Analysis of Waiting lines and queuing system in Banks in India during demonetization: A Case Study

Sarbjit Singh
Department of Mathematics
Guru Nanak Dev University College, NarotJaimal Singh

Abstract

Queuing theory has been fairly a successful tool in the performance analysis of waiting lines.Waiting lines and service efficiency are the important elements for any system.This paper deals with the analysis of waiting lines and queuing system in banks in India during demonetization.Long queues and snaking queues were common sight of banks in India during demonetization.The data for the study was collected by observation in which number of customers arriving at the service window was recorded. Data was collected for a period of six working days from 10.00A.M to 4.00P.M of State Bank of India, NarotJaimal Singh. Data were fitted into the model and the results were computed. Theanalysis of the results shows that the numbers of existing service windows are not adequate for the customer's service. In order to serve the customer better and reduce the waiting lines in the system, the number of service windows should be increased.

KEYWORDS: Arrival Rate, Service Rate, Service Unit, Servers, Performance measures.

1. Introduction

Queues or waiting lines are very common in everyday lifewhereby certain business situations require customers to waiting line for a service, namely: - telephone exchange, at a bank, in public transportation or in a traffic jam, in a supermarket, at a petrol station, atcomputer systems, waiting to use an ATM machine, andpaying for groceries at the supermarket [1].Studying how these lines form and how to manage them iscalled Queuing theory. More generally, queuing theory isconcerned with the mathematical modeling and analysis ofsystems that provide service to random demands which dealswith one of the most unpleasant experiences of life, waiting [2].A queuing problem arises when the current service rate offacility falls short of the current service rate of customers.

Delays and queuing problems are most common features ofour daily-life situations. Queuing theory was born in the early1900s with the work of A. K. Erlang of the CopenhagenTelephone Company, who derived several important formulasfor teletraffic engineering that today bear his name [2]. Erlangwas the first who treated congestion problems caused bytelephone calls where the company requested him to work onthe holding times in a telephone switch. He identified that thenumber of telephone conversations and telephone holdingtime fit into Poisson distribution and exponentiallydistributed [3]. This was the beginning of the study of queuingtheory.

On November 8, 2016, government of India announced demonetization of all Rs.500 and Rs.1000 bank notes. The government took this step to curtail the shadow of economy and crack down the use of illicit and counterfeit cash to curb black money. As a result, long queues formed inside as well as outside the bank premises. During this period, withdrawal from the banks was limited. People waited for hours to get cash needed to meet their daily expenses. The objective of this paper to analyze the existing number of service windowsto meet customer`s need for cash using multiserver queuing models.

2. Queuing process

The process in queuing system is the customers arriving for service, waiting for service if it is not immediate, and leaving the system once they are served [4] [5]. Typical measures of system performance are server utilization, length of waiting lines, and delays of customers, for relatively simple systems, compute mathematically, for realistic models of complex systems, simulation is usually required [6].

Key elements of queuing systems are

2.1.1 Customer: refers to anything that arrives at a facility and requires service.

2.1.2 Server: refers to any resource that provides the requested service,

2.1.3 System Capacity: a limit on the number of customers that may bein the waiting line or system.

a) **Limited capacity**, e.g., an automatic car wash only has room for 10 carsto wait in line to enter the mechanism.
b) **Unlimited capacity**, e.g., concert ticket sales with no limit on thenumber of people allowed waiting to purchase tickets.

2.1.4 Calling population: the population of potential customers may be assumed to be finite or infinite.

a) **Finite population model**: if arrival rate depends on the number of customers being served and waiting.
b) **Infinite population model:** if arrival rate is not affected by the number of customers being served and waiting.

2.1.5 Random arrivals: inter-arrival times usually characterized by aprobability distribution.Most important model: Poisson arrival process (with rate λ), where$A_n$represents the inter-arrival time between customer (n-1) and customern, and is exponentially distributed (with mean$1/_{\lambda}$)

2.1.6 Scheduled arrivals: inter-arrival times can be constant or constant plusor minus a small random amount to represent early or late arrivals.

2.1.7 Queue behavior: the actions of customers while in a queue waitingfor service to begin, for example:

- **Balk:** leave when they see that the line is too long,
- **Renege:** leave after being in the line when it's moving too slowly,
- **Jockey:** move from one line to a shorter line.

2.1.8 Queue discipline: The logical ordering of customers in a queue thatdetermines which customer is chosen for service when a serverbecomes free [7]. Some common service disciplines are

First-in first-out (FIFO)

Last-in first-out (LIFO)

Service in random order (SIRO)

Shortest processing time first (SPT)

Service according to priority (PR)

2.1.9 Service times and service mechanism: Service times of successive arrivals are denoted by S_1, S_2, S_3.......... These service times may be constant or random. The sequence $\langle S_1, S_2, S_3 .... \rangle$ is usually characterized as a sequence of independent and identically distributed random variables [8].

3. Queuing Notation

A notation system for parallel server queues: A/B/c/N/K, (due toKendall) [8], where

A represents the inter-arrival-time distribution,

B represents the service-time distribution,

c represents the number of parallel servers,

N represents the system capacity,

K represents the size of the calling population.

Primary performance measures of queuing systems are

P_n : Steady-state probability of having n customers in system,

$P_n(t)$: Probability of n customers in system at time t,

λ: arrival rate,

λ_e : Effective arrival rate,

μ: service rate of one server,

ρ : Server utilization,

A_n : Inter-arrival time between customer n-1 and n,

S_n : Service time of the nth arriving customer,

W_n: Total time spent in system by the nth arriving customer,

W_n^Q: Total time spent in the waiting line by customer n,

L(t): the number of customers in system at time t,

$L_Q(t)$: The number of customers in queue at time t,

L: long-run time-average number of customers in system,

L_Q: Long-run time-average number of customers in queue,

w: long-run average time spent in system per customer,

w_Q: Long-run average time spent in queue per customer.

4. Queuing Models

Queuing models provide the analyst with a powerful tool fordesigning and evaluating the performance of queuing systems.Model as an idealized representation of the real life situation; in order to keep the model as simple as possiblehowever, some assumptions need to be made [9].

Assumptions

a. Single channel queue.
b. There is an infinite population from which customers originate.
c. Poisson arrival (Random arrivals).
d. Exponential distribution of service time.
e. Arrival in group at the same time (i.e. bulk arrival) is treated as single arrival.
f. The queue discipline is First Come First Served (FCFS).

Although several queuing models abound; designed to serve different purposes [5], highlighted the following:

4.1 The M/M/c/∞; there are c servers to serve from a single line customer, if the arrival is less than or equals to c server every customer is being attended to; if z arrival is greater than the c servers, then z-c customers are waiting in the line.

The service utilization for c servers is given by $\rho = \frac{\lambda}{c\mu}$ ……. (1)

The average number in the line is $L_q = \frac{P_0 \left(\frac{\lambda}{\mu}\right)^c \rho}{c!(1-\rho)^2}$ …….. (2)

where $P_0 = \left[\sum_{m=0}^{c-1} \frac{(c\rho)^m}{m!} + \frac{(c\rho)^c}{c!(1-\rho)} \right]^{-1}$ ……. (3)

P_0 denote the probability that there are 0 customers in the system.

The expected number of customers in the system is

$$L_s = L_q + \frac{\lambda}{\mu} \quad \text{..... (4)}$$

Expected number of customers waiting to be served at any‘t’ is

$$L_w = \frac{c\mu}{c\mu - \lambda} \quad \text{........ (5)}$$

The average waiting time of an arrival is

$$W_q = \frac{L_q}{\lambda} \quad \text{...... (6)}$$

Average time an arrival spends in the system is

$$W_s = \frac{L_s}{\lambda} \quad \text{..... (7)}$$

4.2 M/M/1 Systems In M/M/1 (∞/FCFS) queuing system, the arrival and service time both has an exponential distribution, with parameters λ and μ respectively, with one server, queue discipline FCFSand the population size is infinite. The expected inter-arrival time and the expected time to serve one customer are (1/ λ) and (1/μ) respectively. An M/M/1 system is a Poisson birth-death process.

5. Research Method Used

Thequantitative research method was used in this study. Data was collected for a period of six working days. Data were fitted into the model and the results were computed. This model developed was used to predict the required number of servers.

Table I: Queuing system analysis of the servers for six working days of the week

		Server 1	Server 2	Server 3	Server 4
DAY 1 Monday	**Arrival Rate**	159	99	126	143
	Service Rate	50	32	39	47
DAY 2 Tuesday	**Arrival Rate**	112	136	150	107
	Service Rate	33	44	49	35
DAY 3 Wednesday	**Arrival Rate**	98	97	103	99
	Service Rate	32	29	35	40
DAY 4 Friday	**Arrival Rate**	109	149	87	133
	Service Rate	35	47	27	50
DAY 5 Saturday	**Arrival Rate**	116	123	114	129
	Service Rate	39	42	36	40
DAY 6 Sunday	**Arrival Rate**	89	143	135	94
	Service Rate	24	45	43	31
Total	**Arrival Rate**	683	747	715	705
	Service Rate	213	239	229	243

Average System Utilization	$\frac{683}{213} = 3.2065$	$\frac{747}{239} = 3.1255$	$\frac{715}{229} = 3.1222$	$\frac{705}{243} = 2.9012$

Table II: Queuing system analysis of the servers for six working days of the week

		Server 1	Server 2	Server 3	Server 4
DAY 1 Monday	**Average Arrival Rate**	26.5	16.5	21	23.83
	Average Service Rate	8.33	5.33	6.5	7.83
DAY 2 Tuesday	**Average Arrival Rate**	18.66	22.66	25	17.83
	Average Service Rate	5.5	7.33	8.16	5.83
DAY 3 Wednesday	**Average Arrival Rate**	16.33	16.16	17.16	16.5
	Average Service Rate	5.33	4.83	5.83	6.66
DAY 4 Friday	**Average Arrival Rate**	18.16	24.83	14.5	22.16
	Average Service Rate	5.83	7.83	4.5	8.33
DAY 5 Saturday	**Average Arrival Rate**	19.33	20.5	19	21.5
	Average Service Rate	6.5	7	6	6.66
DAY 6 Sunday	**Average Arrival Rate**	14.83	23.83	22.5	15.66
	Average Service Rate	4	7.5	7.16	5.16

Table III

	Server 1	Server 2	Server 3	Server 4
DAY 1 (Monday)	3.1812	3.0956	3.2307	3.0434
DAY 2 (Tuesday)	3.3927	3.0914	3.0637	3.0583
DAY 3 (Wednesday)	3.0637	3.3457	3.0463	2.4774
DAY 4 (Friday)	3.1149	3.1711	3.2222	2.6602
DAY 5 (Saturday)	2.9738	2.9285	3.1666	3.2282
DAY 6 (Sunday)	3.7075	3.1773	3.1424	3.0348

Table IV

	Server 1	**Server 2**	**Server 3**	**Server 4**	
Customer Arrival Rate (λ_i)	18.9683	20.7467	19.86	19.58	Average Arrival Rate λ = 19.7888
Customer Service Rate (μ_i)	5.915	6.6366	6.3583	6.745	Average Service Rate μ = 6.4137
Average No. of Customers served $\left(\frac{\lambda_i}{\mu_i}\right)$	3.2068	3.1261	3.1234	2.9028	$\frac{\lambda}{\mu}$ = 3.0853

Table V: Results of Performance measures of three server analysis

c	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
L_Q	-4.5649	-5.3213	-38.898	1.8414	0.4155	0.1170	0.0339	0.0095	0.0025
P_o	-2.0854	-0.2134	-0.0061	0.0331	0.0424	0.0448	0.0455	0.0457	0.0457
L_S	-1.4795	-2.2359	-35.044	4.9268	3.5009	3.2024	3.1192	3.0949	3.0879
W_Q	-0.2307	-0.2689	-1.9248	0.0931	0.0210	0.0059	0.0017	0.0005	0.0001
W_S	-0.0748	-0.1130	-1.7689	0.2490	0.1769	0.1618	0.1576	0.1564	0.1560

6. DISCUSSION OF RESULTS

From the above table, it has been observed that existing number of service windows required to serve the customers is not sufficient. The suitable number of servers that can serve the customers as and at when necessary without waiting forlong before customers are been served at the actual time should be more than four. This increase in number of service windows reduces thewaiting time.

7. CONCLUSIONS

The resultanalysis of the above queuing system shows the need to increase the number of the service windows. The increase in thenumber of service windows will reduce the time that customers have to wait in line before been served.

REFERENCES

[1] William J. Stewart.,“Probability, Markov Chains, Queues and Simulation”, ISBN 978-0-691-14062-9.

[2] Sundarapandian, V. (2009). "7. Queuing Theory".Probability, Statistics and Queuing Theory.PHI Learning.ISBN 8120338448.

[3] Lawrence W. Dowdy, Virgilio A.F. Almeida, Daniel A. Menasce (2004). "Performance by Design: Computer Capacity Planning By Example". p. 480

[4] Cooper, R.P., ‘‘Introduction to Queuing Theory’’, 1981 Elesevier, New York.

[5] Bose S.J., Chapter 1 - An Introduction to Queuing Systems, Kluwer/Plenum Publishers, 2002.

[6] Sheldon, M. Ross, “ Introduction to Probability Models”, Academic Press, New York.

[7] Van Dijk, N. M(1993)., “On the arrival theorem for Communication Networks” Computer Networks & ISDN 25(10); 1135-2013

[8] Kendall, D. G. (1953). "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain". The Annals of Mathematical Statistics 24 (3): 338. doi:10.1214/aoms/1177728975. JSTOR 2236285.edit

[9] B. F. Adam, I. J. Boxma, and J. A. C. Resing, Queuing models with multiple waiting lines queuing systems, 37, 2001, 65-9