EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Efficient SDN-based Task offloading in fog-assisted cloud environment

Bibhuti Bhusan Dash[1,2], Rabinarayan Satpathy[2] and Sudhansu Shekhar Patra[1,*]

[1]School of Computer Applications, KIIT Deemed to be University, Bhubaneswar, India
[2]Faculty of Emerging Technologies, Sri Sri University, Cuttack, India

Abstract

A distributed computing model called "fog computing" provides cloud-like services which is closer to end devices, and is rapidly gaining popularity. It offers cloud-like computing including storage capabilities, but with less latency and bandwidth requirements, thereby improving the computation capabilities of IoT devices and mobile nodes. In addition, fog computing offers advantages such as support for context awareness, scalability, dependability, and node mobility. Fog computing is frequently used to offload tasks from end devices' applications, enabling quicker execution utilizing the fog nodes' capabilities. Because of the changing nature of the fog environment, task offloading is challenging and the multiple QoS criteria that depend on the type of application being used. This article proposes an SDN-based offloading technique to optimize the task offloading technique for scheduling and processing activities generated by the Internet of Space Things (IoST) devices. The proposed technique utilizes Software-Defined Networking (SDN) optimization to dynamically manage network resources and to facilitate the deployment and execution of offloaded tasks. To model the system which computes the optimal virtual machines (VM) to be allocated in the fog network in order to actively process the offloaded tasks, the GI/G/r queueing model is utilised. This approach minimizes the delay-sensitive task queue and minimises the necessary number of VMs while minimising the waiting time for the fog layer. The findings of the simulation are used to verify the effectiveness of the proposed model.

Keywords: Task Offloading, Software defined networking, Energy Optimization, Queueing model, GI/G/r

Received on 24 September 2023, accepted on 06 December 2023, published on 13 December 2023

doi: 10.4108/eetiot.4591

*Corresponding Author. Email: sudhanshupatra@mail.com

1. Introduction

In the present era, the use of mobile nodes and sensor devices has become quite popular for collecting rich data and executing complex tasks. However, these devices often face resource constraints, such as limited storage and computing power. In order to address these issues, jobs have been "offloaded," or transferred, to external servers using cloud computing in order to maintain these devices and transfer data to other servers.

Although cloud computing has several benefits, Centralized processing and resulting latency in cloud computing may not be suitable for latency-sensitive tasks. Fog computing has surfaced as a distributed computing paradigm to address this issue, enabling IoT devices to be in close proximity to wireless nodes, sensors, and cloud-like services. In comparison to the cloud servers, fog devices are found nearer to the edge devices in the nearby proximity. By providing processing, networking, and storage capabilities to devices with limited resources, fog computing provides cloud-like services closer to IoT devices, sensors, or mobile nodes. As of edge computing, fog computing seeks to solve the issues that standard cloud computing presents with IoT applications.

The proposed paper aims to present an SDN-based offloading technique that optimizes task offloading in the fog environment. As the SDN controller can gather network information using southbound APIs and makes the best task offloading decisions based on its overall view of the network, the technique makes use of Software-Defined Networking (SDN) optimisation to dynamically manage network resources and facilitate the deployment and execution of offloaded tasks. The GI/G/r queueing system can be

employed to model the system and determine the ideal number of VMs to employ in the fog network.

One popular model that is often used to research queuing systems is the GI/G/r queuing model. It's a mathematical model that may be used to predict how many jobs should be in a queue and how long they should take to finish. Assumed by the model, tasks come according to a general independent inter-arrival time and are served by r virtual machines (VMs), with service times falling within a general service time. The study includes numerical analyses and simulations to support the suggested model. The suggested approach is successful in optimising task offloading in the fog environment, as shown by the numerical results of the simulations that are run.

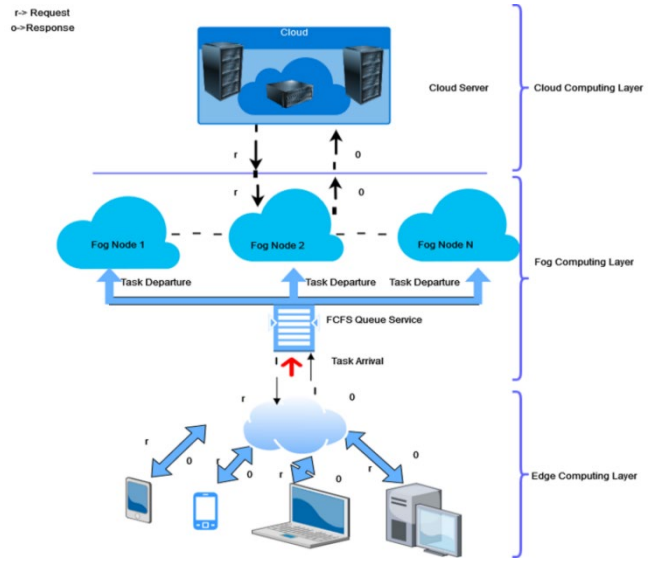

Figure 1: Fog assisted cloud model

Numerical evaluations and simulations confirm the proposed model and show that it is useful in optimising task offloading in the fog layer. The proposed approach can be useful for IoT applications and other resource-constrained environments.

This article is arranged as follows. Section 2 provides a list of earlier works in this field, the system model is shown in section 3. , section 4 the energy consumption in SDN-assisted fog center, section 5 the numerical results and section 6 provides a conclusion of the article and an outline for further research..

2. Related Work

The fog model architecture is built upon the 3-layer concept of cloud, fog, and edge, with the fog layer located closer to the edge, consisting of multiple Fog Nodes (FNs) that may share resources and provide services to clients while maintaining service quality. But since FNs come in many varieties and capabilities, it's likely that some of them won't be able to perform computationally demanding jobs, which might result in delays and unexpected processing outcomes, among other quality of service (QoS) problems. Fog computing design makes it difficult to offload and provide timely service as FNs must work together to overcome these obstacles. Thus, in order to ascertain if the fog layer is capable of accomplishing the offloading mission, a thorough investigation is necessary. To improve system performance, certain FNs must cooperate well with one another. Recent research studies have focused on workload offloading in edge or fog computing, with many examining the efficiency of energy and resource allocation. In certain experiments, researchers created models like the M/M/1 queue model to reduce resource costs and response times while reducing network power consumption and maintaining delay restrictions. By employing the M/M/1 and the M/M/n model in the fog centres, the delay related to queuing models and computing is computed.

This Research on an M/M/m/K queuing system using extra servers was done by authors in [8]. The concept of "no passing" for a multi-server queueing model with dissuasion and two customer kinds was first presented in [9]. Authors in [10] established the $G/G^{y}/m$ queuing model and discourage customers using the diffusion approximation, the authors in [11] analyzed the discouragement-based diffusion approximation for a G/G/m queuing model. In [12] authors looked at a Markovian queuing model that included server expansion and discouragement to shorten wait times. A finite buffer multi-server queuing system with retained reneged requests' steady-state likelihood of reneging behavior clients was determined in [13,14] determined the probabilities of the size of the stable system for the multi-server queuing system's balking behaviour. The general response times for cloud server farms are covered in [15-19], which applies the M/G/1, M/G/m, and M/G/c/k queuing models to get performance measures including energy usage, infrastructure costs, and mean response times.

3. System Model

Many tasks are offloaded to the multi-server fog computing centres so they can receive service. In accordance with user needs, the fog centre offers services. Figure 1 depicts the fog computing system's implementation mechanism.

Consider a fog centre with r parallel virtual machines. Randomly and individually, tasks offloaded to the fog centres. The service times also follow a general distribution, and the inter-arrival times are independently and identically distributed (IID). Inter-arrival and service timings are independent of one another as well. First come, first served is the system's guiding principle for how it treats users. The request must be placed in the buffer and left to wait if the task delegated to the fog centre cannot discover any spare VMs. When a virtual machine (VM) is vacant, a task request is sent to it and is promptly processed there. Figure 2 illustrates how the fog centre is viewed as a GI/G/c queuing system. The tasks are offloaded to the fog centre with an arrival rate of λ and the service rate of each VM is $\mu=1/s$ where s is the average service time of the offloaded tasks, the

square of the coefficient of inter-arrival time is Cv_a^2 and the square of the coefficient of service time is Cv_s^2. The VM utilization is given by $\rho = \lambda/c\,\mu$. $P < 1$.

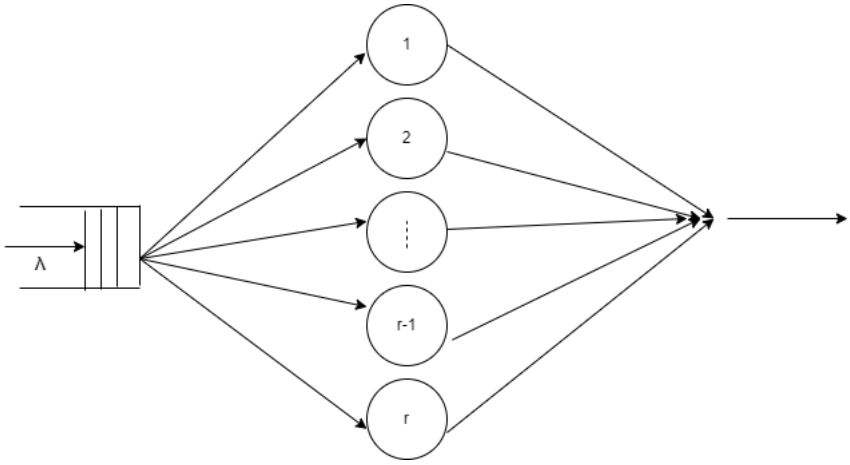

Figure 2: Queueing Model of fog layer

Based on GI/G/m, the fog centers are analyzed.

Let
Λ= average number of tasks offloaded (arrived) to the fog center
W= waiting time of a task offloaded to the fog center.
E(W)= mean waiting time of the offloaded task in the fog center.
$E(L_q)$= The mean number of tasks waiting in the queue in the fog center.
$E(L_s)$= mean number of tasks in the VMs.
E(L)= mean number of tasks in the system
T= waiting time of a task or time a task spends in the fog center before leaving the system, also called the sojourn time.

$$E(L) = E(Lq) + E(Ls)$$
$$E(T) = E(W) + s \quad (1)$$

Using Little's law

$$E(T) = \frac{E(L)}{\lambda}, E(W) = \frac{E(L_q)}{\lambda}, \quad (2)$$
$$E(L_s) = \lambda s = r\rho$$

The average waiting time in the GI/G/r fog system is:

$$E(W) = \phi(\rho, Cv\_a^2, Cv\_s^2, r) \times \left(\frac{Cv\_a^2 + Cv\_s^2}{2}\right) EW(M/M/r) \quad (3)$$

$$\phi(\rho, Cv\_a^2, Cv\_s^2, r) = \begin{cases} \left(\frac{4(Cv\_a^2 - Cv\_s^2)}{4Cv\_a^2 - 3Cv\_s^2}\right)\phi_1(r,\rho) \\ +\left(\frac{Cv\_s^2}{4Cv\_a^2 - 3Cv\_s^2}\right)\psi\left(\frac{Cv\_a^2 + Cv\_s^2}{2}, r, \rho\right), \\ \quad when \quad Cv\_a^2 \geq Cv\_s^2 \\ \left(\frac{Cv\_s^2 - Cv\_a^2}{2Cv\_a^2 + 2Cv\_s^2}\right)\phi_3(r,\rho) + \\ \left(\frac{Cv\_s^2 + 3Cv\_a^2}{2Cv\_a^2 + 2Cv\_s^2}\right)\psi\left(\frac{Cv\_a^2 + Cv\_s^2}{2}, r, \rho\right) \\ \quad when \quad Cv\_a^2 \leq Cv\_s^2 \end{cases} \quad (4)$$

$$\phi_1(r,\rho) = 1 + \gamma(r,\rho) \quad where\ \gamma(r,\rho) = \min\left\{0.24, \frac{(m-1)(sqrt(4+5r)-2)}{16r}.\frac{1-\rho}{\rho}\right\} \quad (5)$$

$$\phi_2(r,\rho) = 1 - 4\gamma(r,\rho) \quad (6)$$

$$\phi_3(r,\rho) = \phi_2(r,\rho)e^{\frac{-2(1-\rho)}{3\rho}} \quad (7)$$

$$\phi_4(r,\rho) = \min\left\{1, \frac{\phi_1(r,\rho) + \phi_3(r,\rho)}{2}\right\} \quad (8)$$

$$\psi(c^2, r, \rho) = \begin{cases} 1, & c^2 \geq 1 \\ \phi_4(r,\rho)^{2(1-c^2)}, & 0 \leq c^2 \leq 1 \end{cases} \quad (9)$$

EW(M/M/r) is the average waiting time of M/M/r queueing model.

$$EW(M/M/r) = \frac{(r\rho)^r}{r!}((1-\rho)\sum_{i=0}^{r-1}\frac{(r\rho)^i}{i!}+\frac{(r\rho)^i}{i!})^{-1}.\frac{1}{1-\rho}.\frac{1}{r\mu} \quad (10)$$

4. Energy Consumption in SDN-assisted fog center

A computing center that employs SDN-assisted fog computing can be characterized by its energy consumption, which is composed of dynamic E_d and static E_s components over a given time period. The static E_s component, which represents the fundamental element of total energy consumption and has a fixed value, is mainly determined by the center's hardware components and circuit layout, and cannot be modified. Meanwhile, dynamic processes account for the majority of the energy consumed by fog center VMs, comprising approximately 80% of the total energy consumption.

A single VM's dynamic power usage per unit of time is given by $E_d \approx k\mu^{\alpha}$ watts, where $\alpha \geq 3$, k is the power usage scaling factor.

The VM utilization is $\rho = \frac{r}{\lambda\mu}$.

Therefore, the overall energy usage per unit of time for the fog center, taking into account the dynamic energy consumption, can be expressed as,

$$E = r\rho k\mu^{\alpha} = k\lambda\mu^{\alpha-1} \quad (11)$$

The total power used in the fog centre per unit of time is denoted by E_{total}.

$$E_{total} = rE_s + E = rE_s + k\lambda\mu^{\alpha-1} \quad (12)$$

5. Numerical Results

Using python 3.9 and by generating the synthetic data, we illustrate the numerical outcomes produced by the proposed approach. Table 1 shows the various performance measures of the fog center of the offloaded tasks by keeping the fixed ρ and λ. The constant parameters is taken as μ=1.5, r=8, α=3, k=1, Cv_a=0.25 and Cv_s=1, k=1, Es=0.4, α=3. The arrival rate λ= [6,7.2,8.4,9.6,10.8,11.4]. According to the table, when the VM utilization rate (ρ) of the fog center increases, there is an increase in the mean waiting length of the queue(E_q), mean number of client's request in the VMs ($E(L_s)$), average number of client's request in the system (E(L)), mean waiting time (E(W)), mean sojourn time (E(T)), and the total energy usage (E_{total}). In essence, a fog center's capacity to process service requests through each VM remains constant. Once the values of r and μ are determined, the overall capacity of the service center becomes fixed, and an increase in the arrival rate of λ leads to an increase in the system's load.. When a task is offloaded into the fog center and discovers that every VM is occupied, it must wait in the buffer. The length of the waiting queue and waiting time increase with the system load. The mean waiting time grows as VM utilization rises. When the VM utilisation of the fog centre is high, it is required to expand the number of VM in order to better improve the service level.

Table 2 depicts the various performance of the system for different λ and r by keeping fixed μ=1.5, r=8, Cv_a^2=0.25 and Cv_s^2=1, k=1, ρ=0.95, α=3. As the number of VMs increases for a fixed ρ, the measures E (L_s), E(L), and E_{total} increase, while the measures E(L_q), E(W), and E(T) decrease. E(Ls) rises as r rises while the fog center's VM utilisation stays constant. It indicates that more offloaded tasks are handled concurrently by the VMs. As the number of VMs r increases, conversely, the average waiting time decreases as well.. The average waiting time E(W) exhibits a declining trend as r continues to rise. However, there is an increase in overall energy consumption (P_{total}).

Table 3 gives the effect of the square of the coefficient of variation of the service time. Cv_s^2 with various performance matrices of the fog environment where the tasks are offloaded. Table 3 shows that as Cv_s^2 increases E(L_q), E(L), E(W), and E(T) increase, while E(L_s) and E_{total} remain constant. E(W) and E(L_q) increase as Cv_s^2 increases.

Table 4 shows when Cv_a^2 increases E(L_q), E(L), E(W), and E(T). But E(L_s) and E_{total} is constant as they are independent of Cv_a^2.

Fig 3 shows the total energy consumed E_{total} is increasing as ρ and λ and r and λ increases as shown in fig (a) and (b). As the squared coefficient of variation $C_v\_s^2$ increases E_{total} remains constant. Similarly the with the increase in squared coefficient of variation of the request inter-arrival time $C_v\_s^2$ the total energy consumption E_{total} remains unchanged. It is shown in fig (c) and (d).

6. Conclusions and Future Work

The utilization of fog computing can aid in performing latency sensitive by offloading the responsibilities of the edge devices to fog layers for processing. This approach offers shared and flexible computing and communication resources, in addition to providing an affordable computing communication infrastructure. When the edge devices offloaded the tasks to the fog layer and all the virtual machines (VMs) are busy fulfilling other client requests, they adhere to the GI/G/r queueing paradigm and remain in the waiting queue. These tasks are then processed using the FCFS scheduling approach by the VMs. The system validation process includes multiple numerical examples presented in the form of figures to assist the service provider in modeling the system. In the future, the application of auction theory may enable the transfer of work from the cloud layer to the fog layer.

Table 1. Performance analysis of the fog center for various ρ and λ.

ρ	λ	$E(L_q)$	$E(L_s)$	$E(L)$	$E(W)$	$E(T)$	E_{total}
0.5	6	0.0185	4.0000	4.0185	0.0031	0.6697	16.7
0.6	7.2	0.0774	4.8000	4.8774	0.0107	0.6774	19.4
0.7	8.4	0.2767	5.6000	5.8767	0.0329	0.6996	22.1
0.8	9.6	0.9244	6.4000	7.3244	0.0963	0.7630	24.8
0.9	10.8	3.5832	7.2000	10.7832	0.3318	0.9984	27.5
0.95	11.4	9.5739	7.6000	17.1739	0.8398	1.5065	28.85

Table 2. Performance analysis of fog center for various r and λ.

r	λ	$E(L_q)$	$E(L_s)$	$E(L)$	$E(W)$	$E(T)$	E_{total}
4	5.7	10.2352	3.8000	14.0352	1.7956	2.4623	14.425
8	11.4	9.5739	7.6000	17.1739	0.8398	1.5065	28.85
16	22.8	8.6926	15.2000	23.8926	0.3813	1.0479	57.7
32	45.6	7.5555	30.4000	37.9555	0.1657	0.8324	115.4
64	91.2	6.1523	60.8000	66.9523	0.0675	0.7341	230.8
128	182.4	4.5326	121.6000	126.1326	0.0248	0.6915	461.6

Table 3. Performance analysis of fog center for various Cv_s^2 and λ.

Cv_s^2	λ	$E(L_q)$	$E(L_s)$	$E(L)$	$E(W)$	$E(T)$	E_{total}
0	11.6	2.0356	7.6020	9.6319	0.1781	0.8452	28.87
0.3	11.6	4.0124	7.6020	11.6097	0.3519	1.0187	28.87
0.6	11.6	5.7132	7.6020	13.3025	0.5003	1.1669	28.87
0.80	11.6	7.6298	7.6020	15.2288	0.6695	1.3362	28.87
1.1	11.6	9.5756	7.6020	17.1742	0.8392	1.5065	28.87
1.35	11.6	11.5398	7.6020	19.1382	1.0126	1.6792	28.87
1.55	11.6	13.5223	7.6020	21.1223	1.1862	1.8537	28.87

Table 4. Performance analysis of fog center for various Cv_a^2 and λ.

Cv_a^2	Λ	$E(L_q)$	$E(L_s)$	$E(L)$	$E(W)$	$E(T)$	E_{total}
0	11.6	7.5292	7.6000	15.1292	0.6605	1.3271	28.87
0.3	11.6	9.5739	7.6000	17.1739	0.8398	1.5065	28.87
0.6	11.6	11.6733	7.6000	19.2733	1.0240	1.6906	28.87
0.80	11.6	13.8281	7.6000	21.4281	1.2130	1.8797	28.87
1.1	11.6	16.0392	7.6000	23.6392	1.4069	2.0736	28.87
1.35	11.6	18.1644	7.6000	25.7644	1.5934	2.2600	28.87
1.55	11.6	20.2273	7.6000	27.8273	1.7743	2.4410	28.87

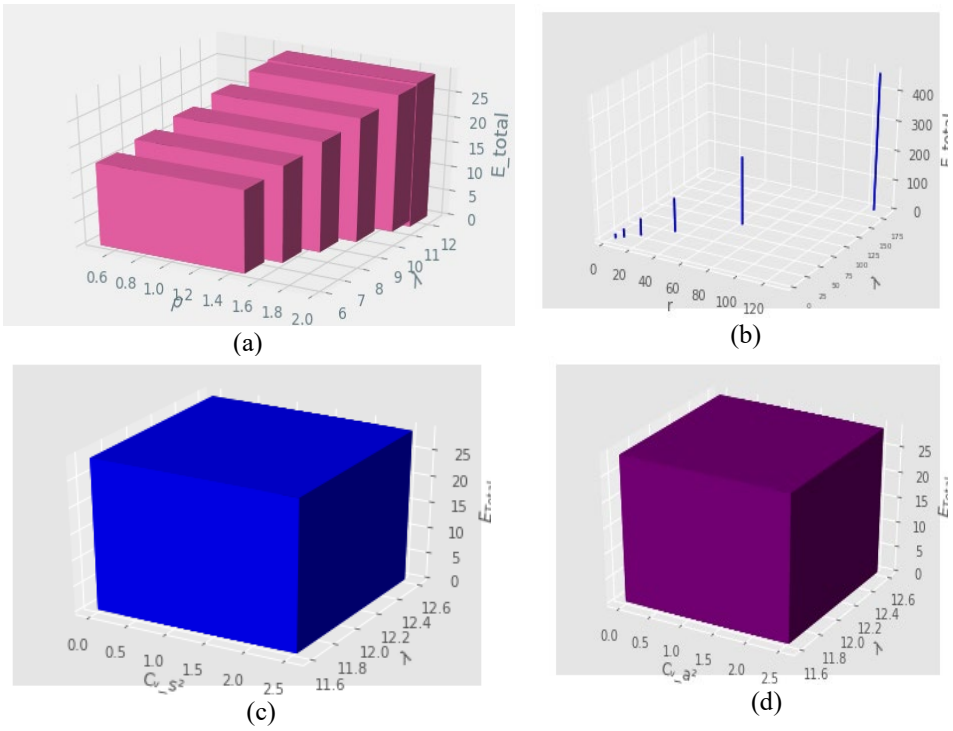

Figure 3: a) Total Energy Consumed w.r.t ρ and λ b) w.r.t r and λ c) w.r.t C_v_s^2 and λ d) w.r.t C_v_a^2 and λ

References

[1] Journal article: Samizadeh Nikoui, T, Rahmani, A. M, Balador, A, & Haj Seyyed Javadi, H. Internet of Things architecture challenges: A systematic review. International Journal of Communication Systems. 2021; 34(4): e4678.

[2] Journal article: Mukherjee M, Shu L, Wang D. Survey of fog computing: Fundamental, network applications, and research challenges. IEEE Communications Surveys & Tutorials. 2018; 20(3): 1826-1857.

[3] Journal article: Sotomayor, B, Montero, R. S, Llorente, I. M, Foster, I: Virtual infrastructure management in private and hybrid clouds. IEEE Internet computing. 2009; 13(5):14–22.

[4] Journal article: Sarkar S, Misra S: Theoretical modelling of fog computing: a green computing paradigm to support IoT applications, IET Networks. 2016; 5(2): 23–29.

[5] Journal article: Patra.S.S: Energy-efficient task consolidation for cloud data center. International Journal of Cloud Applications and Computing (IJCAC).2018; 8(1): 117-142.

[6] Conference: Tassi A, Mavromatis I, Piechocki R, Nix A, Compton C, Poole T, Schuster W: Agile data offloading over novel fog computing infrastructure for CAVs. In 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring); 28 Apr-1 May 2019, Kuala Lumpur, Malaysia: IEEE; 2019,1-6.

[7] Journal article: Misra S, Saha N: Detour: Dynamic task offloading in software-defined fog for IoT applications. IEEE Journal on Selected Areas in Communications. 2019; 37(5):1159-1166.

[8] Conference: Varshney K, Jain M, Sharma G. C: The M/M/m/K Queuing System with Additional Servers for a Large Queue. Proceeding of the seminar 65th Birthday celebration of Prof. S.C. Dasgupta, 1988, 277-282.

[9] Journal article: Jain M, Sharma B, Sharma G. C: On No passing Multiserver Queuing Model with Two Types of Customers and Discouragement. Journal of Mathematical Physics Scimago. 1989; 23(4): 319 - 329.

[10] Journal article: Garg K. M, Jain M, Sharma G. C: "G/Gy/m Queuing System with Discouragement Via diffusion Approximation. Microelectronics Reliability. 1993; 33(7): 1057-1059.

[11] Journal article: Varshney K, Jain M, Sharma G. C: Diffusion Approximation for G/G/m Queuing System with Discouragement. Journal of the Indian Statistical Association. 1987; 25: 91- 96.

[12] Journal article: Jain M: M/M/m Queue with Discouragement and Additional Servers. Gujarat Statistical Review. 1998; 25(1-2): 31-42.

[13] Journal article: Kumar R, Sharma S. K: An M/M/c/N queuing system with reneging and retention of reneged customers. International Journal of Operational Research. 2013; 17: 333–344.

[14] Journal article: Kumar R, Sharma S. K: A Markovian multi-server queuing model with retention of reneged customers and balking. International Journal of Operational Research. 2014; 20(4): 427–438.

[15] Journal article: Khazaei H, Misic J, Misic V. B: Performance analysis of cloud computing centers using $/g/m/m + r$ queuing systems. IEEE Transactions on Parallel and Distributed Systems. 2011; 23 (5): 936–943.

[16] Conference: Outamazirt A, Barkaoui K, Aissani D: Maximizing profit in cloud computing using $M/G/c/k$ queuing model. In 2018 International Symposium on Programming and Systems, (ISPS); 24-26 Apr 2018, Algeria: IEEE, 2018, 1–6.

[17] Conference: Khazaei H, Misic J, Misic V. B: Modelling of cloud computing centers using $M/G/m$ queues. In 2011 31st International Conference on Distributed Computing Systems Workshops; 20-24 June 2011, Minneapolis, Minnesota USA: IEEE, 2011, 87-92.

[18] Conference: Goswami V, Patra S. S, Mund G. B: Performance analysis of cloud with queue-dependent virtual machines. In 2012 1st International conference on recent advances in information technology (RAIT); 15-17 Mar 2012, ISM Dhanbad, India: IEEE, 2012, 357-362.

[19] Journal article: Patra S. S, Govindaraj R, Chowdhury S, Shah M. A, Patro R, Rout S: Energy Efficient End Device Aware Solution Through SDN in Edge-Cloud Platform. IEEE Access. 2022; 10: 115192-115204.