International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 23 (2015) pp 43716-43721

The Apartment service dispatch strategy and load estimation

Kakoli Bora

Assistant Professor, Department of Information Science and Engineering,
PESIT-South campus, Bangalore-560100 k_bora@pes.edu

Sumana Sinha

Assistant Professor, Department of Information Science and Engineering
PESIT-South campus, Bangalore-560100 sumanasinha@pes.edu

Abstract
Waiting in lines is a part of daily life. Usually long queues cause customer dissatisfaction, which in turn, affect the economy of an organization. This paper demonstrates the effective management of service requests originating from the residents of Multi-unit Apartments using Queuing theory under a set of feasible assumptions. The primary focal point is on the design and execution of an M/M/1 queuing model capable of handling service requests according to the order of arrival and dispatching them to the server for further processing. A First Come First Serve (FCFS) service discipline is employed with minimal loss of requests. The inter-arrival time and the service time are exponentially distributed with a continuous window. The paper analyzes the Quality of Service (QoS) parameters such as the usage of the system, expected number of service requests and the mean waiting time of the service requests in the system using the queue model.

Keywords: Apartment service management system; Queuing theory; utilization; service manager; exponential distribution.

Introduction

The problems arise from the length of the waiting lines (e.g. waiting in line at the Bank for a cashier, in a ticket counter, at billing counter in shopping malls, at hospitals, waiting for the doctor, at the server side during the arrival of HTTP requests, arrival of telephone calls at the switchboard, at logistics management or at factories) and solution is sought to handle them in cost effective manner. If the waiting times and service times are longer then customers may leave the queue resulting in customer dissatisfaction. The fundamental problem in virtually every waiting line is a trade-off decision. Queuing system is considered as the flow of units towards a system in the hope of getting service, forming or joining a queue, if service is not immediately available, leaving the system after being served.

The increase in number of Multi-unit residential buildings in every city of India/South Asia and subsequent portfolio of service requests generated by every unit efficiently for the facility manager has become very challenging. According to a study prepared by a strategic advisory group reported in [8], 20600 residential units (Apartments as well as villa/row houses) were absorbed and 44000 residential units were established in Bangalore in the year 2011. Granting to the report of LJ Hooker's 470 residential projects (128689 units) in the categories of apartments, villas, row houses, villament's and plots at the end of year 2012. This statistic demonstrates the development of real estate business in Bangalore. Thus a strict measure should be taken to ensure residents optimal satisfaction and to avoid the total collapse of the system. Everyday a huge amount of service requests related to carpentry, electricity, plumbing, seepage etc. are received by the facility management team who tediously perform the task of meeting the service requests, manage them efficiently and follow up diligently after it has been completed. This paper concentrates on an automated system to evaluate the effectiveness of queuing models in handling the incoming service request used by the service manager with the following assumptions-

- All service requests are legitimate
- Service requests related to common blocks (block outage) are not counted
- No shows of service personnel are not counted
- No Inventory management is covered
- "Not honoured" service requests need to be reformed
- Any service request sent off but not delivered is not a priority.

In this paper, a solution has been presented to efficiently cover the service requests that arrive from a finite bit of flat owners in First Come First Serve basis with a continuous window of service (24 hours) using M/M/1 queuing models . The strategy estimates the average number of requests in the queue, the average waiting time of a service request and the system utilization rate. If utilization or traffic intensity is high there is a high probability of buffer overflow. This work presents a solution to this problem with a conservative deployment by the way of one server only.

Queuing models have been used in a variety of applications. In [2], M/M/1 queuing model and Little's theorem are used to study the waiting lines in a Bank's ATM in Gujarat and to improve the service rate of the ATM. Using queuing theory, utilization factor, average waiting time in the queue and the median number of customers in the queue are calculated. The outcome indicated that there was no significant difference between the actual service charge per unit and calculated service rate. An another application of queuing theory in health care management described in [1], which applies a single server and multiple queue model. They used Chi-square and Erlang's queuing and made some recommendations to

reduce the waiting time in line. In [3], multiple server queuing model was employed to improve the waiting time of a customer in a Bank in Nigeria. The analysis suggested the Bank to increase the number of servers from 3 to 5. Birth-death process model is used in [4] to optimize logistics processes i.e. delivering the orders efficiently. A fuzzy decision making algorithm is suggested to assign components to orders randomly which optimizes the logistic process. The queuing theory is used in Remote Medical monitoring described in [6]. The vital signs are gathered from sensors attached to human bodies and are transmitted to the remote server present in Hospitals. They have designed an M/M/1 queuing model to handle the readings of the remote patients and send them in turn to the medical personnel when the demand arises. In [5] various statistical properties are studied to determine which distribution is suitable for Mobile Communication Network. It is assumed that arrival rate is Poisson distribution and service time is exponentially distributed in a traffic situation in mobile communication networks. These statistical properties since to the best suitable in mobile communication networks because of their unique parameters and are mere to dissect. Samuel and Jeffrey summarizes in [8] the contribution and application of queuing theory in the area of health care. This paper gives a range of queuing theory results in the areas like: waiting time and system utilization analysis, system design and appointment system.

Methodology

This study concentrates on the probability distribution of arrival rate of the service requests and their service times. All service requests waiting to be dispatched are kept in a queue and there is a service manager to manage the requests. The inter-arrival time and service time are exponentially distributed; the system is memory less. Memory less is defined as, at any point of time, the future of the scheme does not depend on its past. The service request to the system arrives one at a time and is independent of each other. A birth-death process is useful in modelling systems in which jobs arrive one at a time. The state of such system can be represented as a function of 'n' denoted as f(n) where n is the number of service requests in the system. When a new request arrives, the state of the system changes to (n+1) hence the birth of a process. Likewise, when a request is completed, state of the system changes to (n – 1) causing death in the process. Hence the scheme manager is positioned as an M/M/1 Queue.

Service Equilibrium Probability theorem

Assume, the arrival and dispatch of service requests are managed by one queue i.e. service manager in the system. Further, assume that the "arrival-dispatch" process imitates the birth-death operation.

Then, the equilibrium probability of such a process in an arbitrary state 'n' is given by

$$p_n = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{j=1}^{n} \mu_j} p_0$$

where,

$$P\left\{n(t+\delta t) = {}^{j}\!/_{n(t)} = j\right\} = 1 - \lambda_j \delta t - \mu_j \delta t$$

and $n \le Z_{\max}, Z_{\max} \in Z - \{Z^-\}$;

Z^- is the set of negative integers.

P_0 = prob(n=0); i.e. the probability of being in the zeroth state.

Proof: Assume, $\delta t \to 0$;

prob (two arrivals OR two departures OR one arrival and one departure) = 0.

Thus, the probability that the system will stay in steady state, i.e. state "j" (no arrivals or departures) is,

$$P\{n(t+\delta t) = \frac{j}{n(t)} = j\} = 1 - \lambda_j \delta t - \mu_j \delta t$$

Let, $P_j(t)$ be the probability of being in the state j, the following system of equation represents the state defined probabilities,

$$p_j(t+\delta t) = \lambda_{j-1}\delta t p_{j-1}(t) + (1 - \mu_j \delta t - \lambda_j \delta t) p_j(t) + \mu_{j+1}\delta t p_{j+1}(t);$$

$j = 1, 2...$

Considering the limit of δt on either side, we obtain

$$\lim_{\delta t \to 0} = \frac{p_j(t+\delta t) - p_j(t)}{\delta t} = \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t)$$

i.e.

$$\frac{dp_j(t)}{dt} = \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t) \qquad \textbf{(1)}$$

Using steady state conditions, $\lim_{t\to\infty} p_j(t) = p_j$ and

$$\lim_{t\to\infty} \frac{dp_j(t)}{dt} = 0,$$

we obtain

$$\lambda_{j-1} p_{j-1} - (\mu_j + \lambda_j) p_j + \mu_{j+1} p_{j+1} = 0$$

$$\Rightarrow p_{j+1} = \left(\frac{\mu_j + \lambda_j}{\mu_{j+1}}\right) p_j - \frac{\lambda_{j-1}}{\mu_{j+1}} p_{j-1}, \; j = 1, 2, ...$$

$$p_1 = \frac{\lambda_0}{\mu_1} p_0$$

$$\Rightarrow p_{j+1} = \left(\frac{\mu_j + \lambda_j}{\mu_{j+1}}\right) p_j - \frac{\lambda_{j-1}}{\mu_{j+1}} p_{j-1}, \; j = 1, 2, ...$$

$$p_1 = \frac{\lambda_0}{\mu_1} p_0 \qquad \textbf{(2)}$$

Applying the principle of Mathematical Induction, it can be demonstrated that the solution to (2) is,

$$p_n = \frac{\prod_{i=0} \lambda_i}{\prod_{j=1} \mu_j}; n = 1, 2...Z_{\max}$$

and $Z_{\max} < \infty$, p_0 can be easily computed as,

$$p_0 = \frac{1}{1+\sum_{n=1}^{Z_{max}}\prod_{j=0}^{n-1}\frac{\lambda_j}{\mu_{j+1}}}$$

QoS measures

The QoS measures are elucidated under the assumptions of the Single Service Manager i.e. M/M/1 queue. The following assumptions of M/M/1 queue which seem fair enough in the context of the problem are outlined:

- One service request processor (service manager)-conservative design
- No limits on buffer size or population
- Inter-arrival time and service time are exponentially distributed
- Service discipline is First Come First Serve
- State of the queue is a function of the number of service requests in the system
- Mean arrival time $\lambda_n = \lambda$ and mean service time $\mu_n = \mu$; n = 1,2,… with 24 hour service window.

Thus, the Probability (n service requests in the system); $p_n = \left(\frac{\lambda}{\mu}\right)^n p_0, n = 1, 2, ...$ and $\frac{\lambda}{\mu}$ is traffic intensity.

Let,

$\rho = \frac{\lambda}{\mu}$, then $p_n = \rho^n p_0$;

and $p_0 = \frac{1}{1+\rho+\rho^2+...} = \frac{1}{1-\rho}$

Hence, $p_n = (1-\rho)\rho^n$; and the number of service requests is geometrically distributed.

QoS parameters

1. Utilization: one or more service requests in the system; $\mu = 1 - p_0 = \rho$. Utilization of the resource (Service Manager) is directly proportional to the traffic intensity.
2. Expected number of service requests in the system, $E[n] = \sum_{n=1}^{\infty} np_n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$
3. Variance of the number of service requests in the system, $\mathrm{Var}[n] = \frac{\rho}{(1-\rho)^2}$
4. Prob(number of requests ≥ n) = $\sum_{j=n}^{\infty} p_j = \rho^n$
5. Mean response time, $E[r]\frac{1/\mu}{1-\rho}$ (using Little's law) **(3)** and CDF of the response time, $F(r) = 1 - e^{-r\mu(1-\rho)}$ which is exponentially distributed.
6. q-percentile of the response time: $r_q = \frac{1}{\mu(1-\rho)}\ln\left(\frac{100}{100-q}\right)$
7. CDF of the waiting time, $F(w) = 1 - \rho e^{-\mu w(1-\rho)}$

Mean waiting time, $E[w] = \rho\frac{1/\mu}{1-\rho}$

Variance of the waiting time, $Var[w] = \frac{(2-\rho)\rho}{\mu^2(1-\rho)^2}$

8. Prob(number of jobs in busy period) = $\frac{1}{n}\left({}^{2n-2}C_{n-1}\right)\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}$
9. $L \equiv$ average number of customer in the system.

$\equiv \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\left(\frac{\lambda}{\mu}\right)^n\left(1-\frac{\lambda}{\mu}\right) = \frac{\lambda}{\mu-\lambda} = \left[\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}\right], 0 < x < 1$

$w \equiv$ average amount of time a customer spends in the system

$= \frac{L}{\lambda} = \frac{1}{1-\lambda}$

$w_Q \equiv$ average amount of time a customer spends in the queue

$L \equiv$ average number of customer in the queue

$= \lambda w_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} \equiv Queuelength$

Probability that an arbitrary customer spends some amount of time in the system,

$$P\{w^* \le a\} = \sum_{n=0}^{\infty} P\{w^* \le | N = n\}P(N = n),$$

where $P(N = n)$ is the probability that there exist "n" customer in the system already. Consider the amount of time the customer must spend in the system given that there are "n" customer already.

$N = 0$, is service time trivially

$N \ge 1$, One customer in service and $(n-1)$ waiting in queue. Since the arrival pattern is memoryless(exponential), the arrival would have to wait an exponential amount of time with "μ ". The customer would have to wait an exponential amount of time for each of the $(n-1)$ customer in line, $(n-1) \Rightarrow$ the sum of $(n+1)$ independent and identically distributed random variable with rate"μ" $\Rightarrow$ Gamma distribution with parameters $n+1, \mu$

$$P\{w^* \le a \mid N = n\} = \int_0^a \mu e^{-\mu t}\frac{(\mu t)^n}{n!}\,dt$$

$$P\{w^* \le a\} = \sum_{n=0}^{\infty}\int_0^a \mu e^{-\mu t}\frac{(\mu t)^n}{n!}\left(\frac{\lambda}{\mu}\right)^n\left(1-\frac{\lambda}{\mu}\right)dt$$

$$= \int_0^a (\lambda - \mu) e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\mu t)^n}{n!} dt$$

$$= \int_0^a (\lambda - \mu) e^{-\mu t} e^{\lambda t} dt = 1 - e^{-(\mu - \lambda)dt}$$

$\Rightarrow w^*$ the amount of time a customer spends in the system is exponential random variable with rate $(\mu - \lambda)$.

Toy Example

The above mentioned QoS parameters are discussed below with an example.

In a 1000 flat apartment complex, it is observed that service requests arrive at a mean rate of 150 per day and the feeder takes about 1 minutes to forward those. Let's analyze the service feeder as well as the probability of buffer overflow if the feeder has 50 buffers. How many buffers are needed to keep the "information loss" below 1 request per 10000?

Given, λ = 150, μ will be $\mu = \frac{24 \times 60}{1} = 1440$

Intensity, $\rho = \frac{150}{1440}$; $E[r] = \frac{1/\mu}{1-\rho}$

Prob (n requests in the service manager) = $(1-\rho)\rho^n$

Expected number of requests = $\rho/(1-\rho)$

Expected time in the service manager = $\frac{1/\mu}{1-\rho}$

Prob (buffer overflow) = Prob (more than 50 requests in the system) = ρ^{50}

Minimize "request loss": $\rho^n < 10^{-4} \Rightarrow n > \frac{\log(10^{-4})}{\log(\rho)} = 4.07$

The above equation estimates the minimum required buffers to keep the information loss below 1 per 10000. This estimate of number of buffer is a good approximation.

High level System design of Service Request in Multi-unit Apartment

Facility management system in multi unit residential apartment comprises of service request sent by a naive user, i.e. residents of the apartment. The class of the service request sent by the user is marked off with a list of services menu stored in the system. If the request is of type block outage like transformer outage, pipeline leakage, and so forth, the request will be discarded otherwise the request is considered as Independent outage like fuse blown, water tap broken, door knob broken, etc.. The independent outage request is appended to the queue in FCFS basis; hence a birth of a process. The scheduler selects a request from the queue and gives it to the service manager for processing. Once a request is processed, it leaves the queue; hence death of a process.

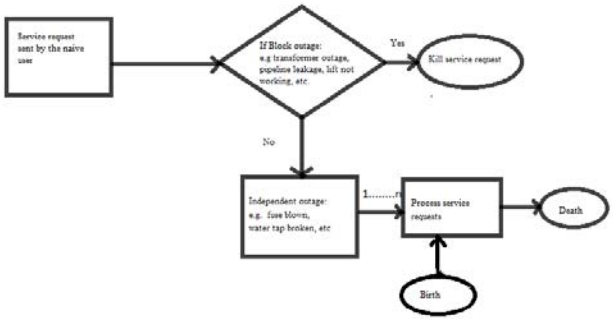

Fig.1. High level System Design

Results and Discussion

In M/M/1 queue models, arrival rate cannot be greater than service rate which states that as the response time increases the utilization approaches to one. Thus, utilization should be always less than one in the given model.

The response time as a function of utilization is shown in Figure 2. Thus with only one server, the system can afford maximum utilization. The waiting time W, of the service requests, is a random variable which is non negative and it's mean is E[W] ≤ 0.

The data set can be downloaded from, https://drive.google.com/a/pes.edu/file/d/0B_Eh9X8o9ckMSXdyS1EyN3R1ZTA/view?usp=sharing which is collected from a local residential apartment complex with 4800 units.

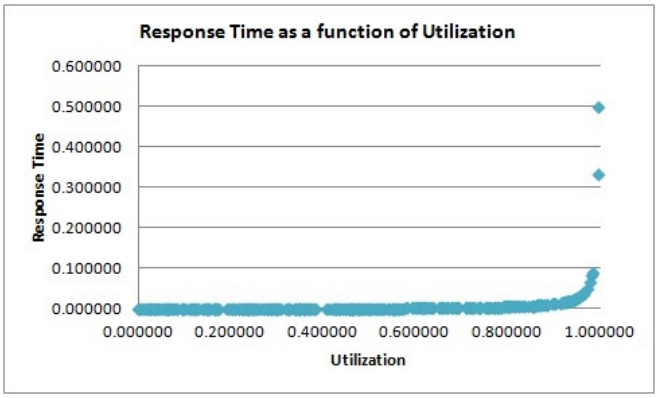

Fig.2. Response time vs. utilization for one year

Theorem:

The probability of waiting time to be greater than a certain threshold is reasonably low i.e. if w is a nonnegative random variable and c is the threshold than, $P\{w \geq c\} \leq \frac{E[w]}{c}$

Implication:

The system thus modeled is indeed robust with utilization less than 1 and waiting time practically low. Proof of the Concept:

Since $w \geq c$

$\Rightarrow w \geq c.1$ i.e $w \geq I\{w \geq c\}$

$\Rightarrow E[w] \geq E[c.I\{w \geq c\}$

$= cE[I\{w \geq c\}]$

$= c.Pw \geq c\}.1$

$\Rightarrow \frac{E[w]}{c} = P\{w \geq c\}$

Where, $I(w \geq c) = 1, w \geq c$

$= 0$, otherwise

Illustration:

$E[w]$ =0.007540 (for given data set)

$Var[w] = 0.000078$

If $c = 1$ hour, say then $P\{w \geq c\} \equiv P\{waiting\ time \geq 1hour\}$ ≤ 0.000078 i.e. the probability of large waiting time is really small.

Figure 3 shows that as utilization of the system approaches to one, mean waiting time of the service requests asymptotically increases. Hence, by keeping the utilization less than one, waiting time of the system can be minimized. It is also observed that if the number of requests sent during busy period is high, probability of serving those requests during that period is much low. From Figure 4 & 5, it is seen that if arrival rate is less than the service rate, mean waiting time as well as mean response time of the system are much lower. Thus, our model is best suited for this application with one server so that all the user requests can be efficiently handled by keeping utilization of the system below one with available resources.

Exponential smoothing is applied to data to produce smoothed data as seen in Figure 6. The predicted arrival also works well within the traffic intensity of less than one rendering stability to the system under the conservative design.

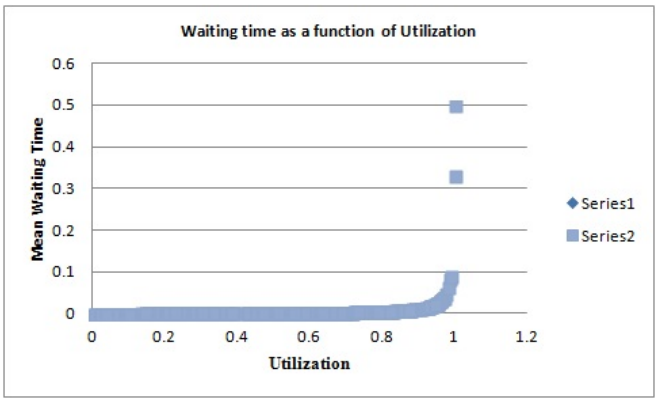

Fig.3. Mean waiting Time vs. Utilization for one year

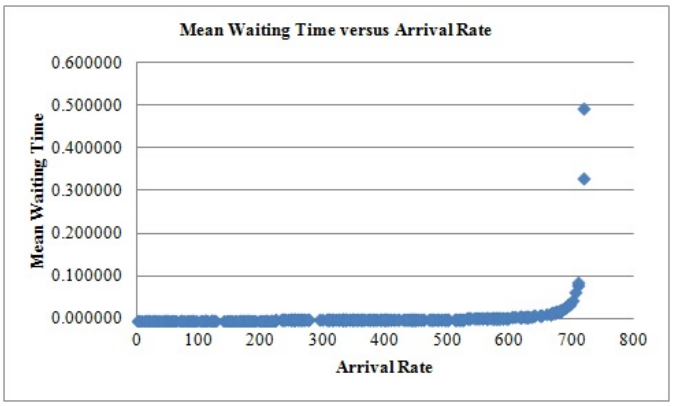

Fig.4. Arrival Rate as a function of mean waiting time

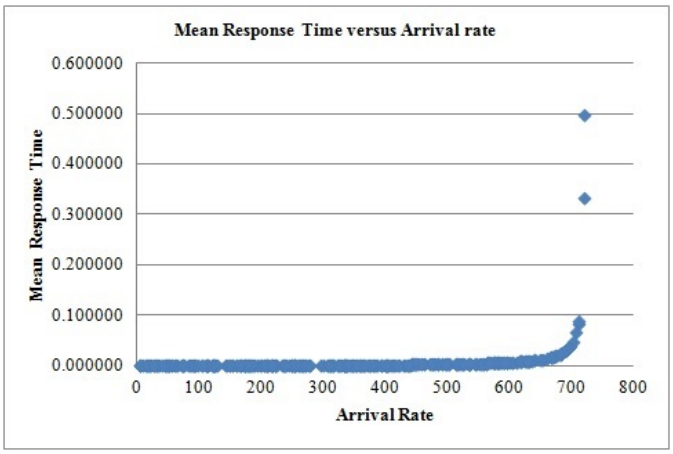

Fig.5. Arrival Rate as a function of mean response time

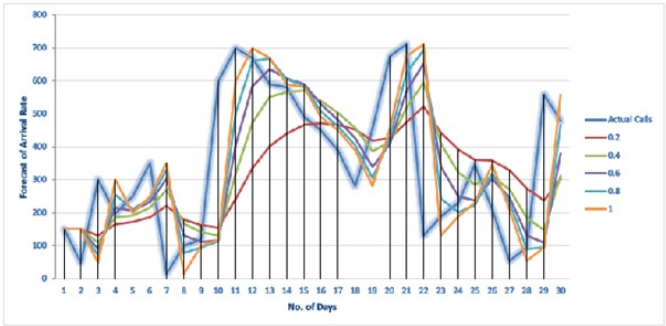

Fig.6. Forecast of arrival rate of requests

Conclusion and Future work

This paper presents handling of service requests in a multi-unit residential apartment using queue model parameters to determine the performance of the system under a set of assumptions. The service request queue is examined using the M/M/1 queuing model. The arrival rate and service rate of the service requests are analyzed, both being exponentially distributed, fitting the problem well. A result is applied to feed the utilization below one so that response time can be reduced and average waiting times of the requests can be minimized. The estimation of the queue capacity reduces the number of service request loss.

The future plans include design of a service scheduler to handle "No show" cases and to generate and distribute the tokens for a service request. As the utilization approaches to one, the response time as well as waiting time asymptotically increases. Therefore it is a smart strategy to feed the utilization below one and asymptotic behavior does not kick in. This is precisely the focus of the work i.e. to check the system's robustness. The future goal is to design a system in such a way (end to end solution) that the response time is reasonably minimum and utilization is optimal. In this context, this method serves the purpose except for a few outliers. The solution to this problem may be addressed via M/M/k queuing model with finite buffers.

References

[1] Imahsunu, Albert Felix. Queuing theory for Health Care Operations Management: A Case Study of University of Benin Health Center and Faith Mediplex.

[2] Bhavin Patel, Pravin Bhathawala. Case Study for Bank ATM Queuing Model: *International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622*, Vol. 2, Issue 5, September-October 2012, pp. 1278-1284.

[3] Dr. Engr. Chuka Emmanuel Chinwuko, Ezeliora Chukwuemeka Daniel , Okoye Patrick Ugochukwu, Obiafudo Obiora J.Analysis of a queuing system in an organization (a case study of First Bank PLC, Nigeria): *American Journal of Engineering Research (AJER)e-ISSN: 2320-0847* p-ISSN: 2320-0936 Volume-03, Issue-02, pp-63-72.

[4] João Miguel da Costa Sousa, Rainer Palm, Carlos Silva, Thomas A. Runkler. Optimizing Logistic Processes Using a Fuzzy Decision Making Approach: *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: HUMANS AND SYSTEMS,* VOL. 33, NO. 2, MARCH 2003.

[5] Osahenvemwen O.A., Edeko F.O., Emagbetere J. Elusive Statistical Property of Arrival Rate and Holding Time used in Mobile Communication Networks: *International Journal of Computer Applications (0975 – 8887)* Volume 59– No.2, December 2012

[6] Ifeoma Oji, Osuagwu O.E., Design and Implementation of an M/M/1 Queuing Model Algorithm and its Applicability in Remote Medical Monitoring: *West African Journal of Industrial and Academic Research:* Vol.4, No. 1 (2012).

[7] Samuel F. Fomundam , Jeffrey W. Herrmann. A Survey of Queuing Theory Applications in Healthcare: ISR Technical Report 2007-24

[8] Bangalore: Residential Market Report, April 2012

Bibliographical Sketch

Sumana Sinha is currently an Assistant Professor with PESIT-BSC of Department Of Information Science and Engineering and got her M.Tech. degree in Computer Science and Engineering from Tripura University (Central University) in 2009. Her research interest is in Ad hoc networks and application of queuing model in different areas.

Kakoli Bora received the M. Tech. degree in Information Technology from Tezpur University, Assam, India in the year 2003. She is currently an Assistant Professor with PESIT Bangalore South Campus, Department of Information Science and Engineering, Bangalore, India. Her research interest includes Algorithm Analysis and Design, Data Mining, Big data Analysis and Astroinformatics.