AN $M/G/2$ QUEUE WITH A FINITE BUFFER
UNDER DUAL CONTROL:
STATIONARY LOST AND COST ANALYSIS

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

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

[2]Department of Mathematics & Computer Science
Umaru Musa Yar'Adua University
Katsina, NIGERIA

Abstract: This paper considers a Poisson arrival queuing system with an exponential server (server-1) and a general server (server-2) operating under dual control. We extend our analysis of the same queuing system with an infinite buffer to the finite case. The case presented here has many applications in manufacturing, computer centers and telecommunications systems.

1. Introduction

In queuing systems, there are two reasons why analysis is carried out under finite capacity assumption. First, the assumption of infinite capacity does not exist in real-life situations, [1]. Most physical systems with queuing tendencies

────────────────────────────────
[§]Correspondence author

for instance, manufacturing systems, assembly lines, material handling systems and service shops are known to harbor finite capacities. Additionally, queuing spaces in malls, telecommunications and computer business centers especially, centers with control procedures are known to have finite waiting spaces for customers. Similarly, in telecommunications and computer systems, finite controlled telephone lines, computer networks and ATM switches exist, [9] and [8]. Secondly, unlike in infinite capacity analysis, an account of lost customers is taken into consideration in finite capacity analysis as rejection or loss probability. This probability (often called the blocking probability) is essential in designing optimal buffers in areas of application. For instance in communication networks, it is well known that increase utilization of network servers can be achieved through resource sharing of link capacities and buffer spaces among different customer sessions. On the other hand, this benefit potentially has led to increase in congestion and degradation in quality of service (QoS). Consequently, understanding the dynamics between offered traffic load, perceived QoS measures, link capacities and buffer spaces is essential for efficient design and provision of network switching elements.

In the last three decades, a lot of analysis has been carried out on queuing systems with finite buffers. For a survey, [5], [6], [9], [8], [17], [1] and [20]. In particular, [21] proved vital results and performance measures for a queuing model called the M/M,G/2 with an exponential server and a general server under controlled procedures that direct customers to service. Basically as in [7], the purpose of applying control procedures in queuing analysis is to find the optimal operating policy for turning the servers on and off that provides the lowest long-run cost. Several papers have been published in this line and broadly, can be divided into two categories according to whether the system is considered to control the service or the arrival process of customers in the system. For service control policies, [22], [11], [3], [16], [15], [12], [2] and [18] are good examples. On the other hand, controlling the arrival process is also not a new practice. The first investigation was carried out by [10]. Recently, [23] studied the optimal control of a randomized control policy in the M/G/1 queue with second optional service and general start-up times in which the server is typically subject to unpredictable breakdowns. [13] studied the operating characteristics of an M/G/1 queuing system with a randomized control policy and at most J vacations. However to the best of our knowledge, few if any research explore the dual control problem that is; a control procedure for both the arrival and service processes of customers in a queuing system.

The dual control procedures implies that when the number of customers reaches a finite value K, no further arriving customers are allowed into the

system and the rule for service is solely determined by the service provider. As for works dealing with optional service procedures, [14] and [19] are good examples. Extensively, [21] extended [19] model to include a fixed control policy that controls the usage of servers in the system under finite and infinite arrival process. Unfortunately, no suitable linear cost structure was derived. Though in [14] a cost structure was given, the assumption of infinite capacity is supposed. The aforementioned papers and many other ones revealed that no works on cost analysis for a queue with dual control under finite customer flow assumption exist. This is against the fact that in many service stations with finite capacities, both the arrival and the service process are controlled away from the customer. In the presented model, we attempt to fill this gap. We suppose that when the number of customers reaches a threshold value K (the system capacity), no further arriving customers are allowed into the system. Similarly, for service customers, the choice of servers is in accordance with a set of predefined rules that are service provider dependent. In addition, we characterized that the exponential server provides service to the first and second arriving customer category and the general server serves the remaining customer category. Similar to what is found in [17], our analysis is based on the finite capacity approximation of the general service distribution to the exponential distribution so that existence of stationary solution is guaranteed.

The rest of the paper is organized as follows. In Section 2, the model is described along side a controlled queue discipline. Section 3 provides some finite behavior of the model under the stated queue discipline. In Section 4, the mathematical model is formulated. Section 5 derives the lost probability function for a finite K under the dual control procedures employed in this work. In Section 6 (Appendix), we study the cost function for this queuing system under the dual control policy assumption and derive vital management results. Finally, some concluding remarks are given.

2. Model Formulation

We consider an $M/G/2$ queuing system with an exponential server (server-1) and a general server (server-2). Customers arrive according to a Poisson process at a rate λ for service on either of the two servers in the system. The job tendency, the nature of servers and the queue discipline are as in [21] however in this case, we make the additional assumption that the available queuing space is limited to K customers only, (including the service customer) where K is an integer. Thus, a customer who upon arrival meets a system size $j = K$ is

lost. The service time distribution depends on the server that provides service. For customers served by server-1, the service time S_1 is exponential with rate μ_1, i.e. $F_1(t) = P(S_1 < t) = 1 - e^{-\mu_1 t}$, probability density function (PDF) $f_1(t)dt = dF_1(t)$ 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 $B(t) = P[S_2 < t]$ is general with PDF $b(t)$, a mean $\beta = E[S_2]$ and a LST $b^*(s)$ given by $b^*(s) = \int_0^\infty e^{-st} dB(t)$. For reasons to do with stability, we suppose that $\lambda \leq \mu_1 + \frac{1}{\beta}$. Now, if an arriving customer finds:

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.

5. **The system capacity accommodates a customer size $j \leq K \in \mathbf{Z}$.**

2.1. The Stationary Customer Distribution

Denote by $\{X(t), \zeta(t)\}_{t \geq 0}$ the customer process, where $X(t)$ represents 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 that the service time of customers is continuous and that the system is empty at time zero. That means one can apply the supplementary variable technique to analyze the process $\{X(t), \zeta(t)\}_{t \geq 0}$, see [4]. Define for $t \geq 0$:

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

Given that $\lambda < \mu_1 + \frac{1}{\beta}$ 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. More precisely, $\{X(t), \zeta(t)\}_{t \geq 0} \to \{X, \zeta\}$. Suppose that the rate-equality principle holds.

If R_j is the stationary probability that there are j customers in the system, then it is trivial under this assumption that $R_0 = R_{0,0}$, $R_1 = R_{1,0}$, $R_2 = R_{1,1,0}$. And for $3 \leq j \leq K$, $R_j(\eta)$ gives the stationary probability that there are three up to K customers in the system.

By the ergodic theorem 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 differential equations below:

$$\lambda R_{0,0} = \mu_1 R_{1,0}, \ j = 0 < K, \tag{1}$$

$$(\lambda + \mu_1)R_{1,0} = \lambda R_{0,0} + \mu_1 R_{1,1,0}, \ j = 1 \leq K, \tag{2}$$

$$(\lambda + \mu_1)R_{1,1,0} = \lambda R_{1,0} + \left(\mu_1 + \frac{1}{\beta}\right) R_{1,1,1}(\eta); \ j = 2 \leq K. \tag{3}$$

And for $3 \leq j \leq K$, we have

$$\left(\lambda + \mu_1 + \frac{1}{\beta}\right) R_{1,1,1}(\eta) = \lambda R_{1,1,0} + \left(\mu_1 + \frac{1}{\beta}\right) R_4(\eta), \tag{4}$$

$$\left(\lambda + \mu_1 + \frac{1}{\beta}\right) R_j(\eta) = \lambda R_{j-1}(\eta) + \left(\mu_1 + \frac{1}{\beta}\right) R_{j+1}(\eta). \tag{5}$$

From (1) and (2), we have

$$R_1 = R_{1,0} = \left(\frac{\lambda}{\mu_1}\right) R_{0,0} = \left(\frac{\lambda}{\mu_1}\right) R_0, \tag{6}$$

$$R_2 = R_{1,1,0} = \left(\frac{\lambda}{\mu_1}\right)^2 R_0. \tag{7}$$

For R_j; $3 \leq j \leq K$, we take advantage of Lemma 6 in Appendix and solve the accompanying differential equations for a special case of $b(t) = 1$. For a complete discussion on the methodology, see [21]. Thus, for $2 \leq j \leq K$, we obtain that

$$R_j = \left(\frac{\lambda}{\mu_1 + \frac{1}{\beta}}\right)^{(j-2)} \left(\frac{\lambda}{\mu_1}\right)^2 R_0, \tag{8}$$

where R_0 is the idle state probability.

Put $\frac{\lambda}{\mu_1 + \frac{1}{\beta}} = \rho_1$ and $\frac{\lambda}{\mu_1} = \rho$ and apply the normalization condition. Consequently,

$$1 = R_{0,0} + R_{1,0} + R_{1,1,0} + \sum_{j=3}^{K} R_j. \tag{9}$$

Upon further simplification, one obtains that

$$R_0 = R_{0,0} = \frac{(1-\rho_1)}{(1+\rho)(1-\rho_1) + \rho^2(1-\rho_1^K)}. \tag{10}$$

Inserting (10) in (8), R_j for the finite capacity $M/M, G/2$ queue under the control formulation adopted here is obtained. This is given by

$$R_1 = R_{1,0} = \frac{\rho(1-\rho_1)}{(1+\rho)(1-\rho_1) + \rho^2(1-\rho_1^K)}, \quad K = 1, \tag{11}$$

and[1] for $2 \leq j \leq K$ customers, we have

$$R_j = \frac{\rho_1^{(j-2)}\rho^2(1-\rho_1)}{(1+\rho)(1-\rho_1) + \rho^2(1-\rho_1^K)}. \tag{12}$$

Now, if we substitute K for j in (12), then the probability that the system is full to capacity for such $K \in Z$ is obtained. This is equivalent to the probability that an incoming customer is blocked and therefore lost. Thus,

$$R_K = \frac{\rho_1^{(K-2)}\rho^2(1-\rho_1)}{(1+\rho)(1-\rho_1) + \rho^2(1-\rho_1^K)}. \tag{13}$$

3. The Lost Customer Analysis

Denote by R_{K+r} the stationary probability that $r = 0, 1, 2, ...$ customers are lost upon arrival when there are K customers in the system. The Lebesgue count for this lost corresponds to the sequence

$$\{K+r\} = \{K, K+1, K+2, K+3, ...K+n+...\}. \tag{14}$$

---

[1]$K = 0$ is unrealistic.

If Q is such that $Q[N = K + r] = R_{K+r}$, then by conditioning on the threshold K we have

$$Q[K + r \,|\, K, K - 1, K - 2, ..., 0] \tag{15}$$
$$= Q[K + r \,|\, K] = Q[K, r \,|\, K] \tag{16}$$
$$= Q[K]Q[r] = R_K R_r. \tag{17}$$

Here, R_r gives the stationary probability that exactly $r = 0, 1, 2, 3, ...$ customers are lost when there are already K customers in the system. Thus,

$$R_{K+r} = R_K R_0 + R_K R_1 + R_K R_2 + R_K R_3 + ... = R_K \sum_{j=0}^{j=\infty} R_j. \tag{18}$$

The summation component in (18) could be determined by letting $j \to \infty$. Consequently, for a fixed value of K, the probability that r customers are lost by the system is given by

$$R_r = \frac{\rho_1^{(K-2)} \rho^2 (1 - \rho_1)}{(1 + \rho)(1 - \rho_1) + \rho^2 (1 - \rho_1^K)} \frac{\rho_1^{(K+r-2)} \rho^2 (1 - \rho_1)}{(1 + \rho)(1 - \rho_1) + \rho^2}. \tag{19}$$

Theorem 1. *Suppose K is minimal. Then the lost probability in (19) reduces to*

$$R_r = \frac{\rho^{4+r}(1 - \rho)^2}{1 - \rho^4}. \tag{20}$$

Proof. In [21], we have shown that K is minimal if

$$R_K = \frac{\rho^2 (1 - \rho_1)}{[(1 + \rho)(1 - \rho_1) + \rho^2] - (\rho \rho_1)^2}. \tag{21}$$

The minimality condition ensures that $\rho_1 \to \rho$. Consequently, the stationary customer process (X, ζ) is saddled on server-1. Given that $\lambda < \mu_1 + \frac{1}{\beta}$ holds, it is trivial that $\lambda < \mu_1$ for the $M/M/1/K$ 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 $\rho_1 = \frac{\lambda}{\mu_1 + \frac{1}{\beta}}$ where ρ_1 is the occupation rate of the two servers in the $M/M, G/2$ model. Then $\rho_1 = \frac{\lambda}{\mu_1}$. Denote by ρ the occupation rate of the classical $M/M/1/K$ model. By definition $\rho = \frac{\lambda}{\mu_1}$. This means $\rho_1 = \rho$. Finally, the assertion follows if K_{min} is substituted in (19) upon further simplification. $\square$

Corollary 2. *The lost probability in terms of the blocking probability is given by*

$$R_r = \rho^{r+2}(1 - \rho)R_K, \tag{22}$$

where ρ is the occupation rate of a system with a single exponential server.

Proof. This is straightforward, in view of (20). □

Theorem 3. *The expected number of lost customers in the $M/M/1/2$ queue is given by*

$$E[N] = \rho^2 E[N]_{M/M/1} R_K. \tag{23}$$

Proof. Denote by $\omega(s)$ the generating function for customers who arrive the system when there are already K customers. By definition,

$$\omega(s) = \sum_{r=0}^{\infty} R_r s^r = \rho^2(1 - \rho)R_K(1 + \rho s + \rho^2 s^2 + \rho^3 s^3 + ...). \tag{24}$$

The limit of this infinite series attached to (24) exists. Consequently,

$$\omega(s) = \frac{\rho^2(1 - \rho)R_K}{1 - \rho s}. \tag{25}$$

Differentiating $\omega(s)$ and applying some basic properties of generating functions gives

$$\omega'(1) = \frac{\rho^3}{(1 - \rho)} R_K. \tag{26}$$

Finally, the theorem follows upon re arranging (26) in the light of stationary mean number of customers in the M/M/1 queue. □

Corollary 4. *The expected number of lost customers in the $M/M/1/2$ system increases with increase in the blocking probability.*

Proof. This holds in view of the proportionality relationship between the two measures. More precisely, the increase in the size of the lost expectation grows maximally for higher values of ρ for a given blocking probability. □

4. Cost Analysis

In this section, we study the cost of running the M/M,G/2 queuing system under the dual control policy employed in this work. The cost is studied in the light of the general server. We define the bounds under which the cost goes to that of two queuing models namely; the M/M/1/K and the parallel server M/M/2/K with an arrival control.

Proposition 5. *The cost f of running a system of servers is inversely proportional to the service rate. In addition, f is injective.*

Proof. The proposition holds directly since the failure rates (slow server rates) climb as servers aged. Consequently, one may conclude that aging servers with slow service rate have a substantial cost-adding role in the overall cost dynamics of the system. Similarly, the injective part follows since f is a cost function. □

Now, suppose that for some i-servers, the cost f is given by the equation

$$f\left(\sum_{i=1}^{i=M} \mu_i\right) = \sum_{i=1}^{M}[C_i + \mu_i v_i(1 - P_i)] + \lambda E[T]C_0, \qquad (27)$$

where:

C_i = fixed cost per unit time for server i

r_i = cost incurred per unit service in the ith server

P_i = proportion of time server i is idle per unit time

C_0 = penalty cost per unit time wait

λ = arrival rate

μ_i = mean service rate of the ith server

E[T] = mean total time spent in the system.

Assuming that customer renewal is possible so that $P_i \to 1$. Let $C_0 = 1$. Then a simplified form of (27) for a two-heterogeneous server queuing system will take the form

$$f\left(\sum_{i=1}^{i=2} \mu_i\right) = C_1 + C_2 + (\mu_1 + \mu_2)E[T]. \qquad (28)$$

If $E[T]$ is stationary, and C_1 and C_2 are fixed numbers, we have for:

1. $\mu_2 = \mu_1$:
   The cost of running the M/G/2 queue with dual control studied in this paper is equivalent to that of a two-exponential server homogenous queuing system with service rate $2\mu_1$ similar to that of [14].

2. $\mu_2 \to 0$:
   The work load of the system is saddled on server-1 entirely. This phenomenon is equivalent to a long service with an average of $\beta = \infty$ units of service time taking place on server-2. Consequently, the effective work load leading to the cost of the system is determined by that of server-1. The cost function under this condition is equal to that of running a queuing system with a single exponential server.

3. $\mu_2 > \mu_1$:
   Here, the cost is greater than the cost of running a two server homogenous system with exponential servers. The upward increase is about ϵ cost unit. For $\epsilon > 0$, the effective increase is significant.

5. Concluding Remarks

We have presented an analysis of two measures needed for a broader system analysis and operations management for the M/M,G/2 queuing system working under a dual control policy namely; the lost probability and the cost function. It was shown for the trivial buffer $M/M, G/2$ queue that the lost probability increases with increase in the blocking probability. Moreover, this lost is pronounced for high values of the occupation rate. Similarly, for the cost function, we have shown that depending on the size of the service rate of the general server the cost goes to that of the M/M/1, the M/M/2 and the M/M/2+ queuing systems respectively. The cost analysis of these known systems are available in the literature.

One can apply the results proved here for designing suitable buffers for relevant controlled systems for better application and optimal use of investment resources. As a scope for further work, analysis of lost customer distribution under the general K is possible and interesting.

Acknowledgement

6. Appendix

Lemma 6. *Given that the traffic condition $\lambda < \mu_1 + \frac{1}{\beta}$ holds, then in a busy period*

$$Q_j^*(0) = \tilde{Q}_j, \qquad j = 3, 4, ...K. \qquad (29)$$

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 $T_n = t_n - t_{n-1}$, $n = 1, 2, 3, ..., ...$ 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[2] at t is $N(t)$ such that $3 \leq N(0) \leq K$, then $N(t)$ would reach steady state[3] starting at $t = 0$. Consequently, $N(t)$ is a regenerative process over t on state space $\mathbf{S} = 3, 4, ...$ 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 $\lambda < \mu_1 + \frac{1}{\beta}$ holds, then upon service completion on server-2, the state probability $R_j(t)$ can be expressed as

$$R_j(t) = P[N(t) = j, \qquad j = 3, 4, \ldots, K]. \qquad (30)$$

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

$$R_j(t, \eta) = P[N(t) = j \mid t = \eta, \qquad j = 3, 4, \ldots, K]. \qquad (31)$$

Now, let

$$Q_j(t, \eta) = \frac{R_j(t, \eta)}{1 - B(t, \eta)}, \qquad (32)$$

so that

---

[2]$N(t)$ is equivalent to $X(t)$

[3]Precisely at $(0+)$

$$Q_j(t)(1 - B(t)) = R_j(t) = P[N(t) = j \mid t_1 > t], \tag{33}$$

then

$$\sum_{j=3}^{K} Q_j(t) = P[t_1 > t] = 1 - B(t) \tag{34}$$

and

$$Q_j(t) = \int_0^\infty P[N(t) = j, t_1 > t \mid t_1 = \eta] \tag{35}$$

which simplifies to the expression

$$Q_j(t) = \int_\eta^\infty P[N(t) = j \mid t_1 = \eta]dB(\eta) \tag{36}$$

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

$$R_j(t) = \int_0^\infty P[N(t) = j, t_1 = \eta]dB(\eta). \tag{37}$$

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

$$R_j(t) = Q_j(t) + \int_0^t R_j(t - x)dB(x). \tag{38}$$

This renewal equation has a unique solution of the form

$$R_j(t) = Q_j(t) + \int_0^t Q_j(t - x)dM(x), \tag{39}$$

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

$$\lim_{t \to \infty} R_j(t) \to \frac{1}{\beta} \int_0^\infty Q(x)dx. \tag{40}$$

The integral in (40) is the probabilistic version of $\tilde{Q}_j$ when the mean service time on server-2 is β. Thus,

$$\tilde{R}_j\beta = \tilde{Q}_j = Q_j^*(0). \tag{41}$$

Thus, the lemma holds.   □

References

[1] V.N. Angel, Finite capacity queue with multiple poisson arrivals and generally distributed service times, *International Journal of Applied Mathematics,* **26**, No 2 (2013), 233-240.

[2] R. Arumuganathan and S. Jeyakumar, Steady state analysis of a bulk queue with multiple vacations, setup times with N-policy and close down times, *Applied Mathematical Modeling,* **29** (2005), 972-986.

[3] K.R. Balachandran, Control policies for a single server system, *Management Science,* **19** (1973), 1013-1018.

[4] 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.

[5] P.J. Courtois, The M/G/1 finite capacity queue with delays, *IEEE Trans., COM* **28** (1980), 165-172.

[6] A. Frey, Y. Takahashi, A note on an M/G/1/N queue with vacation time and exhaustive service discipline, *Operational Research Letters*, **21** (1997), 95-100.

[7] M.C. Fu and C.K. Jau, Cost Analysis of a two-phase queue system with randomized control policy, *Procedia-Social and Behavioral Sciences*, **25** (2011), 137-146.

[8] E. Gelembe, X. Mang and Y. Feng, Diffusion cell loss estimates for atm with multiclass bursty traffic, *Computer System Science and Engineering*, **11**, No 6 (1996), 325-334.

[9] E. Gelembe, X. Mang and R. Onruval, Diffusion based call admission control in ATM, *Performance Evaluation,* **27**, No 28 (1996), 411-436.

[10] S.M. Gupta, Interrelationship between controlling arrival and service in queuing systems, *Computers and Operations Research,* **22** (1995), 1005-1014.

[11] D.P. Heyman, The T policy for the M/G/1 queue, *Management Science,* **23** (1977), 775-778.

[12] J.C. Ke, The optimal control of an M/G/1 queuing system with server vacations, startup and breakdowns, *Computers and Industrial Engineering,* **44**, No 4 (2003), 567-579.

[13] J.C. Ke, K.B. Huang and W.L. Pearn, Randomized policy of a Poisson input queue with *J* vacations, *Journal of System Science and System Engineering,* **19**, No 1 (2010), 50-71.

[14] B. Krishnamoorthi, On Poisson *queue with two heterogeneous servers*, *Operations Research,* **11**, No 3 (1963), 321-330.

[15] S.S. Lee, H.W. Lee and K.C. Chae, Batch arrival queue with *N* policy and single vacation, *Computers and Operations Research,* **22** (1995), 173-189.

[16] H.W. Lee, S.S. Lee, J.O. Park and K.C. Chae, Analysis of M[x]/G/1 queue with N policy and multiple vacations, *Journal of Applied Probability,* **31** (1994), 467-496.

[17] J.S. MacGregor, M/G/c/K blocking probability models and system performance, *Performance Evaluation,* **52** (2002), 237-267.

[18] P. Moreno, Analysis of a Geo/G/1 queuing system with a generalized *N* policy and setup-close down times, *Quality Technology & Quantitative Management,* **5**, No 2 (2008), 111-128.

[19] 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,* **240** (2015), 140-146.

[20] S.L. Stephen, The steady state queuing time distribution for the M/G/1 finite capacity queue, *Management Sciences,* **21**, No 5 (1975), 501-506.

[21] S. Sulaiman and O.A. Daman, An M/G/2 queue with heterogeneous servers under controlled service schedule: Stationary performance analysis, *IAENG International Journal of Applied Mathematics,* **45** (2015), 31-40.

[22] M. Yadin and P. Naor, Queuing systems with a removable service station, *Operational Research Quarterly,* **14** (1963), 393-405.

[23] D.H. Yang, K.H. Wang, J.C. Ke and W.L. Pearn, Optimal randomized control policy of an unreliable server system with second optional service and startup, *Engineering Computations,* **25**, No 8 (2010), 783-800.