ISSN (e) 2520-7393
ISSN (p) 2521-5027
www.estirj.org

Transient Time Analysis of Discrete Time System with load Balancing Technique

Amir Akhtar[1],Wajiha Shah[2], Qudsia Memon[3],

[1,2,3]*Department of Electronic Engineering,MUET Jamshoro*

Abstract: Many investigators have deliberated queueing systems with load balancing. Such systems have prospective applications in sculpting the functioning of production systems, Computer and telecommunication setups with energy saving contrivance based on cyclical observing the queue state (Internet of Things, wireless sensors networks), Load balancing techniques in software defined Cloud Computing, power redeemable systems, wellbeing upkeep systems, etc. Unfortunately, conventional queuing network models are difficult to use in the continuous time transient regime because of the computational burden associated with their analysis. The purpose of this work is to study analytically the transient mode of the load balancing system in discrete time domain. This analysis is of great importance for describing the behavior of this system in a transient mode. Using load balancing technique, we will develop a double threshold model for discrete time system in transient domain. A morkov chain will be constructed from the model and state equations will be obtained analytically from the markov chain. The main motivation of the paper is to find explicit formulae and graphical illustration for queue-size distribution in the considered model in the transient (non-stationary) case.

Keywords: *Queuing system, Markov Chain, Load Balancing, Transient Time, Discrete Time System*

1. Introduction

The main indicators of the system in the transient mode are the probability of losses, the time of the transitory mode, output, and the number of customers in the system. Transient analysis of the system having queue is dependent on the time which expresses the state of the system at dissimilar times where number of clienteles are entered in the scheme. When the scheme has limited capacity and have one or fewer service facilities, then the system reaches its maximum capacity as soon as it is filled with specified capacity. The lessons on queuing systems with multi-servers generally expect the servers to serve the customers at the same rate. But such development occurs only if the service system is automatic in environment, otherwise the servers work with different rates. We cannot expect that the work will be carried out at same rate in a queuing system with human servers. In our daily life, we usually face situations of this kind, e.g., at checkout registers in retail stores, in hospitals etc.

The queuing systems with servers possessing identical service rate are appropriate only to the automatic system. Generally, discrete time queuing systems are analyzed in steady-state condition for the performance of the system. Mostly, systems are analyzed on the same parameters for the arrival rate and service rate due to equilibrium situation consideration. And other system parameters like system size are independent of starting condition when system behavior monitor for a long time performance. In terms of real time or practical situations discrete time queueing system depends on time and conditions are changeable. In various situations, the queuing system parameters may change over time period. In this situation, discrete queuing system are not operating under the steady state condition. [2] [3] [4].

The system behavior is called transient state of the system when it varies with the time. Blocking or congestion is the most important consideration in finite queuing system. When this situation occurs then the data or information becomes unable to enter in the queuing system until the system provides a space after data or information being serviced [5]. There is always a requirement to avoid this congestion or blocking situation through any suitable technique by considering the arrival and service rate with special phenomenon [6].

Queueing theory is used to study queueing systems. A queueing system is, roughly stated, any system where congestion occurs [7]. Queueing systems consist of one or multiple queues, each queue further consisting of one or multiple service stations (servers) and a buffer. Load balancing has proven to be an effective method to shrink fault incident and to address scalable user requests by balancing the incoming load. Fault tolerance is the process of making machines to work continuously even after the occurrence of fault. Load balancing is an important fault tolerant technique used in cloud computing [8]. Queueing theory plays a crucial role in modelling systems with congestion. It has been long applied in analyzing and improving the performance of communication systems. As modern communication systems often are composed of multiple heterogeneous resources, the analysis of such large-scale systems using traditional queueing theory can be

++Electronic Engineering Department Mehran University Jamshoro,
***Department of Information and Communication Technology IIU Malaysia,
****Department of Electrical and Electronic Engineering Universiti Teknologi PETRONAS, Malaysia,

prohibitive [9]. Limited capacity means finite capacity, specified capacity means assigned [10]. Discrete distributions are used to describe random variables with a finite or countably infinite number of outcomes. Examples include surveying a random person on the street for this person’s age in years, number of owned cars, number of children [11]. Simple queueing systems can often be studied directly by using Markov chains. These systems include a variety of systems where the number of servers N is small [12]. For many load balancing policies, the rate at which jobs arrive to the first queue (or any other queue) in the large scale limit, depends on the number of jobs in said queue [13]. Queueing model in which clusters of packets arrive according to a load balancing process with intensity α, can receive two types of services: usual (classic) and distinctive, for which the average duration times differ significantly. As time passes, the probability decreases until the system reaches equilibrium. The nature of the decreasing probability depends on the arrival intensity: the higher α, the faster the system reaches equilibrium.

2. Methodology

Analyze a Finite Discrete Queening system with load balancing technique to avoid the blocking. Construction of graphical representation of overall system behavior. Develop analytical equations of the system through transient domain. Transient analysis results for various system sizes.

2.1 Load Balancing System

Load balancing is the method of reallocation of load among processors to expand the performance of system. The load balancing technique, as a tool of redundancy, can considerably improve the system reliability and has important uses in plentiful engineering arenas, in queuing systems is a crucial concept for ensuring optimum resource employment, curtailing intervals, and providing a good user experience. Queuing systems involve the managing of incoming jobs or demands when resources are narrow, and load balancing helps distribute these tasks efficiently across available resources. Efficient load balancing enhances system performance, reduces response times, and improves consumer involvement.

2.2 Load Balancing Techniques

2.2.1 **Round Robin:** Assigns tasks in a circular manner, ensuring equal task distribution.

2.2.2 **Weighted Round Robin:** Assigns tasks proportionally based on resource capacities.

2.2.3 **Least Connections:** Assigns tasks to resources with the fewest active connections.

2.2.4 Weighted Least Connections: Assigns tasks based on resource capacities and active connections.

2.2.5 Random Load Balancing: Assigns tasks randomly to available resources.

2.2.6 Least Response Time: Assigns tasks to resources with the shortest response time.

2.2.7 Adaptive Load Balancing: Dynamically adjusts load distribution based on real-time resource performance.

2.2.8 Content-Based Load Balancing: Assigns tasks based on content attributes (e.g., web server farm).

2.2.9 Queue Prioritization: Processes tasks based on predefined priorities.

2.2.10 Geographic Load Balancing: Assigns tasks based on client proximity in distributed systems.

2.3 Proposed Systems

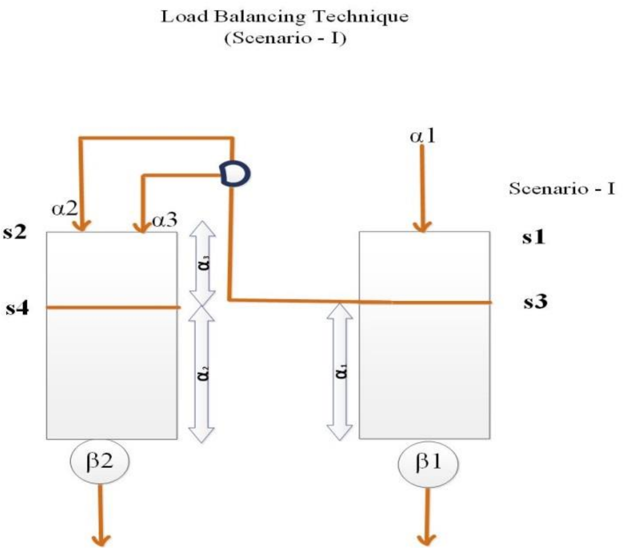


Figure.1. A graphical representation of Load Balancing Technique Model 1 (Scenario_1 (Variable Arrival rate))

2.3.1 System – I

- System consists of two queues
  1. Main System
  2. Secondary System

2.3.2 Main System

- It has two thresholds/ Limits
- S1 External
- S3 Internal
- It accepts arrivals α1 upto S3 threshold
- Uptp S3 system provides service β1
- After S3 threshold, arrivals will be switched to secondary system.
-

2.3.3 We studied the impact of load balancing on the behavior of a discrete-time dual-server queueing system under general probability distributions for both the number of customer arrivals during a slot and the length of the service time of a customer. A limited waiting room for customers. Where β is the called the rate at which jobs get serviced from the queue, i.e. the number of customers that get served per slot. Every queue has a single server. Every server provides service to the customers (jobs) that arrive to the server's queue. When studying a queueing system, we have to specify several things. First, the arrival process of the customers and the way they are distributed among the queues. The latter is called a "**load balancing policy**".

a. α>β = sys continue to grow and reached at max level →unstable →congested / blocking
b. α<β = sys stable

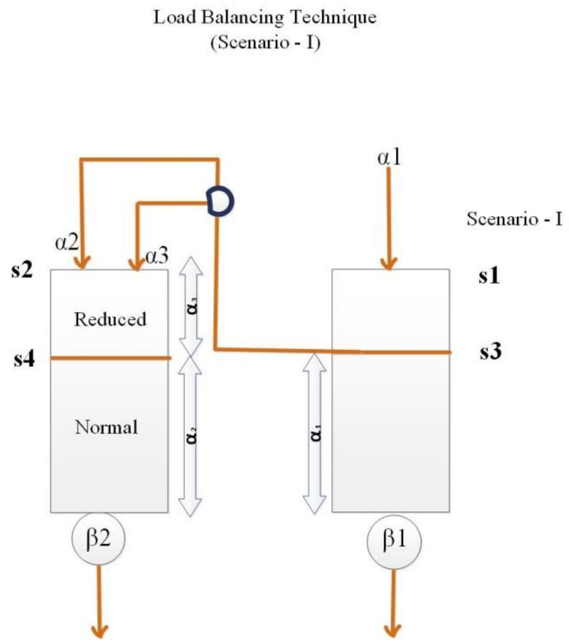


Figure.2. Load Balancing Technique with Reduced Arrivals

2.3.4 Both Queues should be congestionless

a. If arrivals reach S3 threshold, then arrivals switch / directed to the second queue
b. In second queue if arrivals reach upto S4 then arrival rate is normal
c. In second queue if arrivals reach upto S4+1 rate of arrivals becomes reduced i-e. α3
d. If arrivals in second queue reduced upto S4 then the arrival rate again becomes normal i-e. α2

3 Markov Chain

State transition diagram or markovian queuing model is a heuristic approach and measured model used in queueing theory to define the conduct of a queuing system over time. It is a memoryless process where the previous history does not help in forecasting the future state of the system depends only on its present state, not on the succession of events that steered to that state. Distribution of the time until the next arrival is independent of when the last arrival occurred. In queueing theory, Markov chains are used to evaluate and model the dynamics of how entities (such as customers, tasks, or requests) move through a queue and interact with available resources. Noted that the Markov chain is irreducible and nonexplosive.

3.2.1 Markov chain of EAS (Early Arrival System)

EAS (Early arrival system). EAS strategy is used to develop a Markov chain. In EAS, arrivals occur in the system just after starting the time slot. Or the arrival precedes the service completion epoch. Using a two-plane markovian state depiction of the system, we acquire terminologies for the α; arrival, β; service, "load" of the queue denoted by ρ of both the system 1 and 2 for arbitrary customers under a Random order of service (ROS): jobs are executed in a random order, irrespective of their arrival times as described in section 2.2.5. In words, the number of customers at queue 1 and the number at queue 2 at the same time instant are independent random variables.

When the arrival precedes the service completion eon, called Arrival First (AF) or the contrary, called Departure First (DF).

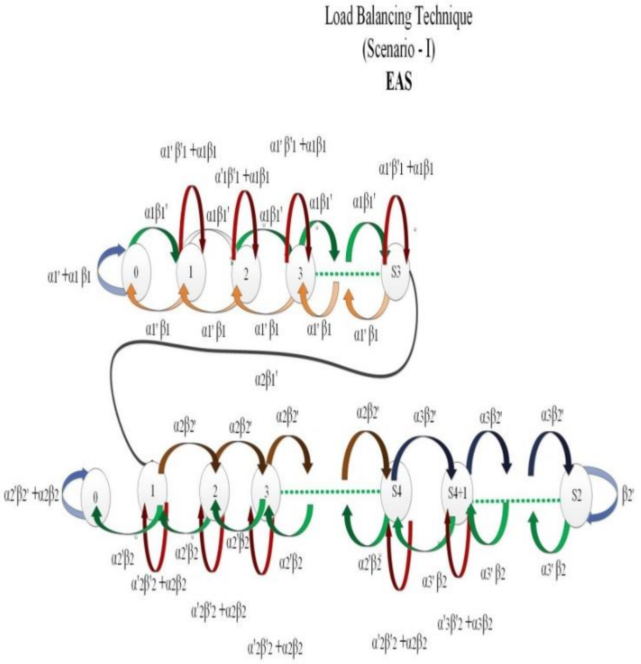


Figure.3. Markov chain of EAS with LOAD BALANCING

3.2.2 Markov chain of LAS (Late Arrival System)

In a late arrival system (LAS) arrivals occur late, just earlier to the end of a time interval, with services ending at the start of time intervals. In a late arrival system with deferred access the arriving customer is obstructed from entering an empty service facility until the servicing interval ends.

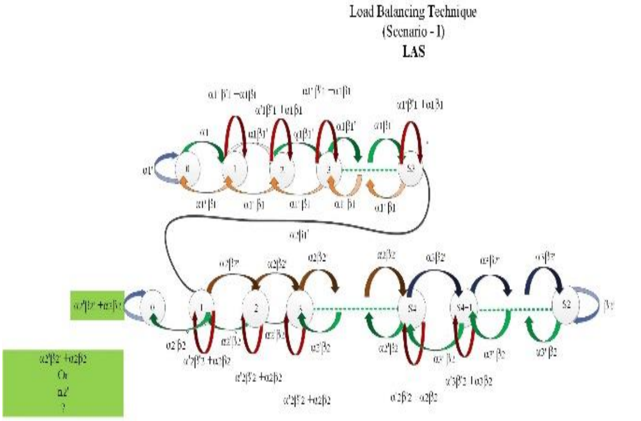


Figure.4. Markov chain of EAS with LOAD BALANCING

4. State Equations

Now it is important to find the general analytical form of the solution for the queuing system under consideration. By deriving the state equations, we can now analyze the queueing behavior. Numerical algorithm employs the theory of discrete time Markov chains of quasi-birth-death type. State equations have been derived from above illustrated markov chains for each model respectively as per their title mentioned against each. Critical expressions for computing the probabilities of system states in a transient mode have been calculated and results are mentioned below.

4.1 System - I

State Equations for Load balancing technique *Scenario I* EAS

State 0

$$P_1 = \rho_1 P_0 \qquad (1)$$

State 1

$$P_2 = \rho_1^2 P1 \qquad (2)$$

State s$_3$

$$P_{S3} = \rho_1^3 P_0 \qquad (3)$$

4.2 System - II

State Equations for Load balancing technique *Scenario I* EAS

State 0

$$P_1 = 1/\ \alpha_2'\beta_2 \qquad (1)$$

State 1

$$P_2 = (\rho_2 + 1)\ P_1\ - [\ \alpha_2\ \beta_1'/\alpha_2'\beta_2\ ]\ \rho_1^3\ P_0 \qquad (2)$$

State S$_4$

$$P_{S4} = P_2\ \ (1 + \rho_2\ ) - \rho_2\ P_1 \qquad (3)$$

4.3 System – I

State Equations for Load balancing technique *Scenario I* LAS

State 0

$$P_1 = [\ \alpha_1\ /\ \alpha_1'\ \beta_1\ ]P_0 \qquad (1)$$

State 1

$$P_2 = [\ \rho_1\ \alpha_1\ /\ \alpha_1'\ \beta_1\ \ ]P_0 \qquad (2)$$

State S$_{3\text{-}1}$

$$PS_{3\text{-}1} = [\ \rho_1^2\ \alpha_1\ /\ \alpha_1'\ \beta 1\ ]P_0 \qquad (3)$$

$$PS_{3\text{-}1} = \rho_1^3\ /\ \beta_1 \qquad (4)$$

4.4 System - II

State Equations for Load balancing technique *Scenario I* LAS

State 0

$$P_1 = 1/\alpha_2'\beta_2 \qquad (1)$$

State 1

$$P_2 = 1+\rho_2/\alpha_2'\beta_2 \qquad (2)$$

State S$_{4\text{-}1}$

$$PS_{4\text{-}1} = 1+\rho_2+\rho_2^2/\alpha_2'\beta_2 \qquad (3)$$

5. Results and Discussion

The results are carried out on MATLAB software. Moreover, a software package called QSA (Queueing Systems Assistance) developed in 2021 has been utilized to calculate and visualize the main performance measures. In addition, it helps to depict a quite general mean total service time with arrival rate and server utilization. This is the easiest way to verify the results. All results have been validated through Simulink as well.

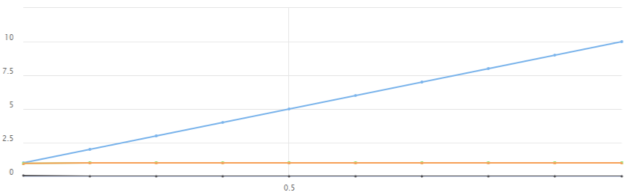


Figure.5. Mean System Content

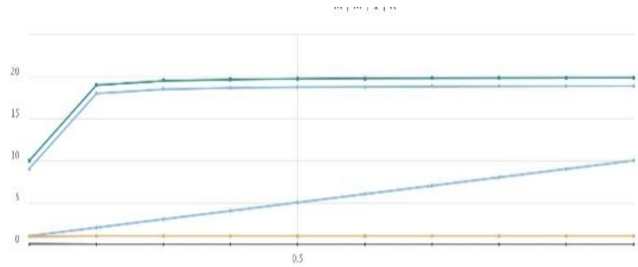


Figure.6. Queue size, system size and delay, alpha =0.1

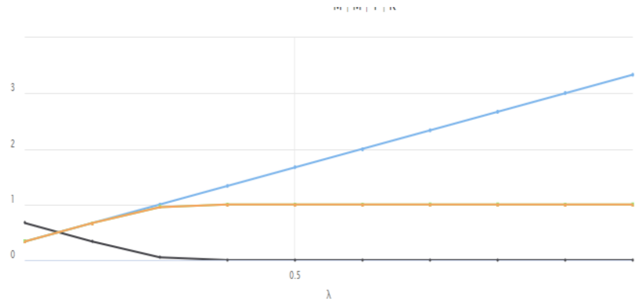


Figure.9. When value of α is set to 0.3

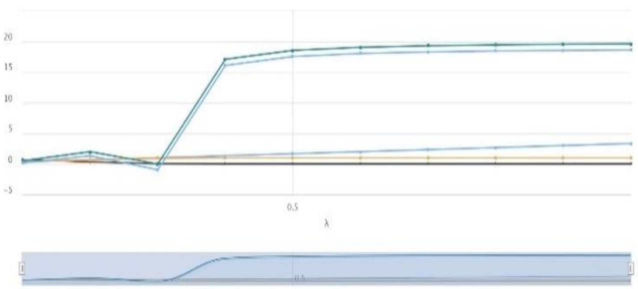


Figure.10 .Service rate of the system

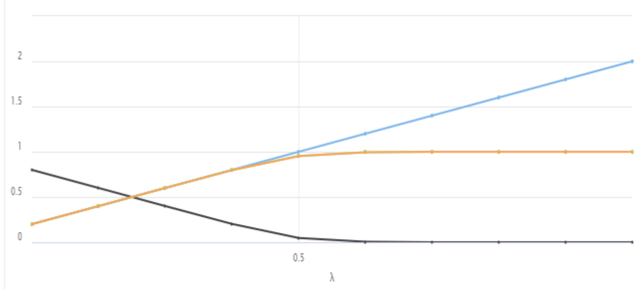


Figure.11. When value of α is set to 0.5

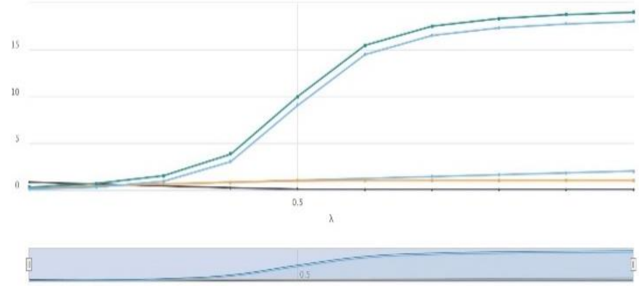


Figure.12. N=α/β-α N_s= Mean no in the sys

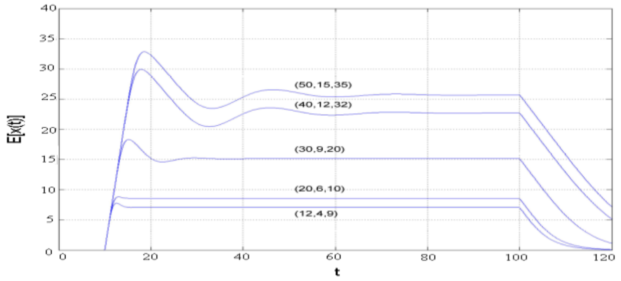
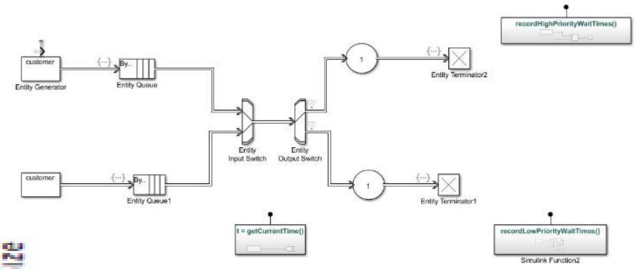
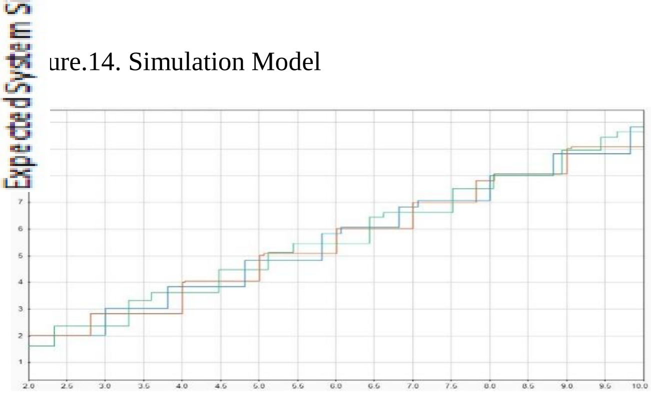


Figure.13 Reliance of the probabilities on the time. Mean system content E [x(t)] versus the number of customers in a system at a time t



ure.14. Simulation Model



]The comparison of two queuing systems with respect to average retention rate

6. Conclusion

This research considers a discrete system queueing model with the transient time. We used load balancing technique with double threshold method to illustrate the queue's arrival process. Through numerical derivations, the impact of load balancing on the system characteristics was evaluated. A comparative analysis is performed which shows that the heterogeneity in service improves the system performance. Transient distributions of the queue size and waiting time were obtained there using quasi-birth-and-death markovian chain and state equations. In the discrete-time system, for each info source, there is a core ROS (Random Order of Service) process that is randomly sampled, Discrete-time Markov forming is fairly dissimilar than that of continuous time as in the latter, only one event can happen at a given time. However, in discrete time, manifold events can happen at the same time instant. This research delved into the intricate realm of transient time analysis for discrete-time systems with a specialized focus on integrating load balancing techniques within the framework of queuing theory. State expressions for obtaining the probabilities of system states in a transient mode have been calculated. The investigation aimed to unravel the intricate relationship between system dynamics, queuing behavior, and load distribution, ultimately shedding light on how load balancing strategies can significantly enhance the transient response of such systems. By effectively managing the allocation of resources and tasks, load balancing strategies demonstrated their potential to expedite response times, minimize waiting periods, and mitigate potential bottlenecks. This was exemplified through the reduction of queue lengths, shorter customer wait times, and improved overall system efficiency. The exploration of transient time analysis in discrete-time systems with load balancing techniques within the realm of queuing theory has unveiled a promising avenue for enhancing system dynamics and responsiveness. The integration of queuing theory and load balancing techniques offers a novel perspective in the realm of system optimization.

Finally, in a discrete-time system, the service rate is improved with arrival modals, demonstrating that as the number of consumers in the system grows, so does the efficiency.

References

[1] Marn-Go Kim, “Discrete time domain modeling and design of current mode controlled flyback LED driver”, 2023.

[2] Gorkem Gok, Ozgul Salor, “Transient event classification using pmu data with deep learning techniques and synthetically supported training-set” 2023.

[3] Wojciech M. Kempa , and Rafał Marjasz, “Study on Transient Queue-Size Distribution in the Finite-Buffer Model with Batch Arrivals and Multiple Vacation Policy” 2023.

[4] Grzegorz Kielanski, “Analysis of load balancing and scheduling policies in large-scale systems” 2023.

[5] Nail Akar, “Discrete-Time Queueing Model of Age of Information With Multiple Information Sources” 2023.

[6] Ridhima Mehta, “Discrete-time simulation for performance modelling of FIFO single-server queuing system” 2022.

[7] Shensheng Tang, “Transient Analysis of a finite Queuing System with bulk arrivals in IoT –Bassed Edge computing system ”, 2022.

[8] Daniela Hurtado-Lange1 ·Siva Theja Maguluri, “A load balancing system in the many-server heavy-traffic asymptotics” 2022.

[9] Vladimir Vishnevsky, Konstantin Vytovtov , Elizaveta Barabanova and Olga Semenova, “Transient Behavior of the MAP/M/1/N Queuing System” 2021.

[10] SUNIL SUBEDI, MANISHA RAUNIYAR, TIMOTHY M. HANSEN, “Review of Dynamic and Transient Modeling of Power Electronic Converters for Converter Dominated Power Systems”, 2021.

[11] CHINMAY SHAH, CAMPO-OSSA, GURUWACHARYA, TREVIZAN, “Review of Dynamic and Transient Modeling of Power Electronic Converters for Converter Dominated Power Systems.” 2021

[12] Madhu Jain and Mayank Singh, “Transient Analysis of Markov Model with Feedback, Discouragement and disaster” International Journal of Applied amd computational Mathematcis, 2020.

[13] Vijiya Laxmi and T. Windewosen Kassahun, “Transient analysis of multi-server Markovian queueing system with synchronous multiple working vacations and impatience of customers” 2020.

[14] Q.Wang, V.S. Frost, “A new solution technique for discrete queuing analysis of ATM system”, IEEE GLOBECOM’91, 2019, USA.

[15] ALFONSO PARREÑO TORRES, PEDRO RONCERO-SÁNCHEZ, JAVIER VÁZQUEZ, JAVIER LÓPEZ-ALCOLEA AND EMILIO J. MOLINA-MARTÍNEZ, “A Discrete-Time Control Method for Fast Transient Voltage-Sag Compensation in DVR.” (2019).

[16] R. Sudhesh and R. Sebasthi Priya, “An analysis of discrete-time Geo/Geo/1 queue with feedback, repair and disaster” 2019.

[17] Wojciech M. Kempa, “Transient study of a discrete-time queueing model with feedback mechanism” 2019.

[18] Flip Paluncic at el, “Queuing models for cognitive radio networks: Servey”, IEEE Access, 2018.

[19] Ramupillai Sudhesh and Arumugam Vaithiyanathan, “Analysis of state-dependent discrete-time queue with system disaster” 2018.

[20] Kunjie Xu, David Tipper, “Time-Dependent Performance Analysis of IEEE 802.11p Vehicular Networks”, IEEE Transactions on Vehicular Technology (Volume: 65, Issue: 7, July 2016).

About Authors

Amir received his B.E degree from Mehran University of Engineering and Technology, Jamshoro Sindh and is enrolled in M.E at Mehran University of Engineering and Technology, Jamshoro Sindh. His research area is transient time analysis of discrete time systems

Dr. Wajiha Shah received her B.E and M.E degree from Mehran University of Engineering and Technology, Jamshoro, Pakistan, and her Ph.D. degree from Vienna University of Engineering and Technology, Vienna, Austria.

Qudsia Memon received her B.E and M.E degree from Mehran University of Engineering and Technology, Jamshoro, Pakistan. Her research area is neural network.