DELHI TECHNOLOGICAL UNIVERSITY

OPERATIONS RESEARCH PROJECT REPORT

QUEUING THEORY APPLICATION IN BANK SERVICE OPTIMISATION

SUBMITTED BY:

JAI GARG (DTU/2K18/MC/044)

JATIN PAPREJA (DTU/2K18/MC/049)

INDEX

INTRODUCTION

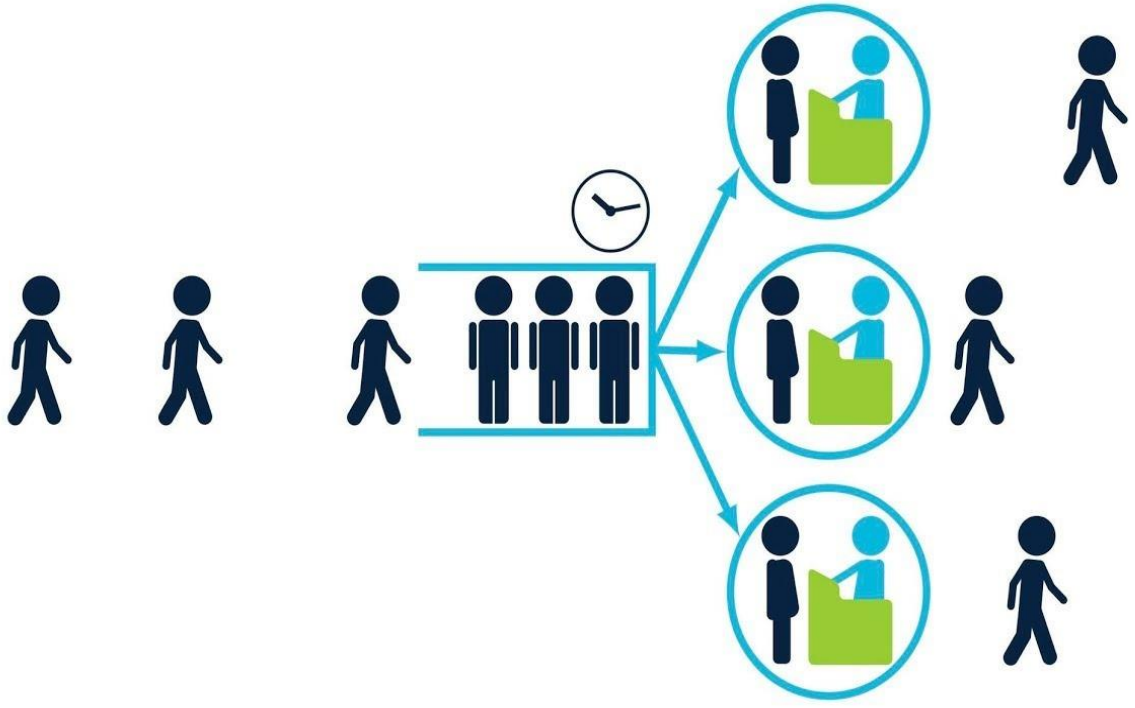
In our lives, we all have faced the situation in which we need to stand in long queues, waiting in order to get served. As more and more organisations are moving towards service based model as opposed to product based model, there is a dire need to provide quality and efficient services. **Customer satisfaction** is of great importance for any **service-based organisation**. **Queuing Theory** can help to provide efficient solutions to such scenarios. **Queues** are a fair and essential way of dealing with the flow of customers when there are limited resources. Negative outcomes arise if a queue process isn't established to deal with overcapacity. For example, when too many visitors navigate to a website, the website will slow and crash if it doesn't have a way to change the speed at which it processes requests or a way to queue visitors.

Queuing Theory:

Queuing theory refers to the mathematical study of the formation, function, and congestion of waiting lines, or queues.

At its core, a queuing situation involves two parts.

- Someone or something that **requests a service**—usually referred to as the **customer, job, or request**.
- Someone or something that **completes or delivers the services**—usually referred to as the **server**.



Queuing theory examines every component of waiting in line to be served, including the arrival process, service process, number of servers, number of system places, and the number of customers—which might be people, data packets, cars, etc. As a branch of operations research, queuing theory can help users make informed business decisions on how to build efficient and cost-effective workflow systems.

Queues happen when resources are limited. In fact, queues make economic sense; no queues would equate to costly overcapacity. Queuing theory helps in the design of balanced systems that serve customers quickly and efficiently but do not cost too much to be sustainable. All queuing systems are broken down into the entities queuing for an activity. At its most elementary level, queuing theory involves the analysis of arrivals at a facility, such as a bank or fast food restaurant, then the service requirements of that facility, e.g., tellers or attendants.

The **origin of queuing theory** can be traced back to the early 1900s, found in a study of the **Copenhagen telephone exchange by Agner Krarup Erlang**, a Danish engineer, statistician and, mathematician. His work led to the **Erlang theory of efficient networks and the field of telephone network analysis**. His mathematical analysis culminated in his **1920 paper "Telephone Waiting Times"**. The international unit of telephone traffic is called the **Erlang**.

Applications of Queuing Theory:

By applying queuing theory, a business can develop more efficient queuing systems, processes, pricing mechanisms, staffing solutions, and arrival management strategies to reduce customer wait times and increase the number of customers that can be served.

Queuing Theory is powerful because the ubiquity of queue situations means there are countless and diverse applications of queuing theory.

Queuing Theory has been applied to various fields such as:

- Telecommunications
- Transportation
- Logistics
- Finance
- Emergency Services
- Computing
- Industrial Engineering
- Project Management

BACKGROUND KNOWLEDGE

Notation for Queuing Theory:

Queuing theory uses the **Kendall notation** to classify the different types of queuing systems, or nodes. Queuing nodes are classified using the notation:

A/S/c/K/N/D

where:

A is the arrival process

S is the mathematical distribution of the service time

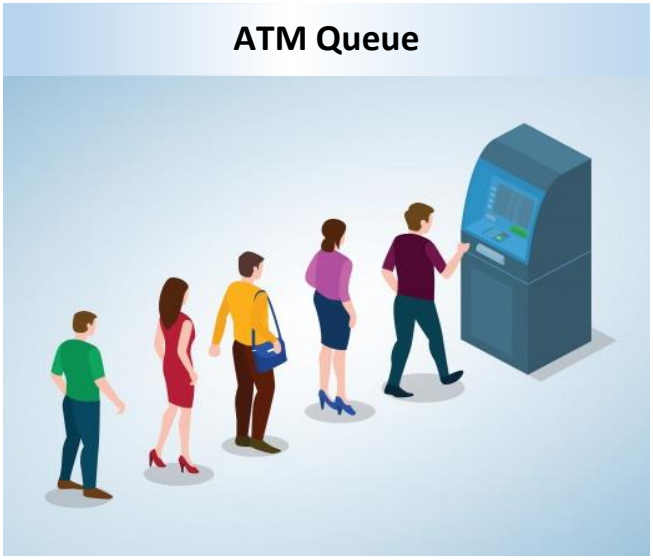
c is the number of servers

K is the capacity of the queue, omitted if unlimited

N is the number of possible customers, omitted if unlimited

D is the queuing discipline, assumed first-in-first-out if omitted

For example, in case of an ATM, it can serve one customer at a time in a first-in-first-out order with a randomly-distributed arrival process and service distribution time unlimited queue capacity and unlimited number of possible customers. Queuing theory would describe this system as a **M/M/1: ∞/∞/FCFS** queue ("**M**" here stands for Markovian, a statistical process to describe randomness).



Parameters for Queuing Theory:

In Queuing Theory, we are concerned with some important parameters that can be listed as:

- Arrival rate in the system (**λ**)
- Service rate of the servers (**μ**)
- Expected Number of Busy Servers / Server Utilization (**ρ**) = λ/μ
- Expected Number of Customers in the System (**Ls**)
- Expected Number of Customers waiting in the Queue (**Lq**)
- Expected Waiting Time in the System (**Ws**)
- Expected Waiting Time in the Queue (**Wq**)

Little's Theorem:

Little's Law states that the long-term average number of customers in a stable system L is equal to the long-term average effective arrival rate, λ, multiplied by the average time a customer spends in the system, W.

$$\mathbf{Ls = \lambda Ws}$$

$$\mathbf{Lq = \lambda Wq}$$

Although it looks intuitively easy, it is quite a remarkable result, as the relationship is "**not influenced by the arrival process distribution, the service distribution, the service order, or practically anything else**."

Single Server Queuing Model (M/M/1):

In queueing theory, a **M/M/1 queuing model** represents a system having a **single server**, where arrivals are determined by a **Poisson process** and job service times have an **exponential distribution**.

Probability of having zero customers in the system (**P_0**) = 1 – ρ

Expected Number of Customers in the System (**Ls**) = ρ/(1 – ρ)

Expected Number of Customers waiting in the Queue (**Lq**) = Ls - ρ

Expected Waiting Time in the System (**Ws**) = Ls/ λ

Expected Waiting Time in the Queue (**Wq**) = Lq/ λ

Expected Number of Busy Servers in the system ($\bar{\mathbf{c}}$) = ρ

Multiple Server Queuing Model (M/M/c):

In queueing theory, a discipline within the mathematical theory of probability, the **M/M/c queue (or Erlang–C model)** is a **multi-server queueing model**.

Probability of having zero customers in the system (**P_0**)

$$P_0 = \frac{1}{\sum_{n=0}^{c-1}\frac{\rho^n}{n!} + \frac{\rho^c}{c!(1-\frac{\rho}{c})}}$$

Expected Number of Customers waiting in the Queue (**Lq**)

$$Lq = \frac{\rho}{(c-1)!(c-\rho)^2} P_0$$

Expected Number of Customers in the System (**Ls**) = Lq + ρ

Expected Waiting Time in the System (**Ws**) = Ls/ λ

Expected Waiting Time in the Queue (**Wq**) = Lq/ λ

Expected Number of Busy Servers in the system ($\bar{\mathbf{c}}$) = ρ/c

MATLAB CODE FOR QUEUING MODEL

- **Code for M/M/1: ∞/∞/FCFS Queuing Model:**

```
function [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] =
Queuing_MM1(lamda, mu)
   rho = lamda/mu;
   p0 = 1 - rho;
   Ls = rho/(1 - rho);
   Lq = Ls - rho;
   Ws = Ls/lamda;
   Wq = Lq/lamda;
   busy_server = rho;
   Prob_queuing = 1 - p0;
```

- **Code for M/M/c: ∞/∞/FCFS Queuing Model:**

```
function [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] =
Queuing_MMc(lamda, mu, c)
   rho = lamda/mu;
   val = 0;
   for i = 0:c-1
      val = val + (rho.^i)/factorial(i);
   end
   val = val + (rho.^c)/(factorial(c)*(1 - (rho/c)));
   p0 = 1./val;
   Lq = ((rho.^(c+1))*p0)/(factorial(c-1)*((c-rho).^2));
   Ls = Lq + rho;
   Ws = Ls/lamda;
   Wq = Lq/lamda;
   busy_server = rho/c;
   val2 = 0;
   for i = 0:c-1
      val2 = val2 + ((rho.^i)*p0)./(factorial(i));
   end
   Prob_queuing = 1 - val2;
```

CASE STUDY

Indian Commercial Banks have done very well in the marketing sector. More and more people now trust banks and are happy to be associated to a particular bank for multiple purposes. Banks provide various services such as **money deposit, withdrawal, account management**, etc. To avail these services, customers visit banks in large numbers on a daily basis. A large number of customers lead to the problem of **Customer Queuing**. Customer Queuing is a problem which involves **low service rate of the bank, poor business environment** and **a number of high-quality potential customers loss**. A survey conducted shows that the service quality of bank branches including service convenience and waiting time in queues are the biggest factors affecting customer experience. Large waiting times can lead to **jockeying, reneging** as well as **unforced balking**. This problem can be analyzed by applying appropriate queuing models which can in turn help to minimize the cost in order to achieve **high level of customer experience**.

In the project, we have considered a branch of a well known bank. The customer arrival is taken as a **Poisson Distribution** with **arrival rate (λ) 40 customers per hour.** The service time is **distributed exponentially** with **mean 5 minutes per service**. We consider that there is **no forced balking** i.e. the length of the queue is infinite. In order to represent this problem in terms of queuing theory, we will use **M/M/c: ∞/∞/FCFS queuing model**. The **'M'** in queuing model represents the **Markovian property**. We are given the cost of per unit service per unit time ($\mathbf{c_s}$) as Rs 70 per hour. The waiting cost of per unit time per waiting customer ($\mathbf{c_w}$) as Rs 15 per customer. Let x represent the service level, the cost model can be expressed as:

$$\mathbf{ETC(x) = EOC(x) + EWC(x)}$$

where, **ETC** = Expected total cost per unit time

EOC = Expected cost of operating the facility per unit time ($\mathbf{c_s \mu c}$)

EWC = Expected cost of waiting per unit time ($\mathbf{c_w Ls}$)

Objectives of this study:

- To study the impact of number of queues on the various parameters of queuing model and customer experience.
- To study the impact and find the optimal number of servers based on the cost model.
- To find the optimal service rate at which the service should be provided to improve the efficiency.

RESULT & ANALYSIS

A. Optimal Number of Queues:

In this case, we take different values for **number of queues (N_q)** in order to study various parameters related to queuing theory. We take three values **(1, 2, 4)** for number of queues and analyze their impact. We are given arrival rate (**λ**) as 40 customers per hour. The service rate (**μ**) can be calculated as 12 (60/5) customers per hour. We consider number of servers as 4.

Case I: $N_q = 1$

```
>> [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] = Queuing_MMc(40, 12, 4)
Ls =
    6.6219
Lq =
    3.2886
Ws =
    0.1655
Wq =
    0.0822
busy_server =
    0.8333
Prob_queuing =
    0.6577
```

Case II: $N_q = 2$

```
>> [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] = Queuing_MMc(20, 12, 2)
Ls =
    5.4545
Lq =
    3.7879
Ws =
    0.2727
Wq =
    0.1894
busy_server =
    0.8333
Prob_queuing =
    0.7576
```

Case III: $N_q = 4$

```
>> [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] = Queuing_MM1(10, 12)
Ls =
    5.0000
Lq =
    4.1667
Ws =
    0.5000
Wq =
    0.4167
busy_server =
    0.8333
Prob_queuing =
    0.8333
```

Analysis:

N_q	λ	μ	Ls	Lq	Ws	Wq
1	40	12	6.6219	3.2886	0.1655	0.0822
2	20	12	5.4545	3.7879	0.2727	0.1894
4	10	12	5.0000	4.1667	0.5000	0.4167

In the above table, we have the values for key parameters for three different values of **number of queues (N_q)**. The data obtained shows that the number of queues has a big impact on all the parameters specifically the **waiting time in the system (Ws)** and **waiting time in the queue (Wq)**. The values of Wq shows that the waiting time in the queue increases from 0.0822 hrs (≈ 5 min) to 0.1894 hrs (≈ 11.3 min) when the number of queues increases from one to two. Further, as the number of queues is increased to four, the waiting time (Wq) increases to 0.4167 hrs (≈ 25 min). This shows that in banking services, in case of "FCFS", **one queue is better than more queues**. This happens because, in case of more than one queue, all the customers are equally divided among the servers which leads to larger value of average waiting time in the queue because of the inefficient server. Whereas, in case of one queue, the number of customers serviced by an inefficient server reduces which gives lower value of average waiting time in the queue.

B. Optimal Number of Servers:

In this case, we take different values for **number of servers (N_s)** in order to study various parameters related to queuing theory. We take three values **(4, 5, 6)** for number of servers and analyze their impact. We are given arrival rate (**λ**) as 40 customers per hour. The service rate (**μ**) can be calculated as 12 (60/5) customers per hour. We consider number of servers as 4. We are given the cost of per unit service per unit time (c_s) as Rs 70 per hour. The waiting cost of per unit time per waiting customer (c_w) as Rs 15 per customer. Using the cost model and the cost per percent decrease in probability of queuing, we will find the optimal number of servers.

Case I: $N_s = 4$

```
>> [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] = Queuing_MM1(10, 12)
Ls =
    5.0000
Lq =
    4.1667
Ws =
    0.5000
Wq =
    0.4167
busy_server =
    0.8333
Prob_queuing =
    0.8333
```

Case II: $N_s = 5$

```
>> [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] = Queuing_MMc(40, 12, 5)
Ls =
    3.9867
Lq =
    0.6533
Ws =
    0.0997
Wq =
    0.0163
busy_server =
    0.6667
Prob_queuing =
    0.3267
```

Case III: $N_s = 6$

```
>> [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] = Queuing_MMc(40, 12, 6)
Ls =
    3.5186
Lq =
    0.1853
Ws =
    0.0880
Wq =
    0.0046
busy_server =
    0.5556
Prob_queuing =
    0.1482
```

Analysis:

N_s	Ls	Lq	Ws	Wq	$\bar{c}$	P[q]
4	6.6219	3.2886	0.1655	0.0822	0.8333	0.6577
5	3.9867	0.6533	0.0997	0.0163	0.6667	0.3267
6	3.5185	0.1853	0.0880	0.0046	0.5556	0.1482

In the above table, we have the values for key parameters for three different values of **number of servers (N_s)**. The data obtained shows that the **length of the system and waiting time** decreases with increase in number of servers. The **server utilization** drops from 83% to 55% as number of servers increase from four to six. The **probability of queuing (P[q])** drops from 0.65 to 0.14 as the number of servers increases. We find the cost (**ETC**) for each value of N_s as:

ETC(4) = Rs. 3459.32 /hr ETC(5) = Rs. 4259.80 /hr ETC(6) = Rs. 5092.78/hr

Using these costs, we find the cost per percent decrease in P[q]. This cost comes out to be Rs. 24.18 as we go from four servers to five servers. Whereas, as we go from five servers to six servers, this cost increases to Rs. 46.66. This clearly shows that increasing the number of servers to five has a much greater impact when compared against increasing the servers from five to six. So**, the optimal number of servers should be five** as it decreases the length of system by approximately 50% while keeping the cost operation in check.

C. Optimal Service Rate:

In this case, we consider the **M/M/1: ∞/∞/FCFS queuing model**. We take the **Expected Total Cost per unit time (ETC)** as objective function given as:

$$\mathbf{ETC(x) = EOC(x) + EWC(x)}$$

where, **ETC** = Expected total cost per unit time

EOC = Expected cost of operating the facility per unit time ($\mathbf{c_s\,\mu\,c}$)

EWC = Expected cost of waiting per unit time ($\mathbf{c_w\,Ls}$)

We are given the cost of per unit service per unit time ($\mathbf{c_s}$) as Rs 70 per hour. The waiting cost of per unit time per waiting customer ($\mathbf{c_w}$) as Rs 15 per customer. This implies

$$\mathbf{ETC(x) = c_s\,\mu + c_w\,(\lambda/(\mu - \lambda))}$$

We need to **maximize ETC** i.e. $\frac{d(ETC)}{d\mu} = 0$. This implies

$$c_s - c_w \frac{\lambda}{(\mu - \lambda)^2} = 0$$

This gives **optimal service rate (μ^*)** as

$$\mu^* = \lambda + \sqrt{\frac{C_w\,\lambda}{C_s}}$$

Using the above equation, we get **optimal service rate (μ^*)**

$$\mathbf{\mu^* = 42.93\ customers/hr}$$

Matlab Result for optimal Service Rate:

```
>> [Ls, Lq, Ws, Wq, busy_server, Prob_queuing] = Queuing_MM1(40, 42.93)
Ls =
   13.6519
Lq =
   12.7201
Ws =
    0.3413
Wq =
    0.3180
busy_server =
    0.9317
Prob_queuing =
    0.9317
```

This gives minimum **ETC = 70*42.93 + 15*13.65**

= Rs. 3210/hr

CONCLUSION

In this project, we highlighted the **problems of customer queuing** for Bank Services. The customer queuing has a huge impact on the customer experience. It can lead to **balking, reneging and jockeying** on the part of the customer. We have tried to provide a solution to this problem by completing three objectives:

- To study the impact of number of queues on the various parameters of queuing model and customer experience.
- To study the impact and find the optimal number of servers based on the cost model.
- To find the optimal service rate at which the service should be provided to improve the efficiency.

We found that in a multiple server model, **one queue** decreases the expected waiting time in the queue (Wq) as compared to the same given by more than one queue. We also studied the impact of **number of servers** on the various parameters of queuing model and found the **optimal number of servers (N_s)** using **the cost model and cost per percent decrease in probability of queuing (P[q])** for a particular queuing situation. Last but not the least, we found the **optimal service rate** at which service should be provided for a **single server model to improve the efficiency**.

We demonstrated the results for all the objectives through **MATLAB Code**. We made functions for **M/M/1: ∞/∞/FCFS queuing model** and **M/M/c: ∞/∞/FCFS queuing model**. The function for M/M/1: ∞/∞/FCFS takes arrival rate (**λ**) and service rate (**μ**) **as input** and gives **key parameters: [Ls, Lq, Ws, Wq, $\bar{c}$, $P[q]$] as output**. The function for **M/M/c: ∞/∞/FCFS** takes arrival rate (**λ**), service rate (**μ**) and number of servers (**c**) **as input** and gives **key parameters: [Ls, Lq, Ws, Wq, $\bar{c}$, $P[q]$] as output**.

REFERENCES

- https://www.investopedia.com/terms/q/queuing-theory.asp
- https://queue-it.com/blog/queuing-theory/#:~:text=Queuing%20theory%20(or%20queueing%20theory,customer%2C%20job%2C%20or%20request
- https://fraser.stlouisfed.org/files/docs/publications/frbatlreview/pages/66546_1980-1984.pdf
- https://en.wikipedia.org/wiki/Queueing_theory
- https://www.statisticshowto.com/queuing-theory/
- https://www.whitman.edu/Documents/Academics/Mathematics/berryrm.pdf
- **Operations Research: An Introduction Tenth Edition by Hamdy A. Taha.**