Neural, Parallel, and Scientific Computations 25 (2017) 385-394

FINITE CAPACITY $M/M/r/N$ QUEUEING MODEL WITH ADDITIONAL SERVERS

SUSHIL GHIMIRE[1], GYAN BAHADUR THAPA[2], AND RAM PRASAD GHIMIRE[3]

[1,2]Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal
[3]Department of Mathematical Sciences, School of Science, Kathmandu University, Nepal

ABSTRACT. In this paper, we study multi server finite capacity queueing systems. There are r regular servers available to serve N number of customers. Customers arrive in the system with Poisson process and are served exponentially. In this scenario, one server is permanently available. All r regular servers are not available in the service system at the same time. They come to the service facility according to the number of customers present in the system. The first server serves the customers in number less than N_1, the second servers serves the customers in number between N_1 and N_2 and in general l^{th} server ($l = 1, 2, 3, ..., r$) serves the customers in number between N_{l-1} and N_l ($N_{l-1} < N_l$). When the number of customers exceed N_r, an additional number of m servers will start providing the service which can be expressed in the general way as s^{th} additional server will be used when number of customers remain N_{r+s} to N_{r+s+1} ($s = 0, 1, 2, ..., m-1$) and $N_{r+s} < N_{r+s+1}$. We derive the probability distribution function for i customers in the system, the explicit formula for the expected number of customers in the system, probability that there is k^{th} server in operation in the system and probability that there are r servers in operation in the system. Some numerical results have also been obtained so as to show applicability of the model investigated.

Key Words: Queueing; Server; Customer; Finite Capacity; Steady-State.

AMS (MOS) Subject Classification. 39A10.

1. INTRODUCTION

It can be found in several literatures that many researchers have been studying queueing system and related theory in various perspectives in different interval of time. There are different situations in which a fixed number of customers are served. Ke et al. [1] considered an infinite capacity $M/M/r$ queueing system in which queue length determines the number of working servers. They constructed a cost function deriving the steady-state probability distributions and the expected number of customers in the system using a genetic algorithm technique. Baba [2] studied a batch arrival $M^X/M/1$ queue with multiple working vacation where a server serves customers at a lower rate rather than completely stopping service using a quasi upper triangular transition probability matrix of two-dimensional Markov chain and matrix

Received December 31, 2016

analytic method. Kim et al. [3] proposed a potentially applicable model in slotted digital telecommunication systems and other related areas. They considered a discrete-time multi-server finite-capacity queueing system with correlated batch arrivals and deterministic service times to present the steady-state distributions and the transient distributions of the system length. Brill and Hlynka [4] derived $M/M/1$ queue where arrivals occur single or in pairs dividing into two groups namely Primary and Secondary to obtain the steady-state probability density function of the workload and related quantities. Ghimire et al. [5] developed a mathematical formulae for mean waiting time in the queue, mean time spent in the system, mean number of customers in the queue and in the system with the provision of fixed arrival batch size b. Herbon and Khmelnitsky [6] derived a relation between the means and variances of the measures in transient time using steady state condition for a First Come First Serve queue discipline with exponentially distributed service time. They also proposed a formula similar to Littles law for the means of the queue measures. Abdollahi and Rad [7] studied heterogeneous k-phases single server $M/G/1$ queueing system to obtain the distribution of response time, the means of response time, number of customers in the system and busy period using a steady-state probability generating function technique. Grag [8, 9] explored a single server queueing model with a constant time-dependent arrival rate and service rate to obtain an explicit expression for the state probability distribution using unit step function. Knessl [10] computed asymptotic approximation to the mean response time for a processor sharing queue with finite capacity K. Gong and Batta [11] developed a two dimensional generating funcrtion in single server two-priority, pre-emptive queueing model to obtain the average number of customers for each priority class. Ghimire, Basnet [12] and Ghimire, Ghimire [13] dealt with heterogeneous arrival and departure $M/M/1$ queue for finite and infinte capacity with vacation and service breakdown respectively to calculate the various performance measures by using a probability generating function method. Tonui et al. [14] studied the stability of the single server Markovian queuing system by the means of sensitivity analysis. Baumann and Sandmann [15] derived formulae for loss blocking probabilities, expectations and higher moments of numbers of customers in the queues and in the whole system of phase-type queueing system. Ghimire et al. [16] studied the transient multi-server queuing system subject to breakdowns without queue of the waiting customers. Ke and Wang [17] investigated finite capacity $G/M/1$ queueing model with removable server under N policy by using supplementary variable technique. Kerbache and Smith [18] modelled manufacturing facilities as an open finite queueing networks to examine the optimal routing in layout and location problems from a network optimization perspective. Kim [19] dealt with a single server inventory control problem to model a queueing system with finite waiting room and non-instantaneous replenishment process. Kim

et al. [20] analysed steady-state distribution of the system states of a tandem queueing system with infinite and finite intermediate buffers, heterogeneous customers and generalized phase-type service time distribution.

In this paper, we develop the mathematical model of a queueing system where N number of customers are serverd by r number of servers. Arrival of the customers follow the Poisson stream whereas the service time is distributed exponentially. The number of servers will increase gradually depending upon the number of customers present in the system. One server is permanently available and the second server will start to serve only after the number of customers exceed N_1 in the queue. If the queue length becomes less than N_1 then the second server will stop serving. Similarly, after the number of customers exceed N_2, the third server will start serving. For example, second servers serves the customers in between N_1 and N_2 and in general l^{th} server $(l = 1, 2, 3, ..., r)$ serves the customers in number between N_{l-1} and N_l $(N_{l-1} < N_l)$. When the number of customers exceed N_r, an additional number of m servers will start providing the service which can be expressed in the general way as s^{th} additional server will be used when number of customers remain N_{r+s} to N_{r+s+1} $(s = 0, 1, 2, ..., m-1)$ and $N_{r+s} < N_{r+s+1}$. These additional servers will be in use only in some special cases when all the regular r servers are not able to serve N number of customers. We derive the explicit formula for the expected number of customers in the system, the probability distribution function for i customers in the system, probability that there is k^{th} server in operation in the system and probability that there are r servers in operation in the system. Some numerical results have also been obtained so as to show applicability of the model. The model under study may have many ubiquitous applications in circuit designing, internet service, assembly line in manufacturing system where the finite capacity has great importance. The model we have developed can rarely be found in the existing field of queueing system. Jain [21] developed finite capacity $M/M/r$ queueing system with the provision of removable servers but she didn't take m additional servers in the model. In some special situations arrival rate becomes so high that operation of all the r servers can not handle the arriving customers. We are addressing this type of conditions by adding another m number of servers. As an example, we have experienced some extra mobile towers in some occasitions in some of the particualar areas. Provision of additional servers makes the system more efficient, sustainable and economic which we have taken into account in our model. If $m = 0$, our model is identical to the model proposed by Jain [21]. So it is meaningful to say that our model is more general.

The remainder of the paper is organized as follows: Section 2 describes the mathematical model with all the notations used in the paper along with the probability distributions for different conditions. Section 3 explains all of the numerical results obtained by MATLAB simulations, and Section 4 is the conclusion of this paper.

2. MATHEMATICAL MODEL

In this section, we derive some of the mathematical formulas for the proposed queueing model. We have used some important assumptions and the notations, which are as follows:

- **•:** Arrival follows in Poisson process.
- **•:** Service times are exponentially distributed.
- **•:** There are r number of total servers.
- **•:** There are m number of additional servers.
- **•:** The first server is permanently available.
- **•:** The second server starts serving after exceeding N_1 customers in the system.

Under these assumptions, following notations are used to describe the mathematical model.

λ = Arrival rate

μ_k= Service rate for different state, $1 \leq k \leq N$

ν_t=Service rate for the additional server, $1 \leq t \leq m$

r =Number of servers

$\phi_j = \sum_{k=1}^{j} \mu_k$

$\psi_m = \sum_{t=1}^{m} \nu_t$

$\rho_{j,m} = \frac{\lambda}{\phi_j + \psi_m}$, $\psi_m = 0$ for m=0

Using the mathematical notations above, the following balanced equations can be established:

$$\lambda P_0 = \mu_1 \ P_1 \tag{2.1}$$

$$(\lambda + \mu_1)P_i = \lambda P_{i-1} + \mu_1 P_{i+1} \tag{2.2}$$

$$(\lambda + \phi_j)P_{N_j-1} = \lambda P_{N_j-2} + \phi_{j+1} P_{N_j}, \qquad 1 \leq j \leq r-1 \tag{2.3}$$

$$(\lambda + \phi_j)P_i = \lambda P_{i-1} + \phi_j P_{i+1}, \qquad 2 \leq j \leq r-1, \ \ N_{j-1} \leq i \leq N_j - 2 \tag{2.4}$$

$$(\lambda+\phi_{r-1}+\psi_m)P_i = \lambda P_{i-1}+(\phi_{r-1}+\psi_m)P_{i+1}, \qquad N_{r-2} \leq i < N, \ \ 1 \leq m \leq r-1 \tag{2.5}$$

$$(\lambda + \phi_r + \psi_r)P_{N-1} = \lambda P_{N-2} + (\phi_r + \psi_r)P_N, \qquad i = N \tag{2.6}$$

where, $\phi_j = \sum_{k=1}^{j} \mu_k$ and $\psi_m = \sum_{t=1}^{m} \nu_t$

Using the equations above, we can write P_i in terms of P_0 for different conditions as follows:

$$P_i = \rho_{1,0}^i P_0 \qquad where \;\; \rho_{1,0}^i = \left(\frac{\lambda}{\mu_1}\right)^i, \qquad 1 \le i \le N_1 - 1 \tag{2.7}$$

$$P_i = \left[\prod_{j=1}^{k-1} \rho_{j,0}^{N_j - N_{j-1}}\right] \rho_{k,0}^{i-N_{k-1}+1} P_0, \qquad k = 2,3,...,r-1 \;\; N_{k-1} \le i \le N_k - 1 \tag{2.8}$$

$$P_i = \left[\prod_{j=1}^{r-1} \rho_{j,t}^{N_j - Nj-1}\right] \rho_{r,t}^{i-N_{r-1}+1} P_0, \qquad N_{r-1} \le i \le N-1, \;\; 1 \le t \le m \tag{2.9}$$

and

$$P_N = \left[\prod_{j=1}^{r-1} \rho_{j,0}^{N_j - Nj-1}\right] \left(\frac{\lambda}{\mu_1 + \mu_2 + \mu_3 + ... + \mu_r}\right)^{N-N_{r-1}+1} P_0, \qquad i = N \tag{2.10}$$

(2.11)

$$P_i = \begin{cases} \rho_{1,0}^i P_0, & 1 \le i \le N_1 - 1 \\ \left[\prod_{j=1}^{k-1} \rho_{j,0}^{N_j - N_{j-1}}\right] \rho_{k,0}^{i-N_{k-1}+1} P_0, & k = 2,3,...r-1 \;\; N_{k-1} \le i \le N_k - 1 \\ \left[\prod_{j=1}^{r-1} \rho_{j,t}^{N_j - Nj-1}\right] \rho_{r,t}^{i-N_{r-1}+1} P_0, & N_{r-1} \le i \le N-1, \;\; 1 \le t \le m \\ \left[\prod_{j=1}^{r-1} \rho_{j,0}^{N_j - Nj-1}\right] \left(\frac{\lambda}{\mu_1+\mu_2+\mu_3+...+\mu_r}\right)^{N-N_{r-1}+1} P_0, & i = N \end{cases}$$

The probability normalizing condition is

$$\sum_{i=0}^{N_1-1} P_i + \sum_{k=2}^{r-1} \left[\sum_{i=N_{k-1}}^{N_k-1} P_i\right] + \sum_{t=1}^{m} \left[\sum_{i=N_{r-1}}^{N-1} P_i\right] + P_N = 1$$

$$\begin{aligned} P_0 = \Bigg[& \sum_{i=0}^{N_1-1} \rho_{1,0}^i + \sum_{k=2}^{r-1} \left[\sum_{i=N_{k-1}}^{N_k-1} \left[\prod_{j=1}^{k-1} \rho_{j,0}^{N_j - N_{j-1}}\right] \rho_{k,0}^{i-N_{k-1}+1}\right] \\ & + \sum_{t=1}^{m} \left[\sum_{i=N_{r-1}}^{N-1} \left[\prod_{j=1}^{r-1} \rho_{j,t}^{N_j - N_{j-1}}\right] \rho_{r,t}^{i-N_{r-1}+1}\right] \\ & + \left[\prod_{j=1}^{r-1} \rho_{j,0}^{N_j - N_{j-1}}\right] \left(\frac{\lambda}{\mu_1 + \mu_2 + ... + \mu_r}\right)^{N-N_{r-1}+1} \Bigg]^{-1} \end{aligned} \tag{2.12}$$

Hence, we can write

$$P_0 = \frac{1}{\alpha + \beta + \gamma + \sigma}$$

where

$$\alpha = \sum_{i=0}^{N_1-1} \rho_{1,0}^{i} = \begin{cases} N_1 & \rho_{1,0} = 1 \\ \frac{1-\rho_{1,0}^{N_1}}{1-\rho_{1,0}} & \rho_{1,0} \neq 1 \end{cases}$$

$$\beta = \sum_{k=2}^{r-1} \left[\sum_{i=N_{k-1}}^{N_k-1} \left[\prod_{j=1}^{k-1} \rho_{j,0}^{N_j-N_{j-1}} \right] \rho_{k,0}^{i-N_{k-1}+1} \right] = \sum_{k=2}^{r-1} \left[\prod_{j=1}^{k-1} \rho_{j,t}^{N_j-N_{j-1}} \right] \omega_k$$

where

$$\omega_k = \begin{cases} N_k - N_{k-1} & \rho_{k,0} = 1 \\ \frac{\rho_{k,0}\left(1-\rho_{k,0}^{N_k-N_{k-1}}\right)}{1-\rho_{k,0}} & \rho_{k,0} \neq 1 \end{cases}$$

$$\gamma = \sum_{t=1}^{m} \left[\sum_{i=N_{r-1}}^{N-1} \left[\prod_{j=1}^{r-1} \rho_{j,t}^{N_j-N_{j-1}} \right] \rho_{r,t}^{i-N_{r-1}+1} \right] = \sum_{t=1}^{m} \left[\prod_{j=1}^{r-1} \rho_{j,t}^{N_j-N_{j-1}} \right] \omega_r$$

where

$$\omega_k = \begin{cases} N_r - N_{r-1} & \rho_{r,t} = 1 \\ \frac{\rho_{r,t}\left(1-\rho_{r,t}^{N_r-N_{r-1}}\right)}{1-\rho_{r,t}} & \rho_{r,t} \neq 1 \end{cases}$$

and

$$\sigma = \left[\prod_{j=1}^{r-1} \rho_{j,0}^{N_j-N_{j-1}} \right] \left(\frac{\lambda}{\mu_1 + \mu_2 + ... + \mu_r} \right)^{N-N_{r-1}+1}$$

(i) The average number of customers in the system is

$$L_s = \sum_{i=0}^{N} i.P_i = \sum_{i=0}^{N_1-1} i.P_i + \sum_{k=2}^{r-1} \left[\sum_{i=N_{k-1}}^{N_k-1} i.P_i \right] + \sum_{t=1}^{m} \left[\sum_{i=N_{r-1}}^{N-1} i.P_i \right] + N.P_N$$

(ii) Probability that the first server is in operation in the system

$$P(1) = Pr(i \geq 1) = \sum_{i=1}^{N} P_i$$

(iii) Probability that k^{th} server is in operation in the system

$$P(k) = Pr(i \geq N_{k-1}) = \left(\sum_{j=k}^{r-1} \omega_j + \gamma \right) P_0$$

(iv) Probability that there are r servers in operation in the system

$$P(r) = Pr(i \geq N_{r-1}) = (\gamma + \sigma) P_0$$

3. NUMERICAL RESULTS

All the results we obtained here are verified by means of computer simulations using MATLAB programming. Figure 1 shows the relations between probability and different arrival rates. Whenever the arrival rate increases, the probability of getting service decreases. At the same time, when service rate increases probability of getting served increases which is realistic in nature.

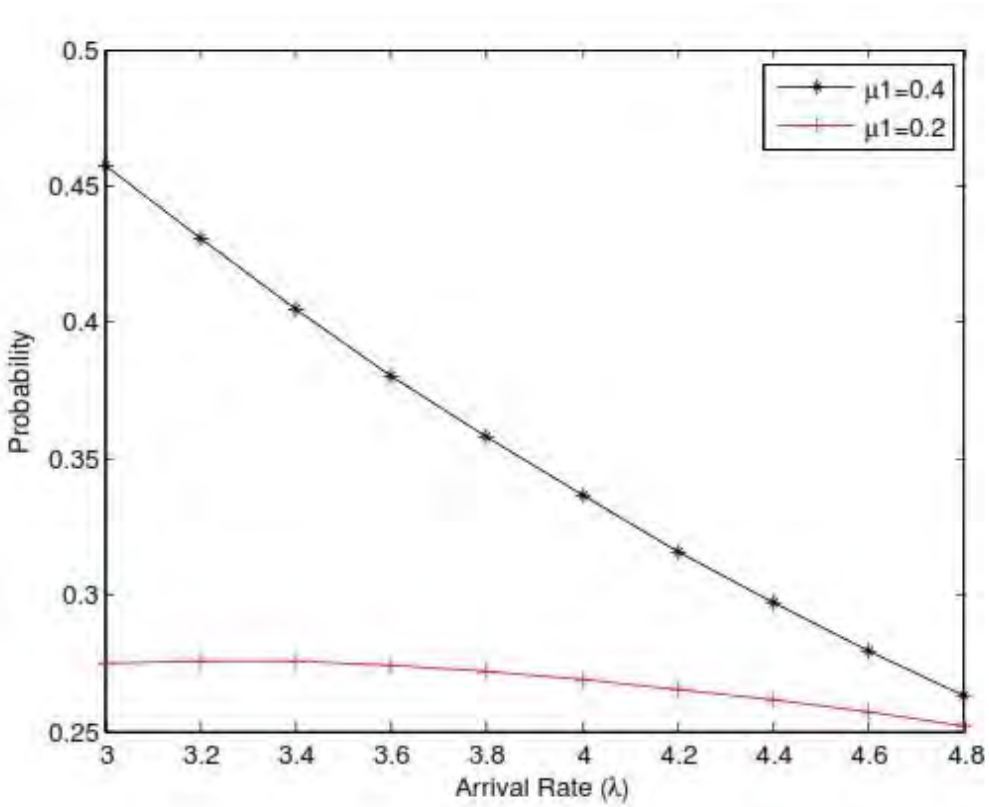

FIGURE 1. Probability vs Arrival Rate

Similarly, Figure 2 represents the relationship between mean queue length and arrival rate. As the arrival rate increases, the queue length also increases. It is quite obvious in nature that when the arrival rate becomes faster, the queue length increases and people have to wait for a longer time to get served. In real life, this type of situation arises whenever some of the necessary and daily consumed items become shortage in the market. People have the tendency of reserving those items for the future without caring how long they have to wait in a queue.

Finally, Table 1 shows the different arrival rates and their corresponding probabilities. Arrival has been varied from 3 to 4.8 at an interval of 0.2. and P_i have been observed for different conditions. P_0 combines all the probabilities for all possible P_i and hence it decreasing as the values of arrival rate is increasing. On the other

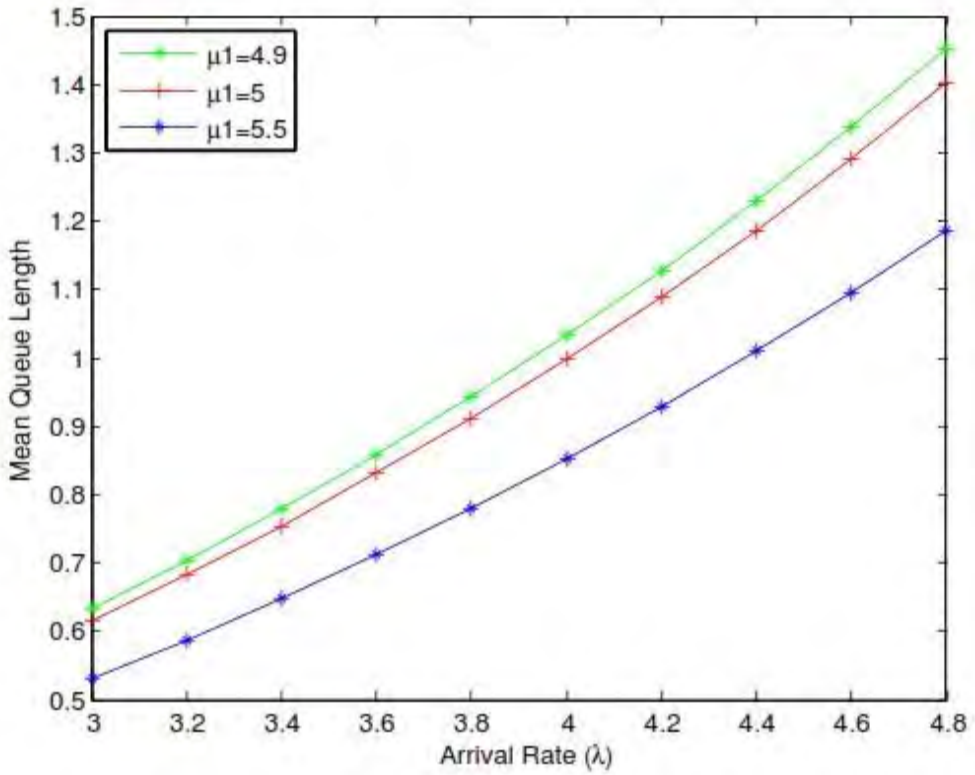

FIGURE 2. Mean Queue Length vs Arrival Rate

λ	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
P0	0.4576	0.4305	0.4048	0.3805	0.3577	0.3362	0.316	0.297	0.2793	0.2627
P1	0.2746	0.2755	0.2753	0.274	0.2718	0.2689	0.2654	0.2614	0.2569	0.2522
P2	0.1647	0.1763	0.1872	0.1973	0.2006	0.2151	0.223	0.23	0.2364	0.2421
P3	0.0011	0.0015	0.0019	0.0023	0.0029	0.0035	0.0042	0.005	0.0058	0.0068
P4	0.00018	0.00024	0.00033	0.00044	0.00057	0.00073	0.00092	0.0012	0.0014	0.0017
P5	0.0011	0.0015	0.0019	0.0023	0.0029	0.0035	0.0042	0.005	0.0058	0.0068
P6	0.00018	0.00024	0.00033	0.00044	0.00057	0.00073	0.00092	0.0012	0.0014	0.0017
P7	0.000028	0.000041	0.000059	0.000083	0.00012	0.00015	0.0002	0.00027	0.00034	0.00043
P8	0.000005	0.000009	0.000013	0.000019	0.000028	0.00004	0.000056	0.000077	0.0001	0.00013

TABLE 1. Probability distribution for different arrival rates

hand, P_i are increasing for the increasing arrival rate which is because of the inverse relations with P_0.

4. CONCLUSIONS

In many of the realistic situations, it is not possible to provide service for all of the customers. Keeping these conditions in mind, there are a number of areas where service providers put restrictions and decide to provide the service only for limited quota. If these quotas are fulfilled, no additional incoming customers can get service. But in this paper, we have studied multi-server queueing system where second server starts serving whenever the queue length becomes longer than N_1. Here, arrival is heterogeneous which is realistic in nature. If the queue length becomes longer than expected without exceeding the capacity, some additional servers are subjected to provide service. In some conditions, the queue length becomes so long that all r servers can not handle the customers. We have added another m number of servers to manage these kinds of difficulties. This kind of model is applicable in circuit designing, internet service and also in assembly line in manufacturing system. If transient condition is added in the model, the scenario becomes more challenging and interesting too.

ACKNOWLEDGMENTS

The first author is thankful to Erasmus Mundus Smart Link project for financial support as a PhD exchange student to carry out the work at Burgas Free University, Bulgaria from Sep 2016 to Sep 2017.

The second author is thankful to the Erasmus Mundus LEADERS Project for funding him as a Post Doc Research Fellow to carry out the work in Department of Mathematics, University of Evora, Portugal from Nov 2016 to Aug 2017.

REFERENCES

[1] J. B. Ke, J. C. Ke and C. H. Lin, Cost Optimization of an $M/M/r$ Queueing System with Queue-Dependent Servers: Genetic Algorithm, *QTNA*: 82–86, 2010.

[2] Y. Baba, The $M^X/M/1$ Queue with Multiple Working Vacation, *American Journal of Operations Research*, Vol. 2: 217–224, 2012.

[3] N. M. Kim, L. M. Chaudhary, B. K. Yoon and K. Kim, A Complete and Simple Solution to a Discrete-Time Finite-Capacity $BMAP/D/c$ Queue, *Applied Mathematics*, Vol. 3: 2169–2173, 2012.

[4] H. B. Percy and H. Myron, Server Workload in an $M/M/1$ Queue with Bulk Arrivals and Special Delays, *Applied Mathematics*, Vol. 3: 2174–2177, 2012.

[5] S. Ghimire, R. P. Ghimire and G. B. Thapa, Mathematical Models of $M^b/M/1$ Bulk Arrival Queueing System, *Journal of the Institute of Engineering*, Vol. 10, No. 1: 184–191, 2014.

[6] A. Herbon and E. Khmelnitsky, Transient Littles Law for the First and Second Moments of $G/M/1/N$ Queue Measures, *Journal of Service Science and Management*, Vol. 3: 512–519, 2010.

[7] S. Abdollahi and M. R. S. Rad, On an $M/G/1$ Queueing Model with k-Phase Services and Bernoulli Feedback, *Journal of Service Science and Management*, Vol. 5: 280–288, 2012.

[8] D. Garg, Approximate Analysis of an $M/M/1$ Markovian Queue Using Unit Step Function, *Open Access Library Journal*, Vol. 1: 1–5, 2014.

[9] D. Garg, Transient Solution of $M/M/2/N$ System Subjected to Catastrophe cum Restoration, *Open Access Library Journal*, Vol. 2: 1–8, 2015.

[10] C. Knessl, On Finite Capacity Processor-Shared Qusues, *Society for Industrial and Applied Mathematics*, Vol. 50, No. 1: 264–287, 1990.

[11] Q. Gong and R. Batta, A Queus-Length Cutoff Model for a Preemptive Two-Priority $M/M/1$ System, *Society for Industrial and Applied Mathematics*, Vol. 67, No. 1: 99–115, 2006.

[12] R. P. Ghimire and R. Basnet, Finite Capacity Queueing System with Vacations and Servers Breakdown, *International Journal of Engineering*, Vol. 24, No. 4: 387–394, 2011.

[13] R. P. Ghimire and S. Ghimire, Heterogenemous Arrival and Departure $M/M/1$ Queue with Vacation and Service Breakdown, *Management Science and Engineering*, Vol. 5, No. 3: 61–67, 2011.

[14] B. C. Touni, C. R. Langat and M. J. Gichengo, On Markovian Queuing Models, *International Journal of Science and Research*, Vol. 3, No. 11: 93–96, 2014.

[15] H. Baumann and W. Sandmann, Multi-Server Tandem Queue with Markovian Arrival Process, Phase-Type Service Times, and Finite Buffers, *European Journal of Operational Research*, Vol. 256: 187–195, 2017.

[16] S. Ghimire, R. P. Ghimire and G. B. Thapa, Performance Evaluation of Unreliable $M(t)/M(t)/n/n$ Queueing System, *British Journal of Applied Science and Technology*, Vol. 7, No. 4: 412–422, 2015.

[17] J. C. ke and K. H. Wang, A Recursive Method for the N Policy $G/M/1$ Queueing System with Finite Capacity, *European Journal of Operational Research*, Vol. 142: 577–592, 2002

[18] L. Kerbache and J. M. Smith, Multi-Objective Routing within Large Scale Facilities Using Open Finite Queueing Networks, *European Journal of Operational Research*, Vol. 121: 105–123, 2000.

[19] E. kim, Optimal Inventory Replenishment Policy for a Queueing System with Finite Waiting Room Capacity, *European Journal of Operational Research*, Vol. 161: 256–274, 2005.

[20] C. Kim, A. Dudin, O. Dudina and S. Dudin, Tandem Queueing System with Infinite and Finite Intermediate Buffers and Generalized Phase-Type Service Time Distribution, *European Journal of Operational Research*, Vol. 235: 170–179, 2014.

[21] M. Jain, Finite Capacity $M/M/r$ Queueing System with Queue-Dependent Servers, *Computers and Mathematics with Applications*, Vol. 50: 187–199, 2005.