QUEUING THEORY

SUBMITTED BY

C. SUCHITRA

Reg No. 170021032403

SEETA.K.J

Reg No. 170021032427

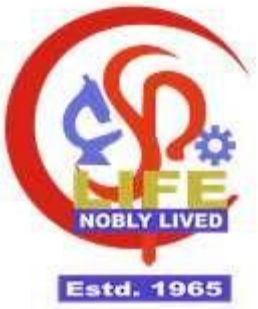
MELVIN FRANCIS

Reg No. 170021032421

IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE
BACHELOR DEGREE OF SCIENCE IN MATHEMATICS

2017-2020

ST. PAUL'S COLLEGE, KALAMASSERY
(AFFILIATED TO M.G.UNIVERSITY, KOTTAYAM)



CERTIFICATE

This is to certify that the project report title "QUEUING THEORY" submitted by C.SUCHITRA(Reg No.170021032403), SEETA.K.J(Reg No.170021032427), MELVIN FRANCIS(Reg No.170021032421)towards partial fulfilment of the requirements for the award of degree of Bachelor of Science in Mathematics is a bonfide work carried out by them during the academic year 2017-2020

Project Supervisor

Mr. ARAVIND KRISHNAN.P

Guest Faculty

Department of Mathematics

Head of the Department

Dr.SAVITA. K.S

Assistant Professor

Department of Mathematics

DECLARATION

We, C.SUCHITRA, SEETA.K.J, MELVIN FRANCIS hereby declare that this project entitled "QUEUING THEORY" is an original work done by us under the supervision and guidance of Mr. ARAVINDKRISHNAN P , Guest Faculty, Department of Mathematics, St. Paul's college Kalamassery in partial fulfilment for the award of The Degree of Bachelor of Science in Mathematics under Mahatma Gandhi University. I further declare that this project is not partly or wholly submitted for any other purpose and the data included in the project is collected from various sources and are true to the best of my knowledge.

C.SUCHITRA

Place: KALAMASSERY

SEETA.K.J

MELVIN FRANCIS

ACKNOWLEDGEMENT

We express our heartfelt gratitude to our project supervisor Mr. ARAVIND KRISHNAN P, Guest Faculty, Department of Mathematics, for providing us necessary stimulus for the preparation of this project.

We would like to acknowledge our deep sense of gratitude to Dr.SAVITHA. K.S, Head of the Department of Mathematics and all other teachers of the department and classmates for their help at all stages.

We also express our sincere gratitude to Ms.VALENTINE D'CRUZ, Principal of St. Paul's College, Kalamassery for the support and inspiration rendered to us in this project.

CONTENTS

CHAPTER-4

Single Server Models

- (M/M/1):(GD/ ∞/∞)
- Waiting Time Distribution for M/M/1):(FCFS/∞/∞)
- (M/M/1):(GD/N/∞)

CHAPTER-5

Multiple Server Models

- (M/M/c) : (GD/∞ /∞)
- (M/M/c):(GD/N/∞)
- (M/M/∞):(GD/ ∞/∞)-Self ServiceModel
- Machine Servicing Model-(M/M/R):(GD/K/K), R <K

Application Of Queuing Theory

Limitation Of Queuing Theory

Conclusion

Reference

INTRODUCTION

Modern information technologies require innovations that are based on Modeling, analyzing, designing and finally implementing new systems. The whole developing process assumes a well organized team work of experts including engineers, computer scientists, mathematicians, physicist just to mention some of them. Queuing theory is one of the most commonly used mathematical tools for the performance evaluation of complex systems.

Queuing theory is the mathematical study of queues or waiting in lines. Queue contains customers or items such as people or information. Queue forms where there are limited resources for providing a service.

Waiting for service is a part of our life. We wait to eat in restaurant, we queue up at checkout counter in the grocery store, we line for service in post office and so on. And the waiting phenomenon is not an experience limited to human beings only. Jobs waiting to be proceeded on a machine, plane circle in the stack before given permission to land on an airport and car stops at traffic lights… Waiting cannot be eliminated completely without incurring inordinate expense and the goal is to reduce its adverse impact to 'tolerable levels'.

The queuing theory owes its developments to A.K Erlang. He, in 1903, took up the problem on congestion of telephone traffic. A.K Erlang directed his first efforts at finding the delay for one operator and later on the results were extended to find the delay for several operators.

A basic queuing system consist of an arrival process, howcustomers arrive at the queue, how many costumers are present in total, the queue itself, the service process for attending to those customers and departure from the system. Mathematical queuing model are often used in software and business to determine the best way of using limited resources.

The study of queue deals with quantifying the phenomenon of waiting in lines using representative measures of performance, such as average queue length, average waiting time in queue and average facility utilization.

Queuing theory is concerned with the statistical description of the behaviors of queues. The queuing system can be described by the input (or arrival pattern), the service mechanism (or service pattern), the queue discipline and customers behavior. In a specified queuing system, the problem is to determine the probability distribution of queue length, probability distribution of waiting time of customers and the busy period distribution. A queuing model is specified completely by the following six main characteristics:

1) Input or arrival(inter-arrival) distribution
2) Output or departure(service) distribution
3) Service channels
4) Service discipline
5) Maximum number of customers allowed in the system
6) Calling source or population

CHAPTER – 1

ELEMENTS OF A QUEUING MODEL

CUSTOMERS AND SERVER

The principle actors in queuing situations are the customers and the server. Customers are generated from a source. On arrival at the facility, they can start service immediately or wait in queue if the facility is busy. When a facility completes a service, it automatically pulls a waiting customer, if any, from the queue. If the queue is empty, the facility becomes idle until a new customer arrives.

From the viewpoint of analyzing queues, the arrival process is represented by the **interarrival time** between successive customers, and the service is described by the **service time** per customer. Generally, the interarrival and the service can be probabilistic as in case of a Post Office or deterministic as in the arrival of applications for job interviews.

QUEUE SIZE

Queue size plays a role in the analysis of queues, and it may have a finite size as in the buffer area between two successive machines, or it may be infinite as in mail order facility.

QUEUE DISCIPLINE

Queue discipline, which represents the order in which customers are selected from a queue, is an important factor in the analysis of queuing models. The most common discipline is [First Come First Served-FCFS] other disciplines includes [Last Come First Served-LCFS] and [Service In Random Order-SIRO].Customers are also selected from the queue based on some order of priority.

FINITE SOURCE AND INFINITE SOURCE

The source from which customers are generated may be finite or infinite. A finite source limits the customers arriving for services. An infinite source is forever abandoned.

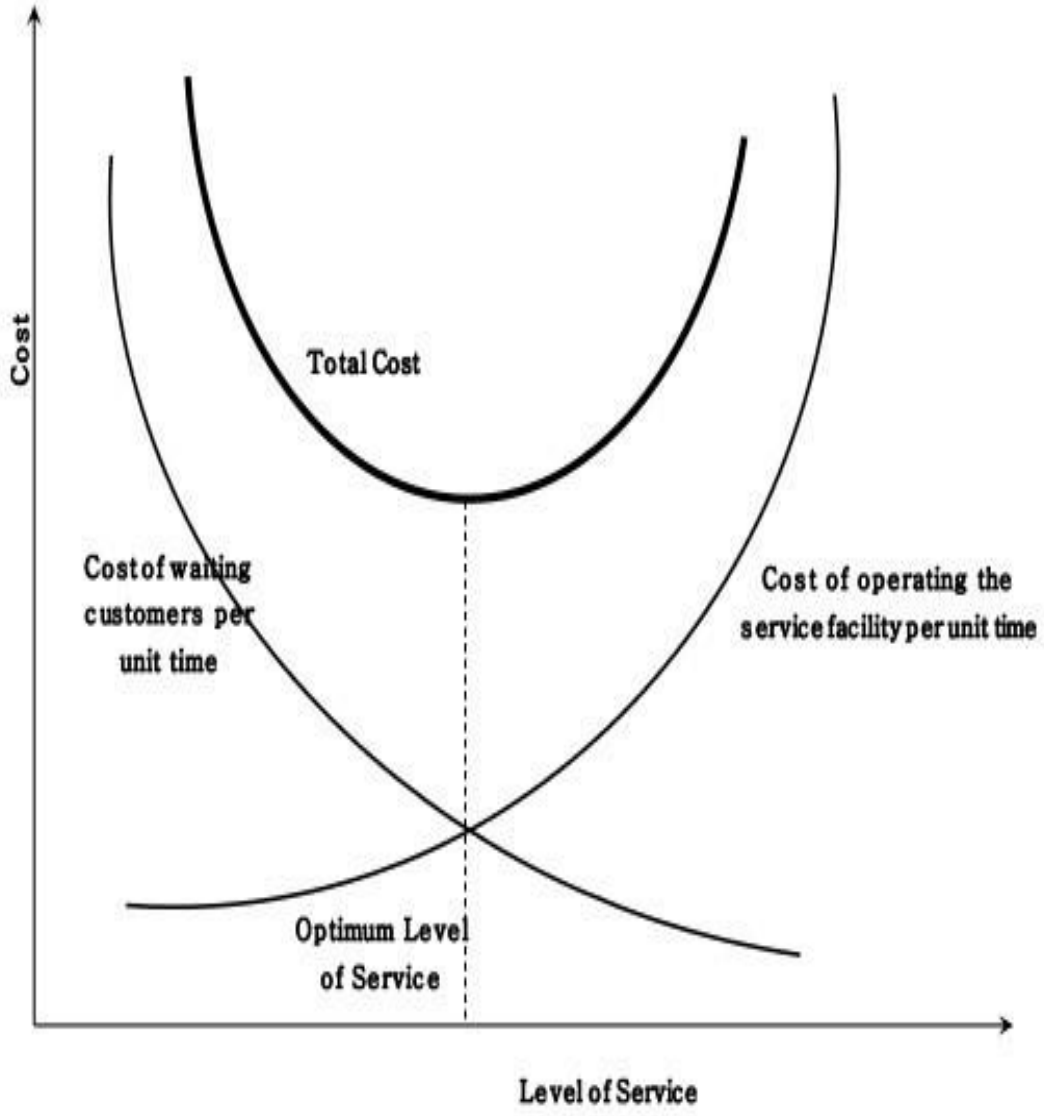
Varying the elements of a queuing situation give rise to a variety of queuing models.

Queuing theory can be used to determine the level of service that balances the following two conflicting costs.

1) Cost of offering the service
2) Cost incurred due to the delay in offering the service.

The first cost is associated with the service facilities and their operation. And the second represent the cost of customer's waiting time. We know that an increase in the existing service facilities would reduce the customer's waiting time. That is an increase (or decrease) in the level of service increases (or decreases) the cost of operating service facilities and decreases (or increases) the cost of waiting. The optimum service level is one that minimizes the sum of two costs.

The following figure illustrates both types of cost as a function of level of service.


Cost
Total Cost
Cost of waiting customers per unit time
Cost of operating the service facility per unit time
Optimum Level of Service
Level of Service

CHAPTER 2

PROBABILITY DISTRIBUTION IN QUEUING SYSTEMS

ROLE OF EXPONENTIALDISTRIBUTION

In most queuing situations, the arrival of customers occurs in a totally random fashion. Randomness here means that the occurrence of an event (e.g., arrival of a customer or completion of a service) is not influenced by the length of time that has elapsed since the occurrence of the last event.

Random interarrival and service times are described quantitatively in queuing models by the **exponential distribution**, which is defined as

$$f(t)= \lambda e^{-\lambda t} , t>0$$

For the exponential distribution,

$$E(t)=\frac{1}{\lambda}$$

$$P(t\leq T)=\int_0^T \lambda e^{-\lambda t} dt$$

$$=1-e^{-\lambda T}$$

The definition of E{t} shows that λ is the rate per unit time at which events (arrivals or departures) are generated. The fact that the exponential distribution is completely random is illustrated by the following example. If the time now is 8:20 A.M. and the last arrival has occurred at 8:02 AM, the probability that the next arrival will occur by 8:29 is a function of the interval

from 8:20 to 8:29 only, and is totally independent of the length of time that has elapsed since the occurrence of the last event (8:02 to 8:20). This result is referred to as the **forgetfulness or lack of memory** of the exponential.

Let the exponential distribution, f(t), represent the time, t , between successive events. If S is the interval since the occurrence of the last event, then the forgetfulness property implies that

$$P(t>T+S \mid t>S)=P(t>T)$$

To prove this result, we note that for the exponential with mean $1/\lambda$,

$$P(t>Y)=1-P(t<Y)=e^{-\lambda Y}$$

Thus,

$$P(t>T+S \mid t>S) = \frac{P(t>T+S, t>S)}{P(t>S)} = \frac{P(t>T+S)}{P(t>S)}$$

$$= \frac{e^{-\lambda(T+S)}}{e^{-\lambda S}} = e^{-\lambda T} = P(t>T)$$

PURE BIRTH AND DEATH MODELS (RELATIONSHIP BETWEEN THE EXPONENTIAL AND POISSON DISTRIBUTIONS)

This section presents two queuing situations: the pure birth model in which arrivals only are allowed, and the pure death model in which departures only can take place. An example of the pure birth model is the creation of birth certificates for newly born babies. The pure death model may be demonstrated by the random withdrawal of a stocked item in a store.

The exponential distribution is used to describe the interarrival time in the pure birth model and the interdeparture time in the pure death model.

A by-product of the development of the two models is to show the close relationship between the exponential and the Poisson distributions, in the sense that one distribution automatically defines the other

It is assumed that customers joining the queuing system arrived in random manner and follow a Poisson distribution and the inter arrival time follows exponential distribution. In most of the cases, service time is also assumed to be exponentially distributed. It implies that the probability of service completion in any short time period is constant. And it is independent of length of time that the service has been in progress.

The probability distributions in queuing systems are based on the following axioms:

A*1*) The number of arrivals on non overlapping intervals are statistically independent.

A*2*) The probability of more than one arrival between time t and$(t + \Delta t)$is $O(\Delta t)$ where $O(\Delta t)$is negligible.

That is the probability of two or more arrivals in small time intervalΔt is negligible. Thus$P_0(\Delta t) + P_1(\Delta t) + O(\Delta t) = 1$

A*3*) The probability that an arrival occur between t and$(t + \Delta t)$is

$$P_1(\Delta t) = \lambda \Delta t + O(\Delta t)$$

Where λ is a constant independent of the total number of arrivals upto time t, Δt is time interval and

$$\lim_{\Delta t \to 0} \frac{O(\Delta t)}{\Delta t} = 0$$

PURE BIRTH MODEL

The model in which only arrivals are counted and no departure takes place are called pure birth model.

Define,

$P_o(t)$= Probability of no arrivals during a period of

time t

Given that the interarrival time is exponential and that the arrival rate is λ customers per unit time, then

$p_o(t)$= P{interarrival time $\geq$ t}

=1- P{interarrival time $\leq$ t}

$= 1 - (1 - e^{-\lambda t})$

$= e^{-\lambda t}$

For a sufficiently a small time interval h > 0, we have

$p_o(h) = e^{-\lambda h} = 1 - \lambda h + \frac{(\lambda h)^2}{2!} - \ldots$

$= 1 - \lambda h + 0(h^2)$

The exponential distribution is based on the assumption that during h > 0, at most one event (arrival) can occur. Thus, as $h \rightarrow 0$,

$p_1(h) = 1 - p_o(h) \approx \lambda h$

This result shows that the probability of an arrival during h is directly proportional to h, with the arrival rate, λ, being the constant of proportionality.

To derive the distribution of the number of arrivals during a period t when the interarrival time is exponential with mean $\frac{1}{\lambda}$, define

$p_n(t)$ = Probability of n arrivals during t

For a sufficiently small h > 0,

$p_n(t + h) \approx p_n(t)(1 - \lambda h) + p_{n-1}(t)\lambda h, n > 0$

$p_o(t + h) \approx p_o(t)(1 - \lambda h)$, n = 0

In the first equation, n arrivals will be realized during t + h if there are n arrivals during t and no arrivals during h, or n-1 arrivals during t and one arrival during h. All other combinations are not allowed because, according to the exponential distribution, at most one arrival can occur during a very small period h. The product law of probability is applicable to the right-hand side of the equation because arrivals are independent. For the second equation, zero arrivals during t + h can occur only if no arrivals occur during t and h.

Rearranging the terms and taking the limits as h →0, we get

$p'_n(t) = \lim_{h\to 0}\frac{P_n(t+h)-P_n(t)}{h} = -\lambda p_n(t) + \lambda p_{n-1}(t), n > 0$

$p'_o(t) = \lim_{h\to 0}\frac{P_0(t-h)-P_0(t)}{h} = -\lambda p_o(t), n=0$

where $p'_n(t)$ is the first derivative of $p_n(t)$ with respect to t.

The solution of the preceding difference-differential equations yields

$p_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$, n = 0,1,2, ...

This is a Poisson distribution with mean E(n | t) = λt arrivals during t

The preceding result shows that if the time between arrivals is exponential with mean $\frac{1}{\lambda}$then the number of arrivals during a specific period t is poisson with mean λt.The converse is true also.

Pure Death Model

In the pure death model, the system starts with N customers at time 0 and non a new arrivals are allowed. Departures occur at the rate μ customers per unit time. To develop the difference-differential equations for the probability $p_n(t)$ of n customers remaining after t time units, we follow the arguments used with the pure birth model. Thus,

$$P_N(t + \Delta t) = P_N(t)(1 - \mu\Delta t)$$

$$P_n(t + \Delta t) = P_n(t)(1 - \mu\Delta t) + P_{n+1}(t)\mu\Delta t, \qquad 0 < n < N$$

$$P_0(t + \Delta t) = P_0(t)(1) + P_1(t)\mu\Delta t$$

As Δt→ 0, we get

$$P'_N(t) = -\mu P_N(t)$$

$$P'_n(t) = -\mu P_n(t) + \mu P_{n+1}(t), 0 < n < N$$

$$P'_0(t) = \mu P_1(t)$$

The solution for these equations yields the following Truncated Poisson distribution:

$$P_n(t) = \frac{(\mu t)^{N-n} e^{-\mu t}}{(N-n)!}, n = 1,2,3 \ldots\ldots N$$

$$P_0(t) = 1 - \sum_{n=1}^{N} P_n(t)$$

CHAPTER 3

GENERALIZED POISSON QUEUING MODEL

This section develops a general queuing model that combines both arrivals and departures based on the Poisson assumptions-that is, the interarrival and the service times follow the exponential distribution.

The development of the generalized model is based on the long-run or **steady-state** behavior of the queuing situation, which is achieved after the system has been in operation for a sufficiently long time. This type of analysis contrasts with the **transient** (or warmup) behavior that prevails during the early operation of the system.

The generalized model assumes that both the arrival and departure rates are **state dependent**, meaning that they depend on the number of customers in the service facility. For example, at a highway toll booth, attendants tend to speed up toll collection during rush hours. Another example occurs in a shop with a given number of machines where the rate of breakdown decreases as the number of broken machines increases(because only working machines are capable of generating new breakdowns).

Define,

n =Number of customers in the system

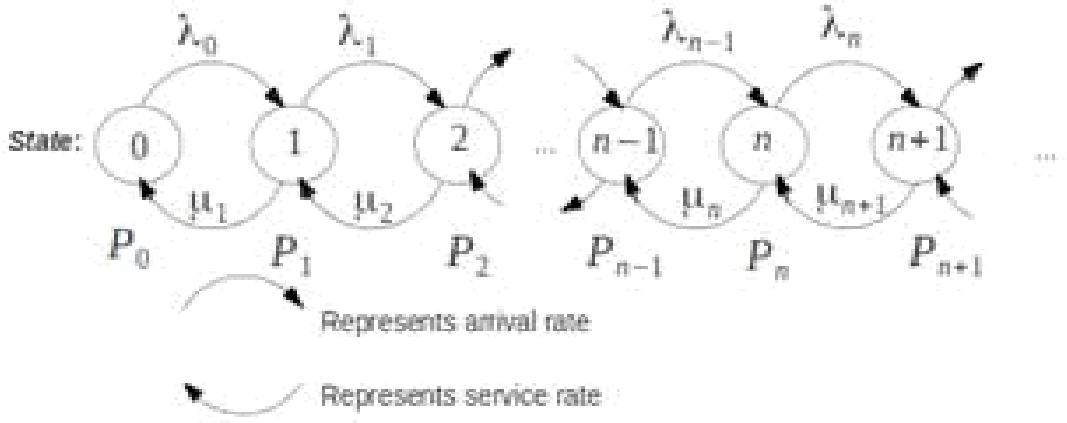
λ_n =Arrival rate given n customers in the system

μ_n =Departure rate given n customers in the system

P_n =Steady-state probability of n customers in the system

The generalized model derives p_n, as a function of λ_n and μ_n.These probabilities are then used to determine the system's measures of performance, such as the average queue length, the average waiting time, and the average utilization of the facility.

The probabilities p_n are determined by using the **transition-rate diagram**



The queuing system is in state n when the number of customers in the system is n.The probability of more than one event occurring during a small interval h tends to zero as $h \rightarrow 0$. This means that for $n > 0$, state n can change only to two possible states: n-1 when a departure occurs at the rate μ_n and n+1 when an arrival occurs at the rate λ_n. State 0 can only change to state 1 when an arrival occurs at the rate λ_o. μ_o is undefined because no departures can occur if the system is empty.

Under steady-state conditions, for n>0, the expected rates of flow into and out of state n must be equal. Based on the fact that state n can be changed to states n-1 and n + 1 only, we get

$$\begin{pmatrix} Expected\ rate\ of \\ flow\ into\ state\ n \end{pmatrix} = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}$$

Similarly,

$$\begin{pmatrix} Expected\ rate\ of \\ flow\ out\ of\ state\ n \end{pmatrix} = (\lambda_n + \mu_n)p_n$$

Equating the two rates , we get the following balance equation:

$\lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} = (\lambda_n + \mu_n)p_n$,n = 1,2,....

From figure 15.2 , the balance equation associated with n=0 , is

$$\lambda_0 P_0 = \mu_1 P_1$$

The balance equation are solved recursively in terms of P_0 as follows: For n=0 ,

we have

$$P_1 = \left(\frac{\lambda_0}{\mu_1}\right) P_0$$

Next , for n=1 , we have

$$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$$

Substituting $P_1 = \left(\frac{\lambda_0}{\mu_0}\right) P_0$ and simplifying, we get

$$P_2 = \left(\frac{\lambda_1 \lambda_0}{\mu_2 \mu_1}\right) P_0$$

In general , we can show by induction that

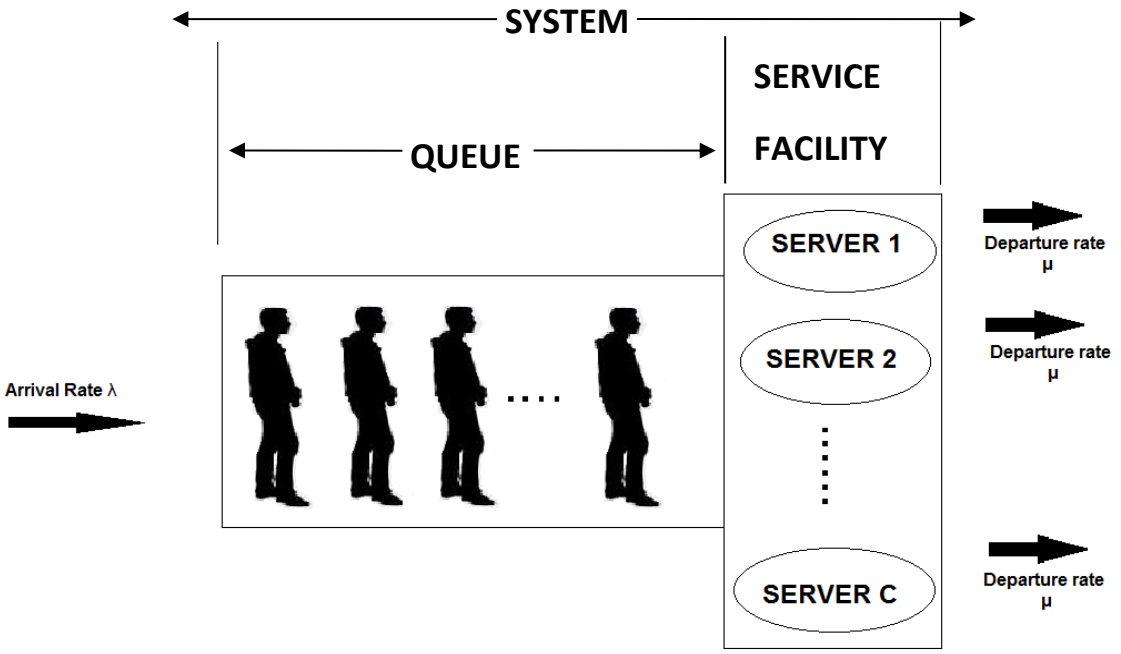
$P_n = \left(\frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n \mu_{n-1}\ldots\mu_1}\right) P_0$, n = 1,2,...

The value of P_0 is determined from the equation $\sum_{n=0}^{\infty} P_n$=1

SPECIALIZED POISSON QUEUES

The specialized Poisson queuing situation with c parallel servers. A waiting customer is selected from the queue to start service with the first available server.The arrival rate at the system is A customers per unit time. All parallel servers are identical, meaning that the service rate for any server is customers per unit time. The number of customers in the system is defined to include those in service and those waiting in queue.

The given figure shows the specialized Poisson queuing situation with c parallel servers



A convenient notation for summarizing the characteristics of the queuing situation in the given figure is given by the following format:(a/b/c):(d/e/f)

where

a = Arrivals distribution

b= Departures (service time)

distribution

c= Number of parallel servers

(=1,2....)

d = Queue discipline

e= Maximum number (finite or. infinite) allowed in the system (in-queue plus in- service)

f= Size of the calling source (finite or infinite)

The standard notation for representing the arrivals and departures distributions (symbols a and b) is

M=Markovian (or Poisson) arrivals or departures distribution

(or equivalently exponential interarrival or service time distribution)

D= Constant (deterministic) time

E_k= Erlang or gamma distribution of time (or, equivalently, the sum of independent exponential distributions)

GI = General (generic) distribution of interarrival time

G= General (generic) distribution of service time

The queue discipline notation (symbol d) includes

FCFS = First come, first served

LCFS = Last come, first served

SIRO= Service in random order

GD = General discipline (i.e., any type of discipline)

The first three Clements of the notation (a/b/c), were devised by D. G. Kendall in 1953 and are known in the literature as the Kendall notation.

Steady-State Measures of Performance

The most commonly used measures of performance in a queuing situation are

L_s= Expected number of customers in system

L_q=Expected number of customers in queue

W_s= Expected waiting time in system

W_q= Expected waiting time in queue

$\bar{C}$=Expected number of busy servers

These relationships are valid under rather general conditions. The parameter Aeff is the effective arrival rate at the system. It equals the (nominal) arrival rateA when all arriving customers can join the system. Otherwise, if some customers cannot join because the system is full (e.g., a parking lot), then $\lambda_{eff} < \lambda$. direct relationship also exists betweenW_s and W_q. By definition,

$$\begin{pmatrix} Expected\ waiting \\ time\ in\ system \end{pmatrix} = \begin{pmatrix} Expected\ waiting \\ time\ in\ queue \end{pmatrix} + \begin{pmatrix} Expected\ service \\ time \end{pmatrix}$$

This translates to

$$W_s = W_q + \frac{1}{\mu}$$

Next, we can relate L_s, to L_q by multiplying both sides of the last formula by λ_{eff}, which together with Little's formula gives

$$L_s = L_q + \frac{\lambda_{eff}}{\mu}$$

By definition, the difference between the average number in the system, $L_s$and the average number in the queue, L_q, must equal the average number of busy servers,$\bar{C}$. We thus have

$$\bar{C} = L_s - L_q = \frac{\lambda_{eff}}{\mu}$$

It follows that

$$\binom{Facility}{utilization} = \frac{\bar{c}}{c}$$

Chapter 4

SINGLE SERVER MODELS

This section presents two models for the single server case. The first model sets no limit on the maximum number in the system and the second model assumes a finite system limit. Both models assume an infinite-capacity source. Arrivals occur at the rate customers per unit time and the service rate is customers per unit time.

(M/M/1):(GD/ ∞/∞). Using the notation of the generalized model,we have

$$\lambda_n = \lambda$$

$$\mu_n = \mu \quad , n=0,1,2,...$$

Also, λeff = λ and λlost =0, because all arriving customers can join the system. Letting ρ= $\frac{\lambda}{\mu}$,the expression for Pn in the generalized model then reduces to

$$Pn = \rho^n P_0, \ n=0,1,2,...$$

To determine the value of P0 ,we use the identity

$$P_0(1+\rho+\rho^2+...) = 1$$

Assuming ρ<1, the geometric series will have the finite sum ($\frac{1}{1-\rho}$)thus

$$P_0 = 1-\rho, \text{ provided } \rho<1.$$

The general formula for Pn is thus given by the following geometric distribution

$$Pn = (1-\rho)\rho^n,\ n=1,2,\ldots(\rho<1)$$

The mathematical derivation of Pn imposes the condition $\rho<1$,or $\lambda<\mu$. If $\lambda\geq\mu$,the geometric series will not convergeand the steady -state probabilities will not exist. This result makes intuitive sense , because unless the service rate is larger than the arrival rate,queue length will continually increase and no steady state can be reached.

The measure of performance Lq can be derived in the following manner:

$$LS = \sum_{n=0}^{\infty} npn = \sum_{n=0}^{\infty}(1-\rho)\,\rho^n$$

$$= (1-\rho)\rho\frac{d}{dp}\sum_{n=0}^{\infty} \rho^n$$

$$= (1-\rho)\rho\frac{d}{dp}\left(\frac{1}{1-p}\right) = \frac{P}{1-\rho}$$

Because $\lambda eff = \lambda$ for the present situation , the remaining measures of performance are computed using the relationships .

Thus,

$$WS = \frac{Ls}{\lambda} = \frac{1}{\mu(1-p)} = \frac{1}{\mu-\lambda}$$

$$Wq = WS - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$$

$$Lq = \lambda Wq = \frac{\rho^2}{1-\rho}$$

$$\bar{c} = LS - Lq = \rho$$

Waiting Time Distribution for (M/M/1):(FCFS/∞/∞)

Although the average waiting time is independent of the queue discipline, its probability density function is not.

Let τ be the amount of time a person just arriving must be in the system(ie,until the service is completed). Based on the FCFS discipline, if there are n customers in the system ahead of an arriving customer,then

$\tau = t'1 + t2 + \ldots + t_{n+1}$

Where t'1 is the time needed for the customer currently in service to complete service and t2,t3,.....tn are the service times for the n-1 customers in the queue . The time tn+1 represents the service times for the arriving customer.

Define ω(τ| n+1) as the conditional density function of τ given n customers in the system ahead of the arriving customer. Because the distribution of the service time is exponential,the forgetfulness property tells us that t'1 is also exponential with the same distribution. Thus,τ is the sum of n+1 identically distributed and independent exponential random variables. From probability theory, ω(τ|n+1) follows a gamma distribution with parameters μ and n+1. We thus have

$$\omega(\tau) = \sum_{n=0}^{\infty} \omega(\tau|n+1)Pn$$

$$= \sum_{n=0}^{\infty} \frac{\mu}{n!} [\ (\mu\tau)^n\ e^{-\mu\tau}\ ](1-\rho)\rho^n$$

$$= (1-\rho)\mu e^{-\mu\tau} \sum_{n=0}^{\infty} \frac{(\lambda\tau)^n}{n!}$$

$$= (1-\rho)\mu e^{-\mu\tau} e^{\lambda\tau}$$

$$= (\mu-\lambda)e^{-(\mu-\lambda)\tau}\ ,\ \tau > 0$$

Thus, ω(τ) is an exponential distribution with mean

$W_s = \frac{1}{(\mu-\lambda)}$,

(**M/M/1**): (**GD/N**/∞). This model differs from (M/M/1) ： (GD/ ∞/∞) in that there is a limit N on the number in the system (maximum queue length = N-1). Examples include manufacturing situations in which a machine may have a limited buffer area, and a one-lane drive -in window in a fast- food restaurant.When the number of customers in the system reaches N, no more arrivals are allowed. Thus, we have

$$\lambda n = \begin{cases} \lambda, n = 0,1, \dots N^{-1} \\ 0, n = N, N+1 \end{cases}$$

μn = μ, n = 0,1,...

Using ρ = $\frac{\lambda}{\mu}$, the generalized model

$$Pn = \begin{cases} p^n p_\circ & n \leq N \\ 0, & n > N \end{cases}$$

The value of P0 is determined from the equation

$\sum_{n=0}^{\infty} P$n =1 ,which yields

P_0(1+ p + p^2+ ... + p^N) = 1

Or

$$P_0 = \begin{cases} \frac{(1-\rho)}{1-\rho^{N+1}}, & \rho \neq 1 \\ \frac{1}{N+1}, & \rho = 1 \end{cases}$$

Thus,

$$P_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{N+1}}, & \rho \neq 1 \\ \frac{1}{N+1}, & \rho = 1 \end{cases}$$,n = 0,1,...N

The value of $\rho = \frac{\lambda}{\mu}$ need not be less than 1 in this model, because arrivals at the system are controlled by the system limit N. This means that λ_{eff}, rather than λ,is the rate that matters in this case. Because customers will be lost when there are N in the system. Then,

$$\lambda_{lost} = \lambda P_N$$

$$\lambda_{eff} = \lambda - \lambda_{lost} = \lambda(1 - P_N)$$

In this case ,$\lambda_{eff} < \mu$.

The expected number of customers in the system is computed as

$L_s = \sum_{n=1}^{N} npn$

$$= \left(\frac{1-\rho}{1-\rho^{N+1}}\right)\sum_{n=0}^{N} n\rho^n$$

$$= \left(\frac{1-\rho}{1-\rho^{N+1}}\right)\rho\frac{d}{d\rho}\sum_{n=0}^{N}\rho^n$$

$$= \frac{(1-\rho)\rho}{1-\rho^{N+1}}\frac{d}{d\rho}\left(\frac{1-\rho^{N+1}}{1-\rho}\right)$$

$$= \frac{\rho[1-(N+1)\rho^N+N\rho^{N+1}]}{(1-\rho)(1-\rho^{N+1})}, \rho \neq 1$$

CHAPTER 5

MULTIPLE SERVER MODELS

This section considers three queuing models with multiple parallel servers.The first two models are the multi-server versions of the models . The third model treats the self-service case, which is equivalent to having an infinite number of parallel servers.

(M/M/c) : (GD/∞ /∞). In this model, there are c parallel servers. The arrival rate is λ and the service rate per server is μ . Because there is no limit on the number in the system, λeff = λ.

The effect of using parallel servers is a proportionate increase in the facility service rate. In terms of the generalized model ,λ_n andμ_n are thus defined as

$$\lambda_n = \lambda , \ n \geq 0$$

$$\mu_n = \begin{cases} n\mu, & n < c \\ c\mu, & n \geq c \end{cases}$$

Thus ,

$$p_0 = \begin{cases} \frac{\lambda^n}{\mu(2\mu)(3\mu)\dots(n\mu)} p_0 = \frac{dy^n}{n!\mu^n} p_0 = \frac{\rho^n}{n!} p_0 , n < c \\ \frac{\lambda^n}{(\prod_{i=1}^{c} i\mu)(c\mu)^{n-c}} p_0 = \frac{\lambda^n}{c!c^{n-c}\mu^n} p_0 = \frac{\rho^n}{c!c^{n-c}} p_0 \ n \geq c \end{cases}$$

Letting ρ = $\frac{\lambda}{\mu}$,and assuming $\frac{\rho}{c}$<1 ,the value of p_0 is determined from $\sum_{n=0}^{\infty} p_n$ =1,

which gives,

$$P_0 = \left\{\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!}\sum_{n=c}^{\infty} \frac{\rho^{n-c}}{c}\right\}.^{-1}$$

$$= \left\{\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!}\left(\frac{1}{1-\frac{\rho}{c}}\right)\right\}.^{-1} \quad ,\frac{\rho}{c} < 1$$

The expression for L_q can be determined as follows :

$$L_q = \sum_{n=c}^{\infty}(n-c)\, p_n$$

$$= \sum_{k=0}^{\infty} k\, p_{k+c}$$

$$= \sum_{k=0}^{\infty} k \frac{\rho^{k+c}}{c^k c!} p_0$$

$$= \frac{\rho^{c+1}}{c!c} p_0 \sum_{k=0}^{\infty} k\left(\frac{\rho}{c}\right)^{k-1}$$

$$= \frac{\rho^{c+1}}{c!c} p_0 \frac{d}{d\left(\frac{\rho}{c}\right)} \sum_{k=0}^{\infty}\left(\frac{\rho}{c}\right)^k$$

$$= \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} p_0$$

Because $\lambda_{eff} = \lambda$, $L_s = L_q + \rho$. The values of W_s and W_q can be determined by dividing L_s and L_q by λ.

(M/M/c):(GD/N/∞),c≤N.This model differs from that of the (M/M/c):(GD/∞ /∞) model in that the system limit is finite and equal to N. This means that the maximum queue size is N-c . The arrival and service rates are λ and μ. The effective arrival rate λ_{eff} is less than λbecause of the system limit N.

In terms of the generalized model ,λ_n and μ_n for the current model are defined as

$$\lambda_n = \begin{cases} \lambda, & 0 \le n \le N \\ 0, & n > N \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 0 \le n \le c \\ c\mu, & c \le n \le N \end{cases}$$

Substituting λ_n and μ_n in the general expression and noting that $\rho = \frac{\lambda}{\mu'}$ we get

$$P_n = \begin{cases} \frac{\rho^n}{n!} p_0, & 0 \le n < c \\ \frac{\rho^n}{c! c^{n-c}} p_0, & c \le n \le N \end{cases}$$

Where

$$P_0 = \begin{cases} \left(\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c (1 - (\frac{\rho}{c})^{N-c+1})}{dx} \right)^{-1}, \frac{\rho}{c} \neq 1 \\ \left(\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} (N - c + 1) \right)^{-1}, \quad \frac{\rho}{c} = 1 \end{cases}$$

Next, we compute L_q for the case where $\frac{\rho}{c} \neq 1$ as

$L_q = \sum_{n=c}^{N} (n - c)\, p_n$

$= \sum_{j=0}^{N-c} j\, p_{j+c}$

$= \frac{\rho^c \rho}{c! c} p_0 \sum_{j=0}^{N-c} j (\frac{\rho}{c})^{j-1}$

$= \frac{\rho^{c+1}}{c c!} p_0 \frac{d}{d(\frac{\rho}{c})} \sum_{j=0}^{N-c} (\frac{\rho}{c})^j$

$= \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} \left\{ 1 - (\frac{\rho}{c})^{N-c+1} - (N - c + 1)(1 - \frac{\rho}{c})(\frac{\rho}{c})^{N-c} \right\} p_0$

It can be shown that for $\frac{\rho}{c} = 1$, L_q reduces to

$$L_q = \frac{\rho^c (N-c)(N-c+1)}{2c!} p_0 \,, \frac{\rho}{c} = 1$$

To determine W_q and hence W_s and L_s ,we compute the value of λ_{eff} as

$$\lambda_{lost} = \lambda P_N$$

$$\lambda_{eff} = \lambda - \lambda_{lost} = (1 - P_N)\lambda$$

(M/M/∞):(GD/ ∞/∞)- Self-Service Model.

The number of servers in unlimited because the customer is also the server. A typical example is taking the written part of a drivers license test. Self - service gas stations and 24-hour ATM banks do not fall under this model's description because the servers in these cases are actually the gas pumps and the ATM machines. The model assumes steady arrival and service rates, λ and μ ,respectively.

In terms of the generalized model, we have

$$\lambda_n = \lambda, \quad n = 0,1,2,...$$

$$\mu_n = n\mu, \quad n = 0,1,2,...$$

Thus,

$$P_n = \frac{\lambda^n}{n!\mu^n} p_0 = \frac{\rho^n}{n!} p_0, \quad n = 0,1,2,...$$

Because $\sum_{n=0}^{\infty} P_n = 1$, it follows that

$$P_0 = \frac{1}{1+\rho+\frac{\rho^2}{2!}+\cdots} = \frac{1}{e^{\rho}} = e^{-\rho}$$

As a result,

$$P_n = \frac{e^{-\rho}\rho^n}{n!}, \quad n = 0,1,2,...$$

Which is Poisson with mean $L_s = \rho$. As should be expected, L_q and W_q are zero because it is a self-service model.

Machine Servicing Model-(M/M/R):(GD/K/K), R <K

The setting for this model is a shop with K machines. When a machine breaks down one of R available repairpersons is called upon to do the repair. The rate of breakdown per machine is λ breakdowns per unit time, and a repairperson will service broken machines at the rate of μ machines per unit time. All breakdowns and services are assumed to follow the Poisson distribution.

This model differs from all the preceding ones because it has a finite calling source. We can see this point by realizing that when all the machines in the shop are broken, no more calls for service can be generated. In essence, only machines in working order can break down and hence can generate calls for service.

Given the rate of breakdown per machine,λ the rate of breakdown for the entire shop is proportional to the number of machines that are in working order. In terms of the queuing model, having n machines in the system signifies that n machines are broken. Thus, the rate of breakdown for the entire shop is

$$\lambda_n = (K - n)\lambda, 0 \leq \text{n} \leq \text{k}$$

In the terms of the generalized model, we have

$$\lambda_n = \begin{cases} (K - n)\lambda, & 0 \leq n \leq K \\ 0, & n \geq k \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 0 \leq n \leq R \\ R\mu, & R \leq n \leq K \end{cases}$$

From the generalized model, we can then obtain,

$$P_n = \begin{cases} C_n^K \rho^n p_0, & 0 \le n \le R \\ C_n^K \dfrac{n!\,\rho^n}{R!\,R^{n-R}} p_0, & R \le n \le K \end{cases}$$

$$P_0 = \left(\sum_{n=0}^{R} C_n^K \rho^n + \sum_{n=R+1}^{K} C_n^K \frac{n!\,\rho^n}{R!\,R^{n-R}} \right)^{-1}$$

There is no closed form expression for L_s , and hence it must be computed using the following basic definition:

$$L_s = \sum_{n=0}^{K} n p_n$$

The value of λ_{eff} is computed as

$$\lambda_{eff} = \mathrm{E}\{\lambda(K - n)\} = \lambda(\mathrm{K} - L_s)$$

APPLICATIONS OF QUEUING THEORY

Queuing theory can be applied to a wide variety of operational situations. In particular, the technique of queuing theory is applied for solutions of large number of problems such as

1. Scheduling of air craft at landing and takeoff from busy airports.
2. Scheduling of issue and return of tools by workmen from tool cribs in factories.
3. Scheduling and distribution of scarce war material.
4. Scheduling of works and jobs in production control.
5. Minimization of congestion due to traffic delay at toll booths.
6. Scheduling of components to assembling lines.
7. Scheduling and routing of salesman.

LIMITATIONS OF QUEIUNG THEORY

1) Most of the queuing models are complex and cannot be easily understood. There is always the element of uncertainty in all queuing situations. There, the probability distribution to be applied for arrivals or services may not be clearly known.
2) Queue discipline also imposes some limitations. We assume first come first served service discipline. If this assumption is not true, Queuing analysis becomes more complex.
3) In multichannel queuing, several times the departure from one queue forms the departure for another. This makes the analysis more complex.

CONCLUSION

Our Project is a sincere attempt to study the topic **Queuing Theory** and its applications in day today life. Queuing theory is a major system in our society. Every person has had to stand in line at one point in their lives. Understanding queuing theory helps compensate for these waiting periods. For electronically functioning devices installing a queuing program is the only way it can function and perfume in banks, air plane landing queues and in this vast world of electronics

Overall, queuing theory can be used to help reduce waiting times and where waiting times are inevitable. Applying the queuing theory in daily life increases the time efficiency in all aspects.

REFERENCE

- HAMDY A.TAHA, *OPERATIONS RESEARCH, AN INTRODUCTION-EIGHTH EDITION*.
  - Bose, S., *An Introduction to Queuing Systems,* Kluwer Academic Publishers, Boston, 2001.
  - Hall, R, *Queuing Methods for Service and Manufacturing,* Prentice Hall, Upper Saddle River, NJ, 1991.
  - Lipsky, L, *Applied Queuing Theory,* A Linear Algebraic Approach, Macmillan, New York, 1958.
  - Morse, P, *Queues, Inventories, and Maintenance,* Wiley, New York, 1958.
  - Parszen, E, *Stochastic Processes,* Holden-Day, San Francisco, 1962.
  - Saaty, T, *Elements of Queuing Theory with Applications,* Dover, New York, 1983.
  - Tanner, M, *Practical Queuing Analysis,* McGraw-Hill, New York, 1995.
  - Tijms, H.C, *Stochastic Models-An Algorithmic Approach,* Willey, New York, 1994.