Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-8, Issue-5, May 2026
DOI: 10.53469/jrse.2026.08(05).07

A Queuing Theory Approach to Optimizing Service Efficiency in Restaurants

Raghvendra Alok Gupta, Sorabh Singal

Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

Abstract: ***In the restaurant business, effective queue management is essential for increasing client happiness and boosting profits. This study describes a queuing model created for a restaurant to enhance the system and reduce client wait times. The model simulates and analyses the queuing dynamics in the restaurant environment by taking into account a variety of variables, including arrival rates, service times, and customer behaviour. This paper makes an effort to show that, when considered in a real-world setting, queuing theory conforms to the model. We obtained the information from Venus, a dining establishment in Bhubaneswar, Odisha. By the application of Little's law and the M/M/1 queuing model, we have produced significant metrics such as arrival rate, service rate, utilization rate, waiting time in the queue, and the rate of prospective clients departing.***

Keywords: Queue, arrival rate, service rate, Poisson distribution, exponential distribution

1. Introduction

The study of modelling of queues or lines of waiting is the primary objective of the mathematical area known as queuing theory. It has several uses, including in telecommunications, transportation, healthcare, customer service, and other areas. Agner Krarup Erlang, who published an important work on telephone traffic congestion in 1909, may be able to be utilized to trace the development of queuing theory. Erlang's contributions gave queuing theory a strong foundation by introducing fundamental presumptions and analytical methods that are still applied in a variety of computer and communications systems today. Other pioneers have made important contributions to queuing theory in addition to Erlang's work. There are queueing mechanisms everywhere in daily life, not only in human interactions. Queues can be seen, for instance, when a machine is processing a work, a plane is waiting to land at an airport, or a car is halted at a traffic light. The objective is to reduce the adverse effects of waiting to practical levels, even though complete eradication of waiting is frequently impractical without incurring excessive expenditures. By optimizing variables including the quantity of service channels, workforce levels and system capacity to meet desired performance goals, they assist in determining the most effective allocation of scarce resources. People are frequently depending on restaurants for their meals, whether it be for lunch or dinner in today's fast –paced world. Numerous elements that affect a restaurant's quality have a significant impact on its performance. The factors to consider include the food's flavour, the restaurant's cleanliness, the level of service received, the diversity of menu options, and the ambiance in general. Restaurants can draw a lot of people if these factors are managed carefully. But even after being able to draw customers, another important factor—the time consumers must wait before being served—comes into play. Studying how people, things, or information flow in a line or queue is a subject known as "queuing theory." The duration of the line wait, the expected queue length, the typical amount of time spent utilizing the system, and the expected number of clients present at any given time are just a few of the assessments that may be obtained using queuing theory.

Amin et al.(2014) used queuing theory to arrive at the best possible solution for real-time situations, Dhari and Rahman (2013) have studied queuing model for bank ATM. Nair et.al (2021) proposed an approach to apply the mathematical queuing model to shorten the wait period in the railway ticket window. Abdel-Aal (2020) has studied a queuing model in parking entry. According to Yadav et.al. (2022),a tertiary care hospital's waiting and service cost were examined using a multi-server queuing model. Yaduvanshi et al. (2019) have applied queuing theory in hospital operations to optimize wating time. Khalili and Khah (2020) have applied a new mathematical model in hotel capacity. Akintunde et al. (2023) have applied Queuing theory for congestion problem during Covid -19. Chakravarty et al. (2020) have explained a backup server during vacation period. Lakshmi and Iyer (2013) have applied queuing theory in health care.

The purpose of the study is to use queuing theory to analyse the queue length at the "Venus Inn Restaurant" in Bhubaneswar, Odisha. The goal is to show how the M/M/1 model, a particular queuing model, can be used in a real-world scenario. By employing queuing theory, this study seeks to provide insights into managing waiting times, optimizing resource allocation, and enhancing customer satisfaction in restaurant operations.

2. Methodology

A queuing system is a mathematical model used to analyze and simulate the action of queue length. It involves the study of entities, such as customers' tasks, arrive at a service facility, wait in line and are eventually served or processed. Queuing measures have six essential characteristics such as

a) Distribution of arrival time - There are three different models for inter arrival times: a deterministic distribution, a Poisson distribution, a generic distribution. Inter arrival times vary independently and without recollection, which are the features of Poisson distribution.
b) Pattern of the servers - Most of this is based on the distribution of service time. Constant, exponential, hypo exponential, hyper exponential or general temporal distributions are possible for the service. Inter arrival time does not impact the service time.

c) Queuing Discipline - The arrangement of the queue is represented by the queuing discipline. There are numerous methods for serving clients in order, including SIRO (Selection in random order), FCFS (First come First service), LCFS (Last come First service) which states that the last person to arrive would be served first.
d) System Capacity - It is also called maximum queue size. The queue mechanism can accommodate a maximum of clients.
e) Queue Length -The number of customers in a queue or present in the system is known as the queue length. A system's queue can be modelled as either finite or infinite.
f) Number of servers: -The computation for waiting line depends on whether the queue is supplied by one server or by multiple servers. In a one server queuing system, customers are served sequentially from the front of the line, while in a multiple server queuing system, numerous service facilities are operating simultaneously to provide one or more services that are identical.

2.1 Little's Theorem:

According to little's theorem the average number of things in a system is exactly related to the average rate at which items enter and leave the system as well as the average amount of time that an item is held up in the system.

$$L \alpha \lambda W$$

i.e. we can state that the average number of items within a system is equal to the product of average rate at which the new items arrive and the average time that they spend in the system being held.

Where,
L= Average number of items within a system
λ = Average rate of arrival of items into and out of the system.
W= holding up time of the items in the system.

By applying queuing theory, we can make the queue length more attractive. In this study data from a restaurant is gathered, and both *M/M/1* model and Little's law is used to analyze it.

Queuing models might be totally indicated in the symbol form(a/b/c) :(d/e), where,
a= Inter arrival time distribution
b= service time distribution
c= Number of service stations
d= Maximum number of jobs that can be there in the system
e= Discipline in the queue

We refer to such notation as Kendal's Notation.

If we state the following letter as
M= Poisson arrival or departure distribution
E_k= Erlangian or Gamma inter arrival for service time distribution
GI= General input distribution
G= General Service time distribution

Then $(M/E_k/1) :(\infty/FIFO)$ describes a queuing system in which arrivals follow a Poisson distribution, and the service time is governed by an Erlangen single server with an infinite capacity.

Therefore, a Poisson distribution-also referred to as a poisson process-will determine the number of customers that appear and are served per unit of time. We can calculate the Poisson probabilities as below.

$$f(x) = (e^{-\lambda} \lambda^x)/x! \quad \forall \lambda>=0, x>=0$$

2.2 Notations Used for Queuing Theory

In queuing theory, the following notations are commonly used:
n = Represents the no of patrons or units in the system at a given time.
$P_n(t)$ = Refers to the transient state probability.
p_n = Represents the steady state probability.
λ_n = Represents the average rate at which customers arrive for n customers.
μ_n = Represents the average rate at which customers are served for n customers.
λ = It is the average rate at which customers arrive.
μ = It is the average rate at which customers are served in the long run.
ρ = Denotes the traffic intensity or utilization factor, which is defined as the ratio of the arrival rate (λ) to the service rate (μ). It represents $\rho = \lambda/\mu$.

These notations are commonly used in queueing theory to analyze and model the behavior of queuing systems. They help in understanding the performance and characteristics of queues, such as waiting times, queue lengths, and system efficiency.

2.3 Transient & Steady State

In transient state the behavior of the system is dependent on time meanwhile, in a steady state, the behavior of the system is independent of time. This means that the probabilities of different states of the system, denoted by P_n, remain constant over time. In this case, as time approaches infinity ($t \rightarrow \infty$), the probability $P_n(t)$ of having n customers in the system at time t also approaches the steady-state probability P_n. So, in mathematical terms, we can express this as

$$\lim_{t \to \infty} P_n(t) = Pn$$

This equation states that as time goes to infinity, the probability of having n customers in the system converges to the steady-state probability of having n customers.

Now, let's consider the derivative of $P_n(t)$ with respect to time (t). The rate of change of $P_n(t)$ with respect to time represents how the probability distribution of the system is evolving over time. In a steady state, since the behavior of the system is independent of time, the derivative of $P_n(t)$ with respect to time approaches zero as time goes to infinity.

Mathematically, we can express this as:

$$\frac{dP_n(t)}{dt} = \frac{dP_n}{dt}$$

$$\lim_{t\to\infty} P_n(t)=0$$

This equation indicates that as time approaches infinity, the rate of change of $P_n(t)$ with respect to time approaches zero.

Therefore, in a steady state, both the probability $P_n(t)$ and its rate of time change relative to time $\frac{dP_n(t)}{dt}$, approach the steady-state probability P_n and zero, respectively, as infinity time passes.

2.4 Utilization Factor

The traffic intensity, also known as the utilization factor, is an important measure in queuing theory. It shows the proportion between a system's average customer arrival rate (λ) and the average service rate (μ). The formula for traffic intensity (ρ) is indeed expressed by:

$$\rho = \lambda / \mu$$

In which
λ = the average rate of customers arrival
μ = the average rate of service.

The unit of traffic intensity is Erlang, named after the Danish mathematician A.K. Erlang. It is a dimensionless unit that represents the traffic load on a system. The traffic intensity value can range from 0 to infinity. A traffic intensity of less than 1 ($\rho < 1$) indicates that the system is not fully loaded, meaning the average arrival rate is lower than the average service rate. In this case, the system can handle the incoming customer traffic without significant delays. However, if the traffic intensity exceeds 1 ($\rho > 1$), it implies that the average arrival rate is higher than the average service rate, indicating an overloaded system. In such cases, the queue will tend to grow indefinitely, causing delays and potentially leading to congestion. Analyzing and managing traffic intensity is crucial in understanding system performance, determining resource requirements, and ensuring efficient operation in scenarios such as telecommunications networks, computer systems, transportation networks, and more.

3. Venus Queuing Theory

3.1 Daily customers count at the restaurant and all other data were obtained through the interview with the Venus restaurant manager. The restaurant has recorded the data as part of its end-of-day procedure. The data leads us to the conclusion that M/M/1, which depicts an exponential distribution (Poisson process) for arrival and service time, is the queuing model that best represents Venus's functioning. According to the restaurant's system, there is just one server. In actual waiting queue, there are a number of waitresses, but only one chef serves every customer, in our sight. To examine the Venus M/M/1 queuing model we can use some variables given below:
λ = the mean
customers
arrival rate.
μ = the mean
service rate.
$\rho = \lambda/\mu$ = Utilization factor.

i) The average no. of customers in the system

$$L_s=\frac{\rho}{1-\rho}=\frac{\lambda}{\mu-\lambda}$$

ii) The average no. of customers in the queue

$$L_q = \rho L_s=\frac{\lambda^2}{\mu(\mu-\lambda)}$$

iii) The probability of *n* customers in the system

$$P_n= (1-\rho)\ \rho^n$$

iv) The probability of zero clients in the system

$$P_0= 1-\rho$$

v) The average time spent in the system

$$w_s=\frac{1}{\mu-\lambda}$$

vi) The average time spent in the line

$$w_q=\frac{\lambda}{\mu\,(\mu-\lambda)}$$

3.2. Modelling Arrival Rate & Service Rate

To model the arrival rate and service rate of a queuing model for a restaurant, you need to gather data and make certain assumptions. Here are the steps we have followed.

1) Define the time interval: We have determined the time interval that we want to analyse, such as per hour, per day, or per week. This helped in precisely determine the rate of arrival and rate of service accurately.
2) Gather arrival data: We Collected data on the number of customers arriving at the restaurant during the chosen time interval and Kept track of the number of arrivals for each time unit (e.g., customers per hour).
3) Calculate arrival rate: To determine the arrival rate, we divided the total number of arrivals by the length of the time interval. For example, if you observed 200 customers over a 4-hour period, the arrival rate would be 200/4 = 50 customers per hour.
4) Gather service time data: We have to collect data on the time taken to serve each customer. This data can also be obtained from your POS system or by timing individual customer transactions.
5) Calculate service rate: To determine the service rate, we have calculated the average service time for each customer. Again, we can divide the total service time by the total number of customers served. For example, if the total service time for 200 customers is 400 hours, the average service time would be 400/200 = 2 hours per customer. In this case, the service rate would be the inverse of the average service time, i.e., 1/2 = 0.5 customers per hour.
6) Validate assumptions: We ensured that the data we collected and the calculated arrival rate and service rate were representative of the restaurant's normal operations because considering factors like seasonality, special events, or any other conditions may affect customer arrivals and service times. Adjustments may be necessary to account for variations in different time periods.
7) Refine the model: As we gathered more data over time, we could refine our model by updating the arrival rate and service rate estimates. This helped for improving the accuracy of our queuing model and aiding in making better operational decisions for the restaurant.

It's worth noting that queuing models can become more complex depending on the specific features of the system, such as multiple service channels, customer priorities, or different customer classes. However, the steps outlined above provide a basic approach to estimate the arrival rate and service rate for a simple queuing model of a restaurant.

4. Result & Discussion

The provided table displays the restaurant's one month data for the establishment:

Table 1: Daily Customer counts of one month

	Count of Monthly Customer						
Day	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
1st Week	920	420	429	510	400	490	440
2nd Week	1010	357	485	525	390	555	1092
3rd Week	985	475	395	670	495	632	890
4th Week	1150	440	525	705	500	856	995

From the above table we can found that on the weekends the numbers of people are double of the people of the other weekdays. The weekends Lunch time are the most crowed time of the restaurant. So, we will proceed our study during this time.

4.1 Calculation

The restaurant at lunch time was conducted by our team. On average 220 customers visit the restaurant in 3 hours. From the above statement we can derive the arrival rate. In the restaurant, each patron stays for an average of 40 minutes, as we have seen. On average 14 people are in the queue length (L_q) and waiting time is around 10 minutes. Now we can observe that actual waiting time doesn't affect by much when actual compared to theoretical waiting time as shown below:

$$\lambda = 1.22 \text{ customer per minute}$$

$$w_q = \frac{L_q}{\lambda} = \frac{14\ customers}{1.22} = 11.48 \text{ minutes}$$

Now we calculate the average numbers of customers in the restaurant

$$L = 1.22*40 \text{ min} = 49 \text{ customers app.}$$

By using this we can calculate the utilization rate and the service rate

$$\mu = \frac{\lambda(1+L)}{L} = \frac{1.22(1+49)}{49} = 1.24 \text{ customer per minute}$$

Hence,

$$\rho = \frac{\lambda}{\mu} = \frac{1.22}{1.24} = 0.98$$

Due to the high utilization rate at lunch, it is extremely unlikely that there will be no customers at all, as shown below:

$$P_0 = 1-\rho = 1-0.98 = 0.02$$

Now the probability of n customers can be calculated as follows:

$$P_n = (1-\rho)\,\rho^n$$
$$= (1-.98)(0.98)^n = (0.02)(0.98^n)$$

When the potential customers see more than 10 people are waiting for the dining they will start refuse. The maximum tolerates waiting length of the potential customer is 25 customers whereas the restaurant can accommodate 70 people at once when fully occupied.

Then we have to compute the probability of 17 people in the queue as the probability of 75 people in the system (36 in the restaurant and 17 or more queuing) as follows:

$$P_{36\text{-}72} = \sum_{n=80}^{95} P_n = 6.05\%$$

Table 2: Implementation portion of *M/M/1* model

Attributes	Symbol	Value
Total number of customers	n	220
Average Service rate	μ	1.24
Customers arrival rate	λ	1.22
Average waiting time in a queue	w_q	11.48 min
Average waiting time in system	w_s	50 min
Average no of clients in queue	L_q	14
Average no of clients in system	L_s	83

4.2 Assessment

a) The claim that utilization directly relates to the average number of consumers is true. As the utilization increases, it indicates that a larger proportion of the restaurant's capacity is being utilized, which implies that more customers are present on average. This relationship is commonly observed in queuing theory and operations management.

b) It is correct to state that the findings of this research's conclusion can be used as a conductor to assess the existing system and enhance the one that comes after. By studying the mean waiting customers count in a queue and the leaving customers count each day, the restaurant can gain insights into the system's performance and make informed decisions to enhance efficiency and customer satisfaction in the future.

c) It makes sense to plan for a large number of clients and set a target profit based on daily estimate. By understanding the expected influx and out flux of customers, the restaurant can establish revenue goals and align operational strategies accordingly.

4.3 Profits

Here are the profits associated with the research described:

a) Increased Quality of Service (QOS): The research helps Venus anticipate the number of customers in the queue. By having this information, the restaurant can better prepare and allocate resources accordingly, ensuring a smoother and more efficient service. This leads to an improved QOS as customers experience shorter wait times and prompt service.

b) Analysis and system improvement: The findings from this research can serve as a reference point for analysing the current system at Venus. By understanding the estimated waiting customers count in the queue and the leaving customers count each day, the restaurant can identify areas for improvement and make adjustments to their operations. This leads to an optimized system that can better handle customer demand.

c) Target profit setting: With the ability to anticipate the influx and outflow of customers on a daily basis, Venus is able to define a profit objective that is consistent with the anticipated number of clients. By considering the estimated number of customers and their behaviour, the restaurant can strategize and plan their operations to achieve the desired profit margin.
d) Applicability to future research and development: The formulas developed during this research can serve as a foundation for further studies and investigations. Other researchers and practitioners in the field can leverage these formulas to build upon and develop more complex theories or models related to restaurant queue management or similar areas of research.

5. Conclusion

The queuing model of a restaurant refers to the analysis and optimization of the queuing system within the restaurant, with the aim of improving customer satisfaction and operational efficiency. It involves studying customer arrival patterns, service times, and techniques for handling the queue that reduce wait times and optimise the use of restaurant resources.

Here is a conclusion of the queuing model of a restaurant.

1) Customer Arrival Patterns: The queuing model considers the different patterns of customer arrivals, such as random arrivals or time-dependent arrivals. By understanding these patterns, the restaurant can allocate appropriate staffing levels and resources to handle varying customer demand throughout the day.
2) Service Times: The model analyzes the duration of service, including order taking, meal preparation and delivery. By evaluating service times, the restaurant can identify bottlenecks and inefficiencies in its operations and make improvements to enhance overall service speed.
3) Queue Management: Efficient queue management is crucial to ensure smooth customer flow and minimize waiting times. The model examines strategies like single queue or multiple queues, self- service options, and prioritization techniques (e.g., first-come-first- served or priority for certain customers). It helps the restaurant determine the most effective approach based on its specific circumstances.

By applying a queuing model to a restaurant, management can gain valuable insights into customer behavior, service times, and resource utilization. This analysis enables the restaurant to use methods and make decisions that enhance the overall dining experience and operational efficiency.

In conclusion, this study has examined the utilization of queuing theory to analyze the operations of Venus Restaurant. The focus was on two key variables: customer arrival rate and service rate, which were found to be 0.83 customers per minute and 0.84 customers per minute, respectively. The results indicate that because there are fewer consumers throughout the week that on the weekends, the arrival rate is lower and the service rate is greater. Noting that several of the data utilized in this study were based on assumptions or approximations, which could have introduced inaccuracies in the results. Therefore, further research, particularly the development of a simulation model, is recommended to validate the findings and account for more complex factors that influence the restaurant's operations.

This study's goal was to enhance Venus restaurant's customer management techniques. By utilizing queuing theory and potentially implementing a simulation model, the restaurant can gain insights into optimizing its operations to enhance customer satisfaction and minimize waiting times. This research paper serves as a starting point for future studies that can develop deeper into the intricacies of the restaurant's operations and provide more accurate and comprehensive recommendations.

References

[1] Akintunde, A.K., Adamu. and Afolabi, S.A. (2023) 'Congestion Problem during Covid -19 in the University College Hospital, Ibadan, Oyo State, Nigeria: An Application of Queuing Theory', *International Journal of Mathematics and Statistics Studies*, Vol. 11, No. 1, pp.61–66.
[2] Amin, A., Mehta, P., Sahay, A., Kumar. and Kumar, A. (2014) 'Optimal Solution of real time problems using Queuing Theory', *International journal of Engineering and Innovative Technology*, Vol. 3, No. 10.
[3] Chakravarthy, S.R., Shruti. and Kulshrestha, R. (2020)'A queuing model with server breakdowns, repairs, vacations and backup server', *Operation s Research Perspectives*, Vol. 7, DOI: 10.1016/j.orp.2019.100131.
[4] Dhar, S.K and Rahman, T. (2013)' Case Study for Bank ATM Queuing Model', *IOSR Journal of Mathematics*, Vol.7, issue.1, pp 01-05.
[5] Khalili, S. and Khah, M. (2020) 'A new queuing – based mathematical model for hotel capacity planning: a genetic algorithm solution', *Journal of Applied Research on Industrial Engineering, Vol.* 7, No. 3, pp.203–220, DOI: 10.22105/jarie.2020.244708.1187.
[6] Lakshmi. and Iyer, S.A. (2013)'Application of queuing theory in health care: A literature review', *Operation Research for Health care*, Volume 2, No. 1-2, pp.25-39, DOI: 10.1016/j.orhc.2013.03.002.
[7] Nair, A.M, Sreelatha, K.S, Ushakumari, P.V. (2021),'Application of Queuing Theory to a Railway ticket window', IEEEXplore.
[8] Thi, M.N and Manh, C.D, (2021), The Application of Queuing Theory in the Parking Lot: a literature review, *proceeding of the international conference on emerging challenges, Business Transformation and circular economy*, Vol.196.
[9] Yaduvanshi, D., Sharma, A. and More, P.V. (2019) 'Application of Queuing Theory to Optimize Waiting Time ion Hospital Operations', *Operations and Supply Chain Management*, Vol. 12, No. 3, pp.165–174.
[10] Yadav, C., Sharma, R.C. and Shankar, U. (2022)'Analysis of waiting and service cost for a multiserver queuing model in a tertiary care Hospital, International journal of Health System, Vol.6(S8), pp-5140-5148, DOI:10.53730/ijhs. v6nS8.133393.