European Journal of Operational Research 240 (2015) 140–146

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

An $M/G/2$ queue where customers are served subject to a minimum violation of FCFS queue discipline

Sivasamy Ramasamy [a], Onkabetse A. Daman [b], Sulaiman Sani [b,*]

[a] Department of Statistics, University of Botswana, Private Bag 0022, Gaborone, Botswana
[b] Department of Mathematics, University of Botswana, Private Bag 0022, Gaborone, Botswana

ARTICLE INFO

Article history:
Received 7 November 2013
Accepted 30 June 2014
Available online 19 July 2014

Keywords:
The M/G/2 queue
The M/(M+G)/2 queue
The M/M,G/2 queue

ABSTRACT

This article discusses the steady state analysis of the $M/G/2$ queuing system with two heterogeneous servers under new queue disciplines when the classical First Come First Served '(FCFS)' queue discipline is to be violated. Customers are served either by server-I according to an exponential service time distribution with mean rate μ or by server-II with a general service time distribution $B(t)$. Sequel to some objections raised in the literature on the use of the classical FCFS queue discipline in heterogeneous service systems, two alternative queue disciplines (*Serial* and *Parallel*) are considered in this work with the objective that if the FCFS is violated then the violation is a minimum in the long run. Using the embedded method under the serial queue discipline and the supplementary variable technique under the parallel queue discipline, we present an exact analysis of the steady state number of customers in the system and most importantly, the actual waiting time expectation of customers in the system. Our work shows that one can obtain all stationary probabilities and other vital measures for this queue under certain simple but realistic assumptions.

1. Introduction

We consider the $M/G/2$ queuing system with an exponential server (server-I) and a general server (server-II). Customers arrive according to a Poisson process at a rate λ for service with much expectation of spending the least waiting time in the system. Server-I has a faster service rate compared with server-II. This information is not new to prospective customers. Thus, server-I is always busy when there is at least a customer in the system. The service times of customers[1] are assumed to be a sequence of mutually independent and identically distributed random variables with finite moments. In addition, services are without preemption. For customers serviced by server-I, the service time T_1 follows the exponential distribution with 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) $f_1^*(s) = \int_0^\infty e^{-st} dF_1(t)dt$. Similarly, for customers serviced by server-II, 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 supposed that $\mu_2 = \frac{1}{\beta}, \rho = \frac{\lambda}{\mu}$ and the servers occupation rate (servers utilization) $\rho_1 = \frac{\lambda}{\mu+\mu_2} < 1$, so that all steady state results are tractable, see Boxma, Deng, and Zwart (2002).

Generally, a number of literature sources on queuing systems raises some objections to the use of the classical First Come First Served (FCFS) queuing discipline in systems with heterogeneous structures (Krishnamoorthy, 1962; Alexander, Marcus, & Cristobal, 2014[2]). This stand can be justified. For instance, if clerks in a reservation counter provide service with varying speeds 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 FCFS queue discipline unrealistic in queuing systems with embedded heterogeneity because of the high probability of violation therein. 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 with that of the FCFS. Similarly in a business center, one may come across a scenario where both the salesman and the boss[3] (owner) are providing

* Corresponding author.
E-mail addresses: ramasamysr@mopipi. ub.bw (S. Ramasamy), damanoa@mopipi. ub.bw (O.A. Daman), man15j@yahoo.com (S. Sani).

[1] Which are independent of inter-arrival times.

[2] Implied in this work also.

[3] Or even a salesman and a supervisor or a senior colleague.

http://dx.doi.org/10.1016/j.ejor.2014.06.048
0377-2217/

services together. If there is only one customer then he is serviced by the salesman provided that no any other one arrives during his service period. On the other hand, if at least an arrival occurs then the boss joins the salesman either serially (working jointly with the salesman to serve the initial customer) or in parallel (attending a different customer). As a schedule, if there are up to two customers in the center then the boss will join the salesman in service[4] so that customers do not get bored by excessive waiting and choose to renege or balk thereby making the business loose market. We refer to the first service approach as *serial service* as implied in *the serial queue discipline* below and the later as *parallel service* as in *parallel queue discipline* also below. Note that the class of serial queuing systems is not extensively studied in the literature in contrast to the parallel case systems (see Emrah, Ceyda, & Irem, 2013) even though, significant areas of application are found in reality. Our motivation stemmed from these numerous physical applications of the proposed models (both serial and parallel) in shops, malls, supermarkets, offices, banks and several other business outfits where heterogeneity of servers is embedded. An in-depth analysis therefore is an excellent tool for decision making relative to congestion management, better service provision, etc. For instance, a business mogul owning two shops each with two staff (servers) of varying work speed may wish to understand which of the two models above[5] minimizes waiting times better.[6] Similarly, he may wish to understand whether an equilibrium point exists under which one model is to be preferred to the other or even when the two models are identical. These kind of questions are vital for performance evaluation and better management practice since heterogeneity is a natural embedding in reality and is the necessitating factor leading to server discrimination. For instance between a well experienced shop seller and an apprentice, a senior doctor and a junior doctor, a professor and a bachelor, etc. Basically, customers hate to wait for a longer time arising from the in-effectiveness of a slow server and in several instances may prefer to wait for the remaining service time of a customer being served by a fast server even when the slow server is idle. Over the years, a lot have been written on homogeneous service systems for instance, Hoksad (1978), Hoksad (1979), Senthamaraikannan and Sivasamy (1997), Tijms, Vaan Hoorn, and Federgruen (1981), etc. The reader is referred to these and many others to refresh. Similarly research works on heterogeneous service systems has grown tremendously in the last two decades; Kim, Ahn, and Righter (2011), Kumar, Madheswari, and Venkatakrishnan (2007), Krishnamoorthy (1962), Shenkar and Weinrib (1989), Singh (1968), etc. In the models described in these works, the heterogeneity structure is saddled on servers following relatively the same distributions. A model of the general service type is studied by Boxma et al. (2002) unfortunately, due to complex structuring, formations and assumptions, it could not estimate certain areas in the general case. Part of this complexity may be saddled on the assumption of the FCFS queue discipline adopted in such a heterogeneous structure.[7]

In this article, we have introduced two queue disciplines (*serial* and *parallel*) whose effects on the two models described below can all be computed numerically. Most importantly, the serial queue discipline is relatively close to that of Boxma et al. (2002) and the parallel queue discipline is that of Krishnamoorthy (1962). Thus, our work in this sense is a base for comparing the effects of these queue disciplines on the models for better use and adoption in real life business applications. The rest of the article is organized as follows: in Section 2, we describe the model together with the preliminary assumptions employed. Section 3 deals with the steady state analysis of the $M/(M+G)/2$ queue with two heterogeneous servers operating under the serial queue discipline. Here, the Probability Generating Function (PGF) of the number of customers in the system, the LST of the waiting time distribution and their mean values have been obtained. Similarly, a numerical illustration is provided to support the results on mean waiting times. Section 4 provides an analysis via the supplementary variable technique and LST methods on the $M/M,G/2$ queue where a necessary condition under which the steady state behavior of the $M/(M+G)/2$ and that of the $M/M,G/2$ are identical. In particular; when the mean queue length and the mean waiting time values are almost equal. Section 5 highlights the various special features of the proposed methodology and its future scope.

2. Modeling and preliminary assumptions

A representation of the $M/G/2$ queuing system under the serial queue discipline with servers is modeled as an $M/(M+G)/2$ queue with a Poisson arrival process and a general service time process on the two servers in the system.[8] Similarly, an $M/G/2$ queuing system under the parallel queue discipline is modeled as the $M/M,G/2$ queue with parallel servers.

Two alternative queue disciplines (*serial* and *parallel*) are proposed in this work. We suppose that an arbitrary shop whose queuing features are that of the $M/M/1$ type (here server-I) is experiencing an increase in demand resulting from the increasing needs of customers. As a remedy, it can be decided that an additional general server (server-II) be put in place to operate jointly[9] with the existing server in series or be placed in parallel to the initial server. In each case, one can infer that, some degree of service improvements will be experienced generally.

Lemma 2.1. *Suppose* $T_1 \sim \exp(\mu)$ *and* $T_2 \sim B(t)$ *denote the service times of customers in the* $M/(M+G)/2$ *queuing system with the number of customers* $N(t) \geqslant 2$. *Let* $T = min(T_1, T_2)$ *and* $D(t) = P[T > t]$ *with PDF* $f_{min}(t)$. *Then the departure rate* $r(t)$ *of the serialized servers is given by*

$$r(t) = \frac{f_{min}(t)}{D(t)} = \mu + \frac{B'(t)}{1 - B(t)} \tag{2.1}$$

Proof. Given that $T_1 \sim \exp(\mu)$ and $T_2 \sim B(t)$, let $D_1(t) = 1 - F_1(t)$ and $D_2(t) = 1 - B(t)$ be the tail service time distributions for T_1 and T_2 respectively. Given that $D(t) = P[\min(T_1, T_2) > t] = P[T > t]$ where T_1 and T_2 are independent random variables then

$$D(t) = P[\min(T_1, T_2) > t] = P[T_1 > t] + P[T_2 > t] = \Sigma_{i=1}^{2} D_i(t) \tag{2.2}$$

Consequently, the departure rate $r(t)$ of the complementary distribution function $D(t)$ with PDF $f_{min}(t)$ is

$$r(t) = \frac{f_{min}(t)}{D(t)} = \frac{f_1(t)}{1 - F_1(t)} + \frac{B'(t)}{1 - B(t)} = \mu + \frac{B'(t)}{1 - B(t)} \qquad \square \tag{2.3}$$

[4] Jointly but independently (one customer at a time) in the first model but independently paralleled in the second model.

[5] Suppose that one of his shops adopts the $M/(M+G)/2$ model under the serial queue discipline while the other shop adopts the parallel service order of the $M/M,G/2$ with service schedule following a parallel queue discipline

[6] This is one of the many physical scenarios involving the applications of the two models under the designed serial and parallel queue disciplines.

[7] For instance, the unknown function $Q_1(x)$.

[8] A realistic scenario for the serial model occurs for instance in a shop with two servers when a customer is being serviced by the first server. Upon arrival of another customer the second server joins the first server to service the first customer to hasten his service process there by reducing the waiting time of the second and subsequent customers. The other model i.e. the $(M/M,G/2)$ comes into picture if the second server decides to serve the second customer independently.

[9] In a manner similar to that of a device functioning with two components.

The Serial Queue Discipline:

Suppose that the decision reached above favors the installation of an additional general server (server-II) jointly in series with the exponential server (server-I) such that:

1. If a customer arrives during the idle state of the system, his service is immediately initiated by server-I (since server-I is faster than server-II). This customer receives service at exponential rate μ if no other customer arrives during his ongoing service period; otherwise if at least one more customer arrives, then the initial customer is served jointly by both servers but independently according to the service time distribution $F_{min}(t)$ defined in item 2(i) below.
2. As long as the system size $N(t) \geqslant 2$ at any time t, then server-II joins server-I to serve a customer in service jointly in series, otherwise server-I is the only available server in the system.

To conceptualize the above type of server interaction process between the two servers during an operational period of the $M/(M+G)/2$ model proposed in this work, suppose that there exist a shop selling distinct products with two human servers and system size $N(t) \geqslant 2$ at anytime t, such that:

i. The queuing system is busy if and only if at least one of the two servers is busy with service time distribution $F_{min}(t)$ where $F_{min}(t) = P(T < t); T = min(T_1, T_2)$ with PDF $f_{min}(t)$ and LST $f^*_{min}(s)$ given by the integral $f^*_{min}(s) = \int_0^\infty e^{-st} dF_{min}(t) = f^*_1(s) + b^*(s+\mu) - f^*_1(s)\ b^*(s+\mu)$.

ii. If the system has only one customer then that customer is served by server-I entirely at a constant rate μ without being interrupted until his service is completed.

The Parallel Queue Discipline:

For the $M/M, G/2$ model proposed here, we adopt the parallel queue discipline of Krishnamoorthy (1962) subject to the condition[10] that the mean service rates of server-I and server-II are respectively μ and μ_2.

A customer arrives to find:

1. Both servers free; he occupies server-I (assuming that server-I gives faster service on average).
2. Server-I is engaged; he waits for service from server-I whether or not server-II is free. But if the number of customers waiting for service from server-I becomes m (a positive integer), he goes to server-II for service if that server is free; otherwise he waits as the $(m+1)$th customer in the queue. Note that the first m customers in the queue will be getting service from server-I and the $(m+1)$th customer in the queue will go to server-II if that server becomes free prior to the finishing of service of the customer in server-I. Otherwise he will move up as the mth customer in the queue. Hence may decide to take service from server-I.
3. Both servers are engaged and a queue of length 'n greater than or equal to m is formed. He joins the queue as the $(n+1)$th customer. All customers after the mth customer 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 server-I engaged above.

The positive integer m is to be chosen such that it is one less than the greatest integer in the ratio $\frac{\mu}{\mu_2}$. It is clear that for this choice of m the following happens: When there are m customers waiting for service from server-I, an incoming customer finds it profitable to go to server-II if that server is free since $(m+2)\mu^{-1} < \mu_2^{-1}$. Similarly, when there are only $(m-1)$ customers waiting for service from server-I, an incoming customer will find it profitable to join the queue for service from server-I, even if server-II is free since $(m+1)\mu^{-1} < \mu_2^{-1}$. In case $\mu\mu_2^{-1}$ is an integer then $m = \mu\mu_2^{-1} - 1$ so that joining the queue for service from server-I is not any more or any less profitable than going to server-II if the server is free. But there is no harm in assuming that even in this case the customer joins the queue for service from server-I when there are only $(m-1)$ customers waiting for service. Thus, this queue discipline achieves the objective that the least amount of waiting time is spent in the system according to the conditions present upon its arrival (for this specific choice of m) and also, it reduces the violation of first-in first-out principle. Generally, results on methodologies for choosing m under any specific queue discipline is limited. However, the work of Efrosinin and Sztrik (2011) is worth mentioning here. Efrosinin and Sztrik (2011) have shown that, if m is chosen such that $m = \frac{(\mu-\lambda+\sqrt{(\mu-\lambda)^2+4\mu_2\lambda})}{(2\mu_2)}$, then an optimal value for the mean number of customers in the system under the parallel queue discipline will be attained.[11]

3. Steady state characteristics of the $M/(M+G)/2$ queue under the serial queue discipline

Consider 'the embedded time points' generated at departure instants of customers just after a service is completed by either server-I or server-II. By analogy, a Markov chain can be discovered at these points. Let this chain represent the state of the system $N_j = N(t_j)$ left behind by the jth customer upon departure epoch t_j. Then the discrete time process $\{N_j\}$ constitutes a Markov chain on the discrete state space $\mathcal{S} = \{0, 1, 2, \ldots \infty\}$. Let q_j be the stationary probability that j customers are left behind by a departing customer with a z-transform $V(z) = \sum_{j=0}^\infty q_j z^j$ and let p_j be the steady state probability that there are j customers in the system at an arbitrary instant between successive embedded points with z-transform $P(z) = \sum_{j=0}^\infty p_j z^j$. Since the number of customers in the system changes at most by ± 1 at a transition (arrival or departure) epoch, we can claim that $P(z)$ is the generating function of the system states at departure instants of customers.[12] Suppose $R_j(t) = P[N(t) = j, j \geqslant 2, t < \zeta < t + \Delta t]$ such that $p_j = \int_0^\infty \frac{R_j(t)}{1-F_{min}(t)} dt$ and $q_j = \int_0^\infty \frac{R_j(t)}{1-F_{min}(t)} dF_{min}(t)$ respectively, then application of the PASTA property of the system yields that $P(z) = V(z)$. Let α_j denote the probability that j customers arrive at departure instants of customers with PDF $f_{min}(t)$ and let δ_j denote the probability that j customers arrive at arbitrary instants when a service is in progress with PDF $f_1(t)$. Since arrivals follow a Poisson process at a steady rate λ, then for $j = 2, 3, \ldots \infty$, one can have $\alpha_j = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j f_{min}(t)}{j!} dt$ and $\delta_j = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j f_1(t)}{j!} dt$. Denote by $A(z)$ and $A_1(z)$ the z-transforms of the probability distributions $\{\alpha_j\}$ and $\{\delta_j\}$ respectively such that $A(z) = \sum_{j=0}^\infty \alpha_j z^j = f^*_{min}(\lambda - \lambda z)$ and $A_1(z) = \sum_{j=0}^\infty \delta_j z^j = f^*_1(\lambda - \lambda z)$. Now, focusing on the embedded points under equilibrium conditions, let the unit step conditional transition probability of the system going from state 'i' of the $(k-1)th$ embedded point to state j of the kth embedded point be $q_{ij} = P(N_k = j / N_{k-1} = i)$ for $i, j \in \mathcal{S}$. Then the transition probabilities will form the unit step transition probability matrix Q=(q_{ij}) below:

[10] We provide an analysis for $m = 1$ customer who might prefer to wait for server-I even when server-II is idle.

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

[12] In view of PASTA.

$$Q = \begin{pmatrix} \delta_0 & \delta_1 & \delta_2 & \delta_3 & . & . & . & . \\ \delta_0 & \delta_1 & \delta_2 & \delta_3 & . & . & . & . \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & . & . & . \\ 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & . & . \\ 0 & 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & . \end{pmatrix} \tag{3.1}$$

If we denote by $q_j = \lim_{n\to\infty} q_{ij}^n$ the equilibrium state probabilities at departure instants where q_{ij}^n represents the n-step probability of moving from state i to j such that $Q = q_{ij}$ with $\mathbf{q} = (q_0, q_1, q_2, \ldots)$ and $\mathbf{e} = (1, 1, 1, \ldots)'$, then a numerical evaluation of $\mathbf{q}$ can be done by solving the following matrix equations:

$$\mathbf{q}Q = \mathbf{q}, \mathbf{qe} = 1 \tag{3.2}$$

Now, multiplying the jth equation of $\mathbf{q}Q = \mathbf{q}$ in (3.2) by z^j and summing all the left-hand sides and the right-hand sides from $j = 0$ to $j = \infty$, one obtains that

$$V(z) = \frac{q_1 z[A_{min}(z) - A_1(z)] + q_0[A_{min}(z) - A_1(z)z]}{A_{min}(z) - z} \tag{3.3}$$

Let $\rho = \frac{\lambda}{\mu}$. Since $q_1 = \rho q_0$, $A'_{min}(1) = -\lambda f'^*_{min}(0) = \frac{\lambda}{\mu+\mu_2} = \rho_1 < 1$, $A'_1(1) = -\lambda f'^*_1(0) = \frac{\lambda}{\mu} = \rho$. At $z = 1$, we derive from (3.3) that

$$q_0 = \frac{1 - \rho_1}{1 + (\rho - \rho_1)(1 + \rho)} \tag{3.4}$$

And the mean number of customers $E(N)$ present in the system at a random point or at a departure epoch is

$$E(N) = \rho + \frac{\rho_1^2}{(1 - \rho_1)} + \frac{(\rho - \rho_1)[\rho + \rho_1(1 + \rho)]}{1 + (\rho - \rho_1)(1 + \rho)} \tag{3.5}$$

Similarly, the mean waiting time $\overline{W}$[13] of customers in the system can be obtained using the well-known Little's formula $\lambda\overline{W} = E(N)$.

For a numerical illustration, suppose that λ varies from 5*to*12 as in Table 1 below while $\mu = 8.0, \mu_2 = 7.5$. The following numerical results corresponding to $\rho, \rho_1, q_0, E(N)$, and $\overline{W}$ are obtained[14]:

Table 1
Mean queue length $E(N)$ and mean waiting time $\overline{W}$ $\mu = 8.0$ and $\mu_2 = 7.5$.

λ	ρ	ρ_1	q_0	$E(N)$	$\overline{W}$
5	0.625	0.322581	0.454208	1.011633193	0.202326639
6	0.750	0.387097	0.374846	1.311295347	0.218549224
7	0.875	0.451613	0.305704	1.653291986	0.236184569
7.2	0.9	0.464516	0.293027	1.727752316	0.239965599
7.5	0.9375	0.483871	0.274697	1.843814969	0.245841996
7.7	0.9625	0.496774	0.26292	1.924332864	0.249913359
7.9	0.9875	0.509677	0.25149	2.007571644	0.254122993
8	1	0.516129	0.245902	2.050273224	0.256284153
8.2	1.025	0.529032	0.234975	2.137977356	0.260728946
10	1.25	0.645161	0.150299	3.115150221	0.311515022
11	1.375	0.709677	0.112522	3.898952161	0.354450196
12	1.5	0.774194	0.080229	5.04031928	0.420026607

From Table 1, it can be seen that both the mean queue length $E(N)$ and the mean waiting time $\overline{W}$ steadily increases with increase in λ. Similarly, the stationary q_0 values decrease with increase in λ as expected.

4. Steady state characteristics of the $M/M, G/2$ queue under the parallel queue discipline and $m = 1$

We will now discuss the steady state analysis of the $M/M, G/2$ queue under the parallel queuing discipline outlined in section two given that $m = 1$ customer. The analysis is carried out using the past service time of the customer being served by server-II as a supplementary variable.

Denote by N the steady state number of customers in the system and by ζ the steady state past service time[15] of the current customer on server-II. Looking at the system at departure instants, then the bi-variate process $\{N, \zeta\}$ is a Markov process with state space $\Im = \{0, 1, 2, \ldots\} \times [0, \infty)$. Suppose that P is a probability measure such that

$$R_0 = P[\textit{Both servers are idle}] \tag{4.1}$$
$$R_{1,0} = [\textit{Server-I is busy; server-II is idle; } N = 1] \tag{4.2}$$
$$R_{0,1}(\eta) = P[\textit{Server-I is idle, Server-II is busy } N = 1; \quad \eta \leqslant \zeta < \eta + d\eta] \tag{4.3}$$
$$R_{1,1}(\eta) = P[\textit{Both servers are busy}; N = 2, \ \eta \leqslant \zeta < \eta + d\eta] \tag{4.4}$$
$$R_{1,1,0} = P[\textit{servers-I is busy, } N = 2, \textit{ and server-II is idle}] \tag{4.5}$$

and that

$$R_{1,1,1}(\eta) = P[\textit{Both servers are busy, } N = 3, \textit{ and } \eta \leqslant \zeta < \eta + d\eta] \tag{4.6}$$

Note: We assign η to R_j only when server-II is busy. Given that two or more customers are present in the system and that their past service time lies in $(\eta, \eta + d\eta)$ then in steady state, $R_j(\eta) \to R_j$ as seen in (4.7)–(4.10) and (4.11) below.

Now, if $R_j = P[N = j]$ is the stationary probability that there are j-customers in the system. Then

$$R_0 = P[N = 0] = R_{0,0} \tag{4.7}$$
$$R_1 = P[N = 1] = R_{1,0} + R_{0,1} \tag{4.8}$$
$$R_2 = P[N = 2] = R_{1,1,0} + R_{1,1} \tag{4.9}$$
$$R_3 = P[N = 3] = R_{1,1,1} \tag{4.10}$$
$$R_j = P(N = j) \tag{4.11}$$

Since if $j \geqslant 2$, *we have* $j = (1, 1)$ *or* $(1, 1, 0)$, $j = (1, 1, 1)$, *and* $j = 4, j = 5, \ldots$ then for $\zeta > 0$, we can write

$$R_j(\eta) = P[N = j, \quad \eta \leqslant \zeta \leqslant \eta + d\eta] \tag{4.12}$$

Furthermore, let

$$Q_j(\eta) = \frac{R_j(\eta)}{1 - B(\eta),} \quad j \geqslant 2. \tag{4.13}$$

[16]such that

$$Q_j^*(s) = \int_0^\infty e^{-s\eta} \frac{R_j(\eta)}{1 - B(\eta)} d\eta, \quad \widetilde{Q_j}(\eta) = \beta \widetilde{R_j}(\eta) \tag{4.14}$$

where

$$\widetilde{R_j}(\eta) = \int_0^\infty Q_j(\eta) dB(\eta) \tag{4.14a}$$

Here, the service time distribution is either $F_1(\eta)$ or $B(\eta)$ throughout since each service period depends on the server providing service with departure rates

$r_1(\eta) = \frac{f_1(\eta)}{D_1(\eta)} = \mu$ or $r_2(\eta) = \frac{b(\eta)}{D_2(\eta)} = \frac{B'(\eta)}{1 - B(\eta)}$.

PGF of the Stationary Customer Distribution P(z)

In this subsection, we explain a methodology for computing the sequence of probability functions $\{R_j\}$ representing the steady state number of customers in the system leading to the generating function $P(z) = \sum_0^\infty R_j z^j$ but not $P(z)$ itself.[17] The standard argument discussed above shows that the sequence $\{R_j\}$ satisfies a set of

[13] Inclusive of the service time.
[14] The range of the arrival rate is varied from low to high so that the results of this work cover all cases of the utilization parameter up to a value sufficiently close to unity.
[15] The supplementary variable.
[16] Strictly, for $j \geqslant 2$ when server-II is busy only.
[17] The equivalent of P(z).

steady state equations given in Appendix A. Using (4.14) and (4.14a) in that sense, we can rewrite them as given below:

$$\lambda R_0 = \mu R_{1,0} + \frac{1}{\beta}\widetilde{Q}_{0,1} \tag{4.15}$$

$$(\lambda+\mu)R_{1,0} = \lambda R_0 + \mu R_{1,1,0} + \frac{1}{\beta}\widetilde{Q}_{1,1} \tag{4.16}$$

$$(\lambda+\mu)R_{1,1,0} = \lambda R_{1,0} + \frac{1}{\beta}\widetilde{Q}_{1,1,1} \tag{4.17}$$

$$Q'_{0,1}(\eta) = -\lambda Q_{0,1}(\eta) + \mu Q_{1,1}(\eta) \tag{4.18}$$

$$Q_{0,1}(0+) = 0, \quad j = 1 \tag{4.19}$$

$$Q'_{1,1}(\eta) = -(\lambda+\mu)Q_{1,1}(\eta) + \lambda Q_{0,1}(\eta) + \mu Q_{1,1,1}(\eta) \tag{4.20}$$

$$Q_{1,1}(0+) = 0, \quad j = 2 \tag{4.21}$$

$$Q'_{1,1,1}(\eta) = -(\lambda+\mu)Q_{1,1,1}(\eta) + \lambda Q_{1,1}(\eta) + \mu Q_4(\eta) \tag{4.22}$$

$$Q_{1,1,1}(0+) = \lambda R_{1,1,0} + \frac{1}{\beta}\widetilde{Q}_4, \quad j = 3 \tag{4.23}$$

For $j \geqslant 4$, we have

$$Q_j(\eta) = -[\lambda+\mu]Q_j(\eta) + \lambda Q_{j-1}(\eta) + \mu Q_{j+1}(\eta) \tag{4.24}$$

$$Q_j(0+) = \frac{1}{\beta}\widetilde{Q}_{j+1}, \quad j \geqslant 4 \tag{4.25}$$

To solve (4.18)–(4.24) and (4.25), apply the Laplace operator on the mentioned differential equations coupled with the appropriate boundary conditions. Then one obtains that

$$sQ^*_{0,1}(s) + \lambda Q^*_{0,1}(s) = \mu Q^*_{1,1}(s) \tag{4.26}$$

And for $j \geqslant 2$, we have

$$sQ^*_{1,1}(s) + (\lambda+\mu)Q^*_{1,1}(s) = \lambda Q^*_{0,1}(s) + \mu Q^*_{1,1,1}(s) \tag{4.27}$$

$$sQ^*_{1,1,1}(s) + (\lambda+\mu)Q^*_{1,1,1}(s) = \lambda Q^*_{1,1}(s) + \mu Q^*_4(s) + \lambda R_{1,1,0} + \frac{1}{\beta}\widetilde{Q}_4 \tag{4.28}$$

$$sQ^*_4(s) + (\lambda+\mu)Q^*_4(s) = \lambda Q^*_{1,1,1}(s) + \mu Q^*_5(s) + \frac{1}{\beta}\widetilde{Q}_5 \tag{4.29}$$

$$sQ^*_j(s) + (\lambda+\mu)Q^*_j(s) = \lambda Q^*_{j-1}(s) + \mu Q^*_{j+1}(s) + \frac{1}{\beta}\widetilde{Q}_{j+1}\ j \geqslant 5 \tag{4.30}$$

To solve (4.26)–(4.29) and (4.30), we take advantage of the following lemma.

Lemma 4.1. *Given that the traffic condition $\lambda < \mu + \frac{1}{\beta}$ holds, then in a busy period*

$$Q^*_j(0) = \widetilde{Q}_j, \quad j = 2,3,\ldots, \tag{4.31}$$

Proof. Suppose that a busy period is in progress such that the time T_n between two successive departures from server-II is given as $T_n = t_n - t_{n-1},\ n = 1,2,3,4,\ldots$ Then for any number of departure $n \geqslant 1$ during this busy period, the service period is a probabilistic replication of the initial period t_1. More so, if the queue length process at any time t during this period is $N(t)$, we are assured that $N(t)$ would reach steady state starting at $t = 0, N(0) \geqslant 2$ in the long run. Consequently, $N(t)$ is a regenerative process over t on state space $\mathbb{S}_2 = 2,3,\ldots$ and $T_n = t_n - t_{n-1}$ is an underlying renewal process at t_j each time a departure occurs on server-II. Now, given that $\lambda < \mu + \frac{1}{\beta}$ holds, then upon service completion on server-II, the state probability is

$$R_j(t) = P[N(t) = j,\ j = 2,3,\ldots] \tag{4.32}$$

and if the past service time is η at a time t, then the conditional probability that there are j customers in the system is

$$R_j(t,\eta) = P[N(t) = j | t = \eta,\ j = 2,3,\ldots] \tag{4.33}$$

Let

$$Q_j(t,\eta) = \frac{R_j(t,\eta)}{1 - B(t,\eta)} \tag{4.34}$$

So that

$$Q_j(t)(1 - B(t)) = R_j(t) = P[N(t) = j | t_1 > t] \tag{4.35}$$

Then

$$\sum_{j=2}^{\infty} Q_j(t) = P[t_1 > t] = 1 - B(t) \tag{4.36}$$

and

$$Q_j(t) = \int_0^{\infty} P[N(t) = j | t_1 > t] = \int_0^{\infty} P[N(t) = j, t_1 > t | t_1 = \eta] \tag{4.37}$$

which can be simplified to

$$Q_j(t) = \int_{\eta}^{\infty} P[N(t) = j | t_1 = \eta] dB(\eta) \tag{4.38}$$

Here, it is seen that

$$R_j(t) = \int_0^{\infty} P[N(t) = j, t_1 = \eta] dB(\eta) \tag{4.39}$$

Thus, by conditioning on t_1, under steady state conditions, it can be shown that the following renewal equation is satisfied.

$$R_j(t) = Q_j(t) + \int_0^t R_j(t-x) dB(x) \tag{4.40}$$

This renewal equation has a unique solution of the form

$$R_j(t) = Q_j(t) + \int_0^t Q_j(t-x) dM(x) \tag{4.41}$$

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

$$\lim_{t\to\infty} R_j(t) \to \frac{1}{\beta}\int_0^{\infty} Q(x) dx \tag{4.42}$$

The integral in (4.42) is the probabilistic version of $\widetilde{Q}_j$ when the mean service time on server-II is β. Thus,

$$\widetilde{R}_j\beta = \widetilde{Q}_j = Q^*_j(0) \tag{4.43}$$

Thus, the lemma holds. □

The Stationary PGF P(z)

If Lemma 4.1 is applied in (4.15)–(4.29) and (4.30) and then simplified as $s \to 0$, one can obtain a compact expression for each member of the sequence $\{R_j\}$ subject to the condition that $\lambda \leqslant \mu + \frac{1}{\beta}$. The results are reported in Appendix B. A summarized version for the results is given below for two real values $a = (\mu^2 + \lambda\mu + \lambda^2)$ and $b = (a + \lambda\mu + 2\lambda^2)$.

$$R_1 = R_{0,1} + R_{1,0} = \left(\frac{\lambda}{b}\right)\left[\frac{\lambda^2}{\mu_2} + \left(\frac{\lambda}{\mu}\right)[a + \lambda^2 + \lambda\mu]\right]R_0 \tag{4.44}$$

$$R_2 = R_{1,1,0} + R_{1,1} = \left(\frac{\lambda}{b}\right)\left[\frac{\lambda}{\mu}\right]\left(\frac{1}{\mu_2}\right)\left[\frac{a}{\mu} + \lambda^2\right]R_0 \tag{4.45}$$

$$R_{1,1,1} = \left(\frac{\lambda}{b}\right)\left[\frac{\lambda^2}{\mu_2}\right]\left(\frac{\lambda}{\mu}\right)^2 R_0 \tag{4.46}$$

$$R_4 = \left[\frac{\lambda^3 - \mu_2 a}{\mu + \mu_2}\right]\left[\frac{\lambda^3}{\mu_2 b}\right]R_0 \tag{4.47}$$

$$R_j = \left[\rho_1 R_{(j-1)}\right] \text{ for } j \geqslant 5 \tag{4.48}$$

$$R_j = \left[(\rho_1)^{(j-4)}\right]\left(\frac{\lambda^3 - \mu_2 a}{\mu + \mu_2}\right)\left[\frac{\lambda^3}{\mu_2 b}\right]R_0 \text{ for } j \geqslant 4 \tag{4.49}$$

$$\sum_{j=4}^{\infty} R_j = \frac{R_4}{1-\rho_1} = \left[\frac{\lambda^3 - \mu_2 a}{\mu + \mu_2}\right]\left[\frac{\lambda^3}{\mu_2 b}\right]\frac{R_0}{(1-\rho_1)} \tag{4.50}$$

Similarly, the generating function $P(z)$, the mean queue length $E(N)$ and the mean waiting time $\overline{W}$ are respectively given by

$$P(z) = \sum_{j=0}^{\infty} R_j z^j = R_0 + R_1 z + R_2 z^2 + R_3 z^3 + \frac{R_4 z^4}{(1-\rho_1 z)} \tag{4.51}$$

$$E(N) = R_1 + 2R_2 + 3R_3 + R_4\left[\frac{4-3\rho_1}{(1-\rho_1)^2}\right] \tag{4.52}$$

$$\overline{W} = \frac{(R_1 + 2R_2 + 3R_3) + R_4\left[\frac{4-3\rho_1}{(1-\rho_1)^2}\right]}{\lambda} \tag{4.53}$$

Lemma 4.2. *Suppose $\lambda = \mu$. Then the underlying realization $\{N_k = j\}$ of the Markov Chain $\{N_k\}$ is ergodic if and only if $\mu > 3\mu_2$.*

Under the stability condition $\lambda < \mu + \mu_2$ i.e. $\rho_1 < 1$, the realization $\{N_k = j\}$ of the Markov chain $\{N_k\}$ is ergodic if and only if each $P(N = j) = R_j$ is positive inclusive of $R_4 = \left[\frac{\lambda^3 - \mu_2 a}{\mu + \mu_2}\right]\left[\frac{\lambda^3}{\mu_2 b}\right]R_0$. This implies that $\frac{\lambda^3 - \mu_2 a}{\mu + \mu_2} > 0$. Now, given that $\lambda = \mu$, then the lemma holds.

Lemma 4.3. *The stationary distribution $\{R_j = P(N = j)\}$ of the system size of the $M/M, G/2$ queue exists if and only if $(\lambda^3 > \mu_2 a)$ holds where $a = \lambda^2 + \lambda\mu + \mu^2$.*

Proof. Since R_4 is proportional to $\lambda^3 - \mu_2 a$ and is positive definite (being a probability value), it is trivial that the stationary distribution $\{R_j = P(N = j)\}$ of the system size of the $M/M, G/2$ queue exist if $(\lambda^3 > \mu_2 a)$ holds. Conversely, suppose $(\lambda^3 > \mu_2 a) > 0$. Then R_4 is positive definite and so it is proportional to $(\lambda^3 - \mu_2 a)$. □

5. Numerical approximations

For a comparative study on the mean number of customers $E[N]$ and the mean waiting times $\overline{W}$ of the two models namely; the $M/(M+G)/2$ and the $M/M, G/2$ queues, we suppose that λ varies from $15.11 to 15.81$ while $\mu = 8.4$ and $\mu_2 = 7.5$. Tables 2a and 2b below summarize the approximate values for ρ, ρ_1 leading to $E(N)$ and $\overline{W}$ for the two models studied in this work.[18]

6. Discussions and scope

Note that under equilibrium conditions, there is an insignificant difference between

(i) $E(N)_{M/M,G/2}$ and $E(N)_{M/(M+G)/2}$.
(ii) $\overline{W}_{M/M,G/2}$ and $\overline{W}_{M/(M+G)/2}$.

Thus, one can conclude that though, some violations of the FCFS principle occurred because of heterogeneity of servers in $M/G/2$ queues generally as pointed out by Krishnamoorthy (1962), the two alternative queue disciplines *serial* and *parallel* here minimize such violations in the long run. This is because the steady state characteristics[19] for the $M/M, G/2$ queue under the parallel queue discipline and that of the $M/(M+G)/2$ queue under the serial queue discipline differ insignificantly as observed in Tables 2a and 2b.

[18] The stability condition $\lambda < (\mu + \mu_2)$ holds since $(\mu + \mu_2) = 15.9 > 15.81 = \lambda_{max}$.
[19] Both the mean queue length and the mean waiting time.

Table 2a
Mean queue length distributions $E[N]$.

λ	ρ	ρ_1	$E(N)_{M/M,G/2}$	$E(N)_{M/(M+G)/2}$
15.11	1.7988	0.9503	19.9475864	21.09605988
15.21	1.8107	0.9566	23.2146864	24.02767247
15.31	1.8226	0.9629	27.4100119	27.94805302
15.41	1.8345	0.9692	33.1532281	33.4625759
15.51	1.8464	0.9755	41.6807379	41.79751239
15.61	1.8583	0.9818	55.9176152	55.87054254
15.71	1.8702	0.9881	84.9295996	84.7418309
15.81	1.8821	0.9943	178.049126	177.7389403

Table 2b
Mean waiting time distributions $\overline{W}$.

λ	ρ	ρ_1	$\overline{W}_{M/M,G/2}$	$\overline{W}_{M/(M+G)/2}$
15.11	1.7988	0.9503	1.32015792	1.396165445
15.21	1.8107	0.9566	1.52627785	1.57972863
15.31	1.8226	0.9629	1.79033387	1.82547701
15.41	1.8345	0.9692	2.15140997	2.171484484
15.51	1.8464	0.9755	2.68734605	2.694875074
15.61	1.8583	0.9818	3.5821662	3.579150707
15.71	1.8702	0.9881	5.40608519	5.394133094
15.81	1.8821	0.9943	11.2618041	11.24218471

Similarly, we infer from these results that, if ρ_1 is relatively far from one, then it is operationally better to allocate a customer to a server instead of joint service when another customer is present. As can be seen in the tables above, both the mean queue length and the mean waiting time in the latter under the parallel queue discipline is stationary smaller than that of the former. Hence, the parallel queue discipline is a better alternative when arrival rates is not approaching the combined server rates.

Similarly, if $\rho_1 \to 1$, then it is operationally better to join service than allocating servers to distinct customers. This is evident from Tables 2a and 2b that when $\rho_1 \to 1$, the expected behavior of the serial model becomes better than that of the parallel model. The results obtained here can be applied in service systems where customer distribution are required for better practice. Most importantly is the new result of our work that under serial queue discipline applied on the two connected servers as in the $M/(M+G)/2$ and parallel queue discipline applied as in the $M/M, G/2$, given that $\lambda < (\mu + \mu_2)$ holds, then the behavior of the $M/(M+G)/2$ and that of the $M/M, G/2$ are identical if and only if $\lambda^3 > \mu_2(\mu^2 + \lambda\mu + \lambda^2)$ holds. This ensures that the associated Markov chain for the customer distribution is ergodic. There is a scope to studying the models discussed here via Markov-renewal theory. This will ensure an equality relationship between the arrival distribution and those specified in Lemma 4.1. Finally, we are grateful to all the sources of literature used and to those that encourage us to see the great work of Boxma et al. (2002) model in a different way.

Acknowledgements

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

Appendix A

It can easily be verified that the steady state probability functions $R_0, R_1 \ldots$ and $R_j(\eta)$ for $(\eta \geqslant 0)$ satisfy the below differential equations:

$$\lambda R_0 = \mu R_{1,0} + \int_0^\infty R_{0,1}(\eta)\frac{dB(\eta)}{1-B(\eta)} \tag{6.1}$$

$$(\lambda+\mu)R_{1,0} = \lambda R_0 + \mu R_{1,1,0} + \int_0^\infty R_{1,1}(\eta)\frac{dB(\eta)}{1-B(\eta)} \tag{6.2}$$

$$R'_{0,1}(\eta) = -(\lambda + \frac{dB(\eta)}{1-B(\eta)})R_{0,1}(\eta) + \mu R_{1,1}(\eta) \tag{6.3}$$

$$R_{0,1}(0+) = 0, \quad j = 1 \tag{6.4}$$

$$(\lambda+\mu)R_{1,1,0} = \lambda R_{1,0} + \int_0^\infty R_{1,1,1}(\eta)\frac{dB(\eta)}{1-B(\eta)} \tag{6.5}$$

$$R'_{1,1}(\eta) = -\left(\lambda+\mu+\frac{dB(\eta)}{1-B(\eta)}\right)R_{1,1}(\eta) + \lambda R_{0,1}(\eta) + \mu R_{1,1,1}(\eta) \tag{6.6}$$

$$R_{1,1}(0+) = 0, \quad j = 2 \tag{6.7}$$

$$R'_{1,1,1}(\eta) = -\left(\lambda+\mu+\frac{dB(\eta)}{1-B(\eta)}\right)R_{1,1,1}(\eta) + \lambda R_{1,1}(\eta) + \mu R_4(\eta) \tag{6.8}$$

$$R_{1,1,1}(0+) = \lambda R_{1,1,0} + \int_0^\infty R_4(\eta)\frac{dB(\eta)}{1-B(\eta)}, \quad j = 3 \tag{6.9}$$

For $j \geqslant 4$, we have

$$R'_j(\eta) = -\left(\lambda+\mu+\frac{dB(\eta)}{1-B(\eta)}\right)R_j(\eta) + \lambda R_{j-1}(\eta) + \mu R_{j+1}(\eta) \tag{6.10}$$

$$R_j(0+) = \int_0^\infty R_{j+1}(\eta)\frac{dB(\eta)}{1-B(\eta)} \tag{6.11}$$

Appendix B

$$R_0 = R_0 \tag{6.12}$$

$$R_{1,0} = \left[\frac{\mu_2}{\lambda}\left(\frac{\lambda^2+\lambda\mu+a}{\lambda\mu}\right)\right]R_{0,1} \tag{6.13}$$

$$R_{1,1,0} = \left[\left[\frac{\mu_2}{\lambda}\right]\left(\frac{a}{\mu^2}\right)\right]R_{0,1} \tag{6.14}$$

$$R_{0,1} = \left[\frac{\lambda^3}{\mu_2[a+2\lambda^2+\lambda\mu]}\right]R_0 \tag{6.15}$$

$$R_2 = R_{1,1,0} + R_{1,1} = \left[\frac{\lambda^2}{\mu\mu_2[a+2\lambda^2+\lambda\mu]}\left(\frac{a}{\mu}+\lambda^2\right)\right]R_0 \tag{6.16}$$

$$R_1 = R_{0,1} + R_{1,0} = \left[\frac{\lambda^3}{\mu_2[a+2\lambda^2+\lambda\mu]} + \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda^2+\lambda\mu+a}{2\lambda^2+\lambda\mu+a}\right)\right]R_0 \tag{6.17}$$

$$R_{1,1,1} = \left(\frac{\lambda}{\mu}\right)^2\left[\frac{\lambda^3}{\mu_2[a+2\lambda^2+\lambda\mu]}\right]R_0 \tag{6.18}$$

$$R_4 = \left[\frac{\lambda^3-\mu_2 a}{\mu+\mu_2}\right]\left[\frac{\lambda^3}{\mu_2[a+2\lambda^2+\lambda\mu]}\right]R_0 \tag{6.19}$$

$$\sum_{j=4}^{\infty} R_j = \frac{R_4}{1-\rho_1} = \left[\frac{\lambda^3-\mu_2 a}{\mu+\mu_2}\right]\left[\frac{\lambda^3}{\mu_2[a+2\lambda^2+\lambda\mu]}\right]\frac{R_0}{(1-\rho_1)} \tag{6.20}$$

References

Alexander, J. B., Marcus, R., & Cristobal, M. (2014). Flow shop scheduling with heterogeneous workers. *European Journal of Operational Research, 237*(2), 713–720. SciVerse ScienceDirect, <www.elsevier.com/locate/ejor>.

Boxma, O. J., Deng, Q., & Zwart, A. P. (2002). Waiting time asymtotics of the M/G/2 queue with heterogeneous servers. *Queuing Systems, 40*, 5–31.

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

Emrah, B. E., Ceyda, O., & Irem, O. (2013). Parallel machine scheduling with additional resources: Notation, classification, models and solution methods. *European Journal of Operational Research, 230*, 449–463.

Hoksad, P. (1978). Approximation for the M/G/m queue. *Operations Research, 26*, 511–523.

Hoksad, P. (1979). On the steady state solution of the M/G/2 queue. *Advances in Applied Probability, 11*, 240–255.

Kim, J. H., Ahn, H.-S., & Righter, R. (2011). Managing queues with heterogeneous servers. *Journal of Applied Probability, 48*(2), 435–452.

Krishnamoorthy, B. (1962). On Poisson queue with two heterogeneous servers. *Operations Research, 2*(3), 321–330.

Kumar, K. B., Madheswari, P. S., & Venkatakrishnan, K. S. (2007). Transient solution of an M/M/2 queue with heterogeneous servers subject to catastrophes. *Information and Management Sciences, 18*(1), 63–80.

Senthamaraikannan, K., & Sivasamy, R. (1997). Embedded processes of a Markov renewal bulk service queue. *Asia Pacific Journal of Operational Research, 11*(1), 51–65.

Shenkar, S., & Weinrib, A. (1989). The optimal control of heterogeneous queuing systems: A paradigm for load-sharing and routing. *IEEE Transactions on Computers, 38*(12), 1724–1736.

Singh, V. P. (1968). Two-server Markovian queues with balking: Heterogeneous vs. homogeneous servers. *Operations Research, 18*(1), 145–159.

Tijms, H. C., Vaan Hoorn, M. H., & Federgruen, A. (1981). Approximation for the steady state probabilities in the M/G/C queue. *Advances in Applied Probability, 13*(1), 186–206.