International Journal of Applied Mathematics

Volume 27 No. 5 2014, 461-472
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v27i5.4

A QUEUE DISCIPLINE DEVOID OF THE LITERATURE

Sulaiman Sani[1] §, Onkabetse A. Daman[2]

[1]Department of Mathematics
University of Botswana
PB 0022 Gaborone, BOTSWANA

[2]Department of Mathematics
Block 208 Room 203
University of Botswana
Gaborone, BOTSWANA

Abstract: In queuing literature, several queue and service disciplines have been extensively studied along side various queuing systems. However, a realistic queue and service discipline alludes the literature since 1909. In this paper, we describe the discipline on the $M/G/2$ queue and give an account of its stationary behavior under the assumption that the mean service rate of the two servers are nowhere equal.

AMS Subject Classification: 60K25
Key Words: $M/G/2$ queue, $M/(M+G)/2$ queue, serial service, varying service rates

1. Introduction

Since 1909 when the Danish mathematician Agner Krarup Erlang published his fundamental paper on congestion in telephone traffic to date, one assumption is normally employed by researchers in queuing theory. This is the assumption that *during the service process of a customer only one server provides him service*. In some instances, the assumption is diversified to include a group of customers for a server as in bulk service models developed over the years. One

Received: July 20, 2014

§Correspondence author

may think that it is the sole schedule possibly realistic going by the quantum of models developed with this one customer (or a group) one server assumption. While this assumption holds good in physical systems of business, communications and telecommunications systems where servers are most often in parallel, there are other service schedules in real life business applications that necessarily do not allocate a single server to serve a customer. For instance there are service shops where one may come across a service process involving both the salesman and the boss (owner) providing services together such that 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 serially that is; working jointly and independently with the salesman to serve the initial customer. As a schedule, if there are up to two customers in the shop then the boss joins the salesman in service jointly but independently (one customer at a time being served by more than one server) so that customers do not get bored by excessive waiting and choose to renege or balk thereby making the business loose market.

Clearly, this service process is distinctively different from the well known parallel server setup in that in the schedule exemplified above, two servers simultaneously serve a customer anytime whenever the number of customers becomes equal to or exceeds the number of servers in the system. One may refer to this service approach as *serial service*. A remarkable feature of the serial service schedule is that, the service process does not violate the classical First Come First Served (FCFS) queue discipline. Mostly found in shops, malls, supermarkets, offices, banks[1] and other business outfits where heterogeneity of servers is evident owing to factors such as degree of usage, experience, age, preferences etc. These factors are the building blocks for server discrimination amongst customers generally. When a businessman engages an experienced salesman and an apprentice, a senior doctor and a junior doctor, a professor and a bachelor, etc as servers in a queuing system with a parallel structure, then the FCFS queue discipline is violated, see Krishnamoorthy [2] and Benavides et al. [3][2]. This is because if the less reliable server (inferior server) is randomly chosen by a customer ahead of the more reliable one (superior server), then there is a tendency that customers that entered the system after him will check out earlier thereby making him wait longer than expected. This type of service schedule creates dissatisfaction amongst customer groups affected by such randomized choice of servers and eventually will lead to abandonment, reneging, balking, etc thereby reducing profits. Thus, there is the need to construct appropriate

[1]For instance, in a shopping mall during training session.
[2]Implied in this work also.

service schedules that share the excess service times amongst customers equivalently given that the FCFS discipline is maintained. A serial service schedule describe in this work is a realistic alternative because of its adherence to the FCFS discipline.

2. Traces in the Literature

Unfortunately, not so many works are found in the literature on this kind of service schedules[3] when compared with the volumes of work [For a survey see Kim et al. [4], Kumar et al. [6], Shenkar and Weinrib [10], Singh [11], etc.] on one server-one customer as in parallel case systems (see Emrah et al. [1]) even though the service time process in a large number of production and business centers are realistically in series. From the onset, the serial service schedule appeared first in the literature about 50 years after the pioneering works of A.K. Erlang[4] in Krishnamoorthy [2] as a viable alternative for reducing waiting times in uni-server queuing systems experiencing increasing service capacity. Unfortunately, it was not analyzed. Between 1963-2012, it appears there are no works in the literature on queuing systems following this schedule until early 2013 when Sivasamy and Kgosi [9] studied a queuing model under the serial service structure and provided the stationary mean analysis. It was shown that the analysis carried out could be used in designing the shortest processing time in queuing systems found in operating machines. Recently in Sivasamy et al. [8], we compared the steady state mean performance of the $M/G/2$ queuing system under the serial and parallel service schedules. 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. It was shown that for certain values of the server rates and under heavy traffic conditions,[5] it is operationally better to serially join service than allocating servers to distinct customers as in the one customer one server case.

Our aim in this paper is to broaden the scope of the serial service discipline by providing an in-depth analysis of its stationary behavior under any interestingly realistic condition. This will define a robustness structure for the service

[3]It deludes the literature. For, our investigation shows that there are no works in queuing system modeling that adopts such serial service as schedules for servicing customers.

[4]Implicitly in the work of Krishnamoorthy [2] on the Poisson queue with two heterogeneous servers.

[5]When the joint servers occupation rate $\rho_1 \to 1$.

process in question similar to others found in the literature. In addition, we will carry out performance analysis to provide a base for selecting appropriate service schedules for bettering performance, evaluating problem situations for better management practice. For instance, the owner of the shop above may wish to understand whether the serial service schedule under relatively tightening realistic service conditions can minimize waiting times better when compared with the parallel service schedule[6] commonly adopted in business applications and under what condition can the serial service be preferred, substituted, relaxed, etc.

As a justification, customers hate to wait especially if the excess waiting time is due to the in-effectiveness of a service process. Consequently, they may prefer to be served in any manner that will minimize their waiting time expectations. Similarly, business owners will be happy to understand new strategies and schedules necessary for maximizing profits. Thus, our work stands to benefit both the customer and the service provider, owners, etc.

3. Serial Service on the $M/G/2$ Queue: Modeling

In this section, we describe the serial service discipline on a two-heterogeneous server queuing system called the M/G/2 queue. The section is derived from our work in Sivasamy et al [8] on the same queue under serial service.

By the $M/G/2$ queue with a serial service process, we refer to a representation of the form $M/(M+G)/2$ where customers arrive according to a Poisson process with a mean arrival rate λ and receive service from the two serially working heterogeneous servers in the system[7]. The service times of customers is assumed to be independent of the inter-arrival times and without preemption.

For customers served by server-1 (the faster server), the service time T_1 follows the exponential distribution with mean rate α_1 i.e. $G_1(t) = P(T_1 < t) = 1 - e^{-\alpha_1 t}$ with probability density function (PDF) $g_1(t) = \frac{dG_1(t)}{dt}$ and Laplace-Stieltjes Transform (LST) $g_1^*(s)$. Similarly, for customers serviced by server-2 (the slow server), their service time distribution $G_2(t) = P[T_2 < t]$ is generally distributed with PDF $g_2(t)$, a mean $\beta = E[T_2]$ and a LST $g_2^*(s)$ given by $g_2^*(s) = \int_0^\infty e^{-st} g_2(t)dt$. We supposed that $\alpha_2 = \frac{1}{\beta}$, $\rho = \frac{\lambda}{\alpha_1}$. For stability, the condition $\frac{\lambda}{k_1\alpha_1 + k_2\alpha_2} < 1$ is necessary where $k_1, k_2 \in (0, 1)$. In the sequel, we

[6] A schedule where the two servers in the system serviced two different customers at a time.

[7] The serial service process here should not be confused with the one where a customer takes service from a system of servers one at a time. Here, all servers serve a customer simultaneously.

supposed that this condition holds so that results are tractable, see Boxma et al. [7]. In addition,

1. If a customer enters the system during idle state, his service process is initiated by server-1. This customer receives an exponential service at a mean service rate α_1 if no other customer arrives during his ongoing service period; otherwise the initial customer is served jointly by both servers but independently according to the service time distribution $G_{min}(t)$ as defined in item $2(i)$ below.

2. If the number of customers $N(t) \geq 2$ any time, then server-2 serially join server-1 to serve a customer receiving service serially.

To understand the server interaction process of the proposed $M/(M+G)/2$ model presented in this work, suppose there exist a supermarket with two salesmen of varying degree of service experience, where the customer size $N(t)$ is such that $N(t) \geq 2$ anytime with the condition that:

i. The queuing system is busy if and only if at least one of the two salesmen is busy with service time distribution $G_{min}(t)$, PDF $g_{min}(t)$ and LST $g^*_{min}(s) = \int_0^\infty e^{-st} g_{min}(t)d(t) = g_1^*(s) + g_2^*(s+\alpha_1) - g_1^*(s)\, g_2^*(s+\alpha_1)$.

ii. If the system has only one customer, then the customer is served by server-I entirely at a mean service rate α_1 without being interrupted.

4. Stationary Behavior Under Varying Service Rates

This section discusses the stationary behavior of the $M/G/2$ queue under the serial service process given that the service rates of the two servers are unequal[8]. This implies that an arbitrary service is almost surely initiated by both servers but completed by one only. For the case when the probability that a service is completed by both servers is non zero, we refer the reader to our work in Sivasamy et al. [8].

Lemma 4.1. *Suppose $P[N(t) = j]$ denotes the probability that there are j customers in an $M/(M+G)/2$ queuing system at time t. Let $P[N_1(t) = j_1]$ and $P[N_2(t) = j_2]$ denote the probability that j customers are left behind upon*

[8]This may depict servers such as a senior and a junior doctor, an experience salesman and an apprentice, etc. It is evident that the service rates are realistically unequal.

departure event in the $M/M/1$ and the $M/G/1$ queuing systems respectively. In steady state

$$P[N = j] = \upsilon_1 P[N_1 = j_1] + \upsilon_2 P[N_2 = j_2], \tag{1}$$

where $\upsilon_1 + \upsilon_2 = 1$.

Proof. Consider a process $\{N(t), \zeta(t)\}$ where $N(t)$ denotes the number of customers in the $M/(M+G)/2$ queuing system at time t and $\zeta(t)$ is the past service time already received by a customer on service. Looking at the system at departure instants when the past service time $\zeta(t) = 0$. Then $\{N(t), \zeta(t)\}$ is a Markov process, see Medhi [5]. □

Now, if $\{N(t), \zeta(t)\}$ is time continuous, then as $t \to \infty$, $\{N(t), \zeta(t)\} \to \{N, \zeta\}$. The remarks below give the properties of the stationary process $\{N, \zeta\}$.

Remark 4.1. 1. Suppose j denotes the present state of the process $\{N, \zeta\}$. Then the associated Markov chain is two-state since a transition from j resides only in $(j-1)$ and vice versa[9].

2. Similarly, $j \leftrightarrow (j-1)$. Hence, the Markov chain is irreducible[10].

3. Let k denote the number of steps to reach $(j-1)$ given that it is in j. Then the $gcd(k) = 1$. Hence, the Markov chain is aperiodic.

By the ergodic theorem, the stationary probability $P[N = j] = R_j$ will satisfy the Kolmogorov difference equations below

$$\lambda R_0 = \left(\upsilon_1\alpha_1 + \frac{(1-\upsilon_1)}{\beta}\right) R_1, \quad j = 0, \tag{2}$$

$$\left(\lambda + \upsilon_1\alpha_1 + \frac{(1-\upsilon_1)}{\beta}\right) R_j = \lambda R_{j-1} + \left(\upsilon_1\alpha_1 + \frac{(1-\upsilon_1)}{\beta}\right) R_{j+1}, \quad j \geq 1. \tag{3}$$

Applying the Markov property of the system states, we obtained that

$$R_j = \left(\frac{\lambda}{\upsilon_1\alpha_1 + \frac{(1-\upsilon_1)}{\beta}}\right)^j R_0, \tag{4}$$

[9]That is upon departure of a customer.

[10]By the rate-equality principle when a Poisson arrival occurs into the system.

where R_0 is the stationary probability that the system is idle. R_0 can be derived via the normalizing condition by summing R_j over $j = 0$ to $j = \infty$ and equating the sum to one. Upon simplification, one obtains that $R_0 = 1 - \frac{\lambda}{\upsilon_1\alpha_1 + \frac{(1-\upsilon_1)}{\beta}}$. Thus

$$R_j = \left(\frac{\lambda}{\upsilon_1\alpha_1 + \frac{(1-\upsilon_1)}{\beta}}\right)^j (1 - \bar{\rho_1}) = (1 - \bar{\rho_1})\bar{\rho_1}^j, \tag{5}$$

where $\bar{\rho_1} = \frac{\lambda}{\upsilon_1\alpha_1 + \frac{(1-\upsilon_1)}{\beta}}$ is the server utilization parameter under the convexity structure of the $M/(M+G)/2$ model.

Corollary 4.2. *Under the condition that $\alpha_1 \neq \frac{1}{\beta}$, then the stationary number of customers in the $M/(M+G)/2$ queue is unique if and only if $(0 < \upsilon_i < 1)$.*

Proof. Suppose to the contrary. That means either $\upsilon_1 = 1$ or $\upsilon_2 = 1$. Let $\upsilon_1 = 1$. This will occur if $\alpha_1 > \frac{1}{\beta}$. That means, with a unit probability all services are completed by server-1 ahead of server-2. This renders server-2 irrelevant to the stability of the system. Putting $\upsilon_1 = 1$ in (5) one obtains the stationary customer distribution for the $M/M/1$ queue. Hence, the stationary customer distribution is not unique when $\upsilon_1 = 1$. Similarly $\upsilon_1 = 0$ leads to the distribution of the $M/G/1$. Now, suppose $(0 < \upsilon_1 < 1)$. Then with positive probability, certain services are completed by server-1 ahead of server and the converse holds. Thus $(1 - \upsilon_1) > 0$. This proves the uniqueness. To show the 'only if' part, suppose that the customer distribution of the $M/(M+G)/2$ queue is unique. Then the two serial servers are necessary for the stability of the system. Consequently, $(0 < \upsilon_1 < 1)$ and $(0 < \upsilon_2 > 1)$. □

Corollary 4.3. *A stronger stability condition for the $M/(M+G)/2$ model is that the joint servers occupation rate $\rho_1 = \frac{\lambda}{\upsilon_1\alpha_1 + \upsilon_2\alpha_2}$ where υ_1 and υ_2 are non-zero probabilities.*[11]

Proof. This is trivial in view of Corollary 4.1. □

Lemma 4.4. *Denote by $E[N]$ the expected number of customers in an*

[11]Strictly greater than zero and less than one.

$M/(M+G)/2$ queuing system. Then

$$E[N] = \frac{\bar{\rho}}{1-\bar{\rho}}. \tag{6}$$

Proof. The lemma follows directly from the definition of E[N] upon further simplification[12]. □

5. Numerical Analysis

In this section we carry a numerical analysis on the stationary performance of the $M/(M+G)/2$ queuing model compared with that of two queuing models namely; the FCFS $M/M/1$ and the parallel server $M/M,G/2$ when the FCFS is slightly violated owing to heterogeneity of servers. We let λ to vary from 5.0 *to* 7.9, $\upsilon_1\alpha_1 = 8.4$ and $\frac{(1-\upsilon_1)}{\beta} = 7.5$ in the first comparison[13] and λ varies from 15.11 *to* 15.81 in the second comparison. The approximate values for ρ (server-1), ρ_1 (joint servers), the stationary mean number of customers $E(N)$ and the stationary mean waiting time $E[W]$ are given[14] below.

Table-1a: Mean queue Length Distributions $E[N]$

λ	ρ	ρ_1	$E(N)_{M/(M+G)/2}$	$E(N)_{M/M/1}$	$E(W)_{M/(M+G)/2}$	$E(W)_{M/M/1}$
5.0	0.5952	0.3144	0.4586	2.2920	0.0912	0.45840
6.0	0.7143	0.3774	0.6062	3.7500	0.1010	0.62500
7.0	0.8333	0.4403	0.7867	7.8750	0.1124	1.12500
7.2	0.8571	0.4528	0.8275	9.9000	0.1149	1.37500
7.5	0.8929	0.4717	0.8929	15.9380	0.1191	2.12507
7.7	0.9167	0.44843	0.9385	26.6292	0.1219	3.45834
7.9	0.9405	0.4969	0.9877	157.0125	0.1250	19.87500

[12]The expected waiting time could be obtained by the application of the well known Little's theorem.

[13]Here, $\upsilon_1 = 0.51, \alpha_1 = 16.5$. Similarly, $\frac{1}{\beta} = 15.3$.

[14]Including the customer on service and his service time.

Table-1b: Stationary Mean Queue length $E[N]$ and Mean Waiting Time E[W]

λ	ρ	ρ_1	$E(N)_{M/M,G/2}$	$E(N)_{M/(M+G)/2}$	$E(W)_{M/M,G/2}$	$E(W)_{M/(M+G)/2}$
15.11	1.7988	0.9503	19.94759	19.1263	1.32016	1.2658
15.21	1.8107	0.9566	23.21469	22.0434	1.52628	1.4492
15.31	1.8226	0.9629	27.41001	25.9542	1.79033	1.6952
15.41	1.8345	0.9692	33.15323	31.4676	2.15141	2.0420
15.51	1.8464	0.9755	41.68074	39.8163	2.68735	2.5671
15.61	1.8583	0.9818	55.91762	53.9451	3.58217	3.4558
15.71	1.8702	0.9881	84.92960	83.0336	5.40609	5.2854
15.81	1.8821	0.9943	178.04913	174.4386	11.26180	11.0334

6. Discussions and Scope

Tables 1a and 1b provide summaries of the stationary mean performance of the $M/(M+G)/2$ queuing model compared with that of the FCFS-M/M/1 and the $M/M,G/2$ models respectively. With reference to these tables, it can be seen that:

1. Both $E(N)_{M/(M+G)/2}$ and $E(W)_{M/(M+G)/2}$ are significantly lower than those of the FCFS-M/M/1 queue. Most importantly, the disparity between the means increases with increase in the arrival rate of customers in the system. As can be seen from table 1a, as the arrival rate λ increases from 5 to 7.9, both $E(N)_{M/(M+G)/2}$ and $E(W)_{M/(M+G)/2}$ values are strictly lower than $E(N)_{M/M/1}$ and $E(W)_{M/M/1}$ respectively. Similarly, when the server-1 occupation rate $\rho \to 1$, the mean values for $E(N)_{M/(M+G)/2}$ and $E(W)_{M/(M+G)/2}$ are stationary in contrast with that of the FCFS-M/M/1 model that shows large deviations. Thus, the mean performance of the serial service $M/(M+G)/2$ model is higher than that of the FCFS-M/M/1 model under similar service conditions and assumptions.

2. On the other hand, when the $M/(M+G)/2$ model is compared with the slightly violated FCFS $M/M,G/2$ model, it appears there is a difference in performance between the means of the two models however not significant[15]. In fact, one can see that when there is no possibility of completing a service by both servers at the same time the serial service model performs better than the parallel server $M/M,G/2$. As can be seen from table 1b, when $\lambda \geq 15.6$, both $E(N)_{M/(M+G)/2}$ and $E(W)_{M/(M+G)/2}$

[15]Relatively, when compared with the difference between the model and the earlier model compared.

are relatively lower than $E(W)_{M/M,G/2}$ and $E(W)_{M/M,G/2}$ respectively. Hence, one can conclude that, the serial service process with varying service rates[16] is a better alternative to the parallel service process dominant in the literature. From the numerical results above, one can conclude that

Lemma 6.1. *For a given arrival rate, the following in-equalities hold*

1. $E(N)_{M/(M+G)/2} < E(N)_{M/M/1}$,
2. $E(W)_{M/(M+G)/2} < E(W)_{M/M/1}$.

Proof. For $\lambda > 0$, $1 \Rightarrow 2$ directly. Table-1a shows the relationship between the expected values $E(N)_{M/(M+G)/2}$, $E(N)_{M/M/1}$ and the arrival rate λ. It can be seen that the $E(N)_{M/(M+G)/2}$ values are strictly less than $E(N)_{M/M/1}$. Thus, $E(N)_{M/(M+G)/2}$ is smaller than $E(N)_{M/M/1}$. Hence, 1 holds. By extension, 2 holds. □

Lemma 6.2. *For* $\lambda > 0$:

1. $E(N)_{M/(M+G)/2} < E(N)_{M/M,G/2}$,
2. $E(W)_{M/(M+G)/2} < E(W)_{M/M,G/2}$.

Proof. Table-1b shows the stationary values of $E(N)_{M/(M+G)/2}$ and $E(N)_{M/M,G/2}$ for certain values of the arrival rate λ. It can be seen that a difference exists between the mean values simulated. Thus $E(N)_{M/M,G/2} < E(N)_{M/(M+G)/2}$. By extension, 2 holds. □

There is a scope in providing an in-depth analysis on the asymptotics of the waiting time of the queuing model under the service process in question. This will define a maximum bound under which the arrival rate λ relates to the service rate α_1 of the most reliable server. Similarly, it will be interesting to provide analysis for the stationary behavior of this model via the Lindley's integral equation.

We are grateful to all the literature sources used in this work and to the anonymous reviewers.

[16]Though, with relatively close means.

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** (2013), 449-463.

[2] B. Krishnamoorthy, On Poisson queue with two heterogeneous servers, *Operations Research,* **2**, No 3 (1962), 321-330.

[3] J.B. Alexander, R. Marcus and M. Cristobal, Flow shop scheduling with heterogeneous workers, *European Journal of Operational Research,* (2014); Available at SciVerse ScienceDirect, www.elsevier.com/locate/ejor; http://dx.doi.org/10.1016/j.ejor.2014.02.012.

[4] J.H. Kim, H.S. Ahn and R. Righter, Managing queues with heterogeneous servers, *Journal of Applied Probability,* **48**, No 2 (2011), 435-452.

[5] J. Medhi, *Stochastic Models in Queuing Theory,* Academic Press, California (2003).

[6] 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**, No 1 (2007), 63-80.

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

[8] R. Sivasamy, O.A. Daman, and S. Sulaiman, An M/G/2 Queue subject to a minimum violation of the FCFS queue discipline, *European Journal of Operational Research,* (2014), Available at http://www.sciencedirect.com/science/article/pii/S0377221714005529; DOI: 10.1016/j.ejor.2014.06.048.

[9] R. Sivasamy and P.M. Kgosi, An $M/(M+G)/2$ Queue with heterogeneous machines operating under the FCFS queue discipline, *International Journal of Engineering and Technical Research,* **2**, No 2 (2014), 1-4.

[10] S. Shenkar and A. Weinrib, The optimal control of heterogeneous queuing systems: A paradigm for load-sharing and routing, *IEEE Transactions on Computers,* **38**, No 12 (1989), 1724-1736.

[11] V.P. Singh, Two-server Markovian queues with balking: Heterogeneous vs. homogeneous servers, *Operations Research,* **18**, No 1 (1968), 145-159.