# Modeling and Performance Evaluation of Early Arrival Discrete Time Queueing System with Load Balancing Using Geometrical Distribution

SYED ASIF ALI SHAH*, WAJIHA SHAH**, AND ABDUL SATTAR LARIK***

RECEIVED ON 10.09.2011 ACCEPTED ON 01.10.2011

## ABSTRACT

**Load balancing is an efficient technique used to maximize throughput, optimal resource utilization, minimized response time and avoiding congestion. This can be achieved by distributing the workload evenly across two or more network stations, nodes or buffers, links, central processing units, hard drives, or other resources. In this paper, we have modeled and developed a load balancing approach in a discrete-time domain to analyze and evaluate the system of finite network buffers using an early arrival system. Our approach of modeling such a system consists of two steps. The first step is the determination of all system-state stages and their corresponding transition probabilities. Next, we compute various performance measures by utilizing the system state transition probabilities for its steady-state behavior.**

**Key Words:** **Load Balancing, Discrete-Time Queueing System, Early Arrival System.**

## 1. INTRODUCTION

The load balancing system is vastly considered and evaluated through the continuous time queueing theory [1-2]. The evaluation of load balancing system in discrete time domain using discrete time queueing system become efficient for the system which are digital in nature [3-5]. In comparison with the continuous time queueing system, a digital system is needed to be treated with some specialized techniques in order to understand and obtain its exact behavior [4]. The operation of digital systems is based on time slotting, thus discrete time queueing system has a potential applications and it provides flexibility to model a fixed length packet system [7-8].

The analysis of discrete time queueing system depends on the slots of time. In the discrete time system, the time scale is divided into equal number of slots in which the customer arrives in and departs from the system. According to the arrival and departure of the customer in the slot, the discrete time queueing system is categorized into two systems. These systems are called early arrival and late arrival systems and performance evaluation of each system depends on the arrival epochs.

In [9] various load balancing algorithms results are discussed. These algorithms relates to the area of server and routing. The many authors studied to find the balance

* Assistant Professor, Department of Electrical Engineering, Mehran University of Engineering & Technology, Jamshoro.
** Assistant Professor, Department of Electronic Engineering, Mehran University of Engineering & Technology, Jamshoro.
** Associate Professor, Department of Electrical Engineering, Mehran University of Engineering & Technology, Jamshoro.

processor load for optimal routing policies in continuous time domain. [10] Discusses the continuous time load balancing model with single server by implementing the two admission control decisions. A two server system with scheduling policies is analyzed in [11] to balance the load in which the inter-arrival and service time is exponentially distributed.

Contrary to above approaches in this paper, we have modeled and developed a load balancing approach in a discrete-time domain to analyze and evaluate the system of finite network buffers using an early arrival system modeling approach for discrete-time queueing systems [12].

The remaining of the paper is organized as follows: Discrete time queueing, early arrival system and system model are briefly discussed in Sections 2. In Section 3, modeling discrete time queueing system with load balancing and classification and formulation of system state stages are discussed. The Sections 4 presents performance measure and results respectively. Finally the conclusions are presented in Section 5.

## 2. DEFINITIONS AND SYSTEM MODEL

In this section we first describe the considered discrete time queueing model. Next, we describe the early arrival system. Later, we discuss the considered system model.

### 2.1 Discrete Time Queueing Model

The discrete-time queueing is based on the assumption that the time is divided into fixed length of intervals (equidistant) called as slots (Fig. 1). In these systems, both arriving customers and departing customers are geometrically distributed.

The discrete-time system allows multiple events during one time unit called a slot. The common causes of the change that takes place in any queueing system are the events when a customer arrive and/or departs from the system during a given time unit. An exact system can only be modeled according to the occurrence of an arriving customer in a given slot at some observing point. Mainly, the choice of observation point from where a system can be observed to analyze is a slot edge or boundary. In these systems, starting edges of slots are called arrival epochs, where any arriving data begin to enter in the system, whereas, the data can only departs from the system at the ending edges of slots, called departure epochs.

Hence, discrete time queueing systems can be modeled in two ways based on the nature of arriving customer whether it come at the start (beginning) of a slot or at the end of a slot, called as an Early Arrival System or Late Arrival System, respectively. The Bernoulli process with binomial and geometric distributions is used to model the arrival and service process of these systems. The probability generating function based on the z-transform can be used to determine different performance measures.

### 2.2 Early Arrival Model

In an early arrival system, we assume that an arrival occurs soon after the slot edge, that is, at the beginning of a slot as shown in Fig. 2.

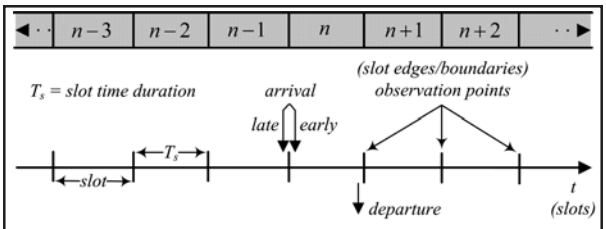



*FIG. 1. REPRESENTATION OF DISCRETE-TIME SCALE AND EVENT OCCURRENCE IN A SLOTTED SYSTEM*

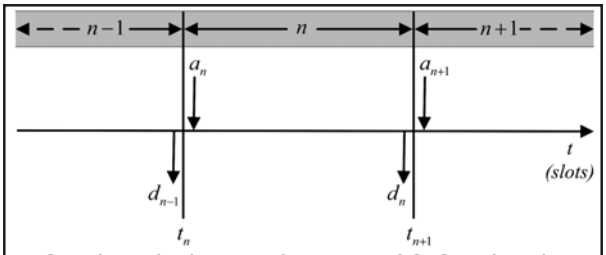



*FIG. 2. ARRIVAL AND DEPARTURE EPOCHS IN AN EARLY ARRIVAL SYSTEM*

In an early arrival system, any arriving customer may start entering the system in a slot in which it arrives unless the server is not busy, otherwise it waits and queued in the buffer which yields increased queue length upon each arriving customer until the system is completely full. Meanwhile, if a customer arrives and finds the system completely full including one in server it will be prevented entering the system. The new arriving data are only allowed to enter the system up to S-1 state of the system. The queue length decreases, whenever one service completion (departure) occurs and no customer arrives within that time slot. The system remains in same state either if there is no arrival and departure takes place or if there is one arrival and one departure takes place in a given slot for all consecutive states, excluding idle (empty) and the state when system is completely full. For an idle state, the system remains in that state either if no customer arrives or if a customer arrives and served. Whereas for the state when system is full, it remains in that state until the service completion (departure) takes place.

## 2.3 System Model

As shown in Fig. 3, the system under consideration consists of two queues, let's say first queue and second queue ($Q_1$ and $Q_2$ respectively), given that:

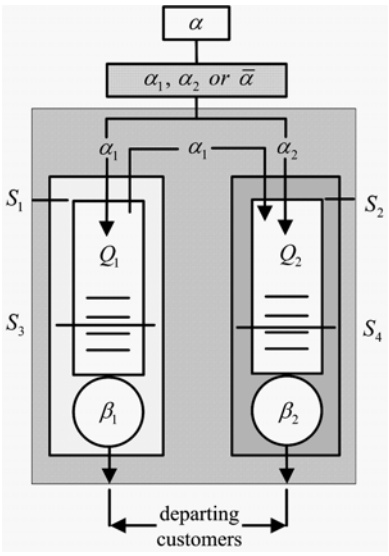



*FIG. 3. QUEUEING MODEL*

Maximum capacity of $Q_1 = S_1 = 6$, and

Maximum capacity of $Q_2 = S_2 = 5$

In addition to their finite capacity, where both queues prevents incoming customers from the output of a link when they are completely full (including service), $Q_2$ also conditionally accepts the data ($\alpha_1$) diverting from $Q_1$ due to the thresholds value $S_3$ and $S_4$ set on $Q_1$ and $Q_2$ respectively, given that:

Diverting arrivals from $Q_1$ to $Q_2$; $S_3 < 3$,

Disallowing arrivals by $Q_2$ from $Q_1$; $S_4 < 2$

The arrivals in each queue are served by two individual servers. Each queue is subject to receive its data from the output link of a source, such as network router, that generates two different types of arrivals, $\alpha_1$ and $\alpha_2$. It is important to note that the link can only forward any one or none of the two arrivals during one time slot, for an example, $\alpha_1$, $\alpha_2$, or $\alpha$. For the selection of service to the arrivals, an arrival $\alpha_1$ conditionally occupies first or second queue, where as an arrival of type $\alpha_2$ only joins the second queue, unless the system is not completely full (including any in service). Both, $\alpha_1$ and $\alpha_2$ arrives according to a Bernoulli process, whereas their inter-arrival and service time is geometrically distributed. In the case, if any of the arriving customer find all waiting places of buffer full (including service), it is prevented from entering the system (lost) otherwise it enters if there is one service completion takes place before its arrival. It is also defined, if the threshold value of first queue is less than $S_3$, it permits the arriving customers (only $\alpha_1$) in its queue, whereas second queue permits all incoming data ($\alpha_2$) that arrives in its own queue irrespective of any limit until it is not completely full.

## 3. Modeling Discrete Time Queueing System with Load Balancing

In order to model the discrete time queueing system with load balancing we take help from state transitions of the system. The state transition diagram and probabilities of the load balancing system with discrete time queueing system are shown in Fig. 4(a-b). The system state of each state is represented by two variable (i,j), where i represents

$Q_1$ and j represents the $Q_2$. The system state (0,0) represents the both queues are idle. When the customer arrives in $Q_1$, the system state change from (0,0) to (1,0) and similarly when the customer arrives in $Q_2$ the system state change from (0,0) to (0,1). When customers in $Q_1$ are less than or equal to the threshold $S_3$ if the $Q_2$ is idle or empty, the system state are (1,0), (2,0) and (3,0). The system states from (4,0) to (7,0) represents state of the system when customers in queue 1 are greater than the threshold $S_3$ and customers in queue 2 are less than or equal to $Q_1$ the threshold $S_4$ then the customers from $Q_1$ are diverted to $Q_2$. The block of system states from (0,4) to (7,6) represents that the no diversion is occurs from $Q_1$ to $Q_2$ due to the customers in $Q_2$ are reached at threshold $S_4$. The system states from (7,0) to (7,5) represents that the $Q_1$ is fully occupied but $Q_2$ is not full similarly the system states from (0,6) to (6,6) represents that the $Q_2$ is fully occupied but $Q_1$ is not full. The system state (7,6) represents that the both queues are full and system become in blocking state in which no new customer is allowed to join any queue. Fig. 4(a-b) shows the labels and its corresponding state transitions for each transition.

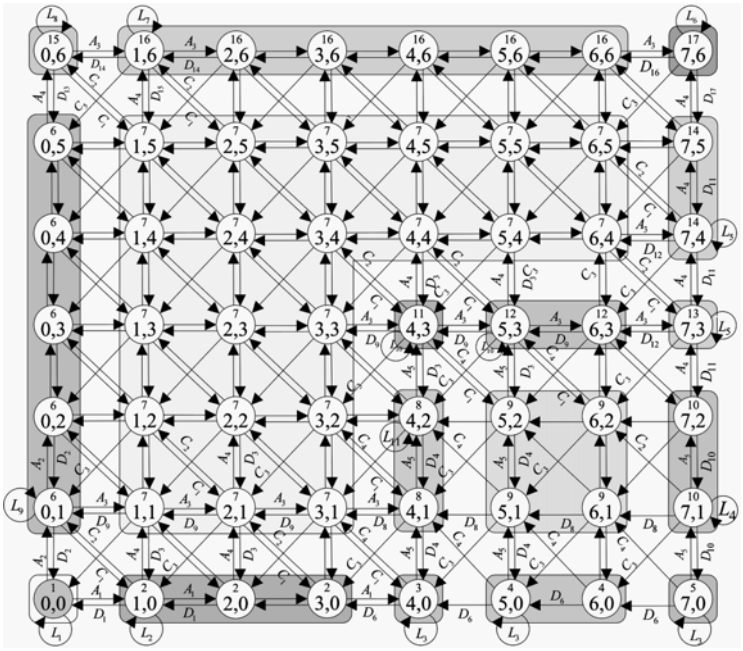


*FIG. 4(a). SYSTEM STATE TRANSITION DIAGRAM*

**Labels and their corresponding state transitions of above queueing diagram**

| | | | | | |
|---|---|---|---|---|---|
| $A_1$ | $\alpha_1\bar{\beta}_1$ | $D_2$ | $\alpha_1\beta_1\beta_2 + \bar{\alpha}\ \beta_2$ | $D_{16}$ | $\beta_1\bar{\beta}_2$ |
| $A_2$ | $\alpha_2\bar{\beta}_2$ | $D_3$ | $\alpha_1\beta_1\beta_2 + \bar{\alpha}\ \bar{\beta}_1\beta_2$ | $D_{17}$ | $\bar{\beta}_1\beta_2$ |
| $A_3$ | $\alpha_1\bar{\beta}_1\bar{\beta}_2$ | $D_4$ | $\bar{\alpha}\bar{\beta}_1\beta_2$ | $L_1$ | $\bar{\alpha} + \alpha_1\beta_1 + \alpha_2\beta_2$ |
| $A_4$ | $\alpha_2\bar{\beta}_1\bar{\beta}_2$ | $D_5$ | $\bar{\alpha}\bar{\beta}_1\beta_2 + \alpha_1\bar{\beta}_1\bar{\beta}_2$ | $L_2$ | $\alpha_1\beta_1 + \alpha_2\bar{\beta}_1\beta_2 + \bar{\alpha}\bar{\beta}_1$ |
| $A_5$ | $\alpha_1\bar{\beta}_1\bar{\beta}_2 + \alpha_2\bar{\beta}_1\bar{\beta}_2$ | $D_6$ | $\alpha_1\beta_1\beta_2 + \alpha_2\beta_1\beta_2 + \bar{\alpha}\beta_1$ | $L_3$ | $\bar{\alpha}\bar{\beta}_1 + \alpha_2\bar{\beta}_1\beta_2 + \alpha_1\bar{\beta}_1\beta_2$ |
| $A_6$ | $\alpha_2\bar{\beta}_1\bar{\beta}_2 + \alpha_1\bar{\beta}_2$ | $D_7$ | $\alpha_2\beta_1\beta_2 + \bar{\alpha}\beta_1\bar{\beta}_2$ | $L_4$ | $\bar{\alpha}\bar{\beta}_1\bar{\beta}_2 + \alpha_1\beta_2 + \alpha_2\bar{\beta}_1\beta_2$ |
| $A_7$ | $\alpha_1\beta_1\beta_2 + \bar{\alpha}_1\beta_2$ | $D_8$ | $\alpha_1\beta_1\beta_2 + \alpha_2\beta_1\beta_2 + \bar{\alpha}\beta_1\bar{\beta}_2$ | $L_5$ | $\alpha_2\beta_1\beta_2 + \bar{\alpha}_2\bar{\beta}_1\bar{\beta}_2$ |
| $C_1$ | $\alpha_1\bar{\beta}_1\beta_2$ | $D_9$ | $\alpha_2\beta_1\beta_2 + \bar{\alpha}\beta_1\bar{\beta}_2$ | $L_6$ | $\bar{\beta}_1\bar{\beta}_2$ |
| $C_2$ | $\alpha_2\beta_1\bar{\beta}_2$ | $D_{10}$ | $\bar{\alpha}\bar{\beta}_1\beta_2 + \alpha_1\beta_1\beta_2$ | $L_7$ | $\alpha_1\beta_1\bar{\beta}_2 + \bar{\alpha}_1\bar{\beta}_1\bar{\beta}_2$ |
| $C_3$ | $\bar{\alpha}\ \beta_1\beta_2$ | $D_{11}$ | $\bar{\alpha}_2\bar{\beta}_1\beta_2$ | $L_8$ | $\alpha_1\beta_1\bar{\beta}_2 + \bar{\alpha}_1\bar{\beta}_2$ |
| $C_4$ | $\alpha_1\beta_1\bar{\beta}_2 + \alpha_2\beta_1\bar{\beta}_2$ | $D_{12}$ | $\alpha_2\beta_1\beta_2 + \bar{\alpha}_2\beta_1\bar{\beta}_2$ | $L_9$ | $\alpha_1\beta_1\bar{\beta}_2 + \bar{\alpha}\bar{\beta}_2 + \alpha_2\beta_2$ |
| $C_5$ | $\bar{\alpha}_1\beta_1\beta_2$ | $D_{13}$ | $\alpha_1\beta_1\beta_2 + \bar{\alpha}_1\beta_2$ | $L_{10}$ | $\bar{\alpha}\bar{\beta}_1\bar{\beta}_2 + \alpha_2\bar{\beta}_1\beta_2 + \alpha_1\beta_1\bar{\beta}_2$ |
| $C_6$ | $\beta_1\beta_2$ | $D_{14}$ | $\bar{\alpha}_1\beta_1\bar{\beta}_2$ | $L_{11}$ | $\alpha_1\bar{\beta}_1\beta_2 + \bar{\alpha}\bar{\beta}_1\bar{\beta}_2 + \alpha_2\bar{\beta}_1\beta_2$ |
| $D_1$ | $\alpha_2\beta_1\beta_2 + \bar{\alpha}\beta_1$ | $D_{15}$ | $\alpha_1\beta_1\beta_2 + \bar{\alpha}_1\bar{\beta}_1\beta_2$ | A, C, D, L = Transition probabilities | |

*FIG. 4(b). SYSTEM STATE TRANSITION PROBABILITIES*

### 3.1 Classification and Formulation of System-State Stages

To understand the operation of model properly, in order to construct the state transition diagram representing the behavioral structure of such system, we divide it in 17 different system state stages. Each of these stages defines explicitly different sets of system state transition probabilities. We present them in a very systematic way along with the transition probabilities and their relation with their neighboring system state stages. A resultant simplified or general equation of each system state stage is obtained by equating their outgoing and incoming transition probabilities. Each stage defines explicitly different characteristics and set of system-state transition probabilities. All 17 system state stages are shown in Fig. 5.

## 4. PERFORMANCE EVALUATION

In this section we first describe the evaluation metrics. Next the simulation results are presented.

### 4.1 Performance Measure

The performance measure of the $Q_1$ can be obtained by manipulating the transition diagram of the system and equations obtained from each system state stage.

The total number of customers present in a queue is defined by mean number in the queue. The mean number in the queue is given by Equation (18) and is obtained by the sum of product of the probabilities $P(i_1, i_2)$ and the number of customers in the queues minus the customer in the server.

$$E[Q_1] = \sum_{i_2=0}^{s_2} \sum_{i_1=2}^{s_1+1} (i_1 - 1) P(i_1, i_2) \quad (18)$$

### 4.2 Results

Analytical program is written in visual C$^{++}$ for obtaining the various performance measures of the load balancing system with discrete time queueing system.

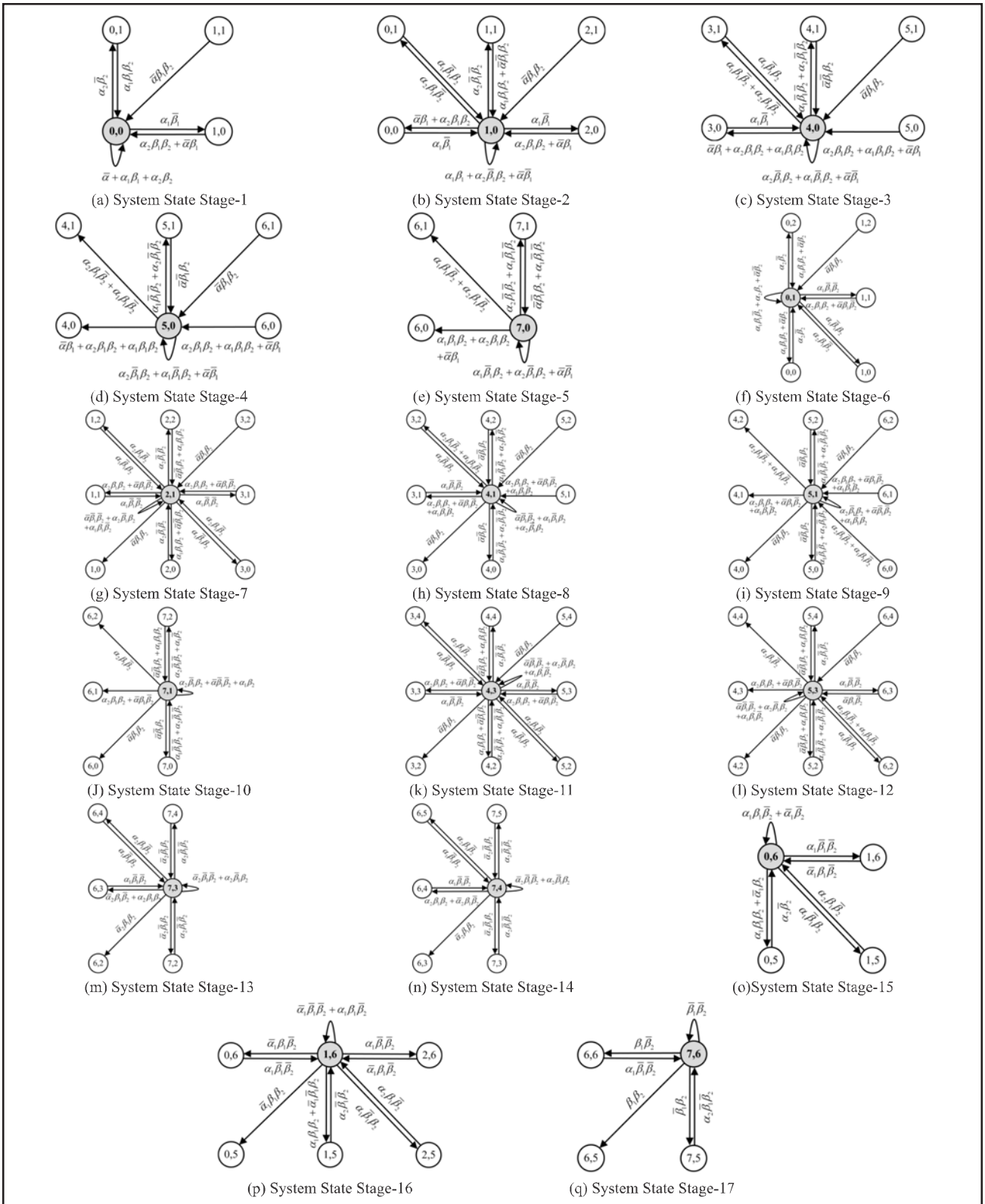
Fig. 6 shows that by changing the threshold values the mean number in the first system decreases by comparatively increasing the threshold value of second system due to the diversion of the customers towards $Q_2$. As increase in the second internal threshold, $Q_2$ is capable to handle the more customers from the $Q_1$. Figs. 7-8 show the results of mean number of customers in first system against varying probability of arrival, varying capacity of system and fixed threshold value on each queue. Fig. 7 shows the effect of keeping the internal thresholds constant and varying the queue capacity of $Q_1$ but the capacity of $Q_2$ is constant. Fig. 8 shows the number of customers in the $Q_1$ by keeping the capacity of both internal thresholds constant and varying the capacity of both queues. Results show that if we increase the threshold value to accept the customers by the second system from first system then mean number in the first system are relatively decreased.

## 5. CONCLUSIONS

Load balancing is an efficient technique used to maximize throughput when a system has limited capacity and resources to serve the arriving customers.

We have modeled the system in discrete time queueing domain using an early arrival system modeling approach by constructing its state transition diagram which features a modeling environment besides a high-level description of the system.

We show and analyze different state stages of the system along with their corresponding transition probabilities and obtained various performance results which shows second system have a very little effect due to this diversion as it only accepts customers from another system when it is in an idle state or when only minimum number of customers are present in this system. Results show that if we increase the threshold value to accept the customers by the second system from first system then mean number in the first system are relatively decreased.



(a) System State Stage-1 (b) System State Stage-2 (c) System State Stage-3

(d) System State Stage-4 (e) System State Stage-5 (f) System State Stage-6

(g) System State Stage-7 (h) System State Stage-8 (i) System State Stage-9

(J) System State Stage-10 (k) System State Stage-11 (l) System State Stage-12

(m) System State Stage-13 (n) System State Stage-14 (o)System State Stage-15

(p) System State Stage-16 (q) System State Stage-17

*FIG. 5. 17 DIFFERENT SYSTEM STATE STAGES*

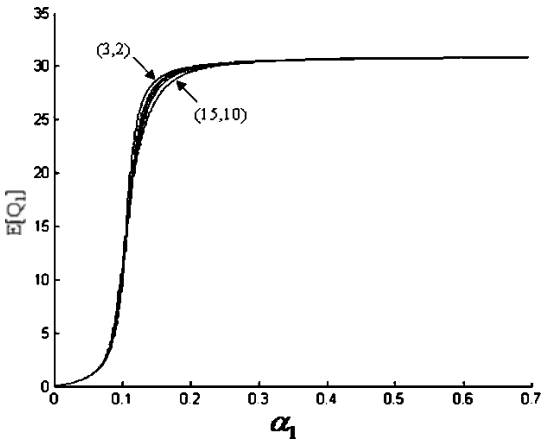



*FIG. 6. MEAN NUMBER IN $Q_1$ VERSUS THE ARRIVAL PROBABILITY AT DIFFERENT THRESHOLD VALUE OF $\alpha_1$, $\alpha_1$=0 TO 0.7, $\alpha_2$=0.3, $\alpha_1$=0.3, $\alpha_2$=0.5, $S_1$, $S_2$=30,20*

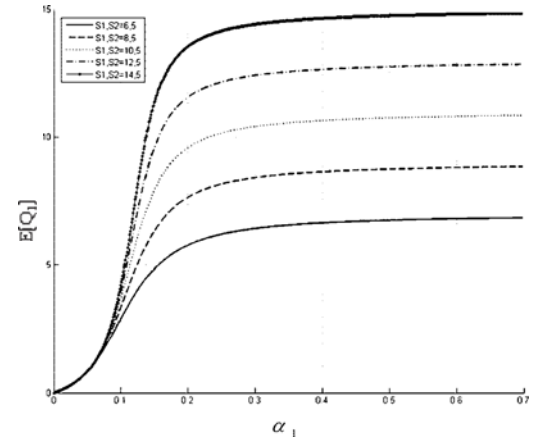



*FIG. 7. MEAN NUMBER IN $Q_1$ VERSUS ARRIVAL PROBABILITY $\alpha_1$, $\alpha_1$=0 TO 0.7, $\alpha_2$=0.3, $\alpha_1$=0.1, $\alpha_2$=0.3, $S_1$, $S_2$=(6,5), (8,5), (10,5), (12,5) AND (14,5), $S_3$, $S_4$=3,2*

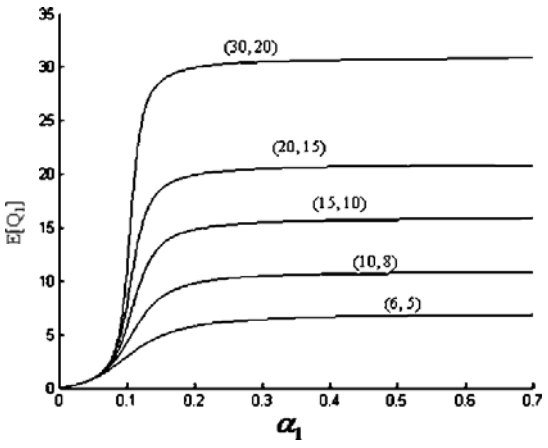



*FIG. 8. MEAN NUMBER IN $Q_1$ VERSUS ARRIVAL PROBABILITY $\alpha_1$, $\alpha_1$=0 TO 0.7, $\alpha_2$=0.3, $\alpha_1$=0.3, $\alpha_2$=0.5, $S_1$, $S_2$=(6,5), (10,8), (15,10), (20,15) AND (30,20), $S_3$, $S_4$=3,2*

## ACKNOWLEDGEMENTS

The authors are thankful to Mehran University of Engineering & Technology, Jamshoro, Pakistan, and Institute of Broadband Communication, Vienna University of Technology, Austria, for providing necessary funding and research facilities during their Ph.D. research studies.

## REFERENCES

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

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

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

[4] Bruneel, H., "Performance of Discrete-Time Queueing Systems", Computers and Operations Research, Volume 20, No. 3, pp. 303-320, Elsevier Science Ltd. Oxford, UK, 1993.

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

[6] Dattatreya, G.R., and Singh, L.N., "Relationships Among Different Models for Discrete-Time Queues", 2008.

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

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

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

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

[11] He, Q., and Nuets, M.F., "Two M/M/1 Queues with Transfer of Customers", Queueing Systems, Volume 42, pp. 377-400, 2002.

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