Bibliotheque de Humanisme et Renaissance | ISSN : 0006-1999
Volume 81, Issue 3, 2021

Level Dependent Quasi-Birth and Death Processes of $M/(M_1,M_2)/2/(B_1,B_2)$ queues with Stalling

N. Paranjothi
Department of Statistics, Annamalai University, Annamalai nagar, Tamil Nadu, India

G. Paulraj
Department of Statistics, Annamalai University, Annamalai nagar, Tamil Nadu, India

Abstract- This Paper analyses a two server $M/(M_1,M_2)/2/(B_1,B_2)$ queueing system with stalling where customers are served by one fast server and one slow server. The servers are allowed to work in parallel. It has provided a finite buffer B $_1$ of size `K $<\infty$ ' to stall customers in queue-1 which is meant to feed the fast sever only. There is one more buffer B $_2$ of infinite capacity, called the waiting space to accept further arrivals in queue-2 when the buffer B $_1$ is full. The primary task of B $_2$ is to feed customers to queue-1 and to the slow server as and when it is warranted. An arriving customer who finds the queue-1 is full and queue-2 is empty joins the slow server. If the queue-2 is non-empty at a time epoch when the slow server finishes a service, he accepts a customer from the head of the queue-2. Both arrival and service rates of customers are assumed to be state dependent parameters for the first (K+2) states starting from state `0'. Formulating the queue length (queue +service) process of the whole system as a QBD processes, steady state results to state probabilities and mean queue length have been obtained using matrix-analytical methods.

Keywords- Quasi-Birth and Death Processes, Fast server, Slow server, Stalling and Level dependent.

1. INTRODUCTION

Let $X_1(t)$ be the number of customers present with Server-1 and in queue-1 at time t>0. Also let $X_2(t)$ for t>0 be the number of customers present with Server-2. Thus X $_1$(t)+ X $_2$(t) gives the total number of customers present in the whole system at time t>0. Further the vector process { X(t)=(X $_1$(t), X $_2$(t): t≥0) } is a continuous time non-homogeneous Markov process on the two-dimensional space { (n,j): n=0, 1,2,… and j=0,1 } that is portioned into levels L(0), L(1), … where

$$L(n)=\{(n,0),(n,1)\} \quad for \quad n=0,1,\ldots,K,(K+1),(K+2), \tag{1}$$

Assumptions: The transitions out of the state {(n,j):n=0, 1,2,{…} and j=0,1 } is dependent on the level L(n). Arrivals occur according to a Poisson process with mean arrival rate λ_n. Servers are working with exponential service rates μ_1(n) and μ_2(n), (μ $_1$(n) > μ_2(n)) respectively. Two buffers B1 and B2 are installed to accommodate customer arriving in queue-1 and queue-2 respectively: Buffer B1 is bounded by a maximum of 'K' which stalls customers in queue-1 and if the buffer B1 is full then the buffer B2 which is unbounded receives further arrivals to form queue-2. Thus queue-1 is the queue of stalling customers who must go to the faster server only, to get served. However queue-2 feeds both queue-1 and the slow server whichever could first accept the head-of-the-line customer of queue-2. Each customer who arrives when the system is idle is served by the fast server. If the fast server is busy at the time of a customer's arrival, then that customer joins queue-1 if it has less than K customers, is served by the slow server if queue-1 has exactly K customers, or joins queue-2 if the system has more than K customers. After finishing a service, fast server chooses the next customer from queue-1 (if there is one; otherwise it idles).

The customer standing at the head of queue-2 (if there is one) leaves queue-2 and joins queue-1 provided it has less than K customers at that time point. When the slow server finishes a service, it serves the next customer in queue-2 (if there is one; otherwise it idles). This system informs its customers that one server is faster than the other and all arriving customers have to stall when the fast server is busy. This paper provides some applications with numerical values to a few performance measures such as the expected queue length, the probability that each server is busy etc., and illustrates how these values are being used to determine the optimal value of K and the service rate for the slow server.

This queueing system is denoted as M/(M1(n), M2(n))/2/(B1, B2) and is solved by matrix-analytic methods. It is noted that if the positive integer K is chosen such that server-1 is K times faster than server-2, then the First-come First-served (FCFS) policy is not violated or insignificantly violated (see Krishnamoorthy, 1963). The steady state behavior of queue length and its characteristics are then discussed for cases of homogeneous QBD to check if they agree with the steady state characteristics of the system that was investigated by Sivasamy et al. (2015 and 2016) and Kim et al. (2011). Replacing the value of μ_2 by 'zero' in the steady state results of M/(M1,M2)/2/(B1,B2) queues, Queue length distribution of the M/M/1 system under T-policy studied by Xuelu Zhang et al. (2015) is obtained. Rubinovitch (1985a, 1985b) studied a Markovian queueing system with two channels (one fast and one slow) and stalling for informed and uninformed customers and computed steady state probabilities and characterized a few policies for overall optimization. Further, assuming that the heterogeneous channels operate under a general discipline, Abou-El-Ata and Shawky (1999) discussed in detail when and how to discard the slow server based on simple conditions. Again Fabricio Bandeira Cabari (2005) considered the slow server problem of M/M/n queues for uninformed customers and established a threshold value λ^* on the arrival rate λ and recommends that if $\lambda < \lambda^*$ the slow server should not be removed. Gumbel (1960) has initiated the first work on Poisson queues with heterogeneous servers and measured the errors occurring due to an assumption that all service rates are equal. Singh (1971) established an optimal combination of service rates to minimize the cost associated with a few performance measures of M/Mi/3 queues.

This paper is organized as follows: Section 2 is devoted to study the QBD processes of the M/(M1,M2)/2/(B1, B2) queueing model with stalling and to obtain the steady state probability distribution of queue length process. Section 3 provides an application with a view to solve an optimization problem through numerical illustrations. Section 4 gives a summary and scope for further study.

2. $M/(M_1,M_2)/2/(B_1,B_2)$ QUEUE WITH STALLING

2.1 MODIFIED LOSS SYSTEM: $M/\left(M_1,M_2\right)/2\ /\left(\left(K+1\right),B_1\right)$ queues with stalling and no waiting space.

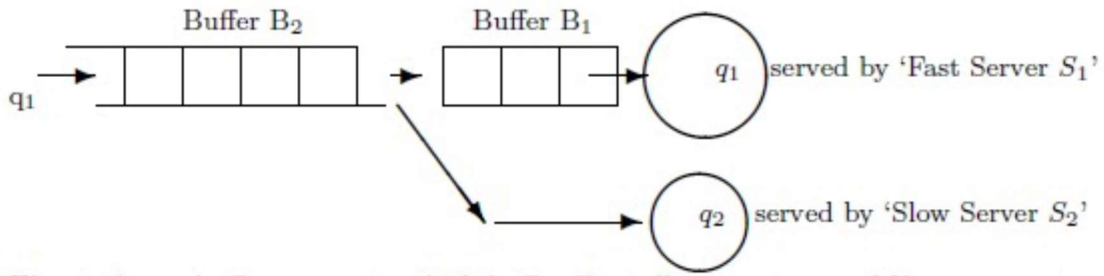

Fig. 1 A two-buffer queue, in which buffer B_1 stalls a maximum of K customers to be served by server S_1. Infinite buffer B_2 accommodates other customers when the system size exceeds $(K + 2)$.

If the buffer B2 is removed from the Figure 1, the result is the modified $M/\left(M_1,M_2\right)/2\ /\left(K+1\right)$ (loss) system or a two-server system with stalling and no waiting space in queue-2. Here, for n=0, 1, 2, … ,(K+1), L(n) is referred to as level 'n' and each level has two states '0 and 1' only. Thus the modified vector process defined by X(t)=(X1(t), X2(t): t≥0) constitutes a finite non-homogeneous QBD process on the finite state space S0 =U0≤n≤(K+1)L(n) with infinitesimal generator Q1 i.e.,

$$\mathbf{Q_1} = \begin{array}{c|ccccccc} & L(0) & L(1) & L(2) & L(3) & ... & L(K) & L(K+1) \\ \hline L(0) & \mathbf{A}_1^{(0)} & \mathbf{A}_0^{(0)} & 0 & 0 & ... & 0 & 0 \\ L(1) & \mathbf{A}_2^{(1)} & \mathbf{A}_1^{(1)} & \mathbf{A}_0^{(1)} & 0 & ... & 0 & 0 \\ L(2) & 0 & \mathbf{A}_2^{(2)} & \mathbf{A}_1^{(2)} & \mathbf{A}_0^{(2)} & ... & 0 & 0 \\ ... & ... & ... & \ddots & \ddots & \ddots & ... & ... \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ L(K) & 0 & 0 & 0 & 0 & \ddots & \mathbf{A}_1^{(K)} & \mathbf{A}_0^{(K)} \\ L(K+1) & 0 & 0 & 0 & 0 & ... & \mathbf{A}_2^{(K+1)} & \mathbf{A}_1^{(K+1)} \end{array} \tag{2}$$

for n=0,1, . . . ,k

$$A_0^{(n)} = \begin{pmatrix} \lambda_0 & 0 \\ 0 & \lambda_0 \end{pmatrix} \tag{3}$$

for n=0,1, . . . ,K and (K+1),

$$A_1^{(0)} = \begin{pmatrix} -(\lambda_0) & 0 \\ \mu_2(0) & -(\lambda_0+\mu_2(0)) \end{pmatrix}, A_1^{(n)} = \begin{pmatrix} -(\lambda_n+\mu_1(n)) & 0 \\ \mu_2(n) & -(\lambda_n+\mu_1(n)+\mu_2(n)) \end{pmatrix},$$

$$A_1^{(k+1)} = \begin{pmatrix} -(\lambda_{k+1}+\mu_1(k+1)) & \lambda_{k+1} \\ \mu_2(k+1) & -(\mu_1(k+1)+\mu_2(k+1)) \end{pmatrix} \tag{4}$$

and for n=1,2, . . . , (k+1),

$$A_2^{(n)} = \begin{pmatrix} \mu_1(n) & 0 \\ 0 & \mu_1(n) \end{pmatrix} \qquad (5)$$

Let the joint probability function of the level dependent QBD process {X(t)} of the modified M/(M1,M2)/2 / (K+1) system be π_{ij}(t)=P(X_1(t)=i, X_2 (t)=j) for i$\in$ $\{$0, 1,2,…,(K+1) $\}$ and j$\in$ $\{0,1\}$. It is then easy to verify that $\pi_{ij} = \lim_{t\to\infty}$ P(X_1(t)=i, X_2 (t)=j) always exists. Let $\mathbf{\Pi_n}$=(π_{n0}, π_{n1}) be a row vector of order 2 for n=0, 1, 2, ..., (K+1). Denote the stationary probability vector $\mathbf{\Pi}$ of the QBD process { X(t)} over the levels L(0), L(1) … , L(K+1) by $\mathbf{\Pi}$=($\mathbf{\Pi_0}, \mathbf{\Pi_1}, \mathbf{\Pi_2}$,…, $\mathbf{\Pi_{K+1}}$). Let $\{$ a_n = $\lim_{t\to\infty}$ P(X_1(t)=i: n=0, 1, …, (K+1) $\}$ be the marginal steady-state queue length distribution. It is then computed by

$$a_n = \sum_{j=0}^{1} \pi_{nj}, \quad \text{n=0,1,2,...(K+1)} \qquad (6)$$

Let $b_{\text{j}}(t) = P(X_2(t) = j)$ be the marginal probability that the slow server is in state j at time t ≥0 for j=0, 1. Then it can be verified that $\lim_{t\to\infty} b_j(t) = b_j$ exists. Thus,

$$b_j = \sum_{n=0}^{K+1} \pi_{nj} \quad , \quad \sum_{j=0}^{1} b_j = 1 \qquad (7)$$

Remark: A quasi-birth-and-death (QBD) process is a bivariate Markov process which is an extension of the standard birth-and-death process. When the transitions of a QBD process are independent of the states, it is termed a homogeneous or level-independent QBD process. Otherwise, it is termed as inhomogeneous or level-dependent QBD (LDQBD) process.

2.2 LINEAR LEVEL REDUCTION

This section applies the linear level reduction method and develops a computational algorithm for the evaluation of the stationary probability vector $\mathbf{\Pi}$ of the non-homogeneous QBD process **X(t)** under study. The procedure developed by Latouche and Ramasamy (1999) for transition matrices of finite quasi-birth and death processes (QBDs) is extended to the respective generators of the same QBDs. This extension needs to formulate restricted QBD process $\left\{X^i(t)\right\}$ on the space $S_i = U_{i\le n\le(k+1)} L(n)$ by removing the first `i' levels, one level at each step where i=1, 2,…, K.

Lemma 1

For i=0, 1, 2, . . . , K, the generator matrix of the restricted process $\left\{X^i(t)\right\}$on the set $S_i = U_{i\le n\le(k+1)} L(n)$ is given by

$$\mathbf{Q}^{(\mathrm{i})} = \begin{pmatrix} \mathbf{C}_{(\mathrm{i})} & \mathbf{A}_0^{(\mathrm{i})} & & 0 & . & . & 0 & 0 \\ \mathbf{A}_2^{(1)} & \mathbf{A}_1^{(1)} & \mathbf{A}_0^{(1)} & 0 & . & . & 0 & 0 \\ 0 & \mathbf{A}_2^{(2)} & \mathbf{A}_1^{(2)} & \mathbf{A}_0^{(2)} & . & . & 0 & 0 \\ 0 & 0 & \mathbf{A}_2^{(3)} & \mathbf{A}_1^{(3)} & \mathbf{A}_0^{(3)} & . & 0 & 0 \\ . & . & . & . & \ddots & \ddots & \ddots & . \\ 0 & 0 & 0 & 0 & . & . & \mathbf{A}_2^{(\mathrm{K+1})} & \mathbf{A}_1^{(\mathrm{K+1})} \end{pmatrix} \tag{8}$$

and $\mathbf{Q}^{(\mathbf{K+1})} = \mathrm{C}_{K+1}$, where the matrix C_i, for i=1, 2, . . . , K , records the rates of returning to the level \`i' before reaching the level \`i+1' starting from the level \`i' and are recursively defined as follows:

$$C_0 = \mathbf{A}_1^{(0)} \;,\; \mathbf{C}_\mathrm{i} = \mathbf{A}_1^{(\mathrm{i})} + \mathbf{A}_2^{(\mathrm{i})}(-\mathbf{C}_{\mathrm{i-1}})^{-1}\mathbf{A}_0^{(\mathrm{i-1})} \qquad \text{for} \quad \text{i=1, 2, . . . , K, (K+1)} \tag{9}$$

It is noticed that the matrix $(-\mathbf{C}_\mathrm{i}^{-1}\mathbf{A}_0^{(\mathrm{i})})$ records the first passage probabilities from level $L(i)$ to $L(i+1)$.

Theorem 1 Suppose the stationary probability vector $\mathbf{\Pi} = (\; \mathbf{\Pi_0}, \mathbf{\Pi_1}, \mathbf{\Pi_2}, \;\ldots,\; \mathbf{\Pi_{K+1}} \;)$ is determined by

$$\mathbf{\Pi_{K+1}} \;\; \mathbf{C_{K+1}} = \mathbf{0} \;.\; \Pi_i = \mathbf{\Pi_{i+1}} \;\; \mathbf{A}_2^{(\mathrm{i+1})} \;\; (-\mathbf{C}_\mathrm{i})^{-1} \quad \text{for} \quad \text{i=K,K-1, . . . ,2,1, and 0} \tag{10}$$

Let e=[1,1] be a column vector of order 2 of unit elements. Then

$$\sum\nolimits_{0 \le n \le (k+1)} \Pi_n \;\; \mathbf{e} = 1 \tag{11}$$

Proof: Modifying the arguments of Latouche and Ramasamy (1999) to the case of generators of finite QBDs, Lemma 1 and Theorem 1 can be proved. Equation (10) characterizes the stationary distribution $\mathbf{\Pi} = (\; \mathbf{\Pi_0}, \mathbf{\Pi_1}, \mathbf{\Pi_2}, \;\ldots,\; \mathbf{\Pi_{K+1}})$ up to a multiplicative constant which is determined by (11). Linear reduction algorithm for the computation of the stationary probability vector $\mathbf{\Pi}$.

Now, for a given set of $(\lambda_n, \mu_1(n), \mu_2(n))$ values, one can compute numerical values of each π_{n1} and π_{n0} , for n=0, 1, 2,. . . ,(K+1). The mean number of customers in the system is

$$L_{(K+1)} = b_1 + \sum_{n=0}^{K+1} n \;\mathrm{a}_\mathrm{n} \tag{12}$$

Theorem 2 If $\mu_2(n) = 0$ for n=0, 1, 2,. . . ,K+1 then the sequence $\{\mathrm{a}_n$:n=0 ,1,2, . . . , (K+1)} of probability values defined by (6) and the set $\{b_j\}$ defined by (7) are given by

$$a_0 = [1 + \sum_{b_0=1}{}_{j=1}^{k+1} \prod_{\mathbf{i=1}}^{\mathbf{j}} \frac{\lambda_{i-1}}{\mu_1(i)}]^{-1}, \;\; a_n = \prod_{b_1=0}{}_{j=1}^{n} \frac{\lambda_{j-1}}{\mu_{1(j)}} a_0 \tag{13}$$

Proof: Substituting \$ $\mu_2(n) = 0$ for n=0, 1, 2,{\dots},K in (8), it is found that the matrix (-C $_n$)$^{-1}$ A$_0^{(n)}$ becomes an identity matrix of order 2 and hence each of $\mathbf{C_n} = \mathbf{A_1^{(n)}} + \mathbf{A_2^{(n)}}$ becomes a diagonal matrix of order 2 with diagonal elements $'-\lambda_n'$ while

$$C_{K+1} = \mathbf{A_1^{(K+1)}} + \mathbf{A_2^{(K+1)}}(-\mathbf{C_K})^{-1}\mathbf{A_0^{(K)}} = \begin{pmatrix} -\lambda_{k+1} & \lambda_{k+1} \\ 0 & 0 \end{pmatrix} \tag{14}$$

Solving of $\mathbf{\Pi_{K+1}} \;\; \mathbf{C_{K+1}} = \mathbf{0}$ and $\pi_{(K+1)0} + \pi_{(K+1)1}$=1, leads to the fact that

$$\begin{pmatrix} \pi_{(k+1)0} \\ \pi_{(k+1)1} \end{pmatrix} = \begin{pmatrix} -\lambda_{(k+1)} & 1 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{15}$$

$\pi_{(K+1)0}$=0 and $\pi_{(K+1)1}$=1. Now computing vectors of (9)

i.e. $\Pi_i = \Pi_{i+1} \, A_2^{(i+1)} \, (-C_i)^{-1}$ *for* $i = K, K-1, ..., 2, 1$ and 0 for $i = K, K-1, ..., 2, 1$ and 0 recursively, it can be shown that

$$\pi_{k0} = 0 \qquad \pi_{k1} = \frac{\mu_1(k+1)}{\lambda_k}$$

$$\pi_{(k-1)0} = 0 \qquad \pi_{(k-1)1} = \frac{\mu_1(k)}{\lambda_{k-1}} \pi_{k1} = \frac{\mu_1(k+1)}{\lambda_k} . \frac{\mu_1(k)}{\lambda_{k-1}}$$

$$\pi_{(k-2)0} = 0 \qquad \pi_{(k-2)1} = \prod_{j=k-2}^{k} \frac{\mu_1(j+1)}{\lambda_j}$$

$$\pi_{(k-3)0} = 0 \qquad \pi_{(k-3)1} = \prod_{j=k-3}^{k} \frac{\mu_1(j+1)}{\lambda_j}$$

………………… …………………………

$$\pi_{00} = 0 \qquad \pi_{01} = \prod_{j=0}^{k} \frac{\mu_1(j+1)}{\lambda_j} \tag{16}$$

Multiplying each π_{nj} , j=0,1 of (16) by $\pi_{01} \prod_{j=1}^{K+1} \frac{\lambda_{j-1}}{\mu_1(j)}$, and simplifying the result, we get

$$\pi_{01} = \pi_{01}$$

$$\pi_{n1} = \pi_{01} \cdot \prod_{j=1}^{n} \frac{\lambda_{(j-1)}}{\mu_{1(j)}} \quad \text{for } n = 1, 2, ..., K \tag{17}$$

$$\pi_{(k+1)1} = \pi_{01} \prod_{j=1}^{k+1} \frac{\lambda_{j-1}}{\mu_{1(j)}}$$

Remark:

Remark : In this case $a_n = \pi_{n1}$ for n=0,1, . . . ,(k+1).

Lemma 2 The steady state probability b_1 of Eq. (7) can also be obtained by a direct argument as

$$b_1 = \frac{\lambda}{\mu_2} \pi_{(K+1)0} \tag{18}$$

Proof: The time dependent marginal probability $b_j(t) = P(X_2(t) = j)$ that the slow server is in state j at time t ≥0 for j=0, 1 satisfies

$$\frac{d}{dt} b_1(t) = -\mu_2 \, b_1(t) + \lambda \pi_{(K+1)0}(t) \tag{19}$$

Result of Eq. (18) follows from Eq. (19) since the $\lim_{t \to \infty} b_j(t) = b_j$ exists. For $\mu_2(n) = 0$ and n=0, 1, 2, . . . ,K+1 then the mean queue length of the state dependent M/M/1/(K+1) queue is obtained as below:

$$L^0_{(K+1)} = \sum_{n=0}^{K+1} n \, \mathrm{a_n}$$

Theorem 3

Suppose $\lambda_n = \lambda$, $\mu_1(n) = \mu_1$ and $\mu_2(n) = \mu_2$ =0 for n=0, 1, … (K+1) the sequence $\{a_n\}$ tends to the (state independent) queue length (queue + service) distribution of the M/M/1/K+1 queue,

i.e.,

$$a_n = \frac{(1-\rho_1)\rho_1^n}{1-\rho_1^{(K+2)}}, \quad \text{n=0,1,2,...(K+1)} \tag{20}$$

Proof: Substitutions of $\lambda_n = \lambda$, $\mu_1(n) = \mu_1$ and $\mu_2(n) = \mu_2 = 0$ for n=0, 1, . . . ,(K+1) in (15) ensure that

$$\pi_{n1} = \prod_{j=1}^{n} \frac{\lambda_{(j-1)}}{\mu_1(j)} = \pi_{01} \; (\frac{\lambda}{\mu_1})^n \quad for \quad n = 1, 2, \ldots, K$$

Implementation of (14), $\rho_1 = \dfrac{\lambda}{\mu_1}$, and the normalizing condition $\sum_{n=0}^{K+1} a_n$ =1, establish the statement (16). Further it follows that

$$a_{K+1} = \pi_{(K+1)1} = \frac{(1-\rho_1)\rho_1^{K+1}}{1-\rho_1^{(K+2)}}, \quad \text{i=0,1,2,...(K+1)} \tag{21}$$

We also note that $X_1(t)$ represents the queue length process in an M/M/1/(K+1) loss system (see Medhi (1975), Prabu (1965), and Sivazlian (1975)) and $\{a_n = \lim_{t\to\infty} P[X_1(t) = n : n = 0,1,2,...,(k+1)]\}$ is the marginal steady-state queue length distribution as found in Baily(1957). It is interesting to remark that all these results $\{\pi_{ij}; i = 0,1,2,...,(k+1)$ and j=0,1} , {b$_j$;j=o,1} and $\{a_n : n = 0,1,...,(k+1)\}so$ far derived by the matrix analytic methods agree with the corresponding expressions obtained by Sivasamy(2016) through scalar analytical methods.

2.3 STEADY STATE RESULTS FOR $M/(M_1,M_2)/2(B_1,B_2)$ QUEUE WITH STALLING

Assume that $\lambda_n > 0$, $\mu_1(n) > 0$, and $\mu_2(n) > 0$, for n=0,1,2,. . . ,(K+1). Let $\lambda = \lambda_n = \lambda_{K+1}$, $\mu_1 = \mu_1(n) = \mu_1(K+1)$, and $\mu_2 = \mu_2(n) = \mu_2(K+2)$ for n=(K+2),(K+3), . . . , ∞ .

Let r = μ_2/μ_1, $\mu = \mu_2 + \mu_1$, $\rho = \lambda/\mu$ and $\rho_1 = \lambda/\mu_1$. Let $\rho < 1$ and we continue with the same notations as in the M/(M_1,M_2)/2 /((K+1),B_1) modified system. Replacing the notation $\mathbf{X(t)} = (X_1$ $(t), X_2(t) : t \geq 0)$ defined in the preceding modified system by $Y(t) = (Y_1(t), Y_2(t) : t \geq 0)$ to monitor the transitions of $M/(M_1,M_2)/2/(B_1,B_2)$ queue with stalling described in Figure 1, this section presents the analysis of the underlying queueing process $\{\mathbf{Y(t)}\}$ on the full two dimensional state space {L(n); n=0,1,2,…,(K+1)} ⋃ { (K+2),(K+3), . . . , ∞ } through a linking mechanism.

Let the sequence of steady state probabilities be $\{p_i = \lim_{t\to\infty} P(Y_1(t)) = i : i = 0,1,...,\}$ and $\alpha_{ij} = \lim_{t\to\infty} P(y_1(t) = i, y_2(t) = j)$ for $i \in \{0,1,2,\ldots,(k+1)\}$ and $j \in \{0,1\}$. For this extension, there exists a proportionality constant, say $\beta = \beta$(K), expected to be a function of K such that

$$\alpha_{ij} = \beta\, \pi_{ij} \;\; for\, i \in \{0,1,2,\ldots(K+1)\}\, and\; j \in \{0,1\} \tag{22}$$

$$p_i = \alpha_{i0+}\alpha_{i1=}\beta\, a_i \;\; for\, i \in \{0,1,2,\ldots(K+1)\} \tag{23}$$

$$p_i = \alpha_{(K+1)1}\, \rho^{i-(K+1)} \;\; for\, i \geq (K+2) \tag{24}$$

Hence the normalizing condition $\sum_{i=0}^{\infty} p_i$ =1$ yields that

$$\beta = \frac{1-\rho}{1-\rho+\rho\,\pi_{(\mathrm{K+1})1}} \qquad (25)$$

Further fraction of the time `Slow Server is busy', say D_1, is given by

$$D_1 = \beta[b_1 + \frac{\pi_{(\mathrm{K+1})1}\,\rho}{1-\rho}] \qquad (26)$$

The mean number L_3 of customers of the system can now be calculated which is

$$L_3 = \sum_{n=0}^{\infty} n\,\mathrm{p_n} + D_1 \qquad (27)$$

Further, D_2 the probability that fraction of the time the fast server is busy is

$$D_2 = \beta(\sum_{n=0}^{K+1} \pi_{n1}) + \sum_{i=K+3}^{\infty} p_i \qquad (28)$$

Substituting for b_1 of (7) and $\pi_{(K+1)1}$ in the value D_1 of (26) one can obtain the same expression as in D_2 of (28) since each gives the probability of the same event. Now the mean number L_3 of customers inclusive of those are being served is $\mathrm{L}_3 = \sum_{n=0}^{\infty} n\,\mathrm{p_n} + D_1$, which simplifies to

$$L_3 = \beta\,[\,\mathrm{L}_{(\mathrm{K+1})} + \frac{\rho\,\pi_{(\mathrm{K+1})1}\{(K+2)(1-\rho)+\rho\}}{(1-\rho)^2}] \qquad (29)$$

Lemma 3 If $\lambda_n > 0$, $\mu_1(n+1) > 0$,for n=0, 1, 2, . . . ,(K+1) and $\mu_2(n) = 0$ for n=0, 1, 2, . . .,(K+1) and $\lambda_n = \lambda_{K+1} = \lambda > 0$, $\mu_1(n) = \mu_1(K+2) = \mu_1 > 0$, $\mu_2(n) = \mu_2(K+1) = \mu_2 > 0$ for $n = (K+2), (K+3), \ldots, \infty$ then substitution of $b_1 = 0$ into L(k+1) of (12) and (29) gives the mean queue length of the state dependent $M/M_{1(n)} + M_{2(n)}/1$ queue with a T(=K+2) policy as

$$L_2 = \beta[\sum_{n=0}^{K+1} n\,\mathrm{a_n} + \frac{\rho_1 \quad a_{K+1}\{(K+2)(1-\rho)+\rho\}}{(1-\rho)^2}] \qquad (30)$$

Lemma 4 If $\lambda_n > 0$, $\mu_1(n) > 0$, $\mu_2(n) = 0$ for n=0, 1, 2, . . . ,(K+1) and $\lambda_n = \lambda_{K+1} = \lambda > 0$, $\mu_1(n) = \mu_1(K+1) = \mu_1 > 0$, $\mu_2(n) = \mu_2$ =0 for n=(K+2),(K+3), . . . ,∞ then substitution of $b_1 = 0$ into L_{k+1} of (12) and (30) gives the mean queue length of the state dependent $M/M_{1(n)}/1$ queue as

$$L_1 = \beta[\sum_{n=0}^{K+1} n\,\mathrm{a_n} + \frac{\rho_1 \quad a_{K+1}\{(K+2)(1-\rho_1)+\rho_1\}}{(1-\rho_1)^2}] \qquad (31)$$

It is expected that the mean values L_1 , L_2 and L_3 stated by (29),(30)and (31) respectively satisfy an intra-relationship $L_1 > L_2 > L_3$. Using this inequality statement, it is possible to do a comparison and draw conclusion with regard to advantages of preferring one model over the remaining models.

Consider the state independent $Y_1(t)$ process representing the queue length process of an M/M/1 system.

Lemma 5 Let $\lambda_n = \lambda$, $\mu_1(n) = \mu_1$ and $\mu_2(n) = \mu_2 = 0$ for n=0, 1, … (K+1), in (29) through (31) then the sequence $\{p_n\}$ tends to the queue length (queue +service) distribution of the M/M/1 queue.

Proof: If $\lambda_n = \lambda$, $\mu_1(n) = \mu_1$ and $\mu_2(n) = \mu_2 = 0$ for n=0, 1, … (K+1), value of $\pi_{(K+1)1}$ is given by (21). Using (21) in the β value (25), it reduces to

$$\beta = 1 - \rho_1^{K+2} \tag{32}$$

Substituting the β of (32) in those values of p_i of (23) and (24) for $i = 0,1,2,\ldots,\infty$, this lemma is proved, i.e.,

$$p_n = \lim_{t\to\infty} P(Y_1(t) = n : n = 0,1,\ldots,\} = (1-\rho_1)\rho_1^n \tag{33}$$

Expected number (queue +service) of customers (L_1 say) of the M/M/1 system is then given by

$$L_1 = \frac{\rho_1}{1-\rho_1} \tag{34}$$

2.3.1 RESULTS FOR $M / M_1 + M_2 / 1$ QUEUES WITH A T (=K+2) POLICY

Xuelu Zhang et al., (2015) considered a (state independent) single server queueing system with a threshold control policy and studied the monotonicity, convexity or concavity properties of the key performance measures of the system. If the number of customers in the system is less than a threshold value T say, the service rate is set in a low value $\mu_1 > 0$ and can be switched to a high value $(\mu_1 + \mu_2)$ $> \mu_1 > 0$ once the number of customers reaches the threshold.

Let q_n be the steady state probability to find `n' customers in the system subject to the condition $\rho = \lambda / (\mu_1 + \mu_2) < 1$ for $n = 0,1,2,\ldots,\infty$. The queue length distribution of the $M / M_1 + M_2 / 1$ queue under T-policy due to Xuelu Zhang et al.,(2015) can be obtained by assigning $\lambda_n = \lambda$, $\mu_1(n) = \mu_1$ and $\mu_2(n) = 0$ for n=0, 1, … (K+1) but $\mu_2(n) = \mu_2 > 0$ for n=(K+2),(K+3),. . . ,∞ and T=(K+2) in those π_{ij} values of the modified Loss system of $M / (M_1, M_2)/2 / ((K+1), B_1)$ queues with stalling and no waiting space when $\lambda \neq \mu_1$, $\rho = \dfrac{\lambda}{\mu_1 + \mu_2}$ and $\rho_1 = \dfrac{\lambda}{\mu_1}$:

$$q_0 = \frac{(1-\rho)(1-\rho_1)}{(1-\rho)(1-\rho_1^{(K+2)}) + \rho\, \rho_1^{(K+1)}(1-\rho_1)} \tag{35}$$

$$q_i = q_0\ \rho_1^{\,i}\ \ for\ i \in \{0,1,2,\ldots(K+1)\} \tag{36}$$

$$q_{(K+1+i)=}q_0\ \rho_1^{(K+1)}\ \ \rho^i\ \ for\ \ i > 0 \tag{37}$$

Expected number (queue + service) E(Q)= L_2 say, of customers in the system of $M\ /\ M_1 + M_2\ /1$ queue with a T =(K+2) policy is

$$L_2 = q_0\,\rho_1\left(\frac{(1-\rho_1^{K+1})}{(1-\rho_1)^2} + \frac{(\rho_1^{K+1})}{(1-\rho)^2} + \frac{(K+1)\rho_1^{K+1}(\rho-\rho_1)}{(1-\rho_1)(1-\rho)}\right) \tag{38}$$

When μ_2=0 further for n=(K+1), (K+2), . . . , ∞, value of L_2 of (38) reduces to L_1 of (34).

3. APPLICATIONS: OPTIMIZATION PROBLEMS

There are several two-server applications where customers know the reasons for stalling of customers. Such scenarios can be seen with agents in an airport check-in, tellers in a bank, cashiers in a supermarket, etc. The proposed model $M\ /\ (M_1, M_2)\ /2/(\mathrm{B}_1,\mathrm{B}_2)$ has a variety of applications in computer communication networks and local area networks (LANs) which have channels to transmit packets of data to different destinations.

3.1 BEST VALUE OF K FOR A GIVEN INPUT SET (λ, μ_1, μ_2) *WHILE* $\rho = \dfrac{\lambda}{\mu_1 + \mu_2} < 1$

Lin and Kumar (1984) investigated a generalized type of M/M/2 queueing system with two heterogeneous servers and have shown that the optimal policy which minimizes the mean sojourn time of customers in the system is of a `threshold policy', which is equivalent to fixing the optimum buffer size K(finite) of the buffer B_1 of the model $M\ /\ (M_1, M_2)\ /2/(\mathrm{B}_1,\mathrm{B}_2)$ with stalling. For an application, assume that the two servers at a petrol filling station follow all the regulations of the queueing system $M\ /\ M_1,\ \ M_2\ /2/(\mathrm{B}_1, \mathrm{B}_2)$ discussed so far. Further all customers who come to this service center have the chance of paying the fuel price either by cash or with a credit card. Following points are informed to the customers:

i) Buffer B_1 area is inside the petrol selling and filling station which accepts credit card and it stalls customers in queue-1 of size K like 5 or 6.

ii) Buffer B_2 area is located adjacent to the B_1 stall of the filling station (which does not accept credit card) where customers form queue-2 like that of queue of vehicles waiting at the national highway of a country (adjacent to the filling stations) without disturbing the road traffic.

iii) No customer can leave the system after entering the Queue and there is no jockeying in the system.

It aims to fix the optimum buffer size K_0 of buffer B_1 by comparing the mean queue lengths, L_3 of the $M/(M_1, M_2)/2/(B_1, B_2)$ system ,state dependent mean queue length L_2 of $M/M_1+M_2/1$ with a T =(K+2) policy given by (26) and the state dependent mean queue length L_1 i.e. (27) of M/M/1 system. An optimization problem can be stated as `minimize K value (from below) and find K_0 subject to the constraint $L_1^{(n)} > L_2^{(n)} > L_3$ for n $\leq K_0$, corresponding to a given set of input ($\lambda_n, \mu_1(n), \mu_2(n)$) values'. For obtaining such a maximum value K_0 of K, required condition is

$\rho = \dfrac{\lambda_{K+1}}{\mu_1(K+1)+\mu_2(K+2)} < 1$ for n $\geq$ (K+2) . For an illustration, one set of input values is selected by λ_n =10.0375-(n/1.75), $\mu_1(n)$=12.5+(n/40.0), $\mu_2(n)$=1.25 and the corresponding results computed and recorded in Table 1:

Table -1: Input values on $\lambda_n = 10.0375-(n/1.75)$, $\mu_1(n) = 12.5+(n/40.5), \mu_2(n) = 1.25$ to check if $L_1 >$ $L_2 > L_3$ n=0,1,2, . . . ,(K+1) and K≥1 is an integer variable.

K	λ (k+1)	μ_1 (k+1)	μ_2 (k+1)	L_1	L_2	L_3	D_1
1	8.8946428	12.550	1.25	2.580	2.163	2.208	0.6601
2	8.3232145	12.575	1.25	2.256	2.035	2.036	0.4598
3	7.7517862	12.600	1.25	2.075	1.965	1.946	0.2924
4	7.1803575	12.625	1.25	1.979	1.298	1.906	1.1713
5	6.6089292	12.650	1.25	1.932	1.909	1.892	1.0924
6	6.0375004	12.675	1.25	1.909	1.901	1.889	0.0457

The results from the Table 1 are not surprising. Among the state dependent parameters selected, arrival rate decreases and service rate of the fast server increases with increasing K values. One of the findings is that all mean queue lengths L_1, L_2 and L_3 decrease as K value increases. It is of interest to note that $L_1 > L_2 < L_3$ for K=1 and 2 and the condition $L_1 > L_2 > L_3$ is satisfied for K=3,4,5 and 6; a local minimum size of K for buffer B_1 is 3 i.e. K_0 = 3. In this case, chances of finding slow server in its busy state is reported under column D_1 which decreases steadily with increasing K values. Hence, for the existing state dependent rates of arrival and service patterns, as long as buffer B_1 is bounded with K=1 and 2 it is advisable to avoid the slow server. However if the size K of the buffer B_1 is 3 or more, in order to reduce the queue length considerably, the slow server of the $M/(M_1, M_2)/2/(B_1, B_2)$ service facility is to be installed.

4. SUMMARY

This paper analyses a two-server $M/(M_1, M_2)/2/(B_1, B_2)$ queueing system with stalling where customers are served by one fast server and one slow server. The servers are allowed to work in parallel. It has provided a finite buffer B_1 of size `K<∞' to stall customers in queue-1 which is meant to feed the fast sever only. There is one more buffer B_2 of infinite capacity, called the waiting space to accept further arrivals in queue-2 when the buffer B_1 is full. The primary task of B_2 is to feed customers to queue-1 and to the slow server as and when it is warranted. An arriving customer who finds the queue-1 is full and queue-2 is empty joins the slow server. If the queue-2 is non-empty at a time epoch when the slow server finishes a service, he accepts a customer from the head of the queue-2. Both arrival and service rates of customers are assumed to be state dependent parameters for the first (K+2) states starting from state `0'. Formulating the queue length (queue +service) process of the whole system as a QBD processes, steady state results to state probabilities and mean queue length have been obtained using matrix-analytical methods. One special feature is that the stationary queue length distribution is obtained in two stages: Stage-1 deals with the finite capacity queue $M/(M_1,M_2)/2/B_1$ and for the determination of the `K+2' boundary characteristics of stage-1 substantial effort is shown in terms of generator matrices. In addition, it outlines interesting computational properties through an accompanying numerical illustration. Exploiting the Markov property of the QBD process, stage-1 results are linked to the stage-2 which helps to obtain the whole queue length distribution and its characteristics in compact and closed form expressions.

After the completion of linking process in these two stages, the queue length distribution of the M/M/1 queue under T-policy (see Xuelu Zhang, et al. (2015)) is deducted as a specific case from the results of the $M/(M_1,M_2)/2/(B_1,B_2)$ system with stalling. To support the viable applications of the proposed queueing system with stalling in the area of computer networks and manufacturing industries, numerical illustrations are provided. one optimization problem for maximizing the value of K subject to a constraint involving the mean queue lengths of $M/(M_1,M_2)/2/(B_1,B_2)$, M/M+M/1 and M/M/1 queues is formulated and solved. As a direction for future research, the authors wish to extend the proposed methodology of this study to other service industries like three-server and multi-server queues with stalling and to the M/G,M/2 / (B_1,B_2) and M/M,G/2/ (B_1,B_2) queues with stalling.

ACKNOWLEDGEMENTS

Second author wishes to thank **UGC** (India) for providing Rajiv Gandhi National Fellowship (**RGNF**) for research scholars.

REFERECES

[1]. Abou-El-Ato, M.O.and A.L. Shawky,1999,` A simple Approach for the Slow Server Problem', *Commum.fac.Univ.Ank,Series* A,V,48., pp 1-6.

[2]. Fabricio Bandeira Cabari ,2005,`The slow server problem for uninformed Customers', Queuing systems, 50, 353-370.

[3]. Gumbel H,1960,`Waiting Lines with Heterogeneous Servers', Source: *Operations Research,* Vol. 8, No. 4 , pp. 504-511, Published by: INFORMS.

[4]. Kim, J.H., Ahn H.S and R.Righter,2011, ` Managing queues with heterogeneous servers', *Journal of Applied Probability*,48,No2435-452.

[5]. Krishnamoorthi. B,1963,`On Poisson Queue with Two Heterogeneous Servers', Source: *Operations Research,* Vol. 11, No. 3, pp. 321-330, Published by: INFORMS.

[6]. Michael Rubinovitch,1985, `The Slow Server Problem, Source: Journal of Applied Probability', Vol. 22, No. 1 pp. 205-213 Published by: *Applied Probability Trust Stable* URL: https://www.jstor.org/stable/3213760.

[7]. Michael Rubinovitch,1985,`The Slow Server Problem: A Queue With Stalling', Technion - *Israel Institute of Technology J. Appl. Prob.* 22, 879-892.

[8]. Singh, V.P.,1971,` Markovian queues with three heterogeneous servers', *AIIE Transations*, vol.3, no.1, pp.45-48.

[9]. Sivasamy, R., Daaman, O.A. and S.Sulaiman,2015, `An M/G/2 Queue subject to a minimum violation of the FCFS queue discipline', *European Journal of Operational Research,* 240, pp 140 146.

[10]. Sivasamy, R.,Paulraj, G.,Kalaimani, S.and N. Thillaigovindan,2016, A two server Poission Queue Operating under FCFS Discipline with an 'm' Policy', *Singapore SG* January 07-08, 2016,18 part-1.

[11]. Xuelu Zhang, Jinting Wamg, Tien Van Do,2015, ` Threshold properties of the M/M/1 queue under T-policy with applications', *Applied Mathematics and Computation*, volume 261, 15, pages 284-301.