AN M/M+G/2 QUEUE WITH HETEROGENEOUS MACHINES OPERATING UNDER FCFS QUEUE DISCIPLINE

R Sivasamy and P M Kgosi

Abstract— **This article discusses the steady analysis of an M/M+G/2 queue. All arriving customers are served either by server-1 according to an exponential service time density $f_1(x)=\mu e^{-\mu x}$ with mean rate μ or by server-2 with a general service time distribution B(t) or density function b(t)dt=dB(t) with mean rate $\mu_2=1/\beta$ or mean service time is β. Sequel to some objections raised on the use of the classical 'First Come First Served (FCFS)' queue discipline when the two heterogeneous servers operate as parallel service providers, alternative queue disciplines in a serial configuration of servers are considered in this work; the objective is that if, in a single-channel queue in equilibrium, the service rate suddenly increases and exceeds the present service capacity, install a new channel to work serially with the first channel as suggested by Krishnamoorthy (1968). Using the embedded method subject to different service time distributions we present an exact analysis for finding the 'Probability generating Function (PGF)'of steady state number of customers in the system and most importantly, the actual waiting time expectation of customers in the system. This work shows that one can obtain all stationery probabilities and other vital measures for this queue under certain additional and simple but realistic assumptions. (Abstract)**

***Index Terms*— Poisson arrival, service time distribution; PGF of queue length distribution, waitime distribution, mean queue length and mean waiting time**

I. INTRODUCTION

We analyse a class of M/G/2 queueing models to study the needs for designing serially connected system with two heterogeneous servers/machines which does not violate the FCFS principle in place of paralleled heterogeneous servers which will violate the classical FCFS queue discipline.

For instance, if there are two parallel clerks in a reservation counter who provide service with varying speeds then customers might prefer to choose the fastest among the free servers. other hand if the customer chooses the slowest server among the free servers then the one who could come subsequently may clear out earlier by obtaining service from the server providing service with faster rates what is exactly called a violation of the FCFS discipline.

Hence there is a need for the designing of alternative ways to the parallel service providers that would reduce the impacts of violating the FCFS so that the resulting waiting times of customers are identical.

Customers arrive according to a Poisson process at a rate λ for service on these servers with an expectation of spending least waiting time.

For customers serviced by server-1, their service time T_1 follows exponential distribution with rate μ i.e. $F_1(t)=P(T_1<t))=1-e^{-\mu t}$ and the probability density function(PDF) is $f_1(t)=dF_1(t)$ and hence the 'Laplace Transform(LT)' of $f_1(t)$ is $f_1^*(s)= \int_0^\infty e^{-st} f_1(t)dt = \frac{\mu}{s+\mu}; \text{Re } s>0$. Let the service times T_2 of customers serviced by server-2 follow a common distribution B(t) with PDF b(t) and mean β and the LT of b(t) is $b^*(s) = \int_0^\infty e^{-st} b(t)dt$. Both of these service time distributions are assumed to be mutually independent and each is independent of inter-arrival time distribution also.

A simple way of connecting the servers in series subject to servicing of customers according to the FCFS queue discipline is proposed below with three types of service time distributions. For a single server queue of M/G/1 type under equilibrium conditions, expecting greater demand for service in the long run it is decided to increase the service capacity of the primary single server-1 by installing an additional server-2 as detailed below:

- If an arriving customer enters into the idle system, his service is immediately initiated by server-1. This customer is then served by the server-1 at an exponential rate μ if no other customer arrives during the on-going service period;
- Otherwise i.e. if at least one more customer arrives before the on-going service is completed then the same customer is served jointly by both servers according to the service time distribution $F_{min}(t)=P(T_{min}<t)$, where $T_{min} = Min(T_1, T_2)$ and the PDF of T_{min} is $f_{min}(t)$.

Thus if system size N(t) *i.e.* the number of customers present in the system at time 't' is greater than or equal to 2, then server-2 joins server-1 to server the customer according to the service time distribution $F_{min}(t)$; otherwise the server-1 is alone available at the service counter. The LT of $F_{min}(t)$ is

$$\int_0^\infty e^{-st} f_{min}(t)dt$$

$$= f_1^*(s) + b^*(s+\mu)\{1 - f_1^*(s)\}$$

It is attempted here to study the two-server (heterogeneous) M/M+G/2 queues in the light of the above queue discipline with different service time distributions $F_{min}(t)$, $F_{max}(t)$ and

Manuscript received Feb. 09, 2014.

Prof. R Sivasamy, Statistics Department Name, University of Botswana, Gaborone, BOTSWANA. Phone 00267-3552707/Mobile No: 00267-74491917.

Dr P M Kgosi, holds both an M.Sc and a Ph D in Statistics from the University of Keny,

$F_{min+max}(t)$ while the servers are connected in a series system. We are motivated by the numerous physical applications and contributions since the above M/M+G/2 model can conceptualize well in the analysis of queues found in modern telecommunications, computer business centres, banking systems and similar service operations systems where human nature is evident. Over the years, a lot have been written on homogeneous service systems as in Hoksad [3]. The reader is referred to and [4] through [8] and many others to refresh on the heterogeneous service systems. A model of the general service type is studied by Boxma, Deng and Zwart [2], unfortunately, due to complex structuring, formations and assumptions; it could not estimate certain areas in the general case.

The probability generating function (PGF) of the number of customers present in the system, LT of the waiting time distribution and their mean values have been obtained in section II. A numerical illustration is also provided to support the results on mean waiting times. Subsection C highlights the various special features of the proposed methodology and its future scope.

II. STEADY STATE CHARACTERISTICS OF M/M+G/2 QUEUES WITH SERVICE TIME DISTRIBUTION

The number N(t) of customers present in the system at time t, does not now constitute a Markov process, see [1]. We consider here 'the embedded time points' generated at the departure instants of customers just after a service completion either by server-1 or by server-2. We consider the Markov Chain of system states at these embedded points where the state of the system is represented by the number, $N_k=N(t_k)$, of customers left behind in the queue by the k^{th} departing customer at the departure epoch 't_k'. The discrete time process $\{N_k\}$ constitutes a Markov chain on the state space $S=\{0,1,\ldots \infty\}$.

A. PGF of {Nk}

Let q_j be the steady state probability of finding 'j' customers in the system as observed by a departing customer and the z-transform of the probability distribution of $\{q_j\ ; j=0, 1, ,\ldots \infty\}$ be $\mathbf{V}(z) = \sum_{j=0}^{\infty} q_j\ z^j$.

Let α_j denote the probability of 'j' arrivals in a service time PDF $f_{min}(x)$ and δ_j denote the probability of 'j' arrivals in the exponential service time PDF $f_1(x)$. Since the arrivals come from a Poisson process at rate λ, we get, for j=2, 3, ... ∞ that

$$\alpha_j = \int_0^{\infty} \frac{e^{-\lambda x}(\lambda x)^j}{j!} f_{min}(x)\, dx \qquad (1)$$

and

$$\delta_j = \int_0^{\infty} \frac{e^{-\lambda x}(\lambda x)^j}{j!} f_1(x)\, dx \qquad (2)$$

Let the respective z-transforms of the probability distributions $\{\alpha_j\}$ and $\{\delta_j\}$ be $A_{min}(z)$ and $A_1(z)$:

$$A_{min}(z) = \sum_{j=0}^{\infty} \alpha_j z^j = f_{min}^{\ *}(\lambda - \lambda z) \quad (3)$$

and

$$A_1(z) = \sum_{j=0}^{\infty} \delta_j\ z^j = f_1^{\ *}(\lambda - \lambda z) \quad (4)$$

Focusing on the embedded points under equilibrium conditions, let the unit step conditional transition probability of the system going from state 'i' of the $(k-1)^{st}$ embedded point to state 'j' in the k^{th} embedded point be $q_{ij}=P(N_k = j\ /N_{k-1}=i)$ for i, j $\in S$. These transition probabilities will form the unit step transition probability matrix $Q=(q_{ij})$ as below:

$$Q = \begin{pmatrix} \delta_0 & \delta_1 & \delta_2 & \delta_3 & . & . & . & . \\ \delta_0 & \delta_1 & \delta_2 & \delta_3 & . & . & . & . \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & . & . & . \\ 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & . & . \\ 0 & 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & . \end{pmatrix} \quad (5)$$

Thus equilibrium state probabilities at the departure instants are given by $q_j = \lim_{n\to\infty} q_{ij}^{\ n}$ where $q_{ij}^{\ n}$ represents the n-step probability of moving from state 'i to j'. Let $\mathbf{q}=(q_0,q_1,q_2,\ldots)$ be a row vector and let $\mathbf{e}=(1,1,\ldots,1)$ be a column vector of unit elements of infinite order. Assume that $A_{min}'(1)<1$, then the stationary distribution of the state transition matrix Q, exits and is given by the unique solution of the following system of equations:

$$\mathbf{q}\, Q = \mathbf{q} \text{ and } \mathbf{q}\ \mathbf{e} = 1 \qquad (6)$$

Multiplying the j^{th} (j=0, 1, 2, …) equation of $\mathbf{q}Q=\mathbf{q}$, of (6) by z^j and summing all the left-hand sides and the right-hand sides from j=0 to j= ∞, we get the PGF(Probability generating Function) $V_{min}(z)$ of the queue length distribution $\{q_j\}$ of the sequence$\{N_k\}$.

$$V_{min}(z) = \frac{q_1 z[A_{min}(z) - A_1(z)] + q_0[A_{min}(z) - A_1(z)z]}{A_{min}(z) - z} \qquad (7)$$

Let $\mu_2 \sim \frac{1}{\beta}$ and $\rho = \frac{\lambda}{\mu}$ be used in the sequel. It is remarked that closed form expression to $b^*(s+\mu)$ cannot be obtained unless particular cases like LT of exponential, Erlangian, Phase (PH) type or hyper exponential distributions. Assume that the service time distribution of the server-2 is one of these distributions such that $b^*(\mu)$ is finite as a function of μ and μ_2 for our discussions to follow. Since $q_1= \rho q_0$, $A_{min}'(1)<1$, $A_1'(1)=\rho$ and V(1)=1, we derive from (7) that

$$q_0 = \frac{1 - A'_{min}(1)}{1 + (\rho - A'_{min}(1))(1+\rho)} \qquad (8)$$

Thus $V_{min}(z)=$

$$\frac{(z-1)A_1(z) + (1+\rho z)[A_1(z) - A_{min}(z)]}{z - A_{min}(z)} .$$

$$\frac{1 - A'_{min}(1)}{1 + (\rho - A'_{min}(1))(1+\rho)} \qquad (9)$$

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

$$\left\{ \begin{array}{l} \dfrac{\{\rho - A_{min}'(1)\}[\rho + A_{min}'(1)(1+\rho)]}{1 + \{\rho - A_{min}'(1)\}(1+\rho)} \\ + \rho + \dfrac{A_{min}'(1)^2}{1 - A_{min}'(1)} \end{array} \right. \qquad (10)$$

Denote the LT of the waiting time W distribution W(t)=P(W\(< t) by $w^*(s)$ for 0<s<1/λ.. Then replacing z by (1-s/λ) in (9) and using the fact $w^*(s)$=V(1-s/λ), we derive $w^*(s)$ as

$$w^*(s) = \frac{1-A'_{\min}(1)}{1+(\rho - A'_{\min}(1))(1+\rho)}. \qquad (11)$$
$$\frac{\{1+\rho(1-s/\lambda)\}\{\lambda f_{\min}^*(s) - \lambda f_1^*(s)\} + s f_1^*(s)}{s-\lambda\{1-f_{\min}^*(s)\}}$$

The mean waiting time $\overline{W}$ of a customer in the system obtained from (11) is found to satisfy the well-known Little's formula $\lambda\overline{W}$ =E(N).

B. Design of Shortest Delay Time Environment for M/M+G/2 Queues

The forgoing embedded methodology of analysing the M/M+G/2 queues can also be extended to design a shortest processing environment as discussed below.

Let $F_{\max}(t)=P(T_{\max}<t)$, where $T_{\max} = \mathrm{Max}(T_1, T_2)$ and the PDF of $T_{\max}$ is $f_{\max}(t)$. The LT of $f_{\max}(t)$ is

$$\int_0^\infty e^{-st} f_{\max}(t)dt$$
$$= b^*(s) - b^*(s+\mu)\{1-f_1^*(s)\} \qquad (12)$$

Considered is the joint service time PDF 'f(t)' of the two servers

$$f(t) = \pi_1 f_{\min}(t) + \pi_2 f_{\max}(t) \qquad (13)$$

in place of $f_{\min}(t)$ where π_1 and π_2 are the probability values of choosing the service time PDFs $f_{\min}(t)$ and $f_{\max}(t)$ respectively.

Let the corresponding z-transform of number of arrivals during the joint service PDF f(t) be

$$A(z) = \sum_{j=0}^{\infty} \alpha_j z^j = \pi_1 f_{\min}^*(\lambda - \lambda z) \qquad (14)$$
$$+ \pi_2 f_{\max}^*(\lambda - \lambda z)$$

Let $\rho_2 = \frac{\lambda}{\mu_2}$ and $\rho = \frac{\lambda}{\mu}$. Assume that A'(1) < 1 of (14), i.e.

$$A'(1) = \pi_1\ \rho\{1-b^*(\mu)\}$$
$$+\pi_2\{\rho_2 + \rho b^*(\mu)\} < 1 \qquad (15)$$

Now on substituting A(z) of (14) in place of $A_{\min}(z)$ of (9), one can verify that (V(z) replaces $V_{\min}(z)$)

$$V(z) = \frac{(z-1)A_1(z) + (1+\rho z)[A_1(z) - A(z)]}{z - A(z)}.$$
$$\frac{1-A'(1)}{1+(\rho - A'(1))(1+\rho)} \qquad (16)$$

As before, to get the closed form expression to the w*(s)=V(1-s/λ) *i.e.* LT of the waiting time distribution, it is obvious that the factor z must be replaced by (1-s/λ) on both sides of (16). Thus both the queue length and the waiting time distributions of the M/M+G/2 queues are completely determined even if the joint service time PDF of the two servers is f(t) of (13).

Lemma: If $\lambda \neq \mu \neq \mu_2$ and $\pi_1 < \pi_2$ then A'(1)<1 of (15) implies that $\rho + \rho_2 < \frac{1}{\pi_1}$.

Let w_1 and w_2 be the mean waiting times corresponding to $\mu < \mu_2$ and $\mu > \mu_2$ values respectively. Thus w_1=E(N)/λ when $\mu < \mu_2$ and w_2=E(N)/λ when $\mu > \mu_2$ and

$$E(N)= \begin{cases} \rho + \dfrac{A'(1)^2}{1-A'(1)} + \\ \dfrac{\{\rho - A'(1)\}[\rho + A'(1)(1+\rho)]}{1+\{\rho - A'(1)\}(1+\rho)} \end{cases} \qquad (17)$$

From the design criterion of the above model, it is expected that if the service rate μ of server-1 is larger than that μ_2 of server-2 then $w_2 < w_1$ with increasing values of π_1. For an illustration, w_1 is calculated fixing λ=4.8, μ=4.5 and μ_2=8 (exponential case) and w_2 is computed fixing λ=4.8, μ=8 and μ_2=4.5 while π_1 varies between 0.4 and 1. Fig 1 below shows the graph of these variations on $1000w_1$ and $1000w_2$ which also ensures that $w_2<w_1$ uniformly.

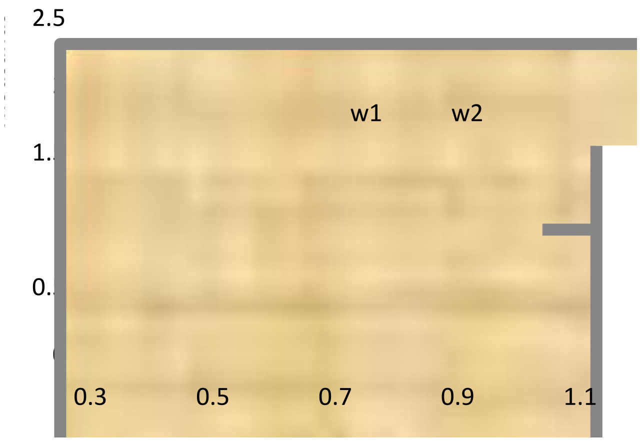

Fig 1: Graph showing π_1 values versus mean waiting times w_1 and w_2.

This is a kind of unequal service rates queueing application where the event of shortest mean delay happens only if the service rate μ of server-1 is larger than that μ_2 of server-2.

III. Conclusion

As the analysis of the above M/M+G/2 queues seems to be simpler and it is an easy procedure to implement if any real life application warrants. We conclude that the proposed method so far discussed is a good alternative for use instead of fitting of M/G/2 queues. Most importantly there are few new results produced from above work that could be applied on the two-serially connected machines in the field of engineering problems rather than applied science areas. To extract more information like arrival epoch probabilities one can employ Markov Renewal theory as in Senthamarikannan and Sivasamy [9] and [10].

.

ACKNOWLEDGMENT

This work has been supported by the authorities of University of Botswana.

REFERENCES

[1] **J. Medhi,** Stochastic Models in Queueing Theory**, 1: 01—457, 2003.**

[2] **O.J Boxma, Q. Deng and A.P Zwart, "**Waiting time asymptotic for the M/G/2 queue with heterogeneous servers', Queueing Systems, 40:5—31, 2002.

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

[4] **D. Efrosinin,** "Controlled Queueing Systems with Heterogeneous Servers: Dynamic Optimization and Monotonicity Properties of Optimal Control Policies in Multiserver Heterogeneous Queues", 2008.

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

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

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

[8] **Krishnamoorthy, B. and Sreenivasan, S.** "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.

[9] **K. Senthamaraikannan, and R. Sivasamy,** "Markov Renewal Queues of a M/M(a,d)/1/(b,N) System-Part I and Part II" Optimization, , Vol. 40, pp 121-145, 1997.

[10] **K. Senthamaraikannan and R. Sivasamy,** "Embedded Processes of a Markov Renewal Bulk Service Queue", Asia Pacific Journal of Operational research, 11(1), 51-65, 1997.

Prof R Sivasamy, holds an M.Sc Degree in 'Statistics' from Annamalai University and a Ph D from the same University. He has had his training in Statistical Methods as Lecturer since 1977. He joined the Statistics department at University of Botswana on the 7th February 2008 as Professor after Three decades of teaching experience at Universities level.
His special interest in the specialty is on the applications of Stochastic Processes and he has been directing research scholars interested in the study of Markov Chain and help them solve problems relating to ensuring product quality and Inventory Management. He is a dedicated teacher and a researcher and as a teacher educator has prepared several research monographs and Books. He has conducted Seminars in thrust areas of Statistical applications and software developments. He is author and co-author of more than 40 publications in international journals. Professionally, he has been invited as an Associate Editor of a few Statistical journals, and has already reviewed several journal article submissions.

Dr P M Kgosi, holds both an M.Sc and a Ph D in Statistics from the University of Keny, UK. His interests are in the areas of Probability, Inference and Stochastic Processes. He authored and co-authored at least 16 refereed journal articles. He is a dedicated teacher and researcher.