The M/G/2 Queue with Heterogeneous Servers Under a Controlled Service Discipline: Stationary Performance Analysis

Sulaiman Sani and Onkabetse A. Daman *

Abstract—**We study the $M/G/2$ queuing system with an exponential server (server-1) and a general server under a controlled queue discipline. Such a queuing model depicts a service system where servers are allocated to customers rather than chosen as in many telecommunications and computer business centers across the globe. The First Come First Served (FCFS) queue discipline is violated to achieve the least waiting time of customers in the system. Using the remaining service time of the customer on server-2 (the slower server) as a supplementary variable and given that every customer who first finds server-1 busy is allocated server-1, we derived the steady state distribution for the number of customers in the system. Furthermore, closed form expressions for the mean waiting time, the mean queue length and the blocking probability for the queuing system are derived. Finally, mean performance measures are computed numerically and results compared with existing models.**

Keywords: ***The*** $M/G/2$ ***queue, the*** $M/M,G/2$ ***queue, queue discipline, controlled queue discipline.***

1 Introduction

We study the $M/G/2$ queuing system modeled as the $M/M,G/2$ queue with an exponential server (server-1) and a general server (server-2) whose queueing order is controlled. The aim is to study the queue length and the waiting time processes of a service system having two heterogeneous servers that are allocated as in banks, hospitals, telecommunications and several business centers instead of chosen by customers. Customers arrive according to a Poisson process at a rate λ for service on either of the two servers in the system. The servers have the tendency to serve the same type of job[1] as in the case of two digital typing jobs or different jobs as in the case of a digital typing job and a printing job. We suppose that the nature of servers anytime is not new to service providers. Thus, if a customer arrives when both servers are idle, he is asked to take service from server-1 being the faster server. Similarly, if upon arrival a customer finds server-1 busy, **he is asked to wait for the unfinished service of the customer on server-1**. Furthermore, if server-1 is busy and a waiting customer is present, an arriving customer is asked to take service from server-2. Subsequent customers join the queue and wait for their service turn. This schedule is renewed each time there are three or more customers in the system. The service time distribution of customers depends on the server that provides service. For customers served by server-1, the service time T_1 is exponential with a rate μ i.e. $F_1(t) = P(T_1 < t) = 1 - e^{-\mu\, t}$ with probability density function (PDF) $f_1(t) = \frac{dF_1(t)}{dt}$ and Laplace-Stiltjes Transform (LST) $f_1^*(s) = \int_0^\infty e^{-st} dF_1(t)$. Similarly, for customers serviced by server-2, their service time distribution $B(t) = P[T_2 < t]$ is general with PDF $b(t)$, a mean $\beta = E[T_2]$ and a LST $b^*(s)$ given by $b^*(s) = \int_0^\infty e^{-st} dB(t)$. We suppose that the stability condition $\lambda < \mu + \frac{1}{\beta}$ holds and that the hazard rate denoting the departure rate of customers is given by

$$r(t) = \mu + \frac{B'(t)}{1 - B(t)}, \quad t \to \infty. \tag{1}$$

Our interest is on the stationary behavior and performance of the $M/M,G/2$ model under the controlled queue schedule described above: i.e, how does waiting for the unfinished job on server-1 (the faster server) ahead of taking service from server-2 affects the entire waiting time distribution and expectations of customers in the system given that server-2 (the slower server) is idle and ready for job. More precisely, the operational performance of the model relative to the model in Krishnamoorthy [2] where some mass of customers finding server-1 busy may choose to receive service from server-2 if the server is free[2].We are motivated by the many applica-

*S. Sani is a Research Scholar in the Department of Mathematics,University of Botswana-Gaborone. O.A Daman is a Senior Lecturer in the Department of Mathematics, University of Botswana-Gaborone. Corresponding Author Email: man15j@yahoo.com

[1]Though, at distinct server rates

[2]The Krishnamoorthy [2] queue discipline is customer dependent. There are many service systems where the choice of servers is entirely controlled by the service providers. For instance in some ATM centers, hospitals, call business centers etc, customers are directed by agents working for the service provider in the direction of the faster machine, doctor, mobile line etc. The other one is put to use only if a certain customer size is waiting ahead of the faster server.

tions of the M/M,G/2 model working under a controlled service schedule for instance in telecommunication centers and other service systems such as banks, hospitals, shops, etc. A realistic application for this kind of queuing system is a service station with two machines, servers, clerks, etc; a reliable one (exponential server) and a less reliable one (general server) in the presence of a controlling staff. Upon arrival of a customer, the staff directs the usage of the machines according to a set of pre-defined rules that ensures the smooth ordering of the system. There are many reasons why owners of such business outfits (banks, hospitals, telecommunications and computer business centers) may require the analysis presented in this work. For instance, for profit reasons an owner of a commercial call center using two heterogeneous network lines for business may require the understanding of the customer distribution in his center when a certain customer group is to be served by the faster network line. This category of customers is called the control group. The business owner may wish to understand the impact of assigning this customer category to the faster server on the entire waiting time distribution and expectations of customers in the system. Similarly, for quality assurance purposes, a bank manager may require the understanding of the number of customers steadily remaining in the queue if a certain customer group is attached to be served by a unique faster counter clerk in the presence of a slowly working clerk even if the slow clerk is idle. An office executive may need to understand the impact a certain staff is making on the client's distribution given that he is preferred to give service compared with other staff. The above problems come down to modeling a service system with heterogeneous servers in the presence of a controlled customer class attached to the faster server. The control here generally implies that servers are allocated and therefore free from customers' preferential choice, their knowledge of servers and other variables attached to customers. A control policy design completely free from customers' preferential choices etc may improve a lot of variables such as business satisfaction, motivation and patronage, owner based service approach and most importantly, business returns. Consequently, our intended analysis will provide solution to these kind of inquiries whose aim is to simplify problems connected with customer distributions and expectations for better management and business practices, standardizations, improvements, etc.

2 Existing Literature

The literature on heterogeneous queuing systems under various policies[3] has been extensively studied, Emrah et al [1]. The bulk of the literature consists of models whose servers follows the exponential or phase-type distribution working under the first come first served (FCFS) queue discipline. A model of the general type distribution under non FCFS is to our knowledge scarce in the literature. More precisely, models with general type distribution with control queue schedules bias to the business owner[4] is totally lacking. For a survey of the FCFS aligned models; Yue et al [9], Kumar et al [14], Fiems et al [8], Boxma et al [16], Hoksad [17], etc. First, the adoption of the FCFS may not be realistic in modeling service systems with heterogeneous structures. For instance, if counter clerks in a bank provide services with varying speeds (essentially, no two servers can work at the same rate for several reasons), then customers might prefer to choose the fastest clerk for service. On the other hand, if one chooses the slowest clerk randomly then customers that entered the system after him may clear out earlier by obtaining service from a clerk with a faster working rate. Apparently in this case, ***the FCFS queue discipline is violated*** due to heterogeneity in service speeds of the clerks. This and similar real life scenarios make the assumption of the FCFS queue discipline really unrealistic in queuing systems with embedded heterogeneity because of the high probability of violation. Hence, there is the need for designing alternative queue disciplines that can reduce the impact of the violation so that the resulting waiting times of customers are almost identical or even better compared with that of the FCFS. Similarly, quite a large number of service systems[5] for instance in banks, telecommunication call centers, ATM machines, hospitals, etc allocate servers to customers and so models where customers choose servers may be unsuitable to describe such systems because of the absence of variables such as customer preferences, nature, knowledge of servers etc therein. Owner controlled schedules are completely devoid of most or all variables attached to customers and are relatively scarce in the literature. Moreover, it is well known that control problems are difficult to solve. Particularly, their analytical solutions are in many cases out of the question; see Mohammad and Ali [15]. Consequently, it is interesting to study this class of physical systems for better management practice. This stand is key to studying controlled queuing systems under new queuing disciplines generally. To buttress further on the need to study queuing systems with controlled policies, Ekamura et al [10] indicated that queuing control is one of the most important problems in the research field of operations research and management science. Efrosinin and Sztrik [7] proved that performance wise, if a service process of a queuing system with embedded inhomogeneity is controlled such that the slower server is initiated for service only if a certain threshold value is crossed then, the system performance becomes better and faster. This was earlier highlighted by Krishnamoorthy [2] who proposed a non FCFS queue discipline that controlled the service process flexibly by assigning a threshold policy under which the slow server is to be put

[3]Both the $M/M/2$ and the $M/G/2/$ models

[4]Not the customer.

[5]With finite or infinite buffer sizes

to use without necessarily imposing stronger conditions on the controlled class. This discipline is a viable alternative for reducing waiting time expectations in such systems. For more on the benefits of threshold control policies see Efrosinin and Rykov [6], [5] and Efrosinin and Breuer [4]. In a recent work [6] on the $M/G/2$ queue with a violation of the FCFS queue discipline and a flexible customer dependent control policy, we have shown that one can obtain the equivalent of the FCFS waiting time expectations when the violation is minimal owing to some degree of control on the service and queuing disciplines[7]. This work together with those of Krishnamoorthy [2] and Efrosinin and Sztrik [7] motivate us to further still investigate the $M/G/2$ queuing system in the light of a complete control policy of the customer routines in the system to study the impact of the policy on the waiting time expectations for the benefit of service systems.[8] As stated in section 1, the model presented in this work is called the $M/M, G/2$ queue with one exponential and one general server working under a controlled service process. We initiate a control service policy under which the general server can be put to use to study the performance of the $M/M, G/2$ queuing model for use in owner-controlled service systems. Our contributions could be summarized as follows:

1. Designing the $M/M, G/2$ model with an owner controlled service schedule.
2. Deriving the stationary behavior of the model and relevant generating functions.
3. Performance measures for the queuing systems in question.
4. Comparative analysis test.

2.1 The Queue Discipline

If a customer arrives and find:

1. **Both servers free:** He is asked to take service from server-1.
2. **Server-1 is busy and server-2 is idle:** The customer is asked to wait for the unfinished job on server-1.
3. **Server-1 is busy, one customer is waiting for server-1, and server-2 is idle:** The customer is asked to take service from server-2.
4. **Both servers are engaged:** the customer is asked to join the queue and wait for his service turn as a second waiting customer. This schedule is renewed each time there are three or more customers in the system[9].

3 The Stationary Distribution under 2.1

The goal of this section is to compute the generating function for the stationary number of customers in the system. To accomplish this goal, we use the supplementary variable technique on a process $\{X(t), \zeta(t)\}_{t\geq 0}$. For the application of this technique to the $M/G/1$ queuing system, see Cohen [13].

Suppose that a process $\{X(t), \zeta(t)\}_{t\geq 0}$ is given where $X(t)$ denotes the number of customers in the system at time t and $\zeta(t)$ is the past service time of a customer on server-2. Looking at the system at departure instants when a service is completed on server-2, then the bi-variate process $\{X(t), \zeta(t)\}_{t\geq 0}$ is a Markov process. Suppose also that the service time of customers is continuous and that the system is empty at time zero, then one can apply the supplementary variable technique to analyze the process $\{X(t), \zeta(t)\}_{t\geq 0}$. Now, define for $t \geq 0$:

$$R_{0,0}(t) = P(X(t) = 0)$$
$$R_{1,0}(t) = P(X(t) = 1, \ \ server - 2 \ idle)$$
$$R_{1,1,0}(t) = P(X(t) = 2, \ server - 2 \ idle)$$
$$R_j(t,\eta)d\eta = P(X(t) = j, \ \ \eta \leq \zeta(t) < \eta + d\eta), \ \eta > 0, \ j = 3, 4, ...$$

Given that $\lambda < \mu + \frac{1}{\beta}$ holds, then as $t \to \infty$, $R_{0,0}(t)$, $R_{1,0}(t)$, $R_{1,1,0}(t)$ and $R_j(t,\eta)$ will converge to $R_{0,0}$, $R_{1,0}$, $R_{1,1,0}$ and $R_j(\eta)$ respectively[10]. More precisely, $\{X(t), \zeta(t)\}_{t\geq 0} \to \{X, \zeta\}$. Let R_j denote the stationary probability that there are j customers in the system. Given that every departure is followed by an arrival in that order so that the rate-equality principle holds, then one can conclude that $R_0 = R_{0,0}$, $R_1 = R_{1,0}$, $R_2 = R_{1,1,0}$. And for $j \geq 3$ customers, $R_j(\eta)$ gives the stationary probability that there are j customers in the system when both servers are busy. By the arguments of ergodicity satisfied by the process $\{X, \zeta\}$, the stationary probabilities $R_{0,0}$, $R_{1,0}$, $R_{1,1,0}$ and $R_j(\eta)$ will satisfy the following differential equations

$$\lambda R_{0,0} = \mu R_{1,0}; \qquad j = 0 \tag{2}$$

[6] European Journal of Operational Research (EJOR), 240, 140-146.

[7] See the parallel queue discipline in Sivasamy et al [18].

[8] In a two-server heterogeneous retrial queue with threshold policy.

[9] Thus, it is apparent that the whole service schedule here is controlled. Moreover, the unit and the second customer classes receive service from server-1 only.

[10] In steady state, every $R_j(\eta) \to R_j$. This ensures that the associated Markov chain for the process $\{X, \zeta\}$ is ergodic. Henceforth, whenever $R_j(\eta)$ comes into picture, it implies a stationary probability for a j when both servers are busy only.

$$(\lambda+\mu)R_{1,0} = \lambda R_{0,0} + \mu R_{1,1,0}; \quad j=1 \tag{3}$$

$$(\lambda+\mu)R_{1,1,0} = \lambda R_{1,0} + \left(\mu+\frac{1}{\beta}\right)R_{1,1,1}(\eta); \quad j=2 \tag{4}$$

For $j=3$ we have

$$\left(\lambda+\mu+\frac{1}{\beta}\right)R_{1,1,1}(\eta) = \lambda R_{1,1,0} + \left(\mu+\frac{1}{\beta}\right)R_4(\eta) \tag{5}$$

Generally for $j \geq 3$

$$\left(\lambda+\mu+\frac{1}{\beta}\right)R_j(\eta) = \lambda R_{j-1}(\eta) + \left(\mu+\frac{1}{\beta}\right)R_{j+1}(\eta) \tag{6}$$

From which

$$R_1 = R_{1,0} = \left(\frac{\lambda}{\mu}\right)R_{0,0} = \left(\frac{\lambda}{\mu}\right)R_0 \tag{7}$$

$$R_2 = R_{1,1,0} = \left(\frac{\lambda}{\mu}\right)^2 R_0 \tag{8}$$

To solve for R_j; $\quad j \geq 3$, let

$$Q_j(\eta) = \frac{R_j(\eta)}{1-B(\eta)}, \tag{9}$$

where $R_j(\eta)$ is the probability that there are j customers in the system when $\eta \leq \zeta \leq \eta + d\eta$ and $1-B(\eta)$ is the complementary distribution of service times on server-2 during this service period [11]such that

$$Q_j^*(s) = \int_0^\infty e^{-s\eta} Q_j(\eta) d\eta \tag{10}$$

Define

$$\tilde{Q_j}(\eta) = \beta \tilde{R_j}(\eta) \tag{11}$$

where

$$\tilde{R_j}(\eta) = \int_0^\infty Q_j(\eta) dB(\eta) \tag{12}$$

If $Q_j^*(s)$ is evaluated as $s \to 0$, then by the following lemma the resulting integral in (10) is equivalent to (11). Thus, under this added condition, one can take $Q_j(\eta)$ to R_j when there are three or more customers in the system.

Lemma 3.1 *Given that the traffic condition $\lambda < \mu + \frac{1}{\beta}$ holds, then in a busy period*

$$Q_j^*(0) = \tilde{Q}_j, \quad j = 3, 4, ..., \tag{13}$$

[11]Strictly, for $j \geq 3$ when server-2 is busy only.

Proof Suppose that a busy period is in progress such that the time T_n between any two successive departures on server-2 is given by $T_n = t_n - t_{n-1}, \quad n = 1, 2, 3, 4, \ldots$ Then for $n \geq 1$, the service period is a probabilistic replication of the initial period T_1 starting at $t = 0$. Now, if the queue length process[12] at t is $N(t)$ such that $N(0) \geq 3$, then $N(t)$ would reach steady state[13] starting at $t = 0$. Consequently, $N(t)$ is a regenerative process over t on state space $S = 3, 4, ...$ and $T_n = t_n - t_{n-1}$ is the underlying renewal process at time epoch t_j each time a departure occurs on server-2. Now, given that $\lambda < \mu + \frac{1}{\beta}$ holds, then upon service completion on server-2, the state probability $R_j(t)$ can be expressed as

$$R_j(t) = P[N(t) = j, \quad j = 3, 4, \ldots]. \tag{14}$$

In addition, if η units of service time elapsed in a busy period at any time t, then the conditional probability that there are j customers in the system is equal to

$$R_j(t, \eta) = P[N(t) = j \mid t = \eta, \quad j = 3, 4, \ldots] \tag{15}$$

Let

$$Q_j(t, \eta) = \frac{R_j(t, \eta)}{1 - B(t, \eta)} \tag{16}$$

so that

$$Q_j(t)(1 - B(t)) = R_j(t) = P[N(t) = j \mid t_1 > t] \tag{17}$$

then

$$\sum_{j=3}^{\infty} Q_j(t) = P[t_1 > t] = 1 - B(t) \tag{18}$$

and

$$Q_j(t) = \int_0^\infty P[N(t) = j \mid t_1 > t] dB(\eta) = \int_0^\infty A dB(\eta) \tag{19}$$

[14] which simplifies to the expression

$$Q_j(t) = \int_\eta^\infty P[N(t) = j \mid t_1 = \eta] dB(\eta) \tag{20}$$

and $R_j(t)$ now will equal to the equation

$$R_j(t) = \int_0^\infty P[N(t) = j, t_1 = \eta] dB(\eta) \tag{21}$$

Thus, by conditioning on T_1 under steady state conditions, it can be shown that the following renewal equation below is satisfied

$$R_j(t) = Q_j(t) + \int_0^t R_j(t - x) dB(x). \tag{22}$$

[12]$N(t)$ is equivalent to $X(t)$
[13]Precisely at (0+)
[14]$A = P[N(t) = j, t_1 > t \mid t_1 = \eta]$

This renewal equation has a unique solution of the form

$$R_j(t) = Q_j(t) + \int_0^t Q_j(t-x)dM(x) \tag{23}$$

where $M(x)$ is the renewal function of a renewal process with inter-renewal time distribution $B(t)$. Thus, the application of the key-renewal theorem yields that

$$\lim_{t\to\infty} R_j(t) \to \frac{1}{\beta}\int_0^\infty Q(x)dx \tag{24}$$

The integral in (24) is the probabilistic version of $\tilde{Q}_j$ when the mean service time on server-2 is β. Thus,

$$\tilde{R}_j\beta = \tilde{Q}_j = Q_j^*(0) \tag{25}$$

Thus, the lemma holds.

3.1 The Stationary Probability Distribution

To solve (5) and (6), apply the Laplace operator on the transformed differential equations (48) and (50) coupled with the initial conditions (49) and (51). One obtains that[15]

$$sQ^*_{1,1,1}(s) + (\lambda + B)Q^*_{1,1,1}(s) = \mu Q_4^*(s) + \lambda R_{1,1,0} + \frac{1}{\beta}\tilde{Q}_4 \tag{26}$$

And $j \geq 4$ we have

$$sQ_j^*(s) + (\lambda + B)Q_j^*(s) = \lambda Q^*_{j-1}(s) + \mu Q^*_{j+1}(s) + \frac{1}{\beta}\tilde{Q}_{j+1} \tag{27}$$

Application of lemma 3.1 together with equations (10) and (11) as $s \to 0$, coupled with the Markov property of the system, $R_j, \quad j \geq 3$ simplifies to the equations below

$$R_3 = R_{1,1,1} = \left(\frac{\lambda}{\mu + \frac{1}{\beta}}\right)\left(\frac{\lambda}{\mu}\right)^2 R_0 \tag{28}$$

$$R_4 = \left(\frac{\lambda}{\mu + \frac{1}{\beta}}\right)^2\left(\frac{\lambda}{\mu}\right)^2 R_0 \tag{29}$$

Generally, if $j \geq 2$, then any R_j can be evaluated from (30) below

$$R_j = \left(\frac{\lambda}{\mu + \frac{1}{\beta}}\right)^{(j-2)}\left(\frac{\lambda}{\mu}\right)^2 R_0 \tag{30}$$

where R_0 is the idle state probability. Put $\frac{\lambda}{\mu+\frac{1}{\beta}} = \rho_1$ and $\frac{\lambda}{\mu} = \rho$ and apply the normalization condition, then

[15]We write $B = \left(\mu + \frac{1}{\beta}\right)$ in longer equations

$$1 = R_{0,0} + R_{1,0} + R_{1,1,0} + \sum_{j=3}^{\infty} R_j \tag{31}$$

Upon further simplification, one obtains that

$$R_0 = R_{0,0} = \frac{(1-\rho_1)}{(1+\rho)(1-\rho_1)+\rho^2} \tag{32}$$

Inserting (32) in (7) and (30), one obtains the stationary probability R_j for the $M/M,/G/2$ queue under the control service process here. Thus

$$R_1 = R_{1,0} = \frac{\rho(1-\rho_1)}{(1+\rho)(1-\rho_1)+\rho^2} \tag{33}$$

And for $j \geq 2$ customers, we have

$$R_j = \frac{\rho_1^{(j-2)}\rho^2(1-\rho_1)}{(1+\rho)(1-\rho_1)+\rho^2} \tag{34}$$

Lemma 3.2 *Suppose $j = K \in \Re$ denote maximum waiting capacity of the M/G/2 queue in question. Then the stationary blocking probability is given by*

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

Proof Sum the right hand side of (31) over $K < \infty$. Upon simplification, one obtains that

$$R_0 = R_{0,0} = \frac{(1-\rho_1)}{(1+\rho)(1-\rho_1)+\rho^2(1-\rho_1^K)} \tag{36}$$

Inserting (36) in (30), R_j for the K-capacity $M/M,G/2$ queue under the controlled queue discipline here is obtained. This is given by

$$R_1 = R_{1,0} = \frac{\rho(1-\rho_1)}{(1+\rho)(1-\rho_1)+\rho^2(1-\rho_1^K)}, \qquad K = 1 \tag{37}$$

and[16] for $2 \leq j \leq K$ customers we have

$$R_j = \frac{\rho_1^{(j-2)}\rho^2(1-\rho_1)}{(1+\rho)(1-\rho_1)+\rho^2(1-\rho_1^K)} \tag{38}$$

Finally, the lemma follows if every j in (38) is replaced by K.

Lemma 3.3 *Suppose server-2 breaks down such that $\rho_1 \to \rho$ in (38). Then the stationary blocking probability R_K reduces to*

$$R_K = \frac{(1-\rho)\rho^K}{1-\rho.\rho^{K+1}} \tag{39}$$

[16]$K = 0$ is unrealistic.

Proof $\rho_1 \to \rho$ implies that a certain service of mean $\beta \to \infty$ is in progress on server-2 such that the mean service rate $\frac{1}{\beta} \to 0$. This reduces the system to a uni server $M/M/1/K$. Trivially, the lemma holds if ρ substitutes ρ_1 in (35) upon simplification.

Remark The blocking probability in (39) corresponds to the reduced form of the well known blocking probability for the $M/M/1/K$ model in MacGregor [12] for finite values of K and $\rho < 1$. Comparing (39) with the one for the $M/M/1/K$ in MacGregor [12], the percentage reduction in the size of R_K is typically about seven percent in favor of the model here for $\rho \leq 0.5$ and up to twenty percent as $\rho \to 1$. However, for smaller values of ρ the difference between the two probabilities is insignificant.

3.2 Generating Functions & Expectations:

Now, denote by $V(z) = \Sigma_{j=0}^{\infty} R_j z^j$ the generating function for the number of customers j (including the service customer) with stationary probability R_j such that[17]

$$V(z) = \Sigma_{j=0}^{\infty} R_j z^j = R_0 + R_1 z + R_2 z^2 + R_3 z^3 + ... \quad (40)$$

then

$$V'(z) = R_1 + 2R_2 z + \frac{d}{dz}\Sigma_{j=3}^{\infty} R_j z^j$$

so that

$$V'(z) = R_1 + 2R_2 z + \frac{d}{dz}\left[\frac{R_3 z^3}{(1-\rho_1 z)}\right]$$

and at $z = 1$, we have

$$V'(1) = R_1 + 2R_2 + \frac{R_3(3-2\rho_1)}{(1-\rho_1)^2}$$

Here, $V'(1) = E[X]$, gives the stationary expected number of customers in the system upon departure instances.

Simplifying further, one obtains that

$$E[X] = \left(\frac{(1-\rho_1)}{(1+\rho)(1-\rho_1)+\rho^2}\right)\left(\rho + 2\rho^2 + \frac{\rho_1\rho^2(3-2\rho_1)}{(1-\rho_1)^2}\right) \quad (41)$$

Application of Little's law gives the expected waiting time E[W]. That is

$$E[W] = \left(\frac{(1-\rho_1)}{\lambda[(1+\rho)(1-\rho_1)+\rho^2]}\right)\left(\rho + 2\rho^2 + \frac{\rho_1\rho^2(3-2\rho_1)}{(1-\rho_1)^2}\right) \quad (42)$$

[17]Here, the departure probability R_j is used assuming that server-2 is exponentially distributed so that PASTA property holds.

Lemma 3.4 *Suppose $j \leq 2$ customers in every service epoch. Then, the stationary probability R_j of the infinite capacity $M/M, G/2$ queuing system converges to that of the $M/M/1$ queuing system.*

Proof Intuitively, the stationary customer process (X, ζ) is saddled on server-1. Given that $\lambda < \mu + \frac{1}{\beta}$ holds, it is trivial that $\lambda < \mu$ for the $M/M/1$ is implied. Suppose that a long service of mean β is in progress on server-2 such that $\frac{1}{\beta} = 0$: more precisely, $\beta \to \infty$ onto $\rho_1 = \frac{\lambda}{\mu+\frac{1}{\beta}}$ where ρ_1 is the occupation rate of the two servers in the $M/M, G/2$ model. Then $\rho_1 = \frac{\lambda}{\mu}$. Denote by ρ the occupation rate of the classical $M/M/1$ model. By definition $\rho = \frac{\lambda}{\mu}$. This means that $\rho_1 = \rho$. Hence, the stationary customer distribution for such customer size of the $M/M, G/2$ is that of the $M/M/1$ queue.

4 Designing Optimal Buffers

If we relax the integrality of the buffer K, we can express it in terms of R_K for fixed values of ρ and ρ_1. Then, one will arrive at a closed-form expression for optimal queuing space size/buffer for the model in question. This corresponds to the largest integer K as follows:

$$K = 2 + \frac{\ln\left\{\frac{R_K[(1+\rho)(1-\rho_1)+\rho^2]}{R_K(\rho\rho_1)^2+\rho^2(1-\rho_1)}\right\}}{\ln \rho_1} \quad (43)$$

The value for K in (43) above could be used in designing appropriate queuing space/buffers for known values of the occupation rates under the probabilistic assumption of a finite customer size expected in the system[18]. The table below summarizes some optimal K values for $\lambda = 15.81$, $\mu_1 = 18.5$ and $\frac{1}{\beta} = 7.4$.

Table-1: Optimal Buffers for selected R_K.

R_K	K
10^{-2}	8.05
10^{-3}	13.65
10^{-4}	17.4
10^{-5}	21.99
10^{-6}	26.65

Lemma 4.1 *For the queuing system in question, the minimum buffer size K_{min} exists.*

Proof Suppose $R_K \to \frac{\rho^2(1-\rho_1)}{[(1+\rho)(1-\rho_1)+\rho^2]-(\rho\rho_1)^2}$, where R_K is the stationary probability that the system is full to

[18]$K = 1$ is a telephone kiosk. Similarly, $K = 2$ is a special case

capacity for such $K \in Z$ and ρ, ρ_1 are the occupation rates of server-1 and the joint servers in the system. This ensures that the other component to the right of (43) goes to zero. Consequently, the lemma holds.

Lemma 4.2 *If K is minimal then $R_K(M/M, G/2/K)$ goes to $R_K(M/M/1/2)$.*

Proof By lemma 4.1, K is minimal if

$$R_K = \frac{\rho^2(1-\rho_1)}{[(1+\rho)(1-\rho_1)+\rho^2]-(\rho\rho_1)^2} \tag{44}$$

The minimality condition in (44) ensures that $\rho_1 \to \rho$. Substituting ρ for ρ_1 in (44) for $K = 2$ and comparing with that computed from (39) proves the lemma.

5 Numerical Simulations & Discussions

In this section, we provide numerical simulation for the $M/M, G/2$ queuing system given that every second customer is controlled to wait for the on-going service on server-1[19]*. For operational sake, we similarly compare and contrast the performance of the model with that of the $M/M/2$ under the Krishnamoorthy [2] queue discipline when some mass of the second customer category decide to take service on server-2. For $15.11 \leq \lambda \leq 15.81$, $\mu = 8.4$ and $\frac{1}{\beta} = \mu_2 = 7.5$, the simulated values for ρ, ρ_1 corresponding to $E(N)_{fl.}$, $\bar{W}_{fl.}$, $E(X)_{ctr.}$ and $\bar{W}_{ctr.}$ are summarized*[20] *in the table below*

Table-2: Expectations $E(N)$, $E[X]$ and $\bar{W}$

λ	ρ	ρ_1	$E(N)_{fl.}$	$E(X)_{ctr.}$	$\bar{W}_{fl.}$	$\bar{W}_{ctr..}$
15.11	1.80	0.950	19.95	20.28	1.32	1.34
15.21	1.81	0.957	23.21	23.20	1.53	1.53*
15.31	1.82	0.963	27.41	27.11	1.79	1.77
15.41	1.83	0.970	33.15	32.62	2.15	2.12
15.51	1.85	0.976	41.68	40.94	2.68	2.64
15.61	1.86	0.982	55.92	55.01	3.58	3.52
15.71	1.87	0.988	84.93	83.87	5.41	5.34
15.81	1.88	0.994	178.05	176.86	11.26	11.19

Table-2 gives a summary of mean performances of the $M/M, G/2$ queue when every second customer takes service from server-1 (controlled) and the case when some mass of second customers take service from server-2 (flexible). Referenced to these results, the following conclusions can be drawn.

1 **Existence of a turning point**

This is implied from the case when $\lambda = 15.11$. Looking at the means corresponding to this arrival rate, it can be seen that both the mean queue length and the mean waiting time numbers for the controlled case are greater than that of the flexible case. This phenomenon might have evolved from an arrival point located far from the combined service rate where both means for the flexible case are stationary smaller than that of the controlled case. Interestingly, as λ approaches $15.11 + \epsilon$; $\epsilon > 0$, the behavior of the models change steadily in favor of the controlled case[21]. This is evident in the size of the mean numbers for the controlled case in comparison to the flexible case somewhere above $\lambda = 15.11$. Consequently, it can be concluded that the said arrival rate is a turning point when it is operationally better to exercise more control on the servers in the system by redirecting a given customer category to wait in total for the unfinished job on the faster server than allow the customer go to the slower server. This can be explained by the fact that even with $\frac{1}{\beta}$ close to μ as in the values used for the numerical approximations, the supremacy of the controlled case is clear after the turning point.

Lemma 5.1 *Operationally, it is beneficial to control a given customer category in high speed systems.*

Proof This is observed from the fact that as $\lambda \to (\mu + \frac{1}{\beta})$, the $M/M, G/2$ model under the controlled customer category has stationary smaller values for both the mean queue length and waiting times compared to the flexible case. More so, the disparity between the means keeps increasing significantly in favor of the controlled case. This can be observed from the table above that when $\lambda = 15.81$, the difference in the mean queue length between the means is approximately 3 customers. Thus the controlled model under the stated queue discipline is a better alternative for reducing both the mean queue length and the mean waiting time in high speed systems.

Lemma 5.2 *Suppose two $M/M, G/2$ queuing models are given one with a controlled customer category and the other of flexible customer category. Under heavy traffic conditions,*[22] *the mean performance of the controlled model is higher than that of the flexible model.*

Proof This is expressed by the numerical approximations above. Similarly, By lemma 4.1. Thus, it can be deduced that at a point when ρ_1 is sufficiently close to one, the disparity of the means between the models will

[19] Under our queue discipline.

[20] fl. stands for flexible, ctr. for controlled.

[21] $*$ indicates that the next digit is smaller in the control than in the flexible.

[22] $\rho_1 \to 1$

attain a global maximum. This maximum defines the supremacy of controlling a unique customer category over the flexible case.

Lemma 5.3 *Buffer size decreases with increase in capacity probability.*

Proof This is evident from the numerical approximations expressed in table 1 which shows that the buffer size K increases significantly with decreasing capacity probability R_K.

The relationship in the lemma above holds equivalently for the $M/M/1$ and the $M/D/1$ queuing systems as in Macgregor [12]. Thus, to minimize loss of resources, knowledge of R_K is necessary for constructing buffers for queuing systems with finite waiting spaces generally. More specifically, by minimizing unnecessary spaces arising from poor understanding of the capacity probability.

As a scope for further work on this model, one may wish to analyze the asypmtotics for various ranges of the arrival rate λ compared with the service rate of the exponential server. This will define a bound under which the slow server should be initiated, left idle, replaced, etc for a better joint system performance similar to what is found in Boxma et al [16]. Similarly, one can provide a similar analysis for r-customers ($r \leq j$) waiting for service on server-1 ahead of taking service on server-2.

6 Conclusion

In this article, the $M/M, G/2$ queuing system with heterogeneous servers and a controlled customer service schedule is proposed. The stationary customer distribution is analyzed and performance measures computed. Furthermore, numerical results are compared with those derived for the $M/M/2$ queuing system under the Krishnamoorthy [2] queuing discipline. Our simulation shows that under similar arrival condition, the model performs better compared with the one proposed in Krishnamoorthy [2]. Hence, a good alternative for use in problems connected with customer distributions for better practice. The work affirms that more control on the initiation of the slow server minimize the waiting time expectations as in Efrosinin and Sztrik [7].

Acknowledgements

The authors gratefully acknowledge the referees and the handling editor for sparing time to correct and suggest relevant changes that improve the article to this stage.

7 APPENDIX

7.1 RATE EQUATIONS FOR THE $M/M, G/2$ WITH CONTROL POLICY

$$\frac{d}{dt}R_0(t) = -\lambda R_0(t) + \mu R_{1,0}(t) \tag{45}$$

$$\frac{d}{dt}R_{1,0}(t) = -(\lambda+\mu)R_{1,0}(t) + \lambda R_0(t) + \mu R_{1,1,0}(t) \tag{46}$$

$$\frac{d}{dt}R_{1,1,0}(t) = -(\lambda+\mu)R_{1,1,0}(t) + \lambda R_{1,0}(t) + \frac{1}{\beta}\tilde{Q}_{1,1,1}(\eta) \tag{47}$$

$$Q'_{1,1,1}(\eta) = -\left(\lambda+\mu+\frac{1}{\beta}\right)Q_{1,1,1}(\eta) + \mu Q_4(\eta) \tag{48}$$

$$Q_{1,1,1}(0+) = \lambda R_{1,1,0} + \frac{1}{\beta}\tilde{Q}_4, \quad j=3 \tag{49}$$

For $j \geq 4$, we have

$$Q'_j(\eta) = -\left(\lambda+\mu+\frac{1}{\beta}\right)Q_j(\eta) + \lambda Q_{j-1}(\eta) + \mu Q_{j+1}(\eta) \tag{50}$$

$$Q_j(0+) = \frac{1}{\beta}\tilde{Q}_{j+1}, \quad j \geq 4 \tag{51}$$

7.2 STATIONARY BALANCED EQUATIONS FOR THE $M/M, G/2$ WITH CONTROL POLICY

$$\lambda R_0 = \mu R_{1,0} \tag{52}$$

$$\lambda R_{1,0} = \mu R_{1,1,0} \tag{53}$$

$$\lambda R_{1,1,0} = \left(\mu+\frac{1}{\beta}\right)R_{1,1,1} \tag{54}$$

$$\lambda R_{1,1,1} = \left(\mu+\frac{1}{\beta}\right)R_4 \tag{55}$$

For $j \geq 3$ we have

$$\lambda R_j = \left(\mu+\frac{1}{\beta}\right)R_{j+1} \tag{56}$$

7.3 STATIONARY BALANCED AND RATE EQUATIONS FOR THE $M/M/2$ WITH FLEXIBILITY[23]

$$\lambda R_0 = \mu R_{1,0} + \int_0^\infty R_{0,1}(\eta)\frac{dB(\eta)}{1-B(\eta)} \quad (57)$$

$$(\lambda+\mu)R_{1,0} = \lambda R_0 + \mu R_{1,1,0} + \int_0^\infty R_{1,1}(\eta)C \quad (58)$$

$$R'_{0,1}(\eta) = -(\lambda+\mu+C)R_{0,1}(\eta) + \mu R_{1,1}(\eta) \quad (59)$$

$$R_{0,1}(0+) = 0,\ j=1 \quad (60)$$

$$(\lambda+\mu)R_{1,1,0} = \lambda R_{1,0} + \int_0^\infty R_{1,1,1}(\eta)C \quad (61)$$

$$R'_{1,1}(\eta) = -(\lambda+\mu+C)R_{1,1}(\eta) + \lambda R_{0,1}(\eta) + \mu R_{1,1,1}(\eta) \quad (62)$$

$$R_{1,1}(0+) = 0,\ j=2 \quad (63)$$

$$R'_{1,1,1}(\eta) = -(\lambda+\mu+C)R_{1,1,1}(\eta) + \lambda R_{1,1}(\eta) + \mu R_4(\eta) \quad (64)$$

$$R_{1,1,1}(0+) = \lambda R_{1,1,0} + \int_0^\infty R_4(\eta)\frac{dB(\eta)}{1-B(\eta)}, j=3 \quad (65)$$

For $j \geq 4$, we have

$$R'_j(\eta) = -(\lambda+\mu+C)R_j(\eta) + \lambda R_{j-1}(\eta) + \mu R_{j+1}(\eta) \quad (66)$$

$$R_j(0+) = \int_0^\infty R_{j+1}(\eta)\frac{dB(\eta)}{1-B(\eta)} \quad (67)$$

References

[1] 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**, 449-463, 2013.

[2] B. Krishnamoorthy. On Poisson Queue with two Heterogeneous Servers. *Operations Research,* **2**(3), 321-330, 1962.

[3] B. Krishnamoorthy and S. Sreenivasan. 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, 2012.

[4] D. Efrosinin and B. Breuer. Threshold policies for controlled retrial queues with heterogeneous servers. Annals of Operation Research, **141**, 139-162, 2006.

[5] D. Efrosinin and V. Rykov. Optimal control of queuing systems with heterogeneous servers. Queuing Systems, **46**, 389-407, 2004.

[6] D. Efrosinin and V. Rykov. On performance characteristics for queuing systems with heterogeneous servers. *Automation and Remote Control*, **1**, 64-82, 2008.

[7] D. Efrosinin and J. Sztrik. Performance analysis of a two-server heterogeneous retrial queue with threshold policy. *Quality Technology and Quantitative Management,* **8**(3), 211-236, 2011.

[8] D. Fiems, T. Maertens and H. Bruneel. Queuing systems with different types of server interruptions. *European Journal of Operations Research,* **3**(1), 838-845, 2008.

[9] D. Yue, W. Yue, J. Yu and R. Tian. A Heterogeneous Two-Server Queuing System with Balking and Server Breakdowns. *Proceedings of The Eighth International Symposium on Operations Research and Its Applications (ISORA09),* Zhangjiajie, China, Sept. 20-22, 230-244, 2009.

[10] H. Okamura, T. Dohi and S. Osaki. Optimal Policies for a Controlled Queuing System with Removable Server under a Random Vacation Circumstance. *Computers and Mathematics with Applications,* **39**, 215-227, 2000.

[11] J.H. Kim, H.S. Ahn and R. Righter. Managing queues with heterogeneous servers. *J. Appl. Probab.,* **48**(2), 435-452, 2011.

[12] J.S. MacGregor. M/G/c/K Blocking Probability Models and System Performance. *Performance Evaluation,* **52**, 237-267, 2002.

[13] J.W. Cohen. The single server queue. *2nd ed., North-Holland, Armsterdam, 1982.*

[14] K.B. Kumar, P.S. Madheswari, and K.S. Venkatakrishnan. Transient Solution of an M/M/2 queue with Heterogeneous Servers subject to Catastrophes. *Information and Management Sciences,* **18**(1), 63-80, 2007.

[15] K. Mohammad and M. Ali. A Hybrid Method for Solving Optimal Control Problems. *IAENG International Journal of Applied Mathematics,* **42**(2), 80-86, 2012.

[16] O.J. Boxma, Q. Deng and A.P. Zwart. Waiting time asymtotics for the M/G/2 queue with heterogeneous servers. *Queuing Systems*, **40**:5–31, 2002.

[17] P. Hoksad. On the steady state solution of the M/G/2 queue. *Advanced applied probability,* **11**, 240–255, 1979.

[18] R. Sivasamy, O.A. Daman and S. Sani. An M/G/2 Queue Subject to a Minimum Violation of the FCFS Queue Discipline. *European Journal of Operational Research,* **240**, 140-146. DOI: 10.1016/j.ejor.2014.06.048, 2015.

[23] Here also we use $C = \frac{dB(\eta)}{1-B(\eta)}$ in longer equations.

[19] S. Shenkar and A. Weinrib. The Optimal Control of Heterogeneous Queuing Systems: A Paradigm for Load-Sharing and Routing *IEEE transactions on Computers,* **38**(12), 1724-1736, 1989.

[20] V.P. Singh. Two-Server Markovian Queues with Balking: Heterogeneous vs. Homogeneous Servers. *Operations Research,* **18**(1), 145-159, 1968.