Thailand Statistician
July 2022; 20(3): 636-654
http://statassoc.or.th
Contributed paper

A Two-Server Bulk Service Queuing Model with a Permanent Server and a Temporary Server on Hold

Kuntal Bakuli* [a] and Manisha Pal [b]
[a] Department of Statistics, Banwarilal Bhalotia College, Asansol, West Bengal, India.
[b] Department of Statistics, University of Calcutta, Kolkata, West Bengal, India.
*Corresponding author; e-mail: kuntalbakuli17@gmail.com, kuntal.statistics1986@gmail.com

Received: 28 March 2020
Revised: 17 December 2020
Accepted: 30 May 2021

Abstract

In this paper, we consider a bulk service queuing model with a fixed bulk size and a single permanent server. An additional server is kept on hold and is allowed to serve when the queue length exceeds certain threshold value. The model is analyzed using embedded Markov chain. A comparison of the performance of the model with the following models have also been made – (i) two-server bulk service model, (ii) bulk service model with two independent queues corresponding to two servers and (iii) a single-server bulk service model with double service capacity of the server.

Keywords: Bulk service queue, two servers, permanent server, temporary server.

1. Introduction

Bulk queues are common phenomena in real life. These are observed in telecommunication, transportation, production, airline scheduling, elevators, restaurants, etc. The first bulk service model was studied by Bailey (1954). Over the past few decades, much study has been carried out on bulk service systems, see, for example, Neuts (1967), Jaiswal (1960), Holman et al. (1981), Chaudhry and Templeton (1981), Abolnikov and Dshalalow (1992), Chaudhry and Gupta (1999), Gupta and Goswami (2002), Krishnamoorthy and Ushakumari (2000), Chaudhry and Chang (2004), Banerjee et al. (2014), Banerjee et al. (2015), to name a few. These authors have, however, studied single server bulk queue models.

The two-server bulk service model was introduced by Arora (1964). He studied a queuing system with two servers with single queue, the maximum serving capacities of the servers are different and the service rates of the servers are also different. He considered that the customers are arriving in a Poisson manner and exponential distribution for the distribution of service time for the servers. In his queuing model if any of the server is in idle state then it starts service as soon as it gets at least one customer to serve. He analyzed the model using differential equations and Laplace transformations. The results of Arora (1964) were generalized to the case of multiple servers by Ghare (1968), using standard generating functions and Laplace transform methods. Much later, Ghimre et al. (2017) studied a multiple server model where the server capacity is variable and it depends on the number of customer available in the system. The service rate of any of the servers depends on the number customers are being served by the server. In their queuing model all the servers in a similar manner, all the servers have equal maximum serving capacity and if any of the servers is in idle state then it starts service as soon as it gets at least one customer to serve. After these contributions, the area was not much explored.

In this paper, we consider a two-server bulk service model, where one of the servers is a permanent one while the other is kept on hold and allowed to function when the queue length exceeds certain threshold value and the permanent server is busy. It is assumed that the service capacity of both the servers are equal. The model is analyzed using embedded Markov chain technique. The model has been compared in terms of the performance measures with other comparable queuing models, like (a) the two-server bulk service model where the servers are working in parallel, (b) two independent single-server bulk service models, where an arriving customer randomly joins a queue, and (c) a single-server bulk service model with double capacity of the server. A new performance measure, namely, average number of customers served per unit time, has been introduced. The models have also been compared in terms of the expected expenditure for running the service centre. The paper is organized as follows. Section 2 discusses the model and its analysis is described in Section 3. Section 4 computes the performance measures of the model for different sets of values of the model parameters. In Section 5, the comparable models are analyzed using embedded Markov chain technique, and a numerical comparison of the performance measures of all the models are carried out. In Section 6, we have briefly discussed about the outcome of the study. Section 7 is the conclusion part, where we have explained the usefulness of the study along with the further scope of the study.

2. The Model and its Assumptions

We consider a $M / M^b / 1$ queuing model with a difference. Apart from the permanent server, which we shall call server 1, there is a temporary server – server 2 - kept on hold, and is called for service whenever the queue length reaches a certain specified threshold value, say q and server 1 is busy. The temporary server is called back when the queue falls below the threshold value. Server 1 follows the general bulk service rule, that is, when it finds at least b customers waiting in the line it starts service with the first b customers in the queue. Customers wait in a single queue, and when server 2 comes into action, the two servers work in parallel, each serving b customers at a time.

The assumptions governing the model are as follows:

1) Customers arrive one at a time in a Poisson fashion with mean arrival rate λ.
2) Each server serves b customers at a time.
3) The service time distribution of each server is exponential with mean service rate μ, and these are independent of one another.
4) The system is of infinite buffer.
5) The queue discipline is FIFO.
6) $\lambda < 2b\mu.$

The threshold of the queue length that brings server 2 into action is q.

3. Analysis of the Model

To analyze the model, let us define $Y_1(t)$ as the state of the servers at time point t, where

$Y_1(t)$ = 0, if both the servers are idle, $Y_1(t)$ = 1, if server 1 is busy and server 2 is idle, $Y_1(t)$ = 2, if server 1 is idle and server 2 is busy, $Y_1(t)$ = 3, if both the servers are busy.

We shall consider that the transitions due to service completion and customer arrival only affect the number of busy servers. Hence, a transition will depend on the number of customers waiting in the queue.

Let $Y_2(t)$ denotes the queue length at time point t. Then, $Y(t)=\left(Y_1(t),Y_2(t)\right)$ is a semi-Markov process. We attempt to obtain the steady state distribution of $Y_1(t)$, which will help to compute the performance measures of the system.

Let $\{t_n\}$ be the sequence of epochs at which service completion of any server occurs or any server starts working from idle state due arrival of customer. Then, $Y_n=Y(t_n^+)$ is the embedded Markov chain defined on the state space

$$S=\left\{\left(0,y_2\right):y_2\in\left(0,1,\ldots,b-1\right)\right\}\cup\left\{\left(1,y_2\right):y_2\in\left(0,1,\ldots,q-1\right)\right\}\cup\{\left(2,y_2\right):y_2\in$$
$$\left(0,1,\ldots,b-1\right)\}\cup\left\{\left(3,y_2\right):y_2\in\left(0,1,2,3,..\right)\right\}.$$

Writing $P\left(\boldsymbol{i},\boldsymbol{j}\right)=P[Y_{n+1}=\boldsymbol{j}\,|\,Y_n=\boldsymbol{i}]$, where $\boldsymbol{j}$=$\left(j_1,j_2\right)$ and $i=\left(i_1,i_2\right)$ the transition probabilities are given by

$$\begin{aligned}
P\left(\boldsymbol{i},\boldsymbol{j}\right)&=1, && \text{if } i_1=0,i_2<b \text{ and } j_1=1,j_2=0,\\
&=p^{j_2-i_2}\left(1-p\right), && \text{if } i_1=1,i_2<b,j_1=0,i_2\le j_2<b \text{ or } i_1=2,i_2<b,j_1=0,i_2\le j_2<b,\\
&=p^{j_2+b-i_2}\left(1-p\right), && \text{if } i_1=1,i_2<q \text{ and } j_1=1,i_2\le j_2<q-b,\\
&=p^{q-i_2}, && \text{if } i_1=1,i_2<q \text{ and } j_1=3,j_2=q-b,\\
&=p^{b-i_2}, && \text{if } i_1=2,i_2<b \text{ and } j_1=3,j_2=0,\\
&=\frac{1}{2}p_1^{j_2-i_2}\left(1-p_1\right), && \text{if } i_1=3,i_2<b \text{ and } j_1=2,i_2\le j_2<b, \text{ or } i_1=3,i_2<q \text{ and}\\
& && j_1=1,i_2\le j_2<q,\\
&=\frac{1}{2}p_1^{j_2+b-i_2}\left(1-p_1\right), && \text{if } i_1=3,i_2<q \text{ and } j_1=3,i_2\le j_2+b<q,\\
&=p_1^{j_2+b-i_2}\left(1-p_1\right), && \text{if } i_1=3,i_2\ge 0,j_1=3 \text{ and } j_2+b\ge q,
\end{aligned}$$

where

$$p=\frac{\lambda}{\lambda+\mu} \text{ and } p_1=\frac{\lambda}{\lambda+2\mu}. \tag{1}$$

(See Appendix)

Theorem 3.1. *The embedded Markov chain* $\{Y_n\}$ *is ergodic.*

Proof: Let us define a one-to-one onto function $f:S\to N$, the set of natural numbers, such that $f\left(x,y\right)\le f\left(3,q\right)$ for all $x\le 3,y\le q$ and $f\left(3,q+k+1\right)-f\left(3,q+k\right)=1$ for all $k\in\{0,1,2,...\}$.

Let $f\left(x,y\right)=4y+\left(x+1\right)$, when $y<b$,

$$f\left(x,y\right)=4b+2\left(y-b\right)+\frac{x+1}{2},\text{ when } q>y>\left(b-1\right),$$
$$f(x,y)=4b+2\left(q-b\right)+\left(y-q+1\right),\text{ when } y>q-1.$$

If S^* be the range of f, we can define a (dummy) Markov chain $\{X_n\}$ over S^* with transition probabilities $P'(i',j')=P(f^{-1}(i'),f^{-1}(j'))$.

This Markov chain is aperiodic since for sufficiently large value of i' (when queue is sufficiently large) $P'(i',i')>0$ for some i'. Now, we shall show that $\{X_n\}$ is ergodic, which will imply that the embedded Markov chain $\{Y_n\}$ is also ergodic. To do so, we use the method suggested by Pakes (1969).

Define $\gamma_i=E(X_{n+1}-X_n|X_n=i)$, which is a sum of finite number of terms when $f^{-1}(i)=(0,k)$ or $f^{-1}(i)=(1,k)$ or $f^{-1}(i)=(2,k)$. If $f^{-1}(i)=(3,k)$ then $\gamma_i=E(X_{n+1}-X_n|X_n=i)$ is a sum of infinite number of terms. Now we will consider this situation only.

Let $i<2q+3b+1$, (here $f\left(3,q+b\right)=2q+3b+1$) then $k<q+b$. Then,

$$\gamma_i=\sum_{j\in f\left(3,l\right):l<q+b}\left(j-i\right)P\left(i,j\right)+\sum_{j\in f\left(3,l\right):l\geq q+b}\left(j-i\right)P\left(i,j\right)=sum_1+sum_2,$$

where $\sum_{j\in f\left(3,l\right):l<q+b}\left(j-i\right)P\left(i,j\right)=sum_1$ and $\sum_{j\in f\left(3,l\right):l\geq q+b}\left(j-i\right)P\left(i,j\right)=sum_2$.

Here, sum_1 is the sum of a finite number of terms, while sum_2 is the sum of an infinite number of terms. Further,

$$sum_2\leq\sum_{l\geq q+b}\left(2b+q+l-3\right)p_1^{l+b-k}\left(1-p_1\right)<\infty\text{ (as }\sum_{l\geq q+b}l\,p_1^{\,l}<\infty\text{)}.$$

Thus, sum_2 also has a finite value, so that $\gamma_i<\infty$, where $i<2q+3b+1$. When $i\geq 2q+3b+1$, we have $k\geq q+b$. In this situation, if $X_{n+1}=j$ and $X_n=i$, then $\left(j-i\right)=\left(l-k\right)$ where, $f^{-1}(i)=(3,k)$ and $f^{-1}(j)=(3,l)$. Therefore, $\gamma_i=\sum_{l\geq k-b}\left(l-k\right)p_1^{l+b-k}\left(1-p_1\right)=\frac{p_1}{1-p_1}-b.$

By Pakes (1969), it follows that $\{X_n\}$ is ergodic if $\lim_{i\to\infty}\gamma_i$ is negative. Now, since, $\lambda<2b\mu$, we have $\left(\frac{p_1}{1-p_1}-b\right)<0$. Hence, $\lim_{i\to\infty}\gamma_i$ is negative. Thus, by the theorem, $\{X_n\}$, and therefore $\{Y_n\}$, is ergodic.

3.1. Stationary distribution of {*Y*n}

To obtain the stationary distribution of $\{Y_n\}$, we first find the stationary distribution of $\{X_n\}$ using north-west corner truncation method suggested by Seneta (1968) and Wolf (1980). Consider a special sequence of transition probability matrix (TPM) $\{P_m, m>0\}$, constructed from the original infinite TPM of $\{X_n\}$, as follows:

$$P_m\left(i,j\right)=0,\quad\text{if } i,j\geq m+1,$$
$$P_m(i,j)=p(i,0)+\sum_{j'>m}p(i,j'),\quad\text{if } i<m+1 \text{ and } j=0,$$
$$p_m(i,j)=p(i,j),\quad\text{if } i<m+1 \text{ and } 0<j<m+1.$$

Then, P_m has exactly one stationary distribution, $\boldsymbol{\pi}^m$, and it converges to $\boldsymbol{\pi}$ (cf. Wolf 1980). Define $\pi(k)^{(m)}$ as the steady state probability of state " k " taking " m " as the order of the truncated TPM.

Let $f^{-1}(k)=(i_1,i_2)$. Then, $\lim\limits_{m\to\infty}\pi(i_1,i_2)^{(m)}=\pi(i_1,i_2)$, where $\pi(i_1,i_2)$ denotes the original steady state probability of state (i_1,i_2) of the embedded Markov chain $\{Y_n\}$.

The unconditional expected waiting times under different states of the system are given in Table 1.

Table 1 The unconditional expected waiting times under different states of the system

State	Waiting time	Unconditional expected waiting time
$i_1=0, i_2<b$	$A(b-i_2)$	$A(b-i_2)/\lambda$
$i_1=1, i_2<q$	$\min(T_1, A(q-i_2))$	$(1/\mu)[1-p^{q-i_2}]$
$i_1=2, i_2<b$	$\min(T_1, A(b-i_2))$	$(1/\mu)[1-p^{b-i_2}]$
$i_1=3, i_2\in\{0,1,2,...\}$	$\min(T_1,T_2)$	$1/2\mu$

In Table 1, $A(n)\sim gamma(\lambda,n)$ (sum of " n " independent $\exp(\lambda)$), and $T_i\sim\exp(\mu)$, $i=1,2.$

3.2. Steady state distribution of the semi-Markov process *Y*(*t*)

Let $\lim\limits_{t\to\infty}P[Y(t)=(i_1,i_2)]=v_{(i_1,i_2)}$. Then $v^{(m)}_{(i_1,i_2)}=M_{(i_1,i_2)}\pi(i_1,i_2)^{(m)}\Big/\sum_{k,l}M_{k,l}\pi(k,l)^{(m)}$ will converge to $v_{(i_1,i_2)}$ as $m\to\infty$, where $M_{k,l}$ is the unconditional expected waiting time at state (k,l). We obtain the values of $v_{(i_1,i_2)}$ by numerical computation (performing a finite sum for very large values of " m ").

3.3. Performance measures

Along with the usual performance measures like average queue length, average waiting time in queue, probability of busy period of server 1 and server 2 here we have introduced a new performance measure, namely average number of customers served per unit time:

(i) Average queue length is $\sum\limits_{(i_1,i_2)\in S} v^{(m)}_{(i_1,i_2)}\times i_2$,

(ii) Average waiting time in queue is $(\text{Average queue length})/\lambda$ (By Little's formula),

(iii) Probability that server 1 is busy:

$$P_{11}=P[Y_1=1,3]=\sum_{(i_1,i_2)\in S} v^{(m)}_{(i_1,i_2)}\times I_{\{(i_1,i_2):i_1=1,3\}},$$

Probability that the server 2 is busy:

$$P_{12}=P[Y_1=2,3]=\sum_{(i_1,i_2)\in S} v^{(m)}_{(i_1,i_2)}\times I_{\{(i_1,i_2):i_1=2,3\}},$$

(iv) To get the average number of customers served per unit time by the system, we proceed as follows. Let $N_1(t)$ be a number of service completions by server 1 in $(0,t]$, and $N_2(t)$ be a number of service completions by server 2 in $(0,t]$. As $N_1(t)$ and $N_2(t)$ are both increasing in " t ",

$$\lim_{t\to\infty} N_i(t) = \infty, \ \ i = 1,2.$$

In the steady state, the average number of service completions in unit time by i^{th} server is $\lim\limits_{t\to\infty}\frac{N_i(t)}{t}$. Therefore, in steady state, the average number of customers served in unit time by the i^{th} server is $\lim\limits_{t\to\infty}\frac{N_i(t)}{t}\times b.$ Hence, in steady state, the average number of customer served per unit time is $\lim\limits_{t\to\infty}\frac{N_1(t)+N_2(t)}{t}\times b.$ Now,

$$P_{1i} = P(i^{\text{th}} \text{ server is busy}) = \lim_{t\to\infty}\frac{\sum_{j=1}^{N_i(t)} T_j}{t},$$

where T_j is the service time of server $j,\ j = 1,2.$

For large value of "t", the total busy period of i^{th} server is approximately given by $\sum\limits_{j=1}^{N_i(\text{t})} T_j$. This is the total time taken by the i^{th} server to complete $N_i(t)$ services. Hence,

$$P_{1i} = \lim_{t\to\infty}\frac{\sum_{j=1}^{N_i(t)} T_j}{t} = \left(\lim_{t\to\infty}\frac{\sum_{i=1}^{N_i(t)} T_i}{N_i(t)}\right)\times\left(\lim_{t\to\infty}\frac{N_i(t)}{t}\right) = \left(\lim_{N_i(t)\to\infty}\frac{\sum_{j=1}^{N_i(t)} T_j}{N_i(t)}\right)\times\left(\lim_{t\to\infty}\frac{N_i(t)}{t}\right)$$

$$= \frac{1}{\mu}\times\left(\lim_{t\to\infty}\frac{N_i(t)}{t}\right).$$

Since by strong law of large number, we can say that $\lim\limits_{t\to\infty}\frac{\sum_{i=1}^{N_i(t)} T_i}{N_i(t)} = \frac{1}{\mu}$, as $T_j \sim \exp(\mu),\ \ j = 1,2.$

Thus, $\lim\limits_{t\to\infty}\frac{N_i(t)}{t} = \mu\times P_{1i}$. In steady state, we therefore obtain the average number of customers served in unit time by the system as

$$\lim_{t\to\infty}\frac{N_1(t)+N_2(t)}{t}\times b = \mu\times b\times\left(P_{11}+P_{12}\right).$$

4. Computation

Consider $m = 500$ in north-west corner truncation. It is noted that for higher values of "m", the values of performance measures do not change up to the 6^{th} place of decimal. Tables 2 and 3 give the performance measures for different sets of values of the system parameters and the threshold "q".

Table 2 Performance measures of the model for some sets of values of the model parameters when $q = 2b$

λ	μ	b	$q = 2b$				
			P_{11}	P_{12}	Av. no. of customers served per unit time	Av. waiting time	Av. queue length
4.5	0.5	5	0.97048029	0.829520	4.5	4.9735	22.38062
4	0.5	5	0.93541071	0.664589	4	2.1557	8.622634
3.5	0.5	5	0.89272424	0.507276	3.5	1.3193	4.617509
3	0.5	5	0.83922155	0.360778	3	0.9567	2.878086
4.5	0.33	8	0.95446569	0.733034	4.5	4.1375	18.61892
4	0.33	8	0.91989116	0.580109	4	2.3576	9.430495
3.5	0.33	8	0.87729253	0.435207	3.5	1.6525	5.783895
3	0.33	8	0.82321514	0.301785	3	1.3056	3.916705
4.5	0.25	10	0.97296219	0.827019	4.499	9.1488	41.16968
4	0.25	10	0.94045906	0.659541	4	4.0283	16.11337
3.5	0.25	10	0.90030427	0.499696	3.5	2.5240	8.834105
3	0.25	10	0.76619231	0.433808	3	1.8844	5.653127

Table 3 Performance measures of the model for some sets of values of the model parameters when $q = 3b$

λ	μ	b	$q = 3b$				
			P_{11}	P_{12}	Av. no. of customers served per unit time	Av. waiting time	Av. queue length
4.5	0.5	5	0.988424	0.811576	4.5	5.875967	26.44185
4	0.5	5	0.970825	0.629175	4	3.057840	12.23136
3.5	0.5	5	0.943609	0.456391	3.5	2.189074	7.661758
3	0.5	5	0.900878	0.299122	3	1.747526	5.242579
4.5	0.33	8	0.981580	0.705920	4.5	5.519939	24.83972
4	0.33	8	0.962451	0.537549	4	3.721832	14.88733
3.5	0.33	8	0.932721	0.379779	3.5	2.942220	10.29777
3	0.33	8	0.885972	0.239028	3	2.438167	7.314501
4.5	0.25	10	0.990125	0.809848	4.5	10.98840	49.44781
4	0.25	10	0.974629	0.625371	4	5.879704	23.51882
3.5	0.25	10	0.949830	0.450170	3.5	4.314797	15.10179
3	0.25	10	0.909431	0.290569	3	3.505507	10.51652

It may be noted from Tables 2 and 3 that in the steady-state situation the average number of customers served per unit time comes out to be equal to the average number of arrivals per unit time. This led to Theorem 4.2 which shows that the above is true whatever be system parameters.

Theorem 4.1. *For the queuing model considered, the average number of customers served per unit time equals the average number of arrivals per unit time.*

Proof: Define $\tau_i^{(j)}$ is time to return to the state " j " from state " j " for the i^{th} time. Assume that the system starts from the state "(0,0)". Since "(0,0)" is a positive recurrent state, in the steady state situation, the system will return to state "(0,0)" after a finite time with probability 1. Therefore, $\tau_i^{(0,0)}$, $i=1(1)n$ are independently and identically distributed random variables with finite expectation say, θ.

Let $\gamma_n^{((0,0))}=\sum_{i=1}^{n}\tau_i^{((0,0))}$. Then, $\gamma_n^{((0,0))}\to\infty$ as $n\to\infty$. Now let, the number of customer served by time $\left(0,\gamma_n^{((0,0))}\right]$ be denoted by $D\left(\gamma_n^{((0,0))}\right)$, and the number of customers arriving during this interval be $B\left(\gamma_n^{((0,0))}\right)$. Then, it is obvious that $D\left(\gamma_n^{((0,0))}\right)=B\left(\gamma_n^{((0,0))}\right)$. So, the number of customer served per unit time is $\lim_{n\to\infty}\frac{D\left(\gamma_n^{((0,0))}\right)}{\gamma_n^{((0,0))}}=\lim_{n\to\infty}\frac{B\left(\gamma_n^{((0,0))}\right)}{\gamma_n^{((0,0))}}$.

As the arrival process is Poisson, we have $B\left(\gamma_n^{((0,0))}\right)=\sum_{i=1}^{n}B\left(\tau_i^{((0,0))}\right)$. Then,

$$\lim_{n\to\infty}\frac{B\left(\gamma_n^{((0,0))}\right)}{\gamma_n^{((0,0))}}=\lim_{n\to\infty}\frac{\sum_{i=1}^{n}B\left(\tau_i^{(0,0)}\right)}{\sum_{i=1}^{n}\tau_i^{(0,0)}}=\lim_{n\to\infty}\frac{\left[\sum_{i=1}^{n}B\left(\tau_i^{(0,0)}\right)\right]/n}{\left[\sum_{i=1}^{n}\tau_i^{(0,0)}\right]/n}=\frac{E\left[B\left(\tau_i^{(0,0)}\right)\right]}{\theta},$$

by strong law of large numbers. Now, $E\left[B\left(\tau_i^{(0,0)}\right)\right]=E\left[E[B\left(\tau_i^{(0,0)}\right)|\tau_i^{(0,0)}\right]=E\left[\lambda\tau_i^{(0,0)}\right]=\theta\lambda,$ therefore, $\lim_{n\to\infty}\frac{B\left(\gamma_n^{((0,0))}\right)}{\gamma_n^{((0,0))}}=\frac{\theta\lambda}{\theta}=\lambda.$ Hence, the average number of customers served per unit time by the system is λ.

Remark: In this theorem, we have not used any specific characteristics of the mentioned queuing system. Thus, the above theorem also holds in general for any queuing system, where the system is in the steady state or is in equilibrium.

5. Comparison with Other Models

In this section, we compare the model in Section 3 with other models, like the $M/M^b/2$ model, two independent $M/M^b/1$ models with arriving customer randomly joining a queue, and the

$M\,/\,M^b\,/1$ model with high service capacity of the single server. To facilitate comparability, we first analyze the models using the embedded Markov chain technique.

5.1. *M* / *M* ^{b}/2 model

Here we make the usual assumptions that govern an $M\ /\ M^b\ /\ 2$ queuing model. Defining $Y_1(t)$ and $Y_2(t)$ as before, where $Y_1(t)$ takes the value 0, 1 or 2 according as both the servers are idle, one server is idle and both the servers are busy, respectively, we have that $Y(t)=\left(Y_1(t),Y_2(t)\right)$ is a semi-Markov process.

Defining $\{t_n\}$ to be the sequence of epochs at which service completion of any server occurs or any server starts working from idle state due to a new arrival, and $Y_n=Y(t_n^+)$ is the embedded Markov chain with state space

$$S_1=\ \left\{\{0,1,2\}\times\mathrm{N}\right\}/\left\{\{0\}\times\left\{\mathrm{N}-\{0,1,\ldots,b-1\}\right\}\cup\{1\}\times\left\{\mathrm{N}-\{0,1,\ldots,b-1\}\right\}\right\},$$

the transition probabilities are obtained as

$$\begin{aligned}
P(\boldsymbol{i},\boldsymbol{j})&=1, && \text{if } i_1=0, i_2<b \text{ and } j_1=1, j_2=0,\\
&=p^{j_2-i_2}\left(1-p\right), && \text{if } i_1=1, \text{ if } i_2<b \text{ and } j_1=0, i_2\le j_2<b,\\
&=p^{b-i_2}, && \text{if } i_1=1, \text{ if } i_2<b \text{ and } j_1=2, j_2=0,\\
&=p_1^{j_2-i_2}\left(1-p_1\right), && \text{if } i_1=2, \text{ if } i_2<b \text{ and } j_1=1, i_2\le j_2<b,\\
&=p_1^{j_2+b-i_2}\left(1-p_1\right), && \text{if } i_1=2, \text{if } i_2\ge 0, j_1=2, j_2\ge\max(0,i_2-b),
\end{aligned}$$

where $P(\boldsymbol{i},\boldsymbol{j})=P[Y_{n+1}=\boldsymbol{j}\,|\,Y_n=\boldsymbol{i}]$, $\boldsymbol{j}=(j_1,j_2)$ and $\boldsymbol{i}=(i_1,i_2)$, and p_1 and p are given by (1) (See Appendix).

Theorem 5.1.1. *The embedded Markov chain* $\{Y_n\}$ *is ergodic under the assumption* $\lambda<2b\mu$.

Proof: Consider a function $f:R^2\to R$ such that

$$f(x,y)=3y+(x+1),\ \text{when}\ y<b$$
$$f(x,y)=(y+2b)+1,\ \text{when}\ y>(b-1).$$

If S^{**} be the range of f, we can define a (dummy) Markov chain $\{X_n\}$ over S^{**} with transition probabilities $P'(i',j')=P(f^{-1}(i'),f^{-1}(j'))$. This Markov chain is aperiodic since for sufficiently large value of "i'" (when queue is sufficiently large) $P'(i',i')>0$ for some i'. Now, we shall show that $\{X_n\}$ is ergodic, which will imply that the embedded Markov chain $\{Y_n\}$ is also ergodic. To do so, we use the method suggested by Pakes (1969).

Define $\gamma_i=E(X_{n+1}-X_n|X_n=i)$, which is a sum of finite number of terms when $f^{-1}(i)=(0,k)$ or $f^{-1}(i)=(1,k)$. If $f^{-1}(i)=(2,k)$ then $\gamma_i=E(X_{n+1}-X_n|X_n=i)$ is a sum of infinite number of terms. Now we will consider this situation only.

Let $i<4b+1$, (here, $f(2,2b)=4b+1$) then $k<2b$. Then,

$$\gamma_i=\sum_{j\in f(2,l):l<2b}(j-i)P(i,j)+\sum_{j\in f(2,l):l\ge 2b}(j-i)P(i,j)=sum_1'+sum_2',$$

where $\sum_{j\in f(2,l):l<2b}(j-i)P(i,j)=sum'_1$ and $\sum_{j\in f(2,l):l\geq 2b}(j-i)P(i,j)=sum'_2$. Here, sum'_1 is the sum of a finite number of terms, while sum'_2 is the sum of an infinite number of terms. Further,

$$sum'_2\leq \sum_{j\geq 4b} jp_1^{\,j-2b}\left(1-p_1\right)<\infty \text{ (as } \sum_{l\geq q+b} l\,p_1^{\,l}<\infty\text{).}$$

Thus, sum'_2 also has a finite value, so that $\gamma_i<\infty$, where $i<4b+1$. When $i\geq 4b+1$, we have $k\geq 2b$. In this situation, if $X_{n+1}=j$ and $X_n=i$, then $(j-i)=(l-k)$ where $f^{-1}(i)=(2,k)$ and $f^{-1}(j)=(2,l)$. Therefore, $\gamma_i=\sum_{l\geq k-b}\left(l-k\right)p_1^{\,l+b-k}\left(1-p_1\right)=\frac{p_1}{1-p_1}-b.$

By Pakes (1969), it follows that $\{X_n\}$ is ergodic if $\lim_{i\to\infty}\gamma_i$ is negative. Now, since $\lambda<2b\mu$, we have $\left(\frac{p_1}{1-p_1}-b\right)<0$. Hence, $\lim_{i\to\infty}\gamma_i$ is negative. Thus, by the theorem, $\{X_n\}$ and therefore $\{Y_n\}$ is ergodic.

Table 4 The unconditional expected waiting time for different states

State	Waiting time	Unconditional expected waiting time
$i_1=0,i_2<b$	$A(b-i_2)$	$(b-i_2)/\lambda$
$i_1=1,i_2<b$	$\min(T_1,A(b-i_2))$	$(1/\mu)[1-p^{b-i_2}]$
$i_1=2,i_2\in\{0,1,2,...\}$	$\min(T_1,T_2)$	$1/2\mu$

Here, $A(n)\sim gamma(\lambda,n)$ (sum of "n" independent $\exp(\lambda)$), and $T_i\sim\exp(\mu)$, $i=1,2$. The steady state distribution of $Y(t)$ and the measures of performance of the model are obtained as before.

5.2. Bulk service model with two independent queues

Here we consider two separate queues for the two servers. Both the servers perform with usual bulk service mechanism with fixed bulk size "b". We assume that an arriving customer randomly joins a queue. As both the systems are identical, we first analyze a single server model, i.e., $M/M^b/1$ with usual assumptions.

Defining (X,t) a semi Markov process, where $X(t)$ stands for number of customers in the system at time "t", we have considered embedded Markov chain $\{X_n\}$ over the time epochs where t_n be the n^{th} epoch at which system size changes and $X_n=X(t_n^+)$. According to Bakuli and Pal (2017), this chain is ergodic when $\lambda<b\mu$. Transition probabilities are obtained as,

$$\begin{aligned}
&P(i,j)=1, && \text{if } 0\leq i<b \text{ and } j=i+1,\\
&P(i,j)=\lambda/\left(\lambda+2\mu\right), && \text{if } i\geq b \text{ and } j=i+1,\\
&P(i,j)=2\mu/\left(\lambda+2\mu\right), && \text{if } i\geq b \text{ and } j=i-b,\\
&P(i,j)=0, && \text{otherwise.}
\end{aligned}$$

We have used north-west corner truncation method as before to obtain steady state distribution of the Markov chain $\{X_n\}$.

Table 5 The unconditional expected waiting time for the different states

State	Waiting time	Unconditional expected waiting time
$i < b$	$\exp(\lambda/2)$	$2/\lambda$
$i \geq b$	$\min(\exp(\lambda/2), \exp(\mu))$	$2/(\lambda+2\mu)$

The steady state distribution of $X(t)$ and the measures of performance of the model are obtained as follows: $\lim\limits_{t\to\infty} P\left[X(t)=j\right]=v_j$, then $v_j^m = M_j\pi_j^m \big/ \sum_i M_i\pi_i^m$ will converge to v_j, where M_i is expected unconditional waiting time at state " i ", " m " is the order of the truncated probability matrix and $\boldsymbol{\pi}^m = \left(\pi_1^m, \pi_2^m, \ldots, \pi_m^m\right)$ be the stationary distribution of the truncated probability matrix of order " m ".

Performance measures:

(i) Average queue length = $2\times\left[\lim\limits_{m\to\infty}\left\{\sum\limits_{j=0}^{b-1} jv_j^m + \sum_{j\geq b}\left(j-b\right)v_j^m\right\}\right]$

(ii) Average queue length in queue = $\left(\text{Average queue length}\right)/\lambda$ (by Little's formula).

5.3. Single-server bulk service model with double capacity

Here we have considered a single-server bulk service model with double capacity, i.e., the server capacity is " $2b$ ". Server follows general bulk service rule. The assumptions are usual. Model has been analyzed as a single-server model analyzed in the previous section. We have considered a similar Markov process (X,t) and obtained steady state distribution as it is obtained in the previous section. Derivations of the performance measures are bit different here because the model in the previous section is a two-server two queue model and this is a single-server queue model.

Performance measures:

(i) Average queue length = $\lim\limits_{m\to\infty}\left[\sum\limits_{j=0}^{2b-1} jv_j^m + \sum_{j\geq 2b}\left(j-b\right)v_j^m\right]$,

(ii) Average queue length in queue = $\left(\text{Average queue length}\right)/\lambda$ (by Little's formula).

5.4. Computations of performance measures for different bulk service models

We have considered $m=500$, we checked that if we take higher value of " m " the values are not changing up to 6-th decimal places. In following two tables we have considered different values of system parameters.

Table 6 Performance measures of different bulk service models for some sets of values of the model parameters

λ	μ	b	General two-server bulk service model		Two-server model with two parallel queues		Single-server model with double capacity	
			Av. waiting time	Av. queue length	Av. waiting time	Av. queue length	Av. waiting time	Av. queue length
4.5	0.5	5	4.344	19.549	11.173	50.278	10.313	46.411
4	0.5	5	1.543	6.170	5.304	21.218	4.969	19.878
3.5	0.5	5	0.739	2.587	3.471	12.149	3.325	11.638
3	0.5	5	0.434	1.301	2.691	8.0716	2.657	7.9715
4.5	0.33	8	3.185	14.334	9.841	44.286	9.384	42.231
4	0.33	8	1.441	5.764	6.021	24.084	5.811	23.245
3.5	0.33	8	0.798	2.796	4.472	15.651	4.386	15.353
3	0.33	8	0.549	1.649	3.803	11.409	3.806	11.418
4.5	0.25	10	7.858	35.363	20.627	92.823	19.090	85.906
4	0.25	10	2.768	11.073	9.939	39.756	9.602	38.410
3.5	0.25	10	1.334	4.669	6.651	23.277	6.505	22.770
3	0.25	10	0.813	2.439	5.314	15.943	5.282	15.848

5.5. Comparison of expenditures towards running the systems

To run a system, the organization has to pay the servers and also spend money on the operation of the service facilities being used. If the servers are permanent employees of the organization, each server is paid a fixed salary, irrespective of whether or not he is in idle state. However, he may be paid an incentive when he remains busy. On the other hand, if a server be appointed on temporary basis, he may be paid as much as the permanent employee when he is called for service, and may be paid a token amount when idle. Other costs associated with running the system are the costs of operating the service facilities. To compare the four queuing models, we may consider the total cost associated with running the system per unit time. For the purpose, we consider the following cost structures for the models.

Table 7 Cost structures for the models

Cost per unit time	General two-server bulk service model	Single-server model with additional server	Two-server model with two queues	Single-server model with double capacity
Running cost of each server	200	200	200	400
Cost paid to the main server in busy period	1,000	1,000	1,000	2,000
Cost paid to the main server in idle period	800	800	800	1,600
Cost paid to the temporary server in busy period	_	1,000	_	_
Cost paid to the temporary server in idle period	_	200	_	_

Table 8 The total expenditure per unit time

Model	Expected Cost per unit time
General two-server bulk service model	$(800+800)P(Y_1=0)+(1000+200+800)P(Y_1=1)+2(1000+200)P(Y_1=2)$
Single-server model with additional server	$(800+800)P(Y_1=0)+(1000+200+200)P(Y_1=1)+(1000+200+800)P(Y_1=2)$ $+2(1000+200)P(Y_1=3)$
Two-server model with two queues	$2\times 800P(X<b)+2(1000+200)P(X>b)$
Single-server model with double capacity	$1600P(X<2b)+(2000+400)P(X>2b)$

Table 9 Expected expenditure for running the service system in the different queuing models

λ	μ	b	General two-server bulk service model	Single-server model with additional server, $q = 2b$	Single-server model with additional server, $q = 3b$	Two-server model with two queues	Single-server model with double capacity
			Cost	Cost	Cost	Cost	Cost
4.50	0.50	5.00	2,320.00	2,217.71	2,206.95	2,320.00	2,378.46
4.00	0.50	5.00	2,240.00	2,038.75	2,017.51	2,240.00	2,357.51
3.50	0.50	5.00	2,160.00	1,864.37	1,833.83	2,160.00	2,337.24
3.00	0.50	5.00	2,080.00	1,696.47	1,659.47	2,080.00	2,317.75
4.50	0.33	8.00	2,275.00	2,114.82	2,098.55	2,275.00	2,378.44
4.00	0.33	8.00	2,200.00	1,948.07	1,922.53	2,200.00	2,366.04
3.50	0.33	8.00	2,125.00	1,786.12	1,752.87	2,125.00	2,354.10
3.00	0.33	8.00	2,050.00	1,631.07	1,593.42	2,050.00	2,342.70
4.50	0.25	10.00	2,320.00	2,216.20	2,205.90	2,320.00	2,388.67
4.00	0.25	10.00	2,240.00	2,035.72	2,015.22	2,240.00	2,377.85
3.50	0.25	10.00	2,160.00	1,859.82	1,830.10	2,160.00	2,367.37
3.00	0.25	10.00	2,080.00	1,740.28	1,654.34	2,080.00	2,357.37

From Table 9, it is evident that the use of a temporary server along with a permanent server reduces the cost of operating the system as compared to the two-server bulk service model, two independent bulk service models and a single-server model with double serving capacity of the server. Thus, from the cost point of view the suggested model seems more cost effective than the other models.

6. Discussion

In this paper, we have proposed a service mechanism for a two-server bulk service queuing model and compare its performance with some comparable queuing models. We have used a computational method based on semi-Markov process to obtain the performance measures. For different values of system parameters, we have checked the average waiting time and average queue length of the proposed model and compared with those of the other models. We find that the proposed model performs better.

From the Figures 1 and 2 below, it is clear that only the two-server bulk service model $M/M^b/2$ with general bulk service rule performs slightly better than our proposed model. However, with respect to the considered cost structure, the operational cost of the $M/M^b/2$ model is higher than the proposed model. It is also noted that for the same operational cost, the two-server model with two parallel queues does not perform better than any of the two-server models with single queue.

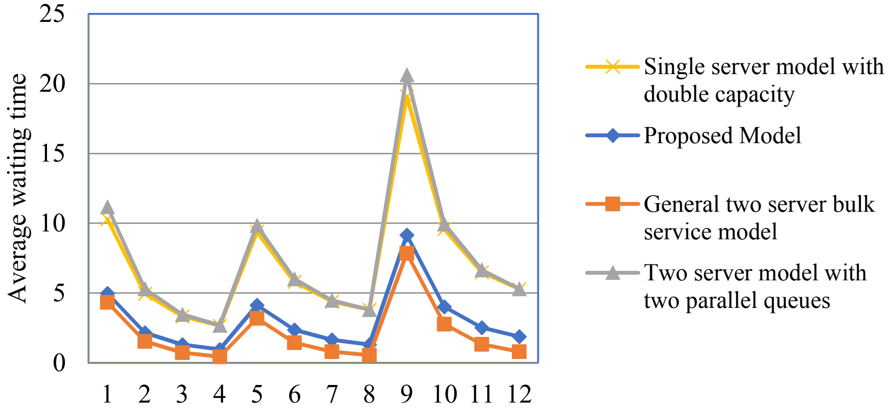

Figure 1 Comparison of waiting time

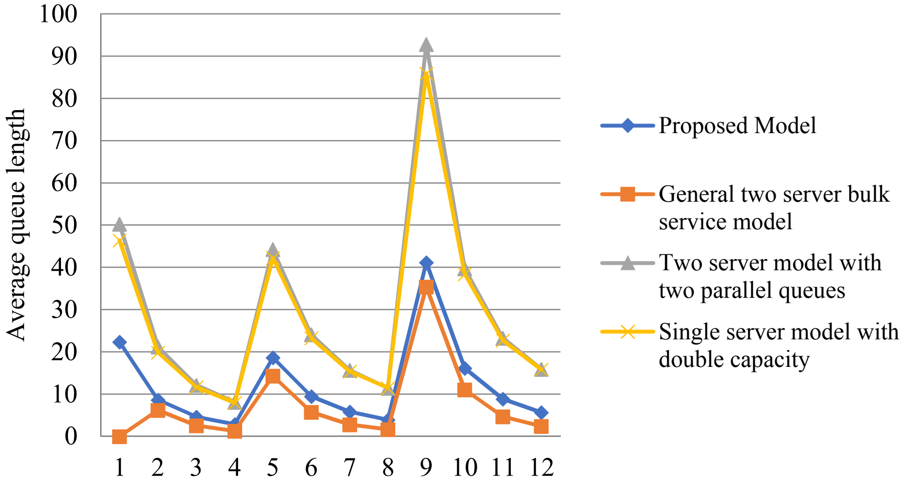

Figure 2 Comparison of queue length

7. Conclusions

We have used the semi-Markov process defined on two-dimensional state space to analyze a two-server model, where both the servers have equal service rate. This method will be also applicable for the systems where the two servers have different service rates. For any specified service mechanism, the method can be applied, and can be extended to the system with multiple servers. However, the only consideration for the applicability of the semi-Markov process is that the service time of a server is exponential distribution.

In Subsection 5.5, we have considered the cost of operating a service system, which is of great importance in real life. The operational cost helps the system designer to make a choice between different models with desired level of performance.

References

Abolnikov L, Dshalalow JH. On a multilevel controlled bulk queuing system Mx /G(r, R)/1. J Appl Math Stoch Anal. 1992; 5: 237-260.

Arora KL. Two-server bulk-service queuing process. Oper Res.1964; 12(2): 286-294.

Bailey NTJ. A continuous treatment of a simple queue using generating functions. J Roy Stat Soc B Met. 1954; 16(2): 288-291.

Banerjee A, Gupta U, Goswami V. Analysis of finite-buffer discrete-time batch-service queue with batch-size-dependent service. Comput Ind Eng. 2014; 75: 121-128.

Banerjee A, Gupta U, Chakravarthy S. Analysis of a finite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service. Comput Oper Res. 2015; 60: 138-149.

Bakuli K, Pal M. Bulk service queuing system with impatient customers: A computational approach. Thail Stat. 2017; 15(1): 1-10.

Chaudhry ML, Templeton JGC. The queuing system M/Gb/l and its ramifications. Eur J Oper Res.1981; 6(1): 56-60.

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

Chaudhry ML, Chang SH. Analysis of the discrete-time bulk-service queue Geo/GY /1/N + B. Oper Res Lett. 2004; 32(4): 355-363.

Ghare P. Multichannel queuing system with bulk service. Oper Res. 1968; 16: 189-192.

Ghimre S, Ghimre RP, Thapa GB, Farnandes S. Multi-server batch service queuing model with variable service rates. Int J Appl Math Stat Sci. 2017; 6(4): 43-54.

Gupta UC, Goswami V. Performance analysis of finite buffer discrete-time queue with bulk service. Comput Oper Res. 2002; 29(1): 1331-1341.

Holman D, Chaudhry M, Ghosal A. Some results for the general bulk service queuing system. Aust Math Soc. 1981; 23(2): 161-179.

Jaiswal NK. Bulk-service queuing problem. Oper Res. 1960; 8(1):139-143.

Krishnamoorthy A, Ushakumari PV. A queuing system with single arrival bulk service and single departure. Math Comput Model. 2000; 31: 99-108.

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

Pakes AG. Some conditions for ergodicity and recurrence of Markov chains. Oper Res. 1969; 17(6): 1058-1061.

Seneta E. Finite approximations to infinite non-negative matrices, II: refinements and applications. Proc Camb Phil Soc. 1968; 64: 465-470.

Wolf D. Approximation of the invariant probability measure of an infinite stochastic matrix. Adv Appl Prob. 1980; 12(3): 710-726.

Appendix

Derivation of transition probabilities for the proposed two-server queuing model

Let $A(n)$ denotes the sum of the inter-arrival times of n customers to the system. Then, $A(n)$ is the sum of n independent i.i.d. random variables each distributed as $\exp(\lambda)$, and hence $A(n) \sim gamma(\lambda, n)$. Further, it is independent of the service time T, which is distributed as $\exp(\mu)$.

In order to get the transition probabilities in a model, we have to primarily compute the following two probabilities:

$$P[A(n)<T]=\int_0^{\infty}\left(\int_x^{\infty}f_T(t)\,dt\right)f_{A(n)}(x)\,dx=\int_0^{\infty}\left(\int_x^{\infty}e^{-\mu t}\mu\,dt\right)\frac{\lambda}{(n-1)!}e^{-\lambda x}(\lambda x)^{n-1}\,dx$$

$$=\int_0^{\infty}e^{-\mu x}\frac{\lambda}{(n-1)!}e^{-\lambda x}(\lambda x)^{n-1}\,dx=\left(\frac{\lambda}{\lambda+\mu}\right)^n,$$

$$P\left[A(n)<T<A(n+1)\right]=P\left[A(n+1)>T\right]-P\left[A(n)>T\right]=\left(\frac{\lambda}{\lambda+\mu}\right)^n\left(\frac{\mu}{\lambda+\mu}\right).$$

Let $p=\frac{\lambda}{\lambda+\mu}$, $p_1=\frac{\lambda}{\lambda+2\mu}$. Let us denote the service times of the permanent and temporary servers by T_1 and T_2, respectively. The transition probabilities of the model are obtained as follows:

Case 1: $i_1=0$ and $i_2<b$

In this case, the next state will be $j_1=1$ and $j_2=0$ for sure. Then, $P\left((1,0),(0,i_2)\right)=1;$ transition t the other state is not possible.

Here unconditional waiting time at the state $(0,i_2)$ is the arrival time of $(b-i_2)$ customers or sum of $(b-i_2)$ inter arrival time, i.e. $A(b-i_2)$.

Case 2: $i_1=1$ and $i_2<b$

If the next state is $j_1=0, j_2<b;$

$$P\left((0,j_2),(1,i_2)\right)=P\left[A(j_2-i_2)<T_1<A(j_2-i_2+1)\right]=p^{j_2-i_2}(1-p).$$

If the next state is $j_1=1, j_2\in\{0,1,...,q-b-1\};$

$$P\left((1,j_2),(1,i_2)\right)=P\left[A(b+j_2-i_2)<T_1<A(b+j_2-i_2+1)\right]=p^{b+j_2-i_2}(1-p).$$

If the next state is $j_1=3, j_2=q-b;$

$$P\left((3,q-b)\,|\,(1,i_2)\right)=P\left[A(q-i_2)<T_1)\right]=p^{q-i_2}.$$

Here transition from this state is occurring due to service completion of the server 1 or arrival of $(q-i_2)$ arrivals. So the unconditional waiting time is $\min\left(T_1,A(q-i_2)\right)$.

Case 3: $i_1=2$ and $i_2<b$

If next state is $j_1=0, j_2<b;$

$$P\left((0,j_2),(1,i_2)\right)=P\left[A(j_2-i_2)<T_2<A(j_2-i_2+1)\right]=p^{j_2-i_2}(1-p).$$

If next state is $j_1=3, j_2=0;$

$$P\left((3,0)\,|\,(1,i_2)\right)=P\left[A(q-i_2)<T_1)\right]=p^{q-i_2}.$$

Here transition from this state is occurring due to service completion of the server 2 or arrival of $(b-i_2)$ arrivals. So the unconditional waiting time is $\min\left(T_2,A(q-i_2)\right)$.

Case 4: $i_1=3$ and $i_2<b$

If the next state is $j_1=1$ and $i_2<j_2<q,$

$$P\big((1,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2-i_2)<\min(T_1,T_2)<A(j_2-i_2+1),\min(T_1,T_2)=T_2\big],$$
$$=P\big[A(j_2-i_2)<\min(T_1,T_2)<A(j_2-i_2+1),\min(T_1,T_2)=T_1\big],$$
$$=\frac{1}{2}P\big[A(j_2-i_2)<\min(T_1,T_2)<A(j_2-i_2+1)\big]=\frac{1}{2}p_1^{j_2-i_2}(1-p_1).$$

If the next state is $j_1=2$ and $i_2\le j_2<b,$

$$P\big((2,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2-i_2)<\min(T_1,T_2)<A(j_2-i_2+1),\min(T_1,T_2)=T_1\big]=\frac{1}{2}p_1^{j_2-i_2}(1-p_1).$$

If the next state is $j_1=3$ and $j_2+b<q,$

$$P\big((3,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2-i_2+b)<\min(T_1,T_2)<A(j_2-i_2+b+1),\min(T_1,T_2)=T_1\big]=\frac{1}{2}p_1^{j_2-i_2+b}(1-p_1).$$

If next state is $j_1=3$ and $j_2+b\ge q,$

$$P\big((3,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2-i_2+b)<\min(T_1,T_2)<A(j_2-i_2+b+1)\big]=p_1^{j_2-i_2+b}(1-p_1).$$

Case 5: $i_1=3$ and $b\le i_2<q$

If the next state is $(1,j_2),$ where $i_2\le j_2<q,$

$$P\big((1,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2-i_2)<\min(T_1,T_2)<A(j_2-i_2+1),\min(T_1,T_2)=T_2\big]=\frac{1}{2}p_1^{j_2-i_2}(1-p_1).$$

If the next state is $(3,j_2)$, where $i_2\le j_2+b<q$, this is possible when $2b<q$,

$$P\big((3,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2+b-i_2)<\min(T_1,T_2)<A(j_2+b-i_2+1),\min(T_1,T_2)=T_1\big]=\frac{1}{2}p_1^{j_2+b-i_2}(1-p_1).$$

If the next state is $(3,j_2),$ where $j_2+b\ge q,$

$$P\big((3,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2+b-i_2)<\min(T_1,T_2)<A(j_2+b-i_2+1)\big]=p_1^{j_2+b-i_2}(1-p_1).$$

Case 6: $i_1=3$ and $q\le i_2$

If the next state is $(3,j_2)$, then

$$P\big((3,j_2)\,|\,(3,i_2)\big)=P\big[A(j_2+b-i_2)<\min(T_1,T_2)<A(j_2+b-i_2+1)\big]=p_1^{j_2+b-i_2}(1-p_1).$$

Here transition occurs only when a service completion happens, so that unconditional waiting time is $\min(T_1,T_2)$.

Derivation of transition probabilities for M/M ^{b}/2:

Case 1: $i_1=0$ and $i_2<b$

Here the only possible transition can occur to state (1,0) owing to the arrival of $(b-i_2)$ customers, which causes the queuing length to be reduced to zero and a service to be started. So, $P\big((1,0)\,|\,(0,i_2)\big)=1.$ The unconditional waiting time at the state $(0,i_2)$ is, therefore, the arrival time of $(b-i_2)$ customers or sum of $(b-i_2)$ inter arrival times, i.e. $A(b-i_2)$.

Case 2: $i_1 = 1$ and $i_2 < b$

If the next state is $j_1 = 0, i_2 \le j_2 < b$ (transition due to completion of service before arrival of (b-i_2) customers),

$$P\big((0, j_2) \mid (1, i_2)\big) = P\big[A(j_2 - i_2) < T_1 < A(j_2 - i_2 + 1)\big] = p^{j_2 - i_2}(1 - p).$$

If the next state is $j_1 = 1, j_2 \in \{0, 1, ..., q - b - 1\}$ (transition due to completion of service after arrival of more than $(b - i_2)$ but less than $(q - i_2)$ customers),

$$P\big((1, j_2) \mid (1, i_2)\big) = P\big[A(b + j_2 - i_2) < T_1 < A(b + j_2 - i_2 + 1)\big] = p^{b + j_2 - i_2}(1 - p).$$

If the next state is $j_1 = 2, j_2 = q - b,$

$$P\big((2, q - b) \mid (1, i_2)\big) = P\big[A(q - i_2) < T_1\big] = p^{q - i_2}.$$

Case 3: $i_1 = 1$ and $b < i_2 < q$

If the next state is $j_1 = 1, j_2 \in \{0, 1, ..., q - b - 1\},$

$$P\big((1, j_2) \mid (1, i_2)\big) = P\big[A(b + j_2 - i_2) < T_1 < A(b + j_2 - i_2 + 1)\big] = p^{b + j_2 - i_2}(1 - p).$$

If the next state is $j_1 = 2, j_2 = q - b,$

$$P\big((2, q - b) \mid (1, i_2)\big) = P\big[A(q - i_2) < T_1\big] = p^{q - i_2}.$$

Here transition from this state is occurring due to service completion of the server or arrival of $(q - i_2)$ arrivals. So unconditional waiting time would be $\min(T_1, A(q - i_2)).$

Case 4: $i_1 = 2$ and $i_2 < q$

If next state is $j_1 = 1, j_2 < q,$

$$P\big((1, j_2) \mid (2, i_2)\big) = P\big[A(j_2 - i_2) < \min(T_1, T_2) < A(j_2 - i_2 + 1)\big] = p_1^{j_2 - i_2}(1 - p_1).$$

If next state is $j_1 = 2, j_2 \ge q - b,$

$$P\big((2, j_2) \mid (2, i_2)\big) = P\big[A(b + j_2 - i_2) < \min(T_1, T_2) < A(b + j_2 - i_2 + 1)\big] = p_1^{b + j_2 - i_2}(1 - p_1).$$

Case 5: $i_1 = 2$ and $i_2 > q$

Then the next state will be $j_1 = 2, j_2 \ge q - b,$

$$P\big((2, j_2) \mid (2, i_2)\big) = P\big[A(b + j_2 - i_2) < \min(T_1, T_2) < A(b + j_2 - i_2 + 1)\big] = p_1^{b + j_2 - i_2}(1 - p_1).$$

Here transition occurs only when a service completion happens, so that the unconditional waiting time is $\min(T_1, T_2).$