP corresponds to entering state -1. By conditioning on entering state -1, we find that

$$P(\text{type } \sigma \text{ server}, n_Q) = \frac{N_\sigma\mu_\sigma\alpha^{n_Q}(1-\alpha)}{N_1\mu_1(1-\alpha^{T_1}) + N_2\mu_2(1-\alpha^{T_2})} \tag{6}$$

for $n_Q = 0, \ldots, T_\sigma - 1$ and $\sigma = 1, 2$.

Free period analysis

In the free period, we not only need to keep track of how many customers are waiting, but also how many servers of each type are busy. The state space for the free period can be described as follows:

$$\begin{aligned}\Big\{(n_1, n_2, n_Q)\,\Big|n_1 = 0, \ldots, N_1 - 1, n_2 = 0, \ldots, N_2, n_Q = 0, \ldots, T_1 - 1\Big\} \cup \\ \Big\{(N_1, n_2, n_Q)\,\Big|n_2 = 0, \ldots, N_2 - 1, n_Q = 0, \ldots, T_2 - 1\Big\} \cup \\ \{(N_1, N_2, 0)\}.\end{aligned}$$

The state $(N_1, N_2, 0)$ represents the BP, with some arbitrary total rate out. As in the analysis of the BP, the sojourn time in this extra state is only used to find the length of the FP. The transition rates are given in Table 1, except for the extra state $(N_1, N_2, 0)$. The transition rates from $(N_1, N_2, 0)$ to $(N_1 - 1, N_2, n_Q)$ and to $(N_1, N_2 - 1, n_Q)$ equal, respectively, $N_1\mu_1\alpha^{n_Q}(1-\alpha)$ and $N_2\mu_2\alpha^{n_Q}(1-\alpha)$ for $n_Q = 0, \ldots, T_\sigma - 1$ and $\sigma = 1, 2$. Note that the rates are proportional to the exit probability of the BP given in (6).

The expected length of a FP can be found analogously to the derivation of the length of the BP, cf. (4),

$$E(T_{\mathrm{FP}}) = \frac{1-\pi(N_1, N_2, 0)}{\pi(N_1, N_2, 0)(N_1\mu_1(1-\alpha^{T_1}) + N_2\mu_2(1-\alpha^{T_2}))}.$$

For the FP, however, no simple expression exists for $\pi(N_1, N_2, 0)$ or for $E(T_{\mathrm{FP}})$. We have to use numerical methods to compute these probabilities.

Combining the results for the busy period and the free period leads to the steady state joint probability of the number of busy servers of each type and the number of waiting customers:

$$\pi(n_1, n_2, n_Q) = \begin{cases} \pi_{\mathrm{BP}}(n_Q)P(\mathrm{BP}), & n_1 = N_1,\ n_2 = N_2, \\ \pi_{\mathrm{FP}}(n_1, n_2, n_Q)P(\mathrm{FP}), & n_1 + n_2 < N_1 + N_2, \end{cases} \tag{7}$$

where $P(\mathrm{BP})$ and $P(\mathrm{FP})$, the probabilities that the system is in a BP, resp. an FP, are given by

$$P(\mathrm{BP}) = \frac{E(T_{\mathrm{BP}})}{E(T_{\mathrm{BP}}) + E(T_{\mathrm{FP}})} \text{ and } P(\mathrm{FP}) = 1 - P(\mathrm{BP}).$$

This joint probability distribution is used in the next section to find the waiting time distribution.

3.2 Waiting time

In this section, we find the steady-state waiting time distribution of a customer. First the special case $T_2 = 1$ is analyzed. In this case a server cannot be idle when a customer is waiting. It is clear that a customer can only wait when all the servers are busy and that a waiting time ends at a service completion. In case that $T_2 > 1$, it might happen that a waiting time ends at an arrival of a new customer. This requires a totally different analysis.

The special case $T_2 = 1$

For the special case where a server will not be free when there are customers waiting, that is, the minimal batch sizes $T_1 = T_2 = 1$, we can find the waiting time distribution as follows.

Consider N_D, the number of arrivals during a waiting time. N_D, given that the waiting time has length d, has a Poisson distribution with $EN_D = \lambda d$. Unconditioning gives us that the z-transform of N_D is given by $E\left(z^{N_D}\right) = \hat{W}(\lambda(1 - z))$, where $\hat{W}$ denotes the Laplace Stieltjes transform (LST) of the steady-state waiting time distribution. An up-down crossing argument gives that N_D is identically distributed to N_A, where N_A denotes the number of customers waiting at an arrival epoch (here we assume that customers enter a batch one by one in order of arrival). By the PASTA property, N_A and N_Q are identically distributed. Using the time stationary probabilities $\pi(n_Q)$ found in the previous subsection, we can find an expression for the LST of the waiting time:

$$\hat{W}(s) = P(\mathrm{FP}) + P(\mathrm{BP})\frac{\lambda(1 - \alpha)}{\lambda(1 - \alpha) + \alpha s}.$$

The general case $T_2 \geq 1$

In the case $T_2 > 1$, it can happen that a type 2 server is free because there are fewer than T_2 customers waiting. Once the threshold is met by an arrival, the server starts working and the waiting of the customers in the queue ends. This phenomenon prevents us from employing the technique used for the case $T_2 = 1$.

The waiting time of an arriving customer depends not only on the number of waiting customers, but also whether there is a type 1 server free or (only) a type 2 server. Suppose there is a type 1 server free, then we know that the number of waiting customers at arrival $n_Q < T_1$. It is readily seen that if $n_Q = T_1 - 1$, the arriving customer has no waiting time. If $n_Q = T_1 - 1 - k$, with $k = 1, \ldots, T_1 - 1$, the waiting time of the customer ends after k new arrivals and therefore has a Erlang-k distribution. If no type 1 server is free but a type 2 server is free, it becomes a bit more complicated. We can have the same reasoning as in the previous case, but now it can happen that a type 1 server becomes free, the number of waiting customers exceeds T_1, and the type 1 server starts a new service. Even then it is possible that the waiting of our tagged customer does not end, because he did not fit in the type 1 batch. Thus, to decide whether its waiting time ends, we need to know both the number of waiting customers which arrived before him and after him. If no server is free at the arrival, the customer first has to wait until he fits in the batch of a server that becomes free. Even if that is the case, it might be possible that he still has to wait, because the threshold batch size of the server is not reached.

To analyze the waiting time of an arbitrary customer, we first concentrate on the remaining waiting time of this tagged customer at some time point. As observed, the remaining waiting time depends on whether a type 1 server is free, or only a type 2 server, or no server at all, and on the number of waiting customers which arrived before, respectively after, the tagged customer.

Remaining waiting time

Suppose we follow a customer during its waiting. We will decompose the remaining waiting time of this customer into two parts, namely the time until a new customer arrives (this event is denoted by A) or a type σ server ends its service (denoted by F_σ), and the remaining waiting time after this event. We denote the remaining waiting time of the customer when there are $n - 1$ waiting customers who arrived before him and ℓ who arrived after him by $W_{n,\ell}$ if there is no server free, by $W^{(1)}_{n,\ell}$ if there is a type 1 server free and by $W^{(2)}_{n,\ell}$ if there is only a type 2 server free.

When a customer has to wait while a type 1 server is free, it has to wait at least until a customer arrives ($\text{Exp}(\lambda)$) and then it either is taken into service or has to wait with one customer more behind him in the queue. This gives us the following relation:

$$W^{(1)}_{n,\ell} = \begin{cases} \text{Exp}(\lambda) + W^{(1)}_{n,\ell+1}, & n + \ell < T_1, \\ 0, & n + \ell \geq T_1, \end{cases}$$

where Exp(λ) denotes a random variable with an exponential distribution and mean $1/\lambda$.

If only a type 2 server is free, there are more relevant events, namely an arrival of a customer or the end of a type 1 service. In the first event (A), the number of customers behind him is increased (recorded until $T_2 - 1$ is reached). In the second event (F_1), the customer might be taken into service (when he is at the head of the line and there are enough customers in queue to fill the batch threshold), or the number of customers in front of him is decreased by B_σ and the type 1 server is busy. This gives

$$W^{(2)}_{n,\ell} = \begin{cases} 0, & n+\ell \geq T_2, \\ \text{Exp}(\lambda + N_1\mu_1) + W^{(2)}_{n,\ell+1} 1_{\{A\}} & \\ \quad + W^{(1)}_{n,\ell} 1_{\{F_1, n \leq B_1\}} + W^{(2)}_{n-B_1,\ell} 1_{\{F_1, n > B_1\}}, & n+\ell < T_2. \end{cases}$$

Finally, for the case where no server is free, the system is in a BP. When a customer waits during a BP, the first part of its waiting time ends either by an arrival, or the end of a service by a type 1 or type 2 server. In the first case, the number of customers behind him is increased. In the second case the customer is either taken into service (when he is at the head of the line and there are enough customers in the queue to fill the batch threshold), or the number of customers in front of him is decreased by B_σ, or an FP starts. This gives

$$W_{n,\ell} = \text{Exp}(M_\lambda) + W_{n,\ell+1} 1_{\{A\}} + \sum_{\sigma=1}^{2} \left(W^{(\sigma)}_{n,\ell} 1_{\{F_\sigma, n \leq B_\sigma\}} + W_{n-B_\sigma,\ell} 1_{\{F_\sigma, n \leq B_\sigma\}} \right),$$

where $M_\lambda = N_1\mu_1 + N_2\mu_2 + \lambda$. Note that $W_{n,\ell} = W_{n,\ell+1}$ for $\ell \geq T_2 - 1$ since the threshold of any batch is always met when $\ell \geq T_2 - 1$.

Now that we have the relations for the remaining waiting time of a customer, we can derive relations for the corresponding Laplace Stieltjes transforms (LST)

$$\phi_{n,\ell}(s) = E\left(e^{-sW_{n,\ell}}\right) \text{ and } \phi^{(\sigma)}_{n,\ell}(s) = E\left(e^{-sW^{(\sigma)}_{n,\ell}}\right), \quad \sigma = 1, 2.$$

If a type 1 server is free we find

$$\phi^{(1)}_{n,\ell}(s) = \begin{cases} \dfrac{\lambda\phi(1)_{n,\ell+1}(s)}{\lambda + s}, & n+\ell < T_1, \\ 1, & n+\ell \geq T_1, \end{cases} \tag{8}$$

If only a type 2 server is free we find

$$\phi^{(2)}_{n,\ell}(s) = \begin{cases} \dfrac{\lambda\phi^{(2)}_{n,\ell+1}(s) + N_1\mu_1\left(\phi^{(1)}_{n,\ell}(s) 1_{\{n \leq B_1\}} + \phi^{(2)}_{n-B_1,\ell}(s) 1_{\{n > B_1\}}\right)}{\lambda + N_1\mu_1 + s}, & \\ & n+\ell < T_2, \\ 1 & n+\ell \geq T_2. \end{cases} \tag{9}$$

$$\begin{aligned}\phi_{n,\ell}(s) = \frac{M_\lambda}{M_\lambda + s}\Bigg(&\frac{\lambda}{M_\lambda}\phi_{n,\ell+1}(s)\\ &+\frac{N_1\mu_1}{M_\lambda}\left(\phi_{n-B_1,\ell}(s)1_{\{n>B_1\}} + \phi^{(1)}_{n,\ell}(s)1_{\{n\le B_1\}}\right)\\ &+\frac{N_2\mu_2}{M_\lambda}\left(\phi_{n-B_2,\ell}(s)1_{\{n>B_2\}} + \phi^{(2)}_{n,\ell}(s)1_{\{n\le B_2\}}\right)\Bigg).\end{aligned}$$

Finally, we define the double generating function $\phi_\ell(z, s)$ for $\ell = 0, 1, \ldots$ by

$$\begin{aligned}\phi_\ell(z, s) &= (1 - z)\sum_{n=1}^{\infty} z^{n-1}\phi_{n,\ell}(s) \qquad (10)\\ &= \frac{(1 - z)\sum_{\sigma=1}^{2} N_\sigma\mu_\sigma \sum_{n=1}^{B_\sigma}\phi^{(\sigma)}_{n,\ell}(s)z^{n-1} + \lambda\phi_{\ell+1}(z, s)}{\sum_{\sigma=1}^{2} N_\sigma\mu_\sigma(1 - z^{B_\sigma}) + \lambda + s}.\end{aligned}$$

Note that $\phi_{\ell+1}(z, s) = \phi_\ell(z, s)$ for $\ell = T_2 - 1, T_2, \ldots$. Therefore, we will use (10) only for $\ell = 0, \ldots, T_2 - 1$ and set $\phi_{T_2}(z, s) = \phi_{T_2-1}(z, s)$.

Waiting time

In this section, we concentrate on the waiting time of an arriving customer. On arrival of a customer, we see that this waiting time is the remaining waiting time of this customer with the same state of the servers, the same queue length and no waiting customers who arrived after him. So we can write

$$W = \begin{cases} W_{n_Q+1,0} & \text{if all servers busy,}\\ W^{(1)}_{n_Q+1,0} & \text{if a type 1 server is free,}\\ W^{(2)}_{n_Q+1,0} & \text{if only a type 2 server is free,}\end{cases}$$

By these equations, we find the LST for W as

$$\begin{aligned}E(e^{-sW}) = P(\mathrm{BP})&\sum_{n_Q=0}^{\infty}\pi_{\mathrm{BP}}(n_Q)\phi_{n_Q+1,0}(s)\\ &+P(\mathrm{FP})\Bigg(\sum_{n_Q=0}^{T_1-2}\sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2}\pi_{\mathrm{FP}}(n_1, n_2, n_Q)\phi^{(1)}_{n_Q+1,0}(s)\\ &+\sum_{n_Q=0}^{T_2-2}\sum_{n_2=0}^{N_2-1}\pi_{\mathrm{FP}}(N_1, n_2, n_Q)\phi^{(2)}_{n_Q+1,0}(s)\Bigg)\\ =P(\mathrm{BP})\,&\phi_0(\alpha, s)\end{aligned}$$

$$+P(\mathrm{FP})\Bigg(\sum_{n_Q=0}^{T_1-2}\sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2}\pi_{\mathrm{FP}}(n_1,n_2,n_Q)\phi^{(1)}_{n_Q+1,0}(s)$$
$$+\sum_{n_Q=0}^{T_2-2}\sum_{n_2=0}^{N_2-1}\pi_{\mathrm{FP}}(N_1,n_2,n_Q)\phi^{(2)}_{n_Q+1,0}(s)\Bigg),$$

where the double generating function $\phi_0(\alpha, s)$ is defined in (10). With this expression for the LST of the waiting time and (8, 9), it is easy to show the next theorem.

Theorem 1 *The waiting time in the* $M_\lambda/M_{\mu_1}^{T_1,B_1} + M_{\mu_2}^{T_2,B_2}/N_1 + N_2$ *queue can be considered as an absorption time of a Markov chain with state space*

$$S = \{0, (1, m, \ell_1), (2, m, \ell_2), (0, \ell_2) | m = 0, \ldots, T_\sigma - \ell_\sigma, \ell_\sigma = 0, \ldots, T_\sigma - 1, \sigma = 1, 2\},$$

where '0' is the absorbing state, with initial probabilities

$$\begin{aligned}
P_I(1, m, 0) &= P(FP)\sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2}\pi_{FP}(n_1, n_2, m-1), && m = 1, \ldots T_1 - 1,\\
P_I(2, m, 0) &= P(FP)\sum_{n_2=0}^{N_2-1}\pi_{FP}(N_1, n_2, m-1), && m = 1, \ldots T_2 - 1,\\
P_I(0, 0) &= P(BP),\\
P_I(0) &= 1 - \sum_{m=1}^{T_1-1} P_I(1, m, 0) - \sum_{m=1}^{T_2-1} P_I(2, m, 0) - P_I(0, 0),
\end{aligned}$$

and transition rates

From	To	Transition rate
$(1, m, \ell)$	$(1, m, \ell + 1)$	$I_{\{m+\ell<T_1-1\}}\lambda,$
	0	$I_{\{m+\ell\geq T_1-1\}}\lambda,$
$(2, m, \ell)$	$(2, m, \ell + 1)$	$I_{\{m+\ell<T_2-1\}}\lambda,$
	$(1, m, \ell)$	$I_{\{m+\ell<T_1\}}N_1\mu_1,$
	$(2, m - B_1, \ell)$	$I_{\{m>B_1\}}N_1\mu_1,$
	0	$I_{\{m+\ell=T_2-1\}}\lambda + I_{\{m+\ell\geq T_1\}}I_{\{m\leq B_1\}}N_1\mu_1,$
$(0, \ell)$	$(0, \ell + 1)$	$I_{\{\ell<T_2-1\}}\lambda,$
	$(1, m, \ell)$	$\alpha^{m-1}(1-\alpha)N_1\mu_1,$ for $m = 1, \ldots, T_1 - 1 - \ell,$
	$(2, m, \ell)$	$\alpha^{m-1}(1-\alpha)N_2\mu_2,$ for $m = 1, \ldots, T_2 - 1 - \ell,$
	0	$\left(\alpha^{[T_1-\ell-1]^+} - \alpha^{B_1}\right)N_1\mu_1 + \left(\alpha^{T_2-\ell-1} - \alpha^{B_2}\right)N_2\mu_2,$

where $[n]^+ = \max(0, n)$.

3.3 Other useful performance measures

Apart from the waiting time, there are also other interesting performance measures such as the throughput of servers per type per period, the throughput of customers per type of server per period, the used capacity per type of server per period, the fraction of used capacity per type of server, and the expected waiting and throughput time. In this section, we give an expression for these performance measures in terms of the steady state probabilities $\pi_{\text{FP}}(n_1, n_2, n_Q)$, $\pi_{\text{BP}}(n_Q)$, $P(\text{FP})$ and $P(\text{BP})$.

To find the throughput of servers per type in the free period ($TH_\sigma(\text{FP})$), we use the interpretation that the steady state distribution also represents the fraction of time the system is in a certain state. For any state we know the departure rate of the servers. This gives

$$TH_1(\text{FP}) = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2} \lambda \pi_{\text{FP}}(n_1, n_2, T_1 - 1) + \sum_{n_2=0}^{N_2-1} \sum_{n_Q=T_1}^{T_2-1} N_1 \mu_1 \pi_{\text{FP}}(N_1, n_2, n_Q),$$

$$TH_2(\text{FP}) = \sum_{n_2=0}^{N_2-1} \lambda \pi_{\text{FP}}(N_1, n_2, T_2 - 1).$$

For the busy period we do the same; here the server only starts a service when at least his threshold is met and we find for the throughput of servers per type in the busy period

$$TH_\sigma(\text{BP}) = N_\sigma \mu_\sigma \alpha^{T_\sigma} \text{ for } \sigma = 1, 2.$$

The throughput of customers per type of server per period ($TH_\sigma^C(P)$, $P = \text{BP}, \text{FP}$) can be found in a similar way as the throughput of servers, but here we also have to count the number of customers being served simultaneously by a server. This gives

$$\begin{aligned} TH_1^C(\text{FP}) &= \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2} \lambda \pi_{\text{FP}}(n_1, n_2, T_1 - 1) T_1 \\ &\quad + \sum_{n_2=0}^{N_2-1} \sum_{n_Q=T_1}^{T_2-1} N_1 \mu_1 \pi_{\text{FP}}(N_1, n_2, n_Q) \min(n_Q, B_1), \\ TH_2^C(\text{FP}) &= \sum_{n_2=0}^{N_2-1} \lambda \pi_{\text{FP}}(N_1, n_2, T_2 - 1) T_2, \end{aligned}$$

and

$$TH_\sigma^C(\text{BP}) = N_\sigma \mu_\sigma \left(B_\sigma \alpha^{B_\sigma} + \sum_{n_Q=T_\sigma}^{B_\sigma-1} \alpha^{n_Q} (1-\alpha) n_Q \right) \text{ for } \sigma = 1, 2.$$

Now we have the throughput of both servers and customers, we can look at the average used capacity per type of server per period:

$$UC_\sigma(P) = \frac{TH_\sigma^C(P)}{TH_\sigma(P)} \text{ for } \sigma = 1, 2 \text{ and } P = \text{BP, FP},$$

and the fraction of used capacity per type of server:

$$UC_\sigma = \frac{TH_\sigma^C(\text{FP})E(T_{\text{FP}}) + TH_\sigma^C(\text{BP})E(T_{\text{BP}})}{(TH_\sigma(\text{FP})E(T_{\text{FP}}) + TH_\sigma(\text{BP})E(T_{\text{BP}}))B_\sigma} \text{ for } \sigma = 1, 2.$$

Finally, we also find the expected waiting time and the expected throughput time without computing their respective distribution. For the expected waiting time, we use Little's Law to find

$$\begin{aligned}E(W) = \frac{1}{\lambda}\Bigg(& P(\text{BP})\frac{1-\alpha}{\alpha} \\ & + P(\text{FP})\left(\sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2}\sum_{n_Q=0}^{T_1-1} n_Q \pi_{\text{FP}}(n_1, n_2, n_Q)\right. \\ & \left. + \sum_{n_2=0}^{N_2-1}\sum_{n_Q=0}^{T_2-1} n_Q \pi_{\text{FP}}(N_1, n_2, n_Q)\right)\Bigg).\end{aligned}$$

To find the expected throughput time, the sum of the waiting time (W) and the service time (T_{S}), we now only need to find the expected service time

$$E(T_{\text{S}}) = \sum_{\sigma=1}^{2} P(\text{served by type } \sigma \text{ server})/\mu_\sigma,$$

where

$$P(\text{served by type } \sigma \text{ server}) = \frac{TH_\sigma^C(\text{FP})E(T_{\text{FP}}) + TH_\sigma^C(\text{BP})E(T_{\text{BP}})}{\sum_{\sigma=1}^{2} TH_\sigma^C(\text{FP})E(T_{\text{FP}}) + TH_\sigma^C(\text{BP})E(T_{\text{BP}})}.$$

4 Analysis of queue with multi-type of servers

In this section, we give some results for the case where $S \geq 2$. We refer to the corresponding formulas and the given arguments in the previous sections. The state space for the queue can be expressed using a vector $(n_1, \ldots, n_S, n_Q)$, with $n_\sigma = 0, \ldots, N_\sigma$, $\sigma = 1, \ldots, S$ and $n_Q = 0, 1, \ldots$, where n_σ denote the number of busy servers of type σ and n_Q denotes the number of waiting customers. Note that if $n_\sigma < N_\sigma$ then $n_Q < T_\sigma$. The rates between these states can be found analogously to the rates given in Table 1.

For the busy period we can derive that

$$\begin{aligned}\pi(-1) &= \frac{\sum_{\sigma=1}^{S} N_\sigma \mu_\sigma (1-\alpha^{T_\sigma})}{\lambda + \sum_{\sigma=1}^{S} N_\sigma \mu_\sigma (1-\alpha^{T_\sigma})},\\ \pi(n) &= \alpha^n (1-\alpha)(1-\pi(-1)),\end{aligned} \tag{11}$$

for $n = 0, 1, \ldots$, with α equal to the unique solution in (0, 1) of the equation

$$\lambda(1-\alpha) = \alpha \sum_{\sigma=1}^{S} N_1 \mu_1 (1-\alpha^{B_\sigma}), \tag{12}$$

for $\lambda < \sum_{\sigma=1}^{S} N_\sigma \mu_\sigma B_\sigma$. We then find, similarly to the derivation in the case $S = 2$, that

$$E(T_{\mathrm{BP}}) = \frac{1}{\sum_{\sigma=1}^{S} N_\sigma \mu_\sigma (1-\alpha^{T_\sigma})}. \tag{13}$$

For the transition rates used in the analysis of the FP, we need to know the probability that a busy period ends with the end of service of a type σ server, finding $n_Q < T_\sigma$ customers in the queue, given by

$$P(\text{type } \sigma \text{ server}, n_Q) = \frac{N_\sigma \mu_\sigma \alpha^{n_Q} (1-\alpha)}{\sum_{\sigma=1}^{S} N_\sigma \mu_\sigma (1-\alpha^{T_\sigma})}. \tag{14}$$

For the free period we describe the state space by

$$\begin{aligned}&\big\{(n_1, \ldots, n_S, n_Q)\,\big|n_1 = 0, \ldots, N_1 - 1, n_\sigma = 0, \ldots, N_\sigma, n_Q = 0, \ldots, T_1 - 1\big\}\\ &\quad \cup \big\{(N_1, n_2, \ldots, n_S, n_Q)\,|n_2 = 0, \ldots, N_2 - 1, n_\sigma = 0, \ldots, N_\sigma,\\ &\qquad\qquad\qquad\qquad n_Q = 0, \ldots, T_2 - 1\big\}\\ &\quad \vdots\\ &\quad \cup \big\{(N_1, \ldots, N_{S-1}, n_S, n_Q)\,\big|n_S = 0, \ldots, N_2 - 1, n_Q = 0, \ldots, T_S - 1\big\}\\ &\quad \cup \{(N_1, \ldots, N_S, 0)\}.\end{aligned}$$

The transition rates from the extra state $(N_1, \ldots, N_S, 0)$ are again proportional to the exit probabilities from the busy period given in Eq. (14). Again, similarly to the previous section, we have

$$E(T_{\mathrm{FP}}) = \frac{1-\pi(N_1, \ldots, N_S, 0)}{\pi(N_1, \ldots, N_S, 0) \sum_{\sigma=1}^{S} N_\sigma \mu_\sigma (1-\alpha^{T_\sigma})}.$$

Finally, we mention some relations for $W^{(\sigma)}_{n,\ell}$, the remaining waiting time of a customer who has position n in the queue, with ℓ customers behind him and the lowest type of free servers σ. In this case the remaining waiting time might be influenced by arriving customers or by the end of a higher priority type of service. We find that

$$W^{(\sigma)}_{n,\ell} = \begin{cases} 0, & n+\ell \geq T_\sigma, \\ \text{Exp}\left(\lambda + \sum_{\tau=1}^{\sigma-1} N_\tau \mu_\tau\right) + W^{(\sigma)}_{n,\ell+1} 1_{\{A\}} & \\ \quad + \sum_{\tau=1}^{\sigma} \left(W^{(\tau)}_{n,\ell} 1_{\{F_\tau, n \leq B_\tau\}} + W^{(\sigma)}_{n-B_\tau,\ell} 1_{\{F_\tau, n > B_\tau\}} \right), & n+\ell < T_\sigma. \end{cases}$$

If no servers are free, we get

$$W_{n,\ell} = \text{Exp}\left(\lambda + \sum_{\sigma=1}^{S} N_\sigma \mu_\sigma\right) + W_{n,\ell+1} 1_{\{A\}} + \sum_{\sigma=1}^{S} \left(W^{(\sigma)}_{n,\ell} 1_{\{F_\sigma, n \leq B_\sigma\}} + W_{n-B_\sigma,\ell} 1_{\{F_\sigma, n \leq B_\sigma\}} \right).$$

Eventually we find that the waiting time in this system has a phase type distribution, as stated in the following theorem:

Theorem 2 *The waiting time in the $M_\lambda / \sum_{\sigma=1}^{S} M^{T_\sigma, B_\sigma}_{\mu_\sigma} / \sum_{\sigma=1}^{S} N_\sigma$ queue can be considered as the absorption time of a Markov chain with state space*

$$S = \{0, (\sigma, m, \ell_\sigma), (0, \ell_\sigma) | m = 1, \ldots, T_\sigma - \ell_\sigma, \ell_\sigma = 0, \ldots, T_\sigma - 1, \sigma = 1, \ldots, S\},$$

where '0' is the absorbing state, with initial probabilities

$$P_I(\sigma, m, 0) = P(FP) \sum_{n_\sigma=0}^{N_\sigma - 1} \sum_{n_{\sigma+1}=0}^{N_{\sigma+1}} \cdots \sum_{n_S=0}^{N_S} \pi_{FP}(N_1, \ldots, N_{\sigma-1}, n_\sigma, \ldots, n_S, m-1),$$
$$\textit{for } m = 1, \ldots, T_\sigma - 1 \textit{ and } \sigma = 1, \ldots, S,$$

$$P_I(0,0) = P(BP),$$
$$P_I(0) = 1 - \sum_{\sigma=1}^{S} \sum_{m=1}^{T_\sigma - 1} P_I(\sigma, m, 0) - P_I(0,0),$$

and transition rates

From	To	Transition rate
(σ, m, ℓ)	$(\sigma, m, \ell+1)$	$1_{\{m+\ell<T_\sigma-1\}}\lambda,$
	(σ_1, m, ℓ)	$1_{\{m+\ell<T_{\sigma_1}\}}N_{\sigma_1}\mu_{\sigma_1},$
	$(\sigma_1, m-B_{\sigma_1}, \ell)$	$1_{\{m>B_{\sigma_1}\}}N_{\sigma_1}\mu_{\sigma_1},$
	0	$\sum_{\tau=1}^{\sigma-1} 1_{\{m+\ell\geq T_\tau\}}1_{\{m\leq B_\tau\}}N_\tau\mu_\tau + 1_{\{m+\ell=T_\sigma-1\}}\lambda,$
$(0, \ell)$	$(0, \ell+1)$	$1_{\{\ell<T_\sigma-1\}}\lambda,$
	(σ, m, ℓ)	$\alpha^{m-1}(1-\alpha)N_1\mu_1,$ for $m = 1, \ldots, T_\sigma - 1 - \ell,$
	0	$\sum_{\sigma=1}^{S}\left(\alpha^{[T_\sigma-\ell-1]^+} - \alpha^{B_\sigma}\right)N_\sigma\mu_\sigma,$

for $\sigma_1 = 1, \ldots, \sigma - 1$ *and* $\sigma = 1, \ldots, S.$

Remark 1 We can also prove that, for an arbitrary customer, the sojourn time in the system has a phase type distribution by adding extra states ‘1’,. . .,‘S’ to the state space of the previous theorem, where ‘σ’ indicates that the tagged customer is being served by a type σ server. Then we split the rate at which the original Markov chain enters the absorbing states over these new states. Furthermore, $P_I(0)$, the initial probability that the waiting time is 0, splits over the states ‘σ’ as

$$P_I(‘\sigma’) = P(\text{FP}) \sum_{n_\sigma=0}^{N_\sigma-1} \sum_{n_{\sigma+1}=0}^{N_{\sigma+1}} \cdots \sum_{n_S=0}^{N_S} \pi_{\text{FP}}(N_1, \ldots, N_{\sigma-1}, n_\sigma, \ldots, n_S, T_\sigma - 1).$$

From state ‘σ’, the rate to the absorbing state ‘0’ is μ_σ. The absorption time of this newly defined Markov chain has the same distribution as the sojourn time of an arbitrary customer.

5 Numerical experiments

To give some results of the method presented in this paper, we consider the system depicted in Fig. 3. In this system, there are two types of servers, say type A and B. The number of type A servers is $N_A = 4$, with processing rate $\mu_A = 0.4$ and capacity $B_A = 5$; the number of type B servers is $N_B = 2$, with characteristics $\mu_B = 0.2$ and $B_B = 9$. The arrival rate $\lambda = 6$. In Table 2 we provide the expected queue length for different threshold values. Note that the server type with the lowest threshold has priority. When the thresholds of both server types are equal, we have a choice to decide which server has priority over the other. In Table 3, we provide the probability of a zero waiting time in the queue. For three combinations of T_A and T_B, we show the graphs of the probability density function and the cumulative distribution function (cdf) of the waiting times in Figs. 4 and 5, respectively. Note the jump of the cdf at $t = 0$ with height the probability of a zero waiting time in the queue.

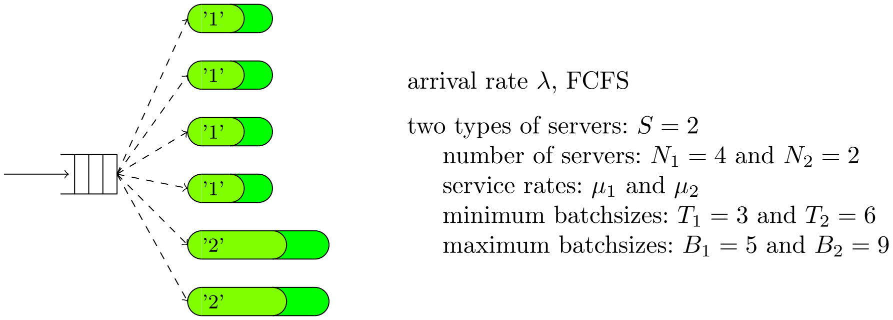
'1'
'1'
'1'
'1'
'2'
'2'
arrival rate λ, FCFS
two types of servers: $S = 2$
number of servers: $N_1 = 4$ and $N_2 = 2$
service rates: μ_1 and μ_2
minimum batchsizes: $T_1 = 3$ and $T_2 = 6$
maximum batchsizes: $B_1 = 5$ and $B_2 = 9$

Fig. 3 The $M_\lambda/M_{\mu_1}^{3,5} + M_{\mu_2}^{6,9}/4 + 2$ Queue

Table 2 The expected queue length in the $M_6/M_{0.4}^{T_A,5} + M_{0.2}^{T_B,9}/4 + 2$. In the upper right corner, type B servers have priority over type A servers

T_A T_B	1	2	3	4	5
1	4.072 4.074	3.604	3.066	2.657	2.506
2	3.953	3.430 3.454	2.928	2.548	2.426
3	3.791	3.258	2.692 2.785	2.445	2.360
4	3.598	3.066	2.526	2.236 2.389	2.337
5	3.391	2.875	2.378	2.148	2.247 2.382
6	3.193	2.707	2.266	2.094	2.230
7	3.074	2.625	2.234	2.099	2.248
8	3.031	2.622	2.266	2.147	2.291
9	3.062	2.682	2.348	2.225	2.350

Table 3 The probability of a zero waiting time in the $M_6/M_{0.4}^{T_A,5} + M_{0.2}^{T_B,9}/4+2$. In the upper right corner, type B servers have priority over type A servers

T_A T_B	1	2	3	4	5
1	0.076 0.077	0.124	0.155	0.165	0.157
2	0.086	0.143 0.146	0.171	0.178	0.168
3	0.094	0.153	0.186 0.194	0.188	0.176
4	0.100	0.157	0.194	0.195 0.205	0.180
5	0.103	0.158	0.193	0.201	0.182 0.188
6	0.103	0.157	0.190	0.196	0.184
7	0.100	0.153	0.185	0.191	0.180
8	0.095	0.146	0.178	0.186	0.176
9	0.089	0.139	0.171	0.180	0.173

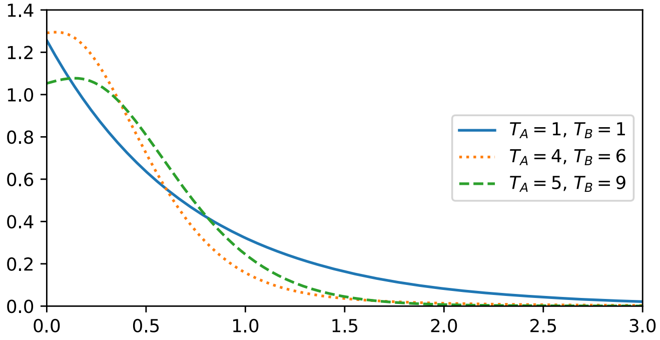

Fig. 4 The probability density functions

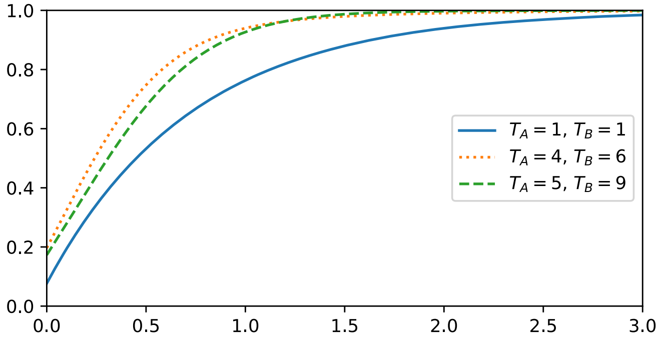

Fig. 5 The cumulative probability functions

To obtain the results, we carried out several experiments. For low and high arrival rates, both the expected queue length and the zero waiting time probability were monotone with increase in threshold. In the case where $\lambda = 6$, we see that there is a threshold setting that minimizes the expected queue length ($T_A = 4$ and $T_B = 6$) and another setting that maximizes the probability of zero waiting time ($T_A = T_B = 4$, type A servers have priority).

6 Conclusion

In this paper, we have shown that the waiting time in a queue with Poisson arrivals and exponential servers of different types has a phase type distribution. To show this result, we split the state space into two parts, the BP and FP, and analyzed these parts separately. This splitting works so well since entering the BP is always at the same state. To find the waiting time distribution, we have to analyze the FP numerically. A possible numerical problem may arise since there are many states corresponding to the FP (in the order of $T_1 * \prod_{\sigma=1}^{S} N_\sigma$ states). Future work could include numerical investigation of the threshold quantity for batch service that can trade-off waiting time vs. used resource capacity. We also hope that this study will find applications in analysis of container terminal systems where there are different types of vehicles for internal container transport, and container handling responsiveness is a key performance measure for the terminal.

References

1. Aalto, S.: Optimal control of batch service queues with finite service capacity and linear holding costs. Math. Methods Oper. Res. **51**(2), 263–285 (2000)
2. Ammar, S.I.: Transient behavior of a two-processor heterogeneous system with catastrophes, server failures and repairs. Appl. Math. Model. **38**(7), 2224–2234 (2014)
3. Arora, K.L.: Two-server bulk service queueing process. Oper. Res. **12**(2), 286–294 (1964). https://doi.org/10.1287/opre.12.2.286
4. Chakka, R., Van Do, T.: The MM$\sum_{k=1}^{K}$ CPP$_k$/GE/c/L G-queue with heterogeneous servers: steady state solution and an application to performance evaluation. Perform. Eval. **64**(3), 191–209 (2007)
5. Chang, Jin-Fu, Harn, Ywh-Pyng: A discrete-time priority queue with two-class customers and bulk services. Queueing Syst. **10**(3), 185–211 (1992)
6. Chen, A., Pollett, P., Li, J., Zhang, H.: Markovian bulk-arrival and bulk-service queues with state-dependent control. Queueing Syst. **64**(3), 267–304 (2010)
7. Gold, H., Tran-Gia, P.: Performance analysis of a batch service queue arising out of manufacturing system modelling. Queueing Syst. **14**(3), 413–426 (1993)
8. Goswami, V., Samanta, S.K.: Discrete-time bulk-service queue with two heterogeneous servers. Comput. Ind. Eng. **56**(4), 1348–1356 (2009)
9. Keaogile, T., Fatai Adewole, A., Ramasamy, S.: Geo (λ)/ Geo (μ) +G/2 queues with heterogeneous servers operating under fcfs queue discipline. Am. J. Appl. Math. Stat. **3**(2), 54–58 (2015)
10. Krishna Kumar, B., Pavai Madheswari, S.: An M/M/2 queueing system with heterogeneous servers and multiple vacations. Math. Comput. Model. **41**(13), 1415–1429 (2005)
11. Mishra, N., Roy, D., van Ommeren, Jan-Kees: A stochastic model for interterminal container transportation. Transp. Sci. **51**(1), 67–87 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.