Neural, Parallel, and Scientific Computations 25 (2017) 375-383

TRANSIENT ANALYSIS OF MARKOVIAN QUEUE WITH FLEXIBLE SERVERS AND BALKING

R.P.GHIMIRE[1], SAMIR SHRESTHA[1], OLIVER TSE[2], AND SHYAM SUNDAR SAH[1]

[1]Department of Mathematical Sciences, Kathemandu University, Nepal

[2]Department of Mathematics, Technical University of Kaiserslautern, Germany

ABSTRACT. This paper deals with the study of time dependent queueing model with balking. In our study, the system has three servers to server the customers and each server has its limited capacity to serve the customers. Initially, only one server joins the system to serve a limited number of customers and once the limit is reached, the second server joins the system. The third server enters the system when second one has reached its maximum limitation. Customers arrive to the system in Poisson fashion and are served exponentially. We compute the numerical results for time-dependent state transition probabilities, the mean number of customers in the system and in the queue at time t. Moreover, the mean time that a customer spent in the system, the mean time that a customer has to wait in the queue and the probability that there are greater than or equal to N customers in the system are also obtained.

Keywords: Queue, Balking, Flexible servers, Markovian

AMS (MOS) Subject Classification. 60k25.

1. Introduction

Tremendous works on queues have been done in steady state. Also, from application view points there are many classes of queueing systems in which a transient analysis is required such as flexible manufacturing system, machining system, assembly lines. These are the practical queueing systems operated over a finite-time horizon whose performance depends upon the system time and initial conditions. So it is relevant to mention some of the works done on the transient queueing systems. Of many results that are obtained in the study of queues, the most significant one is the time-dependent probability that a queue starting with a given number of customers attains a prescribed length at a later time Parthasarathy and Sharafali (1989). Ashour and Jha (1973) did the pioneering works on implementation of non-homogeneous Markov process in the study of transient queueing model in the sense that service parameters taken as the function of time. Narahari and Viswanadham(1994) analyzed $M/M/1$ queueing model with time system performance and showed the applicability of model

Received December 31, 2016

in the performance evaluation of automated manufacturing system with deadlocks and failures. Kumar and Arivudainambi (2000) considered the $M/M/1$ queueing model under catastrophic assumptions and obtained the various measures of performance. Al-Seedy (2004) presented the two-channel $M/M/2$ queue with balking and heterogeneity service rates by choosing the free servers with different probabilities. Al-Seedy et al. (2009) explored the technique to compute the transient probabilities for the $M/M/C$ queue with the involvement of the Bessel functions under the provision of balking and reneging behaviors of the customers. Czachorski et al.(2009) proposed a method of diffusion approximation to solve the problem of tele-traffic networks where the load in Internet protocol (IP) routers is dynamically changing and the traffic is entirely different from Poisson streams. Yang and Liu (2010) studied the transient behavior of general queueing model by means of statistical methodology and they claimed that their model could be widely used in the production planning manufacturing in which more difficulties are encountered due to uncertainty such as demand forecast mismatches, unexpected interruptions in productions and natural disasters. Kaczynski et al. (2012) derived exact distribution of the n^{th} customers sojourn time in an $M/M/s$ queue with k customers initially presented. They also provided algorithms for computing the covariance between sojourn times for an $M/M/1$ queue with the help of Maple code. Chan et al. (2012) proposed boundary value methods to find transient solutions of $M/M/2$ queueing systems with two heterogeneous servers of which numerical approximation was obtained by using Crank-Nicolson to accelerate the convergence. Ayyappan and Shyamala (2013) considered the queueing model wherein the customers are arriving as batches in compound Poisson process with the provision of server vacation. Rusek et al. (2014) obtained new analytical results for the queueing characteristics of the packet buffer limited to a fixed number fed by Markovian arrival process (MAP) with the provision of semi-Markov (SM) service-time process and they claimed that their proposed model is much closer to the results of $MAP/G/1/b$ model. Ammar (2014) derived new explicit formulas for probability distribution function for a two heterogeneous servers subject to balking and reneging for the $M/M/2$ queue. Since the earlier works on the transient behavior of queues in literature were published in the late 1950 and in early 1960, in many cases the system being modeled never reaches steady-state and hence do not accurately portray the system behavior which can be experienced in military air traffic control when the computerized information center suddenly becomes inoperative. To handle such a problem Vinodhini and Vidhya (2014) suggested a solution for transient multi-server queueing model with impatience and disaster for the system prone to down. Singla and Garg (2014) made comprehensive study of feedback queueing model with correlated departure and transition time mark in which they obtained the transient state queue length probabilities and its Laplace transform. Zeng et al.(2014) dealt with a optimization problem of queueing model of railway containers that have to run through internal and externals terminals. Under the investigation they obtained queue length and average waiting time of the railway container terminal gate system and determined the optimal number of service channels within the reasonable computation time. Balking is the behavior of the customer that when the customer entering the system finds joining the queue may take longer time in waiting and he/she decides not to join the queue and leaves the system. This is very common phenomenon in real life and provision of balking in the study of queueing system makes the model realistic due to optimal performance measures. Gong and Li (2014) made the queueing analysis of service call centers with the imposition of balking and reneging and they developed the maximum system utility optimization model with the consideration of customers' psychology behavior. Recently Vijaya et al.(2015) proved that the expected total cost is less for single queue multi-server model as compared with multi-queue multi-server model by using the mathematical equations for both types of models. Zhu and Qi (2016) modeled queueing system that can increase the service efficiency and improved patient satisfaction that pass through the system in different service routes. Kalyanaraman and Nagarajan (2016) derived a steady-state solution of bulk arrival queue with fixed batch service by using probability generating function.

In this paper, we have made the provision of flexible servers in the sense that according to the work load in the system the servers join the service facility to serve the customers and all the servers that work in the system have heterogeneous service rates. In one hand, our investigation differs from the work done on the field by several authors is that we have time varying heterogeneous arrival and service rates, on the other hand, time-dependent analysis of finite capacity queueing model has been made with the help of numerical approximation technique.

2. **Mathematical model and analysis**

For our model, we have the following assumptions:

(i) First server gives the service to the customers with service rate $\alpha(t)$ if there are $J-1$ customers in system. If there are greater than J and less than $K-1$ customers in the system then two servers are employed for the service having service rate $\beta(t)$ of the second server. If there are K to L costumers in the system then three servers are provisioned wherein third server has $\gamma(t)$ service rate.

(ii) When there are K to L customers in the system the customers start balking with the balking rate $(1-b_n)$, where the rate of joining the system is b_n.

(iii) $\alpha(t)=t\mu_1$, $\beta(t)=t\mu_2$, $\gamma(t)=t\mu_3$, $\alpha'(t)=\alpha(t)+\beta(t)$, $\beta'(t)=\alpha'(t)+\gamma(t)$.

(iv) The customers arrive in the system arrival rate $\lambda(t)$ in the Poisson fashion and are served exponentially.

Under these assumptions the probability transition system of equations is given by following system of linear differential difference equations:

$$
\begin{aligned}
\frac{d}{dt}P_0(t) &= -\lambda(t)b_0P_0(t)+\alpha(t)P_1(t)\\
\frac{d}{dt}P_n(t) &= \lambda(t)b_nP_{n-1}(t)-(\lambda(t)b_n+\alpha(t))P_n(t)+\alpha(t)P_{n+1}(t),\\
&\qquad (1\le n\le J-2)\\
\frac{d}{dt}P_{J-1}(t) &= \alpha(t)b_{j-1}P_{J-2}(t)-(\lambda(t)b_{J-1}+\\
&\quad \alpha(t))P_{J-1}+\alpha(t)P_{J(1)}+\beta(t)P_{J(2)}(t)\\
\frac{d}{dt}P_{J(1)}(t) &= \lambda(t)b_{J(1)}P_{J-1}(t)-(\lambda(t)b_{J(1)+\alpha(t)})P_{J(1)}(t)+\beta(t)P_{J+1}(t)\\
\frac{d}{dt}P_{J(2)}(t) &= -(\lambda(t)b_{J(2)}+\beta(t))P_{J(2)}(t)+\alpha(t)P_{J+1}(t)\\
\frac{d}{dt}P_n(t) &= \lambda(t)b_nP_{n-1}(t)-(\lambda(t)b_n+\alpha'(t))P_n(t)+\alpha'(t)P_{n+1}(t),\\
&\qquad (J+1\le n\le K-2)\\
\frac{d}{dt}P_{K-1}(t) &= \lambda(t)b_{K-1}P_{K-2}(t)-(\lambda(t)b_{K-1}+\alpha'(t))P_{K-1}(t)+\\
&\quad \alpha'(t)P_{K(1)}(t)+\gamma(t)P_{K(2)}(t)\\
\frac{d}{dt}P_{K(1)}(t) &= \lambda(t)b_{K(1)}P_{K-1}(t)-(\lambda(t)b_{K(1)}+\alpha'(t))P_{K(1)}+\gamma(t)P_{K+1}(t)\\
\frac{d}{dt}P_{K(2)}(t) &= -\big[\lambda(t)(1-b_{K(2)})+\gamma(t)\big]P_{K(2)}(t)+\alpha'(t)P_{K+1}(t)\\
\frac{d}{dt}P_n(t) &= \lambda(t)(1-b_n)P_{n-1}(t)-\big[\lambda(t)(1-b_n)+\beta'(t)\big]P_n(t)+\beta'(t)P_{n+1}(t)\\
&\qquad (K+1\le n\le L-1)\\
\frac{d}{dt}P_L(t) &= \lambda(t)(1-b_L)P_{L-1}(t)-\beta'(t)P_L(t)
\end{aligned}
\tag{2.1}
$$

The above system can be represented by the matrix equation of the following form:

$$
\frac{d}{dt}\mathbf{P}(\mathbf{t}) = Q(t)\mathbf{P}(t) \tag{2.2}
$$

where $\mathbf{P}(t)=[P_0(t),P_1(t),\cdots,P_{J-1},P_{J(1)},P_{J(2)},\cdots,P_{K-1},P_{K(1)},P_{K(2)},\cdots,P_L]^T$ is the probability vector of size $(L+3)\times 1$ and $Q(t)$ is coefficient matrix of the system size $(L+3)\times(L+3)$.

A few of the transient queueing literatures have explored the exact solutions by using Laplace transform and Modified Bessel function of first kind, Caley Hamilton theorem, continued fraction expansion methods for the solution of homogeneous Markov process. Meadows (1962) obtained the analytical solution for system of n^{th} order linear periodic differential equations with the considerations of periodic coefficients and he used the approximation technique for the numerical computations but our parameters are not the periodic and have to deal with non-homogeneous Markov process (because arrival and service rate parameters vary with time). This paper explores the transient solutions to the non-homogeneous Markov by applying the implicit Euler's method to solve the system (2.2) numerically. The argument for seeking a numerical method for solving a wide class of transient queueing problems has a strong practical appeal.

3. **Performance indices**

Following are the major performance measures of our model under study:

1. The expected number of customers in the system is

$$L_s(t) = \sum_{n=1}^{L} nP_n(t) \tag{3.1}$$

2. The expected number of customers in the queue is

$$L_q(t) = \sum_{n=1}^{J-1}(n-1)P_n(t) + \sum_{n=J}^{K-1}(n-2)P_n(t) + \sum_{n=K}^{L}(n-3)P_n(t) \tag{3.2}$$

3. The expected time spent in the system is

$$W_s(t) = \frac{L_s(t)}{\lambda(t)} \tag{3.3}$$

4. The expected time spent in the queue is

$$W_q(t) = \frac{L_q(t)}{\lambda(t)} \tag{3.4}$$

5. Probability that there are greater than or equal to N number of customers in the system is

$$Prob(\text{Queue size} \geq N) = \sum_{n=N}^{L} P_n(t) \tag{3.5}$$

4. **Numerical results**

Because of the time dependent behavior of both arrival and service rates, queueing model under study yields state probabilities of transient nature. The probability of n units being in the system is a function of time t. The state probabilities as well as rest of the performance indices are best evaluated through the use of numerical methods of the solution. We have used the implicit Euler scheme for solving the system of ordinary differential equations (2.1) numerically by using MATLAB 7.10 (R2010a). For the numerical computations, we have taken $J = 10$; $K = 20$; $L = 30$; $b_n = 0.95$;

$\lambda(t) = \lambda_0 t$; $N = 20$; $\mu_1 = 0.41$; $\mu_2 = 0.42$; $\mu_3 = 0.43$; $t = 80$. At arrival rate $\lambda_0 = 0.7$, figure 1 displays that initially at time $t = 0$, the probability that the system being empty is 1 so at the same time the probabilities that the system has the number of customers greater than 1 are zeros but as the time passes on the probabilities that there are $1, 2, \cdots, 7$ customers have some values other than zero. For various increasing values of the arrival rates, we have computed the probability distributions and has been presented in the figure 1 to figure 4. Figures 2(a,b) demonstrate that higher the arrival rate larger the number of customers in the system and in waiting in the queue and sooner the server being idle, this is due to our provision that higher the accumulation of customers sooner the joining of new server in system to serve the customers which shows that our model has much more practicability in the real-world. Figures 3(a,b) exhibit that increasing arrival rates make the longer waiting time in the queue that inherently yields longer system time. Figure 4 predicts that probabilities of having greater than 20 customers in the system are zeros for a while but as the arrival rates are set to increase the probabilities increase.

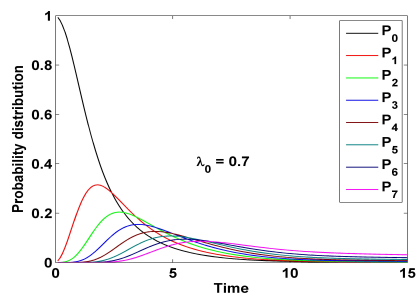

FIGURE 1. Probability distribution.

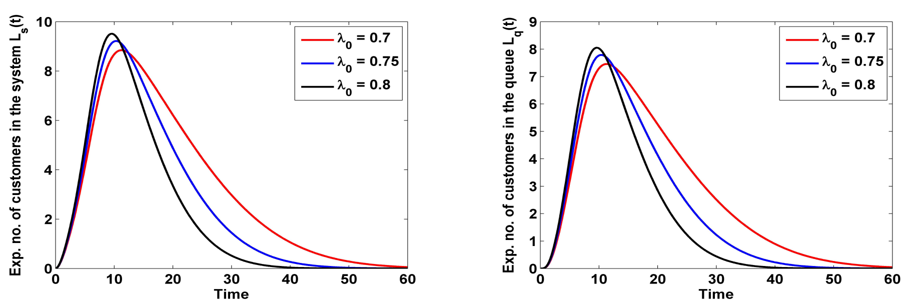

FIGURE 2. Expected number of customers: (a) in the system, and (b) in the queue.

5. **Conclusion and outlook**

Because of the time dependent behavior of both arrival and service rates, queuing model under study yields state probabilities of transient nature. The probability of n units being in the system is a function of time t. The state probabilities as well as rest

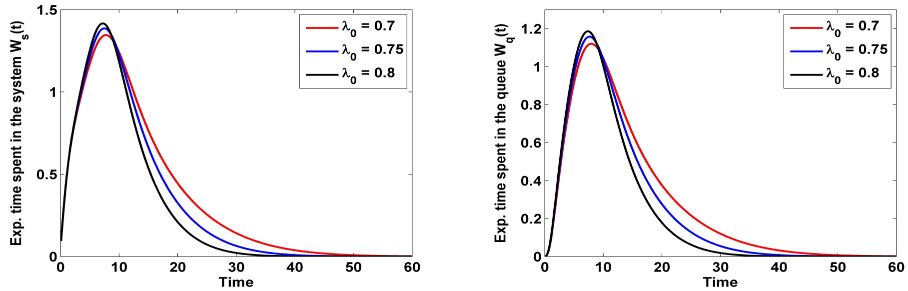

FIGURE 3. (a) Expected time spent in system, and (b) Expected waiting time in queue.

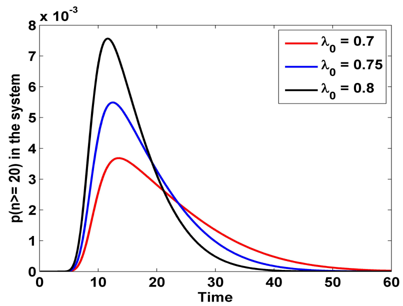

FIGURE 4. Probability that number of customers in the system greater than $N = 20$.

of the performance indices are best evaluated through the use of numerical methods. Intensive search of literatures reveals that explicit transient solution (modified Bessel function or Laplace transform) methods to the most of the queuing models are much more complex to tackle, if not impossible to obtain. The model under study may have applications in designing and operating transportation systems such as airports, freeways, ports, and subways. Applicability of model is also in the performance evaluation of automated manufacturing system with deadlocks and failures. Our model can be extended to more general that makes the provision of n heterogeneous time dependent arrival and service rates for the finite capacity model with balking and reneging discipline that may be more realistic to tackle modern technological complex problems.

ACKNOWLEDGMENTS

Authors are very thankful to unknown reviewers for valuable comments and suggestions due to which this paper could be presented in this form. Also, first author is thankful to the Department of Mathematics, University of Kaiserslautern, Germany for providing the German Academic Exchange Service (DAAD) supported research visit to produce this paper.

REFERENCES

[1] Al-Seedy, 2004: A transient solution of the non-truncated queue $M/M/2$ with balking and an additional server for longer queues (Krishnamoorthi discipline). *Applied Mathematics and Computation*, **156**, 763-769.

[2] Al-Seedy R.O., A.A. El-Sherbiny, S. A .El-Shehawy and S. I. Ammar, 2009: Transient solution of the $M/M/C$ queue with balking and reneging. *Computers and Mathematics with Applications*, **57**, 1280-1285.

[3] Ammar S.I., 2014: Transient analysis of a two heterogeneous servers queue with impatient behavior. *Journal of Egyptian Mathematical Society*, 22, 90 -95.

[4] Ashour S. and R.D. Jha, 1973: Numerical transient-state solutions of queueing systems. *Society for Modelling and Simulation International*, **21**, 117-122.

[5] Ayyappan G. and S. Shyamala, 2013: Time dependent solution of $M^{[X]}/G/1$ queueing model with Bernoulli vacation and balking. *International Journal of Computer Applications*, 61(21), 19-24.

[6] Chan R. -H., S. -T. Lee and H. -W. Sun, 2012: Boundary value methods for transient solutions of queueing networks with variant vacation policy. *Journal of Computational and Applied Mathematics*, **236**, 3948-3955.

[7] Czachorski T., T.N.P. Slaska and F.P. Lipn, 2009: Transient state analysis diffusion approximation as an alternative to Markov models. Fluid-flow approximation and simulation, 2009,13-18.

[8] Gong J. and M. Li, 2014: Queueing time decision model with the consideration on call center customers abandonment behavior. *Journal of Networks*, 9(**9**), 2441-2447.

[9] Kaczynski W.H., L.M. Leemis, and L.H. -Drew, 2012: Transient queueing analysis. *INFORMS Journal on Computing*, 24(**1**), 10-28.

[10] Kumar B.K. and D. Arivudainambi, 2000: Transient solution of an $M/M/1$ queue with catastrophes. *An International Journal of Computers and Mathematics with Applications*, **40** , 1233-1240.

[11] Meadows, H.E., 1962: Solution of system of linear ordinary differential equations with periodic coefficients. *The Bell System Technical Journal*, 1275-1294.

[12] N. Singla and P.C. Garg, 2014: Transient and numerical solution of a feed back queueing system with correlated departures, *American Journal of Numerical Analysis*, 2(**1**), 20-28.

[13] Narahari Y. and N. Viswanadham, 1994: Transient analysis of manufacturing system performance. *IEEE Transactions on Robotics and Automation*, 10(**2**), 230-244.

[14] Parthasarathy P.R. and M. Sharafali, 1989: Transient solution to the many-server Poisson queue: A simple approach. *J. Appl. Prob.* **26**, 584-594.

[15] Rusek K., L. Janowski, and Z. Papir 2014: Transient and stationary characteristics of a packet buffer modeled as an MAP/SM/1Bsystem. *Int. J. Appl. Math. Comput. Sci.*, 24(**2**), 429-442.

[16] Vinodhini, G.A.F. and V. Vidhya, 2014: Transient solution of a multi-server queue with catastrophes and impatient customers when system is down. *Applied Mathematical Sciences*, 8(**92**), 4585-4592.

[17] Yang F. and J. Liu: Transient analysis of general queue systems via simulation based transferred function modeling . *Proceedings of the 2010 Winter Simulation Conference*, USA. 1110-1122.

[18] Zeng M., W. Cheng, and P. Guo, 2014: A transient model for analyzing and optimizing gate congestion of railway container terminals. *Mathematical Problems in Engineering*, 1-14.

[19] Vijaya Prasad S.,V. H. Badshah, and A. K. Tariq,2015: Mathematical analysis of single queue multi-server and multi-queue multi-server queueing models. *Global Journal of Mathematical Analysis*, 3(**3**), http://www.sciencepubco.com/index.php/GJMA/article/view/4689/1837.

[20] Zhu, M. and E. Qi ,2016 : A multi type queueing system network analysis method for controlling server number in the outpatient. *The open Automation and Control Systems Journal*,**8** ,21-33.

[21] Kalyanaraman, R and P. Nagarajan,2016: Batch arrival,fixed batch service queue with unreliable server and with cupulsory vacation. *Indian Journal of Science and Technology* ,9(**38**).1-8.