Analysis of Discrete Time Queueing System with Late Arrival using Load Balancing Technique

Fozia Panhwer[†], Abdul Fattah Chandio[†], M. Mujtaba Shaikh[†], Zeeshan Ali[††], Khalil M. Zuhaib[†††]

[†]Department of Electronic Engineering, QUEST, Nawabshah, Sindh, Pakistan
[††]National Research Council of Italy (IREA- CNR), Italy
[†††]Department of Electronic Engineering, QUCEST, Larkana, Sindh, Pakistan

Summary
Almost in every field of communication, networks are now being digitized. Thus, digital techniques are essentially required to solve and to obtain the performance evaluation of those network problems. Slotted time queueing is one of those latest technique to treat with the digitized networks. In any digital network, the working and transfer of data is based on time slotting as well as entering and leaving of the data packets which are controlled and synchronized by the clock pulses. Network station buffer is only part where each time before transmission to destination point on the network, the data packets are being really received and processed. In this paper, network buffer is focused and examine it with the help of load balancing technique by using late entering system modeling approach in slotted time domain. By focusing on network buffer, this work is enhanced for slotted time modeling approach in order to achieve minimum overloading with limited resources by analyzing the performance measure of mean which consequently improves throughput and maximizes the resource utilization.

Key words:
Slotted time queueing, network Buffer, Load balancing, Late arrival.

1. Introduction

For limited resources queuing networks, with their limited capacity especially in the telecommunication and the data networks, data sharing is known to be effective technique with incoming traffic [1] [2]. As focus is on digital communication network which is done with time slotting, balancing the load is an important technique for these type of networks [3] [4] [5] [6]. In these digital systems which use time slotting technique, every movement of packet i.e., transmitting, processing and reception of the data packets are supervised and synchronized with the clock pulse. With time slotting, clock pulse distributes the whole time into equal number of time intervals and every task is done within that time interval. It is assumed in this paper that the data packets only arrive and get their service at end of the slots only. Moreover, the arrival may get their service in minimum one or multiple number of slots to leave the system before entering to another system until they reach their destination. As compared with continuous time modeling, discrete time modeling can be used more effectively to examine the behavior of network buffer because time slotting is done in the digital systems [7] [8]. Most of the literature work is about the continuous time queueing systems. In [6], the authors have analyzed the load balancing techniques for continuous time but not for discrete time queueing system. Similarly, in [9], load balancing is treated for continuous time and single server. The authors in [10] and [11] had considered the network of finite buffers under discrete time queueing with early arrival approach but did not consider the late arrival approach. Authors in [12] derived the mathematical expressions of orbit and system size distributions for Geo/G/1 discrete time queueing system for preferred and impatient customers. In [13], authors analyzed the Geo/Geo/1 discrete time queueing system for preferred customers with partial buffer sharing. The authors in [12] and [13] also not considered late arrival system.

In this paper, load balancing technique on the limited capacity slotted time queuing system using late arrival modeling approach with two finite buffers is applied which is different from above mentioned approaches in order to examine the behavior of the network/system. For the analysis purpose, by varying the probabilities of entering the data units into system, queue capacity and the threshold, performance figure of an average number of the data packets in the system (mean) is achieved which consequently improves the other performance metrics.

Remaining paper is organized as follows: In section-2, slotted time queueing modeling approaches are discussed. Late entering modeling approach is described in section-3. Proposed load balancing system model with its organization is described in section-4. In section-5, simulation results for the performance metric of mean are presented and discussed. At the last, paper is concluded in section-6.

Manuscript received July 5, 2018
Manuscript revised July 20, 2018

2. Slotted Time Queueing Modeling Approaches

In slotted time queueing system, it is assumed that an equivalent number of interims or intervals known as slots are division of time scale. It is an efficient way to analyze the performance and behavior of the digital system. In discrete time systems, both arrival and/or departure may take place in the same time slots. It is assumed that all the activities (sending, processing and receiving) are synchronized with the clock pulse and may occur only on the slot edge or boundaries [10]. Any slotted system can be observed at the slot boundary. On the basis of the occurrence of the entrance of the data packets at particular check point on the time scale that is mostly the slot boundary, behavior of the system can be effectively analyzed with early entering system (EES) and late entering system (LES). In Fig. 1, 'T' represents the period of time slot and T_n is the time slot at its n^{th} position where as T_{n+1} and T_{n-1} represents the one slot ahead and on slot before its n^{th} position consecutively.

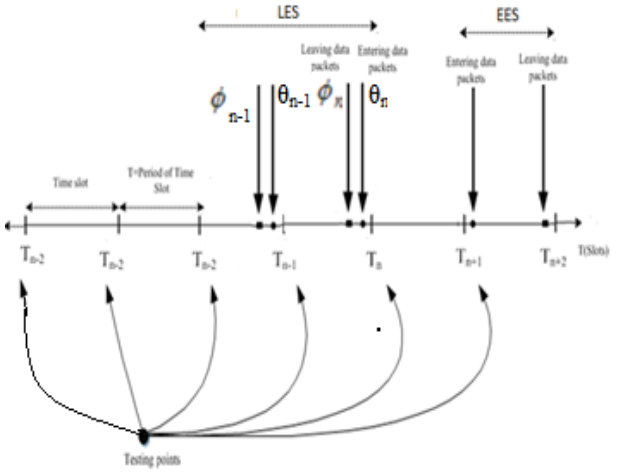

Fig. 1 Time Scale of Early and Late Entrance Slotted Time Queueing System.

Mostly, the system can be checked at the slot edges in order to obtain the exact behavior. In Fig. 1, starting and ending of the slots is called entering and leaving instances at which any data may enter and leave from the system. It is assumed that the data packets will always leave from the system at the end of the slot, therefore, slotted time system can be modeled on the basis of the entrance of the data packets in the queueing system weather they enter at the starting or ending edge of the slot. Based on the entrance of the data packets, slotted time queueing system can be modeled by two approaches, called as early entrance modeling approach and late entrance modeling approach.

As in the slotted time queueing system, more than one event may occur at the same time instance so in the early entrance system, data packets will enter into the system just after the slot instance where they may find the service in the same time slot. On the other hand, in the late entrance system, the data packets will enter into the system just before the end of the slot instance where they will not find the service in that slot in which they enter but they will always find the service in the next consecutive slots. Queue length of the late entrance system is one slot more than that of the early entrance system [10].

3. Late Entrance Modeling Approach

In late entrance system, it is assumed that the data packets enter into the system just before the slot instance or prior to leaving event as shown in Fig. 1 where θ_n and θ_{n-1} show the entering instances and ϕ_n and ϕ_{n-1} show the leaving instances at the slot check points.

In LES, data packets enter into the system and will not find the service in the same time slot in which they enter, however, they will find the service in the next consecutive slots, so the size of the queue will get enlarged except an empty or idle state of the system. If the data packets enter and the system is occupied and the server is also busy then the data packets are prevented to enter into the system. If no any data packets enter into the system but the service completion takes place in particular time instance then the size of the queue or buffer will get decreased by one state for all consecutive states. Empty state of the system is known as the idle state of the system and system will continue to exist in this state in the case only when no any data packet enters into the system at specific time instance otherwise at every entrance of the data packet queue length will get enlarged. Another state of the system is fully occupied state of the system and when it will be reached, the system will continue to exist in this state with the probabilities of either data packets enter and leave from the system or there is no any data packet entering into the system in a given time instance [10]. For the system states from state one to state L-1 (except idle and full state), system will continue to exist in the same state either data packets enter or leave or no any data packet enters into the system in a given time instance [10].

4. Proposed Load Balancing System Model

A model with shared data packets among two finite queues is developed, in which one of the queue can shift its data packets to another queue if it is found having data packets below than the chosen threshold value in order to overcome the network overloading or any congestion later

on. This all is done in slotted time domain with late entering system modeling approach. Both queues having different thresholds to set the condition to allow or not allow the data packets from one queue to another queue to obtain the performance figure using all system state class switching probabilities at all states one by one until fullness of the system is achieved. It is assumed throughout our model that there are two types of the data packets entering the system one by one i.e., θ_1 and θ_2 where they enter according to Bernoulli process with geometrically distributed inter-arrival and service time. It is also assumed that probabilities that neither data packets of type-1 nor data packets of type-2 or none of them enter into the system can be represented by $\overline{\theta_1}$, $\overline{\theta_2}$, and $\overline{\theta}$ respectively. There are two separate servers to provide the service for entering data packets in their queues. The data packets which enter into the first queue get service from the server-1 and the data packets which enter into the second queue get service from the server-2 with the probabilities of ϕ_1 and ϕ_2 respectively but if both servers are busy and cannot provide the service to the incoming data traffic in their respective queues then those probabilities can be given as $\overline{\phi_1}$ and $\overline{\phi_2}$ respectively.

Proposed model consists of two queues C_1 and C_2 respectively having maximum space of storing or holding data packets is 'Th_1 = 6' for queue-1 and 'Th_2 = 5'for queue-2. As the system under discussion is of limited capacity so the data packets may not enter into the system in the case when both queues are entirely occupied including one in server but one condition is set here that the queue-2, 'C_2' can also allow incoming data packets 'θ_1' routed from queue-1. Based on the threshold value 'Th_3' and 'Th_4' settled on 'C_1' and 'C_2' respectively i.e., when $Th_3 \geq 3$ is reached then all incoming data packets are diverted from C_1 to C_2 and when $Th_4 \geq 4$ is reached then diverted data packets from C_1 to C_2 are blocked.

Fig. 2 shows a proposed model in which load balancing technique is applied in a slotted time queuing system using late entering methodology in order to control the overloading and congestion. Two different kinds of data packets are originated by the network station or buffer θ_1 and θ_2 respectively and are served by the two separate servers with the service rate ϕ_1 and ϕ_2 respectively.

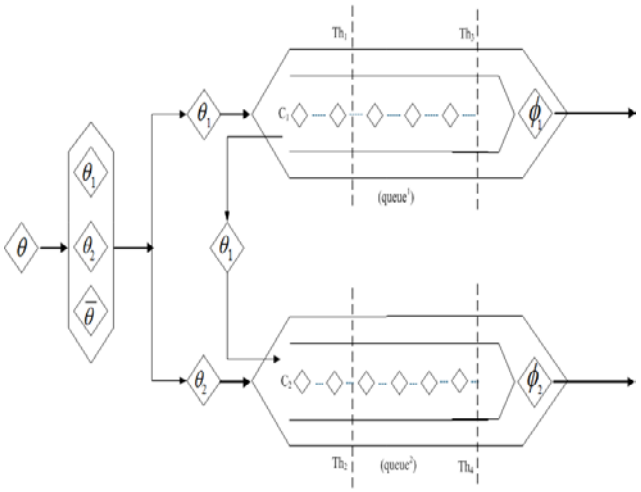

Fig. 2 Load Balancing Model of Late Entering System

Server-2 provides the faster service to the incoming data packets because service rate ϕ_2 is greater than ϕ_1. During one time instant either one or none of the data packet can enter into the system i.e. θ_1, θ_2 or $\overline{\theta}$ respectively. It is assumed that all incoming data packets of type-2 'θ_2' can only enter into the queue-2 regardless of any threshold limit until the system is not fully occupied whereas incoming data packets of type-1 ' θ_1' can join both queues based on the threshold value set on C_1 and C_2 with the condition that if the threshold value of queue-1 is less then Th3 then incoming data packets of type-1 'θ_1' can enter in its own queue-1 but when the threshold value Th_3 is achieved, then further incoming data packets of type-1 'θ_1' in the queue-1 are routed towards queue-2 where they find quick or rapid service from the server. Until the threshold value of Th4 is not achieved, this routing of data packets of type-1 'θ_1' is continued. Once the threshold value of Th4 is achieved in queue-2 then it stops incoming data packets of type-1 'θ_1' routed from queue-1. Till the instance when the threshold value of queue-2 is reached to Th4 or threshold value of queue-1 is less then Th_3, server in both queues will provide the service to the incoming data packets which are generated for them only until they are not fully occupied including one in server. At the instant when both queues are fully occupied including one in server, then the data packets are not allowed to enter into the system (lost) but before the new arrival of the data packets if one service completion takes place then they are allowed to enter into the system.

5. Simulation Results and Discussion

In this work, limited capacity slotted time queuing system is considered using late entrance system modeling approach in order to examine the behavior of the system under discussion. Different results of average number of the data packets in the system (mean) are achieved by varying the probabilities of entering the data units into system, by varying the threshold, and by varying the queue capacity. Mathematically, average number of the data packets in queue-1 is given by,

$$Q_{[1]} = \sum_{p_2=0}^{Th_2} \sum_{p_1=0}^{Th_1+1} P_r(p_1, p_2) \qquad (1),$$

where p_1 and p_2 are the thresholds of queue-1 and queue-2 of network buffers respectively and $P_r(p_1, p_2)$ is the joint probability of both queues.

Load balancing technique is used when a system has limited capacity to serve the entering data packets in order to minimize the congestion and overloading of system which consequently maximizes throughput. The results are shown in Fig. 3 and Fig. 4 respectively. Fig. 3 shows the results of mean vs data packets entering probability of queue-1, denoted by θ1 for different thresholds (shown in Fig. 3) with fixed capacity of 40, 30 for queue-1 and queue-2 respectively. The results show that if the threshold value is increased in order to accept the incoming data packets by queue-2 diverting from the queue-1 then the queue contents in first system will get decreased as indicated by decrease in mean value of queue-1 from blue to yellow in Fig. 3.

Fig. 3 and Fig. 4 show two separate results having fixed size of buffer. They describe that with the variation of the threshold value of both queues or with the variation of the total capacity of the both queue buffers, average number of data packets in the systems are affected.

It can be clearly observed from Fig. 3 that by varying the threshold values of both queues, average number of the data packets (mean) in queue-1 will get minimized by analogously extending the threshed value of queue-2.

Fig.4 describes that if we consider systems or buffers or buffer queues with load balancing technique having constant threshold values for both queues and with different values of the probabilities of entrance of the data.

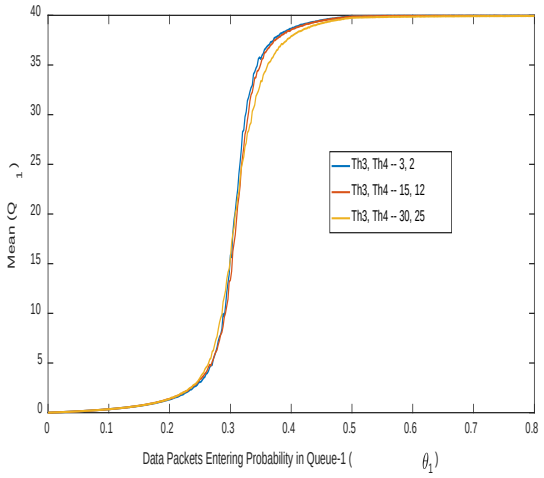

Fig. 3 Mean Vs Data Packets Entering Probability of Queue-1 (θ_1) for different Thresholds with Fixed Capacity of Queue-1 (Th_1 = 40) & Queue-2 (Th_2 = 30).

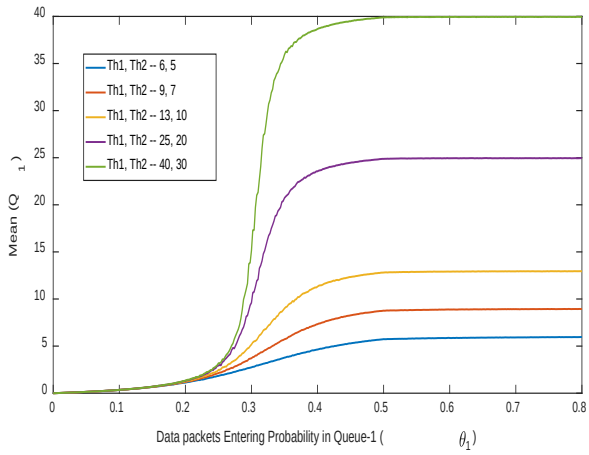

Fig. 4 Mean Vs Data Packets Entering Probability of Queue-1 (θ_1) for different Capacity Values with Fixed Thresholds of Queue-1 (Th_3 = 3) & Queue-2 (Th_4 = 2).

packets into the system and different values of the maximum capacity of both systems, average number of data packets in queue-1 will get increased i.e., with increase in capacity of both queues proportionally, mean (average number of data packets) increase for queue-1. Similarly, for the same condition, if we increase the probability of service that the data packets will exit from the system then the average number of data packets in queue-1 will increase less rapidly as compared with small value of probability of service to the incoming data packets.

6. Conclusion

In this research, a network buffer model was examined in slotted time to reliably investigate a discrete time queuing system having limited capacity with load sharing applied to a late entering data packets. The proposed model was developed with shared data packets among two finite queues where one of the queue could shift its data packets to another queue if it was found having data packets below than the chosen threshold value in order to overcome the network overloading or any congestion. Both queues having different threshold values to set the condition to allow or not allow the data packets from one queue to another queue to obtain the performance measure of mean. This work was done in slotted time domain with late entering system modeling approach using load balancing technique. The results showed that the load is balanced in queue-1 with diverting of data packets to queue-2 by increasing thresholds and mean improves in queue-1 with increase in capacity of both queues which consequently improves throughput and resource utilization by avoiding congestion in the system.

Acknowledgments

We are so much grateful to Quaid-e-Awam University of Engineering, Science, & Technology (QUEST), Nawabshah, Sindh, Pakistan for conducting and finally finishing this research work.

References

[1] A. Allen, "Probability, Statistics and Queueing Theory with Computer Science applications", 1990, Second Edition, Academic Press, New York.

[2] G. Bolch, S. Greiner, H. de Meer, K.S. Trivedi, "Queueing Networks and Markov Chains, Modeling and Performance Evaluation with Computer Science Applications", 2006, Second Edition, John Wiley Sons, Inc., Hoboken, New Jersey.

[3] H. Bruneel, B.G. Kim, "Discrete-Time Models for Communication Systems Including ATM", 1993, Kluwer Academic Publications, Boston.

[4] H. Bruneel, "Performance of Discrete-Time Queueing Systems", Computers and Operations Research, 1993, Vol. 20(3), pp. 303-320, Elsevier Science Ltd. Oxford, UK.

[5] R.B. Cooper, "Queueing Theory. Stochastic Models", Handbook of Operations Research and Management Science, 1990, Vol (2), Chapter 10, pp. 469518, North Holland, Amsterdam.

[6] Y.T. Wang, R.J.T. Morris, "Load Sharing in Distributed Systems", IEEE Transaction on Computers, 1985, Vol.34, pp. 204-217.

[7] U.C. Gupta, S.K. Samanta, R.K. Sharma, "Computing Queueing Length and Waiting Time Distributions Infinite-Buffer Discrete-Time Multi-Server Queues with Late and Early Arrivals", Computers and Mathematics with Applications, 2004, Vol. 48. pp. 1557-1573.

[8] T. Miesling, "Discrete-Time Queuing Theory", Operations Research, 1958, Vol. 6, pp. 96-105.

[9] M.E. Lewis, "Average Optimal Policies in a Controlled Queueing System with Dual Admission Control", Journal of Applied Probability, 2001, Vol. 38, pp. 369-385.

[10] S.A.A. Shah, "Performance Modeling and Congestion Control Through Discrete-Time Queueing", Ph.D. Thesis, 2010, Faculty of Electrical Engineering & Information Technology, Vienna University of Technology, Wien, Austria.

[11] S.A.A. Shah, S. Wajiha, A.S. Larik, "Modeling and Performance Evaluation of Early Arrival Discrete Time Queueing System with Load Balancing Using Geometrical Distribution", 2011, Mehran University Research Journal of Engineering & Technology, Vol. 30 (4), pp. 699-706.

[12] J. B. Wu, J. X. Wang, and Z. M. Liu, "A discrete-time Geo/G/1 retrial queue with preferred and impatient customers," Applied Mathematical Modelling, 2013, vol. 37, no. 4, pp. 2552–2561.

[13] S. Zhou, L. Liu, J. Li, "A Discrete-Time Queue with Preferred Customers and Partial Buffer Sharing," Mathematical problems in Engineering, 2015, Vol. 2015, Article ID 173938, 12 pages.