

## DISCRETE TIME POISSON QUEUES OPERATED BY TWO HETEROGENEOUS SERVERS UNDER 'FIRST COME FIRST SERVED' QUEUE DISCIPLINE

R. SIVASAMY AND K. THAGA

Department of Statistics, Faculty of Social Sciences, University of Botswana,  
Gaborone, BP 00705

**ABSTRACT.** The proposed discrete-time queueing systems are  $\text{Geo}(\lambda)/\text{Geo}(\mu_1)+\text{Geo}(\mu_2)/2$  and  $\text{Geo}(\lambda)/\text{Geo}(\mu_1),\text{Geo}(\mu_2)/2$  that have an infinite number of waiting positions with one faster server i.e. server-1 and a slow server i.e. server-2. If the slow server is free, then according to the classical First Come First Served (FCFS) discipline, an incoming customer is assigned to the slow server. Now since this customer is getting the slowest possible service, customers arriving subsequently might clear out of the system earlier by getting service from the faster server. This is clearly a violation of the FCFS principle. Such a violation is greater for greater heterogeneity of the service capacities of the two servers. For such a situation, this paper proposes how a customer might find it preferable to wait for service at the faster server than to go into the slow server without violating of the FCFS principle through queue discipline-I and queue discipline-II. Further time axis is divided into fixed length intervals or slots. Customers arrive during the consecutive slots, but they can only start service at the beginning of slots. The numbers of customers that arrive in successive slots are independent, identically distributed (i.i.d.) random variables subject to a condition that only one customer can arrive in a slot with probability  $\lambda$  ( $0 < \lambda < 1$ ) and that no customer arrive in a slot with probability  $1 - \lambda$ . Customer service times are integer multiples of the slot length, which implies that customers leave the system at slot boundaries. All customers are served either by server-I according to geometric service time distribution with mean rate  $\mu_1$  or by server-2 with geometric service time distribution where mean rate is  $\mu_2 < \mu_1$ . The steady state analysis is then discussed and numerical values to the steady state expected number of customers  $E(N)_{\text{Geo}/\text{Geo}+\text{Geo}/2}$ , and  $E(N)_{\text{Geo}/\text{Geo},\text{Geo}/2}$  have been computed. To check if the proposed queues operate under FCFS rule and thus satisfy the Little's formula  $\lambda\bar{W} = E(N)$ , the actual expected waiting times  $\bar{W}_{\text{Geo}/\text{Geo}+\text{Geo}/2}$  and  $\bar{W}_{\text{Geo}/\text{Geo},\text{Geo}/2}$  of customers in the system are then calculated numerically. A simple comparison study over these numerical measures proves that there is an insignificant difference between  $E(N)_{\text{Geo}/\text{Geo}+\text{Geo}/2}$ , and  $E(N)_{\text{Geo}/\text{Geo},\text{Geo}/2}$  values and between the values  $\bar{W}_{\text{Geo}/\text{Geo}+\text{Geo}/2}$  and  $\bar{W}_{\text{Geo}/\text{Geo},\text{Geo}/2}$  due to the fact that the proposed two alternative queue disciplines here minimize violations of the FCFS discipline in the long run. Finally  $\text{Geo}(\lambda)/\text{Geo}(\mu_1)+\text{Geo}(\mu_2)/2$  queueing model is applied to model a single computing node as a server to obtain the power consumption and the associated expected cost for a specific set of input values.

**AMS (MOS) Subject Classification.** 39A10.

## 1. INTRODUCTION

This paper discusses a special case of at the most only one arrival occurring (late arrival) in a slot at slot boundaries and service to a job that can only start at a slot boundary. That is, service of customers is synchronized with respect to slot boundaries. Further, customer service times are integer multiples of the slot length, which implies that customers leave the system at slot boundaries. The service duration of a job and the inter-arrival durations between consecutive job arrivals are measured as random number of slot durations. One special feature of this investigation is that it derives those parallel results that have already been obtained by Sivasamy et al. (2015) to the corresponding continuous time version M/G/2 queues.

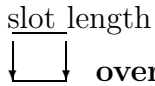
For sufficiently small slot lengths, discrete-time queueing models may also be used as an approximation of corresponding models where the time scale is continuous. In fact, one can obtain results for continuous-time models directly from the equivalent discrete-time results and vice versa. Asynchronous transfer mode (ATM) multiplexers and broadband integrated services digital network (B-ISDN) use to transfer data sets, voice and video communications on a discrete time basis. Hence for studying the important characteristics of such discrete-time queueing of jobs served by a computer or a telecommunication device, the time axis can be divided into slots (fixed-length of continuous intervals called slots of unit length (= right-end boundary-left-end boundary)).

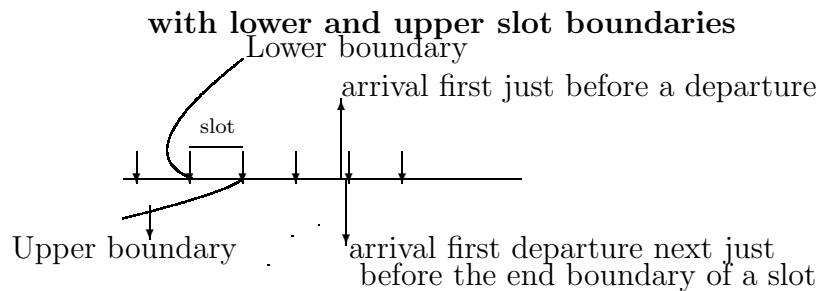
Literature: Over the recent years, several authors Singh (1968), Hoksad (1979), Boxma et al., (2002), Efrasinin (2008), Kim et al., (2011), Krishnamoorthy and Srinivasan (2012) have studied continuous time queue length processes of service systems with either two servers or multi-servers. As the discrete-time queueing systems are the best choices to model computer and communication systems, Takagi (1993), and Bruneel and Kim (1993) have presented most of the basic features of discrete-time parallel queueing systems to that of continuous counterparts. In continuous-time queues, probability of an arrival and departure occurring simultaneously in a very small time interval is zero, it is not so in discrete-time queues. They can occur simultaneously at a boundary epoch of a slot but their order must be taken care of by either arrival first (AF) or departure first (DF) management policies.

Assumptions on slot boundaries: Without loss of generality, we assume the length of each slot as unity. Arrivals are considered to occur on slot boundaries according to an 'Arrival First (AF)' policy. To be more specific, this paper discusses a queue with AF policy by letting the time axis be marked by  $0, 1, 2, \dots, t, \dots$ , and let potential arrivals occur at  $0-, 1-, \dots, t-, \dots$ . Service can start only at slot boundaries and always takes an entire number of slots. The customer leaves the system as soon as she has received the service. It is assumed that potential departures occur at

slot boundaries, i.e. at  $0+, 1+, \dots, t+, \dots$ . The time between successive arrivals is referred to as the inter-arrival time and is denoted by  $A$ . More details on this discrete time concepts can be found in Gupta and Goswami (2002) who have modeled a single-server bulk service queue with finite buffer space in a discrete-time environment and carried out analytic analysis of the model under both AF and DF management policies and distributions of buffer content at various epochs have been obtained. Such management policies play a significant role towards the determination of steady-state probabilities relating to the number of customers in the system (queue) at special epochs (e.g., arrival, and departure) and hence they affect performance measures to a great extent.

1.1. **SLOT Diagram under AF management policy.**

Dividing of time axis into slots between down arrows  over a line



When a businessman engages an experienced salesman and an apprentice or a senior doctor and a junior doctor as parallel servers in a queuing system, then the FCFS queue discipline is violated as it is well justified by Krishnamoorthi (1963) and Sivasamy et al., (2014). This type of violation of FCFS rule creates dissatisfaction among customer groups affected by such random selection of a server and eventually will lead to abandonment, renegeing, balking, etc. thereby reducing profits.

1.2. **Queue Discipline-II that minimizes violation of the FCFS.** The proposed ‘Queue Discipline-II’ of Krishnamoorthi is a refined FCFS policy with an  $m$ -policy to reduce the impacts of violating the FCFS so that the resulting waiting times of customers are identical with that of the FCFS rule subject to the condition that mean service rates of server-1(=Channel I) and server- 2(=Channel II) are  $\mu_1$  and  $\mu_2$  respectively. Further  $\mu_1$  is ‘ $m$ ’ times larger than  $\mu_2$  almost surely:

**A customer arrives to find:**

1. **Both channels free;** it occupies Channel I (assuming that Channel I gives faster service on the average)
2. **Channel I is engaged;** it waits for service before Channel I whether or not Channel II is free. But if the number of units waiting for service before Channel

I becomes ' $m$ ' (a positive integer), it goes to Channel II for service if that is free; otherwise it waits as the  $(m + 1)^{th}$  unit in the queue. It should be noted that the first ' $m$ ' units in the queue will be getting service from Channel I. The  $(m + 1)^{th}$  unit in the queue will go to Channel II if that becomes free prior to the finishing of service of the unit in Channel I. Otherwise it will move up as the  $m^{th}$  unit in the queue, and hence decides to take service from Channel I.

3. **Both channels are engaged;** and a waiting line of length ' $n$  say' greater than or equal to ' $m$ ' joins the queue as the last  $(n + 1)^{st}$  number. All units after the  $m^{th}$  in the queue take a decision only when they reach the  $(m + 1)^{th}$  position in the queue. The decision is taken according to the rule mentioned in 2 of 'Channel I is engaged'.

The positive integer ' $m$ ' is to be chosen such that it is one less than the greatest integer in the ratio of  $(\mu_1/\mu_2)$ . So  $m + 1 =$  an integer just greater than or equal to  $(\mu_1/\mu_2)$ . It is clear that for this choice of ' $m$ ', the following happens: When there are ' $m$ ' units waiting for service in Channel I, an incoming unit finds it profitable to go to Channel-II if that is free since the amount  $(m + 2) (\mu_1)^{-1}$  is larger than  $(\mu_2)^{-1}$ . Similarly when there are only  $(m - 1)$  units waiting for service in Channel I, an incoming unit will find it profitable to join the queue for service in Channel I, even if Channel II is free since  $(m + 1) (\mu_1)^{-1}$  is smaller than  $(\mu_2)^{-1}$ . Thus this queue discipline achieves the objective that the least amount of waiting time is spent in the system according to the conditions present on its arrival at the system due to the fact that this Queue Discipline-II reduces the violation of FCFS principle.

It is remarked that Sivasamy et al. (2014) have pre-fixed the ' $m$ ' as ' $m = 1$ ' (instead of computing the ' $m$ ' value as one less than the greatest integer contained in the ratio of  $(\mu_1/\mu_2)$ ) in their investigations who also compared the steady state mean performance of the M/G/2 queuing systems under the serial and parallel service schedules.

**1.3. Organization of the methodology.** In section 2, difference equation method has been used to analyze the  $\text{Geo}(\lambda)/\text{Geo}(\mu), \text{Geo}(\mu_2)/2$  under 'Queue Discipline-II' with an ' $m$ ' policy. Section 3 proposes a condition on the classical FCFS queue  $\text{Geo}(\lambda)/\text{Geo}(\mu) + \text{Geo}(\mu_2)/2$  with an ' $m$ ' policy to yield the same steady state behavior of  $\text{Geo}(\lambda)/\text{Geo}(\mu), \text{Geo}(\mu_2)/2$  queue; i.e. most performance measures of both  $\text{Geo}(\lambda)/\text{Geo}(\mu), \text{Geo}(\mu_2)/2$  and  $\text{Geo}(\lambda)/\text{Geo}(\mu) + \text{Geo}(\mu_2)/2$  queues become identical and in particular their mean queue length and mean waiting time values are almost equal. Section 4 concludes the various special features of the proposed methodology and its future scope.

2. **Poisson Queue:  $\text{Geo}(\lambda)/\text{Geo}(\mu), \text{Geo}(\mu_2)/2$  under ‘Queue Discipline-II’**

Considered is the discrete version of the two server Poisson queue (Poisson input, geometric service times) of Krishnamoorthi *i.e.*  $M/(M_1, M_2)/2$  under ‘Queue Discipline-II’ with service rates  $\mu_1$  and  $\mu_2 (\leq \mu_1)$  and arrival rate  $\lambda (\neq \mu_1)$  in Channels I and II respectively like that of two suppliers/servers of an inventory house filling the orders with different mean times.

2.1. **Steady-state Probability Distribution of Queue Length Process.** Let the steady-state probability distribution  $\{p(\cdot)\}$  of queue length process over the three mutually exclusive states (a), (b) and (c) of the servers be defined as follows subject to the condition  $\rho = \lambda/(\mu_1 + \mu_2), < 1$ :

(a) **No units waiting for service:**

$$\begin{aligned} p_{00} &= P \text{ \{both the Channels are free\}} \\ p_{01} &= P \text{ \{Channel I is free and Channel II engaged\}} \\ p_{10} &= P \text{ \{Channel I is engaged and Channel II free\}} \\ p_{11} &= P \text{ \{both the Channels are engaged\}} \end{aligned}$$

(b) **Number of units waiting for service is ‘n’ such that  $1 \leq n \leq m$**

$$\begin{aligned} p_{n.10} &= P\{n \text{ units waiting for service, no unit in Channel II and one unit in Channel I}\} \\ p_{n.11} &= P\{n \text{ units waiting for service in the queue, both Channels engaged}\}. \end{aligned}$$

(c) **Number of units in the system is  $n \geq (m + 3)$**

$$p_n = P\{\text{the system has } n \text{ units; } (n - 2) \text{ units wait for service and both channels engaged}\} \text{ for } n = m + 3, m + 4, \dots$$

**Steady state queue length distribution:**

Let  $p_{-1.11} = p_{01}, p_{-1.10} = p_{00}, p_{-2.11} = 0, p_0 = p_{00}$ . Also let,  $p_n = p_{n-1.10} + p_{n-2.11}$  for  $n = 0, 1, 2, 3, \dots (m + 1)$  and  $p_{m+2} = p_{m.11}$ . It is to be noted that  $p_{m+1.10} = 0$  and hence  $p_{m.11} = p_{m+2}$ . Thus,  $\sum_{n=0}^{\infty} p_n = 1$  is called the normalizing condition to the following system of difference equations (A), (B), (C) and (D) satisfied by the steady-state probabilities that can be derived in the usual manner:

$$\left. \begin{aligned} \mu_1 p_{10} + \mu_2 p_{01} &= \lambda p_{00} \\ \mu_1 p_{11} &= (\lambda + \mu_2) p_{01} \\ \mu_1 p_{1.10} + \mu_2 p_{11} + \lambda p_{00} &= (\lambda + \mu_1) p_{10} \\ \mu_1 p_{1.11} + \lambda p_{01} &= (\lambda + \mu_1 + \mu_2) p_{11} \end{aligned} \right\}, \dots \tag{A}$$

$$\left. \begin{aligned} \mu_2 p_{n.11} + \mu_1 p_{n+1.10} + \lambda p_{n-1.10} &= (\lambda + \mu_1) p_{n.10} \\ \mu_1 p_{n+1.11} + \lambda p_{n-1.11} &= (\lambda + \mu_1 + \mu_2) p_{n.11} \end{aligned} \right\}, \tag{B}$$

for  $n = 1, 2, \dots (m - 1) \dots$

$$\mu_2 p_{m.11} + \lambda p_{m-1.10} = (\lambda + \mu_1) p_{m.10} \dots \tag{C}$$

$$\left. \begin{aligned} (\mu_1 + \mu_2 p_{m+3} + \lambda p_{m+1} &= (\lambda + \mu_1 + \mu_2) p_{m.11} \\ (\mu_1 + \mu_2 p_{n+1} + \lambda p_{n-1} &= (\lambda + \mu_1 + \mu_2) p_n \end{aligned} \right\}, \text{ for } n = m + 3, m + 4, \dots \infty \dots \quad (\text{D})$$

**Solution:** Since  $p_{m.11} = p_{m+2}$ , it is observed from (D) that

$$p_n = \rho^{n-(m+1)} p_{m+2} \quad \text{for all } n \geq m + 2 \quad (2.1)$$

Rewrite the first two of equations of (A) and the whole of (C) as (A') below:

$$\left. \begin{aligned} \mu_1 p_{l0} + \mu_2 p_{01} &= \lambda p_{00} \\ \mu_1 p_{l1} &= (\lambda + \mu_2) p_{01} \\ \mu_2 p_{m.11} + \lambda p_{m-1.10} &= (\lambda + \mu_1) p_{m.10} \end{aligned} \right\}, \dots \quad (\text{A}')$$

Let  $P = (p_{n.10}, p_{n.11})$  be a row vector and  $\mathbf{M} = \begin{pmatrix} \lambda + \mu_1 & 0 \\ -\mu_2 & \lambda + \mu_1 + \mu_2 \end{pmatrix}$  be a square matrix of order 2. The following matrix equation (B') is obtained from those equations of (A) that are not accounted in (A') and from (B)

$$P_n \mathbf{M} = \mu_1 P_{n+1} + \lambda P_{n-1} \quad \text{for } n = 0, 1, 2, \dots, (m-1) \dots \quad (\text{B}')$$

Two latent roots of the matrix  $\mathbf{M}$  of (B') are  $(\lambda + \mu_1)$  and  $(\lambda + \mu_1 + \mu_2)$ . Select the corresponding latent vectors as  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$  and  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$  respectively. Letting

$$Q_n = p_{n.10} + p_{n.11} \quad (2.2)$$

$$Q'_n = p_{n.11} \quad (2.3)$$

and multiplying the right of (B') by  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$  and  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$  respectively, for  $n = 0, 1, 2, \dots, (m-1)$ , it is established that

$$(\lambda + \mu_1) Q_n = \mu_1 Q_{n+1} + \lambda Q_{n-1} \quad (2.4)$$

$$(\lambda + \mu_1 + \mu_2) Q'_n = \mu_1 Q'_{n+1} + \lambda Q'_{n-1} \quad (2.5)$$

It can be shown with the first two boundary conditions of (A)' that the solution to (2.4) is

$$Q_n = (p_{00} + p_{01}) w^{n+1} \quad \text{where } w = \frac{\lambda}{\mu_1} \quad \text{for } n = -1, 0, \dots, m \quad (2.6)$$

Let  $w_j$  for  $j = 1$  and  $2$  be the two roots of the following quadratic equation

$$\mu_1 x^2 - (\lambda + \mu_1 + \mu_2) x + \lambda = 0 \quad (2.7)$$

$$w_1 = \frac{(\lambda + \mu_1 + \mu_2) + \sqrt{(\lambda + \mu_1 + \mu_2)^2 - 4\lambda\mu_1}}{2\mu_1} \quad (2.8)$$

$$w_2 = \frac{(\lambda + \mu_1 + \mu_2) - \sqrt{(\lambda + \mu_1 + \mu_2)^2 - 4\lambda\mu_1}}{2\mu_1}$$

Define, for  $n = 1, 2, \dots, m$ , the following quantities:

$$a' = (w_1 - 1)w_1 \tag{2.9}$$

$$b' = w_2(1 - w_2) \tag{2.10}$$

$$r_n = a'w_1^n + b'w_2^n = (w_1 - 1)w_1^{n+1} + (1 - w_2)w_2^{n+1} \tag{2.11}$$

Now it is easy to check that the solution of (2.5) is

$$Q'_n = r_n(p_{01}) \quad \text{for } n = 1, 2, \dots, m \tag{2.12}$$

Application of the normalizing condition  $\sum_{n=-1}^m Q_n + \sum_{n=m+3}^\infty p_n = 1$  together with  $Q'_m = r_m(p_{01}) = p_{m.11}$  and (2.1) leads that

$$(p_{00} + p_{01}) \left( \frac{1 - w^{m+2}}{1 - w} \right) + p_{01}r_m \left( \frac{\rho}{1 - \rho} \right) = 1 \dots \tag{2.13}$$

On using the fact of  $Q_m - Q'_m = p_{m.10}$  and  $Q'_m = p_{m.11}$  into the third boundary condition  $\mu_2 p_{m.11} + \lambda p_{m-1.10} = (\lambda + \mu_1)p_{m.10}$  of (A)', it is seen that

$$\begin{aligned} \mu_2 p_{m.11} + \lambda p_{m-1.10} &= (\lambda + \mu_1)p_{m.10} \\ \Rightarrow \mu_2 Q'_m + \lambda (Q_{m-1} - Q'_{m-1}) &= (\lambda + \mu_1) (Q_m - Q'_m) \\ \Rightarrow (\lambda + \mu_1 + \mu_2)Q'_m - \lambda Q'_{m-1} &= (\lambda + \mu_1)Q_m - \lambda Q_{m-1} \\ \Rightarrow p_{01} [(\lambda + \mu_1 + \mu_2)r_m - \lambda r_{m-1}] &= \lambda(p_{00} + p_{01})w^{m+1} \end{aligned}$$

Converting the above scalar equations into its matrix version, it is found that

$$\begin{pmatrix} \left( \frac{1-w^{m+2}}{1-w} \right) & \left( \frac{1-w^{m+2}}{1-w} \right) + r_m \left( \frac{\rho}{1-\rho} \right) \\ \lambda w^{m+1} & \lambda w^{m+1} + \lambda r_{m-1} - (\lambda + \mu_1 + \mu_2)r_m \end{pmatrix} \begin{pmatrix} p_{00} \\ p_{01} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{2.14}$$

Let  $d$  be the determinant value of the non-singular co-efficient matrix of (2.14) then

$$d = \frac{(1 - w^{m+2})}{1 - w} [\lambda r_{m-1} - (\lambda + \mu_1 + \mu_2)r_m] - \lambda w^{m+1} \left[ \frac{\rho r_m}{1 - \rho} \right] \tag{2.15}$$

Solving for the unknowns of (2.14) by using inversion method, it is obtained that

$$p_{00} = [\lambda w^{m+1} + \lambda r_{m-1} - (\lambda + \mu_1 + \mu_2)r_m] / d \tag{2.16}$$

$$p_{01} = -\lambda w^{m+1} / d \tag{2.17}$$

Since the queue length process is an ergodic process,  $p_{01} > 0$  of which is ensured if  $d < 0$  i.e.  $d$  is a negative real number. Now the steady-state probability distribution

$\{p(\cdot)\}$  of queue length process is completely determined as follows:

$$\left. \begin{aligned}
 d &= \frac{(1-w^{m+2})}{1-w} [\lambda r_{m-1} - (\lambda + \mu_1 + \mu_2)r_m] \\
 &\quad - \lambda w^{m+1} \left[ \frac{\rho r_m}{1-\rho} \right] \\
 p_{00} &= [\lambda w^{m+1} + \lambda r_{m-1} - (\lambda + \mu_1 + \mu_2)r_m] / d, \\
 p_{01} &= -\lambda w^{m+1} / d \\
 p_{10} &= p_{00}w + \left( \frac{\mu_2}{\mu_1} \right) \left( \frac{\lambda w^{m+1}}{d} \right) \\
 p_{11} &= p_{01} \frac{(\lambda + \mu_2)}{\mu_1}, \text{ where } w = \frac{\lambda}{\mu_1} \\
 p_{n.11} &= p_{01}r_n, p_{n.10} = (p_{00} + p_{01})w^{n+1} - p_{01}r_n \\
 &\text{for } n = 1, 2, \dots, m \\
 p_{m.11} &= p_{m+2} = p_{01}r_m, \\
 \text{and } p_n &= \rho^{n-(m+2)}p_{m+2} \text{ for } n \geq (m + 2)
 \end{aligned} \right\} \tag{2.18}$$

From this distribution, all moments can be numerically calculated for a given set of input values on  $\lambda$ ,  $\mu_1$ , and  $\mu_2$  provided  $\rho = \lambda/(\mu_1 + \mu_2) < 1$ . For example the expected number  $E(N)$  of customers in the system is given by

$$E(N) = \sum_{n=0}^{\infty} np_n \text{ where } p_n = p_{n-1.10} + p_{n-2.11} \text{ for } n = -1, 0, \dots, (m + 2) \tag{2.19}$$

**2.2. Formulation of a Cost Balancing Problem.** A simple question arises on how to tackle the problem of determining  $\mu_2$ , the service capacity in the second channel. Assume that the objective is to find the value of  $\mu_2$  such that average waiting time length cost and average service capacity cost must be equal. This objective is expected to create an impression in the mind of each arriving customer that the costs of mean service time and the mean waiting time might be almost equal. Let

$C_0$  =penalty cost per unit time wait,

$\beta_i$  = cost incurred per unit service when the  $i^{th}$  channel is busy for  $i = 1$  and  $2$

$\lambda$  = arrival rate,

$\mu_i$  = number of units serviced per unit per unit time in the  $i^{th}$  channel ( $i = 1, 2$ ).

$E(W)$  = average total time (=  $W$ ) spent in the system =  $E(N)/\lambda$

Now let the cost per unit time incurred by the firm providing the service be  $T_1$  and that of average waiting time be  $T_2$ . Then  $T_1 = \beta_1\mu_1(1 - p_{00} - p_{10}) + \beta_1\mu_2(1 - p_{00} - p_{01})$  and  $T_2 = C_0E(N)/\lambda$ . Since each of the arrival and departure processes follows a Poisson process and the arrival and service rates are known constants, values of  $T_1$  are expected to increase while values of  $T_2$  are expected to decrease linearly with increasing values in  $\mu_2$ . To find one such a value of  $\mu_2$  such that average waiting time cost and average service capacity cost are equal through a numerical exercise, the following values are used:  $\lambda = 10.2$ ,  $\mu_1 = 9.2$ ,  $\beta_1 = 0.12$  and  $\beta_2 = 0.125$  and  $C_0 = 22.432$  while  $\mu_2$  values vary over the closed interval (3.39493012, 3.39512202).



TABLE 1. Values of  $T_1$  and  $T_2$  for  $\lambda = 10.2$ ,  $\mu_1 = 9.2$ ,  $\beta_1 = 0.12$ ,  $\beta_2 = 0.125$  and  $C_0 = 22.432$

$\mu_2$ $m = 2$	$T_1$	$T_2$	$T_1 - T_2$
3.39493012	9.75189192	9.75203801	-0.000146089511
3.39494022	9.75189576	9.75200619	-0.000110431377
3.39495032	9.75189959	9.75197437	-7.47735159E-005
3.39497052*	9.75190727	9.75191073	-3.45860939E-006*
3.39498062	9.75191111	9.75187891	3.21984358E-005
3.39499072	9.75191495	9.75184709	6.7855209E-005
3.39512202	9.75196485	9.75143348	0.000531368505

The corresponding values selected on  $\mu_2$  and the outcomes obtained on  $T_1$  and  $T_2$  are reported in Table-1.

As the largest integer contained in  $(9.2/3.3949301)$  and  $(9.2/3.3949901)$  is 3, the value to be assigned to  $m$  is  $3 - 1 = 2$ . A simple inspection over the values of  $(T_1 - T_2)$  reveals that the best value of  $\mu_2$  that satisfies the objective of the above problem that balances the average waiting time cost and average service capacity cost is found as 3.39497052.

### 3. Poisson Queue Geo/Geo<sub>1</sub>+Geo<sub>2</sub>/2 under Queue Discipline-I

This Queue Discipline-I of the Geo/Geo<sub>1</sub>+Geo<sub>2</sub>/2 queue refers to the classical FCFS rule. Assume that the arrival process follows Poisson law with arrival rate  $\lambda$  ( $\neq \mu_1$ ), service time distribution to each of the two service channels is exponential with service rates of channels I and II being  $\mu_1$  and  $\mu_2$  ( $\leq \mu_1$ ) respectively and are serially connected. Each arriving customer is served jointly by both servers according to an ' $m$ -policy' which perfectly implements the FCFS rule as detailed below.

There are some services of companies in our social and real life business applications that do not allocate a single server to serve some customers. For instance there are service counters where one may come across a service process involving both the salesman and the boss (owner) who provide services together to a customer being served if the queue length value (excluding the one being served) is beyond a threshold level, 'say  $m$ ', in order to speed up the service rate. On the other hand if the queue length is less than or equal to ' $m$ ' customers then each customer is serviced by the salesman only till the queue length crosses the level ' $m$ ' if a few more arrivals occur during the on-going service periods containing less than ' $m$ ' customers in the queue length, then the boss joins the salesman serially as and when the queue length crosses the level ' $m$ ' to provide the service jointly with the salesman as a customer satisfaction measure of reducing the average waiting time. This mechanism ensures

that the second server joins with the first server as long as the system size (including the one being served) is larger than or equal to  $(m + 2)$  or the queue length (including the one being served) is greater than or equal to  $(m + 1)$ . A remarkable feature of the serial service schedule is that the service process does not violate the classical FCFS queue discipline which is known as ‘Queue Discipline-I’ with an  $m$ -policy.

The past literature has accounted several multi-server queueing systems that have investigated into the impacts of heterogeneity of the servers. Since the middle of 1970’s more interesting and useful results have been contributed on homogeneous service systems (see Hoksad (1978), and Tijms et al. (1981)). Also focuses on queues with heterogeneous service channels have been well discussed since 1960’s from the first work made by Gumbel (1960) on a Poisson queue allowing the differences in the service capacity of the servers and measuring the errors occurring due to an assumption that all service rates are equal. Singh (1971) has established an optimal combination of service rates that minimize the performance measures of  $M/M_i/3$  queues. Further investigations about performance evaluation of queueing systems operated by heterogeneous servers are found in Gall (1998), Grassmann and Zhao (2004), Alves *et al.* (2011) and Sivasamy *et al.* (2014). Reviewing over the above contributions and other types of heterogeneous servers, the authors of this paper get motivated to develop a new methodology to analyze Markovian queueing system  $\text{Geo}/\text{Geo}_1+\text{Geo}_2/2$  subjecting it to a serial service process with an ‘ $m$ -policy’. A condition is found for this heterogeneous  $\text{Geo}/\text{Geo}_1+\text{Geo}_2/2$  system to yield the same steady state probability distribution of queue length and other performance measures of the two server Poisson queue  $\text{Geo}/(\text{Geo}_1,\text{Geo}_2)/2$  operating under ‘Queue Discipline-II’ discussed in section 2 above, where servers provide parallel services to all waiting customers with an  $m$ -policy.

**3.1. Equilibrium equations of  $\text{Geo}/\text{Geo}_1+\text{Geo}_2/2$ .** Let  $\pi_i$  be the steady state probability to find ‘ $i$ ’ customers in the system subject to the condition  $\rho = \lambda/(\mu_1 + \mu_2) < 1$ . Since each arriving customer into the  $\text{Geo}/\text{Geo}_1+\text{Geo}_2/2$  under ‘Queue Discipline-I’ is served jointly by both servers according to the ‘ $m$ -policy’ which perfectly implements the FCFS rule, the equilibrium equations of the queue length process can be obtained from the conservation of flow as stated below.

$$\mu_1\pi_1 = \lambda\pi_0 \Leftrightarrow \pi_1 = \left(\frac{\lambda}{\mu_1}\right)\pi_0 \quad (3.1)$$

$$\lambda\pi_{n-1} + \mu_1\pi_{n+1} = (\lambda + \mu_1)\pi_n \text{ for } n = 1, 2, 3, \dots, m \quad (3.2)$$

$$\Leftrightarrow \pi_n = \left(\frac{\lambda}{\mu_1}\right)^n \pi_0 \text{ for } n = 2, 3, \dots, m + 1$$

$$\left. \begin{aligned} \lambda\pi_m + (\mu_1 + \mu_2)\pi_{m+2} &= (\lambda + \mu_1)\pi_{m+1} \\ \lambda\pi_{n-1} + (\mu_1 + \mu_2)\pi_{n+1} &= (\lambda + \mu_1 + \mu_2)\pi_n \\ \text{for } n &\geq (m + 2) \end{aligned} \right\} \quad (3.3)$$

$$\Leftrightarrow \pi_n = \left(\frac{\lambda}{\mu_1}\right)^{m+1} \rho^{(n-m-1)} \pi_0 \text{ for } n \geq m + 1$$

$$\sum_{n=0}^{\infty} \pi_n = 1 \Rightarrow \pi_0 = \frac{(1 - \rho)(1 - \rho_1)}{(1 - \rho)(1 - \rho_1^{m+2}) + (1 - \rho_1)\rho\rho_1^{m+1}} \text{ where } \rho_1 = \frac{\lambda}{\mu_1} \neq 1 \quad (3.4)$$

Expected number E(Q) of customers in the system of Geo/Geo<sub>1</sub>+Geo<sub>2</sub>/2 queue is

$$\begin{aligned} E(Q) &= \sum_{n=0}^{m+1} n\pi_n + \sum_{n=m+2}^{\infty} n\pi_n \quad (3.5) \\ &= \pi_0 \left[ \rho_1 \frac{1 - (m + 2)\rho_1^{m+1} + (m + 1)\rho_1^{m+2}}{(1 - \rho_1)^2} \right. \\ &\quad \left. + \frac{\rho\rho_1^{m+1}((m + 2) - (m + 1)\rho)}{(1 - \rho)^2} \right] \end{aligned}$$

To find a condition for this heterogeneous Geo/Geo<sub>1</sub>+Geo<sub>2</sub>/2 system where servers are serially connected to yield the same steady state results of the Geo/(Geo<sub>1</sub>, Geo<sub>2</sub>)/2 system where servers provide parallel services to all waiting customers discussed in section-2, the following methodology is recommended.

Assume here that there is a dual server to server-2 associated with the heterogeneous Geo/Geo<sub>1</sub>+Geo<sub>2</sub>/2 system who accepts first customer at the head of the queue, if any, to transfer him/her to the real service (to be served either by server-1 or by both servers jointly) as and when the on-going service with a customer is completed. Notice that this dual server is continuously busy during each type of busy period (there are four types of busy periods). Let  $q_n$  = conditional probability that the dual server sees ‘ $n$ ’ number of customers in the queue (excluding the customer receiving service from server-1) while server-1 is busy.

$$q_{n-1} = \frac{\pi_n}{1 - \pi_0} \text{ for } n = 1, 2, \dots, \infty$$

Then, it is claimed that the sequence  $\{q_n\}$  that is obtained from the distribution  $\{\pi_n\}$  of Geo/Geo+Geo/2 system and the sequence  $\{p_n$  for  $n = 0, 1, 2, \dots, \infty\}$  of Geo/(Geo, Geo)/2 system should yield the same average queue lengths or mean sojourn times of customers.

To establish this claim, number of numerical illustrations have been carried out by randomly fixing  $\lambda$ ,  $\mu_1$  and  $\mu_2$  values and a summary over numerical results on  $\{q_n$  for  $n = 0, 1, \dots, m + 1\}$  of Geo/Geo+Geo/2 system and  $\{p_n$  for  $n = 0, 1, 2, \dots, m + 1\}$  of the Geo/ (Geo, Geo) /2 system is provided in Table-2.

**Table 2:** First  $(m + 1)$  probability values of queue length distribution of Geo/(Geo, Geo)/2 and Geo/Geo+Geo/2 queues for a given  $\lambda = 10.5$  and  $\mu_2 = 5.5$  while  $\mu_1$  varies from 5.2

$\mu_1$	Geo/(Geo, Geo)/2 $\{p_n, n = 0, 1, 2, \dots, m+1\}$	$m$	Geo/Geo+Geo/2 $\{q_n, n = 0, 1, 2, \dots, m+1\}$
5.5	$p_0 = 0.00986944333$ $p_1 = 0.0188416656$ $p_2 = 0.0181549849$ $p_3 = 0.0178156394$ $E(N) = 52.9542388$ $E(W)$ $\bar{W} = 5.04326084$	1	$q_0 = 0.00969590128$ $q_1 = 0.018164369$ $q_2 = 0.0178248481$ $q_3 = 0.0174916734$ $E(N) = 51.9909652$ $\bar{W} = 4.95152049$
11.5	$p_0 = 0.249611414$ $p_1 = 0.227906073$ $p_2 = 0.207796376$ $p_3 = 0.116829584$ $p_4 = 0.0734557263$ $E(N) = 2.12049344$ $E(W)$ $\bar{W} = 0.201951756$	2	$q_0 = 0.240470412$ $q_1 = 0.219559942$ $q_2 = 0.126042612$ $q_3 = 0.0792483486$ $q_4 = 0.049826806$ $E(N) = 1.47352654$ $\bar{W} = 0.140335861$
16	$p_0 = 0.382358849$ $p_1 = 0.250922994$ $p_2 = 0.16466807$ $p_3 = 0.108005049$ $p_4 = 0.0474661276$ $p_5 = 0.023509167$ $E(N) = 1.37274162$ $E(W)$ $\bar{W} = 0.130737297$	3	$q_0 = 0.377803591$ $q_1 = 0.247933606$ $q_2 = 0.162706429$ $q_3 = 0.0528843863$ $q_4 = 0.0261927385$ $q_5 = 0.0129728186$ $E(N) = 0.990509072$ $\bar{W} = 0.0943341974$

A simple comparison study over the probability values of Table-2 and other numerical measures proves that there is an insignificant difference between  $E(N)_{Geo/Geo+Geo/2}$ , and  $E(N)_{Geo/Geo,Geo/2}$  values and between the values  $\bar{W}_{Geo/Geo+Geo/2}$  and  $\bar{W}_{Geo/Geo,Geo/2}$ . It is due to the fact that the proposed two alternative queue disciplines here minimize violations of the FCFS discipline in the long run.

**Geo/Geo/1 queue:** It is of some interest that corresponding closed forms of expressions for the classical single server Geo/Geo/1 queue follows from that of the results through by applying  $\mu_2 \rightarrow 0$  i.e.  $\rho = \frac{\lambda}{\mu_1 + \mu_2} \rightarrow \rho_1$ :

$$\pi_n = \left(\frac{\lambda}{\mu_1}\right)^n \pi_0 \text{ for } n \geq 0 \text{ where } \pi_0 = 1 - \frac{\lambda}{\mu_1}$$

Expected number,  $E(Q_1)$  say, of customers in the Geo/Geo/1 system is then given by

$$E(Q_1) = \frac{\rho_1}{1 - \rho_1} \dots \quad (3.6)$$

### 3.2. Application of Geo/Geo+ Geo/2 model to a single computing node.

This section discusses a way out on how to serve the customers of a single computing node as customers of Geo/Geo+ Geo/2 system of the preceding analysis. For, the objective is to balance the power consumption (PC) and the quality of service (QoS) of the Central Processing unit (CPU) of some computing nod. The customers are either the incoming messages or demands into the CPU called the server. It becomes a single server case of Geo/Geo+ Geo/1 type, under the assumption that the single CPU dynamically adjusts the service rates as  $\mu_1$  during the system size is strictly less than  $T = (m + 2)$  say, for  $m = -1, 0, 1, 2, \dots$ , and as  $\mu_1 + \mu_2$  as long as the system size is more than or equal to  $T$ , in order to operate like a Geo/Geo+ Geo/1 system according to the T-policy yielding following results subject to  $\rho_1 = \frac{\lambda}{\mu_1} \neq 1$ ,  $\rho_1 = \frac{\lambda}{\mu_1 + \mu_2} < 1$  and  $T = m + 2$  in those results (3.1) through (3.4) of Geo/Geo+ Geo/2 system. Some of those results needed for further discussion are summarized below:

Steady state probabilities:

$$\pi_0 = \frac{(1 - \rho)(1 - \rho_1)}{(1 - \rho)(1 - \rho_1^T) + (1 - \rho_1)\rho\rho_1^{T-1}} \text{ where } \rho_1 = \frac{\lambda}{\mu_1} \neq 1 \quad (3.7)$$

$$P_T = \sum_{n=0}^{T-1} \pi_n = \pi_0 \frac{(1 - \rho_1^T)}{(1 - \rho_1)} \quad (3.8)$$

Mean Queue Length  $L$  and Mean waiting Time  $\bar{W}$ :

$$L = \pi_0 \left[ \rho_1 \frac{1 - T\rho_1^{T-1} + (T-1)\rho_1^T}{(1 - \rho_1)^2} + \frac{\rho\rho_1^{T-1}(T - (T-1)\rho)}{(1 - \rho)^2} \right] \text{ and } \bar{W} = \frac{L}{\lambda} \quad (3.9)$$

Assume that the CPU is operated between frequency values  $f_{\min} = 0.8$  GHz and  $f_{\max} = 2.7$  GHz. If the number 'n' of queueing demands/customers is less than  $T$ , the CPU is operated at a low frequency value  $f_0 \in [f_{\min} = 0.8, f_{\max} = 2.7]$  while it is to be operated at a higher frequency  $f_1 \in [f_{\min} = 0.8, f_{\max} = 2.7]$  when the number 'n' is greater than or equal to  $T$ . The power consumption function  $P(f_0, f_1, T)$  is dependent on  $C$  = the capacity of the transistor,  $V$  = supply voltage and  $P_{\text{static}}$  the static power to be consumed and is measured through a rule given by

$$P(f_0, f_1, T) = C[f_1 - P_T(f_1 - f_0)]V^2 + P_{\text{static}} \quad (3.10)$$

Let the cost for power consumption per unit is  $\beta$  and that of QoS be  $\eta$  for every one unit of waiting time of a customer. Thus the total expected cost function  $F(f_0, f_1, T)$  per customer is

$$F(f_0, f_1, T) = \beta P(f_0, f_1, T) + \eta \bar{W}(f_0, f_1, T) \quad (3.11)$$

For a numerical study, let  $\lambda = 15.2$ ,  $\mu_1 = 10.2$  and  $\mu_2 = 5.1$  so that  $m = \lceil 10.3/5.1 \rceil - 1 = 2$  and  $T = m + 2 = 4$ . Let the cost values of  $\beta$  and  $\eta$  be \$2 and \$3 respectively and  $C = 14.23$ ,  $V = 1.35$ ,  $P_{\text{static}} = 5$ . Let  $\alpha = (\mu_1 + \mu_2)/f_{\text{max}}$  as in Zhang et al. (2015) be the energy efficiency bench mark value for the CPU under consideration (which can also be determined from the standard performance evaluation corporation (SPEC)). The service rates  $\mu_1$  and  $(\mu_1 + \mu_2)$  are considered as linear functions of  $\alpha$ , such that  $\mu_1 = \alpha f_0$  and  $\mu_1 + \mu_2 = \alpha f_1$ . Values of power to be generated and the corresponding expected cost for varying service rate  $\mu_2 \in [5.1, 6.9]$  while fixing  $f_2 = 2.7$ , have been computed and reported in Table-3.

**Table-3:** Values of Power generated and Expected cost for  $f_2 = 2.7$ ,  $C = 14.23$ ,  $V = 1.35$ ,  $P_{\text{static}} = 5$ ,  $\beta = \$2$ ,  $\eta = \$3$ ,  $\lambda = 15.2$ ,  $\mu_1 = 10.2$  and  $\mu_2 \in [5.1, 6.9]$

$\mu_2$	$\alpha$	$f_1$	Power	Expected Cost
5.1	5.666667	1.8	74.65596	179.6243
5.3	5.740741	1.776774	74.06009	158.2966
5.5	5.814814	1.75414	73.42218	153.0227
5.7	5.888889	1.732075	72.78818	150.0422
5.9	5.962962	1.710559	72.15883	147.8328
6.1	6.037036	1.689571	71.53475	145.9802
6.3	6.111110	1.669091	70.91647	144.3256
6.5	6.185184	1.649102	70.30441	142.7952
6.7	6.259258	1.629586	69.69893	141.3504
6.9	6.333333	1.610526	69.10032	139.9687

Inspecting the values of  $\mu_2$ ,  $\alpha$ ,  $f_1$ , Power, and Expected Cost of Table-3, it is observed that the power and the expected cost values decrease with increase in the service rate  $\mu_2$  as expected while the generated amount of power does not exceed the maximum level 75 units and the frequency value  $f_1$  does not exceed  $f_{\text{max}} = 2.7$  GHz.

**3.3. Conclusion:** This paper obtains the results of the Geo/Geo, Geo/2, a discrete version model of the continuous time Poisson queue  $M/M_1, M_2/2$  with two parallel heterogeneous servers studied by Krishnamoorthy (1963) under queue discipline-II that minimizes the violation of the classical ‘First Come First Served (FCFS)’ queue discipline. Further a new Poisson queue model Geo/Geo+Geo/2 with an  $m$ -policy is proposed here for the two servers to yield the same steady state probability distribution of queue length and other performance measures subject to a serial connection between the servers. Numerical illustration is then provided to support the fact of ‘no violation of the FCFS rule’ through a comparison over appropriate measures.

There are advantages of availing the proposed Poisson queue model Geo/Geo + Geo/2 with an  $m$ -policy in some services providers of companies in our social and real life business applications where a service process involving both the salesman and the boss (owner) who provide services together to a customer being served if the queue length value (excluding the one being served) is beyond a threshold level, 'say  $m$ ', in order to reduce the waiting time of customers. On the other hand if the queue length is less than or equal to ' $m$ ' customers then each customer is serviced by the salesman only till the queue length crosses the level ' $m$ ' if a few more arrivals occur during the on-going service periods containing less than ' $m$ ' customers in the queue length. The boss can join with the salesman serially as and when the queue length crosses the level ' $m$ ' to provide the service jointly as a customer satisfaction measure. A remarkable feature of the serial service schedule is that the service process does not violate the classical FCFS queue discipline which is known as 'Queue Discipline-I' with an  $m$ -policy. This kind of serial configuration facility in providing service to customers could find many applications in retail shops, malls, supermarkets, offices, banks and other business outfits where heterogeneity of servers is evident owing to factors such as degree of usage, experience, age, preferences etc. It is then applied to a single computing node as the server of the queueing model Geo/Geo+ Geo/1 to obtain the power consumption and the associated expected cost for a specific set of input values.

**3.4. Acknowledgment.** The authors wish to thank the office of 'Research & Development', University of Botswana for financial support to enable them to attend the conference held from May 27–30, 2015 at the Morehouse Collage, Atlanta (USA) and present this paper. We are also thankful to colleagues who provided insight and expertise that greatly assisted the outcomes of this research.

### Appendix

Algorithm for computing the steady-state probability distribution of the queue length process of **Geo/(Geo,Geo)/2 queue** using steps 1 through 9 and of **Geo/(Geo+Geo)/2 queue** using the steps 1 through 12.

**Step 1:** Compute the largest integer ' $M$ ' contained in the ratio  $\frac{\mu_1}{\mu_2}$ , then let  $m = M - 1$

$$\text{Step 2: } \begin{aligned} w_1 &= \frac{(\lambda + \mu_1 + \mu_2) + \sqrt{(\lambda + \mu_1 + \mu_2)^2 - 4\lambda\mu_1}}{2\mu_1} \text{ and} \\ w_2 &= \frac{(\lambda + \mu_1 + \mu_2) - \sqrt{(\lambda + \mu_1 + \mu_2)^2 - 4\lambda\mu_1}}{2\mu_1} \end{aligned}$$

$$\text{Step 3: } \begin{cases} a' = (w_1 - 1)w_1, \\ b' = w_2(1 - w_2) \\ \text{and } r_n = a'w_1^n + b'w_2^n \text{ for } n = 1, 2, \dots, m \end{cases}$$

$$\text{Step 4: } d = \frac{(1 - w^{m+2})}{1 - w} [\lambda r_{m-1} - (\lambda + \mu_1 + \mu_2)r_m] - \lambda w^{m+1} \left[ \frac{\rho r_m}{1 - \rho} \right]$$

**Step 5:**  $p_{00} = [\lambda w^{m+1} + \lambda r_{m-1} - (\lambda + \mu_1 + \mu_2)r_m] / d$  and  $p_{01} = \frac{\lambda w^{m+1}}{d}$

**Step 6:**  $p_1 = p_{01} + p_{10} = p_{00}w + (1 - \frac{\mu_2}{\mu_1})p_{01}$

**Step 7:**  $p_2 = p_{11} + p_{1.10} = (p_{00} + p_{01})w^2 + \left[ \frac{\lambda + \mu_2}{\mu_1} - r_1 \right] p_{01}$

**Step 8:**  $p_{n+1} = p_{n-1.11} + p_{n.10} = (p_{00} + p_{01})w^{n+1} + [r_{n-1} - r_n]p_{01}$  for  $n = 2, 3, \dots, m$

**Step 9:**  $p_{m+2} = p_{m.11} = r_m p_{01}$  and  $p_n = \rho^{n-(m+2)}p_{m+2}$  for  $n \geq (m+2)$

### Continued to Geo/(Geo+Geo)/2 queue:

**Step 10:**  $\pi_0 = \frac{(1-\rho)(1-w)}{(1-w^{m+1}\rho)}$  and  $\pi_n = w^n \pi_0$  for  $n = 2, 3, \dots, m+1$ ; where  $w = \frac{\lambda}{\mu_1}$

**Step 11:**  $\pi_n = w^{m+1} \rho^{(n-m-1)} \pi_0$  for  $n \geq m+2$

**Step 12:**  $q_{n-1} = \frac{\pi_n}{1-\pi_0}$  for  $n = 1, 2, \dots, \infty$

## REFERENCES

- [1] F. S. Q. Alves, H. C. Yehia, L. A. C. Pedrosa, F. R. B. Cruz, and Laoucine Kerbache, *Upper Bounds on Performance Measures of Heterogeneous Queues*, Mathematical Problems in Engineering Volume 2011 (2011).
- [2] O. J. Boxma, Q. Deng and A. P. Zwart, *Waiting time asymptotics of the M/G/2 queue with heterogeneous servers*, *Queueing Systems*, 40 (2002), 5–31.
- [3] B. E. Emrah, O. Ceyda and O. Irem, *Parallel machine scheduling with additional resources: Notation, classification, models and solution methods*, *European Journal of Operational Research*, 230 (2013), 449–463.
- [4] D. Efrosinin *Controlled Queueing Systems with Heterogeneous Servers: Dynamic Optimization and Monotonicity Properties of Optimal Control Policies in Multiserver Heterogeneous Queues* 2008.
- [5] P. L. Gall. *The stationary G/G/s queue*, *Journal of Applied Mathematics and Stochastic Analysis*, vol 11, no.1, (1998), 59–71.
- [6] P. L. Gall. *The stationary G/G/s queue with nonidentical servers*, *Journal of Applied Mathematics and Stochastic Analysis*, vol. 11, no. 2, 1998, 163–178.
- [7] K. W. Grassmann and Q. Y. Zhao, *Heterogeneous multi server queues with general input*, Tech. rep., University of Winnipeg, 2004.
- [8] H. Gumbel. *Waiting lines with heterogeneous servers*, *Operations Research*, vol. 8, no. 4, 1960, 504–511.
- [9] J. H. Kim, H. S. Ahn and R. Righter *Managing queues with heterogeneous servers* *Journal of Applied Probability*, 48, No 2 (2011), 435–452.
- [10] Hoksad, P., 1978. *Approximation for the M/G/m Queue*. *Journal of Operation Research*, 26, 511–523.
- [11] B. Krishnamoorthi, *On Poisson queue with two heterogeneous servers*, *Operations Research*, 2, No 3 (1963), 321–330.
- [12] Krishnamoorthy, B. and Sreenivasan, S. (2012). *An M/M/2 queue with Heterogeneous Servers including one with Working Vacations*. *International Journal of Stochastic Analysis*, Hindawi publishing company, doi 10.1155/2012/145867.
- [13] J. Medhi, *Stochastic Models in Queueing Theory*, Academic Press, California (2003)
- [14] V. P. Singh. *Markovian queues with three heterogeneous servers*, *AIIE Transactions*, vol. 3, no. 1, 1971, 45–48.



- [15] V. P. Singh. *Two-server Markovian queues with balking: Heterogeneous vs. homogeneous servers*, Operations Research, vol. 18, no.1, 1970, 145–159,
- [16] R. Sivasamy, O. A. Daman, and S. Sulaiman, *An M/G/2 Queue subject to a minimum violation of the FCFS queue discipline*, European Journal of Operational Research, 240 (2015) 140–146.
- [17] Tijms, H. C., Vaan Hoorn, M. H., and Federgruen, A., 1981. *Approximation for the steady state probabilities in the M/G/C queue*. Advances in Applied Probability, 13(1): 186–206.
- [18] X. Zhang, J. Wang, and Tien Van Do, *Threshold properties of the M/M/1 queue under T-policy with applications*, Applied Mathematics and computation 261, (2015) 284–301.